Provable Algorithms for Reinforcement Learning: Efficiency, Scalability, and Robustness

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Laixi Shi

B.S., Electronic Engineering, Tsinghua University

Carnegie Mellon University

Pittsburgh, PA

August 2023

©Laixi Shi, 2023

All Rights Reserved

Acknowledgments

The research in this thesis is supported in part by the grants ONR N00014-19-1-2404, NSF CCF-2106778, CCF2007911, DMS-2134080, CNS-2148212, and the CAREER award ECCS-1818571; as well as CIT Dean's Fellowship, Leo Finzi Memorial Fellowship, Wei Shen and Xuehong Zhang Presidential Fellowship, and Liang Ji-Dian Graduate Fellowship at Carnegie Mellon University.

I would like to extend my deepest gratitude to everyone who plays a key role in establishing the enriching and vibrant journey towards earning my Ph.D..

First and foremost, I would like to extend my heartfelt appreciation to my extraordinary supervisor, Prof. Yuejie Chi. Yuejie is an all-powerful Doraemon for not only me, but all her students, offering invaluable advice on personal career planning, research view development, and technical guidance. I constantly learn from her pocket, a wellspring of career wisdom, deep sense of responsibility, love for students, zest for research, openness, and remarkable execution abilities. Yuejie's efforts in building an inclusive and liberating environment for her group shape my values of research and a desired community.

Yuejie has steered the enjoying journey of my Ph.D. with theoretical guarantees, as a warm and steadfast backup that has made my Ph.D. endeavor feel both accompanied and replete with endless possibilities. The sense of destiny I felt in working with Yuejie began to form even before our initial interview for the CMU Ph.D. program. As it turned out, the experience was far richer than I could have anticipated. I have immensely benefited from Yuejie's exceptional optimization abilities, where she initialized me at a promising research initial point and mentored the effective ascent direction at each key step.

I would like to express my appreciation to the other members of my thesis committee. Prof. Jiantao Jiao offered me a surge of insightful comments on the research directions. Prof. Gauri Joshi has provided constructive feedback on the thesis writing and presentation style. Prof. Quannan Qu suggested multiple promising directions for future exploration. And Prof. Matthieu Geist, apart from offering me an internship opportunity in Google, Paris, provided invaluable technical support for the works of the thesis.

In addition, I want to extend my sincere gratitude to several mentors during my internships. Robert Datashi, along with Prof. Matthieu Geist, guided me at the Google Brain Paris Team, where we linked theoretical concepts with practical algorithm designs perfectly. Dr. Pablo Samuel Castro hosted me remotely for a productive and wonderful internship in Google Brain, Mountain View, always samshowing support and encouragement for my work. Dr. Dehong Liu mentored me at Mitsubishi Electric Research Laboratories, where we built frameworks for mechanical engineering applications.

Moreover, my sincere thanks go to all the other faculty members with whom I have worked or who have inspired me. Dr. Gen Li, who deserves a special mention, directed my initial foray into reinforcement learning - the topic of this thesis - and has continued to inspire the subsequent productive advancements. Prof. Yuxin Chen and Prof. Yuting Wei offered priceless supervision and guidance for a significant portion of the works in this thesis, including but not limited to the selection of research direction, presentation, paper writing, and valuable discussions. Prof. Pei Zhang, Prof. Hae Young Noh, and Prof. Shijia Pan helped me to see the path to designing practical

prototypes for smart cities, along with the development of remarkable presentation skills. My experience collaborating with Prof. Tze Meng Low, Prof. James C. Hoe, and Prof. Swarun Kumar has been invaluable. My gratitude also extends to Prof. John Wright, who initially inspired me to join the theoretical investigation community. Thanks to the entire group of Prof. Yimin Liu and Prof. Tianyao Huang for hosting me at Tsinghua University, where I gleaned a lot from real-world applications and wrote the first paper draft during my Ph.D. journey.

I'm deeply appreciative of the interdisciplinary research environment cultivated by my fellow students. In addition to my thesis, my Ph.D. journey has been significantly enriched by discussions and collaborations with Mengdi Xu, Wenhao Ding, Rui Chen, and Yiqi Wang, which inspired many of my practical works. My joint project with Peide Huang and Jiacheng Zhu offered me another opportunity to apply theoretical analysis to empirical studies.

Beyond research, I have been fortunate to meet a myriad of friendly, generous, and supportive friends at CMU, a non-exhaustive list of whom I acknowledge here. My appreciation extends to my research family members at CMU — Yuanxin Li, Harlin Lee, Vince Monardo, Tian Tong, Boyue Li, Maxime Ferreira Da Costa, Zhize Li, Shicong Cen, Pedro Valdeira, Jiin Woo, Harry Dong, Lingjing Kong, Zixin Wen, He Wang, and Xingyu Xu. They have not only provided a wealth of research inspiration but also truly enjoyable times during group meetings and hangout events. I also wish to express my gratitude to another supportive network at CMU, including Zuxin Liu, Jielin Qiu, Haohong Lin, Zhepeng Cen, Chengyuan Zhang, Yaru Niu, Miao Li, Junting Deng, Yihang Yao, Hanjiang Hu and Yuyou Zhang. They have enriched my life immeasurably, through climbing, shared meals, skiing, and sports activities. I thank the broader community members — Shuhua Yu, Ran Xin, Yuhang Yao, Zhuoyuan Wang, Meiyi Li, Yingsi Qin, Marie T. Siew, John Shi, Srinivasa Pranav, and Weiyu Chen for their warm companionship in Porter Hall. Gratitude is also due to Yuwei Qiu, Jingxuan Hou, Xuewei Yu, and Yu Wang for making the undergraduate EE home at Webster Hall so welcoming. I thank Wenyu Xia and Yufei Ye for motivating me to join the running group, which led to my first experience of a half Marathon alongside Emma Gao and Ruihan Gao.

In addition, I want to express my gratitude to all the friends I've made outside of CMU, a select few of whom I've mentioned here. Thanks to Jolanta Mozyrska, Yifan Zhu, and Henrik Reinstädtler for the amazing time in Paris. My appreciation extends to Xingmin Wang, Ruxing Fu, Shuwen Qiu, Zihan Yang, He Peng, Zijian Yao, and Zheqing Li for the incredible experiences we shared traveling around the world. I'm eagerly anticipating the life that awaits me with Shuwen Qiu and Xinyi Huang in Caltech, believing it will be just as fulfilling. I'm grateful to Yi Guo for sending me the first set of face masks during the Covid pandemic, and to Jinda Bi, Zhaoyuan Gu, and Han Zhang for their necessary emotional support during challenging times.

Finally, a heartfelt thanks to my parents, who have always stood by me and believed in me. While much of the trustness to me either attributed to my advisor or the friends around me, that is still my fortune to have them. Their unwavering faith, open-mindedness, and positive attitude have been my guiding stars during times of difficulty and uncertainty. Their love constantly motivates me to discover and become the person I truly aspire to be.

Laixi Shi

Abstract

Reinforcement learning (RL), which strives to learn desirable sequential decisions based on trial-and-error interactions with an unknown environment, has achieved remarkable success recently in a variety of domains including games and large language model alignment. In the face of unknown environments with unprecedentedly large dimensionality, making the best use of available samples inevitably lies at the core of RL, especially in ubiquitous sample-starved scenarios such as clinical trials and autonomous driving. To understand and tackle the challenges of sample efficiency, substantial progress has been made recently by developing a finite-sample theoretical framework to analyze the algorithms of interest and design provably optimal algorithms in terms of sample efficiency. Nevertheless, existing results still fall short with regards to the statistical understanding and algorithmic optimality in a wide range of RL settings. Moreover, motivated by countless scenarios with large dimensionality or sim-to-real gaps, sample efficiency needs to be considered along with scalability and robustness — two equally important principles in RL.

This thesis breaks down the sample barriers of various RL formulations, taking additional scalability and robustness into account. Specifically, for online RL that allows for adaptive interactions with the environment, this thesis provides the first provable regret-optimal model-free RL algorithm with a small burn-in cost — an initial sampling burden needed for the algorithm to exhibit the desired performance — while maintaining its memory efficiency for scalability. For offline RL that only has access to historical datasets, this thesis proposes the first provable near-optimal model-free offline RL algorithm without the need of performing model estimation, and settles the sample complexity by establishing the minimax optimality of model-based offline RL algorithms without burn-in cost. Finally, for a robust variant of standard RL — distributionally robust RL, this thesis uncovers a surprising fact that promoting additional distributional robustness in the learned policy is neither necessarily harder nor easier than standard RL in terms of sample requirements, which depends heavily on the prescribed uncertainty set. This thesis closes by providing the first provable near-optimal algorithm for offline robust RL that can learn under simultaneous model uncertainty and limited historical datasets.

Contents

Acknowledgments iii

Abstract v

List of Figures xi

List of Tables xii

List of Algorithms xiv

Chapter 1 Introduction 1

1.1 Overview . . . 1
1.2 Online RL . . . 3
1.2.1 Regret-optimal model-free RL? A sample size barrier . . . 3
1.2.2 Our contributions to model-free online RL . . . 5
1.3 Offline RL . . . 7
1.3.1 Challenges of offline RL: distribution shift and limited data coverage . . . 7
1.3.2 Inadequacy of prior works . . . 8
1.3.3 Our contributions to model-free offline RL . . . 9
1.3.4 Our contributions to model-based offline RL . . . 10
1.4 Robust RL . . . 11
1.4.1 Challenges of robust RL . . . 13
1.4.2 Our contributions to model-based robust RL with a generative model . . . 15
1.4.3 Our contributions to model-based robust offline RL . . . 18
1.5 Related works . . . 19
1.5.1 Online RL . . . 19
1.5.2 Offline RL . . . 21
1.5.3 Robust RL . . . 22
1.5.4 Other related works . . . 23
1.6 Thesis organization and notation . . . 25

Chapter 2 Models and Backgrounds 27

2.1 Preliminaries of standard RL . . . 27
2.1.1 Basics of episodic finite-horizon MDPs . . . 27
2.1.2 Basics of discounted infinite-horizon MDPs . . . 28
2.2 Preliminaries of robust RL . . . 29

2.2.1 Basics of episodic finite-horizon RMDPs . . . 29
2.2.2 Basics of discounted infinite-horizon RMDPs . . . 32

Chapter 3 Model-Free Online RL **35**
3.1 Problem formulation . . . 35
3.2 Algorithm and theory . . . 35
3.2.1 Review: Q-learning with UCB exploration and reference advantage . . . 35
3.2.2 The proposed algorithm: CB-Q-Advantage . . . 37
3.2.3 Theoretical guarantees . . . 40
3.3 Analysis . . . 42
3.3.1 Preliminaries: basic properties about learning rates . . . 42
3.3.2 Additional notation used in the proof . . . 44
3.3.3 Key properties of Q-estimates and auxiliary sequences . . . 45
3.3.4 Main steps of the proof . . . 47
3.4 Discussions . . . 50

Chapter 4 Model-Free Offline RL **51**
4.1 Problem formulation . . . 51
4.2 Algorithms and theory . . . 52
4.2.1 LCB-Q: a natural pessimistic variant of Q-learning . . . 52
4.2.2 Theoretical guarantees for LCB-Q . . . 54
4.2.3 LCB-Q-Advantage for near-optimal offline RL . . . 55
4.2.4 Theoretical guarantees for LCB-Q-Advantage . . . 57
4.3 Analysis . . . 58
4.3.1 Analysis of LCB-Q . . . 58
4.3.2 Analysis of LCB-Q-Advantage . . . 63
4.4 Discussions . . . 66

Chapter 5 Model-Based Offline RL **71**
5.1 Algorithm and theory: episodic finite-horizon MDPs . . . 71
5.1.1 A refined single-policy concentrability $C^{\star}_{\mathsf{clipped}}$ . . . 71
5.1.2 A model-based offline RL algorithm: VI-LCB . . . 72
5.1.3 VI-LCB with two-fold subsampling . . . 74
5.1.4 Theoretical guarantees . . . 75
5.2 Algorithm and theory: discounted infinite-horizon MDPs . . . 78
5.2.1 Problem formulation and assumptions . . . 78
5.2.2 VI-LCB for infinite-horizon MDPs . . . 81
5.2.3 Theoretical guarantees . . . 84

5.3 Analysis: episodic finite-horizon MDPs . . . 87
5.3.1 Preliminary facts and notation . . . 87
5.3.2 A crucial statistical independence property . . . 88
5.3.3 Proof of Theorem 4 . . . 89
5.4 Analysis: discounted infinite-horizon MDPs . . . 95
5.4.1 Preliminary facts . . . 96
5.4.2 Proof of Theorem 9 . . . 98
5.5 Numerical experiments . . . 106
5.6 Discussions . . . 107

Chapter 6 Model-Based Robust RL with a Generative Model 109
6.1 Problem formulation . . . 109
6.2 Distributionally robust value iteration . . . 110
6.3 Theoretical guarantees: sample complexity analyses . . . 111
6.3.1 The case of TV distance: RMDPs are easier to learn than standard MDPs . 111
6.3.2 The case of χ^2 divergence: RMDPs can be harder than standard MDPs . . . 114
6.4 Discussions . . . 116

Chapter 7 Model-Based Robust Offline RL 117
7.1 Algorithm and theory: episodic finite-horizon RMDPs . . . 117
7.1.1 Problem formulation and assumptions . . . 117
7.1.2 Proposed algorithm: a pessimistic variant of robust value iteration . . . 119
7.1.3 Theoretical guarantees . . . 121
7.2 Algorithm and theory: discounted infinite-horizon RMDPs . . . 123
7.2.1 Problem formulation and assumptions . . . 123
7.2.2 DRVI-LCB for discounted infinite-horizon RMDPs . . . 124
7.2.3 Theoretical guarantees . . . 127
7.3 Numerical experiments . . . 129
7.4 Discussions . . . 131

Chapter 8 Conclusion 132

Appendix A Proofs for Chapter 3 134
A.1 Freedman's inequality . . . 134
A.1.1 A user-friendly version of Freedman's inequality . . . 134
A.1.2 Application of Freedman's inequality . . . 134
A.2 Proof of Lemma 1 . . . 139
A.3 Proof of key lemmas in Chapter 3.3.3 . . . 140
A.3.1 Proof of Lemma 2 . . . 140

A.3.2 Proof of Lemma 3 . . . 156
A.3.3 Proof of Lemma 4 . . . 160
A.4 Proof of Lemma 5 . . . 165
A.5 Proof of Lemma 6 . . . 169
A.5.1 Bounding the term $\mathcal{R}_1$ . . . 169
A.5.2 Bounding the term $\mathcal{R}_2$ . . . 170
A.5.3 Bounding the term $\mathcal{R}_3$ . . . 178

Appendix B Proofs for Chapter 4 **188**
B.1 Technical lemmas . . . 188
B.1.1 Preliminary facts . . . 188
B.1.2 Application of Freedman's inequality . . . 189
B.2 Proof of main lemmas for LCB-Q (Theorem 2) . . . 193
B.2.1 Proof of Lemma 7 . . . 193
B.2.2 Proof of Lemma 8 . . . 197
B.2.3 Proof of Lemma 9 . . . 200
B.3 Proof of lemmas for LCB-Q-Advantage (Theorem 3) . . . 202
B.3.1 Proof of Lemma 10 . . . 205
B.3.2 Proof of Lemma 11 . . . 210
B.3.3 Proof of Lemma 12 . . . 217
B.3.4 Proof of Lemma 30 . . . 232

Appendix C Proofs for Chapter 5 **243**
C.1 Proof of auxiliary lemmas: episodic finite-horizon MDPs . . . 243
C.1.1 Proof of Lemma 13 . . . 243
C.2 Proof of auxiliary lemmas: infinite-horizon MDPs . . . 246
C.2.1 Proof of Lemma 14 . . . 246
C.2.2 Proof of Lemma 15 . . . 250
C.2.3 Proof of Lemma 19 . . . 251
C.2.4 Proof of Lemma 20 . . . 252
C.2.5 Proof of Theorem 5 . . . 260
C.3 Proof of minimax lower bounds . . . 273
C.3.1 Preliminary facts . . . 273
C.3.2 Proof of Theorem 7 . . . 274
C.4 Discounted infinite-horizon MDPs with Markovian data . . . 282
C.4.1 Sampling models and assumptions . . . 282
C.4.2 A subsampling trick . . . 283
C.4.3 Performance guarantees . . . 286

Appendix D Proofs for Chapter 6 **291**
D.1 Preliminaries . . . 291
D.1.1 Basic facts . . . 292
D.1.2 Properties of the robust Bellman operator . . . 293
D.1.3 Additional facts of the empirical robust MDP . . . 295
D.2 Proof of the upper bound with TV distance: Theorem 10 . . . 298
D.2.1 Technical lemmas . . . 298
D.2.2 Proof of Theorem 10 . . . 298
D.2.3 Proof of the auxiliary lemmas . . . 310
D.3 Proof of the lower bound with TV distance: Theorem 11 . . . 322
D.3.1 Construction of the hard problem instances . . . 322
D.3.2 Establishing the minimax lower bound . . . 324
D.3.3 Proof of the auxiliary facts . . . 327
D.4 Proof of the upper bound with χ^2 divergence: Theorem 12 . . . 331
D.4.1 Proof of Theorem 12 . . . 332
D.4.2 Proof of the auxiliary lemmas . . . 334
D.5 Proof of the lower bound with χ^2 divergence: Theorem 13 . . . 339
D.5.1 Construction of the hard problem instances . . . 339
D.5.2 Establishing the minimax lower bound . . . 341
D.5.3 Proof of the auxiliary facts . . . 344

Appendix E Proofs for Chapter 7 **351**
E.1 Preliminaries . . . 351
E.1.1 Properties of the robust Bellman operator . . . 351
E.1.2 Concentration inequalities . . . 352
E.1.3 Kullback-Leibler (KL) divergence . . . 352
E.2 Analysis: episodic finite-horizon RMDPs . . . 353
E.2.1 Proof of Theorem 14 . . . 353
E.2.2 Proof of Lemma 61 . . . 358
E.2.3 Proof of Theorem 15 . . . 364
E.3 Analysis: discounted infinite-horizon RMDPs . . . 380
E.3.1 Proof of Lemma 38 . . . 380
E.3.2 Proof of Lemma 41 . . . 381
E.3.3 Proof of Theorem 16 . . . 382
E.3.4 Proof of Theorem 17 . . . 394

List of Figures

1.1 Illustrations of the obtained sample complexity upper and lower bounds for learning RMDPs with comparisons to state-of-the-art and the sample complexity of standard MDPs, where the uncertainty set is specified using the TV distance (a) and the χ^2 divergence (b). 16

4.1 An illustration of the epoch-based LCB-Q-Advantage algorithm. 55

5.1 The performances of the proposed method VI-LCB and the baseline value iteration (VI) in the gambler's problem. It shows that VI-LCB outperforms VI by taking advantage of the pessimism principle and achieves approximately $1/\sqrt{N}$ sample complexity dependency w.r.t. the sample size N. 106

7.1 The performance evaluation of the proposed algorithm DRVI-LCB, where it shows better sample efficiency than the baseline algorithm DRVI without pessimism, as well as better robustness in the learned policy compare to its non-robust counterpart. . 131

List of Tables

1.1 Comparisons between prior art and our results for non-stationary episodic MDPs when $T \geq H^2SA$. The table includes the order of the regret bound, the range of sample sizes that achieve the optimal regret $\widetilde{O}(\sqrt{H^2SAT})$, and the memory complexity, with all logarithmic factors omitted for simplicity of presentation. The red text highlights the suboptimal part of the respective algorithms. 6

1.2 Comparisons with prior results (up to log terms) regarding finding an ε-optimal policy in offline RL. The ε-range stands for the range of accuracy level ε for which the derived sample complexity is optimal. Here, one always has $C^\star_{\mathsf{clipped}} \leq C^\star$; and the parameter $d^{\mathsf{b}}_{\min} := \frac{1}{\min_{s,a,h}\{d^{\mathsf{b}}_h(s,a):\, d^{\mathsf{b}}_h(s,a)>0\}}$ employed in Yin and Wang (2021) could be exceedingly small, with d^{b}_h the occupancy distribution of the dataset. While multiple algorithms are referred to as VI-LCB in the table, they correspond to different variants of VI-LCB. Our results are the first to achieve sample optimality for the full ε-range. 12

1.3 Comparisons between our results and prior arts for finding an ε-optimal robust policy in the infinite-horizon RMDPs, with the uncertainty set measured w.r.t. the TV distance. Here, S, A, γ, and $\sigma \in (0,1)$ are the state space size, the action space size, the discount factor, and the uncertainty level, respectively, and all logarithmic factors are omitted in the table. Our results (Shi et al., 2023b) provide the first matching upper and lower bounds (up to log factors), improving upon all prior results. 15

1.4 Comparisons between our results and prior art on finding an ε-optimal robust policy in the infinite-horizon RMDPs, with the uncertainty set measured w.r.t. the χ^2 divergence. Here, S, A, γ, and $\sigma \in (0,\infty)$ are the state space size, the action space size, the discount factor, and the uncertainty level, respectively, and all logarithmic factors are omitted in the table. Improving upon all prior results, our theory is tight (up to log factors) when $\sigma \asymp 1$, and otherwise loose by no more than a polynomial factor in $1/(1-\gamma)$. 17

1.5 Our results for finding an ε-optimal robust policy in the finite/infinite-horizon robust MDPs with an uncertainty set measured with respect to the KL divergence using history data under partial coverage. The sample complexities included in the table are valid for sufficiently small ε, with all logarithmic factors omitted. Here, σ is the uncertainty level, $C^\star_{\mathsf{rob}}$ is the robust single-policy clipped concentrability coefficient, $P^\star_{\mathsf{min}}$ is the smallest positive state transition probability of the nominal kernel *visited by the optimal robust policy* $\pi^\star$. 20

1.6 Comparisons between our results and prior arts for finding an ε-optimal robust policy in the infinite/finite-horizon robust MDPs with an uncertainty set measured with respect to the KL divergence under full coverage of the history data. The sample complexities included in the table are valid for sufficiently small ε, with all logarithmic factors omitted. Here, σ is the uncertainty level, $P^{\star}_{\mathsf{min}}$ is the smallest positive state transition probability of the nominal kernel *visited by the optimal robust policy* $\pi^{\star}$, and P_{min} is the smallest positive state transition probability of the nominal kernel; it holds $P_{\mathsf{min}} \leq P^{\star}_{\mathsf{min}}$. 21

List of Algorithms

1 CB-Q-Advantage . . . 39
2 Auxiliary functions . . . 41
3 CB-Q-Advantage (a rewrite of Algorithm 1 that specifies dependency on k) . . . 43

4 LCB-Q for offline RL . . . 53
5 Offline LCB-Q-Advantage RL . . . 67
6 Auxiliary functions . . . 68
7 LCB-Q for offline RL (a rewrite of Algorithm 4 to specify dependency on k) . . . 69
8 LCB-Q-Advantage (a rewrite of Algorithm 5 that specifies dependency on k or (m, t).) 70

9 Offline value iteration with LCB (VI-LCB) for finite-horizon MDPs. . . . 73
10 Subsampled VI-LCB for episodic finite-horizon MDPs . . . 75
11 Offline value iteration with LCB (VI-LCB) for discounted infinite-horizon MDPs . . . 83

12 Distributionally robust value iteration (DRVI) for infinite-horizon RMDPs. . . . 111

13 Robust value iteration with LCB (DRVI-LCB) for robust offline RL. . . . 121
14 Robust value iteration with LCB (DRVI-LCB) for infinite-horizon RMDPs. . . . 125

15 Subsampled VI-LCB for discounted infinite MDPs with Markovian data . . . 285

Chapter 1

Introduction

When contemplating the process of learning, especially within an unfamiliar environment, the first thing that comes to mind is probably establishing a loop involving interacting with the real-world environment and updating of knowledge to enhance performance. In this context, reinforcement learning (RL) is a prominent framework that provides a general and mathematical formulation of the learning process including the learner (or called an *agent*), the environments, and their interactions. Specifically, RL introduces several key concepts to formulate the learning process: 1) *action*: how the agent moves to interact with the environment; 2) *state*: the status of the agent and the environment; 3) *policy*: the strategy the agent employs to select an action; 4) *reward*: the immediate feedback the agent receives post-interaction with the environment. Equipped with these principles, the learning problems can be naturally described by RL framework as searching for an optimal sequential decision-making policy to maximize the long-term cumulative rewards gained through trial-and-error within an unknown environment.

1.1 Overview

As a fast-growing subfield of artificial intelligence, RL has achieved remarkable success in diverse areas of human endeavor, such as games (Silver et al., 2017), large language model alignment (OpenAI, 2023; Ziegler et al., 2019), healthcare (Fatemi et al., 2021; Liu et al., 2019), and robotics and control (Kober et al., 2013; Mnih et al., 2013). These noteworthy accomplishments are largely due to the vast volume of interactive data that fuels the learning of the policy. Today, data-driven methodologies are progressively vital in enhancing various aspects of human life. Then it is natural to ask:

In designing data-driven RL algorithms, what should we consider?

Sample efficiency is arguably a cornerstone of contemporary RL that cannot be overlooked. Contemporary RL problems typically involve unprecedentedly large environments and models of the policies (OpenAI, 2023; Silver et al., 2017). Consequently, an agent may need to accumulate vast amounts of data from these extensive environments to learn an effective policy, especially in ubiquitous data-starved applications. This challenge is further magnified as the environment's complexity increases exponentially in terms of the horizon length, a characteristic inherent to RL's sequential problem structure. Moreover, data collection can be limited by privacy, expensive, time-consuming, or even high-stakes issues, for instance, in clinical trials, online advertisements,

and autonomous systems (Best et al., 2018; Fulbright, 2017; McGinnis et al., 2011; Saengkyongam et al., 2023). Consequently, understanding and improving the sample efficiency of RL algorithms inevitably lie at the core of cutting-edge RL research and are the key enabler for future advances.

In order to evaluate and compare the sample efficiency of RL algorithms in high dimension, a recent body of works sought to develop a finite-sample theoretical framework to analyze the algorithms of interest, with the aim of delineating the dependency of algorithm performance on all salient problem parameters in a non-asymptotic fashion (Dann et al., 2017; Kakade, 2003). Such finite-sample guarantees are brought to bear towards understanding and tackling the sample efficiency challenges in the sample-starved regime commonly encountered in practice and have achieved tremendous progress. Nevertheless, existing statistical understanding and provable algorithm performance are still far from adequate in both theory and practice, due to technical challenges and the broadness and diversity of RL world.

In light of this, this thesis concentrates on understanding and breaking sample size barriers for different RL problems using the finite-sample theoretical framework. Prior to proceeding further, it is also worth emphasizing two other vital facets of algorithm performance that we consider in data-driven RL problems, as outlined below:

- *Algorithm scalability.* Given that the dimensions of environments encountered in practical applications are often substantial, the scalability of RL algorithms is of critical importance, particularly when memory and computational resources are constrained.

- *Robustness to uncertainty.* Robustness is highly desirable since the performance of the learned policy in training environment could significantly deteriorate due to the uncertainty and variations of the test environment induced by random perturbation, rare events, or even malicious attacks (Mahmood et al., 2018; Zhang et al., 2021a).

In this thesis, driven by the principles previously mentioned, we concentrate on surmounting sample barriers across various RL tasks by offering statistical insights and designing sample-efficient algorithms with provable non-asymptotic guarantees. These efforts can be paired with two equally significant principles - scalability and robustness. RL problem formulations can be classified in numerous ways, according to the objectives of tasks, task-specific structures, and sampling mechanisms, i.e., available data collection methods. Bearing this in mind, this thesis will focus on three RL settings with distinct sampling mechanisms and objectives - online RL, offline RL, and distributionally robust RL, which we will introduce shortly.

Specifically, this thesis focuses on the widely studied Markov decision processes (MDPs), whose salient problem parameters (i.e., the number of states, actions, and the effective horizon) could be enormous in modern RL applications. We consider two sets of MDPs that have been extensively studied - finite-horizon MDPs and discounted infinite-horizon MDPs, which shall be separately introduced in Chapter 2.1. These two settings are generally alike except for the configuration of the

accumulated reward, which gives rise to distinct technical challenges. For either or both of the two settings, we evaluate and compare the statistical performance of RL algorithms mainly through the lens of sample complexity — namely, the number of samples needed for an algorithm to output, with probability approaching one, a policy whose resultant value function is at most ε away from optimal (called "ε-accuracy" throughout).

The rest of this chapter is organized as follows. Chapter 1.2 to Chapter 1.4 provide an overview of the main results of this thesis in understanding or breaking the sample barriers in a variety of RL settings. Chapter 1.5 summarizes the related works. Finally, Chapter 1.6 describes the organization of the rest of the thesis.

1.2 Online RL

An agent in online RL is only allowed to draw sample trajectories by executing a policy in the unknown Markov decision process (MDP), where the initial states are pre-assigned and might even be chosen by an adversary. Careful deliberation needs to be undertaken when deciding what policies to use to allow for effective interaction with the unknown environment, how to optimally balance exploitation and exploration, and how to process and store the collected information intelligently without causing redundancy. Consequently, simultaneously achieving the desired sample efficiency and memory efficiency for algorithm scalability is particularly challenging when it comes to online RL scenarios.

1.2.1 Regret-optimal model-free RL? A sample size barrier

To facilitate discussion, let us take a moment to summarize the state-of-the-art theory, focusing on minimizing cumulative *regret* — a metric that quantifies the performance difference between the learned policy and the true optimal policy — with the fewest number of samples. For the formal definition of regret, please refer to Chapter 3.1. Here and throughout, we denote by S, A, and H the size of the state space, the size of the action space, and the horizon length of an episodic finite-horizon MDP, respectively, and let T represent the sample size. The immediate reward gained at each time step is assumed to lie between 0 and 1.

Fundamental regret lower bound. Following the arguments in Auer et al. (2002); Jaksch et al. (2010), the recent works Domingues et al. (2021); Jin et al. (2018) developed a fundamental lower bound[1] on the expected total regret for this setting. Specifically, this lower bound claims that: no

[1] It is worth emphasizing that Domingues et al. (2021) adopts the notation T to represent the number of trajectories (with each trajectory containing H samples), while the present chapter employs K to denote the number of sample trajectories and $T = KH$ the total number of samples. Consequently, the lower bound developed in Domingues et al. (2021) for non-stationary finite-horizon MDPs reads $\Omega(\sqrt{H^3SAK})$, or equivalently, $\Omega(\sqrt{H^2SAT})$ using the notation adopted herein.

matter what algorithm to use, one can find an MDP such that the accumulated regret incurred by the algorithm necessarily exceeds the order of

$$\text{(lower bound)} \qquad \sqrt{H^2SAT}, \tag{1.1}$$

as long as $T \geq H^2SA$.[2] This sublinear regret lower bound in turn imposes a sampling limit if one wants to achieve ε average regret.

Model-based RL. Moving beyond the lower bound, let us examine the effectiveness of model-based RL — which can be interpreted as a "plug-in" statistical approach — start by computing an empirical model for the unknown MDP, and output a policy that is (near)-optimal in accordance with the empirical MDP (Agrawal and Jia, 2023; Azar et al., 2017; Efroni et al., 2019; Jaksch et al., 2010; Pacchiano et al., 2021). In order to ensure a sufficient degree of exploration, Azar et al. (2017) came up with an algorithm called UCB-VI that blends model-based learning and the optimism principle, which achieves a regret bound $\widetilde{O}\big(\sqrt{H^2SAT}\big)$ that nearly attains the lower bound (1.1) as T tends to infinity. Caution needs to be exercised, however, that existing theory does not guarantee the near optimality of this algorithm unless the sample size T surpasses

$$T \geq S^3AH^6,$$

a threshold that is significantly larger than the dimension of the underlying model. This threshold can also be understood as the initial *burn-in cost* of the algorithm, namely, a sampling burden needed for the algorithm to exhibit the desired performance. In addition, model-based algorithms typically require storing the estimated probability transition kernel, resulting in a space complexity that could be as high as $O(S^2AH)$ (Azar et al., 2017).

Model-free RL. Another competing solution paradigm is the model-free approach, which circumvents the model estimation stage and attempts to learn the optimal values directly (Bai et al., 2019; Jin et al., 2018; Strehl et al., 2006; Yang et al., 2021). Noteworthily, Q-learning and its variants (Watkins and Dayan, 1992), which apply stochastic approximation (Robbins and Monro, 1951) based on the Bellman optimality condition, are among the most widely used model-free paradigms. In comparison to the model-based counterpart, the model-free approach holds the promise of low space complexity, as it eliminates the need of storing a full description of the model. In fact, in a number of previous works (e.g., Jin et al. (2018); Strehl et al. (2006)), an algorithm is declared to be model-free only if its space complexity is $o(S^2AH)$ regardless of the sample size T.

[2] Given that a trivial upper bound on the regret is T, one needs to impose a lower bound $T \geq H^2SA$ in order for (1.1) to be meaningful.

- *Memory-efficient model-free methods.* Jin et al. (2018) proposed the first memory-efficient model-free algorithm — which is an optimistic variant of classical Q-learning — that achieves a regret bound proportional to $\sqrt{T}$ with a space complexity $O(SAH)$. Compared to the lower bound (1.1), however, the regret bound in Jin et al. (2018) is off by a factor of $\sqrt{H}$ and hence suboptimal for problems with long horizon. This drawback has recently been overcome in Zhang et al. (2020c) by leveraging the idea of variance reduction (or the so-called "reference-advantage decomposition") for large enough T. While the resulting regret matches the information-theoretic limit asymptotically, its optimality in the non-asymptotic regime is not guaranteed unless the sample size T exceeds (see Zhang et al. (2020c, Lemma 7))

$$T \geq S^6 A^4 H^{28},$$

  a requirement that is even far more stringent than the burn-in cost imposed by Azar et al. (2017).

- *A memory-inefficient "model-free" variant.* The recent work Ménard et al. (2021) put forward a novel sample-efficient variant of Q-learning called UCB-M-Q, which relies on a carefully chosen momentum term for bias reduction. This algorithm is guaranteed to yield near-optimal regret $\widetilde{O}\big(\sqrt{H^2SAT}\big)$ as soon as the sample size exceeds $T \geq SA\mathrm{poly}(H)$, which is a remarkable improvement vis-à-vis previous regret-optimal methods (Azar et al., 2017; Zhang et al., 2020c). Nevertheless, akin to the model-based approach, it comes at a price in terms of the space and computation complexities, as the space required to store all bias-value function is $O(S^2AH)$ and the computation required is $O(ST)$, both of which are larger by a factor of S than other model-free algorithms like Zhang et al. (2020c). In view of this memory inefficiency, UCB-M-Q falls short of fulfilling the definition of model-free algorithms in Jin et al. (2018); Strehl et al. (2006). See Ménard et al. (2021, Section 3.3) for more detailed discussions.

A more complete summary of prior results can be found in Table 1.1.

1.2.2 Our contributions to model-free online RL

In brief, while it is encouraging to see that both model-based and model-free approaches allow for near-minimal regret as the sample size T tends to infinity, they are either memory-inefficient, or burn-in sample-inefficient — require the sample size to exceed a threshold substantially larger than the model dimensionality. In fact, no prior algorithms have been shown to be *simultaneously regret-optimal and memory-efficient* unless

$$T \geq S^6 A^4 \,\mathrm{poly}(H),$$

Algorithm	Regret	Range of sample sizes T that attain optimal regret	Space complexity
UCB-VI (Azar et al., 2017)	$\sqrt{H^2SAT} + H^4S^2A$	$[S^3AH^6, \infty)$	S^2AH
UCB-Q-Hoeffding (Jin et al., 2018)	$\sqrt{H^4SAT}$	never	SAH
UCB-Q-Bernstein (Jin et al., 2018)	$\sqrt{H^3SAT} + \sqrt{H^9S^3A^3}$	never	SAH
UCB2-Q-Bernstein (Bai et al., 2019)	$\sqrt{H^3SAT} + \sqrt{H^9S^3A^3}$	never	SAH
UCB-Q-Advantage (Zhang et al., 2020c)	$\sqrt{H^2SAT} + H^8S^2A^{\frac{3}{2}}T^{\frac{1}{4}}$	$[S^6A^4H^{28}, \infty)$	SAH
UCB-M-Q (Ménard et al., 2021)	$\sqrt{H^2SAT} + H^4SA$	$[SAH^6, \infty)$	S^2AH
CB-Q-Advantage **(Theorem 1)**	$\sqrt{H^2SAT} + H^6SA$	$[SAH^{10}, \infty)$	SAH
Lower bound (Domingues et al., 2021)	$\sqrt{H^2SAT}$	n/a	n/a

Table 1.1: Comparisons between prior art and our results for non-stationary episodic MDPs when $T \geq H^2SA$. The table includes the order of the regret bound, the range of sample sizes that achieve the optimal regret $\widetilde{O}(\sqrt{H^2SAT})$, and the memory complexity, with all logarithmic factors omitted for simplicity of presentation. The red text highlights the suboptimal part of the respective algorithms.

which constitutes a stringent sample size barrier constraining their utility in the sample-starved and memory-limited regime. The presence of this sample complexity barrier motivates one to pose a natural question:

> *Is it possible to design an algorithm that accommodates a significantly broader sample size range without compromising regret optimality and memory efficiency?*

We break the sample barrier affirmatively, by designing a new model-free algorithm, dubbed as CB-Q-Advantage, which has a space complexity $O(SAH)$, and achieves near-optimal regret $\widetilde{O}\big(\sqrt{H^2SAT}\big)$ as soon as the sample size exceeds $T \geq SA\,\mathrm{poly}(H)$. As can be seen from Table 1.1, the proposed algorithm is far more memory-efficient than both the model-based approach in Azar et al. (2017) and the UCB-M-Q algorithm in Ménard et al. (2021) (both of these prior algorithms require S^2AH units of space). In addition, the sample size requirement $T \geq SA\,\mathrm{poly}(H)$ of our algorithm improves — by a factor of at least S^5A^3 — upon that of any prior algorithm that is simultaneously regret-optimal and memory-efficient. In fact, this requirement is nearly sharp in

terms of the dependency on both S and A, and was previously achieved only by the UCB-M-Q algorithm at a price of a much higher storage burden.

Let us also briefly highlight the key ideas of our algorithm. As an optimistic variant of variance-reduced Q-learning, CB-Q-Advantage leverages the recently-introduced reference-advantage decompositions for variance reduction (Zhang et al., 2020c). As a distinguishing feature from prior algorithms, we employ an *early-stopped* reference update rule, with the assistance of two Q-learning sequences that incorporate upper and lower confidence bounds, respectively. The design of our early-stopped variance reduction scheme, as well as its analysis framework, might be of independent interest to other settings that involve managing intricate exploration-exploitation trade-offs.

1.3 Offline RL

Limited capability of online data collection in other real-world applications — e.g., clinical trials and online advertising, where real-time data acquisition is expensive, high-stakes, and/or time-consuming, — presents a fundamental bottleneck for carrying such RL success over to broader scenarios. To circumvent this bottleneck, one plausible strategy is to make more effective use of data collected previously, given that such historical data might contain useful information that readily transfers to new tasks (for instance, the state transitions in a historical task might sometimes resemble what happen in new tasks). The potential of this data-driven approach has been explored and recognized in a diverse array of contexts including but not limited to robotic manipulation (Ebert et al., 2018), autonomous driving (Diehl et al., 2021), and healthcare (Tang and Wiens, 2021); see Levine et al. (2020); Prudencio et al. (2023) for overviews of recent development. Nowadays, the subfield of RL using historical data, without further exploration of the environment, is commonly referred to as *offline RL* or *batch RL* (Lange et al., 2012; Levine et al., 2020). A desired offline RL algorithm would achieve the target statistical accuracy using as few samples of the history dataset as possible.

1.3.1 Challenges of offline RL: distribution shift and limited data coverage

In contrast to online exploratory RL, offline RL has to deal with several critical issues resulting from the absence of active exploration. Below we single out two representative issues surrounding offline RL.

- *Distribution shift.* For the most part, the historical data is generated by a certain behavior policy that departs from the optimal one. A key challenge in offline RL thus stems from the shift of data distributions: how to leverage past data to the most effect, even though the distribution induced by the target policy differs from what we have available?
- *Limited data coverage.* Ideally, if the dataset contained sufficiently many data samples for every state-action pair, then there would be hope to simultaneously learn the performance of

every policy. Such a uniform coverage requirement, however, is oftentimes not only unrealistic (given that we can no longer change the past data) but also unnecessary (given that we might only be interested in identifying a single optimal policy).

Whether one can effectively cope with distribution shift and insufficient data coverage becomes a major factor that governs the feasibility and statistical efficiency of offline RL.

In order to address the aforementioned issues, a recent strand of works put forward the *principle of pessimism or conservatism* (e.g., Buckman et al. (2020); Chen et al. (2021a); Cui and Du (2022); Jin et al. (2021); Kumar et al. (2020); Liu et al. (2020); Rashidinejad et al. (2021); Uehara and Sun (2021); Xie et al. (2021b); Yin and Wang (2021); Zanette et al. (2021); Zhong et al. (2022)). This is reminiscent of the optimism principle in the face of uncertainty for online exploration (Azar et al., 2017; Bourel et al., 2020; Jaksch et al., 2010; Jin et al., 2018; Lai and Robbins, 1985), but works for drastically different reasons (as we shall elucidate momentarily). One plausible idea of the pessimism principle, which has been incorporated into offline RL approaches, is to penalize value estimation of those state-action pairs that have been poorly covered. Informally speaking, insufficient coverage of a state-action pair inevitably results in low confidence and high uncertainty in the associated value estimation, and it is hence advisable to act cautiously by tuning down the corresponding value estimate. Proper use of pessimism amid uncertainty brings about several provable benefits (Rashidinejad et al., 2021; Xie et al., 2021b): (i) it allows for a reduced sample size that adapts to the degree of distribution shift; (ii) as opposed to uniform data coverage, it only requires coverage of the part of the state-action space reachable by the target policy.

1.3.2 Inadequacy of prior works

Despite extensive recent activities, however, existing statistical guarantees for the above paradigm remain inadequate, as we shall elaborate on below. In addition, previous works have isolated an important parameter $C^\star \geq 1$ — called the single-policy concentrability coefficient (Rashidinejad et al., 2021; Xie et al., 2021b) — that measures the mismatch of distributions induced by the target policy against the behavior policy; see Chapters 5.1.1 and 5.2.1 for precise definitions. Naturally, the statistical performance of desirable algorithms would degrade gracefully as the distribution mismatch worsens (i.e., as $C^\star$ increases). In the sequel, considering finite-horizon non-stationary (with horizon length H) and discounted infinite-horizon (with discount factor γ) MDPs, we shall discuss the two RL paradigms introduced in Chapter 1.2.1 — model-based RL and model-free RL — in the scope of offline RL separately.

Model-based offline RL. When coupled with the pessimism principle in offline RL, the model-based approach has been shown to enjoy the following sample complexity bounds.

- By incorporating Hoeffding-style lower confidence bounds into value iteration, Rashidinejad

et al. (2021); Xie et al. (2021b) demonstrated that a sample complexity of

$$\begin{cases} \widetilde{O}\big(\frac{H^6SC^\star}{\varepsilon^2}\big) & \text{for finite-horizon MDPs} \\ \widetilde{O}\big(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}\big) & \text{for infinite-horizon MDPs} \end{cases} \tag{1.2}$$

suffices to yield ε-accuracy. Such a sample complexity bound, however, is a large factor of H^2 (resp. $\frac{1}{(1-\gamma)^2}$) above the minimax lower limit derived for finite-horizon MDPs (resp. infinite-horizon MDPs) (Rashidinejad et al., 2021; Xie et al., 2021b; Yin and Wang, 2021).

- In an attempt to optimize the sample complexity, Xie et al. (2021b) leveraged the idea of variance reduction — a powerful strategy originating from the stochastic optimization literature (Johnson and Zhang, 2013) — in model-based RL and obtained a strengthened sample complexity of

$$\widetilde{O}\Big(\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^{6.5}SC^\star}{\varepsilon}\Big) \tag{1.3}$$

for finite-horizon MDPs. This sample complexity bound approaches the minimax lower limit (i.e., the order of $\frac{H^4SC^\star}{\varepsilon^2}$) once the sample size exceeds the order of

$$\text{(burn-in cost)} \qquad H^9SC^\star; \tag{1.4}$$

in other words, an enormous burn-in sample size is needed in order to attain sample optimality.

Model-free offline RL. Before this thesis, it remains unknown whether the pessimism principle can be incorporated into model-free algorithms — another class of popular algorithms that is flexible and performs learning without model estimation — in a provably effective fashion for offline RL.

1.3.3 Our contributions to model-free offline RL

Consider finite-horizon non-stationary MDPs, our work pins down the sample efficiency for pessimistic variants of model-free algorithms, under the mild single-policy concentrability assumption (Rashidinejad et al., 2021; Xie et al., 2021b). Given K episodes of history data each of length H (which amounts to a total number of $T = KH$ samples), our main contributions are summarized as follows.

- We first study a natural pessimistic variant of the Q-learning algorithm, which simply modifies the classical Q-learning update rule by subtracting a penalty term (via certain lower confidence bounds). We prove that pessimistic Q-learning finds an ε-optimal policy as soon as the sample

size T exceeds the order of (up to log factor)

$$\frac{H^6SC^\star}{\varepsilon^2},$$

where $C^\star$ denotes the single-policy concentrability coefficient of the batch dataset. In comparison to the minimax lower bound $\Omega\big(\frac{H^4SC^\star}{\varepsilon^2}\big)$ developed in Xie et al. (2021b), the sample complexity of pessimistic Q-learning is at most a factor of H^2 from optimal (modulo some log factor).

- To further improve the sample efficiency of pessimistic model-free algorithms, we introduce a variance-reduced variant of pessimistic Q-learning. This algorithm is guaranteed to find an ε-optimal policy as long as the sample size T is above the order of

$$\frac{H^4SC^\star}{\varepsilon^2} + \frac{H^5SC^\star}{\varepsilon}$$

up to some log factor. In particular, this sample complexity is minimax-optimal (namely, as low as $\frac{H^4SC^\star}{\varepsilon^2}$ up to log factor) for small enough ε (namely, $\varepsilon \leq (0, 1/H]$).

Regarding the scalability, both of the proposed algorithms achieve low computation cost (i.e., $O(T)$) and low memory complexities (i.e., $O(\min\{T, SAH\})$). In comparison with model-based algorithms, model-free algorithms require drastically different technical tools to handle the complicated statistical dependency between the estimated Q-values at different time steps.

1.3.4 Our contributions to model-based offline RL

Existing offline algorithms either suffer from suboptimal sample complexities, or require sophisticated techniques like variance reduction to approach minimax optimality (cf. Chapter 1.3.2). Even when variance reduction is employed, prior algorithms incur an enormous burn-in cost in order to work optimally, thus posing an impediment to achieving sample efficiency in data-starved applications. All this motivates the studies of the following open questions:

> *Can we develop an offline RL algorithm that achieves near-optimal sample complexity without burn-in cost? If so, can we accomplish this goal by means of a simple algorithm without resorting to sophisticated schemes like variance reduction?*

In this thesis, we settle the sample complexity of model-based offline RL by studying a pessimistic variant of value iteration — called VI-LCB — applied to some empirical MDP. Encouragingly, for both finite-horizon and discounted infinite-horizon MDPs, the model-based algorithms provably achieve minimax-optimal sample complexities for any given target accuracy level ε — namely, any $\varepsilon \in (0, H]$ for finite-horizon MDPs and $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$ for discounted infinite-horizon MDPs.

To be more precise, we introduce a slightly modified version $C^\star_{\mathsf{clipped}}$ of the concentrability coefficient $C^\star$, which always satisfies $C^\star_{\mathsf{clipped}} \le C^\star$ and shall be termed the single-policy clipped concentrability coefficient (see Chapters 5.2.1 and 5.1.1 for more details as well as the advantages of this coefficient). The introduction of this new parameter leads to slightly improved sample complexity compared to the one based on $C^\star$. The main contributions are summarized as follows.

- For finite-horizon MDPs with nonstationary transition kernels, we propose a variant of VI-LCB that adopts the Bernstein-style penalty to enforce pessimism in the face of uncertainty. We prove that for any given $\varepsilon \in (0, H]$, the proposed algorithm yields an ε-optimal policy using

$$\widetilde{O}\left(\frac{H^4 S C^\star_{\mathsf{clipped}}}{\varepsilon^2}\right) \tag{1.5}$$

  samples with high probability. A key ingredient in the algorithm design is a two-fold subsampling trick that helps decouple the statistical dependency along the sample rollouts.

- For γ-discounted infinite-horizon MDPs, we demonstrate that with high probability, the VI-LCB algorithm with Bernstein-style penalty finds an ε-optimal policy with a sample complexity of

$$\widetilde{O}\left(\frac{S C^\star_{\mathsf{clipped}}}{(1-\gamma)^3 \varepsilon^2}\right) \tag{1.6}$$

  for any given accuracy level $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$. Our algorithm reuses all samples across all iterations in order to achieve data efficiency, and our analysis builds upon a novel leave-one-out argument to decouple complicated statistical dependency across iterations.

- To assess the tightness and optimality of our results, we further develop minimax lower bounds, which match the above upper bounds modulo some logarithmic factors.

Remarkably, our algorithms do not require sophisticated variance reduction schemes, as long as suitable confidence bounds are adopted. Detailed theoretical comparisons with prior art can be found in Table 1.2.

1.4 Robust RL

While standard RL has been heavily investigated recently, its use can be significantly hampered in practice due to the sim-to-real gap; for instance, a policy learned in an ideal, nominal environment might fail catastrophically when the deployed environment is subject to small changes in task objectives or adversarial perturbations (Klopp et al., 2017; Mahmood et al., 2018; Zhang et al., 2020a). Consequently, in addition to maximizing the long-term cumulative reward, robustness emerges as another critical goal for RL, especially in high-stakes applications such as robotics,

horizon	algorithm	sample complexity	ε-range to attain sample optimality	type
finite	VI-LCB (Xie et al., 2021b)	$\frac{H^6SC^\star}{\varepsilon^2}$	——	model-based
	LCB-Q **(Theorem 2)**	$\frac{H^6SC^\star}{\varepsilon^2}$	——	model-free
	VPVI (Yin and Wang, 2021)	$\frac{H^5SC^\star}{\varepsilon^2}$	——	model-based
	PEVI-Adv (Xie et al., 2021b)	$\frac{H^4SC^\star}{\varepsilon^2}+\frac{H^{6.5}SC^\star}{\varepsilon}$	$\left(0, \frac{1}{H^{2.5}}\right]$	model-based
	LCB-Q-Advantage **(Theorem 3)**	$\frac{H^4SC^\star}{\varepsilon^2}+\frac{H^5SC^\star}{\varepsilon}$	$\left(0, \frac{1}{H}\right]$	model-free
	APVI/LCBVI (Yin and Wang, 2021)	$\frac{H^4SC^\star}{\varepsilon^2}+\frac{H^4}{d^{\mathsf{b}}_{\min}\varepsilon}$	$(0, SC^\star d^{\mathsf{b}}_{\min}]$	model-based
	VI-LCB **(Theorem 4)**	$\frac{H^4SC^\star_{\mathsf{clipped}}}{\varepsilon^2}$ $\left(\leq \frac{H^4SC^\star}{\varepsilon^2}\right)$	$(0, H]$	model-based
	lower bound **(Theorem 5)**	$\frac{H^4SC^\star_{\mathsf{clipped}}}{\varepsilon^2}$	——	——
infinite	VI-LCB (Rashidinejad et al., 2021)	$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$	——	model-based
	Q-LCB (Yan et al., 2022a)	$\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}$	——	model-free
	VR-Q-LCB (Yan et al., 2022a)	$\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}+\frac{SC^\star}{(1-\gamma)^4\varepsilon}$	$(0, 1-\gamma]$	model-free
	VI-LCB **(Theorem 6)**	$\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}$ $\left(\leq \frac{SC^\star}{(1-\gamma)^3\varepsilon^2}\right)$	$\left(0, \frac{1}{1-\gamma}\right]$	model-based
	lower bound **(Theorem 7)**	$\frac{SC^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}$	——	——

Table 1.2: Comparisons with prior results (up to log terms) regarding finding an ε-optimal policy in offline RL. The ε-range stands for the range of accuracy level ε for which the derived sample complexity is optimal. Here, one always has $C^\star_{\mathsf{clipped}} \leq C^\star$; and the parameter $d^{\mathsf{b}}_{\min} := \frac{1}{\min_{s,a,h}\{d^{\mathsf{b}}_h(s,a): d^{\mathsf{b}}_h(s,a)>0\}}$ employed in Yin and Wang (2021) could be exceedingly small, with d^{b}_h the occupancy distribution of the dataset. While multiple algorithms are referred to as VI-LCB in the table, they correspond to different variants of VI-LCB. Our results are the first to achieve sample optimality for the full ε-range.

autonomous driving, clinical trials, financial investments, and so on. Towards achieving this, distributionally robust RL (Iyengar, 2005; Nilim and El Ghaoui, 2005), which leverages insights from distributionally robust optimization and supervised learning (Bertsimas et al., 2018; Blanchet and Murthy, 2019; Duchi and Namkoong, 2021; Gao, 2022; Rahimian and Mehrotra, 2019), becomes a natural yet versatile framework; the aim is to learn a policy that performs well even when the deployed environment deviates from the nominal one in the face of environment uncertainty.

More concretely, imagine that one has access to samples from a MDP with a nominal transition kernel under some sampling mechanisms. Standard RL aims to learn the optimal policy tailored to the nominal kernel, for which the minimax sample complexity limit has been fully settled (Azar et al., 2013; Li et al., 2023c). In contrast, distributionally robust RL seeks to learn a more *robust* policy using the same set of samples, with the aim of optimizing the worst-case performance when the transition kernel is arbitrarily chosen from some *prescribed* uncertainty set around the nominal kernel; this setting is frequently referred to as robust MDPs (RMDPs).[3] The formal formulation of RMDPs can be referred to Chapter 2.2. Clearly, the RMDP framework helps ensure that the performance of the learned policy does not fail catastrophically as long as the sim-to-real gap is not overly large.

Compared with standard MDPs, the class of RMDPs encapsulates richer models, given that one is allowed to prescribe the shape and size of the uncertainty set. Oftentimes, the uncertainty set is hand-picked as a small ball surrounding the nominal kernel, with the size and shape of the ball specified by some distance-like metric ρ between probability distributions and some uncertainty level σ. To ensure tractability of solving RMDPs, the uncertainty set is often selected to obey certain structures. For instance, a number of prior works assumed that the uncertainty set can be decomposed as a product of independent uncertainty subsets over each state or state-action pair (Wiesemann et al., 2013; Zhou et al., 2021), dubbed as the s- and (s, a)-rectangularity, respectively. This thesis adopts the second choice by assuming (s, a)-rectangularity for the uncertainty set.

1.4.1 Challenges of robust RL

We are interested in how the sample complexity — the number of samples needed for an algorithm to output a policy whose robust value function (the worst-case value over all the transition kernels in the uncertainty set) is at most ε away from the optimal robust one — scales with all these salient problem parameters. Unique challenge with RMDPs arises from distribution shift induced by model uncertainty, where the transition kernel drawn from the uncertainty set can be different from the nominal kernel. This challenge leads to complicated nonlinearity and nested optimization in the problem structure not present in standard MDPs. In sum, it is natural to wonder how the robustness consideration impacts data efficiency: is there a statistical premium that one needs to pay in quest

[3] While it is straightforward to incorporate additional uncertainty of the reward in our framework, we do not consider it here for simplicity, since the key challenge is to deal with the uncertainty of the transition kernel.

of additional robustness, and how to design sample-efficient RL algorithms for RMDPs?

Statistical implications of distributional robustness. It is of fundamental importance to understand about whether, and how, the choice of distributional robustness bears statistical implications in learning a desirable policy, through the lens of sample complexity. Even with the simplest sampling mechanism, i.e. the generative model (also called a simulator), there remained large gaps between the sample complexity upper and lower bounds established in prior literature, regardless of the divergence metric in use. For example, consider two choices of the distance-like metric ρ for the uncertainty set — total variation (TV) distance and the χ^2 divergence, which are motivated by their practical appeals: easy to implement, and already adopted by empirical RL (Lee et al., 2021). For the case w.r.t. the TV distance (see Table 1.3 for a summary), while the state-of-the-art upper bound (Clavier et al., 2023) and lower bound (Yang et al., 2022) coincide when the uncertainty level $\sigma \lesssim 1-\gamma$ is small, the upper bound can be a factor of $\frac{1}{(1-\gamma)^5}$ larger than the lower bound when σ approaches 1. The situation is even worse when it comes to the case w.r.t. the χ^2 divergence (see Table 1.4 for a summary). More specifically, the state-of-the-art upper bound (Panaganti and Kalathil, 2022) scales quadratically with the size S of the state space and linearly with the uncertainty level σ when $\sigma \gtrsim 1$, while the lower bound (Yang et al., 2022) exhibits only linear scaling with S and is, in the meantime, inversely proportional to σ; these lead to unbounded gaps between the upper and lower bounds as σ grows.

Perhaps a more pressing issue is that, past works of robust RL failed to provide an affirmative answer regarding how to benchmark the sample complexity of RMDPs with that of standard MDPs over the full range of uncertainty levels, given the large unresolved gaps mentioned above. In fact, prior works only achieved limited success in this regard — namely, demonstrating that the sample complexity for RMDPs is the same as that of standard MDPs in the case of TV distance when the uncertainty level satisfies $\sigma \lesssim 1-\gamma$. For all the remaining situations, however, existing sample complexity upper (resp. lower) bounds are all larger (resp. smaller) than the sample size requirement for standard MDPs. As a consequence, it remains unclear *whether learning RMDPs is harder or easier than learning standard MDPs.*

Robust RL meets offline data. Providing robustness guarantees becomes even more relevant in the offline setting, which can be formulated as *robust offline RL*, since the history data is often inevitably collected from a timeframe where it is no longer reasonable to assume model stillness, due to the highly non-stationary and time-varying dynamics of many real-world applications. Despite significant amount of recent activities in robust RL and offline RL, addressing model uncertainty and sample efficiency simultaneously remains challenging. Understanding the implications of — and designing algorithms that work around — these challenges play a major role in advancing the state-of-the-art of robust offline RL.

Result type	Reference	Sample complexity	
		$0 < \sigma \lesssim 1-\gamma$	$1-\gamma \lesssim \sigma < 1$
Upper bound	Yang et al. (2022)	$\frac{S^2A}{\sigma^2(1-\gamma)^4\varepsilon^2}$	
	Panaganti and Kalathil (2022)	$\frac{S^2A}{(1-\gamma)^4\varepsilon^2}$	
	Clavier et al. (2023)	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^4\varepsilon^2}$
	Theorem 10	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$
Lower bound	Yang et al. (2022)	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA(1-\gamma)}{\sigma^4\varepsilon^2}$
	Theorem 11	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$

Table 1.3: Comparisons between our results and prior arts for finding an ε-optimal robust policy in the infinite-horizon RMDPs, with the uncertainty set measured w.r.t. the TV distance. Here, S, A, γ, and $\sigma \in (0,1)$ are the state space size, the action space size, the discount factor, and the uncertainty level, respectively, and all logarithmic factors are omitted in the table. Our results (Shi et al., 2023b) provide the first matching upper and lower bounds (up to log factors), improving upon all prior results.

1.4.2 Our contributions to model-based robust RL with a generative model

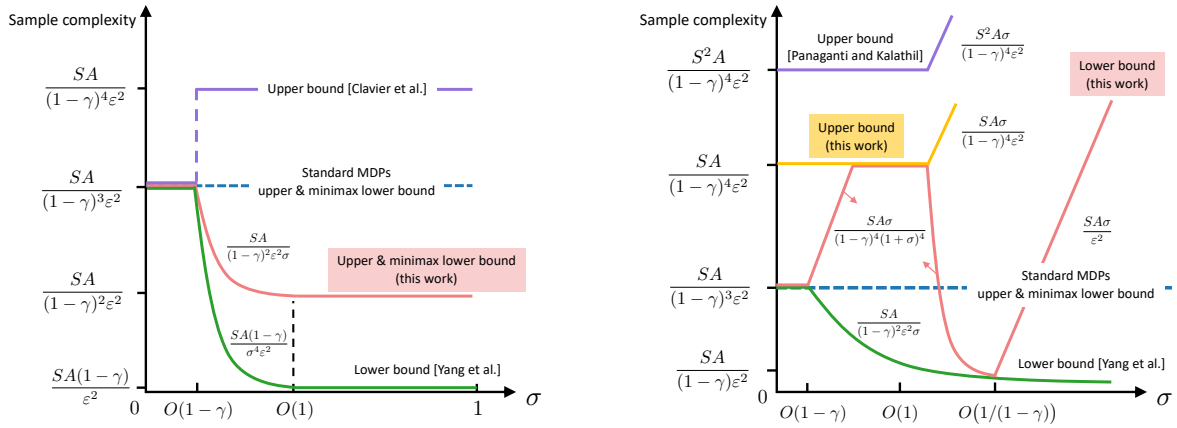
We focuses on developing strengthened sample complexity upper bounds on learning RMDPs with the TV distance and χ^2 divergence in the infinite-horizon setting (with discount factor γ), using a model-based approach called distributionally robust value iteration (DRVI) assuming access to a generative model. Improved minimax lower bounds are also developed to help gauge the tightness of our upper bounds and enable benchmarking with standard MDPs. The novel analysis framework developed herein leads to new insights into the interplay between the geometry of uncertainty sets and statistical hardness.

Sample complexity of RMDPs under the TV distance. We summarize our results and compare them with past works in Table 1.3; see Figure 1.1(a) for a graphical illustration.

- **Minimax-optimal sample complexity.** We prove that DRVI reaches ε accuracy as soon as the sample complexity is on the order of

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\varepsilon^2}\min\left\{\frac{1}{1-\gamma},\frac{1}{\sigma}\right\}\right)$$

  for all $\sigma \in (0,1)$, assuming that ε is small enough. In addition, a matching minimax lower bound (modulo some logarithmic factor) is established to guarantee the tightness of the upper



(a) TV distance (b) χ^2 divergence

Figure 1.1: Illustrations of the obtained sample complexity upper and lower bounds for learning RMDPs with comparisons to state-of-the-art and the sample complexity of standard MDPs, where the uncertainty set is specified using the TV distance (a) and the χ^2 divergence (b).

bound over the full range of the uncertainty level. To the best of our knowledge, this is the *first* minimax-optimal sample complexity for RMDPs, which was previously unavailable regardless of the divergence metric in use.

- **RMDPs are easier to learn than standard MDPs under the TV distance.** Given the sample complexity $\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$ of standard MDPs (Li et al., 2023c), it can be seen that learning RMDPs under the TV distance is never harder than learning standard MDPs; more concretely, the sample complexity for RMDPs matches that of standard MDPs when $\sigma \lesssim 1-\gamma$, and becomes smaller by a factor of $\sigma/(1-\gamma)$ when $1-\gamma \lesssim \sigma < 1$. Therefore, in this case, distributional robustness comes almost for free, given that we do not need to collect more samples.

Sample complexity of RMDPs under the χ^2 divergence. We summarize our results and provide comparisons with prior works in Table 1.4; see Figure 1.1(b) for an illustration.

- **Near-optimal sample complexity.** We demonstrate that DRVI yields ε accuracy as soon as the sample complexity is on the order of

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right)$$

for all $\sigma \in (0,\infty)$, which is the first sample complexity in this setting that scales linearly in the size S of the state space; in other words, our theory breaks the quadratic scaling bottleneck that was present in prior works (Panaganti and Kalathil, 2022; Yang et al., 2022). We have also developed a strengthened lower bound that is optimized by leveraging the geometry

Result type	Reference	Sample complexity		
		$0 < \sigma \lesssim 1-\gamma$	$1-\gamma \lesssim \sigma \lesssim \frac{1}{1-\gamma}$	$\sigma \gtrsim \frac{1}{1-\gamma}$
Upper bound	Panaganti and Kalathil (2022)	$\frac{S^2A(1+\sigma)}{(1-\gamma)^4\varepsilon^2}$		
	Yang et al. (2022)	$\frac{S^2A(1+\sigma)^2}{(\sqrt{1+\sigma}-1)^2(1-\gamma)^4\varepsilon^2}$		
	Theorem 12	$\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}$		
Lower bound	Yang et al. (2022)	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}$	
	Theorem 13	$\frac{SA}{(1-\gamma)^3\varepsilon^2}$	$\frac{SA\sigma}{(1-\gamma)^4(1+\sigma)^4\varepsilon^2}$	$\frac{SA\sigma}{\varepsilon^2}$

Table 1.4: Comparisons between our results and prior art on finding an ε-optimal robust policy in the infinite-horizon RMDPs, with the uncertainty set measured w.r.t. the χ^2 divergence. Here, S, A, γ, and $\sigma \in (0, \infty)$ are the state space size, the action space size, the discount factor, and the uncertainty level, respectively, and all logarithmic factors are omitted in the table. Improving upon all prior results, our theory is tight (up to log factors) when $\sigma \asymp 1$, and otherwise loose by no more than a polynomial factor in $1/(1-\gamma)$.

of the uncertainty set under different ranges of σ. Our theory is tight when $\sigma \asymp 1$, and is otherwise loose by at most a polynomial factor of the effective horizon $1/(1-\gamma)$ (regardless of the uncertainty level σ). This significantly improves upon prior results (as there exists an unbounded gap between prior upper and lower bounds as $\sigma \to \infty$).

- **RMDPs can be harder to learn than standard MDPs under the χ^2 divergence.** Somewhat surprisingly, our improved lower bound suggests that RMDPs in this case can be much harder to learn than standard MDPs, at least for a certain range of uncertainty levels. We single out two regimes of particular interest. Firstly, when $\sigma \asymp 1$, the sample size requirement of RMDPs is on the order of $\frac{SA}{(1-\gamma)^4\varepsilon^2}$ (up to log factor), which is provably larger than the one for standard MDPs by a factor of $\frac{1}{1-\gamma}$. Secondly, the lower bound continues to increase as σ grows and exceeds the sample complexity of standard MDPs when $\sigma \gtrsim \frac{1}{(1-\gamma)^3}$.

In sum, our sample complexity bounds not only strengthen the prior art in the development of both upper and lower bounds, but also unveil that the additional robustness consideration might affect the sample complexity in a somewhat surprising manner. As it turns out, RMDPs are not necessarily harder nor easier to learn than standard MDPs; the conclusion is far more nuanced and highly dependent on both the size and shape of the uncertainty set. This constitutes a curious phenomenon that has not been elucidated in prior analyses.

1.4.3 Our contributions to model-based robust offline RL

Two prevalent algorithmic ideas, distributional robustness and pessimism, are called out as our guiding principles. While these two ideas have been proven useful for robust RL and offline RL *separately*, tackling robust offline RL needs novel ingredients that go significantly beyond a naïve combination of existing techniques. This is because, in robust offline RL, one needs to handle the distribution shift induced not only by the behavior policy, but also by model perturbations, thus the penalty term derived from the pessimism principle in standard offline RL is no longer applicable. In short, while the value function of standard RL depends linearly with respect to the transition kernel, the dependency between the nominal transition kernel and the robust value function unfortunately becomes highly nonlinear — even without a closed-form expression — making the control of statistical uncertainty extremely challenging in robust offline RL. Altogether, this naturally leads to a question:

Can we learn a near-optimal policy which is robust with respect to uncertainties and variabilities of the environments using as few history samples as possible?

Our contributions. We provide an affirmative answer, by developing a provably efficient model-based algorithm that learns a near-optimal *distributionally-robust* policy from a minimal number of offline samples. Specifically, we consider an RMDP with S states, A actions in both the nonstationary finite-horizon setting (with horizon length H) and the discounted infinite-horizon setting (with discount factor γ), where the uncertainty set is taken as a small ball of size σ around a nominal transition kernel with respect to the Kullback-Leibler (KL) divergence. Given some history data drawn by following some behavior policy π^{b} under the nominal transition kernel in the finite-horizon (resp. infinite-horizon) setting, our goal is to learn the optimal robust policy $\pi^\star$ in the maximin sense, which has the best worst-case value for all the models within the uncertainty set (Iyengar, 2005; Nilim and El Ghaoui, 2005). Our main results are summarized below.

- We introduce a notion called *robust single-policy clipped concentrability coefficient* $C^\star_{\mathsf{rob}} \in [1/S, \infty]$ to quantify the quality of history data, which measures the distribution shift between the behavior policy π^{b} and the optimal robust policy $\pi^\star$ in the presence of model perturbations, without requiring full coverage of the entire state-action space by the behavior policy. In contrast, prior algorithms (Panaganti and Kalathil, 2022; Yang et al., 2022; Zhou et al., 2021) — using simulator or offline data — all require full coverage of the entire state-action space.

- We propose a novel pessimistic variant of distributionally robust value iteration with a plug-in estimate of the nominal transition kernel (Iyengar, 2005; Nilim and El Ghaoui, 2005), called DRVI-LCB, by penalizing the robust value estimates with a carefully designed data-driven penalty term. We demonstrate that DRVI-LCB finds an ε-optimal robust policy as soon as the sample size is above $\widetilde{O}\left(\frac{SC^\star_{\mathsf{rob}}H^5}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}\right)$ for the finite-horizon setting and $\widetilde{O}\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}\sigma^2(1-\gamma)^4\varepsilon^2}\right)$ for the

infinite-horizon setting, up to some logarithmic factor after a burn-in cost independent of ε. Here, $P^\star_{\mathsf{min}}$ is the smallest positive state transition probability of the optimal robust policy $\pi^\star$ under the nominal kernel.

- To complement the upper bound, we further develop an information-theoretic lower bound, where there exists some robust MDP such that at least $\Omega\left(\frac{SC^\star_{\mathsf{rob}}H^3}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}\right)$ samples (resp. $\Omega\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}\sigma^2(1-\gamma)^2\varepsilon^2}\right)$ samples) are needed to find an ε-optimal robust policy regardless of the choice of algorithms in the finite-horizon (resp. infinite-horizon) setting. Hence, this corroborates the near-optimality of DRVI-LCB with respect to all key parameters up to a polynomial factor of the horizon length H (resp. the effective horizon length $\frac{1}{1-\gamma}$).

To the best of our knowledge, this work is the first one to execute the principle of pessimism in a data-driven manner for robust offline RL, leading to the first provably near-optimal algorithm that learns under simultaneous model uncertainty and partial coverage of the history dataset. See Table 1.5 and Table 1.6 for summaries.

Comparison with prior art under full coverage. In truth, prior works (Panaganti and Kalathil, 2022; Yang et al., 2022; Zhou et al., 2021) have only addressed the infinite-horizon setting under full coverage of the history data. Specializing our result to this setting, DRVI-LCB finds an ε-optimal robust policy with at most $\widetilde{O}\left(\frac{SA}{P^\star_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}\right)$ samples, which depends linearly with respect to the size of the state space S (ignoring other parameters). In contrast, all prior works (Panaganti and Kalathil, 2022; Yang et al., 2022; Zhou et al., 2021) incur sample complexities that scale at least quadratically with respect to the size of the state space S. In addition, our bound improves the *exponential* dependency on $\frac{1}{1-\gamma}$ of Panaganti and Kalathil (2022); Zhou et al. (2021) to a *polynomial* dependency, as well as the *quadratic* dependency on $1/P_{\mathsf{min}}$ (which satisfies $P_{\mathsf{min}} \leq P^\star_{\mathsf{min}}$) of Yang et al. (2022) to a *linear* one on $1/P^\star_{\mathsf{min}}$. These improvements further corroborate the benefit of the proposed DRVI-LCB even under full coverage. See Table 1.6 for detailed comparisons.

1.5 Related works

We now discuss the related works of all the works proposed in this thesis. We limit our discussions primarily to RL algorithms in the tabular setting with finite state and action spaces, which are the closest to our work.

1.5.1 Online RL

Online RL and the optimism principle. The optimism principle in the face of uncertainty has received widespread adoption from bandits to online RL (Agarwal et al., 2019; Lai and Robbins, 1985; Lattimore and Szepesvári, 2020). In the context of online RL, Jaksch et al. (2010) constructed

Horizon	Algorithm	Coverage	Sample complexity
finite-horizon	DRVI-LCB **(Theorem 14)**	partial	$\frac{SC^\star_{\mathsf{rob}}H^5}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}$
	Lower bound **(Theorem 15)**	partial	$\frac{SC^\star_{\mathsf{rob}}H^3}{P^\star_{\mathsf{min}}\sigma^2\varepsilon^2}$
infinite-horizon	DRVI-LCB **(Theorem 16)**	partial	$\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}$
	Lower bound **(Theorem 17)**	partial	$\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2}$

Table 1.5: Our results for finding an ε-optimal robust policy in the finite/infinite-horizon robust MDPs with an uncertainty set measured with respect to the KL divergence using history data under partial coverage. The sample complexities included in the table are valid for sufficiently small ε, with all logarithmic factors omitted. Here, σ is the uncertainty level, $C^\star_{\mathsf{rob}}$ is the robust single-policy clipped concentrability coefficient, $P^\star_{\mathsf{min}}$ is the smallest positive state transition probability of the nominal kernel *visited by the optimal robust policy* $\pi^\star$.

confidence regions for the probability transition kernel to help select optimistic policies in the setting of weakly communicating MDPs, based on a variant (called UCRL2) of the UCRL algorithm originally proposed in Auer and Ortner (2006); see also Bourel et al. (2020); Filippi et al. (2010); Talebi and Maillard (2018) for other variants of UCRL. When applied to episodic finite-horizon MDPs, the regret bound in Jaksch et al. (2010) was suboptimal by a factor of at least $\sqrt{H^2S}$; see discussion in Azar et al. (2017); Jin et al. (2018). Fruit et al. (2020) developed an improved regret bound for UCRL2 by using empirical Bernstein-style bounds, which however was still suboptimal by a factor of at least $\sqrt{HS}$ when specialized to episodic finite-horizon MDPs. In comparison, a more sample-efficient paradigm is to build Bernstein-style upper confidence bounds (UCBs) for the optimal values to help select exploration policies, which has been recently adopted in both model-based (Azar et al., 2017) and model-free algorithms (Dong et al., 2019; Jin et al., 2018; Liu and Su, 2020; Yang et al., 2021). Note that Bernstein-style uncertainty estimation alone is not enough to ensure regret optimality in model-free algorithms, thereby motivating the design of more sophisticated variance reduction strategies (Li et al., 2023b; Zhang et al., 2020c). As alluded to previously, none of these works was able to achieve optimal sample complexity without incurring a large burn-in sample size requirement; addressing this issue requires development of a new statistical toolbox beyond what is currently available (see more details in our work (Li et al., 2023b)). Finally, the optimism principle has been studied in undiscounted infinite-horizon MDPs too (e.g., Jafarnia-Jahromi et al. (2020); Qian et al. (2019)).

Regret lower bound. Inspired by the classical lower bound argument developed for multi-armed bandits (Auer et al., 2002), the work Jaksch et al. (2010) established a regret lower bound for MDPs

Problem type	Algorithm	Coverage	Sample complexity
infinite	DRVI (Zhou et al., 2021)	full	$\frac{S^2A\exp\left(O(\frac{1}{1-\gamma})\right)}{(1-\gamma)^4\sigma^2\varepsilon^2}$
	REVI/DRVI (Panaganti and Kalathil, 2022)	full	$\frac{S^2A\exp\left(O(\frac{1}{1-\gamma})\right)}{(1-\gamma)^4\sigma^2\varepsilon^2}$
	DRVI (Yang et al., 2022)	full	$\frac{S^2A}{P_{\mathsf{min}}^2(1-\gamma)^4\sigma^2\varepsilon^2}$
	DRVI-LCB **(Theorem 16)**	full	$\frac{SA}{P_{\mathsf{min}}^\star(1-\gamma)^4\sigma^2\varepsilon^2}$

Table 1.6: Comparisons between our results and prior arts for finding an ε-optimal robust policy in the infinite/finite-horizon robust MDPs with an uncertainty set measured with respect to the KL divergence under full coverage of the history data. The sample complexities included in the table are valid for sufficiently small ε, with all logarithmic factors omitted. Here, σ is the uncertainty level, $P_{\mathsf{min}}^\star$ is the smallest positive state transition probability of the nominal kernel *visited by the optimal robust policy* $\pi^\star$, and P_{min} is the smallest positive state transition probability of the nominal kernel; it holds $P_{\mathsf{min}} \leq P_{\mathsf{min}}^\star$.

with finite diameters (so that for an arbitrary pair of states, the expected time to transition between them is assumed to be finite as long as a suitable policy is used), which has been reproduced in the note Osband and Van Roy (2016) with the purpose of facilitating comparison with Bartlett and Tewari (2009). The way to construct hard MDPs in Jaksch et al. (2010) has since been adapted by Jin et al. (2018) to exhibit a lower bound on episodic MDPs (with a sketched proof provided therein). It was recently revisited in Domingues et al. (2021), which presented a detailed and rigorous proof argument with a different construction.

1.5.2 Offline RL

Broadly speaking, at least two families of problems have been investigated in the literature that tackle offline batch data: off-policy evaluation, where the goal is to estimate the value function of a target policy that deviates from the behavior policy used in data collection; and offline policy learning, where the goal is to identify a near-optimal policy (or at least an improved one compared to the behavior policy). Our works (Li et al., 2022a; Shi et al., 2022) falls under the second category. A topic of its own interest, off-policy evaluation has been extensively studied in the recent literature; we excuse ourselves from enumerating the works in that space but only provide pointers to a few examples including Duan and Wang (2020); Duan et al. (2021); Jiang and Huang (2020); Jiang and Li (2016); Kallus and Uehara (2020); Li et al. (2014); Ren et al. (2021); Thomas and Brunskill (2016); Uehara et al. (2020); Xu et al. (2021); Yang et al. (2020).

Offline RL and the pessimism principle. One of the key challenges in offline RL lies in the insufficient coverage of the batch dataset, due to lack of interaction with the environment (Levine et al., 2020; Liu et al., 2020). To address this challenge, most of the recent works can be divided into two lines: 1) regularizing the policy to avoid visiting under-covered state and action pairs (Dadashi et al., 2021; Fujimoto et al., 2019); 2) penalizing the estimated values of the under-covered state-action pairs (Buckman et al., 2020; Kumar et al., 2020). Our work follows the latter line (also known as the principle of pessimism), which has garnered significant attention recently. In fact, pessimism has been incorporated into recent development of various offline RL approaches, such as policy-based approaches (Rezaeifar et al., 2022; Xie et al., 2021a; Zanette et al., 2021), model-based approaches (Jin et al., 2021; Kidambi et al., 2020; Rashidinejad et al., 2021; Uehara and Sun, 2021; Uehara et al., 2022; Xie et al., 2021b; Yan et al., 2022b; Yin et al., 2022; Yin and Wang, 2021; Yu et al., 2021b, 2020), and model-free approaches (Kumar et al., 2020; Shi et al., 2022; Yan et al., 2022a; Yu et al., 2021a).

In addition to the ones discussed in Chapter 1.3.2 that focus on minimax performance, The recent works Yin et al. (2022); Yin and Wang (2021) further developed instance-dependent statistical guarantees for the pessimistic model-based approach. The results in Yin and Wang (2021), however, required a large burn-in sample size, thus preventing it from attaining minimax optimality for the entire accuracy range. On the empirical side, model-based algorithms (Kidambi et al., 2020; Yu et al., 2020) have been shown to achieve superior performance than their model-free counterpart for offline RL. In addition, a number of recent works studied offline RL under various function approximation assumptions, e.g., Jin et al. (2021); Nguyen-Tang et al. (2021); Uehara and Sun (2021); Uehara et al. (2022); Yin et al. (2022); Zanette et al. (2021); Zhan et al. (2022), which are beyond the scope of the current thesis.

1.5.3 Robust RL

Robustness in RL. While standard RL has achieved remarkable success, current RL algorithms still have significant drawbacks in that the learned policy could be completely off if the deployed environment is subject to perturbation, model mismatch, or other structural changes. To address these challenges, an emerging line of works begin to address robustness of RL algorithms with respect to the uncertainty or perturbation over different components of MDPs — state, action, reward, and the transition kernel; see Moos et al. (2022) for a recent review. Besides the framework of distributionally robust MDPs (RMDPs) (Iyengar, 2005) adopted by this thesis, to promote robustness in RL, there exist various other works including but not limited to Han et al. (2022); Qiaoben et al. (2021); Sun et al. (2021); Xiong et al. (2022); Zhang et al. (2021a, 2020a) investigating the robustness w.r.t. state uncertainty, where the agent's policy is chosen based on a perturbed observation generated from the state by adding restricted noise or adversarial attack. Besides, Tan et al. (2020); Tessler et al. (2019) considered the robustness to the uncertainty of the action, namely,

the action is possibly distorted by an adversarial agent abruptly or smoothly.

Distributionally robust RL. Rooted in the literature of distributionally robust optimization, which has primarily been investigated in the context of supervised learning (Bertsimas et al., 2018; Blanchet and Murthy, 2019; Duchi and Namkoong, 2021; Gao, 2022; Rahimian and Mehrotra, 2019), distributionally robust dynamic programming and RMDPs have attracted considerable attention recently (Badrinath and Kalathil, 2021; Derman and Mannor, 2020; Goyal and Grand-Clement, 2022; Ho et al., 2018, 2021; Iyengar, 2005; Kaufman and Schaefer, 2013; Smirnova et al., 2019; Tamar et al., 2014; Wolff et al., 2012; Xu and Mannor, 2012). In the context of RMDPs, both empirical and theoretical studies have been widely conducted, although most prior theoretical analyses focus on planning with an exact knowledge of the uncertainty set (Iyengar, 2005; Tamar et al., 2014; Xu and Mannor, 2012), or are asymptotic in nature (Roy et al., 2017).

Resorting to the tools of high-dimensional statistics, various recent works begin to shift attention to understand the finite-sample performance of provable robust RL algorithms, under diverse data generating mechanisms and forms of the uncertainty set over the transition kernel, where the most related ones to ours prescribe the uncertainty set via the KL divergence, the TV distance and the χ^2 divergence. The KL divergence is a popular choice widely considered, where Blanchet et al. (2023); Panaganti and Kalathil (2022); Wang et al. (2023a); Xu et al. (2023); Yang et al. (2022); Zhou et al. (2021) investigated the sample complexity of both model-based and model-free algorithms under the simulator or offline settings. Finite-sample complexity bounds for RMDPs with the TV distance and the χ^2 divergence are developed for both the infinite-horizon setting (see Table 1.3 and Table 1.4) and the finite-horizon setting in Dong et al. (2022); Xu et al. (2023). In addition, many other forms of uncertainty sets have been considered. For example, Wang and Zou (2021) considered an R-contamination uncertain set and proposed a provable robust Q-learning algorithm for the online setting with similar guarantees as standard MDPs. Xu et al. (2023) considered a variety of uncertainty sets including one associated with Wasserstein distance. Badrinath and Kalathil (2021) considered a general (s, a)-rectangular form of the uncertainty set and proposed a model-free algorithm for the online setting with linear function approximation to cope with large state spaces. Moreover, various other related issues have been explored such as the iteration complexity of the policy-based methods (Kumar et al., 2023; Li et al., 2022b), and regularization-based robust RL (Yang et al., 2023).

1.5.4 Other related works

Model-based RL. This popular paradigm has been deployed and studied under various data collection mechanisms, including but not limited to the generative model (or simulator) setting (Agarwal et al., 2020b; Azar et al., 2013; Li et al., 2023c; Pananjady and Wainwright, 2020) that beats the state-of-the-art model-free algorithms by achieving optimality for the entire sample size

range (Li et al., 2023c), the online exploratory setting Azar et al. (2017); He et al. (2021); Jin et al. (2020), MDPs with bounded total reward (Zanette and Brunskill, 2019; Zhang et al., 2021b), and Markov games (Zhang et al., 2020b). The leave-one-out analysis (and the construction of absorbing MDPs) has been adopted by several recent works Agarwal et al. (2020b); Cui and Yang (2021); Li et al. (2023c); Pananjady and Wainwright (2020).

Model-free RL. Another widely used paradigm is model-free RL, which attempts to learn the optimal value function without explicit construction of the model. Q-learning is arguably among the most famous model-free algorithms developed in the RL literature (Jaakkola et al., 1994; Szepesvári, 1997; Tsitsiklis, 1994; Watkins and Dayan, 1992), which applies the stochastic approximation paradigm to find the fixed point of the Bellman operator and enjoys a low space complexity. Non-asymptotic sample analysis and probably approximately correct (PAC) bounds for Q-learning and its variants have seen extensive developments in the last several years, including but not limited to the works of Azar et al. (2011); Beck and Srikant (2012); Chen et al. (2020); Even-Dar and Mansour (2003); Li et al. (2023a); Wainwright (2019a); Woo et al. (2023); Xiong et al. (2020) for the synchronous setting (the case with access to a generative model or a simulator), the works of Beck and Srikant (2012); Chen et al. (2020, 2021c); Even-Dar and Mansour (2003); Li et al. (2023a, 2021); Qu and Wierman (2020); Wainwright (2019b); Woo et al. (2023); Xiong et al. (2020); Yin et al. (2021a,b) for the asynchronous setting (where one observes a single Markovian trajectory induced by a behavior policy), the works of Bai et al. (2019); Dong et al. (2019); Jafarnia-Jahromi et al. (2020); Jin et al. (2018); Li et al. (2023b); Weng et al. (2020); Yang et al. (2021); Zhang et al. (2021b, 2020c, 2021c) for the online setting via regret analysis, and the works of Shi et al. (2022); Yan et al. (2022a) for the offline setting with the access to a history dataset.

It is worthnoting that the Q-learning in the asynchronous setting shares some similarity with offline RL; note that prior results on vanilla asynchronous Q-learning require a strong uniform coverage requirement (Chen et al., 2021c; Li et al., 2023a; Qu and Wierman, 2020), which is stronger than the single-policy concentrability considered herein. Moreover, Q-learning alone is known to be sub-optimal in terms of the sample complexity in various settings (Bai et al., 2019; Jin et al., 2018; Li et al., 2023a; Shi et al., 2022; Wainwright, 2019a).

Variance reduction in RL. The seminal idea of variance reduction was originally proposed to accelerate finite-sum stochastic optimization, e.g., Gower et al. (2020); Johnson and Zhang (2013); Nguyen et al. (2017). Thereafter, the variance reduction strategy has been imported to RL, which assists in improving the sample efficiency of RL algorithms in multiple contexts, including but not limited to policy evaluation (Du et al., 2017; Khamaru et al., 2021; Wai et al., 2019; Xu et al., 2019), RL with a generative model (Sidford et al., 2018a,b; Wainwright, 2019b), asynchronous Q-learning (Li et al., 2021), and offline RL (Shi et al., 2022; Yin et al., 2021b). Note, however, variance-reduced model-free RL typically requires a large burn-in cost in order to operate in a sample-optimal fashion,

and is hence outperformed by the model-based approach under multiple sampling mechanisms.

1.6 Thesis organization and notation

Organization. The rest of this document is organized as follows.

- Chapter 2 describes the models and backgrounds of RL considered in this thesis, in particular standard MDPs and robust MDPs.
- Chapter 3 describes the proposed algorithm for online RL, together with its theoretical guarantees.
- Chapter 4 and Chapter 5 describe the proposed model-free and model-based algorithms of offline RL, respectively, together with their theoretical guarantees.
- Chapter 6 and Chapter 7 describe the proposed algorithms for robust RL with a generative model and offline data, respectively, together with their theoretical guarantees.
- Chapter 8 concludes the thesis and discusses future directions.
- The proof details are deferred to the Appendix.

Notation. Let us introduce a set of notation that will be used throughout the thesis.

- **Basic notation.** We denote by $\Delta(\mathcal{S})$ the probability simplex over a set $\mathcal{S}$, and introduce the notation $[N] \coloneqq \{1, \cdots, N\}$ for any integer $N > 0$. We adopt the convention that $0/0 = 0$. We use $\mathbb{1}(\cdot)$ to represent the indicator function. Additionally, we denote by e_i the i-th standard basis vector, with the only non-zero element being in the i-th entry and equal to 1. The KL divergence for any two distributions P and Q is denoted as $\mathsf{KL}(P \parallel Q)$.
- **Notation for vectors.** For any vector $x \in \mathbb{R}^{SA}$ (resp. $x \in \mathbb{R}^{S}$) that constitutes certain values for each of the state-action pairs (resp. state), we shall often use $x(s,a)$ (resp. $x(s)$) to denote the entry associated with the (s,a) pair (resp. state s), as long as it is clear from the context. Similarly, we shall denote by $x \coloneqq \{x_h\}_{h\in[H]}$ the set composed of certain vectors for each of the time step $h \in [H]$. In addition, we often overload scalar functions and expressions to take vector-valued arguments, with the interpretation that they are applied in an entrywise manner. For example, for a vector $x = [x_i]_{1\le i\le n}$, we have $x^2 = [x_i^2]_{1\le i\le n}$. For any two vectors $x = [x_i]_{1\le i\le n}$ and $y = [y_i]_{1\le i\le n}$, the notation $x \le y$ (resp. $x \ge y$) means $x_i \le y_i$ (resp. $x_i \ge y_i$) for all $1 \le i \le n$.
- **Big O notation.** Let $\mathcal{X} \coloneqq \big(S, A, \frac{1}{1-\gamma}, \sigma, \frac{1}{\varepsilon}, \frac{1}{\delta}\big)$. Here and throughout, we use the standard notation $f(n) = O(g(n))$ to indicate that $f(n)/g(n)$ is bounded above by a constant as

n grows. The notation $f(n) = o(g(n))$ means that $\lim_{n\to\infty} f(n)/g(n) = 0$. The notation $f(\mathcal{X}) = O(g(\mathcal{X}))$ or $f(\mathcal{X}) \lesssim g(\mathcal{X})$ indicates that there exists a universal constant $C_1 > 0$ such that $f \le C_1 g$, the notation $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ indicates that $g(\mathcal{X}) = O(f(\mathcal{X}))$, and the notation $f(\mathcal{X}) \asymp g(\mathcal{X})$ indicates that $f(\mathcal{X}) \lesssim g(\mathcal{X})$ and $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ hold simultaneously. Additionally, the notation $\widetilde{O}(\cdot)$ is defined in the same way as $O(\cdot)$ except that it hides logarithmic factors.

- **Additional notation.** Following the convention in RL (Agarwal et al., 2019), the norm $\|\cdot\|_1$ of a matrix $P = [P_{ij}]$ is defined to be $\|P\|_1 \coloneqq \max_i \sum_j |P_{ij}|$. For any vector $V = [V_i]_{1\le i\le n}$, we define its ℓ_∞ norm as $\|V\|_\infty \coloneqq \max_{1\le i\le n} |V_i|$. For any probability vector $q \in \mathbb{R}^{1\times S}$ (which is a row vector) and any vector $V \in \mathbb{R}^S$, define

$$\mathsf{Var}_q(V) \coloneqq q\big(V \circ V\big) - (qV)^2 \ \in \mathbb{R} \tag{1.7}$$

with $qV = \sum_i q_i V_i$, which corresponds to the variance of V w.r.t. the distribution q.

Chapter 2

Models and Backgrounds

2.1 Preliminaries of standard RL

In this subchapter, we introduce two widely-used models for standard RL, i.e., finite-horizon MDPs and discounted infinite-horizon MDPs, respectively.

2.1.1 Basics of episodic finite-horizon MDPs

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$ represent a finite-horizon MDP, where $\mathcal{S} \coloneqq \{1, \cdots S\}$ is the state space of size S, $\mathcal{A} \coloneqq \{1, \cdots, A\}$ is the action space of size A, H denotes the horizon length, and $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ (resp. $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$) represents the probability transition kernel (resp. reward function) at the h-th time step, $1 \le h \le H$, respectively. More specifically, $P_h(\cdot \,|\, s, a) \in \Delta(\mathcal{S})$ stands for the transition probability vector from state s at time step h when action a is taken, while $r_h(s, a)$ indicates the immediate reward received at time step h for a state-action pair (s, a) (which is assumed to be deterministic and fall within the range $[0, 1]$). The MDP is said to be non-stationary when the P_h's are not identical across $1 \le h \le H$. A policy of an agent is represented by $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h : \mathcal{S} \to \mathcal{A}$ the action selection rule at time step h, so that $\pi_h(s)$ specifies which action to execute in state s at time step h. Throughout this sub-chapter, we concentrate on deterministic policies.

Value functions, Q-functions, and Bellman equations. The value function $V_h^\pi(s)$ of a (deterministic) policy π at step h is defined as the expected cumulative rewards received between time steps h and H when executing this policy from an initial state s at time step h, namely,

$$V_h^\pi(s) \coloneqq \mathop{\mathbb{E}}_{s_{t+1} \sim P_t(\cdot | s_t, \pi_t(s_t)), t \ge h} \left[\sum_{t=h}^{H} r_t\big(s_t, \pi_t(s_t)\big) \,\Big|\, s_h = s \right], \tag{2.1}$$

where the expectation is taken over the randomness of the MDP trajectory $\{s_t \mid h \le t \le H\}$. The action-value function (a.k.a. the Q-function) $Q_h^\pi(s, a)$ of a policy π at step h can be defined analogously except that the action at step h is fixed to be a, that is,

$$Q_h^\pi(s, a) \coloneqq r_h(s, a) + \mathop{\mathbb{E}}_{\substack{s_{h+1} \sim P_h(\cdot | s, a), \\ s_{t+1} \sim P_t(\cdot | s_t, \pi_t(s_t)), t > h}} \left[\sum_{t=h+1}^{H} r_t\big(s_t, \pi_t(s_t)\big) \,\Big|\, s_h = s, a_h = a \right]. \tag{2.2}$$

In addition, we define $V_{H+1}^{\pi}(s) = Q_{H+1}^{\pi}(s,a) = 0$ for any policy π and any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. By virtue of basic properties in dynamic programming (Bertsekas, 2017), the value function and the Q-function satisfy the following Bellman equation:

$$Q_h^{\pi}(s,a) = r_h(s,a) + \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a)} \left[V_{h+1}^{\pi}(s')\right]. \tag{2.3}$$

Additionally, when the initial state is drawn from a given distribution ρ, the expected value of a given policy π and that of the optimal policy at the initial step are defined respectively by

$$V_1^{\pi}(\rho) \coloneqq \mathop{\mathbb{E}}_{s_1 \sim \rho} \left[V_1^{\pi}(s_1)\right] \qquad \text{and} \qquad V_1^{\star}(\rho) \coloneqq \mathop{\mathbb{E}}_{s_1 \sim \rho} \left[V_1^{\star}(s_1)\right]. \tag{2.4}$$

A policy $\pi^{\star} = \{\pi_h^{\star}\}_{h=1}^{H}$ is said to be an optimal policy if it maximizes the value function simultaneously for all states among all policies. The resulting optimal value function $V^{\star} = \{V_h^{\star}\}_{h=1}^{H}$ and optimal Q-functions $Q^{\star} = \{Q_h^{\star}\}_{h=1}^{H}$ satisfy

$$V_h^{\star}(s) = V_h^{\pi^{\star}}(s) = \max_{\pi} V_h^{\pi}(s) \qquad \text{and} \qquad Q_h^{\star}(s,a) = Q_h^{\pi^{\star}}(s,a) = \max_{\pi} Q_h^{\pi}(s,a) \tag{2.5}$$

for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$; here and throughout, we denote $[H] \coloneqq \{1, \cdots, H\}$. It is well known that the optimal policy always exists (Puterman, 2014), and satisfies the Bellman optimality equation:

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]: \qquad Q_h^{\star}(s,a) = r_h(s,a) + \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a)} \left[V_{h+1}^{\star}(s')\right]. \tag{2.6}$$

Before proceeding, we shall also let

$$P_{h,s,a} = P_h(\cdot \,|\, s,a) \in \mathbb{R}^{1 \times S} \tag{2.7}$$

abbreviate the transition probability vector given the (s,a) pair at time step h.

2.1.2 Basics of discounted infinite-horizon MDPs

Consider a discounted infinite-horizon MDP (Bertsekas, 2017) represented by a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, \gamma, r\}$. The key components of $\mathcal{M}$ are: (i) $\mathcal{S} = \{1, 2, \cdots, S\}$: a finite state space of size S; (ii) $\mathcal{A} = \{1, 2, \cdots, A\}$: an action space of size A; (iii) $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: the transition probability kernel of the MDP (i.e., $P(\cdot \,|\, s,a)$ denotes the transition probability from state s when action a is executed); (iv) $\gamma \in [0,1)$: the discount factor, so that $\frac{1}{1-\gamma}$ represents the effective horizon; (v) $r : \mathcal{S} \times \mathcal{A} \to [0,1]$: the deterministic reward function (namely, $r(s,a)$ indicates the immediate reward received when the current state-action pair is (s,a)). Without loss of generality, the immediate rewards are normalized so that they are contained within the interval $[0,1]$. Throughout, we

introduce the convenient notation

$$P_{s,a} \coloneqq P(\cdot \,|\, s,a) \in \mathbb{R}^{1\times S}. \tag{2.8}$$

Policy, value function and Q-function. A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a possibly randomized action selection rule; that is, $\pi(a \,|\, s)$ represents the probability of choosing a in state s. When π is a deterministic policy, we abuse the notation by letting $\pi(s)$ represent the action chosen by the policy π in state s. A sample trajectory induced by the MDP under policy π can be written as $\{(s_t, a_t)\}_{t\geq 0}$, with s_t (resp. a_t) denoting the state (resp. action) of the trajectory at time t. To proceed, we shall also introduce the value function V^π and Q-value function Q^π associated with policy π. Specifically, the value function $V^\pi : \mathcal{S} \to \mathbb{R}$ of policy π is defined as the expected discounted cumulative reward as follows:

$$\forall s \in \mathcal{S}: \qquad V^\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s; \pi\right], \tag{2.9}$$

where the expectation is taken over the sample trajectory $\{(s_t, a_t)\}_{t\geq 0}$ generated in a way that $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ for all $t \geq 0$. Given that all immediate rewards lie within $[0, 1]$, it is easily verified that $0 \leq V^\pi(s) \leq \frac{1}{1-\gamma}$ for any policy π. The Q-function (or action-state function) of policy π can be defined analogously as follows:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad Q^\pi(s,a) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a; \pi\right], \tag{2.10}$$

which differs from (2.9) in that it is also conditioned on $a_0 = a$.

Let $\rho \in \Delta(\mathcal{S})$ be a given state distribution. If the initial state is randomly drawn from ρ, then we can define the following weighted value function of policy π:

$$V^\pi(\rho) \coloneqq \mathop{\mathbb{E}}_{s\sim\rho}\big[V^\pi(s)\big]. \tag{2.11}$$

2.2 Preliminaries of robust RL

Abusing the notation in standard RL, we introduce two models of robust RL — finite-horizon robust MDPs and discounted infinite-horizon robust MDPs (RMDPs), respectively.

2.2.1 Basics of episodic finite-horizon RMDPs

Recall that $\pi = \{\pi_h\}_{h=1}^H$ is the policy or action selection rule of an agent, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the action selection probability over the action space. Slightly abusing the notation, the

value function $V^{\pi,P} = \{V_h^{\pi,P}\}_{h=1}^H$ of policy π with a transition kernel P is defined by

$$\forall (h,s) \in [H] \times \mathcal{S}: \qquad V_h^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_t\big(s_t,a_t\big) \,\Big|\, s_h = s\right], \tag{2.12}$$

where the expectation is taken over the randomness of the trajectory $\{s_h, a_h, r_h\}_{h=1}^H$ generated by executing policy π, namely, $a_t \sim \pi_t(s_t)$, and $s_{t+1} \sim P_t(\cdot \,|\, s_t, a_t)$. Similarly, the Q-function $Q^{\pi,P} = \{Q_h^{\pi,P}\}_{h=1}^H$ of policy π is defined as

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad Q_h^{\pi,P}(s,a) \coloneqq r_h(s,a) + \mathbb{E}_{\pi,P}\left[\sum_{t=h+1}^{H} r_t(s_t,a_t) \,\Big|\, s_h = s, a_h = a\right], \tag{2.13}$$

where the expectation is again taken over the randomness of the trajectory.

Moreover, when the initial state s_1 is drawn from a given distribution φ, let $d_h^{\pi,P}(s \,|\, \varphi)$ and $d_h^{\pi,P}(s,a \,|\, \varphi)$ denote respectively the state occupancy distribution and the state-action occupancy distribution induced by π at time step $h \in [H]$, i.e.,

$$\forall (h,s) \in [H] \times \mathcal{S}: \qquad d_h^{\pi,P}(s) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \varphi, \pi, P), \tag{2.14a}$$

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad d_h^{\pi,P}(s,a) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \varphi, \pi, P)\, \pi_h(a \,|\, s), \tag{2.14b}$$

which are conditioned on $s_1 \sim \varphi$ and the event that all actions and states are drawn according to policy π and transition kernel P. In particular, we often drop the dependency with respect to φ whenever it is clear from the context, by simply writing $d_h^{\pi,P}(s) \coloneqq d_h^{\pi,P}(s \,|\, \varphi)$ and $d_h^{\pi,P}(s,a) \coloneqq d_h^{\pi,P}(s,a \,|\, \varphi)$.

Finite-horizon distributionally robust MDPs. Consider a finite-horizon distributionally robust MDP (RMDPs), denoted by $\mathcal{M}_{\mathsf{rob}} = \big(\mathcal{S}, \mathcal{A}, H, \mathcal{U}_\rho^\sigma(P^0), \{r_h\}_{h=1}^H\big)$. Different from standard MDPs, we now consider an ensemble of probability transition kernels or models within an uncertainty set centered around a nominal one $P^0 = \{P_h^0\}_{h=1}^H$, where the distance between the transition kernels is specified using some distance metric ρ of radius $\sigma > 0$. Specifically, the uncertainty set around P^0 with the divergence metric $\rho : \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \to \mathbb{R}^+$ is specified as

$$\mathcal{U}_\rho^\sigma(P^0) \coloneqq \otimes\, \mathcal{U}_\rho^\sigma(P_{h,s,a}^0) \qquad \text{with} \quad \mathcal{U}_\rho^\sigma(P_{h,s,a}^0) \coloneqq \big\{P_{h,s,a} \in \Delta(\mathcal{S}) : \rho\left(P_{h,s,a}, P_{h,s,a}^0\right) \le \sigma\big\}, \tag{2.15}$$

where $\otimes$ denote the Cartesian product. In words, the divergence between the true transition probability vector and the nominal one at each state-action pair is at most σ; moreover, the RMDP reduces to the standard MDP when $\sigma = 0$.

Instead of evaluating a policy in a fixed MDP, the performance of a policy in the RMDP is evaluated based on its worst-case — i.e., smallest — value function over all the instances in the

uncertainty set. That is, we define the *robust value function* $V^{\pi,\sigma} = \{V_h^{\pi,\sigma}\}_{h=1}^H$ and the *robust Q-function* $Q^{\pi,\sigma} = \{Q_h^{\pi,\sigma}\}_{h=1}^H$ respectively as

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad V_h^{\pi,\sigma}(s) \coloneqq \inf_{P \in \mathcal{U}_\rho^\sigma(P^0)} V_h^{\pi,P}(s), \qquad Q_h^{\pi,\sigma}(s,a) \coloneqq \inf_{P \in \mathcal{U}_\rho^\sigma(P^0)} Q_h^{\pi,P}(s,a),$$

where the infimum is taken over the uncertainty set of transition kernels.

Optimal robust policy and the robust Bellman operator. For finite-horizon RMDPs, it has been established that there exists at least one deterministic policy that maximizes the robust value function and Q-function simultaneously (Iyengar, 2005; Nilim and El Ghaoui, 2005). In view of this, we shall denote a deterministic policy $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ as an optimal robust policy throughout this chapter. The resulting *optimal robust value function* $V^{\star,\sigma} = \{V_h^{\star,\sigma}\}_{h=1}^H$ and *optimal robust Q-function* $Q^{\star,\sigma} = \{Q_h^{\star,\sigma}\}_{h=1}^H$ are denoted by

$$\forall (h,s) \in [H] \times \mathcal{S}: \quad V_h^{\star,\sigma}(s) \coloneqq V_h^{\pi^\star,\sigma}(s) = \max_\pi V_h^{\pi,\sigma}(s), \tag{2.16a}$$

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \quad Q_h^{\star,\sigma}(s,a) \coloneqq Q_h^{\pi^\star,\sigma}(s,a) = \max_\pi Q_h^{\pi,\sigma}(s,a). \tag{2.16b}$$

Similar to (4.1), we adopt the following short-hand notation for the occupancy distributions associated with the optimal policy:

$$\forall (h,s) \in [H] \times \mathcal{S}: \qquad d_h^{\star,P}(s) \coloneqq d_h^{\pi^\star,P}(s), \tag{2.17a}$$

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \qquad d_h^{\star,P}(s,a) \coloneqq d_h^{\pi^\star,P}(s,a) = d_h^{\star,P}(s)\, \mathbb{1}\{a = \pi_h^\star(s)\}. \tag{2.17b}$$

It turns out the Bellman's principle of optimality can be extended naturally to its robust counterpart (Iyengar, 2005; Nilim and El Ghaoui, 2005), which plays a fundamental role in solving the RMDP. To begin with, for any policy π, the robust value function and robust Q-function satisfy the following *robust Bellman consistency equation*:

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \qquad Q_h^{\pi,\sigma}(s,a) = r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^0_{h,s,a})} \mathcal{P} V_{h+1}^{\pi,\sigma}. \tag{2.18}$$

Additionally, the optimal robust Q-function obeys the *robust Bellman optimality equation*:

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \qquad Q_h^{\star,\sigma}(s,a) = r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^0_{h,s,a})} \mathcal{P} V_{h+1}^{\star,\sigma}, \tag{2.19}$$

which can be solved efficiently via a robust variant of value iteration when the RMDP is known (Iyengar, 2005; Nilim and El Ghaoui, 2005).

2.2.2 Basics of discounted infinite-horizon RMDPs

We now turn to the definition of discounted infinite-horizon RMDPs. To characterize the cumulative reward, with a slight abuse of the notation, the value function $V^{\pi,P}$ for any policy π under the transition kernel P is defined by

$$\forall s \in \mathcal{S}: \qquad V^{\pi,P}(s) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s\right], \tag{2.20}$$

where the expectation is taken over the randomness of the trajectory $\{s_t, a_t\}_{t=0}^{\infty}$ generated by executing policy π under the transition kernel P, namely, $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ for all $t \geq 0$. Similarly, the Q-function $Q^{\pi,P}$ associated with any policy π under the transition kernel P is defined as

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^{\pi,P}(s,a) \coloneqq \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r\big(s_t, a_t\big) \,\Big|\, s_0 = s, a_0 = a\right], \tag{2.21}$$

where the expectation is again taken over the randomness of the trajectory under policy π.

Letting φ be some initial state distribution, we denote $d^{\pi,P}(s \,|\, \varphi)$ and $d^{\pi,P}(s,a \,|\, \varphi)$ respectively as the state occupancy distribution and the state-action occupancy distribution induced by policy π, namely

$$\forall s \in \mathcal{S}: \qquad d^{\pi,P}(s) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \,|\, s_0 \sim \varphi, \pi, P), \tag{2.22a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad d^{\pi,P}(s,a) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \,|\, s_0 \sim \varphi, \pi, P)\, \pi(a \,|\, s). \tag{2.22b}$$

Here, the occupancy distributions are conditioned on $s_0 \sim \varphi$ and the sequence of actions and states are generated based on policy π and transition kernel P.

Discounted infinite-horizon distributionally robust MDPs. We now introduce the distributionally robust MDP (RMDP) tailored to the discounted infinite-horizon setting, denoted by $\mathcal{M}_{\mathsf{rob}} = \{\mathcal{S}, \mathcal{A}, \gamma, \mathcal{U}_\rho^\sigma(P^0), r\}$, where $\mathcal{S}, \mathcal{A}, \gamma, r$ are identical to those in the standard MDP. A key distinction from the standard MDP is that: rather than assuming a fixed transition kernel P, it allows the transition kernel to be chosen arbitrarily from a prescribed uncertainty set $\mathcal{U}_\rho^\sigma(P^0)$ centered around a *nominal* kernel $P^0 : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where the uncertainty set is specified using some distance metric ρ of radius $\sigma > 0$. In particular, given the nominal transition kernel P^0 and some uncertainty level σ, the uncertainty set — with the divergence metric $\rho : \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \to \mathbb{R}^+$

— is specified as

$$\mathcal{U}_\rho^\sigma(P^0) \coloneqq \otimes\, \mathcal{U}_\rho^\sigma(P^0_{s,a}) \qquad \text{with} \quad \mathcal{U}_\rho^\sigma(P^0_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \rho\left(P_{s,a}, P^0_{s,a}\right) \le \sigma\right\}, \tag{2.23}$$

where we recall that a vector of the transition kernel P or P^0 at state-action pair (s,a) is denoted respectively as

$$P_{s,a} \coloneqq P(\cdot \,|\, s,a) \in \mathbb{R}^{1\times S}, \qquad P^0_{s,a} \coloneqq P^0(\cdot \,|\, s,a) \in \mathbb{R}^{1\times S}. \tag{2.24}$$

In other words, the uncertainty is imposed in a decoupled manner for each state-action pair, obeying the so-called (s,a)-rectangularity (Wiesemann et al., 2013; Zhou et al., 2021).

In RMDPs, we are interested in the worst-case performance of a policy π over all the possible transition kernels in the uncertainty set. This is measured by the *robust value function* $V^{\pi,\sigma}$ and the *robust Q-function* $Q^{\pi,\sigma}$ in $\mathcal{M}_{\mathsf{rob}}$, defined respectively as

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad V^{\pi,\sigma}(s) \coloneqq \inf_{P\in\mathcal{U}_\rho^\sigma(P^0)} V^{\pi,P}(s), \qquad Q^{\pi,\sigma}(s,a) \coloneqq \inf_{P\in\mathcal{U}_\rho^\sigma(P^0)} Q^{\pi,P}(s,a). \tag{2.25}$$

Optimal robust policy and the robust Bellman operator. As a generalization of properties of standard MDPs, it is well-known that there exists at least one deterministic policy that maximizes the robust value function (resp. robust Q-function) simultaneously for all states (resp. state-action pairs) (Iyengar, 2005; Nilim and El Ghaoui, 2005). Therefore, we denote the *optimal robust value function* (resp. *optimal robust Q-function*) as $V^{\star,\sigma}$ (resp. $Q^{\star,\sigma}$), and the optimal robust policy as $\pi^\star$, which satisfy

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) \coloneqq V^{\pi^\star,\sigma}(s) = \max_\pi V^{\pi,\sigma}(s), \tag{2.26a}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^{\star,\sigma}(s,a) \coloneqq Q^{\pi^\star,\sigma}(s,a) = \max_\pi Q^{\pi,\sigma}(s,a). \tag{2.26b}$$

A key machinery in RMDPs is a generalization of Bellman's optimality principle, encapsulated in the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*):

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}_\rho^\sigma(P^0_{s,a})} \mathcal{P}V^{\pi,\sigma}, \tag{2.27a}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}_\rho^\sigma(P^0_{s,a})} \mathcal{P}V^{\star,\sigma}. \tag{2.27b}$$

Applying (2.22) with $\pi = \pi^\star$, we adopt the the following short-hand notation for the occupancy distributions associated with the optimal policy:

$$\forall s \in \mathcal{S}: \qquad d^{\star,P}(s) \coloneqq d^{\pi^\star,P}(s), \tag{2.28a}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \qquad d^{\star,P}(s,a) \coloneqq d^{\pi^\star,P}(s,a) = d^{\star,P}(s)\,\mathbb{1}\{a = \pi^\star(s)\}. \tag{2.28b}$$

The robust Bellman operator (Iyengar, 2005; Nilim and El Ghaoui, 2005) is denoted by $\mathcal{T}^\sigma(\cdot) : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$ and defined as follows:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad \mathcal{T}^\sigma(Q)(s,a) \coloneqq r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^0_{s,a})} \mathcal{P}V, \quad \text{with} \quad V(s) \coloneqq \max_a Q(s,a). \tag{2.29}$$

Given that $Q^{\star,\sigma}$ is the unique fixed point of $\mathcal{T}^\sigma$, one can recover the optimal robust value function and Q-function using a procedure termed *distributionally robust value iteration* (DRVI). Generalizing the standard value iteration, DRVI starts from some given initialization and recursively applies the robust Bellman operator until convergence. As has been shown previously, this procedure converges rapidly due to the γ-contraction property of $\mathcal{T}^\sigma$ w.r.t. the ℓ_∞ norm (Iyengar, 2005; Nilim and El Ghaoui, 2005).

Chapter 3

Model-Free Online RL

3.1 Problem formulation

In this chapter, we investigate the online episodic finite-horizon RL setting, where the agent is allowed to execute the MDP sequentially in a total number of K episodes each of length H. This amounts to collecting

$$T = KH \text{ samples}$$

in total. More specifically, in each episode $k = 1, \ldots, K$, the agent is assigned an arbitrary initial state s_1^k (possibly by an adversary), and selects a policy $\pi^k = \{\pi_h^k\}_{h=1}^H$ learned based on the information collected up to the $(k-1)$-th episode. The k-th episode is then executed following the policy π^k and the dynamic of the MDP $\mathcal{M}$, leading to a length-H sample trajectory.

Goal: regret minimization. In order to evaluate the quality of the learned policies $\{\pi^k\}_{1\le k\le K}$, a frequently used performance metric is the cumulative regret defined as follows:

$$\mathsf{Regret}(T) \coloneqq \sum_{k=1}^{K} \big(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\big). \tag{3.1}$$

In words, the regret reflects the sub-optimality gaps between the values of the optimal policy and those of the learned policies aggregated over K episodes. A natural objective is thus to design a sample-optimal algorithm, namely, an algorithm whose resulting regret scales optimally in the sample size T. Accomplishing this goal requires carefully managing the trade-off between exploration and exploitation, which is particularly challenging in the sample-limited regime.

3.2 Algorithm and theory

In this chapter, we present the proposed algorithm called CB-Q-Advantage, as well as the accompanying theory confirming its sample and memory efficiency.

3.2.1 Review: Q-learning with UCB exploration and reference advantage

This subchapter briefly reviews the Q-learning algorithm with UCB exploration proposed in Jin et al. (2018), as well as a variant that further exploits the idea of variance reduction (Zhang et al.,

2020c). These two model-free algorithms inspire the algorithm design in the current chapter.

Q-learning with UCB exploration (UCB-Q or UCB-Q-Hoeffding). Recall that the classical Q-learning algorithm has been proposed as a stochastic approximation scheme (Robbins and Monro, 1951) to solve the Bellman optimality equation (2.6), which consists of the following update rule (Watkins and Dayan, 1992; Watkins, 1989):

$$Q_h(s,a) \leftarrow (1-\eta)Q_h(s,a) + \eta\Big\{r_h(s,a) + \underbrace{\widehat{P}_{h,s,a}V_{h+1}}_{\text{stochastic estimate of } P_{h,s,a}V_{h+1}}\Big\}. \tag{3.2}$$

Here, Q_h (resp. V_h) indicates the running estimate of $Q_h^\star$ (resp. $V_h^\star$), η is the (possibly iteration-varying) learning rate or stepsize, and $\widehat{P}_{h,s,a}V_{h+1}$ is a stochastic estimate of $P_{h,s,a}V_{h+1}$ (cf. (2.7)). For instance, if one has available a sample (s,a,s') transitioning from state s at step h to s' at step $h+1$ under action a, then a stochastic estimate of $P_{h,s,a}V_{h+1}$ can be taken as $V_{h+1}(s')$, which is unbiased in the sense that

$$\mathbb{E}\big[V_{h+1}(s')\big] = P_{h,s,a}V_{h+1}.$$

To further encourage exploration, the algorithm proposed in Jin et al. (2018) — which shall be abbreviated as UCB-Q or UCB-Q-Hoeffding hereafter — augments the Q-learning update rule (3.2) in each episode via an additional exploration bonus:

$$Q_h^{\mathsf{UCB}}(s,a) \leftarrow (1-\eta)Q_h^{\mathsf{UCB}}(s,a) + \eta\big\{r_h(s,a) + \widehat{P}_{h,s,a}V_{h+1} + b_h\big\}. \tag{3.3}$$

The bonus term $b_h \geq 0$ is chosen to be a certain upper confidence bound for $(\widehat{P}_{h,s,a} - P_{h,s,a})V_{h+1}$, an exploration-efficient scheme that originated from the bandit literature (Lai and Robbins, 1985; Lattimore and Szepesvári, 2020). The algorithm then proceeds to the next episode by executing/sampling the MDP using a greedy policy w.r.t. the updated Q-estimate. These steps are repeated until the algorithm is terminated.

Q-learning with UCB exploration and reference advantage (UCB-Q-Advantage). The regret bounds derived for UCB-Q (Jin et al., 2018), however, fall short of being optimal, as they are at least a factor of $\sqrt{H}$ away from the fundamental lower bound. In order to further shave this $\sqrt{H}$ factor, one strategy is to leverage the idea of variance reduction to accelerate convergence (Johnson and Zhang, 2013; Li et al., 2021; Sidford et al., 2018b; Wainwright, 2019b). An instantiation of this idea for the regret setting is a variant of UCB-Q based on reference-advantage decomposition, which was put forward in Zhang et al. (2020c) and shall be abbreviated as UCB-Q-Advantage throughout this chapter.

To describe the key ideas of UCB-Q-Advantage, imagine that we are able to maintain a collection of reference values $V^{\mathsf{R}} = \{V_h^{\mathsf{R}}\}_{h=1}^H$, which form reasonable estimates of $V^\star = \{V_h^\star\}_{h=1}^H$

and become increasingly more accurate as the algorithm progresses.

At each time step h, the algorithm adopts the following update rule

$$Q_h^{\mathsf{R}}(s,a) \leftarrow (1-\eta)Q_h^{\mathsf{R}}(s,a) + \eta\Big\{r_h(s,a) + \underbrace{\widehat{P}_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathsf{R}}\big)}_{\text{stochastic estimate of } P_{h,s,a}\left(V_{h+1}-V_{h+1}^{\mathsf{R}}\right)} + \widehat{[P_h V_{h+1}^{\mathsf{R}}]}(s,a) + b_h^{\mathsf{R}}\Big\}. \tag{3.4}$$

Two ingredients of this update rule are worth noting.

- Akin to the UCB-Q case, we can take $\widehat{P}_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathsf{R}}\big)$ to be the stochastic estimate $V_{h+1}(s') - V_{h+1}^{\mathsf{R}}(s')$ if we observe a sample transition (s,a,s') at time step h. If V_{h+1} is fairly close to the reference V_{h+1}^{R}, then this stochastic term can be less volatile than the stochastic term $\widehat{P}_{h,s,a}V_{h+1}$ in (3.3).
- Additionally, the term $\widehat{P_h V_{h+1}^{\mathsf{R}}}$ indicates an estimate of the one-step look-ahead value $P_h V_{h+1}^{\mathsf{R}}$, which shall be computed using a batch of samples.

  The variability of $\widehat{P_h V_{h+1}^{\mathsf{R}}}$ can be well-controlled through the use of batch data, at the price of an increased sample size.

Accordingly, the exploration bonus term b_h^{R} is taken to be an upper confidence bound for the above-mentioned two terms combined. Given that the uncertainty of (3.4) largely stems from these two terms (which can both be much smaller than the variability in (3.3)), the incorporation of the reference term helps accelerate convergence.

3.2.2 The proposed algorithm: CB-Q-Advantage

As alluded to previously, however, the sample size required for UCB-Q-Advantage to achieve optimal regret needs to exceed a large polynomial S^6A^4 in the size of the state/action space. To overcome this sample complexity barrier, we come up with an improved variant called CB-Q-Advantage.

Motivation: early settlement of a reference value. An important insight obtained from previous algorithm designs is that: in order to achieve low regret, it is desirable to maintain an estimate of Q-function that (i) provides an optimistic view (namely, an over-estimate) of the truth $Q^\star$, and (ii) mitigates the bias $Q - Q^\star$ as much as possible. With two additional optimistic Q-estimates in hand — Q_h^{UCB} based on UCB-Q, and Q_h^{R} based on the reference-advantage decomposition — it is natural to combine them as follows to further reduce the bias without violating the optimism principle:

$$Q_h(s_h,a_h) \leftarrow \min\Big\{Q_h^{\mathsf{R}}(s_h,a_h),\, Q_h^{\mathsf{UCB}}(s_h,a_h),\, Q_h(s_h,a_h)\Big\}. \tag{3.5}$$

This is precisely what is conducted in UCB-Q-Advantage. However, while the estimate Q_h^{R} obtained with the aid of reference-advantage decomposition provides great promise, fully realizing its potential

in the sample-limited regime relies on the ability to quickly *settle on* a desirable "reference" during the initial stage of the algorithm. This leads us to a dilemma that requires careful thinking. On the one hand, the reference value V^{R} needs to be updated in a timely manner in order to better control the stochastic estimate of $P_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathsf{R}}\big)$. On the other hand, updating V^{R} too frequently incurs an overly large sample size burden, as new samples need to be accumulated whenever V^{R} is updated.

Built upon the above insights, it is advisable to prevent frequent updating of the reference value V^{R}. In fact, it would be desirable to stop updating the reference value once a point of sufficient quality — denoted by $V^{\mathsf{R,final}}$ — has been obtained. Locking on a reasonable reference value early on means that (a) fewer samples will be wasted on estimating a drifting target $P_h V_{h+1}^{\mathsf{R}}$, and (b) all ensuing samples can then be dedicated to estimating the key quantity of interest $P_h V_{h+1}^{\mathsf{R,final}}$.

Remark 1. In Zhang et al. (2020c), the algorithm UCB-Q-Advantage requires collecting $\widetilde{O}\big(SAH^6\big)$ samples *for each state* before settling on the reference value, which inevitably contributes to the large burn-in cost.

The proposed CB-Q-Advantage algorithm. We now propose a new model-free algorithm that allows for early settlement of the reference value. A few key ingredients are as follows.

- *An auxiliary sequence based on LCB.* In addition to the two optimistic Q-estimates Q_h^{R} and Q_h^{UCB} described previously, we intend to maintain another *pessimistic* estimate $Q_h^{\mathsf{LCB}} \le Q_h^\star$ using the subroutine `update-lcb-q`, based on lower confidence bounds (LCBs). We will also maintain the corresponding value function V_h^{LCB}, which lower bounds $V_h^\star$.

- *Termination rules for reference updates.* With $V_h^{\mathsf{LCB}} \le V_h^\star$ in place, the updates of the references (lines 15-18 of Algorithm 1) are designed to terminate when

$$V_h(s_h) \le V_h^{\mathsf{LCB}}(s_h) + 1 \le V_h^\star(s_h) + 1. \tag{3.6}$$

  Note that V_h^{R} keeps tracking the value of V_h before it stops being updated. In effect, when the additional condition in lines 15 is violated and thus (3.6) is satisfied, we claim that it is unnecessary to update the reference V_h^{R} afterwards, since it is of sufficient quality (being close enough to the optimal value $V_h^\star$) and further drifting the reference does not appear beneficial. As we will make it rigorous shortly, this reference update rule is sufficient to ensure that $|V_h - V_h^{\mathsf{R}}| \le 2$ throughout the execution of the algorithm, which in turn suggests that the standard deviation of $\widehat{P}_{h,s,a}(V_{h+1} - V_{h+1}^{\mathsf{R}})$ might be $O(H)$ times smaller than that of $\widehat{P}_{h,s,a} V_{h+1}$ (i.e., the stochastic term used in (3.2) of UCB-Q). This is a key observation that helps shave the addition factor H in the regret bound of UCB-Q.

- *Update rules for Q_h^{UCB} and Q_h^{R}.* The two optimistic Q-estimates Q_h^{UCB} and Q_h^{R} are updated using the subroutine `update-ucb-q` (following the standard Q-learning with Hoeffding bonus

Algorithm 1: CB-Q-Advantage

Parameters: some universal constant $c_{\mathrm{b}} > 0$ and probability of failure $\delta \in (0,1)$;
Initialize $Q_h(s,a), Q_h^{\mathsf{UCB}}(s,a), Q_h^{\mathsf{R}}(s,a) \leftarrow H$; $V_h(s), V_h^{\mathsf{R}}(s) \leftarrow H$; $Q_h^{\mathsf{LCB}}(s,a) \leftarrow 0$; $V_h^{\mathsf{LCB}}(s) \leftarrow 0$; $N_h(s,a) \leftarrow 0$;
$\mu_h^{\mathsf{ref}}(s,a), \sigma_h^{\mathsf{ref}}(s,a), \mu_h^{\mathsf{adv}}(s,a), \sigma_h^{\mathsf{adv}}(s,a), \delta_h^{\mathsf{R}}(s,a), B_h^{\mathsf{R}}(s,a) \leftarrow 0$; and $u_{\mathrm{ref}}(s) = \mathsf{True}$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.
for *Episode* $k = 1$ **to** K **do**
  Set initial state $s_1 \leftarrow s_1^k$.
  for *Step* $h = 1$ **to** H **do**
    Take action $a_h = \pi_h^k(s_h) = \arg\max_a Q_h(s_h,a)$, and draw $s_{h+1} \sim P_h(\cdot \,|\, s_h, a_h)$. `// sampling`
    $N_h(s_h,a_h) \leftarrow N_h(s_h,a_h) + 1$; $n \leftarrow N_h(s_h,a_h)$. `// update the counter`
    $\eta_n \leftarrow \frac{H+1}{H+n}$. `// update the learning rate`
    $Q_h^{\mathsf{UCB}}(s_h,a_h) \leftarrow$ `update-ucb-q()`. `// run UCB-Q; see Algorithm 6`
    $Q_h^{\mathsf{LCB}}(s_h,a_h) \leftarrow$ `update-lcb-q()`. `// run LCB-Q; see Algorithm 6`
    $Q_h^{\mathsf{R}}(s_h,a_h) \leftarrow$ `update-ucb-q-advantage()`. `// estimate` Q_h^{R} `; see Algorithm 6`
    `/* update Q-estimates using all estimates in hand, and update value estimates */`
    $Q_h(s_h,a_h) \leftarrow \min\{Q_h^{\mathsf{R}}(s_h,a_h),\, Q_h^{\mathsf{UCB}}(s_h,a_h),\, Q_h(s_h,a_h)\}$.
    $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$.
    $V_h^{\mathsf{LCB}}(s_h) \leftarrow \max\{\max_a Q_h^{\mathsf{LCB}}(s_h,a), V_h^{\mathsf{LCB}}(s_h)\}$.
    `/* update reference values */`
    if $V_h(s_h) - V_h^{\mathsf{LCB}}(s_h) > 1$ **then**
      $V_h^{\mathsf{R}}(s_h) \leftarrow V_h(s_h)$.
    else if $u_{\mathrm{ref}}(s_h) = \mathsf{True}$ **then**
      $V_h^{\mathsf{R}}(s_h) \leftarrow V_h(s_h), \qquad u_{\mathrm{ref}}(s_h) = \mathsf{False}$.

(Jin et al., 2018)) and `update-ucb-q-advantage`, respectively. Note that Q_h^{R} continues to be updated even after V_h^{R} is no longer updated.

Q-learning with reference-advantage decomposition. The rest of this subchapter is devoted to explaining the subroutine `update-ucb-q-advantage`, which produces a Q-estimate Q^{R} based on the reference-advantage decomposition similar to Zhang et al. (2020c). To facilitate the implementation, let us introduce the parameters associated with a reference value V^{R}, which include six different components, i.e.,

$$\big[\mu_h^{\mathsf{ref}}(s,a), \sigma_h^{\mathsf{ref}}(s,a), \mu_h^{\mathsf{adv}}(s,a), \sigma_h^{\mathsf{adv}}(s,a), \delta_h^{\mathsf{R}}(s,a), B_h^{\mathsf{R}}(s,a)\big], \tag{3.7}$$

for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Here $\mu_h^{\mathsf{ref}}(s,a)$ and $\sigma_h^{\mathsf{ref}}(s,a)$ estimate the running mean and 2nd moment of the reference $\big[P_h V_{h+1}^{\mathsf{R}}\big](s,a)$; $\mu_h^{\mathsf{adv}}(s,a)$ and $\sigma_h^{\mathsf{adv}}(s,a)$ estimate the running (weighted)

mean and 2nd moment of the advantage $\big[P_h(V_{h+1} - V_{h+1}^{\mathsf{R}})\big](s,a)$; $B_h^{\mathsf{R}}(s,a)$ aggregates the empirical standard deviations of the reference and the advantage combined; and last but not least, $\delta_h^{\mathsf{R}}(s,a)$ is the temporal difference between $B_h^{\mathsf{R}}(s,a)$ and its previous value.

As alluded to previously, the Q-function estimation follows the strategy (3.4) at a high level. Upon observing a sample transition (s_h, a_h, s_{h+1}), we compute the following estimates to update $Q^{\mathsf{R}}(s_h, a_h)$.

- The term $\widehat{P}_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathsf{R}}\big)$ is set to be $V_{h+1}(s_{h+1}) - V_{h+1}^{\mathsf{R}}(s_{h+1})$, which is an unbiased stochastic estimate of $P_{h,s,a}\big(V_{h+1} - V_{h+1}^{\mathsf{R}}\big)$.
- The term $\big[P_h V_{h+1}^{\mathsf{R}}\big](s,a)$ is estimated via $\mu_h^{\mathsf{ref,R}}$ (cf. line 9). Given that this is estimated using all previous samples, we expect the variability of this term to be well-controlled as the sample size increases (especially after V^{R} is locked).
- The exploration bonus $b_h^{\mathsf{R}}(s,a)$ is updated using $B_h^{\mathsf{R}}(s_h,a_h)$ and $\delta_h^{\mathsf{R}}(s_h,a_h)$ (cf. lines 5-6 of Algorithm 6), which is a confidence bound accounting for both the reference and the advantage. Let us also explain line 6 of Algorithm 6 a bit. If we augment the notation by letting $b_h^{\mathsf{R},n+1}(s,a)$ and $B_h^{\mathsf{R},n+1}(s,a)$ denote respectively $b_h^{\mathsf{R}}(s,a)$ and $B_h^{\mathsf{R}}(s,a)$ after (s,a) is visited for the n-th time, then this line is designed to ensure that

$$\eta_n b_h^{\mathsf{R},n+1}(s,a) + (1-\eta_n) B_h^{\mathsf{R},n}(s,a) \approx B_h^{\mathsf{R},n+1}(s,a).$$

With the above updates implemented properly, Q_h^{R} provides the advantage-based update of the Q-function at time step h, according to the update rule (3.4).

3.2.3 Theoretical guarantees

Encouragingly, the proposed CB-Q-Advantage algorithm manages to achieve near-optimal regret even in the sample-limited and memory-limited regime, as formalized by the following theorem.

Theorem 1. *Consider any $\delta \in (0,1)$, and suppose that $c_{\mathrm{b}} > 0$ is chosen to be a sufficiently large universal constant. Then there exists some absolute constant $C_0 > 0$ such that Algorithm 1 achieves*

$$\mathsf{Regret}(T) \le C_0 \left(\sqrt{H^2 SAT \log^4 \frac{SAT}{\delta}} + H^6 SA \log^3 \frac{SAT}{\delta} \right) \tag{3.8}$$

with probability at least $1-\delta$.

Theorem 1 delivers a non-asymptotic characterization of the performance of our algorithm CB-Q-Advantage. Several appealing features of the algorithm are noteworthy.

Algorithm 2: Auxiliary functions

Function `update-ucb-q()`:

$Q_h^{\mathsf{UCB}}(s_h,a_h) \leftarrow (1-\eta_n)Q_h^{\mathsf{UCB}}(s_h,a_h) + \eta_n\Big(r_h(s_h,a_h) + V_{h+1}(s_{h+1}) + c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{n}}\Big)$.

Function `update-lcb-q()`:

$Q_h^{\mathsf{LCB}}(s_h,a_h) \leftarrow (1-\eta_n)Q_h^{\mathsf{LCB}}(s_h,a_h) + \eta_n\Big(r_h(s_h,a_h) + V_{h+1}^{\mathsf{LCB}}(s_{h+1}) - c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{n}}\Big)$.

Function `update-ucb-q-advantage()`:

`/* update the moment statistics of` V_h^{R} `*/`

$[\mu_h^{\mathsf{ref}},\sigma_h^{\mathsf{ref}},\mu_h^{\mathsf{adv}},\sigma_h^{\mathsf{adv}}](s_h,a_h) \leftarrow$ `update-moments()`;

`/* update the accumulative bonus and bonus difference */`

$[\delta_h^{\mathsf{R}},B_h^{\mathsf{R}}](s_h,a_h) \leftarrow$ `update-bonus()`;

$b_h^{\mathsf{R}} \leftarrow B_h^{\mathsf{R}}(s_h,a_h) + (1-\eta_n)\frac{\delta_h^{\mathsf{R}}(s_h,a_h)}{\eta_n} + c_{\mathrm{b}}\frac{H^2\log\frac{SAT}{\delta}}{n^{3/4}}$;

`/* update the Q-estimate based on reference-advantage decomposition */`

$Q_h^{\mathsf{R}}(s_h,a_h) \leftarrow (1-\eta_n)Q_h^{\mathsf{R}}(s_h,a_h) + \eta_n\big(r_h(s_h,a_h) + V_{h+1}(s_{h+1}) - V_{h+1}^{\mathsf{R}}(s_{h+1}) + \mu_h^{\mathsf{ref}}(s_h,a_h) + b_h^{\mathsf{R}}\big)$;

Function `update-moments()`:

$\mu_h^{\mathsf{ref}}(s_h,a_h) \leftarrow (1-\frac{1}{n})\mu_h^{\mathsf{ref}}(s_h,a_h) + \frac{1}{n}V_{h+1}^{\mathsf{R}}(s_{h+1})$; `// mean of the reference`

$\sigma_h^{\mathsf{ref}}(s_h,a_h) \leftarrow (1-\frac{1}{n})\sigma_h^{\mathsf{ref}}(s_h,a_h) + \frac{1}{n}\big(V_{h+1}^{\mathsf{R}}(s_{h+1})\big)^2$; `//` 2^{nd} `moment of the reference`

$\mu_h^{\mathsf{adv}}(s_h,a_h) \leftarrow (1-\eta_n)\mu_h^{\mathsf{adv}}(s_h,a_h) + \eta_n\big(V_{h+1}(s_{h+1}) - V_{h+1}^{\mathsf{R}}(s_{h+1})\big)$; `// weighted average of the advantage`

$\sigma_h^{\mathsf{adv}}(s_h,a_h) \leftarrow (1-\eta_n)\sigma_h^{\mathsf{adv}}(s_h,a_h) + \eta_n\big(V_{h+1}(s_{h+1}) - V_{h+1}^{\mathsf{R}}(s_{h+1})\big)^2$. `// weighted` 2^{nd} `moment of the advantage`

Function `update-bonus()`:

$B_h^{\mathsf{next}}(s_h,a_h) \leftarrow c_{\mathrm{b}}\sqrt{\frac{\log\frac{SAT}{\delta}}{n}}\Big(\sqrt{\sigma_h^{\mathsf{ref}}(s_h,a_h) - \big(\mu_h^{\mathsf{ref}}(s_h,a_h)\big)^2} + \sqrt{H}\sqrt{\sigma_h^{\mathsf{adv}}(s_h,a_h) - \big(\mu_h^{\mathsf{adv}}(s_h,a_h)\big)^2}\Big)$;

$\delta_h^{\mathsf{R}}(s_h,a_h) \leftarrow B_h^{\mathsf{next}}(s_h,a_h) - B_h^{\mathsf{R}}(s_h,a_h)$;

$B_h^{\mathsf{R}}(s_h,a_h) \leftarrow B_h^{\mathsf{next}}(s_h,a_h)$.

- *Regret optimality.* Our regret bound (3.8) simplifies to

$$\mathsf{Regret}(T) \le \widetilde{O}\big(\sqrt{H^2SAT}\big) \tag{3.9}$$

as long as the sample size T exceeds

$$T \ge SA\,\mathrm{poly}(H). \tag{3.10}$$

This sublinear regret bound (3.9) is essentially optimal, as it coincides with the existing lower bound (1.1) modulo some logarithmic factor.

- *Sample complexity and substantially reduced burn-in cost.* As an interpretation of our theory (3.9), our algorithm attains ε average regret (i.e., $\frac{1}{K}\mathsf{Regret}(T) \le \varepsilon$) with a sample complexity

$$\widetilde{O}\Big(\frac{SAH^4}{\varepsilon^2}\Big).$$

  Crucially, the burn-in cost (3.10) is significantly lower than that of the state-of-the-art memory-efficient model-free algorithm (Zhang et al., 2020c) (whose optimality is guaranteed only in the range $T \ge S^6A^4\,\mathrm{poly}(H)$).

- *Memory efficiency.* Our algorithm, which is model-free in nature, achieves a low space complexity $O(SAH)$. This is basically un-improvable for the tabular case, since even storing the optimal Q-values alone takes $O(SAH)$ units of space. In comparison, while Ménard et al. (2021) also accommodates the sample size range (3.10), the algorithm proposed therein incurs a space complexity of $O(S^2AH)$ that is S times higher than ours.

- *Computational complexity.* An additional intriguing feature of our algorithm is its low computational complexity. The runtime of CB-Q-Advantage is no larger than $O(T)$, which is proportional to the time taken to read the samples. This matches the computational cost of the model-free algorithm UCB-Q proposed in Jin et al. (2018), and is considerably lower than that of the UCB-M-Q algorithm in Ménard et al. (2021) (which has a computational cost of at least $O(ST)$).

3.3 Analysis

In this chapter, we outline the main steps needed to prove our main result in Theorem 1.

3.3.1 Preliminaries: basic properties about learning rates

Before continuing, let us first state some basic facts regarding the learning rates. Akin to Jin et al. (2018), the proposed algorithm adopts the linearly rescaled learning rate

$$\eta_n = \frac{H+1}{H+n} \tag{3.11}$$

for the n-th visit of a state-action pair at any time step h. For notation convenience, we further introduce two sequences of related quantities defined for any integer $N \ge 0$ and $n \ge 1$:

$$\eta_n^N := \begin{cases} \eta_n \prod_{i=n+1}^{N}(1-\eta_i), & \text{if } N > n, \\ \eta_n, & \text{if } N = n, \\ 0, & \text{if } N < n \end{cases} \qquad \text{and} \qquad \eta_0^N := \begin{cases} \prod_{i=1}^{N}(1-\eta_i) = 0, & \text{if } N > 0, \\ 1, & \text{if } N = 0. \end{cases} \tag{3.12}$$

Algorithm 3: CB-Q-Advantage (a rewrite of Algorithm 1 that specifies dependency on k)

Parameters: some universal constant $c_{\mathrm{b}} > 0$ and probability of failure $\delta \in (0,1)$;

Initialize $Q_h^1(s,a), Q_h^{\mathsf{UCB},1}(s,a), Q_h^{\mathsf{R},1}(s,a) \leftarrow H$; $Q_h^{\mathsf{LCB},1}(s,a) \leftarrow 0$; $N_h^0(s,a) \leftarrow 0$; $V_h^1(s), V_h^{\mathsf{R},1}(s) \leftarrow H$; $\mu_h^{\mathsf{ref}}(s,a), \sigma_h^{\mathsf{ref}}(s,a), \mu_h^{\mathsf{adv}}(s,a), \sigma_h^{\mathsf{adv}}(s,a), \delta_h^{\mathsf{R}}(s,a), B_h^{\mathsf{R}}(s,a) \leftarrow 0$; and $u_{\mathrm{ref}}^1(s) = \mathsf{True}$, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

for *Episode* $k = 1$ **to** K **do**

  Set initial state $s_1 \leftarrow s_1^k$.

  for *Step* $h = 1$ **to** H **do**

    Take action $a_h^k = \pi_h^k(s_h) = \arg\max_a Q_h^k(s_h^k, a)$, and draw $s_{h+1}^k \sim P_h(\cdot \,|\, s_h^k, a_h^k)$. `// sampling`

    $N_h^k(s_h^k, a_h^k) \leftarrow N_h^{k-1}(s_h^k, a_h^k) + 1$; $n \leftarrow N_h^k(s_h^k, a_h^k)$. `// update the counter`

    $\eta_n \leftarrow \frac{H+1}{H+n}$. `// update the learning rate`

    $Q_h^{\mathsf{UCB},k+1}(s_h^k, a_h^k) \leftarrow$ `update-ucb-q()`. `// run UCB-Q; see Algorithm 6`

    $Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) \leftarrow$ `update-lcb-q()`. `// run LCB-Q; see Algorithm 6`

    $Q_h^{\mathsf{R},k+1}(s_h^k, a_h^k) \leftarrow$ `update-ucb-q-advantage()`. `// estimate` Q_h^{R}`; see Algorithm 6`

    `/* update Q-estimates using all estimates in hand, and update value estimates */`

    $Q_h^{k+1}(s_h^k, a_h^k) \leftarrow \min\left\{Q_h^{\mathsf{R},k+1}(s_h^k, a_h^k),\, Q_h^{\mathsf{UCB},k+1}(s_h^k, a_h^k),\, Q_h^k(s_h^k, a_h^k)\right\}$;

    $V_h^{k+1}(s_h^k) \leftarrow \max_a Q_h^{k+1}(s_h^k, a)$.

    $V_h^{\mathsf{LCB},k+1}(s_h^k) \leftarrow \max\left\{\max_a Q_h^{\mathsf{LCB},k+1}(s_h^k, a),\, V_h^{\mathsf{LCB},k}(s_h^k)\right\}$.

    `/* update reference values */`

    if $V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) > 1$ **then**

      $V_h^{\mathsf{R},k+1}(s_h^k) \leftarrow V_h^{k+1}(s_h^k)$, $\quad u_{\mathrm{ref}}^{k+1}(s_h^k) = \mathsf{True}$;

    else if $u_{\mathrm{ref}}^k(s_h^k) = \mathsf{True}$ **then**

      $V_h^{\mathsf{R},k+1}(s_h) \leftarrow V_h^{k+1}(s_h)$, $\quad u_{\mathrm{ref}}^{k+1}(s_h^k) = \mathsf{False}$.

As can be easily verified, we have

$$\sum_{n=1}^{N} \eta_n^N = \begin{cases} 1, & \text{if } N > 0, \\ 0, & \text{if } N = 0. \end{cases} \tag{3.13}$$

The following properties play an important role in the analysis.

Lemma 1. *For any integer $N > 0$, the following properties hold:*

$$\frac{1}{N^a} \le \sum_{n=1}^{N} \frac{\eta_n^N}{n^a} \le \frac{2}{N^a}, \qquad \textit{for all} \quad \frac{1}{2} \le a \le 1, \tag{3.14a}$$

$$\max_{1\le n\le N}\eta_n^N \le \frac{2H}{N}, \qquad \sum_{n=1}^{N}(\eta_n^N)^2 \le \frac{2H}{N}, \qquad \sum_{N=n}^{\infty}\eta_n^N \le 1+\frac{1}{H}. \tag{3.14b}$$

Proof. See Appendix A.2. □

3.3.2 Additional notation used in the proof

In order to enable a more concise description of the algorithm, we have suppressed the dependency of many quantities on the episode number k in Algorithms 1 and 6. This, however, becomes notationally inconvenient when presenting the proof. As a consequence, we shall adopt, throughout the analysis, a more complete set of notation, detailed below.

- (s_h^k, a_h^k): the state-action pair encountered and chosen at time step h in the k-th episode.
- $k_h^n(s,a)$: the index of the episode in which (s,a) is visited for the n-th time at time step h; for the sake of conciseness, we shall sometimes use the shorthand $k^n = k_h^n(s,a)$ whenever it is clear from the context.
- $k_h^n(s)$: the index of the episode in which state s is visited for the n-th time at time step h; we might sometimes abuse the notation by abbreviating $k^n = k_h^n(s)$.
- $P_h^k \in \{0,1\}^{1\times|\mathcal{S}|}$: the empirical transition at time step h in the k-th episode, namely,

$$P_h^k(s) = \mathbb{1}\left(s = s_{h+1}^k\right). \tag{3.15}$$

In addition, for several parameters of interest in Algorithm 1, we introduce the following set of augmented notation.

- $N_h^k(s,a)$ denotes $N_h(s,a)$ *by the end* of the k-th episode; for the sake of conciseness, we shall often abbreviate $N^k = N_h^k(s,a)$ or $N^k = N_h^k(s_h^k, a_h^k)$ (depending on which result we are proving).
- $Q_h^k(s,a)$, $V_h^k(s)$, and $Q_h^{\mathsf{UCB},k}(s,a)$ denote respectively $Q_h(s,a)$, $V_h(s)$ and $Q_h^{\mathsf{UCB}}(s,a)$ *at the beginning* of the k-th episode.
- $Q_h^{\mathsf{LCB},k}(s,a)$ and $V_h^{\mathsf{LCB},k}(s)$ denote respectively $Q_h^{\mathsf{LCB}}(s,a)$ and $V_h^{\mathsf{LCB}}(s)$ *at the beginning* of the k-th episode.
- $Q_h^{\mathsf{R},k}(s,a)$, $V_h^{\mathsf{R},k}(s)$ and $u_{\mathrm{ref}}^k(s)$ denote respectively $Q_h^{\mathsf{R}}(s,a)$, $V_h^{\mathsf{R}}(s)$ and $u_{\mathrm{ref}}(s)$ *at the beginning* of the k-th episode.
- $\left[\mu_h^{\mathsf{ref},k}, \sigma_h^{\mathsf{ref},k}, \mu_h^{\mathsf{adv},k}, \sigma_h^{\mathsf{adv},k}, \delta_h^{\mathsf{R},k}, B_h^{\mathsf{R},k}\right]$ denotes $\left[\mu_h^{\mathsf{ref}}, \sigma_h^{\mathsf{ref}}, \mu_h^{\mathsf{adv}}, \sigma_h^{\mathsf{adv}}, \delta_h^{\mathsf{R}}, B_h^{\mathsf{R}}\right]$ *at the beginning* of the k-th episode.

For any given vector $V \in \mathbb{R}^S$, we define the variance parameter w.r.t. $P_{h,s,a}$ (cf. (2.7)) as follows

$$\mathsf{Var}_{h,s,a}(V) \coloneqq \mathop{\mathbb{E}}_{s' \sim P_{h,s,a}} \Big[\big(V(s') - P_{h,s,a}V\big)^2 \Big] = P_{h,s,a}\big(V^2\big) - \big(P_{h,s,a}V\big)^2. \tag{3.16}$$

3.3.3 Key properties of Q-estimates and auxiliary sequences

In this subchapter, we introduce several key properties of our Q-estimates and value estimates, which play a crucial role in the proof of Theorem 1. The proofs for this subchapter are deferred to Appendix A.3.

Properties of the Q-estimate Q_h^k: monotonicity and optimism. We first make an important observation regarding the monotonicity of the value estimates Q_h^k and V_h^k. To begin with, it is straightforward to see that the update rule in Algorithm 3 (cf. line 12) ensures the following monotonicity property:

$$Q_h^{k+1}(s,a) \le Q_h^k(s,a) \qquad \text{for all } (s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H], \tag{3.17a}$$

which combined with line 13 of Algorithm 3 leads to monotonicity of $V_h(s)$ as follows:

$$V_h^{k+1}(s) = Q_h^{k+1}\big(s, \pi_h^{k+1}(s)\big) \le Q_h^k\big(s, \pi_h^{k+1}(s)\big) \le V_h^k(s). \tag{3.17b}$$

Moreover, by virtue of the update rule in line 12 of Algorithm 3, we can immediately obtain (via induction) the following useful property

$$Q_h^{\mathsf{R},k}(s,a) \ge Q_h^k(s,a) \qquad \text{for all } (k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}. \tag{3.18}$$

In addition, Q_h^k and V_h^k form an "optimistic view" of $Q_h^\star$ and $V_h^\star$, respectively, as asserted by the following lemma.

Lemma 2. *Consider any $\delta \in (0,1)$. Suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$,*

$$Q_h^k(s,a) \ge Q_h^\star(s,a) \qquad \textit{and} \qquad V_h^k(s) \ge V_h^\star(s) \tag{3.19}$$

hold simultaneously for all $(s,a,k,h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$.

Lemma 2 implies that Q_h^k (resp. V_h^k) is a pointwise upper bound on $Q_h^\star$ (resp. $V_h^\star$). Taking this result together with the non-increasing property (3.17), we see that Q_h^k (resp. V_h^k) becomes an increasingly tighter estimate of $Q_h^\star$ (resp. $V_h^\star$) as the number of episodes k increases. This important fact forms the basis of the subsequent proof, allowing us to replace $V_h^\star$ with V_h^k when upper bounding

the regret. Combining Lemma 2 with (3.18), we can straightforwardly see that with probability at least $1-\delta$:

$$Q_h^{\mathsf{R},k}(s,a) \geq Q_h^\star(s,a) \qquad \text{for all } (k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}. \tag{3.20}$$

Properties of the Q-estimate $Q_h^{\mathsf{LCB},k}$: pessimism and proximity. In parallel, we formalize the fact that $Q_h^{\mathsf{LCB},k}$ and $V_h^{\mathsf{LCB},k}$ provide a "pessimistic view" of $Q_h^\star$ and $V_h^\star$, respectively. Furthermore, it becomes increasingly more likely for $Q_h^{\mathsf{LCB},k}$ and Q_h^k to stay close to each other as k increases, which indicates that the confidence interval that contains the optimal value $Q_h^\star$ becomes shorter and shorter. These properties are summarized in the following lemma.

Lemma 3. *Consider any $\delta\in(0,1)$, and suppose that $c_{\mathrm{b}}>0$ is some sufficiently large constant. Then with probability at least $1-\delta$,*

$$Q_h^{\mathsf{LCB},k}(s,a) \leq Q_h^\star(s,a) \qquad \textit{and} \qquad V_h^{\mathsf{LCB},k}(s) \leq V_h^\star(s) \tag{3.21}$$

hold for all $(s,a,k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$, and

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > \varepsilon\right) \lesssim \frac{H^6SA\log\frac{SAT}{\delta}}{\varepsilon^2} \tag{3.22}$$

holds for all $\varepsilon\in(0,H]$.

Interestingly, the upper bound (3.22) only scales logarithmically in the number K of episodes, thus implying the closeness of $Q_h^{\mathsf{LCB},k}$ and Q_h^k for a large fraction of episodes. Note that it is straightforward to ensure the monotonicity property of $V_h^{\mathsf{LCB},k}$ from the update rule in Algorithm 3 (cf. line 14):

$$V_h^{\mathsf{LCB},k+1}(s) \geq V_h^{\mathsf{LCB},k}(s) \qquad \text{for all } (s,k,h)\in\mathcal{S}\times[K]\times[H], \tag{3.23}$$

which in conjunction with (3.21), implies that $V_h^{\mathsf{LCB},k}(s)$ gets closer to $V_h^\star(s)$ as the number of episodes k increases. Together with the monotonicity of V_h^k (cf. (3.17b)), an important consequence is that the reference value V_h^{R} will stop being updated shortly after the following condition is met for the first time (according to lines 15-18 of Algorithm 1)

$$V_h^k(s) \leq V_h^{\mathsf{LCB},k}(s)+1 \leq V_h^\star(s)+1 \qquad \text{for all } s\in\mathcal{S}. \tag{3.24}$$

Properties of the reference $V_h^{\mathsf{R},k}$. The above fact ensures that $V_h^{\mathsf{R},k}$ will not be updated too many times. In fact, its value stays reasonably close to V_h^k even after being locked to a fixed value, which ensures its fidelity as a reference signal. Moreover, the aggregate difference between $V_h^{\mathsf{R},k}$ and the final reference $V_h^{\mathsf{R},K}$ over the entire trajectory can be bounded in a reasonably tight fashion

(owing to (3.22)), as formalized in the next lemma. These properties play a key role in reducing the burn-in cost of the proposed algorithm.

Lemma 4. *Consider any $\delta \in (0,1)$. Suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability exceeding $1-\delta$, one has*

$$\left|V_h^k(s) - V_h^{\mathsf{R},k}(s)\right| \le 2 \tag{3.25}$$

for all $(k,h,s) \in [K] \times [H] \times \mathcal{S}$, and

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathsf{R},k}(s_h^k) - V_h^{\mathsf{R},K}(s_h^k)\right) \\
&\quad\le H^2 S + \sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)\right)\mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) > 1\right) &(3.26)\\
&\quad\lesssim H^6 SA \log\frac{SAT}{\delta}. &(3.27)
\end{aligned}$$

In words, Lemma 4 guarantees that (i) our value function estimate and the reference value are always sufficiently close (cf. (3.25)), and (ii) the aggregate difference between $V_h^{\mathsf{R},k}$ and the final reference value $V_h^{\mathsf{R},K}$ is nearly independent of the sample size T (except for some logarithmic scaling).

3.3.4 Main steps of the proof

We are now ready to embark on the regret analysis for CB-Q-Advantage, which consists of multiple steps as follows.

Step 1: regret decomposition. Lemma 2 allows one to upper bound the regret as follows

$$\mathsf{Regret}(T) := \sum_{k=1}^{K}\left(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \le \sum_{k=1}^{K}\left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right). \tag{3.28}$$

To continue, it boils down to controlling $V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)$. Towards this end, we intend to examine $V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)$ across all time steps $1 \le h \le H$, which admits the following decomposition:

$$\begin{aligned}
V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) &\overset{\text{(i)}}{=} Q_h^k(s_h^k,a_h^k) - Q_h^{\pi^k}(s_h^k,a_h^k)\\
&= Q_h^k(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) + Q_h^\star(s_h^k,a_h^k) - Q_h^{\pi^k}(s_h^k,a_h^k)\\
&\overset{\text{(ii)}}{=} Q_h^k(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) + P_{h,s_h^k,a_h^k}\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right)\\
&\overset{\text{(iii)}}{=} Q_h^k(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) + \left(P_{h,s_h^k,a_h^k} - P_h^k\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right) + V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)
\end{aligned}$$

$$\leq Q_h^{\mathsf{R},k}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) + \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) + V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k). \tag{3.29}$$

Here, (i) holds since π_h^k is a greedy policy w.r.t. Q_h^k and $\pi_h^k(s_h^k) = a_h^k$, (ii) comes from the Bellman equations

$$Q_h^{\pi^k}(s,a) - Q_h^\star(s,a) = \big(r_h(s,a) + P_{h,s,a}V_{h+1}^{\pi^k}\big) - \big(r_h(s,a) + P_{h,s,a}V_{h+1}^\star\big) = P_{h,s,a}\big(V_{h+1}^{\pi^k} - V_{h+1}^\star\big),$$

(iii) follows from $P_h^k(V_{h+1}^\star - V_{h+1}^{\pi^k}) = V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)$ (see the notation (3.15)), whereas the last inequality comes from (3.18). Summing (3.29) over $1 \leq k \leq K$ and making use of Lemma 2, we reach at

$$\begin{aligned}
\sum_{k=1}^K \big(V_h^\star(s_h^k) - V_h^{\pi^k}(s_h^k)\big) &\leq \sum_{k=1}^K \big(V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)\big) \\
&\leq \sum_{k=1}^K \big(Q_h^{\mathsf{R},k}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)\big) + \sum_{k=1}^K \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) \\
&\quad + \sum_{k=1}^K \big(V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big).
\end{aligned} \tag{3.30}$$

This allows us to establish a connection between $\sum_k \big(V_h^\star(s_1^k) - V_h^{\pi^k}(s_h^k)\big)$ for step h and $\sum_k \big(V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big)$ for step $h+1$.

Step 2: managing regret by recursion. The regret can be further manipulated by leveraging the update rule of $Q_h^{\mathsf{R},k}$ as well as recursing over the time steps $h = 1, 2, \cdots, H$ with the terminal condition $V_{H+1}^k = V_{H+1}^{\pi^k} = 0$. This leads to a key decomposition as summarized in the lemma below, whose proof is provided in Appendix A.4.

Lemma 5. *Fix $\delta \in (0,1)$. Suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$, one has*

$$\sum_{k=1}^K \big(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\big) \leq \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3, \tag{3.31}$$

where

$$\mathcal{R}_1 := \sum_{h=1}^H \left(1 + \frac{1}{H}\right)^{h-1} \left(HSA + 8c_{\mathrm{b}}H^2(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta} + \sum_{k=1}^K \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big)\right), \tag{3.32a}$$

$$\mathcal{R}_2 := \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K} B_h^{\mathsf{R},k}(s_h^k,a_h^k), \tag{3.32b}$$

$$\mathcal{R}_3 := \sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_h^k\left((P_h^k - P_{h,s_h^k,a_h^k})(V_{h+1}^{\star}-V_{h+1}^{\mathsf{R},k}) + \frac{\sum_{i=1}^{N_h^k(s_h^k,a_h^k)}\left(V_{h+1}^{\mathsf{R},k_h^i(s_h^k,a_h^k)}(s_{h+1}^{k_h^i(s_h^k,a_h^k)}) - P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k}\right)}{N_h^k(s_h^k,a_h^k)}\right), \tag{3.32c}$$

with

$$\lambda_h^k := \left(1+\frac{1}{H}\right)^{h-1}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\eta^n_{N_h^k(s_h^k,a_h^k)}.$$

This lemma attempts to upper bound the target quantity $\sum_{k=1}^{K}\left(V_1^k(s_1^k)-V_1^{\pi^k}(s_1^k)\right)$ via three terms (see (3.31)). Informally, these terms reflect (i) the influence of the initialization as well as the finite-sample uncertainty of $P_h^k(V_{h+1}^{\star}-V_{h+1}^{\pi^k})$, (ii) the influence of the size of the bonus terms, and (iii) the discrepancy term when the running value iterates are replaced by the reference values. As we shall see in the analysis, the key in obtaining these terms lies in properly expanding the component $\sum_{k=1}^{K}\left(Q_h^{\mathsf{R},k}(s_h^k,a_h^k)-Q_h^{\star}(s_h^k,a_h^k)\right)$ in (3.30), as well as applying induction across all $h=1,\cdots,H$.

Step 3: controlling the terms in (3.32) separately. As it turns out, each of the terms in (3.32) can be well controlled. We provide the bounds for these terms in the following lemma.

Lemma 6. *Consider any $\delta\in(0,1)$. With probability at least $1-\delta$, we have the following upper bounds:*

$$\mathcal{R}_1 \le C_{\mathrm{r}}\left\{\sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^{4.5}SA\log^2\frac{SAT}{\delta}\right\},$$
$$\mathcal{R}_2 \le C_{\mathrm{r}}\left\{\sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^{4}SA\log^2\frac{SAT}{\delta}\right\},$$
$$\mathcal{R}_3 \le C_{\mathrm{r}}\left\{\sqrt{H^2SAT\log^4\frac{SAT}{\delta}} + H^{6}SA\log^3\frac{SAT}{\delta}\right\}$$

for some universal constant $C_{\mathrm{r}}>0$.

In order to derive the above bounds, the main strategy is to apply the Bernstein-type concentration inequalities carefully, and to upper bound the sum of variance in a careful manner. The proofs are deferred to Appendix A.5.

Step 4: putting all this together. We now have everything in place to establish our main result. Taking the preceding bounds in Lemma 6 together with (3.32), we see that with probability

exceeding $1-\delta$, one has

$$\mathsf{Regret}(T) \le \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3 \lesssim \sqrt{H^2SAT\log^4\frac{SAT}{\delta}} + H^6SA\log^3\frac{SAT}{\delta}$$

as claimed.

3.4 Discussions

In this chapter, we have proposed a novel model-free RL algorithm — tailored to online episodic settings — that attains near-optimal regret $\widetilde{O}(\sqrt{H^2SAT})$ and near-minimal memory complexity $O(SAH)$ at once. Remarkably, the near-optimality of the algorithm comes into effect as soon as the sample size rises above $O(SA\,\mathrm{poly}(H))$, which has significantly improved upon the sample size requirements (or burn-in cost) for any prior regret-optimal model-free algorithm (based on the definition of the model-free algorithm in Jin et al. (2018)). Given that online data collection could be expensive, time-consuming, or high-stakes in a variety of contemporary applications (e.g., clinical trials, autonomous driving, online advertisement), reducing burn-in sample sizes compromising sample optimality is crucial in enabling sample-efficient solutions in these sample-constrained applications.

Chapter 4

Model-Free Offline RL

4.1 Problem formulation

In this chapter, we consider offline RL in the episodic finite-horizon setting (introduced in Chapter 2.1.1), which assumes the availability of a history dataset $\mathcal{D}$ containing K episodes each of length H. These episodes are independently generated based on a certain policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$ — called the *behavior policy*, resulting in a dataset

$$\mathcal{D} \coloneqq \Big\{\big(s_1^k,\, a_1^k,\, r_1^k,\, \ldots, s_H^k,\, a_H^k,\, r_H^k\big)\Big\}_{k=0}^{K-1}.$$

Here, the initial states $\{s_1^k\}_{k=1}^K$ are independently drawn from $\rho \in \Delta(\mathcal{S})$ such that $s_1^k \overset{\text{i.i.d.}}{\sim} \rho$, while the remaining states and actions are generated by the MDP induced by the behavior policy μ. The total number of samples is thus given by

$$T = KH.$$

In addition, let $d_h^\pi(s)$ and $d_h^\pi(s,a)$ denote respectively the occupancy distribution induced by π at time step $h \in [H]$, namely,

$$d_h^\pi(s) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi), \qquad d_h^\pi(s,a) \coloneqq \mathbb{P}(s_h = s \,|\, s_1 \sim \rho, \pi)\, \pi_h(a \,|\, s); \tag{4.1}$$

here and throughout, we denote $[H] \coloneqq \{1, \cdots, H\}$. Given that the initial state s_1 is drawn from ρ, the above definition gives

$$d_1^\pi(s) = \rho(s) \qquad \text{for any policy } \pi. \tag{4.2}$$

Goal. With the notation (2.4) in place, the goal of offline RL amounts to finding an ε-optimal policy $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$ satisfying

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon$$

with as few samples as possible, and ideally, in a computationally fast and memory-efficient manner.

Single-policy concentrability. Obviously, efficient offline RL cannot be accomplished without imposing proper assumptions on the behavior policy, which also provide means to gauge the

difficulty of the offline RL task through the quality of the history dataset. Following the recent works Rashidinejad et al. (2021); Xie et al. (2021b), we assume that the behavior policy μ satisfies the following property called *single-policy concentrability*.

Definition 1 (single-policy concentrability)**.** The single-policy concentrability coefficient $C^\star \in [1, \infty)$ of a behavior policy μ is defined to be the smallest quantity that satisfies

$$\max_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)} \leq C^\star, \tag{4.3}$$

where we adopt the convention $0/0 = 0$.

Intuitively, the single-policy concentrability coefficient measures the discrepancy between the optimal policy $\pi^\star$ and the behavior policy μ in terms of the resulting density ratio of the respective occupancy distributions. It is noteworthy that a finite $C^\star$ does not necessarily require μ to cover the entire state-action space; instead, it can be attainable when its coverage subsumes that of the optimal policy $\pi^\star$. This is in stark contrast to, and in fact much weaker than, other assumptions that require either full coverage of the behavior policy (i.e., $\min_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}} d_h^{\mu}(s,a) > 0$ (Li et al., 2021; Yin et al., 2021a,b)), or uniform concentrability over all possible policies (Chen and Jiang, 2019). Additionally, the single-policy concentrability coefficient is minimized (i.e., $C^\star = 1$) when the behavior policy μ coincides with the optimal policy $\pi^\star$, a scenario closely related to imitation learning or behavior cloning (Rajaraman et al., 2020).

4.2 Algorithms and theory

In the current chapter, we present two model-free algorithms — namely, LCB-Q and LCB-Q-Advantage — for offline RL, along with their respective theoretical guarantees. The first algorithm can be viewed as a pessimistic variant of the classical Q-learning algorithm, while the second one further leverages the idea of variance reduction to boost the sample efficiency. In this chapter, we begin by introducing LCB-Q.

4.2.1 LCB-Q: a natural pessimistic variant of Q-learning

Before proceeding, we find it convenient to first review the classical Q-learning algorithm (Watkins and Dayan, 1992; Watkins, 1989), which can be regarded as a stochastic approximation scheme to solve the Bellman optimality equation (2.6). Upon receiving a sample transition (s_h, a_h, r_h, s_{h+1}) at time step h, Q-learning updates the corresponding entry in the Q-estimate as follows

$$Q_h(s_h,a_h) \;\leftarrow\; (1-\eta)Q_h(s_h,a_h) + \eta\Big\{r_h(s_h,a_h) + V_{h+1}(s_{h+1})\Big\}, \tag{4.4}$$

where Q_h (resp. V_h) indicates the running estimate of $Q_h^\star$ (resp. $V_h^\star$), and $0 < \eta < 1$ is the learning rate. In comparison to model-based algorithms that require estimating the probability transition kernel based on all the samples, Q-learning, as a popular kind of model-free algorithms, is simpler and enjoys more flexibility without explicitly constructing the model of the environment. The wide applicability of Q-learning motivates one to adapt it to accommodate offline RL.

Inspired by recent advances in incorporating the pessimism principle for offline RL (Jin et al., 2021; Rashidinejad et al., 2021), we study a pessimistic variant of Q-learning called LCB-Q, which modifies the Q-learning update rule as follows

$$Q_h(s_h, a_h) \leftarrow (1-\eta_n) Q_h(s_h, a_h) + \eta_n \Big\{ r_h(s_h, a_h) + V_{h+1}(s_{h+1}) - b_n \Big\}, \tag{4.5}$$

where η_n is the learning rate depending on the number of times n that the state-action pair (s_h, a_h) has been visited at step h, and the penalty term $b_n > 0$ (cf. line 8 of Algorithm 4) reflects the uncertainty of the corresponding Q-estimate and implements pessimism in the face of uncertainty. The entire algorithm, which is a *single-pass* algorithm that only requires reading the offline dataset once, is summarized in Algorithm 4.

Algorithm 4: LCB-Q for offline RL

Parameters: some constant $c_{\mathrm{b}} > 0$, target success probability $1-\delta \in (0,1)$, and $\iota = \log\big(\frac{SAT}{\delta}\big)$.
Initialize $Q_h(s,a) \leftarrow 0$, $N_h(s,a) \leftarrow 0$, and $V_h(s) \leftarrow 0$ for all $(s,h) \in \mathcal{S} \times [H+1]$; $\widehat{\pi}$ s.t. $\widehat{\pi}_h(s) = 1$ for all $(h,s) \in [H] \times \mathcal{S}$.
for *Episode* $k = 1$ **to** K **do**
  Sample a new trajectory $\{s_h, a_h, r_h\}_{h=1}^H$ from $\mathcal{D}$. `// sampling from batch dataset`
  `// update the policy`
  for *Step* $h = 1$ **to** H **do**
    $N_h(s_h, a_h) \leftarrow N_h(s_h, a_h) + 1$. `// update the counter`
    $n \leftarrow N_h(s_h, a_h)$; $\eta_n \leftarrow \frac{H+1}{H+n}$. `// update the learning rate`
    $b_n \leftarrow c_{\mathrm{b}} \sqrt{\frac{H^3 \iota^2}{n}}$. `// update the bonus term`
    `// run the Q-learning update with LCB`
    $Q_h(s_h, a_h) \leftarrow Q_h(s_h, a_h) + \eta_n \Big\{ r_h(s_h, a_h) + V_{h+1}(s_{h+1}) - Q_h(s_h, a_h) - b_n \Big\}$.
    `// update the value estimates`
    $V_h(s_h) \leftarrow \max\Big\{ V_h(s_h),\, \max_a Q_h(s_h, a) \Big\}$.
    If $V_h(s_h) = \max_a Q_h(s_h, a)$: update $\widehat{\pi}_h(s) \leftarrow \arg\max_a Q_h(s, a)$.
Output: the policy $\widehat{\pi}$.

4.2.2 Theoretical guarantees for LCB-Q

The proposed LCB-Q algorithm manages to achieve an appealing sample complexity as formalized by the following theorem.

Theorem 2. *Consider any $\delta \in (0,1)$. Suppose that the behavior policy μ satisfies Assumption 1 with single-policy concentrability coefficient $C^\star \geq 1$. Let $c_{\mathrm{b}} > 0$ be some sufficiently large constant, and take $\iota := \log\left(\frac{SAT}{\delta}\right)$. Assume that $T > SC^\star\iota$, then the policy $\widehat{\pi}$ returned by Algorithm 4 satisfies*

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \leq c_{\mathrm{a}}\sqrt{\frac{H^6SC^\star\iota^3}{T}} \tag{4.6}$$

with probability at least $1-\delta$, where $c_{\mathrm{a}} > 0$ is some universal constant.

As asserted by Theorem 2, the LCB-Q algorithm is guaranteed to find an ε-optimal policy with high probability, as long as the total sample size $T = KH$ exceeds

$$\widetilde{O}\left(\frac{H^6SC^\star}{\varepsilon^2}\right), \tag{4.7}$$

where $\widetilde{O}(\cdot)$ hides logarithmic dependencies. When the behavior policy is close to the optimal policy, the single-policy concentrability coefficient $C^\star$ is closer to 1; if this is the case, then our bound indicates that the sample complexity does not depend on the size A of the action space, which can be a huge saving when the action space is enormous.

Comparison with model-based pessimistic approaches. A model-based approach — called Value Iteration with Lower Confidence Bounds (VI-LCB) — has been recently proposed for offline RL (Rashidinejad et al., 2021; Xie et al., 2021b). In the finite-horizon case, VI-LCB incorporates an additional LCB penalty into the classical value iteration algorithm, and updates *all* the entries in the Q-estimate simultaneously as follows

$$Q_h(s,a) \;\leftarrow\; r_h(s,a) + \widehat{P}_{h,s,a}V_{h+1} - b_h(s,a), \tag{4.8}$$

with the aim of tuning down the confidence on those state-action pairs that have only been visited infrequently. Here, $\widehat{P}_{h,s,a}$ represents the empirical estimation of the transition kernel $P_{h,s,a}$, and $b_h(s,a) > 0$ is chosen to capture the uncertainty level of $(\widehat{P}_{h,s,a} - P_{h,s,a})V_{h+1}$. Working backward, the algorithm estimates the Q-value Q_h recursively over the time steps $h = H, H-1, \cdots, 1$. In comparison with VI-LCB, our algorithm enjoys enhanced flexibility without the need of specifying the transition kernel of the environment (as model estimation might potentially incur a higher memory burden).

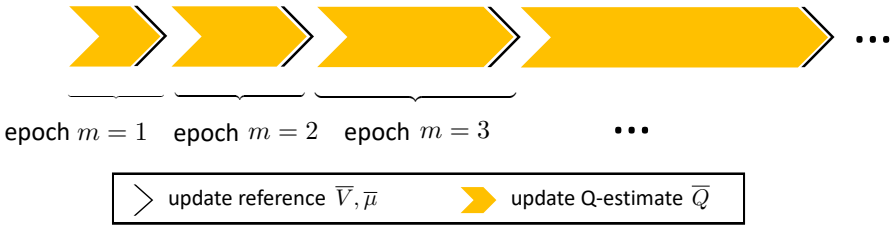


Figure 4.1: An illustration of the epoch-based LCB-Q-Advantage algorithm.

4.2.3 LCB-Q-Advantage for near-optimal offline RL

The careful reader might notice that the sample complexity (4.7) derived for LCB-Q remains a factor of H^2 away from the minimax lower bound (see Table 1.2). To further close the gap and improve the sample complexity, we propose a new variant called LCB-Q-Advantage, which leverages the idea of variance reduction to accelerate convergence (Johnson and Zhang, 2013; Li et al., 2023b, 2021; Sidford et al., 2018b; Wainwright, 2019b; Xie et al., 2021b; Zhang et al., 2020c).

Inspired by the reference-advantage decomposition adopted in (Li et al., 2023b; Zhang et al., 2020c) for online Q-learning, LCB-Q-Advantage maintains a collection of reference values $\{\overline{V}_h\}_{h=1}^H$, which serve as running proxy for the optimal values $\{V_h^\star\}_{h=1}^H$ and allow for reduced variability in each iteration. To be more specific, the LCB-Q-Advantage algorithm (cf. Algorithm 5 as well as the subroutines in Algorithm 6 that closely resemble Li et al. (2023b)) proceeds in an epoch-based style (the m-th epoch consists of $L_m = 2^m$ episodes of samples), where the reference values are updated at the end of each epoch to be used in the next epoch, and the Q-estimates are iteratively updated during the remaining time of each epoch. By maintaining two auxiliary sequences of *pessimistic* Q-estimates — that is, Q^{LCB} constructed by the pessimistic Q-learning update, and $\overline{Q}$ constructed by the pessimistic Q-learning update based on the reference-advantage decomposition — the Q-estimate is updated by taking the maximum over the three candidates (cf. line 16 of Algorithm 5)

$$Q_h(s,a) \leftarrow \max\{Q_h^{\mathsf{LCB}}(s,a),\, \overline{Q}_h(s,a),\, Q_h(s,a)\} \tag{4.9}$$

when the state-action pair (s,a) is visited at the step h. We now take a moment to discuss the key ingredients of the proposed algorithm in further detail.

Updating the references $\overline{V}_h$ and $\overline{\mu}_h$. At the end of each epoch, the reference values $\{\overline{V}_h\}_{h=1}^H$, as well as the associated running average $\{\overline{\mu}_h\}_{h=1}^H$, are determined using what happens during the current epoch. More specifically, the following update rules for $\overline{V}_h$ and $\overline{\mu}_h$ are carried out at the

end of the m-th epoch:

$$\overline{V}_h(s) \leftarrow \overline{V}_h^{\mathsf{next}}(s), \tag{4.10a}$$

$$\overline{\mu}_h(s,a) \leftarrow \frac{\sum_{t=1}^{L_m} \mathbb{1}(s_h^t = s, a_h^t = a)\overline{V}_{h+1}(s_{h+1}^t)}{\max\Big\{\big\{\sum_{t=1}^{L_m} \mathbb{1}(s_h^t = s, a_h^t = a)\big\}, 1\Big\}} \tag{4.10b}$$

for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. Here, $\overline{V}_h(s)$ is assigned by $\overline{V}_h^{\mathsf{next}}(s)$, which is maintained as the value estimate $V_h(s)$ at the end of the $(m-1)$-th epoch, and the update of $\overline{\mu}_h(s,a)$ is implemented in a recursive manner in the current m-th epoch. See also line 21 and line 19 of Algorithm 5.

Learning Q-estimate $\overline{Q}_h$ based on the reference-advantage decomposition. Armed with the references $\overline{V}_h$ and $\overline{\mu}_h$ updated at the end of the previous $(m-1)$-th epoch, LCB-Q-Advantage iteratively updates the Q-estimate $\overline{Q}_h$ in all episodes during the m-th epoch. At each time step h in any episode, whenever (s,a) is visited, LCB-Q-Advantage updates the reference Q-value as follows:

$$\overline{Q}_h(s,a) \leftarrow (1-\eta)\overline{Q}_h(s,a) + \eta\Big\{r_h(s,a) + \underbrace{\widehat{P}_{h,s,a}\big(V_{h+1} - \overline{V}_{h+1}\big)}_{\text{estimate of } P_{h,s,a}(V_{h+1}-\overline{V}_{h+1})} + \underbrace{\overline{\mu}_h}_{\text{estimate of } P_{h,s,a}\overline{V}_{h+1}} - \overline{b}_h(s,a)\Big\}. \tag{4.11}$$

Intuitively, we decompose the target $P_{h,s,a}V_{h+1}$ into a reference part $P_{h,s,a}\overline{V}_{h+1}$ and an advantage part $P_{h,s,a}(V_{h+1} - \overline{V}_{h+1})$, and cope with the two parts separately. In the sequel, let us take a moment to discuss three essential ingredients of the update rule (4.11), which shed light on the design rationale of our algorithm.

- Akin to LCB-Q, the term $\widehat{P}_{h,s,a}\big(V_{h+1} - \overline{V}_{h+1}\big)$ serves as an unbiased stochastic estimate of $P_{h,s,a}\big(V_{h+1} - \overline{V}_{h+1}\big)$ if a sample transition (s,a,s_{h+1}) at time step h is observed. If V_{h+1} stays close to the reference $\overline{V}_{h+1}$ as the algorithm proceeds, the variance of this stochastic term can be lower than that of the stochastic term $\widehat{P}_{h,s,a}V_{h+1}$ in (4.5).

- The auxiliary estimate $\overline{\mu}_h$ introduced in (4.10b) serves as a running estimate of the reference part $P_{h,s,a}\overline{V}_{h+1}$. Based on the update rule (4.10b), we design $\overline{\mu}_h(s,a)$ to estimate the running mean of the reference part $\big[P_{h,s,a}\overline{V}_{h+1}\big]$ using a number of previous samples. As a result, we expect the variability of this term to be well-controlled, particularly as the number of samples in each epoch grows exponentially (recall that $L_m = 2^m$).

- In each episode, the term $\overline{b}_h(s,a)$ serves as the additional confidence bound on the error between the estimates of the reference/advantage and the ground truth. More specifically, $\mu_h^{\mathsf{ref}}(s,a)$ and $\sigma_h^{\mathsf{ref}}(s,a)$ are respectively the running mean and 2nd moment of the reference part $\big[P_{h,s,a}\overline{V}_{h+1}\big]$ (cf. lines 9-10 of Algorithm 6); $\mu_h^{\mathsf{adv}}(s,a)$ and $\sigma_h^{\mathsf{adv}}(s,a)$ represent respectively

the running mean and 2nd moment of the advantage part $\big[P_{h,s,a}(V_{h+1}-\overline{V}_{h+1})\big]$ (cf. lines 11-12 of Algorithm 6); $\overline{B}_h(s,a)$ aggregates the empirical standard deviations of the reference and the advantage parts. The LCB penalty term $\overline{b}_h(s,a)$ is updated using $\overline{B}_h(s,a)$ and $\overline{\delta}_h(s_h,a_h)$ (cf. lines 5-6 of Algorithm 6), taking into account the confidence bounds for both the reference and the advantage.

In a nutshell, the auxiliary sequences of the reference values are designed to help reduce the variance of the stochastic Q-learning updates, which taken together with the principle of pessimism play a crucial role in the improvement of sample complexity for offline RL.

4.2.4 Theoretical guarantees for LCB-Q-Advantage

Encouragingly, the proposed LCB-Q-Advantage algorithm provably achieves near-optimal sample complexity for sufficiently small ε, as demonstrated by the following theorem.

Theorem 3. *Consider any $\delta \in (0,1)$, and recall that $\iota = \log\big(\frac{SAT}{\delta}\big)$ and $T = KH$. Suppose that $c_{\mathrm{b}} > 0$ is chosen to be a sufficiently large constant, and that the behavior policy μ satisfies Assumption 1. Then there exists some universal constant $c_{\mathrm{g}} > 0$ such that with probability at least $1-\delta$, the policy $\widehat{\pi}$ output by Algorithm 5 satisfies*

$$V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \le c_{\mathrm{g}}\left(\sqrt{\frac{H^4SC^{\star}\iota^5}{T}} + \frac{H^5SC^{\star}\iota^4}{T}\right). \tag{4.12}$$

As a consequence, Theorem 3 reveals that the LCB-Q-Advantage algorithm is guaranteed to find an ε-optimal policy (i.e., $V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon$) as long as the total sample size T exceeds

$$\widetilde{O}\left(\frac{H^4SC^{\star}}{\varepsilon^2} + \frac{H^5SC^{\star}}{\varepsilon}\right). \tag{4.13}$$

For sufficiently small accuracy level ε (i.e., $\varepsilon \le 1/H$), this results in a sample complexity of

$$\widetilde{O}\left(\frac{H^4SC^{\star}}{\varepsilon^2}\right), \tag{4.14}$$

thereby matching the minimax lower bound developed in Xie et al. (2021b) up to logarithmic factor. Compared with the minimax lower bound $\Omega\big(\frac{H^4SA}{\varepsilon^2}\big)$ in the online RL setting (Domingues et al., 2021), this suggests that offline RL can be fairly sample-efficient when the behavior policy closely mimics the optimal policy in terms of the resulting state-action occupancy distribution (a scenario where $C^{\star}$ is potentially much smaller than the size of the action space).

4.3 Analysis

In this subchapter, we outline the main steps needed to establish the main results in Theorem 2 and Theorem 3. Before proceeding, let us first recall the following rescaled learning rates

$$\eta_n = \frac{H+1}{H+n} \tag{4.15}$$

for the n-th visit of a given state-action pair at a given time step h, which are adopted in both LCB-Q and LCB-Q-Advantage. For notational convenience, we further introduce two sequences of related quantities defined for any integers $N \geq 0$ and $n \geq 1$:

$$\eta_0^N := \begin{cases} \prod_{i=1}^N (1-\eta_i) = 0, & \text{if } N > 0, \\ 1, & \text{if } N = 0, \end{cases} \qquad \text{and} \qquad \eta_n^N := \begin{cases} \eta_n \prod_{i=n+1}^N (1-\eta_i), & \text{if } N > n, \\ \eta_n, & \text{if } N = n, \\ 0, & \text{if } N < n. \end{cases} \tag{4.16}$$

The following identity can be easily verified:

$$\sum_{n=0}^N \eta_n^N = 1. \tag{4.17}$$

4.3.1 Analysis of LCB-Q

To begin with, we intend to derive a recursive formula concerning the update rule of Q_h^k — the estimate of the Q-function at step h at the beginning of the k-th episode. Note that we have omitted the dependency of all quantities on the episode index k in Algorithm 4. For notational convenience and clearness, we rewrite Algorithm 4 as Algorithm 7 by specifying the dependency on the episode index k and shall often use the following set of short-hand notation when it is clear from context.

- $N_h^k(s,a)$, or the shorthand N_h^k: the number of episodes that has visited (s,a) at step h before the beginning of the k-th episode.
- $k_h^n(s,a)$, or the shorthand k^n: the index of the episode in which the state-action pair (s,a) is visited at step h for the n-th times. We also adopt the convention that $k^0 = 0$.
- $P_h^k \in \{0,1\}^{1\times S}$: a row vector corresponding to the empirical transition at step h of the k-th episode, namely,

$$P_h^k(s) = \mathbb{1}\left(s = s_{h+1}^k\right) \qquad \text{for all } s \in \mathcal{S}. \tag{4.18}$$

- $\pi^k = \{\pi_h^k\}_{h=1}^H$ with $\pi_h^k(s) \coloneqq \arg\max_a Q_h^k(s,a), \forall (h,s) \in [H] \times \mathcal{S}$: the deterministic greedy policy at the beginning of the k-th episode.
- $\widehat{\pi}$: the final output $\widehat{\pi}$ of Algorithms 4 corresponds to π^{K+1} defined above; for notational simplicity, we shall treat $\widehat{\pi}$ as π^K in our analysis, which does not affect our result at all.

Consider any state-action pair (s,a). According to the update rule in line 11 of Algorithm 7, we can express (with the assistance of the above notation)

$$Q_h^k(s,a) = Q_h^{k^{N_h^k}+1}(s,a) = \big(1-\eta_{N_h^k}\big) Q_h^{k^{N_h^k}}(s,a) + \eta_{N_h^k}\Big\{r_h(s,a) + V_{h+1}^{k^{N_h^k}}\big(s_{h+1}^{k^{N_h^k}}\big) - b_{N_h^k}\Big\}, \tag{4.19}$$

where the first identity holds since $k^{N_h^k}$ denotes the latest episode prior to k that visits (s,a) at step h, and the learning rate is defined in (4.15). Note that it always holds that $k > k^{N_h^k}$. Applying the above relation (4.19) recursively and using the notation (4.16) lead to

$$Q_h^k(s,a) = \eta_0^{N_h^k} Q_h^1(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(r_h(s,a) + V_{h+1}^{k^n}\big(s_{h+1}^{k^n}\big) - b_n\Big). \tag{4.20}$$

As another important fact, the value estimate V_h^k is monotonically non-decreasing in k, i.e.,

$$V_h^{k+1}(s) \geq V_h^k(s) \qquad \text{for all } (s,k,h) \in \mathcal{S} \times [K] \times [H], \tag{4.21}$$

which is an immediate consequence of the update rule in line 12 of Algorithm 7. Crucially, we observe that the iterate V_h^k forms a "pessimistic view" of $V_h^{\pi^k}$ — and in turn $V_h^\star$ — resulting from suitable design of the penalty term. This observation is formally stated in the following lemma, with the proof postponed to Appendix B.2.1.

Lemma 7. *Consider any $\delta \in (0,1)$, and suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$,*

$$\left| \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \Big(P_{h,s,a} - P_h^{k^n(s,a)}\Big) V_{h+1}^{k^n(s,a)} \right| \leq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n \tag{4.22}$$

holds simultaneously for all $(k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, and

$$V_h^k(s) \leq V_h^{\pi^k}(s) \leq V_h^\star(s) \tag{4.23}$$

holds simultaneously for all $(k,h,s) \in [K] \times [H] \times \mathcal{S}$.

In a nutshell, the result (4.23) in Lemma 7 reveals that V_h^k is a pointwise lower bound on $V_h^{\pi^k}$ and $V_h^\star$, thereby forming a pessimistic estimate of the optimal value function. In addition,

the property (4.22) in Lemma 7 essentially tells us that the weighted sum of the penalty terms dominates the weighted sum of the uncertainty terms, which plays a crucial role in ensuring the aforementioned pessimism property. As we shall see momentarily, Lemma 7 forms the basis of the subsequent proof.

We are now ready to embark on the analysis for LCB-Q, which is divided into multiple steps as follows.

Step 1: decomposing estimation errors. With the aid of Lemma 7, we can develop an upper bound on the performance difference of interest in (4.12) as follows

$$\begin{aligned} V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &= \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{\pi^K}(s_1)\right] \\ &\overset{\text{(i)}}{\le} \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{K}(s_1)\right] \\ &\overset{\text{(ii)}}{\le} \frac{1}{K}\sum_{k=1}^{K}\left(\mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{k}(s_1)\right]\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s) - V_1^k(s)\right), \end{aligned} \tag{4.24}$$

where (i) results from Lemma 7 (i.e., $V_1^{\pi^K}(s) \ge V_1^K(s)$ for all $s \in \mathcal{S}$), (ii) follows from the monotonicity property in (4.21), and the last equality holds since $d_1^{\pi^\star}(s) = \rho(s)$ (cf. (4.2)).

We then attempt to bound the quantity on the right-hand side of (4.24). Given that $\pi^\star$ is assumed to be a deterministic policy, we have $d_h^{\pi^\star}(s) = d_h^{\pi^\star}(s, \pi^\star(s))$. Taking this together with the relations $V_h^k(s) \ge \max_a Q_h^k(s,a) \ge Q_h^k(s, \pi_h^\star(s))$ (see line 12 of Algorithm 7) and $V_h^\star(s) = Q_h^\star(s, \pi_h^\star(s))$, we obtain

$$\begin{aligned} \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s) - V_h^k(s)\right) &= \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))\left(V_h^\star(s) - V_h^k(s)\right) \\ &\le \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s,\pi_h^\star(s))\Big(Q_h^\star\big(s,\pi_h^\star(s)\big) - Q_h^k\big(s,\pi_h^\star(s)\big)\Big) \\ &= \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left(Q_h^\star(s,a) - Q_h^k(s,a)\right) \end{aligned} \tag{4.25}$$

for any $h \in [H]$, where the last identity holds since $\pi^\star$ is deterministic and hence

$$d_h^{\pi^\star}(s,a) = 0 \qquad \text{for any } a \ne \pi_h^\star(s). \tag{4.26}$$

In view of (4.25), we need to properly control $Q_h^\star(s,a) - Q_h^k(s,a)$. By virtue of (4.17), we can

rewrite $Q_h^\star(s,a)$ as follows

$$\begin{aligned} Q_h^\star(s,a) &= \sum_{n=0}^{N_h^k} \eta_n^{N_h^k} Q_h^\star(s,a) = \eta_0^{N_h^k} Q_h^\star(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} Q_h^\star(s,a) \\ &= \eta_0^{N_h^k} Q_h^\star(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(r_h(s,a) + P_{h,s,a} V_{h+1}^\star\right), \end{aligned} \tag{4.27}$$

where the second line follows from Bellman's optimality equation (2.6). Combining (4.20) and (4.27) leads to

$$\begin{aligned} &Q_h^\star(s,a) - Q_h^k(s,a) \\ &= \eta_0^{N_h^k} \left(Q_h^\star(s,a) - Q_h^1(s,a)\right) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_{h,s,a} V_{h+1}^\star - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + b_n\right) \\ &= \eta_0^{N_h^k} \left(Q_h^\star(s,a) - Q_h^1(s,a)\right) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_n + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\big(V_{h+1}^\star - V_{h+1}^{k^n}\big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(P_{h,s,a} - P_h^{k^n}\big) V_{h+1}^{k^n} \end{aligned} \tag{4.28}$$

$$\leq \eta_0^{N_h^k} H + 2\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_n + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\big(V_{h+1}^\star - V_{h+1}^{k^n}\big), \tag{4.29}$$

where we have made use of the definition in (4.18) by recognizing $P_h^{k^n} V_{h+1}^{k^n} = V_{h+1}^{k^n}(s_{h+1}^{k^n})$ in (4.28), and the last inequality follows from the fact $Q_h^\star(s,a) - Q_h^1(s,a) = Q_h^\star(s,a) - 0 \leq H$ and the bound (4.22) in Lemma 7. Substituting the above bound into (4.25), we arrive at

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_h^k(s)\right) \\ &\leq \underbrace{\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \eta_0^{N_h^k(s,a)} H + 2\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n}_{=: I_h} \\ &\quad + \sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) P_{h,s,a} \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(V_{h+1}^\star - V_{h+1}^{k_h^n(s,a)}\big). \end{aligned} \tag{4.30}$$

Step 2: establishing a crucial recursion. As it turns out, the last term on the right-hand side of (4.30) can be used to derive a recursive relation that connects step h with step $h+1$, as summarized in the next lemma.

Lemma 8. *With probability at least $1-\delta$, the following recursion holds:*

$$\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\big(V_{h+1}^\star - V_{h+1}^{k_h^n(s,a)}\big)$$
$$\le \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right) + 24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}. \tag{4.31}$$

Lemma 8 taken together with (4.30) implies that

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right) \le \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right)$$
$$+ I_h + 24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}. \tag{4.32}$$

Invoking (4.32) recursively over the time steps $h=H,H-1,\cdots,1$ with the terminal condition $V_{H+1}^k = V_{H+1}^\star = 0$, we reach

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s)-V_1^k(s)\right) \le \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)$$
$$\le \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(I_h + 24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}\right), \tag{4.33}$$

which captures the estimation error resulting from the use of pessimism principle.

Step 3: controlling the right-hand side of (4.33). The right-hand side of (4.33) can be bounded through the following lemma, which will be proved in Appendix B.2.3.

Lemma 9. *Consider any $\delta\in(0,1)$. With probability at least $1-\delta$, we have*

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(I_h + 24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}\right) \lesssim H^2SC^\star\iota + \sqrt{H^5SC^\star K\iota^3}, \tag{4.34}$$

where we recall that $\iota \coloneqq \log\left(\frac{SAT}{\delta}\right)$.

Combining Lemma 9 with (4.33) and (4.24) yields

$$\begin{aligned} V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &\le \frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s) - V_1^k(s)\right) \\ &\le \frac{1}{K}\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s) - V_h^k(s)\right) \\ &\le \frac{c_{\mathrm{a}}}{2}\sqrt{\frac{H^5SC^\star\iota^3}{K}} + \frac{c_{\mathrm{a}}}{2}\frac{H^2SC^\star\iota}{K} = \frac{c_{\mathrm{a}}}{2}\sqrt{\frac{H^6SC^\star\iota^3}{T}} + \frac{c_{\mathrm{a}}}{2}\frac{H^3SC^\star\iota}{T} \\ &\le c_{\mathrm{a}}\sqrt{\frac{H^6SC^\star\iota^3}{T}} \end{aligned} \tag{4.35}$$

for some sufficiently large constant $c_{\mathrm{a}} > 0$, where the last inequality is valid as long as $T > SC^\star\iota$. This concludes the proof of Theorem 2.

4.3.2 Analysis of LCB-Q-Advantage

We now turn to the analysis of LCB-Q-Advantage. Thus far, we have omitted the dependency of all quantities on the epoch number m and the in-epoch episode number t in Algorithms 5 and 6. While it allows for a more concise description of our algorithm, it might hamper the clarity of our proofs. In the following, we introduce the notation k to denote the current episode as follows:

$$k := \sum_{i=1}^{m-1} L_i + t, \tag{4.36}$$

which corresponds to the t-th in-epoch episode in the m-th epoch; here, $L_m = 2^m$ stands for the total number of in-epoch episodes in the m-th epoch. With this notation in place, we can rewrite Algorithm 5 as Algorithm 8 in order to make clear the dependency on the episode index k, epoch number m, and in-epoch episode index t.

Before embarking on our main proof, we make two crucial observations which play important roles in our subsequent analysis. First, similar to the property (4.21) for LCB-Q, the update rule (cf. lines 16-17 of Algorithm 8) ensures the monotonic non-decreasing property of $V_h(s)$ such that for all $k \in [K]$,

$$V_h^{k+1}(s) \ge V_h^k(s), \qquad \text{for all } (k,s,h) \in [K]\times\mathcal{S}\times[H]. \tag{4.37}$$

Secondly, V_h^k forms a "pessimistic view" of $V_h^\star$, which is formalized in the lemma below; the proof is deferred to Appendix B.3.1.

Lemma 10. *Let $\delta \in (0,1)$. Suppose that $c_{\mathrm{b}} > 0$ is some sufficiently large constant. Then with probability at least $1-\delta$, the value estimates produced by Algorithm 5 satisfy*

$$V_h^k(s) \le V_h^{\pi^k}(s) \le V^\star(s) \tag{4.38}$$

for all $(k,h,s) \in [K] \times [H+1] \times \mathcal{S}$.

With these two observations in place, we can proceed to present the analysis for LCB-Q-Advantage. To begin with, the performance difference of interest can be controlled similar to (4.24) as follows:

$$\begin{aligned} V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &= \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{\pi^K}(s_1)\right] \\ &\overset{\text{(i)}}{\le} \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{K}(s_1)\right] \\ &\overset{\text{(ii)}}{\le} \frac{1}{K}\sum_{k=1}^{K}\left(\mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^\star(s_1)\right] - \mathop{\mathbb{E}}_{s_1\sim\rho}\left[V_1^{k}(s_1)\right]\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s) - V_1^k(s)\right), \end{aligned} \tag{4.39}$$

where (i) follows from Lemma 10 (i.e., $V_1^{\pi^K}(s) \ge V_1^K(s)$ for all $s \in \mathcal{S}$), (ii) holds due to the monotonicity in (4.37) and the last equality holds since $d_1^{\pi^\star}(s) = \rho(s)$ (cf. (4.2)). It then boils down to controlling the right-hand side of (4.39). Towards this end, it turns out that one can control a more general counterpart, i.e.,

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s) - V_h^k(s)\right) \tag{4.40}$$

for any $h \in [H]$. This is accomplished via the following lemma, whose proof is postponed to Appendix B.3.2.

Lemma 11. *Let* $\delta \in (0,1)$, *and recall that* $\iota := \log\left(\frac{SAT}{\delta}\right)$. *Suppose that* $c_{\mathrm{a}}, c_{\mathrm{b}} > 0$ *are some sufficiently large constants. Then with probability at least* $1-\delta$, *one has*

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s) - V_h^k(s)\right) \le J_h^1 + J_h^2 + J_h^3, \tag{4.41}$$

where

$$J_h^1 := \sum_{k=1}^{K}\sum_{s,a\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left[\eta_0^{N_h^k(s,a)}H + \frac{4c_{\mathrm{b}}H^{7/4}\iota}{\left(N_h^k(s,a)\vee 1\right)^{3/4}} + \frac{4c_{\mathrm{b}}H^2\iota}{N_h^k(s,a)\vee 1}\right],$$

$$J_h^2 := 2\sum_{k=1}^{K}\sum_{s,a\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\overline{B}_h^k(s,a),$$

$$J_h^3 := \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right) + 48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2. \tag{4.42}$$

As a direct consequence of Lemma 11, one arrives at a recursive relationship between time steps h and $h+1$ as follows:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)\\
&\le \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right) + 48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2 + J_h^1 + J_h^2.
\end{aligned} \tag{4.43}$$

Recursing over time steps $h = H, H-1, \cdots, 1$ with the terminal condition $V_{H+1}^k = V_{H+1}^\star = 0$, we can upper bound the performance difference at $h=1$ as follows

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_1^{\pi^\star}(s)\left(V_1^\star(s)-V_1^k(s)\right)\\
&\le \max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)\\
&\le \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2 + J_h^1 + J_h^2\right).
\end{aligned} \tag{4.44}$$

To finish up, it suffices to upper bound each term in (4.44) separately. We summarize their respective upper bounds as follows; the proof is provided in Appendix B.3.3.

Lemma 12. *Fix $\delta\in(0,1)$, and recall that $\iota := \log\left(\frac{SAT}{\delta}\right)$. With probability at least $1-\delta$, we have*

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} J_h^1 \lesssim H^{2.75}(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2 + H^3SC^\star\iota^3, \tag{4.45a}$$

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} J_h^2 \lesssim \sqrt{H^4SC^\star\iota^3\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)} + \sqrt{H^3SC^\star K\iota^5} + H^4SC^\star\iota^4, \tag{4.45b}$$

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2\right) \lesssim \sqrt{H^3C^\star K\log\frac{2H}{\delta}} + H^4C^\star\sqrt{S}\iota^2. \tag{4.45c}$$

Substituting the above upper bounds into (4.39) and (4.44) and recalling that $T = HK$, we arrive at

$$
\begin{aligned}
V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) &\lesssim \frac{1}{K} \max_{h\in[H]} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_h^k(s)\right) \\
&\lesssim \frac{1}{K} \left(\sqrt{H^4 SC^\star \iota^3 \max_{h\in[H]} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_h^k(s)\right)} + \left(\sqrt{H^3 SC^\star K\iota^5} + H^4 SC^\star \iota^4 + H^{2.75}(SC^\star)^{\frac{3}{4}} K^{\frac{1}{4}} \iota^2\right) \right) \\
&\overset{\text{(i)}}{\asymp} \frac{1}{K} \left(\sqrt{H^4 SC^\star \iota^3 \max_{h\in[H]} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_h^k(s)\right)} + \sqrt{H^3 SC^\star K\iota^5} + H^4 SC^\star \iota^4 \right) \\
&\overset{\text{(ii)}}{\lesssim} \frac{1}{K} \left(\sqrt{H^3 SC^\star K\iota^5} + H^4 SC^\star \iota^4\right) \\
&\asymp \sqrt{\frac{H^4 SC^\star \iota^5}{T}} + \frac{H^5 SC^\star \iota^4}{T},
\end{aligned}
$$

where (i) has made use of the AM-GM inequality:

$$
2H^{2.75}(SC^\star)^{\frac{3}{4}} K^{\frac{1}{4}} \le \left(H^{0.75}(SC^\star)^{\frac{1}{4}} K^{\frac{1}{4}}\right)^2 + \left(H^2 (SC^\star)^{\frac{1}{2}}\right)^2 = \sqrt{H^3 SC^\star K} + H^4 SC^\star,
$$

and (ii) holds by letting $x \coloneqq \max_{h\in[H]} \sum_{k=1}^{K} \sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s) \left(V_h^\star(s) - V_h^k(s)\right)$ and solving the inequality $x \lesssim \sqrt{H^4 SC^\star \iota^3 x} + \sqrt{H^3 SC^\star K\iota^5} + H^4 SC^\star \iota^4$. This concludes the proof.

4.4 Discussions

Focusing on model-free paradigms, in this chapter, we developed near-optimal sample complexities for some variants of pessimistic Q-learning algorithms — armed with lower confidence bounds and variance reduction — for offline RL. These sample complexity results, taken together with the analysis framework developed herein, open up a few exciting directions for future research. For example, the pessimistic Q-learning algorithms can be deployed in conjunction with their optimistic counterparts (e.g., Jin et al. (2018); Li et al. (2023b); Zhang et al. (2020c)), when additional online data can be acquired to fine-tune the policy (Xie et al., 2021b).

Algorithm 5: Offline LCB-Q-Advantage RL

Parameters: number of epochs M, universal constant $c_{\mathrm{b}} > 0$, probability of failure $\delta \in (0,1)$, and $\iota = \log\left(\frac{SAT}{\delta}\right)$;

Initialize:

$Q_h(s,a), Q_h^{\mathsf{LCB}}(s,a), \overline{Q}_h(s,a), \overline{\mu}_h(s,a), \overline{\mu}_h^{\mathsf{next}}(s,a), N_h(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;

$V_h(s), \overline{V}_h(s), \overline{V}_h^{\mathsf{next}}(s) \leftarrow 0$ for all $(s,h) \in \mathcal{S} \times [H+1]$;

$\mu_h^{\mathsf{ref}}(s,a), \sigma_h^{\mathsf{ref}}(s,a), \mu_h^{\mathsf{adv}}(s,a), \sigma_h^{\mathsf{adv}}(s,a), \overline{\delta}_h(s,a), \overline{B}_h(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

for *Epoch* $m = 1$ **to** M **do**

  $L_m = 2^m$; // specify the number of episodes in the current epoch

  $\widehat{N}_h(s,a) = 0$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$. // reset the epoch-wise counter

  /* Inner-loop: update value-estimates $V_h(s,a)$ and Q-estimates $Q_h(s,a)$ */

  for *In-epoch Episode* $t = 1$ **to** L_m **do**

    Sample a new trajectory $\{s_h, a_h, r_h\}_{h=1}^H$. // sampling from batch dataset

    for *Step* $h = 1$ **to** H **do**

      $N_h(s_h,a_h) \leftarrow N_h(s_h,a_h) + 1$; $n \leftarrow N_h(s_h,a_h)$. // update the overall counter

      $\eta_n \leftarrow \frac{H+1}{H+n}$; // update the learning rate

      // run the Q-learning update rule with LCB

      $Q_h^{\mathsf{LCB}}(s_h,a_h) \leftarrow$ `update-lcb-q()`.

      // update the Q-estimate with LCB and reference-advantage

      $\overline{Q}_h(s_h,a_h) \leftarrow$ `update-lcb-q-ra()`.

      // update the Q-estimate Q_h and value estimate V_h

      $Q_h(s_h,a_h) \leftarrow \max\left\{Q_h^{\mathsf{LCB}}(s_h,a_h), \overline{Q}_h(s_h,a_h), Q_h(s_h,a_h)\right\}$.

      $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$.

      // update the epoch-wise counter and $\overline{\mu}_h^{\mathsf{next}}$ for the next epoch

      $\widehat{N}_h(s_h,a_h) \leftarrow \widehat{N}_h(s_h,a_h) + 1$;

      $\overline{\mu}_h^{\mathsf{next}}(s_h,a_h) \leftarrow \left(1 - \frac{1}{\widehat{N}_h(s_h,a_h)}\right)\overline{\mu}_h^{\mathsf{next}}(s_h,a_h) + \frac{1}{\widehat{N}_h(s_h,a_h)}\overline{V}_{h+1}^{\mathsf{next}}(s_{h+1})$;

  /* Update the reference $(\overline{V}_h, \overline{V}_h^{\mathsf{next}})$ and $(\overline{\mu}_h, \overline{\mu}_h^{\mathsf{next}})$ */

  for $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H+1]$ **do**

    $\overline{V}_h(s) \leftarrow \overline{V}_h^{\mathsf{next}}(s)$; $\overline{\mu}_h(s,a) \leftarrow \overline{\mu}_h^{\mathsf{next}}(s,a)$. // set $\overline{V}_h$ and $\overline{\mu}_h$ for the next epoch

    $\overline{V}_h^{\mathsf{next}}(s) \leftarrow V_h(s)$; $\overline{\mu}_h^{\mathsf{next}}(s,a) \leftarrow 0$. // restart $\overline{\mu}_h^{\mathsf{next}}$ and set $\overline{V}_h^{\mathsf{next}}$ for the next epoch

Output: the policy $\widehat{\pi}$ s.t. $\widehat{\pi}_h(s) = \arg\max_a Q_h(s,a)$ for any $(s,h) \in \mathcal{S} \times [H]$.

Algorithm 6: Auxiliary functions

Function `update-lcb-q()`:

$Q_h^{\mathsf{LCB}}(s_h,a_h) \leftarrow (1-\eta_n)Q_h^{\mathsf{LCB}}(s_h,a_h) + \eta_n\big(r(s_h,a_h) + V_{h+1}(s_{h+1}) - c_{\mathrm{b}}\sqrt{\frac{H^3\iota^2}{n}}\big)$.

Function `update-lcb-q-ra()`:

`/* update the moment statistics of the interested terms */`

$[\mu_h^{\mathsf{ref}}, \sigma_h^{\mathsf{ref}}, \mu_h^{\mathsf{adv}}, \sigma_h^{\mathsf{adv}}](s_h,a_h) \leftarrow$ `update-moments()`;

`/* update the bonus difference and accumulative bonus */`

$[\overline{\delta}_h, \overline{B}_h](s_h,a_h) \leftarrow$ `update-bonus()`;

$\overline{b}_h(s_h,a_h) \leftarrow \overline{B}_h(s_h,a_h) + (1-\eta_n)\frac{\overline{\delta}_h(s_h,a_h)}{\eta_n} + c_{\mathrm{b}}\frac{H^{7/4}\iota}{n^{3/4}} + c_{\mathrm{b}}\frac{H^2\iota}{n}$;

`/* update the Q-estimate based on reference-advantage */`

$\overline{Q}_h(s_h,a_h) \leftarrow$
$(1-\eta_n)\overline{Q}_h(s_h,a_h) + \eta_n\big(r_h(s_h,a_h) + V_{h+1}(s_{h+1}) - \overline{V}_{h+1}(s_{h+1}) + \overline{\mu}_h(s_h,a_h) - \overline{b}_h\big)$;

Function `update-moments()`:

$\mu_h^{\mathsf{ref}}(s_h,a_h) \leftarrow (1-\frac{1}{n})\mu_h^{\mathsf{ref}}(s_h,a_h) + \frac{1}{n}\overline{V}_{h+1}^{\mathsf{next}}(s_{h+1})$; // mean of the reference

$\sigma_h^{\mathsf{ref}}(s_h,a_h) \leftarrow (1-\frac{1}{n})\sigma_h^{\mathsf{ref}}(s_h,a_h) + \frac{1}{n}\big(\overline{V}_{h+1}^{\mathsf{next}}(s_{h+1})\big)^2$; // 2^{nd} moment of the reference

$\mu_h^{\mathsf{adv}}(s_h,a_h) \leftarrow (1-\eta_n)\mu_h^{\mathsf{adv}}(s_h,a_h) + \eta_n\big(V_{h+1}(s_{h+1}) - \overline{V}_{h+1}(s_{h+1})\big)$; // mean of the advantage

$\sigma_h^{\mathsf{adv}}(s_h,a_h) \leftarrow (1-\eta_n)\sigma_h^{\mathsf{adv}}(s_h,a_h) + \eta_n\big(V_{h+1}(s_{h+1}) - \overline{V}_{h+1}(s_{h+1})\big)^2$. // 2^{nd} moment of the advantage

Function `update-bonus()`:

$B_h^{\mathsf{next}}(s_h,a_h) \leftarrow$
$c_{\mathrm{b}}\sqrt{\frac{\iota}{n}}\Big(\sqrt{\sigma_h^{\mathsf{ref}}(s_h,a_h) - \big(\mu_h^{\mathsf{ref}}(s_h,a_h)\big)^2} + \sqrt{H}\sqrt{\sigma_h^{\mathsf{adv}}(s_h,a_h) - \big(\mu_h^{\mathsf{adv}}(s_h,a_h)\big)^2}\Big)$;

$\overline{\delta}_h(s_h,a_h) \leftarrow B_h^{\mathsf{next}}(s_h,a_h) - \overline{B}_h(s_h,a_h)$;

$\overline{B}_h(s_h,a_h) \leftarrow B_h^{\mathsf{next}}(s_h,a_h)$.

Algorithm 7: LCB-Q for offline RL (a rewrite of Algorithm 4 to specify dependency on k)

Parameters: some constant $c_{\mathrm{b}} > 0$, target success probability $1-\delta \in (0,1)$, and $\iota = \log\left(\frac{SAT}{\delta}\right)$.

Initialize $Q_h^1(s,a) \leftarrow 0$; $N_h^1(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$; $V_h^1(s) \leftarrow 0$ for all $(s,h) \in \mathcal{S} \times [H+1]$; π^1 s.t. $\pi_h^1(s) = 1$ for all $(s,h) \in \mathcal{S} \times [H]$.

for *Episode* $k = 1$ **to** K **do**

  Sample the k-th trajectory $\{s_h^k, a_h^k, r_h^k\}_{h=1}^H$ from $\mathcal{D}$. `// sampling from batch dataset`

  for *Step* $h = 1$ **to** H **do**

    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**

      `// carry over the estimates and policy`

      $N_h^{k+1}(s,a) \leftarrow N_h^k(s,a)$; $Q_h^{k+1}(s,a) \leftarrow Q_h^k(s,a)$; $V_h^{k+1}(s) \leftarrow V_h^k(s)$; $\pi_h^{k+1}(s) \leftarrow \pi_h^k(s)$.

    $N_h^{k+1}(s_h^k, a_h^k) \leftarrow N_h^k(s_h^k, a_h^k) + 1$. `// update the counter`

    $n \leftarrow N_h^{k+1}(s_h^k, a_h^k)$; $\eta_n \leftarrow \frac{H+1}{H+n}$. `// update the learning rate`

    $b_n \leftarrow c_{\mathrm{b}}\sqrt{\frac{H^3\iota^2}{n}}$. `// update the bonus term`

    `// update the Q-estimates with LCB`

    $Q_h^{k+1}(s_h^k, a_h^k) \leftarrow Q_h^k(s_h^k, a_h^k) + \eta_n\Big\{r_h(s_h^k, a_h^k) + V_{h+1}^k(s_{h+1}^k) - Q_h^k(s_h^k, a_h^k) - b_n\Big\}$.

    `// update the value estimates`

    $V_h^{k+1}(s_h^k) \leftarrow \max\Big\{V_h^k(s_h^k),\, \max_a Q_h^{k+1}(s_h^k, a)\Big\}$.

    `// update the policy`

    If $V_h^{k+1}(s_h^k) = \max_a Q_h^{k+1}(s_h^k, a)$: update $\pi_h^{k+1}(s_h^k) = \arg\max_a Q_h^{k+1}(s_h^k, a)$.

Algorithm 8: LCB-Q-Advantage (a rewrite of Algorithm 5 that specifies dependency on k or (m,t).)

Parameters: number of epochs M, universal constant $c_{\mathrm{b}} > 0$, target success probability $1-\delta \in (0,1)$, and $\iota = \log\left(\frac{SAT}{\delta}\right)$.

Initialize:

$Q_h^1(s,a), Q_h^{\mathsf{LCB},1}(s,a), \overline{Q}_h^1(s,a), \overline{\mu}_h^1(s,a), \overline{\mu}_h^{\mathsf{next},1}(s,a), N_h^1(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$;

$V_h^1(s), \overline{V}_h^1(s), \overline{V}_h^{\mathsf{next},1}(s) \leftarrow 0$ for all $(s,h) \in \mathcal{S}\times[H+1]$;

$\mu_h^{\mathsf{ref},1}(s,a), \sigma_h^{\mathsf{ref},1}(s,a), \mu_h^{\mathsf{adv},1}(s,a), \sigma_h^{\mathsf{adv},1}(s,a), \overline{\delta}_h^1(s,a), \overline{B}_h^1(s,a) \leftarrow 0$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.

for *Epoch* $m = 1$ **to** M **do**

  $L_m = 2^m$. // specify the number of episodes in the current epoch

  $\widehat{N}_h^{(m,1)}(s,a) = 0$ for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$. // reset the epoch-wise counter

  /* Inner-loop: update value-estimates $V_h(s,a)$ and Q-estimates $Q_h(s,a)$ */

  for *In-epoch Episode* $t = 1$ **to** L_m **do**

    Set $k \leftarrow \sum_{i=1}^{m-1} L_i + t$. // set the episode index

    Sample the k-th trajectory $\{s_h^k, a_h^k, r_h^k\}_{h=1}^H$. // sampling from batch dataset

    Compute π^k s.t. $\pi_h^k(s) = \arg\max_a Q_h^k(s,a)$ for all $(s,h) \in \mathcal{S}\times[H]$. // update the policy

    for *Step* $h = 1$ **to** H **do**

      for $(s,a) \in \mathcal{S}\times\mathcal{A}$ **do**

        // carry over the estimates

        $N_h^{k+1}(s,a) \leftarrow N_h^k(s,a)$; $\widehat{N}_h^{k+1}(s,a) \leftarrow \widehat{N}_h^k(s,a)$; $V_h^{k+1}(s) \leftarrow V_h^k(s)$;

        $Q_h^{\mathsf{LCB},k+1}(s,a) \leftarrow Q_h^{\mathsf{LCB},k}(s,a)$ $\overline{Q}_h^{k+1}(s,a) \leftarrow \overline{Q}_h^k(s,a)$;

        $Q_h^{k+1}(s,a) \leftarrow Q_h^k(s,a)$; $\overline{V}_h^{k+1}(s) \leftarrow \overline{V}_h^k(s)$ $\overline{V}_h^{\mathsf{next},k+1}(s) \leftarrow \overline{V}_h^{\mathsf{next},k}(s)$;

        $\overline{\mu}^{k+1}(s,a) \leftarrow \overline{\mu}^k(s,a)$.

      $N_h^{k+1}(s_h^k,a_h^k) \leftarrow N_h^k(s_h^k,a_h^k) + 1$; $n \leftarrow N_h^{k+1}(s_h^k,a_h^k)$. // update the overall counter

      $\eta_n \leftarrow \frac{H+1}{H+n}$. // update the learning rate

      // update the Q-estimate with LCB

      $Q_h^{\mathsf{LCB},k+1}(s_h^k,a_h^k) \leftarrow$ `update-lcb-q()`.

      // update the Q-estimate with LCB and reference-advantage

      $\overline{Q}_h^{k+1}(s_h^k,a_h^k) \leftarrow$ `update-lcb-q-ra()`.

      // update the Q-estimate Q_h and value estimate V_h

      $Q_h^{k+1}(s_h^k,a_h^k) \leftarrow \max\left\{Q_h^{\mathsf{LCB},k+1}(s_h^k,a_h^k), \overline{Q}_h^{k+1}(s_h^k,a_h^k), Q_h^k(s_h^k,a_h^k)\right\}$.

      $V_h^{k+1}(s_h^k) \leftarrow \max_a Q_h^{k+1}(s_h^k,a)$.

      // update epoch-wise counter and $\overline{\mu}_h^{\mathsf{next}}(s,a)$ for the next epoch

      $\widehat{N}_h^{(m,t+1)}(s_h^k,a_h^k) \leftarrow \widehat{N}_h^{(m,t)}(s_h^k,a_h^k) + 1$.

      $\overline{\mu}_h^{\mathsf{next},k+1}(s_h^k,a_h^k) \leftarrow$

      $$\left(1 - \frac{1}{\widehat{N}_h^{(m,t+1)}(s_h^k,a_h^k)}\right)\overline{\mu}_h^{\mathsf{next},k}(s_h,a_h) + \frac{1}{\widehat{N}_h^{(m,t+1)}(s_h^k,a_h^k)}\overline{V}_{h+1}^{\mathsf{next},k}(s_{h+1}).$$

  /* Update the reference ($\overline{V}_h$, $\overline{V}_h^{\mathsf{next}}$) and ($\overline{\mu}_h$, $\overline{\mu}_h^{\mathsf{next}}$) */

  for $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H+1]$ **do**

    $\overline{V}_h^{k+1}(s) \leftarrow \overline{V}_h^{\mathsf{next},k+1}(s)$; $\overline{\mu}_h^{k+1}(s,a) \leftarrow \overline{\mu}_h^{\mathsf{next},k+1}(s,a)$. // set $\overline{V}_h$ and $\overline{\mu}_h$ for the next epoch

    $\overline{V}_h^{\mathsf{next},k+1}(s) \leftarrow V_h^{k+1}(s)$; $\overline{\mu}_h^{\mathsf{next},k+1}(s,a) \leftarrow 0$. // set $\overline{\mu}_h^{\mathsf{next}}$ and $\overline{V}_h^{\mathsf{next}}$ for the next epoch

Output: the policy $\widehat{\pi} = \pi^K$ with $K = \sum_{m=1}^M L_m$.

Chapter 5

Model-Based Offline RL

5.1 Algorithm and theory: episodic finite-horizon MDPs

We begin by studying offline RL in episodic finite-horizon MDPs, which follows the same problem formulation as Chapter 4.1. In the following, we shall first introduce a slightly improved version of the single-policy concentrability (cf. Definition 1), followed by algorithm design and main results.

5.1.1 A refined single-policy concentrability $C^\star_{\mathsf{clipped}}$

Let us begin with recalling the formulation of the concrete setting in Chapter 4.1. Throughout this chapter, we denote $\rho = \rho^{\mathsf{b}}$ stands for some predetermined initial state distribution associated with the batch dataset. For notational simplicity, we introduce the following short-hand notation for the occupancy distribution w.r.t. the behavior policy π^{b}:

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]: \qquad d_h^{\mathsf{b}}(s) \coloneqq d_h^{\pi^{\mathsf{b}}}(s) \quad \text{and} \quad d_h^{\mathsf{b}}(s,a) \coloneqq d_h^{\pi^{\mathsf{b}}}(s,a). \tag{5.1}$$

In particular, it is easily seen that $d_1^{\mathsf{b}}(s) = \rho^{\mathsf{b}}(s)$ for all $s \in \mathcal{S}$. Note that the initial state distribution ρ^{b} of the batch dataset might not coincide with the test state distribution ρ.

Then, recall Definition 1, the introduced concentrability coefficient to capture the distribution shift between the desired distribution and the one induced by the behavior policy.

$C^\star$ employs the largest density ratio (using the occupancy distributions defined above) to measure the distribution mismatch; it concerns the behavior policy vs. a single policy $\pi^\star$, and does not require uniform coverage of the state-action space (namely, it suffices to cover the part reachable by $\pi^\star$). We further introduce a slightly modified version of $C^\star$ as follows.

Definition 2 (Single-policy clipped concentrability for finite-horizon MDPs)**.** The single-policy clipped concentrability coefficient of a batch dataset $\mathcal{D}$ is defined as

$$C^\star_{\mathsf{clipped}} \coloneqq \max_{(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]} \frac{\min\left\{d_h^\star(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b}}(s,a)}. \tag{5.2}$$

From the definition above, it holds trivially that

$$C^\star_{\mathsf{clipped}} \le C^\star \qquad \text{and} \qquad C^\star_{\mathsf{clipped}} \ge \frac{1}{S}. \tag{5.3}$$

As we shall see shortly, while all sample complexity upper bounds developed herein remain valid if we replace $C^\star_{\mathsf{clipped}}$ with $C^\star$, the use of $C^\star_{\mathsf{clipped}}$ might yield some sample size reduction when $C^\star_{\mathsf{clipped}}$ drops below 1.

Goal. With the above batch dataset $\mathcal{D}$ in hand, our aim is to compute, in a sample-efficient fashion, a policy $\widehat{\pi}$ that results in near-optimal values w.r.t. a given test state distribution $\rho \in \Delta(\mathcal{S})$. Formally speaking, the current sub-chapter focuses on achieving

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \leq \varepsilon$$

with high probability using as few samples as possible, where ε stands for the target accuracy level. We seek to achieve sample optimality for the full ε-range, i.e., for any $\varepsilon \in (0, H]$.

5.1.2 A model-based offline RL algorithm: VI-LCB

Suppose for the moment that we have access to a dataset $\mathcal{D}_0$ containing N sample transitions $\{(s_i, a_i, h_i, s_i')\}_{i=1}^N$, where (s_i, a_i, h_i, s_i') denotes the transition from state s_i at step h_i to state s_i' in the next step when action a_i is taken. We now describe a pessimistic variant of the model-based approach on the basis of $\mathcal{D}_0$.

Empirical MDP. For each $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we denote by

$$N_h(s,a) \coloneqq \sum_{i=1}^N \mathbb{1}\big\{(s_i, a_i, h_i) = (s, a, h)\big\} \tag{5.4a}$$

$$N_h(s) \coloneqq \sum_{i=1}^N \mathbb{1}\big\{(s_i, h_i) = (s, h)\big\} \tag{5.4b}$$

the total number of sample transitions at step h that transition from (s, a) and from s, respectively. We can then compute the empirical estimate $\widehat{P} = \{\widehat{P}_h\}_{1\leq h\leq H}$ of the transition kernel P as follows:

$$\widehat{P}_h(s' \,|\, s, a) = \begin{cases} \frac{1}{N_h(s,a)} \sum\limits_{i=1}^N \mathbb{1}\big\{(s_i, a_i, h_i, s_i') = (s, a, h, s')\big\}, & \text{if } N_h(s,a) > 0 \\ \frac{1}{S}, & \text{else} \end{cases} \tag{5.5}$$

for each $(s, a, h, s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$.

The VI-LCB algorithm. With this estimated model in place, the VI-LCB algorithm (i.e., value iteration with lower confidence bounds) maintains the value function estimate $\{\widehat{V}_h\}$ and Q-function estimate $\{\widehat{Q}_h\}$, and works backward from $h = H$ to $h = 1$ as in classical dynamic programming

Algorithm 9: Offline value iteration with LCB (VI-LCB) for finite-horizon MDPs.

input: dataset $\mathcal{D}_0$; reward function r; target success probability $1-\delta$.
initialization: $\widehat{V}_{H+1} = 0$.
for $h = H, \cdots, 1$ **do**
  compute the empirical transition kernel $\widehat{P}_h$ according to (5.5).
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    compute the penalty term $b_h(s,a)$ according to (5.9).
    set $\widehat{Q}_h(s,a) = \max\left\{r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a), 0\right\}$.
  for $s \in \mathcal{S}$ **do**
    set $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) \in \arg\max_a \widehat{Q}_h(s,a)$.
output: $\widehat{\pi} = \{\widehat{\pi}_h\}_{1\le h\le H}$.

with the terminal value $\widehat{V}_{H+1} = 0$ (Jin et al., 2021; Xie et al., 2021b). Specifically, the algorithm adopts the following update rule:

$$\widehat{Q}_h(s,a) = \max\left\{r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a), 0\right\}, \tag{5.6}$$

where $\widehat{P}_{h,s,a}$ is the empirical estimate of $P_{h,s,a}$ (cf. (2.7)),

$$\widehat{V}_{h+1}(s) = \max_a \widehat{Q}_{h+1}(s,a), \tag{5.7}$$

and $b_h(s,a) \ge 0$ denotes some penalty term that is a decreasing function in $N_h(s,a)$ (as we shall specify momentarily). In addition, the policy $\widehat{\pi}$ is selected greedily in accordance to the Q-estimate:

$$\forall (s,h) \in \mathcal{S}\times[H]: \qquad \widehat{\pi}_h(s) \in \arg\max_a \widehat{Q}_h(s,a). \tag{5.8}$$

In a nutshell, the VI-LCB algorithm — as summarized in Algorithm 9 — applies the classical value iteration approach to the empirical model $\widehat{P}$, and in addition, implements the principle of pessimism via certain lower confidence penalty terms $\{b_h(s,a)\}$.

The Bernstein-style penalty terms. As before, we adopt Bernstein-style penalty in order to better capture the variance structure over time; that is,

$$\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]: \ b_h(s,a) = \min\left\{\sqrt{\frac{c_{\mathrm{b}}\log\frac{NH}{\delta}}{N_h(s,a)}\mathsf{Var}_{\widehat{P}_{h,s,a}}\big(\widehat{V}_{h+1}\big)} + c_{\mathrm{b}}H\frac{\log\frac{NH}{\delta}}{N_h(s,a)},\ H\right\} \tag{5.9}$$

for some universal constant $c_{\mathrm{b}} > 0$ (e.g., $c_{\mathrm{b}} = 16$). Here, $\mathsf{Var}_{\widehat{P}_{h,s,a}}\big(\widehat{V}_{h+1}\big)$ corresponds to the variance of $\widehat{V}_{h+1}$ w.r.t. the distribution $\widehat{P}_{h,s,a}$ (see the definition (1.7)). Note that we choose $\widehat{P}$ as opposed to P (i.e., $\mathsf{Var}_{P_{h,s,a}}\big(\widehat{V}_{h+1}\big)$) in the variance term, mainly because we have no access to the true

transition kernel P.

Finally, it is worth noting that the Bernstein-style uncertainty estimates have been widely studied when performing online exploration in episodic finite-horizon MDPs (e.g., Azar et al. (2017); Fruit et al. (2020); Jin et al. (2018); Li et al. (2023b); Talebi and Maillard (2018); Zhang et al. (2020c)). Once again, the main purpose therein is to encourage exploration of the insufficiently visited states/actions, a mechanism that is not applicable to offline RL due to the absence of further data collection.

5.1.3 VI-LCB with two-fold subsampling

Given that the batch dataset $\mathcal{D}$ is composed of several sample trajectories each of length H, the sample transitions in $\mathcal{D}$ cannot be viewed as being independently generated (as the sample transitions at step h might influence the sample transitions in the subsequent steps). As one can imagine, the presence of such temporal statistical dependency considerably complicates analysis.

In order to circumvent this technical difficulty, we propose a two-fold subsampling trick that allows one to exploit the desired statistical independence. Informally, we propose the following steps:

- First of all, we randomly split the dataset into two halves $\mathcal{D}^{\mathsf{main}}$ and $\mathcal{D}^{\mathsf{aux}}$, where $\mathcal{D}^{\mathsf{main}}$ consists of $N_h^{\mathsf{main}}(s)$ sample transitions from state s at step h.
- For each $(s,h) \in \mathcal{S} \times [H]$, we use the dataset $\mathcal{D}^{\mathsf{aux}}$ to construct a high-probability lower bound $N_h^{\mathsf{trim}}(s)$ on $N_h^{\mathsf{main}}(s)$, and then subsample $N_h^{\mathsf{trim}}(s)$ sample transitions w.r.t. (s,h) from $\mathcal{D}^{\mathsf{main}}$; this results in a new subsampled dataset $\mathcal{D}^{\mathsf{trim}}$.
- Run VI-LCB on the subsampled dataset $\mathcal{D}^{\mathsf{trim}}$ (i.e., Algorithm 9).

The whole procedure is detailed in Algorithm 10. A few important features are worth highlighting, under the assumption that the sample trajectories in $\mathcal{D}$ are independently generated from the same distribution.

- Given that $\{N_h^{\mathsf{trim}}(s)\}$ are computed on the basis of the dataset $\mathcal{D}^{\mathsf{aux}}$ and that $\mathcal{D}^{\mathsf{trim}}$ is subsampled from another dataset $\mathcal{D}^{\mathsf{main}}$, one can clearly see that $\{N_h^{\mathsf{trim}}(s)\}$ are statistically independent from the sample transitions in $\mathcal{D}^{\mathsf{trim}}$.
- As we shall justify in the analysis, the samples in $\mathcal{D}^{\mathsf{trim}}$ can almost be treated as being statistically independent, a key attribute resulting from the subsampling trick.
- The proposed algorithm only splits the data into two subsets, which is in stark contrast to prior variants of VI-LCB that perform H-fold sample splitting (e.g., Xie et al. (2021b)). Eliminating the H-fold splitting requirement plays a crucial role in enabling optimal sample complexity.

Algorithm 10: Subsampled VI-LCB for episodic finite-horizon MDPs

input: a dataset $\mathcal{D}$; reward function r.

subsampling: run the following procedure to generate the subsampled dataset $\mathcal{D}^{\mathsf{trim}}$.

1) *Data splitting.* Split $\mathcal{D}$ into two halves: $\mathcal{D}^{\mathsf{main}}$ (which contains the first $K/2$ trajectories), and $\mathcal{D}^{\mathsf{aux}}$ (which contains the remaining $K/2$ trajectories); we let $N_h^{\mathsf{main}}(s)$ (resp. $N_h^{\mathsf{aux}}(s)$) denote the number of sample transitions in $\mathcal{D}^{\mathsf{main}}$ (resp. $\mathcal{D}^{\mathsf{aux}}$) that transition from state s at step h.

2) *Lower bounding* $\{N_h^{\mathsf{main}}(s)\}$ *using* $\mathcal{D}^{\mathsf{aux}}$. For each $s \in \mathcal{S}$ and $1 \le h \le H$, compute

$$N_h^{\mathsf{trim}}(s) \coloneqq \max\left\{N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{HS}{\delta}},\, 0\right\}; \tag{5.10}$$

3) *Random subsampling.* Let $\mathcal{D}^{\mathsf{main}'}$ be the set of all sample transitions (i.e., the quadruples taking the form (s,a,h,s')) from $\mathcal{D}^{\mathsf{main}}$. Subsample $\mathcal{D}^{\mathsf{main}'}$ to obtain $\mathcal{D}^{\mathsf{trim}}$, such that for each $(s,h) \in \mathcal{S}\times[H]$, $\mathcal{D}^{\mathsf{trim}}$ contains $\min\{N_h^{\mathsf{trim}}(s), N_h^{\mathsf{main}}(s)\}$ sample transitions randomly drawn from $\mathcal{D}^{\mathsf{main}'}$.

run VI-LCB: set $\mathcal{D}_0 = \mathcal{D}^{\mathsf{trim}}$; run Algorithm 9 to compute a policy $\widehat{\pi}$.

Before proceeding, we formally justify that $N_h^{\mathsf{trim}}(s)$ — as computed in (5.10) — is a valid lower bound on $N_h^{\mathsf{main}}(s)$. Here and below, we denote by $N_h^{\mathsf{trim}}(s,a)$ the number of sample transitions in $\mathcal{D}^{\mathsf{trim}}$ that are associated with the state-action pair (s,a) at step h.

Lemma 13. *Suppose that the K trajectories in $\mathcal{D}$ are generated in an i.i.d. fashion (see Chapter 5.1.1). With probability at least $1-8\delta$, the quantities constructed in (5.10) obey*

$$N_h^{\mathsf{trim}}(s) \le N_h^{\mathsf{main}}(s), \tag{5.11a}$$

$$N_h^{\mathsf{trim}}(s,a) \ge \frac{Kd_h^{\mathsf{b}}(s,a)}{8} - 5\sqrt{Kd_h^{\mathsf{b}}(s,a)\log\frac{KH}{\delta}} \tag{5.11b}$$

simultaneously for all $1 \le h \le H$ and all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

5.1.4 Theoretical guarantees

In what follows, we characterize the sample complexity of Algorithm 10, as formalized below.

Theorem 4. *Consider any $\varepsilon \in (0, H]$ and any $0 < \delta < 1$. With probability exceeding $1-12\delta$, the policy $\widehat{\pi}$ returned by Algorithm 10 obeys*

$$V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon \tag{5.12}$$

as long as the penalty terms are chosen according to the Bernstein-style quantity (5.9) for some

large enough numerical constant $c_{\mathrm{b}} > 0$, and the total number of sample trajectories exceeds

$$K \geq \frac{c_{\mathsf{k}} H^3 S C^\star_{\mathsf{clipped}} \log \frac{KH}{\delta}}{\varepsilon^2} \tag{5.13}$$

for some sufficiently large numerical constant $c_{\mathsf{k}} > 0$, where $C^\star_{\mathsf{clipped}}$ is introduced in Definition 2.

In general, the total sample size characterized by Theorem 4 could be far smaller than the ambient dimension (i.e., S^2AH) of the probability transition kernel P, thus precluding one from estimating P in a reliable fashion. As a crucial insight from Theorem 4, the model-based (or plug-in) approach enables reliable policy learning even when model estimation is completely off. Our analysis of Theorem 4 relies heavily on (i) suitable decoupling of complicated statistical dependency via subsampling, and (ii) careful control of the variance terms in the presence of Bernstein-style penalty.

In order to help assess the tightness and optimality of Theoerem 4, we further develop a minimax lower bound as follows.

Theorem 5. *For any $(H, S, C^\star_{\mathsf{clipped}}, \varepsilon)$ obeying $H \geq 12$, $C^\star_{\mathsf{clipped}} \geq 8/S$ and $\varepsilon \leq c_3 H$, one can construct a collection of MDPs $\{\mathcal{M}_\theta \mid \theta \in \Theta\}$, an initial state distribution ρ, and a batch dataset with K independent sample trajectories each of length H, such that*

$$\inf_{\widehat{\pi}} \max_{\theta \in \Theta} \mathbb{P}_\theta \left\{ V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \geq \varepsilon \right\} \geq \frac{1}{4}, \tag{5.14}$$

provided that the total sample size

$$N = KH \leq \frac{c_4 C^\star_{\mathsf{clipped}} S H^4}{\varepsilon^2}. \tag{5.15}$$

Here, $c_3, c_4 > 0$ are some small enough numerical constants, the infimum is over all estimator $\widehat{\pi}$, and $\mathbb{P}_\theta$ denotes the probability when the MDP is $\mathcal{M}_\theta$.

Implications. In what follows, let us take a moment to discuss several other key implications of Theorem 4.

- *Near-optimal sample complexities.* In the presence of the Bernstein-style penalty, the total number of samples (i.e., KH) needed for our algorithm to yield ε-accuracy is

$$\widetilde{O}\left(\frac{H^4 S C^\star_{\mathsf{clipped}}}{\varepsilon^2} \right). \tag{5.16}$$

  This confirms the optimality of the proposed model-based approach (up to some logarithmic term) when Bernstein-style penalty is employed, since Theorem 5 reveals that at least $\frac{H^4 S C^\star_{\mathsf{clipped}}}{\varepsilon^2}$ samples are needed regardless of the algorithm in use.

- *Full ε-range and no burn-in cost.* The sample complexity bound (5.13) stated in Theorem 4 holds for an arbitrary $\varepsilon \in (0, H]$. In other words, no burn-in cost is needed for the algorithm to work sample-optimally. This improves substantially upon the state-of-the-art results for model-based and model-free offline algorithms, both of which require a significant level of burn-in sample size ($H^9SC^\star$ and $H^6SC^\star$, respectively).

- *Sample reduction and model compressibility when $C^\star_{\mathsf{clipped}} < 1$.* Given that $C^\star_{\mathsf{clipped}}$ might drop below 1, the sample complexity of our algorithm might be as low as $\widetilde{O}\big(\frac{H^4S}{\varepsilon^2}\big)$. In fact, recognizing that $C^\star_{\mathsf{clipped}}$ can be as small as $\frac{1+o(1)}{S}$, we see that the sample complexity can sometimes be reduced to

$$\widetilde{O}\bigg(\frac{H^4}{\varepsilon^2}\bigg), \tag{5.17}$$

  resulting in significant sample size saving compared to prior works. Caution needs to be exercised, however, that this sample size improvement is made possible as a result of certain *model compressibility* implied by a small $C^\star_{\mathsf{clipped}}$. For instance, $C^\star_{\mathsf{clipped}} = O(1/S)$ might happen when a small number of states accounts for a dominant fraction of probability mass in $d^\star_h(s)$, with the remaining states exhibiting vanishingly small occupancy probability (see also the lower bound construction in the proof of Theorem 5); if this happens, then it often suffices to focus on learning those dominant states.

Infeasibility of estimating $C^\star_{\mathsf{clipped}}$. With the sample complexity (5.16) in mind, one natural question arises as to whether it is possible to estimate $C^\star_{\mathsf{clipped}}$ from the batch dataset. Unfortunately, this is in general infeasible, as demonstrated by the following example.

- (A hard example) Consider an MDP with horizon $H = 2$. In step $h = 1$, we have a singleton state space $\mathcal{S}_1 = \{0\}$ and an action space $\mathcal{A}_1 = \{0, 1\}$, whereas in step $h = 2$, we have a state space $\mathcal{S}_2 = \{0, 1\}$ and a singleton action space $\mathcal{A}_2 = \{0\}$. The reward function and the transition kernel are given by:

$$\begin{gathered} r_1(0,0) = 0, \quad r_1(0,1) = 0, \quad r_2(0,0) = 0, \quad r_2(1,0) = 1 \\ P_1(0\,|\,0,0) = 0.5, \quad P_1(1\,|\,0,0) = 0.5, \quad P_1(0\,|\,0,1) = p, \quad P_1(1\,|\,0,1) = 1-p \end{gathered}$$

  for some unknown parameter $p \in (0, 1)$. We have K independent trajectories as usual, and let

$$d^{\mathsf{b}}_1(0,0) = 1 - \frac{1}{K} \qquad \text{and} \qquad d^{\mathsf{b}}_1(0,1) = \frac{1}{K}. \tag{5.18}$$

  Elementary calculation then reveals that: $C^\star_{\mathsf{clipped}} = K$ when $p < \frac{1}{2}$, and $C^\star_{\mathsf{clipped}} = 1 + \frac{1}{K-1}$ when $p > \frac{1}{2}$. Such a remarkable difference in $C^\star_{\mathsf{clipped}}$ depends on the value of p, which is only reflected in $(s, a) = (0, 1)$ at step 1. However, by construction, there is nonvanishing probability

(i.e., $\left(1-d_1^{\mathsf{b}}(0,1)\right)^K \approx 1/e$ for large K) such that the dataset does not visit $(s,a)=(0,1)$ in step $h=1$ at all, which in turn precludes one from distinguishing $C^\star_{\mathsf{clipped}} = 1+\frac{1}{K-1}$ from $C^\star_{\mathsf{clipped}} = K$ given only the available dataset.

Fortunately, implementing our algorithm does not require prior knowledge of $C^\star_{\mathsf{clipped}}$ at all, and the algorithm succeeds once the task becomes feasible. On the other hand, we won't be able to tell how large a sample size is enough *a priori*, but this is in general information-theoretically infeasible as illustrated by the above example.

Comparisons with prior statistical analysis. We now briefly discuss the novelty of our statistical analysis compared with past theory. Perhaps the most related prior work is Xie et al. (2021b), which proposed two algorithms. The first algorithm therein is VI-LCB with H-fold sample splitting and Hoeffding-style penalty, and each of these two features adds an H factor to the total sample complexity. The second algorithm therein combines VI-LCB with variance reduction, which leads to optimal sample complexity for sufficiently small ε (i.e., a large burn-in cost is required). Note, however, that none of the existing statistical tools for variance reduction is able to work without imposing a large burn-in cost, regardless of the sampling mechanism in use (e.g., generative model, offline RL, online RL) (Li et al., 2023b; Sidford et al., 2018a; Xie et al., 2021b; Zhang et al., 2020c). In contrast, our theory makes apparent that variance reduction is unnecessary, which leads to both simpler algorithm and tighter analysis. Additionally, while Bernstein-style confidence bounds have been deployed in online RL for finite-horizon MDPs (Azar et al., 2017; Fruit et al., 2020; Jin et al., 2018; Zhang et al., 2020c), none of these works was able to yield optimal sample complexity without a large burn-in cost (e.g., Azar et al. (2017) incurred a burn-in cost as large as S^3AH^6). This in turn underscores the power of our statistical analysis when coping with the most data-hungry regime.

5.2 Algorithm and theory: discounted infinite-horizon MDPs

Now, we turn attention to the studies of offline RL for discounted infinite-horizon MDPs.

5.2.1 Problem formulation and assumptions

As before, recalling the definition of discounted infinite-horizon MDPs in Chapter 2.1, we shall further introduce additional notations, the sampling model and the goal.

Similar to finite-horizon case, we introduce the *discounted occupancy distributions* associated with policy π as follows:

$$\forall s\in\mathcal{S}: \quad d^\pi(s;\rho) \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s \mid s_0\sim\rho;\pi), \tag{5.19}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad d^{\pi}(s,a;\rho) \coloneqq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s, a_t = a \mid s_0 \sim \rho; \pi), \tag{5.20}$$

where we consider the randomness over a sample trajectory that starts from an initial state $s_0 \sim \rho$ and that follows policy π (i.e., $a_t \sim \pi(\cdot \mid s_t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ for all $t \geq 0$).

Correspondingly, we adopt the notation of the discounted occupancy distributions associated with the optimal policy $\pi^\star$ as:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad d^\star(s) \coloneqq d^{\pi^\star}(s;\rho) \quad \text{and} \quad d^\star(s,a) \coloneqq d^{\pi^\star}(s,a;\rho) = d^\star(s)\, \mathbb{1}\left(a = \pi^\star(s)\right), \tag{5.21}$$

where the last equality is valid since $\pi^\star$ is assumed to be deterministic.

Offline/batch data. Let us work with an independent sampling model as studied in the prior work Rashidinejad et al. (2021). To be precise, imagine that we observe a batch dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{1 \leq i \leq N}$ containing N sample transitions. These samples are independently generated based on a distribution $d^{\mathsf{b}} \in \Delta(\mathcal{S} \times \mathcal{A})$ and the transition kernel P of the MDP, namely,

$$(s_i, a_i) \overset{\text{ind.}}{\sim} d^{\mathsf{b}} \qquad \text{and} \qquad s_i' \overset{\text{ind.}}{\sim} P(\cdot \mid s_i, a_i), \qquad 1 \leq i \leq N. \tag{5.22}$$

In addition, it is assumed that the learner is aware of the reward function.

In order to capture the distribution shift between the desired occupancy measure and the data distribution, we introduce a key quantity previously introduced in Rashidinejad et al. (2021).

Definition 3 (Single-policy concentrability for infinite-horizon MDPs)**.** The single-policy concentrability coefficient of a batch dataset $\mathcal{D}$ is defined as

$$C^\star \coloneqq \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d^\star(s,a)}{d^{\mathsf{b}}(s,a)}. \tag{5.23}$$

Clearly, one necessarily has $C^\star \geq 1$.

In words, $C^\star$ measures the distribution mismatch in terms of the maximum density ratio. The batch dataset can be viewed as expert data when $C^\star$ approaches 1, meaning that the batch dataset is close to the target policy in terms of the induced distributions. Moreover, this coefficient $C^\star$ is referred to as the "single-policy" concentrability coefficient since it is concerned with a single policy $\pi^\star$; this is clearly a much weaker assumption compared to the all-policy concentrability assumption (as adopted in, e.g., Chen and Jiang (2019); Fan et al. (2020); Farahmand et al. (2010); Munos (2007); Ren et al. (2021); Xie and Jiang (2021)), the latter of which assumes a uniform density-ratio bound over all policies and requires the dataset to be highly exploratory.

In the current sub-chapter, we also introduce a slightly improved version of $C^\star$ as follows.

Definition 4 (Single-policy clipped concentrability for infinite-horizon MDPs)**.** The single-policy clipped concentrability coefficient of a batch dataset $\mathcal{D}$ is defined as

$$C^\star_{\mathsf{clipped}} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\min\left\{d^\star(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)}. \tag{5.24}$$

Remark 2. A direct comparison of Conditions (5.23) and (7.25) implies that for a given batch dataset $\mathcal{D}$,

$$C^\star_{\mathsf{clipped}} \le C^\star. \tag{5.25}$$

As we shall see later, while our sample complexity upper bounds will be mainly stated in terms of $C^\star_{\mathsf{clipped}}$, all of them remain valid if $C^\star_{\mathsf{clipped}}$ is replaced with $C^\star$. Additionally, in contrast to $C^\star$ that is always lower bounded by 1, we have a smaller lower bound as follows (directly from the definition (7.25))

$$C^\star_{\mathsf{clipped}} \ge 1/S, \tag{5.26}$$

which is nearly tight.[1] This attribute could lead to sample size saving in some cases, to be detailed shortly.

Let us take a moment to further interpret the coefficient in Definition 4, which says that

$$d^{\mathsf{b}}(s,a) \ge \begin{cases} \frac{1}{C^\star_{\mathsf{clipped}}} d^\star(s,a), & \text{if } d^\star(s,a) \le 1/S \\ \frac{1}{C^\star_{\mathsf{clipped}} S}, & \text{if } d^\star(s,a) > 1/S \end{cases} \tag{5.27}$$

holds for any pair (s,a). Consider, for instance, the case where $C^\star_{\mathsf{clipped}} = O(1)$: if a state-action pair is infrequently (or rarely) visited by the optimal policy, then it is fine for the associated density in the batch data to be very small (e.g., a density proportional to that of the optimal policy); by contrast, if a state-action pair is visited fairly often by the optimal policy, then Definition 4 might only require $d^{\mathsf{b}}(s,a)$ to exceed the order of $1/S$. In other words, the required level of $d^{\mathsf{b}}(s,a)$ is clipped at the level $\frac{1}{C^\star_{\mathsf{clipped}} S}$ regardless of the value of $d^\star(s,a)$.

Goal. Armed with the batch dataset $\mathcal{D}$, the objective of offline RL in this case is to find a policy $\widehat{\pi}$ that attains near-optimal value functions — with respect to a given test state distribution $\rho \in \Delta(\mathcal{S})$

[1] As a concrete example, suppose that $d^\star(s) = \begin{cases} 1 - \frac{S-1}{S^3} & \text{if } s = 1 \\ \frac{1}{S^3} & \text{else} \end{cases}$ and $d^{\mathsf{b}}(s,a) = \begin{cases} 1 - \frac{S-1}{S^2} & \text{if } a = \pi^\star(s) \text{ and } s = 1, \\ \frac{1}{S^2} & \text{if } a = \pi^\star(s) \text{ and } s \neq 1, \\ 0, & \text{else.} \end{cases}$ Then it can be easily verified that $C^\star_{\mathsf{clipped}} = \frac{1}{S-1+\frac{1}{S}}$. Nonetheless, caution should be exercised that an exceedingly small $C^\star_{\mathsf{clipped}}$ requires highly compressible structure of $d^\star$, and the real-world data often do not fall within this benign range of $C^\star_{\mathsf{clipped}}$.

— in a sample-efficient manner. To be precise, for a prescribed accuracy level ε, we seek to identify an ε-optimal policy $\widehat{\pi}$ satisfying

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon \tag{5.28}$$

with high probability, using a batch dataset $\mathcal{D}$ (cf. (5.22)) containing as few samples as possible. Particular emphasis is placed on achieving minimal sample complexity for the entire range of accuracy levels (namely, for any $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$).

Remark 3. The careful reader might remark that i.i.d. sampling as in (5.22) might be too stringent. While our main theory is developed based on this idealistic sampling model, we shall present extensions to Markovian data as well (see Appendix C.4).

5.2.2 VI-LCB for infinite-horizon MDPs

In this subchapter, we introduce a model-based offline RL algorithm that incorporates lower concentration bounds in value estimation. The algorithm, called VI-LCB, applies value iteration (based on some pessimistic Bellman operator) to the empirical MDP, with the key ingredients described below.

The empirical MDP. Recall that we are given N independent sample transitions $\{(s_i, a_i, s_i')\}_{i=1}^N$ in the dataset $\mathcal{D}$. For any given state-action pair (s, a), we denote by

$$N(s,a) \coloneqq \sum_{i=1}^{N} \mathbb{1}\big((s_i, a_i) = (s, a)\big) \tag{5.29}$$

the number of samples transitions from (s, a). We then construct an empirical transition matrix $\widehat{P}$ such that

$$\widehat{P}(s' \,|\, s, a) = \begin{cases} \frac{1}{N(s,a)} \sum_{i=1}^{N} \mathbb{1}\big\{(s_i, a_i, s_i') = (s, a, s')\big\}, & \text{if } N(s,a) > 0 \\ \frac{1}{S}, & \text{else} \end{cases} \tag{5.30}$$

for each $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

The pessimistic Bellman operator. Our offline algorithm is developed based on finding the fixed point of some variant of the classical Bellman operator. Let us first introduce this key operator and eludicate how the pessimism principle is enforced. Recall that the Bellman operator $\mathcal{T}(\cdot) : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$ w.r.t. the transition kernel P is defined such that for any vector $Q \in \mathbb{R}^{SA}$,

$$\mathcal{T}(Q)(s,a) \coloneqq r(s,a) + \gamma P_{s,a} V \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{5.31}$$

where $V = [V(s)]_{s\in\mathcal{S}}$ with $V(s) \coloneqq \max_a Q(s,a)$. We propose to penalize the original Bellman operator w.r.t. the empirical kernel $\widehat{P}$ as follows:

$$\widehat{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a) \coloneqq \max\Big\{r(s,a) + \gamma \widehat{P}_{s,a}V - b(s,a;V), 0\Big\} \qquad \text{for all } (s,a)\in\mathcal{S}\times\mathcal{A}, \tag{5.32}$$

where $b(s,a;V)$ denotes the penalty term employed to enforce pessimism amid uncertainty. As one can anticipate, the properties of the fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$ relies heavily upon the choice of the penalty terms $\{b_h(s,a;V)\}$, often derived based on certain concentration bounds. In this sub-chapter, we focus on the following Bernstein-style penalty to exploit the importance of certain variance statistics:

$$b(s,a;V) \coloneqq \min\left\{\max\left\{\sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(V)}, \frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\}, \frac{1}{1-\gamma}\right\} + \frac{5}{N} \tag{5.33}$$

for every $(s,a)\in\mathcal{S}\times\mathcal{A}$, where $c_{\mathrm{b}} > 0$ is some numerical constant (e.g., $c_{\mathrm{b}} = 144$), and $\delta\in(0,1)$ is some given quantity (in fact, $1-\delta$ is the target success probability). Here, for any vector $V\in\mathbb{R}^S$, we recall that $\mathsf{Var}_{\widehat{P}_{s,a}}(V)$ is the variance of V w.r.t. the distribution $\widehat{P}_{s,a}$ (see (1.7)).

We immediately isolate several useful properties as follows.

Lemma 14. *For any $\gamma\in[\frac{1}{2},1)$, the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$ (cf. (5.32)) with the Bernstein-style penalty (5.33) is a γ-contraction w.r.t. $\|\cdot\|_\infty$, that is,*

$$\big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}(Q_2)\big\|_\infty \le \gamma\|Q_1 - Q_2\|_\infty \tag{5.34}$$

for any $Q_1, Q_2\in\mathbb{R}^{S\times A}$ obeying $Q_1(s,a), Q_2(s,a)\in\big[0,\frac{1}{1-\gamma}\big]$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. In addition, there exists a unique fixed point $\widehat{Q}^\star_{\mathsf{pe}}$ of the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$, which also obeys $0\le\widehat{Q}^\star_{\mathsf{pe}}(s,a)\le\frac{1}{1-\gamma}$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

In words, even though $\widehat{\mathcal{T}}_{\mathsf{pe}}(\cdot)$ integrates the penalty terms, it still preserves the γ-contraction property and admits a unique fixed point, thereby resembling the classical Bellman operator (5.31).

The VI-LCB algorithm. We are now positioned to introduce the VI-LCB algorithm, which can be regarded as classical value iteration applied in conjunction with pessimism. Specifically, the algorithm applies the Bernstein-style pessimistic operator $\widehat{\mathcal{T}}_{\mathsf{pe}}$ (cf. (5.32)) iteratively in order to find its fixed point:

$$\widehat{Q}_\tau(s,a) = \widehat{\mathcal{T}}_{\mathsf{pe}}\big(\widehat{Q}_{\tau-1}\big)(s,a) = \max\Big\{r(s,a) + \gamma\widehat{P}_{s,a}\widehat{V}_{\tau-1} - b\big(s,a;\widehat{V}_{\tau-1}\big), 0\Big\}, \qquad \tau = 1,2,\cdots \tag{5.35}$$

We shall initialize it to $\widehat{Q}_0 = 0$, implement (5.35) for $\tau_{\max}$ iterations, and output $\widehat{Q} = \widehat{Q}_{\tau_{\max}}$ as the final Q-estimate. The final policy estimate $\widehat{\pi}$ is chosen on the basis of $\widehat{Q}$ as follows:

$$\widehat{\pi}(s)\in\arg\max_a \widehat{Q}(s,a) \qquad \text{for all } s\in\mathcal{S}, \tag{5.36}$$

Algorithm 11: Offline value iteration with LCB (VI-LCB) for discounted infinite-horizon MDPs

input: dataset $\mathcal{D}$; reward function r; target success probability $1-\delta$; max iteration number $\tau_{\max}$.
initialization: $\widehat{Q}_0 = 0$, $\widehat{V}_0 = 0$.
construct the empirical transition kernel $\widehat{P}$ according to (5.30).
for $\tau = 1, 2, \cdots, \tau_{\max}$ **do**
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    compute the penalty term $b\big(s, a; \widehat{V}_{\tau-1}\big)$ according to (5.33).
    set $\widehat{Q}_\tau(s,a) = \max\big\{r(s,a) + \gamma \widehat{P}_{s,a}\widehat{V}_{\tau-1} - b\big(s,a;\widehat{V}_{\tau-1}\big), 0\big\}$.
  for $s \in \mathcal{S}$ **do**
    set $\widehat{V}_\tau(s) = \max_a \widehat{Q}_\tau(s,a)$.
output: $\widehat{\pi}$ s.t. $\widehat{\pi}(s) \in \arg\max_a \widehat{Q}_{\tau_{\max}}(s,a)$ for any $s \in \mathcal{S}$.

with the whole algorithm summarized in Algorithm 11.

Let us pause to explain the rationale of the pessimism principle on a high level. If a pair (s,a) has been insufficiently visited in $\mathcal{D}$ (i.e., $N(s,a)$ is small), then the resulting Q-estimate $\widehat{Q}_\tau(s,a)$ could suffer from high uncertainty and become unreliable, which might in turn mislead value estimation. By enforcing suitable penalization $b(s,a;\widehat{V}_{\tau-1})$ based on certain lower confidence bounds, we can suppress the negative influence of such poorly visited state-action pairs. Fortunately, suppressing these state-action pairs might not result in significant bias in value estimation when $C^\star_{\mathsf{clipped}}$ is small; for instance, when the behavior policy π^{b} resembles $\pi^\star$, the poorly visited state-action pairs correspond primarily to suboptimal actions (as they are not selected by $\pi^\star$), making it acceptable to neglect these pairs.

Interestingly, Algorithm 11 is guaranteed to converge rapidly. In view of the γ-contraction property in Lemma 14, the iterates $\{\widehat{Q}_\tau\}_{\tau\geq 0}$ converge linearly to the fixed point $\widehat{Q}^\star_{\mathsf{pe}}$, as asserted below.

Lemma 15. *Suppose $\widehat{Q}_0 = 0$. Then the iterates of Algorithm 11 obey*

$$\widehat{Q}_\tau \leq \widehat{Q}^\star_{\mathsf{pe}} \qquad \text{and} \qquad \big\|\widehat{Q}_\tau - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \leq \frac{\gamma^\tau}{1-\gamma} \qquad \text{for all } \tau \geq 0, \tag{5.37}$$

where $\widehat{Q}^\star_{\mathsf{pe}}$ is the unique fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}$. As a consequence, by choosing $\tau_{\max} \geq \frac{\log \frac{N}{1-\gamma}}{\log(1/\gamma)}$ one fulfills

$$\big\|\widehat{Q}_{\tau_{\max}} - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \leq 1/N. \tag{5.38}$$

Algorithmic comparison with Rashidinejad et al. (2021). VI-LCB has been studied in the prior work Rashidinejad et al. (2021). The difference between our algorithm and the version therein

is two-fold:

- *Sample reuse vs. $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$-fold sample splitting.* Our algorithm reuses the same set of samples across all iterations, which is in sharp contrast to Rashidinejad et al. (2021) that employs fresh samples in each of the $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$ iterations. This results in considerably better usage of available information.

- *Bernstein-style vs. Hoeffding-style penalty.* Our algorithm adopts the Bernstein-type penalty, as opposed to the Hoeffding-style penalty in Rashidinejad et al. (2021). This choice leads to more effective exploitation of the variance structure across time.

Pessimism vs. optimism in the face of uncertainty. The careful reader might also notice the similarity between the pessimism principle and the optimism principle utilized in online RL. A well-developed paradigm that balances exploration and exploitation in online RL is optimistic exploration based on uncertainty quantification (Lai and Robbins, 1985). The earlier work Jaksch et al. (2010) put forward an algorithm called UCRL2 that computes an optimistic policy with the aid of Hoeffding-style confidence regions for the probability transition kernel. Later on, Azar et al. (2017) proposed to build upper confidence bounds (UCB) for the optimal values instead, which leads to significantly improved sample complexity; see, e.g., He et al. (2021); Wang et al. (2019) for the application of this strategy to discounted infinite-horizon MDPs. Note, however, that the rationales behind optimism and pessimism are remarkably different. In offline RL (which does not allow further data collection), the uncertainty estimates are employed to identify, and then rule out, poorly-visited actions; this stands in sharp contrast to the online counterpart where poorly-visited actions might be more favored during exploration.

5.2.3 Theoretical guarantees

When the Bernstein-style concentration bound (5.33) is adopted, the VI-LCB algorithm in Algorithm 11 yields ε-accuracy with a near-minimal number of samples, as stated below.

Theorem 6. *Suppose $\gamma \in [\frac{1}{2}, 1)$, and consider any $0 < \delta < 1$ and $\varepsilon \in \big(0, \frac{1}{1-\gamma}\big]$. Suppose that the total number of iterations exceeds $\tau_{\max} \geq \frac{1}{1-\gamma} \log \frac{N}{1-\gamma}$. With probability at least $1 - 2\delta$, the policy $\widehat{\pi}$ returned by Algorithm 11 obeys*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon, \tag{5.39}$$

provided that c_{b} (cf. the Bernstein-style penalty term in (5.33)) is some sufficiently large numerical constant and the total sample size exceeds

$$N \geq \frac{c_1 S C^{\star}_{\mathsf{clipped}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3 \varepsilon^2} \tag{5.40}$$

for some large enough numerical constant $c_1 > 0$, where $C^\star_{\mathsf{clipped}}$ is introduced in Definition 4.

In general, the total sample size characterized by Theorem 6 could be far smaller than the ambient dimension (i.e., S^2A) of the transition kernel P, thus precluding one from estimating P in a reliable fashion. As a crucial insight from Theorem 6, the model-based (or plug-in) approach enables reliable offline learning even when model estimation is completely off.

Before discussing key implications of Theorem 6, we develop matching minimax lower bounds that help confirm the efficacy of the proposed model-based algorithm.

Theorem 7. *For any $(\gamma, S, C^\star_{\mathsf{clipped}}, \varepsilon)$ obeying $\gamma \in \left[\frac{2}{3}, 1\right)$, $S \geq 2$, $C^\star_{\mathsf{clipped}} \geq \frac{8\gamma}{S}$, and $\varepsilon \leq \frac{1}{42(1-\gamma)}$, one can construct two MDPs $\mathcal{M}_0, \mathcal{M}_1$, an initial state distribution ρ, and a batch dataset with N independent samples and single-policy clipped concentrability coefficient $C^\star_{\mathsf{clipped}}$ such that*

$$\inf_{\widehat{\pi}} \max\left\{\mathbb{P}_0\big(V^\star(\rho) - V^{\widehat{\pi}}(\rho) > \varepsilon\big),\, \mathbb{P}_1\big(V^\star(\rho) - V^{\widehat{\pi}}(\rho) > \varepsilon\big)\right\} \geq \frac{1}{8},$$

provided that

$$N \leq \frac{c_2 S C^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}$$

for some numerical constant $c_2 > 0$. Here, the infimum is over all estimator $\widehat{\pi}$, and $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) denotes the probability when the MDP is $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

Implications. In the following, we take a moment to interpret the above two theorems and single out several key implications about the proposed model-based algorithm.

- *Optimal sample complexities.* In the presence of the Bernstein-style penalty, the total number of samples needed for our algorithm to yield ε-accuracy is

$$\widetilde{O}\left(\frac{S C^\star_{\mathsf{clipped}}}{(1-\gamma)^3\varepsilon^2}\right). \tag{5.41}$$

  This taken together with the minimax lower bound asserted in Theorem 7 confirms the optimality of the proposed model-based approach (up to some logarithmic factor). In comparison, the sample complexity derived in Rashidinejad et al. (2021) exhibits a worse dependency on the effective horizon (i.e., $\frac{1}{(1-\gamma)^5}$). Theorem 7 also enhances the lower bound developed in Rashidinejad et al. (2021) to accommodate the scenario where $C^\star_{\mathsf{clipped}}$ can be much smaller than $C^\star$, i.e., $C^\star_{\mathsf{clipped}} = O(1/S)$.

- *No burn-in cost.* The fact that the sample size bound (5.40) holds for the full ε-range (i.e., any given $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$) means that there is no burn-in cost required to achieve sample optimality. This not only drastically improves upon, but in fact eliminates, the burn-in cost of the best-known sample-optimal result (cf. Table 1.2), the latter of which required a burn-in cost

at least on the order of $\frac{SC^\star}{(1-\gamma)^5}$. Accomplishing this requires one to tackle the sample-hungry regime, which is statistically challenging to cope with.

- *No need of sample splitting.* It is noteworthy that prior works typically required sample splitting. For instance, Rashidinejad et al. (2021) analyzed the VI-LCB algorithm with fresh samples employed in each iteration, which effectively split the data into $\widetilde{O}\big(\frac{1}{1-\gamma}\big)$ disjoint subsets. In contrast, the algorithm studied herein permits the reuse of all samples across all iterations. This is an important feature in sample-starved applications to effectively maximize information utilization, and is a crucial factor that assists in improving the sample complexity compared to Rashidinejad et al. (2021).

- *Sample size saving when $C^\star_{\mathsf{clipped}} < 1$.* In view of Theorem 6, the sample complexity of the proposed algorithm can be as low as

$$\widetilde{O}\bigg(\frac{1}{(1-\gamma)^3\varepsilon^2}\bigg)$$

when $C^\star_{\mathsf{clipped}}$ is on the order of $1/S$. This might seem somewhat surprising at first glance, given that the minimax sample complexity for policy evaluation is at least $\widetilde{O}\big(\frac{S}{(1-\gamma)^3\varepsilon^2}\big)$ even in the presence of a simulator (Azar et al., 2013). To elucidate this, we note that the condition $C^\star_{\mathsf{clipped}} = O(1/S)$ implicitly imposes special — in fact, highly compressible — structure on the MDP that enables sample size reduction. As we shall see from the lower bound construction in Theorem 7, the case with $C^\star_{\mathsf{clipped}} = O(1/S)$ might require $d^\star(s,a)$ to concentrate on one or a small number of important states, with exceedingly small probability assigned to the remaining ones. If this occurs, then it often suffices to focus on what happens on these important states, thus requiring much fewer samples.

Comparisons with prior statistical analysis. Before concluding this subchapter, we highlight the innovations of our statistical analysis compared to past theory when it comes to discounted infinite-horizon MDPs. To begin with, our sample size improvement over Rashidinejad et al. (2021) stems from the two algorithmic differences mentioned in Chapter 5.2.2: the sample-reuse feature allows one to improve a factor of $\frac{1}{1-\gamma}$, while the use of Bernstein-style penalty yields an additional gain of $\frac{1}{1-\gamma}$. In addition, while the design of data-driven Bernstein-style bounds has been extensively studied in online RL in discounted MDPs (e.g., He et al. (2021); Zhang et al. (2021c)), all of these past results were either sample-suboptimal, or required a huge burn-in sample size (e.g., $\frac{S^3A^2}{(1-\gamma)^4}$ in He et al. (2021)). In other words, sample optimality was not previously achieved in the most data-hungry regime. In comparison, our theory ensures optimality of our algorithm even for the most sample-constrained scenario, which relies on much more delicate statistical tools. In a nutshell, our statistical analysis is built upon at least two ideas: (i) a leave-one-out analysis framework that allows to decouple complicated statistical dependency across iterations without losing statistical

tightness; (ii) a delicate self-bounding trick that allows us to simultaneously control multiple crucial statistical quantities (e.g., empirical variance) in the most sample-starved regime.

5.3 Analysis: episodic finite-horizon MDPs

5.3.1 Preliminary facts and notation

We first collect a few preliminary facts that are useful for the analysis. The first fact determines the range of our estimates $\widehat{Q}_h$ and $\widehat{V}_h$.

Lemma 16. *The iterates of Algorithm 9 obey*

$$0 \le \widehat{Q}_h(s,a) \le H-h+1 \quad \text{and} \quad 0 \le \widehat{V}_h(s) \le H-h+1 \qquad \text{for all } (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]. \tag{5.42}$$

Proof. The non-negativity of $\widehat{Q}_h$ (and hence $\widehat{V}_h$) follows directly from the update rule (5.6). Regarding the upper bound, we suppose for the moment that $\widehat{V}_{h+1}(s) \le H-h$ for step $h+1$. Then (5.6) tells us that

$$\widehat{Q}_h(s,a) \le 1 + \big\|\widehat{V}_{h+1}\big\|_\infty \le 1 + H - h,$$

which together with $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$ justifies the claim (5.42) for step h as well. Taking this together with the base case $\widehat{V}_{H+1} = 0$ and the standard induction argument concludes the proof. □

The second fact is concerned with the vector $d_h^\star := [d_h^\star(s)]_{s\in\mathcal{S}} \in \mathbb{R}^S$. For any $h \in [H]$, denote by $P_h^\star \in \mathbb{R}^{S\times S}$ a matrix whose s-th row is given by $P_h\big(\cdot \,|\, s, \pi_h^\star(s)\big)$. Then the Markovian property of the MDP indicates that: for any $j > h$, one has

$$\big(d_j^\star\big)^\top = \big(d_h^\star\big)^\top P_h^\star \cdots P_{j-1}^\star. \tag{5.43}$$

Notation. We remind the reader that $P_{h,s,a} \in \mathbb{R}^{1\times S}$ represents the probability transition vector $P_h(\cdot \,|\, s,a)$, and the associated variance parameter $\mathsf{Var}_{P_{h,s,a}}(V)$ is defined to be the (h,s,a)-th row of $\mathsf{Var}_P(V)$ (cf. (1.7)), namely,

$$\mathsf{Var}_{P_{h,s,a}}(V) := \sum_{s'\in S} P_h(s' \,|\, s,a)\big(V(s')\big)^2 - \bigg(\sum_{s'\in S} P_h(s' \,|\, s,a)V(s')\bigg)^2 \tag{5.44}$$

for any given vector $V \in \mathbb{R}^S$. The vector $\widehat{P}_{h,s,a} \in \mathbb{R}^{1\times S}$ and the variance parameter $\mathsf{Var}_{\widehat{P}_{h,s,a}}(V)$ are defined analogously.

5.3.2 A crucial statistical independence property

This subchapter demonstrates that the subsampling trick introduced in Chapter 5.1.3 leads to some crucial statistical independence property. To be precise, let us consider the following two data-generating mechanisms; here and below, a sample transition refers to a quadruple (s, a, h, s') that indicates a transition from state s to state s' when action a is taken at step h.

- **Model 1 (augmented dataset).** Augment $\mathcal{D}^{\mathsf{trim}}$ to yield a dataset $\mathcal{D}^{\mathsf{trim,aug}}$ via the following steps. For every $(s,h) \in \mathcal{S} \times [H]$:

  1) Add to $\mathcal{D}^{\mathsf{trim,aug}}$ all $N_h^{\mathsf{trim}}(s)$ sample transitions in $\mathcal{D}^{\mathsf{trim}}$ that transition from state s at step h;
  2) If $N_h^{\mathsf{trim}}(s) > N_h^{\mathsf{main}}(s)$, then we further add to $\mathcal{D}^{\mathsf{trim,aug}}$ another set of $N_h^{\mathsf{trim}}(s) - N_h^{\mathsf{main}}(s)$ independent sample transitions $\big\{\big(s, a_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big)\big\}$ obeying

$$a_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \pi_h^{\mathsf{b}}(\cdot \,|\, s), \qquad s_{h,s}^{\prime(i)} \overset{\text{i.i.d.}}{\sim} P_h\big(\cdot \,|\, s, a_{h,s}^{(i)}\big), \qquad N_h^{\mathsf{main}}(s) < i \le N_h^{\mathsf{trim}}(s). \tag{5.45}$$

- **Model 2 (independent dataset).** For every $(s,h) \in \mathcal{S} \times [H]$, generate $N_h^{\mathsf{trim}}(s)$ independent sample transitions $\big\{\big(s, a_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big)\big\}$ as follows:

$$a_{h,s}^{(i)} \overset{\text{i.i.d.}}{\sim} \pi_h^{\mathsf{b}}(\cdot \,|\, s), \qquad s_{h,s}^{\prime(i)} \overset{\text{i.i.d.}}{\sim} P_h(\cdot \,|\, s, a), \qquad 1 \le i \le N_h^{\mathsf{trim}}(s). \tag{5.46}$$

  This forms the following dataset:

$$\mathcal{D}^{\text{i.i.d.}} := \Big\{\big(s, a_{h,s}^{(i)}, h, s_{h,s}^{\prime(i)}\big) \mid s \in \mathcal{S}, 1 \le h \le H, 1 \le i \le N_h^{\mathsf{trim}}(s)\Big\}. \tag{5.47}$$

In words, the dataset $\mathcal{D}^{\mathsf{trim,aug}}$ generated in Model 1 differs from $\mathcal{D}^{\mathsf{trim}}$ only if $N_h^{\mathsf{trim}}(s) > N_h^{\mathsf{main}}(s)$ occurs; this data generating mechanism ensures that $\mathcal{D}^{\mathsf{trim,aug}}$ comprises exactly $N_h^{\mathsf{trim}}(s)$ sample transitions from state s at step h. Two key features are: (a) the samples in $\mathcal{D}^{\mathsf{trim,aug}}$ are statistically independent, and (b) $\mathcal{D}^{\mathsf{trim,aug}}$ is essentially equivalent to $\mathcal{D}^{\mathsf{trim}}$ with high probability, as asserted below.

Lemma 17. *The above two datasets $\mathcal{D}^{\mathsf{trim,aug}}$ and $\mathcal{D}^{\text{i.i.d.}}$ have the same distributions. In addition, with probability exceeding $1 - 8\delta$, $\mathcal{D}^{\mathsf{trim,aug}} = \mathcal{D}^{\mathsf{trim}}$.*

Proof. Both $\mathcal{D}^{\mathsf{trim,aug}}$ and $\mathcal{D}^{\text{i.i.d.}}$ contain exactly $N_h^{\mathsf{trim}}(s)$ sample transitions from state s at step h. where $\{N_h^{\mathsf{trim}}(s)\}$ are statistically independent from the randomness of the samples. It is easily seen that: given $\{N_h^{\mathsf{trim}}(s)\}$, the sample transitions in $\mathcal{D}^{\mathsf{trim,aug}}$ across different steps are statistically independent. As a result, $\mathcal{D}^{\mathsf{trim}}$ and $\mathcal{D}^{\text{i.i.d.}}$ both consist of independent samples and are of the same distribution.

Furthermore, Lemma 13 tells us that with probability at least $1-8\delta$, $N_h^{\mathsf{trim}}(s) \le N_h^{\mathsf{main}}(s)$ holds for all $(s,h) \in \mathcal{S} \times [H]$, implying that that all data in $\mathcal{D}^{\mathsf{trim,aug}}$ come from $\mathcal{D}^{\mathsf{main}}$ and hence $\mathcal{D}^{\mathsf{trim,aug}} = \mathcal{D}^{\mathsf{trim}}$. □

5.3.3 Proof of Theorem 4

We first demonstrate that Theorem 4 is valid as long as the following theorem can be established.

Theorem 8. *Consider the dataset $\mathcal{D}_0$ described in Chapter 5.1.2, and any $0 < \delta < 1$. Suppose that $\mathcal{D}_0$ contains N sample transitions, and that the non-negative integers $\{N_h(s,a)\}$ defined in (5.4) obey*

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]: \qquad N_h(s,a) \ge \frac{K d_h^{\mathsf{b}}(s,a)}{8} - 5\sqrt{K d_h^{\mathsf{b}}(s,a) \log \frac{NH}{\delta}}, \tag{5.48}$$

with K some quantity obeying $K \ge 3872 H S C^\star_{\mathsf{rob}} \log \frac{NH}{\delta}$. Assume that conditional on $\{N_h(s,a)\}$, the sample transitions $\{(s,a,h,s'_{(i)}) \,|\, 1 \le i \le N_h(s,a), (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\}$ are statistically independent. The penalty terms are taken to be (5.9), where $c_{\mathrm{b}} \ge 16$ is chosen to be some constant. Then with probability at least $1-4\delta$, one has

$$\sum_s d_h^\star(s)\big(V_h^\star(s) - V_h^{\widehat{\pi}}(s)\big) \le 80\sqrt{\frac{2c_{\mathrm{b}} H^3 S C^\star_{\mathsf{rob}} \log \frac{NH}{\delta}}{K}}, \qquad 1 \le h \le H. \tag{5.49}$$

By construction, $\{N_h^{\mathsf{trim}}(s,a)\}$ are computed using $\mathcal{D}^{\mathsf{aux}}$, and hence are independent from the empirical model $\widehat{P}_h$ generated based on $\mathcal{D}^{\mathsf{trim}}$. Additionally, Lemma 17 permits us to treat the samples in $\mathcal{D}^{\mathsf{trim}}$ as being statistically independent. Recalling that the lower bound (5.11b) holds with probability at least $1-8\delta$, we can readily invoke Theorem 8 by taking $N_h(s,a) = N_h^{\mathsf{trim}}(s,a)$ and the property (4.2) to show that

$$\sum_{s\in\mathcal{S}} \rho(s)\big(V_1^\star(s) - V_1^{\widehat{\pi}}(s)\big) = \sum_{s\in\mathcal{S}} d_1^\star(s)\big(V_1^\star(s) - V_1^{\widehat{\pi}}(s)\big) \le 80\sqrt{\frac{2c_{\mathrm{b}} H^3 S C^\star_{\mathsf{rob}} \log \frac{NH}{\delta}}{K}} \tag{5.50}$$

with probability at least $1-12\delta$, provided that $K \ge 3872 H S C^\star_{\mathsf{rob}} \log \frac{KH}{\delta}$. Setting the right-hand side of (5.50) to be smaller than ε immediately concludes the proof of Theorem 4, where we have used the fact that $N \le KH$ in $\mathcal{D}_0$. As a consequence, it suffices to establish Theorem 8. In the sequel, we shall assume without loss of generality that we are working on the high-probability event (5.11).

5.3.3.1 Proof of Theorem 8

Step 1: showing that $\widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi}}(s,a)$. This part relies crucially on the following lemma.

Lemma 18. *Consider any $1 \le h \le H$, and any vector $V \in \mathbb{R}^S$ independent of $\widehat{P}_h$ obeying $\|V\|_\infty \le H$. With probability at least $1 - 4\delta/H$, one has*

$$\left|\left(\widehat{P}_{h,s,a} - P_{h,s,a}\right)V\right| \le \sqrt{\frac{48\mathsf{Var}_{\widehat{P}_{h,s,a}}(V)\log\frac{NH}{\delta}}{N_h(s,a)}} + \frac{48H\log\frac{NH}{\delta}}{N_h(s,a)} \tag{5.51}$$

$$\mathsf{Var}_{\widehat{P}_{h,s,a}}(V) \le 2\mathsf{Var}_{P_{h,s,a}}\left(V\right) + \frac{5H^2\log\frac{NH}{\delta}}{3N_h(s,a)} \tag{5.52}$$

simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Proof. The proof follows from exactly the same argument as that of Lemma 31, except that the assumed upper bound on $\|V\|_\infty$ is now H (as opposed to $\frac{1}{1-\gamma}$) and δ is replaced with δ/H. We thus omit the proof details for brevity. □

Additionally, we make note of the crude bound $\left|(\widehat{P}_{h,s,a} - P_{h,s,a})\widehat{V}_{h+1}\right| \le \|\widehat{V}_{h+1}\|_\infty \le H$. Also, given that Algorithm 9 works backwards, the iterate $\widehat{V}_{h+1}$ does not use $\widehat{P}_h$, and is hence statistically independent from $\widehat{P}_h$. Thus, we can readily apply Lemma 18 to obtain

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]: \quad \left|\left(\widehat{P}_{h,s,a} - P_{h,s,a}\right)\widehat{V}_{h+1}\right| \le b_h(s,a) \tag{5.53}$$

in the presence of the Bernstein-style penalty (5.9), provided that the constant $c_{\mathrm{b}} > 0$ is sufficiently large.

In the sequel, we shall work with the high-probability events (5.53) and (5.52), in addition to (5.11). We intend to prove the following relation

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]: \quad \widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi}}(s,a) \quad \text{and} \quad \widehat{V}_h(s) \le V_h^{\widehat{\pi}}(s) \tag{5.54}$$

hold with probability exceeding $1 - 4\delta$. Note that the latter assertion concerning $\widehat{V}_h$ is implied by the former, according to the following relation:

$$\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a) = \widehat{Q}_h\big(s, \widehat{\pi}_h(s)\big) \le Q_h^{\widehat{\pi}}\big(s, \widehat{\pi}_h(s)\big) = V_h^{\widehat{\pi}}(s). \tag{5.55}$$

Therefore, we focus on the first assertion and will show it by induction. First of all, the claim (5.54) holds trivially for the base case with $h = H+1$, given that $\widehat{Q}_{H+1}(s,a) = Q_{H+1}^{\widehat{\pi}}(s,a) = 0$. Next, suppose that $\widehat{Q}_{h+1}(s,a) \le Q_{h+1}^{\widehat{\pi}}(s,a)$ holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and some step $h+1$. We would like to show that the claimed inequality holds for step h as well. If $\widehat{Q}_h(s,a) = 0$, then the claim holds trivially; otherwise, our update rule (5.6) reveals that

$$\begin{aligned}
\widehat{Q}_h(s,a) &= r_h(s,a) + \widehat{P}_{h,s,a}\widehat{V}_{h+1} - b_h(s,a) \\
&= r_h(s,a) + P_{h,s,a}\widehat{V}_{h+1} + \left(\widehat{P}_{h,s,a} - P_{h,s,a}\right)\widehat{V}_{h+1} - b_h(s,a)
\end{aligned}$$

$$\overset{\text{(i)}}{\leq} r_h(s,a) + P_{h,s,a} V_{h+1}^{\widehat{\pi}} \overset{\text{(ii)}}{=} Q_h^{\widehat{\pi}}(s,a),$$

with probability at least $1-\delta/2$, where (i) results from (5.53) and (5.55) (i.e., $\widehat{V}_{h+1}(s) \leq V_{h+1}^{\widehat{\pi}}(s)$), and (ii) arises from the Bellman equation. We have thus established (5.54) via a standard induction argument.

Step 2: bounding $V_h^\star(s) - V_h^{\widehat{\pi}}(s)$. In view of (5.55), we make the observation that

$$0 \leq V_h^\star(s) - V_h^{\widehat{\pi}}(s) \leq V_h^\star(s) - \widehat{V}_h(s) \leq Q_h^\star\big(s,\pi_h^\star(s)\big) - \widehat{Q}_h\big(s,\pi_h^\star(s)\big), \tag{5.56}$$

where the last inequality holds true since $V_h^\star(s) = Q_h^\star(s,\pi_h^\star(s))$ and $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a) \geq \widehat{Q}_h(s,\pi_h^\star(s))$. Recognizing that

$$\begin{aligned}
Q_h^\star\big(s,\pi_h^\star(s)\big) &= r\big(s,\pi_h^\star(s)\big) + P_{h,s,\pi_h^\star(s)} V_{h+1}^\star, \\
\widehat{Q}_h\big(s,\pi_h^\star(s)\big) &= \max\Big\{r\big(s,\pi_h^\star(s)\big) + \widehat{P}_{h,s,\pi_h^\star(s)} \widehat{V}_{h+1} - b_h\big(s,\pi_h^\star(s)\big),\, 0\Big\},
\end{aligned}$$

we can continue the derivation of (5.56) to obtain

$$\begin{aligned}
V_h^\star(s) - \widehat{V}_h(s) &\leq r\big(s,\pi_h^\star(s)\big) + P_{h,s,\pi_h^\star(s)} V_{h+1}^\star - \Big\{r\big(s,\pi_h^\star(s)\big) + \widehat{P}_{h,s,\pi_h^\star(s)} \widehat{V}_{h+1} - b_h\big(s,\pi_h^\star(s)\big)\Big\} \\
&= P_{h,s,\pi_h^\star(s)} V_{h+1}^\star - \widehat{P}_{h,s,\pi_h^\star(s)} \widehat{V}_{h+1} + b_h\big(s,\pi_h^\star(s)\big) \\
&= P_{h,s,\pi_h^\star(s)}\big(V_{h+1}^\star - \widehat{V}_{h+1}\big) - \Big(\widehat{P}_{h,s,\pi_h^\star(s)} - P_{h,s,\pi_h^\star(s)}\Big)\widehat{V}_{h+1} + b_h\big(s,\pi_h^\star(s)\big) \\
&\leq P_{h,s,\pi_h^\star(s)}\big(V_{h+1}^\star - \widehat{V}_{h+1}\big) + 2b_h\big(s,\pi_h^\star(s)\big)
\end{aligned} \tag{5.57}$$

with probability at least $1-\delta$, where the last inequality is valid due to (5.53). For notational convenience, let us introduce a sequence of matrices $P_h^\star \in \mathbb{R}^{S\times S}$ $(1\leq h\leq H)$ and vectors $b_h^\star \in \mathbb{R}^S$ $(1\leq h\leq H)$, with their s-th rows given by

$$\big[P_h^\star\big]_{s,\cdot} \coloneqq P_{h,s,\pi_h^\star(s)} \qquad \text{and} \qquad b_h^\star(s) \coloneqq b_h\big(s,\pi_h^\star(s)\big). \tag{5.58}$$

This allows us to rewrite (5.57) in matrix/vector form as follows:

$$0 \leq V_h^\star - \widehat{V}_h \leq P_h^\star\big(V_{h+1}^\star - \widehat{V}_{h+1}\big) + 2b_h^\star. \tag{5.59}$$

The inequality (5.59) plays a key role in the analysis since it establishes a connection between the value estimation errors in step h and step $h+1$.

Given that $b_h^\star$, $P_h^\star$ and $V_h^\star - \widehat{V}_h$ are all non-negative, applying (5.59) recursively with the

boundary condition $V^\star_{H+1} = \widehat{V}_{H+1} = 0$ leads to

$$\begin{aligned} 0 \le V^\star_h - \widehat{V}_h &\le P^\star_h\big(V^\star_{h+1} - \widehat{V}_{h+1}\big) + 2b^\star_h \\ &\le P^\star_h P^\star_{h+1}\big(V^\star_{h+2} - \widehat{V}_{h+2}\big) + 2P^\star_h b^\star_{h+1} + 2b^\star_h \le \cdots \\ &\le 2\sum_{j=h}^{H}\bigg(\prod_{k=h}^{j-1} P^\star_k\bigg) b^\star_j, \end{aligned}$$

where we adopt the following notation for convenience (note the order of the product)

$$\prod_{k=h}^{h-1} P^\star_k = I \qquad \text{and} \qquad \prod_{k=h}^{j-1} P^\star_k = P^\star_h \cdots P^\star_{j-1} \quad \text{if } j > h.$$

With this inequality in mind, we can let $d^\star_h \coloneqq [d^\star_h(s)]_{s\in\mathcal{S}}$ be a S-dimensional vector and derive

$$\begin{aligned} \big\langle d^\star_h, V^\star_h - V^{\widehat{\pi}}_h\big\rangle \le \big\langle d^\star_h, V^\star_h - \widehat{V}_h\big\rangle &\le \bigg\langle d^\star_h, 2\sum_{j=h}^{H}\bigg(\prod_{k=h}^{j-1} P^\star_k\bigg) b^\star_j\bigg\rangle \\ &= 2\sum_{j=h}^{H}\big(d^\star_h\big)^\top\bigg(\prod_{k=h}^{j-1} P^\star_k\bigg) b^\star_j = 2\sum_{j=h}^{H}\big\langle d^\star_j, b^\star_j\big\rangle, \end{aligned} \tag{5.60}$$

where we have made use of (5.56) and the elementary identity (5.43).

Step 3: using concentrability to bound $\langle d^\star_j, b^\star_j\rangle$. To finish up, we need to make use of the concentrability coefficient. In what follows, we look at two cases separately.

- *Case 1:* $Kd^{\mathsf{b}}_j\big(s,\pi^\star_j(s)\big) \le 4c_{\mathrm{b}}\log\frac{NH}{\delta}$. Given that $b_h(s,a) \le H$ (cf. (5.9)), we necessarily have

$$b^\star_j(s) \le H \le H\cdot\frac{4c_{\mathrm{b}}\log\frac{NH}{\delta}}{Kd^{\mathsf{b}}_j\big(s,\pi^\star_j(s)\big)} \le \frac{4c_{\mathrm{b}}C^\star_{\mathsf{rob}}H\log\frac{NH}{\delta}}{K\min\big\{d^\star_j(s),\frac{1}{S}\big\}} \tag{5.61}$$

  in this case, where the last inequality arises from Definition 2.

- *Case 2:* $Kd^{\mathsf{b}}_j\big(s,\pi^\star_j(s)\big) > 4c_{\mathrm{b}}\log\frac{NH}{\delta}$. It follows from the assumption (5.48) that

$$\begin{aligned} N_j\big(s,\pi^\star_j(s)\big) &\ge \frac{Kd^{\mathsf{b}}_j\big(s,\pi^\star_j(s)\big)}{8} - 5\sqrt{Kd^{\mathsf{b}}_j\big(s,\pi^\star_j(s)\big)\log\frac{N}{\delta}} \ge \frac{Kd^{\mathsf{b}}_j\big(s,\pi^\star_j(s)\big)}{16} \\ &\ge \frac{K\min\big\{d^\star_j\big(s,\pi^\star_j(s)\big),\frac{1}{S}\big\}}{16C^\star_{\mathsf{rob}}} = \frac{K\min\big\{d^\star_j(s),\frac{1}{S}\big\}}{16C^\star_{\mathsf{rob}}}, \end{aligned} \tag{5.62}$$

  as long as $c_{\mathrm{b}} > 0$ is sufficiently large. Here, the last line results from Definition 2 and the assumption that $\pi^\star$ is a deterministic policy (so that $d^\star_j(s) = d^\star_j\big(s,\pi^\star_j(s)\big)$).

This further leads to

$$
\begin{aligned}
b_j^\star(s) &\le \sqrt{\frac{c_{\mathrm{b}} \log \frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)} \mathsf{Var}_{\widehat{P}_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + c_{\mathrm{b}} H \frac{\log \frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)} \\
&\overset{\text{(i)}}{\le} \sqrt{\frac{2c_{\mathrm{b}} \log \frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)} \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + 3c_{\mathrm{b}} H \frac{\log \frac{NH}{\delta}}{N_j\big(s,\pi_j^\star(s)\big)} \\
&\overset{\text{(ii)}}{\le} \sqrt{\frac{32c_{\mathrm{b}} C^\star_{\mathsf{rob}} \log \frac{NH}{\delta}}{K \min\big\{d_j^\star(s), \frac{1}{S}\big\}} \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + 48c_{\mathrm{b}} C^\star_{\mathsf{rob}} H \frac{\log \frac{NH}{\delta}}{K \min\big\{d_j^\star(s), \frac{1}{S}\big\}}.
\end{aligned}
$$

Here, (i) comes from (5.52) and the elementary inequality $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for any $x, y \ge 0$, provided that c_{b} is large enough; and (ii) relies on (5.62).

Putting the above two cases together, we arrive at

$$
\begin{aligned}
\sum_s d_j^\star(s) b_j^\star(s) &\le \sum_s d_j^\star(s) \sqrt{\frac{32c_{\mathrm{b}} C^\star_{\mathsf{rob}} \log \frac{NH}{\delta}}{K \min\big\{d_j^\star(s), \frac{1}{S}\big\}} \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + 48c_{\mathrm{b}} H \sum_s d_j^\star(s) \frac{C^\star_{\mathsf{rob}} \log \frac{NH}{\delta}}{K \min\big\{d_j^\star(s), \frac{1}{S}\big\}} \\
&\le \sum_s d_j^\star(s) \sqrt{\frac{32c_{\mathrm{b}} C^\star_{\mathsf{rob}} \log \frac{NH}{\delta}}{K \min\big\{d_j^\star(s), \frac{1}{S}\big\}} \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + \frac{96c_{\mathrm{b}} HSC^\star_{\mathsf{rob}} \log \frac{NH}{\delta}}{K}, \qquad (5.63)
\end{aligned}
$$

where the last inequality holds since

$$
\sum_s \frac{d_j^\star(s)}{\min\big\{d_j^\star(s), \frac{1}{S}\big\}} \le \sum_s d_j^\star(s) \left\{\frac{1}{d_j^\star(s)} + \frac{1}{1/S}\right\} \le \sum_s 1 + S \sum_s d_j^\star(s) \le 2S.
$$

In addition, we make the observation that

$$
\begin{aligned}
&\sum_{j=h}^H \sum_s d_j^\star(s) \sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)}{\min\big\{d_j^\star(s), \frac{1}{S}\big\}}} \le \sum_{j=h}^H \sum_s d_j^\star(s) \sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)}{d_j^\star(s)}} + \sum_{j=h}^H \sum_s d_j^\star(s) \sqrt{\frac{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)}{1/S}} \\
&= \sum_{j=h}^H \sum_s \sqrt{d_j^\star(s) \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + \sqrt{S} \sum_{j=h}^H \sum_s d_j^\star(s) \sqrt{\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} \\
&\le \sqrt{HS} \cdot \sqrt{\sum_{j=h}^H \sum_s d_j^\star(s) \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} + \sqrt{S} \sqrt{\sum_{j=h}^H \sum_s d_j^\star(s)} \sqrt{\sum_{j=h}^H \sum_s d_j^\star(s) \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} \\
&= 2\sqrt{HS} \cdot \sqrt{\sum_{j=h}^H \sum_s d_j^\star(s) \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big)} \\
&\le 4\sqrt{HS\left(H^2 + H \sum_{j=h}^H \big\langle d_j^\star, b_j^\star \big\rangle\right)} \le 4\sqrt{H^3 S} + 4\sqrt{H^2 S \sum_{j=h}^H \big\langle d_j^\star, b_j^\star \big\rangle},
\end{aligned}
$$

where the third line makes use of the Cauchy-Schwarz inequality, and the last line would hold as long as we could establish the following inequality

$$\sum_{j=h}^{H}\sum_{s} d_j^\star(s)\mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big) \le 4H^2 + 4H\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle \tag{5.64}$$

for all $h \in [H]$ with probability exceeding $1-4\delta$. Substitution into (5.63) yields

$$\begin{aligned}
\sum_{j=h}^{H}\sum_{s} d_j^\star(s)b_j^\star(s) &\le \sqrt{\frac{32c_{\mathrm{b}}C^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K}}\left\{4\sqrt{H^3S} + 4\sqrt{H^2S\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle}\right\} + \sum_{j=h}^{H}\frac{96c_{\mathrm{b}}HSC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K} \\
&\le 16\sqrt{\frac{2c_{\mathrm{b}}H^2SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K}}\sqrt{\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle} + 16\sqrt{\frac{2c_{\mathrm{b}}H^3SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K}} + \frac{96c_{\mathrm{b}}H^2SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K} \\
&\le \frac{1}{2}\sum_{j=h}^{H}\big\langle d_j^\star, b_j^\star\big\rangle + \frac{256c_{\mathrm{b}}H^2SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K} + 16\sqrt{\frac{2c_{\mathrm{b}}H^3SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K}} + \frac{96c_{\mathrm{b}}H^2SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K},
\end{aligned}$$

where the last inequality follows from the elementary inequality $2xy \le x^2+y^2$. Rearranging terms, we are left with

$$\begin{aligned}
\sum_{j=h}^{H}\sum_{s} d_j^\star(s)b_j^\star(s) &\le 32\sqrt{\frac{2c_{\mathrm{b}}H^3SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K}} + \frac{704c_{\mathrm{b}}H^2SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K} \\
&\le 40\sqrt{\frac{2c_{\mathrm{b}}H^3SC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}}{K}},
\end{aligned}$$

provided that $K \ge 3872HSC^\star_{\mathsf{rob}}\log\frac{NH}{\delta}$. This taken collectively with (5.60) completes the proof of Theorem 8, as long as the inequality (5.64) can be validated.

Proof of inequality (5.64). First of all, we observe that

$$\begin{aligned}
\widehat{V}_j(s) + 2b_j^\star\big(s,\pi_j^\star(s)\big) - P_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} &= \widehat{V}_j(s) - \widehat{P}_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} + 2b_j^\star\big(s,\pi_j^\star(s)\big) + \big(\widehat{P}_{j,s,\pi_j^\star(s)} - P_{j,s,\pi_j^\star(s)}\big)\widehat{V}_{j+1} \\
&\overset{\text{(i)}}{\ge} \widehat{V}_j(s) - \widehat{P}_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} + b_j^\star\big(s,\pi_j^\star(s)\big) \\
&\ge \widehat{V}_j(s) - \Big\{r\big(s,\pi_j^\star(s)\big) + \widehat{P}_{j,s,\pi_j^\star(s)}\widehat{V}_{j+1} - b_j^\star\big(s,\pi_j^\star(s)\big)\Big\} \\
&\ge \max_a \widehat{Q}_j(s,a) - \widehat{Q}_j\big(s,\pi_j^\star(s)\big) \ge 0
\end{aligned}$$

for any $s \in \mathcal{S}$, where (i) is a consequence of (5.53), and the last line arises from (5.6) and (5.7). This implies the non-negativity of the vector $\widehat{V}_j + 2b_j^\star - P_j^\star \widehat{V}_{j+1}$, which in turn allows one to deduce that

$$\begin{aligned}
\widehat{V}_j \circ \widehat{V}_j - \big(P_j^\star \widehat{V}_{j+1}\big) \circ \big(P_j^\star \widehat{V}_{j+1}\big) &= \big(\widehat{V}_j + P_j^\star \widehat{V}_{j+1}\big) \circ \big(\widehat{V}_j - P_j^\star \widehat{V}_{j+1}\big) \\
&\le \big(\widehat{V}_j + P_j^\star \widehat{V}_{j+1}\big) \circ \big(\widehat{V}_j + 2b_j^\star - P_j^\star \widehat{V}_{j+1}\big) \\
&\le 2H\big(\widehat{V}_j + 2b_j^\star - P_j^\star \widehat{V}_{j+1}\big), \qquad (5.65)
\end{aligned}$$

where the last line relies on Lemma 16. Consequently, we can demonstrate that

$$\begin{aligned}
&\sum_{j=h}^{H} \sum_s d_j^\star(s) \mathsf{Var}_{P_{j,s,\pi_j^\star(s)}}\big(\widehat{V}_{j+1}\big) = \sum_{j=h}^{H} \big\langle d_j^\star, P_j^\star\big(\widehat{V}_{j+1} \circ \widehat{V}_{j+1}\big) - \big(P_j^\star \widehat{V}_{j+1}\big) \circ \big(P_j^\star \widehat{V}_{j+1}\big)\big\rangle \\
&= \sum_{j=h}^{H} \big(d_j^\star\big)^\top P_j^\star \widehat{V}_{j+1} \circ \widehat{V}_{j+1} - \big\langle d_j^\star, \big(P_j^\star \widehat{V}_{j+1}\big) \circ \big(P_j^\star \widehat{V}_{j+1}\big)\big\rangle \\
&\overset{\text{(i)}}{\le} \sum_{j=h}^{H} \Big(\big\langle d_{j+1}^\star, \widehat{V}_{j+1} \circ \widehat{V}_{j+1}\big\rangle - \big\langle d_j^\star, \widehat{V}_j \circ \widehat{V}_j\big\rangle + 2H\big\langle d_j^\star, \widehat{V}_j + 2b_j^\star - P_j^\star \widehat{V}_{j+1}\big\rangle\Big) \\
&\overset{\text{(ii)}}{=} \sum_{j=h}^{H} \Big(\big\langle d_{j+1}^\star, \widehat{V}_{j+1} \circ \widehat{V}_{j+1}\big\rangle - \big\langle d_j^\star, \widehat{V}_j \circ \widehat{V}_j\big\rangle\Big) + 2H \sum_{j=h}^{H} \Big(\big\langle d_j^\star, \widehat{V}_j\big\rangle - \big\langle d_{j+1}^\star, \widehat{V}_{j+1}\big\rangle\Big) + 4H \sum_{j=h}^{H} \big\langle d_j^\star, b_j^\star\big\rangle \\
&= \big\langle d_{H+1}^\star, \widehat{V}_{H+1} \circ \widehat{V}_{H+1}\big\rangle - \big\langle d_h^\star, \widehat{V}_h \circ \widehat{V}_h\big\rangle + 2H\Big(\big\langle d_h^\star, \widehat{V}_h\big\rangle - \big\langle d_{H+1}^\star, \widehat{V}_{H+1}\big\rangle\Big) + 4H \sum_{j=h}^{H} \big\langle d_j^\star, b_j^\star\big\rangle \\
&\le \big\|d_{H+1}^\star\big\|_1 \big\|\widehat{V}_{H+1} \circ \widehat{V}_{H+1}\big\|_\infty + 2H\big\|d_h^\star\big\|_1 \big\|\widehat{V}_h\big\|_\infty + 4H \sum_{j=h}^{H} \big\langle d_j^\star, b_j^\star\big\rangle \\
&\overset{\text{(iii)}}{\le} 3H^2 + 4H \sum_{j=h}^{H} \big\langle d_j^\star, b_j^\star\big\rangle,
\end{aligned}$$

where (i) arises from (5.65) as well as the basic property $\big(d_j^\star\big)^\top P_j^\star = \big(d_{j+1}^\star\big)^\top$, (ii) follows by rearranging terms and using the property $\big(d_j^\star\big)^\top P_j^\star = \big(d_{j+1}^\star\big)^\top$ once again, and (iii) holds due to the fact that $\|\widehat{V}_h\|_\infty \le H$ and $\|d_h^\star\|_1 = 1$. This concludes the proof of (5.64).

5.4 Analysis: discounted infinite-horizon MDPs

This subchapter is devoted to establishing Theorem 6. Towards this end, we claim that it is sufficient to prove the following theorem.

Theorem 9. *Consider any $0 < \delta < 1$ and any $\gamma \in [\frac{1}{2}, 1)$. Suppose that the penalty terms are set to be (5.33) for any numerical constant $c_{\mathrm{b}} \ge 144$. Then with probability exceeding $1 - 2\delta$, for any*

estimate $\widehat{Q}$ *obeying* $\left\|\widehat{Q}-\widehat{Q}^{\star}_{\mathsf{pe}}\right\|_{\infty} \leq 1/N$ *one has*

$$V^{\star}(\rho)-V^{\widehat{\pi}}(\rho) \leq 120\sqrt{\frac{c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3N}}+\frac{3464c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^2N}, \tag{5.66}$$

where $\widehat{\pi}(s)\in\arg\max_a \widehat{Q}(s,a)$ *for any* $s\in\mathcal{S}$.

As we have demonstrated in Lemma 15, the output of Algorithm 11 satisfies $\left\|\widehat{Q}-\widehat{Q}^{\star}_{\mathsf{pe}}\right\|_{\infty}\leq 1/N$ once the iteration number exceeds $\tau_{\max}\geq\frac{\log\frac{N}{1-\gamma}}{\log(1/\gamma)}$, thus making Theorem 9 applicable. Taking the right-hand side of (5.66) to be no larger than ε reveals that: $V^{\star}(\rho)-V^{\widehat{\pi}}(\rho)\leq\varepsilon$ holds as long as N exceeds

$$N\geq\frac{21000c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3\varepsilon^2}, \tag{5.67}$$

given that $\varepsilon\in\big(0,\frac{1}{1-\gamma}\big]$.

The remainder of this subchapter is thus dedicated to establishing Theorem 9. Throughout the proof, it suffices to focus on the case where

$$N\geq\frac{c_3SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{1-\gamma} \tag{5.68}$$

for some large constant $c_3\geq 2880000$; otherwise the claim (5.66) follows directly since $V^{\star}(\rho)-V^{\widehat{\pi}}(\rho)\leq\frac{1}{1-\gamma}$.

5.4.1 Preliminary facts

Before embarking on the proof, we collect a couple of preliminary facts that will be used multiple times.

Properties of $N(s,a)$. To begin with, the quantity $N(s,a)$ — the total number of sample transitions from (s,a) — can be bounded through the following lemma; the proof is provided in Appendix C.2.3.

Lemma 19. *Consider any* $\delta\in(0,1)$. *With probability at least* $1-\delta$, *the quantities* $\{N(s,a)\}$ *in* (5.29) *obey*

$$\max\left\{N(s,a),\frac{2}{3}\log\frac{N}{\delta}\right\}\geq\frac{Nd^{\mathsf{b}}(s,a)}{12} \tag{5.69}$$

simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

Properties about $\widehat{V}$ and $\widehat{V}^\star_{\mathsf{pe}}$. First of all, note that the assumption

$$\big\|\widehat{Q} - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \le \frac{1}{N} \tag{5.70}$$

has the following direct consequence:

$$\big\|\widehat{V} - \widehat{V}^\star_{\mathsf{pe}}\big\|_\infty = \max_s \Big|\max_a \widehat{Q}(s,a) - \max_a \widehat{Q}^\star_{\mathsf{pe}}(s,a)\Big| \le \big\|\widehat{Q} - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \le \frac{1}{N}. \tag{5.71}$$

Given the proximity of $\widehat{V}$ and $\widehat{V}^\star_{\mathsf{pe}}$, we can bound the difference of the corresponding variance terms as follows:

$$\begin{aligned}
\Big|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{\mathsf{pe}}\big) - \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\Big| &\overset{\text{(i)}}{=} \Big|\widehat{P}_{s,a}\big(\widehat{V}^\star_{\mathsf{pe}} \circ \widehat{V}^\star_{\mathsf{pe}}\big) - \widehat{P}_{s,a}\big(\widehat{V} \circ \widehat{V}\big) + \big(\widehat{P}_{s,a}\widehat{V}\big)^2 - \big(\widehat{P}_{s,a}\widehat{V}^\star_{\mathsf{pe}}\big)^2\Big| \\
&= \Big|\widehat{P}_{s,a}\Big(\big(\widehat{V} + \widehat{V}^\star_{\mathsf{pe}}\big) \circ \big(\widehat{V}^\star_{\mathsf{pe}} - \widehat{V}\big)\Big) + \Big(\widehat{P}_{s,a}\big(\widehat{V} + \widehat{V}^\star_{\mathsf{pe}}\big)\Big)\Big(\widehat{P}_{s,a}\big(\widehat{V}^\star_{\mathsf{pe}} - \widehat{V}\big)\Big)\Big| \\
&\le \big\|\widehat{P}_{s,a}\big\|_1 \big\|\widehat{V} + \widehat{V}^\star_{\mathsf{pe}}\big\|_\infty \big\|\widehat{V}^\star_{\mathsf{pe}} - \widehat{V}\big\|_\infty + \big\|\widehat{P}_{s,a}\big\|_1^2 \big\|\widehat{V} + \widehat{V}^\star_{\mathsf{pe}}\big\|_\infty \big\|\widehat{V}^\star_{\mathsf{pe}} - \widehat{V}\big\|_\infty \\
&\le \Big(\big\|\widehat{P}_{s,a}\big\|_1 + \big\|\widehat{P}_{s,a}\big\|_1^2\Big)\Big(2\big\|\widehat{V}\big\|_\infty + \big\|\widehat{V}^\star_{\mathsf{pe}} - \widehat{V}\big\|_\infty\Big)\big\|\widehat{V}^\star_{\mathsf{pe}} - \widehat{V}\big\|_\infty \\
&\le \frac{2}{N}\left(\frac{2}{1-\gamma} + \frac{1}{N}\right) \le \frac{6}{(1-\gamma)N}.
\end{aligned} \tag{5.72}$$

Here, (i) follows from the definition (1.7), the penultimate inequality follows from (5.71) and the basic facts $\|\widehat{P}_{s,a}\|_1 = 1$ and $\|\widehat{V}\|_\infty \le \frac{1}{1-\gamma}$, while the last line relies on (5.68).

Armed with (5.72), one can further control the difference of the associated penalty terms. Note that the definition of $b(s,a;V)$ in (5.33) tells us that

$$\begin{aligned}
\Big|b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big) - b\big(s,a;\widehat{V}\big)\Big| = \Bigg|&\min\left\{\max\left\{\sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{\mathsf{pe}}\big)}, \frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\}, \frac{1}{1-\gamma}\right\} \\
&- \min\left\{\max\left\{\sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)}, \frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\}, \frac{1}{1-\gamma}\right\}\Bigg|.
\end{aligned} \tag{5.73}$$

If at least one of the variance terms is not too small in the sense that

$$\max\Big\{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{\mathsf{pe}}\big), \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\Big\} \ge \frac{4c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a)}, \tag{5.74}$$

then (5.73) implies that

$$(5.73) \le \sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}}\Big|\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{\mathsf{pe}}\big)} - \sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)}\Big| = \sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}}\frac{\big|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{\mathsf{pe}}\big) - \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\big|}{\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{\mathsf{pe}}\big)} + \sqrt{\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)}}$$

$$\overset{\text{(i)}}{\leq} \frac{1-\gamma}{2}\left|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}_{\mathsf{pe}}^{\star}\big) - \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}\big)\right| \overset{\text{(ii)}}{\leq} \frac{3}{N}, \tag{5.75}$$

where (i) results from (5.74), and (ii) holds due to (5.72). On the other hand, if (5.74) is not satisfied, then one clearly has $b\big(s,a;\widehat{V}_{\mathsf{pe}}^{\star}\big) = b\big(s,a;\widehat{V}\big)$. In conclusion, in all cases we have

$$\left|b\big(s,a;\widehat{V}_{\mathsf{pe}}^{\star}\big) - b\big(s,a;\widehat{V}\big)\right| \leq \frac{3}{N}. \tag{5.76}$$

5.4.2 Proof of Theorem 9

Armed with the preceding preliminary facts, we can readily turn to the proof of Theorem 9. By virtue of Lemma 19, our proof shall — unless otherwise noted — operate on the high-probability event that

$$\forall(s,a)\in\mathcal{S}\times\mathcal{A}: \qquad \max\left\{N(s,a), \frac{2}{3}\log\frac{SN}{\delta}\right\} \geq \frac{Nd^{\mathsf{b}}(s,a)}{12}. \tag{5.77}$$

In addition, from the sampling model (5.22), the sample transitions employed to form $\widehat{P}$ are statistically independent conditional on $\{N(s,a)\}$. Our proof consists of four steps as detailed below.

Step 1: Bernstein-style inequalities and leave-one-out decoupling argument. We are in need of tight control of the size of $\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}$. However, this becomes challenging due to the statistical dependency between $\widehat{P}$ and the value estimate $\widehat{V}$ (given that we reuse samples in all iterations of Algorithm 11). In order to circumvent this difficulty, we resort to a leave-one-out argument to decouple the statistical dependency, as motivated by Agarwal et al. (2020b); Li et al. (2023c). The result stated below establishes Bernstein-style inequalities despite the complicated dependency.

Lemma 20. *Suppose that $\gamma\in[\frac{1}{2},1)$, and consider any $\delta\in(0,1)$. With probability at least $1-\delta$, we have*

$$\left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widetilde{V}\right| \leq 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big)} + \frac{74\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \tag{5.78a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widetilde{V}\big) \leq 2\mathsf{Var}_{P_{s,a}}\big(\widetilde{V}\big) + \frac{41\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2N(s,a)} \tag{5.78b}$$

simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and all $\widetilde{V}$ with $\big\|\widetilde{V}-\widehat{V}_{\mathsf{pe}}^{\star}\big\|_{\infty}\leq\frac{1}{N}$ and $\|\widetilde{V}\|_{\infty}\leq\frac{1}{1-\gamma}$.

High-level proof ideas. In short, the proof consists of contructing a finite collection of auxiliary MDPs $\{\widehat{\mathcal{M}}^{s,u}\}$ for each state s obeying the following properties: (i) each $\widehat{\mathcal{M}}^{s,u}$ is constructed without

using any sample transition that comes from state s, and is hence statistically independent from $\widehat{P}_{s,a}$ for all $a \in \mathcal{A}$ (instead, the useful information is embedded into the corresponding immediate reward, which is a low-dimensional object and easier to control); (ii) at least one of the MDPs in $\{\widehat{\mathcal{M}}^{s,u}\}$ is extremely close to the true MDP in terms of the resulting value function. With the aid of these leave-one-out auxiliary MDPs, one can control $\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}$ by first exploiting the statistical independence between $\widehat{P}_{s,a}$ and $\{\widehat{\mathcal{M}}^{s,u}\}$ and then transferring the concentration bound back to the original MDP using the proximity property (ii). The construction of these auxiliary MDPs and the proof details can be found in Appendix C.2.4. □

Note that (5.78a) has been derived only for those pairs (s,a) with $N(s,a) > 0$. For every (s,a) with $N(s,a) = 0$, one can directly obtain

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}\Big| = \big|P_{s,a}\widetilde{V}\big| \le \big\|P_{s,a}\big\|_1 \big\|\widetilde{V}\big\|_\infty \le \frac{1}{1-\gamma}.$$

Putting these bounds together with the definition (5.33) of $b(s,a;V)$ reveals that

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widetilde{V}\Big| + \frac{5}{N} \le b\big(s,a;\widetilde{V}\big) \qquad \text{for all } (s,a) \in \mathcal{S}\times\mathcal{A} \tag{5.79}$$

for all $\widetilde{V}$ obeying $\big\|\widetilde{V} - \widehat{V}^\star_{\mathsf{pe}}\big\|_\infty \le \frac{1}{N}$ and $\|\widetilde{V}\|_\infty \le \frac{1}{1-\gamma}$, provided that the constant c_{b} is sufficiently large. The remainder of the proof should then also operate on the high-probability events (5.79) and (5.78b), in addition to assuming that the event (5.77) occurs.

Step 2: showing that $\widehat{Q}(s,a)$ is a lower bound on $Q^{\widehat{\pi}}(s,a)$. We now justify that $\widehat{Q}(s,a)$ (resp. $\widehat{V}(s)$) is a "pessimistic" estimate of $Q^{\widehat{\pi}}(s,a)$ (resp. $V^{\widehat{\pi}}(s)$); this is enabled by the pessimism principle (so that the algorithm effectively seeks lower estimates of the value iteration) and the Bernstein-style bounds in Lemma 20 (so that the penalty term always dominates the uncertainty incurred by using the empirical MDP).

To begin with, recall that $\widehat{Q}^\star_{\mathsf{pe}}(s,a)$ is the unique fixed point of the pessimistic Bellman operator that obeys

$$\widehat{Q}^\star_{\mathsf{pe}}(s,a) = \max\Big\{r(s,a) + \gamma\widehat{P}_{s,a}\widehat{V}^\star_{\mathsf{pe}} - b\big(s,a;\widehat{V}^\star_{\mathsf{pe}}\big), 0\Big\}. \tag{5.80}$$

In the sequel, we divide the set of state-action pairs (s,a) into two types.

- *Case 1:* $\widehat{Q}^\star_{\mathsf{pe}}(s,a) = 0$. Given that $\widehat{Q}_0 = 0$, Lemma 15 tells us that

$$\widehat{Q}(s,a) = \widehat{Q}_{\tau_{\max}}(s,a) \le \widehat{Q}^\star_{\mathsf{pe}}(s,a) = 0.$$

This combined with the basic fact $Q^{\widehat{\pi}} \ge 0$ immediately yields $0 = \widehat{Q}(s,a) \le Q^{\widehat{\pi}}(s,a)$.

- *Case 2:* $\widehat{Q}^{\star}_{\mathsf{pe}}(s,a) = r(s,a) + \gamma \widehat{P}_{s,a}\widehat{V}^{\star}_{\mathsf{pe}} - b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big) > 0$. It is first observed that

$$\begin{aligned}
\widehat{Q}(s,a) &\overset{\text{(i)}}{\leq} \widehat{Q}^{\star}_{\mathsf{pe}}(s,a) + \frac{1}{N} \overset{\text{(ii)}}{=} r(s,a) - b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big) + \gamma \widehat{P}_{s,a}\widehat{V}^{\star}_{\mathsf{pe}} + \frac{1}{N} \\
&\leq r(s,a) - b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big) + \gamma \widehat{P}_{s,a}\widehat{V} + \frac{1}{N} + \gamma \big\|\widehat{P}_{s,a}\big\|_1 \big\|\widehat{V} - \widehat{V}^{\star}_{\mathsf{pe}}\big\|_{\infty} \\
&\overset{\text{(iii)}}{\leq} r(s,a) - b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big) + \gamma \widehat{P}_{s,a}\widehat{V} + \frac{2}{N} \\
&\leq r(s,a) - b\big(s,a;\widehat{V}\big) + \gamma P_{s,a}\widehat{V} + \frac{2}{N} + \gamma \Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)\widehat{V}\Big| + \Big|b\big(s,a;\widehat{V}^{\star}_{\mathsf{pe}}\big) - b\big(s,a;\widehat{V}\big)\Big| \\
&\overset{\text{(iv)}}{\leq} r(s,a) + \gamma P_{s,a}\widehat{V}. \qquad (5.81)
\end{aligned}$$

  Here, (i) and (iii) arise from the assumption (5.70), (ii) relies on the fact that $\widehat{Q}^{\star}_{\mathsf{pe}}$ is the fixed point of the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}$, whereas (iv) takes advantage of (5.76) and (5.79). Combining (5.81) with the Bellman equation $Q^{\widehat{\pi}} = r + \gamma P V^{\widehat{\pi}}$ results in

$$Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a) \geq r(s,a) + \gamma P_{s,a}V^{\widehat{\pi}} - \big(r(s,a) + \gamma P_{s,a}\widehat{V}\big) = \gamma P_{s,a}\big(V^{\widehat{\pi}} - \widehat{V}\big). \qquad (5.82)$$

  Suppose for the moment that there exists some (s,a) obeying $Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a) < 0$ (which clearly cannot happen in Case 1), then $\arg\min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big]$ must belong to Case 2. Thus, taking the minimum over (s,a) and using the above inequality (5.82) give

$$\begin{aligned}
\min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big] &\geq \min_{s,a}\big[\gamma P_{s,a}\big(V^{\widehat{\pi}} - \widehat{V}\big)\big] \overset{\text{(i)}}{\geq} \gamma \min_{s}\big[V^{\widehat{\pi}}(s) - \widehat{V}(s)\big] \\
&= \gamma \min_{s}\big[Q^{\widehat{\pi}}\big(s,\widehat{\pi}(s)\big) - \widehat{Q}\big(s,\widehat{\pi}(s)\big)\big] \geq \gamma \min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big], \qquad (5.83)
\end{aligned}$$

  where (i) holds since $P_{s,a} \in \Delta(\mathcal{S})$. Given that $1 > \gamma > 0$, inequality (5.83) holds only when $\min_{s,a}\big[Q^{\widehat{\pi}}(s,a) - \widehat{Q}(s,a)\big] \geq 0$. We therefore conclude that in this case, one also has $Q^{\widehat{\pi}}(s,a) \geq \widehat{Q}(s,a)$.

With the arguments for the above two cases in place, we arrive at

$$Q^{\widehat{\pi}}(s,a) \geq \widehat{Q}(s,a) \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \qquad (5.84)$$

and evidently,

$$V^{\star}(s) \geq V^{\widehat{\pi}}(s) = Q^{\widehat{\pi}}\big(s,\widehat{\pi}(s)\big) \geq \widehat{Q}\big(s,\widehat{\pi}(s)\big) = \max_{a}\widehat{Q}(s,a) = \widehat{V}(s) \qquad \text{for all } s \in \mathcal{S}. \qquad (5.85)$$

Step 3: bounding $V^\star(s) - V^{\widehat{\pi}}(s)$. Recall that the Bellman optimality equation gives

$$V^\star(s) = r\big(s, \pi^\star(s)\big) + \gamma P_{s,\pi^\star(s)} V^\star. \tag{5.86}$$

Before continuing, we make note of the following lower bound on $\widehat{V}$:

$$\begin{aligned}
\widehat{V}(s) = \max_a \widehat{Q}(s,a) \geq \widehat{Q}\big(s, \pi^\star(s)\big) &\overset{\text{(i)}}{\geq} \widehat{Q}^\star_{\mathsf{pe}}\big(s, \pi^\star(s)\big) - \frac{1}{N} \\
&\overset{\text{(ii)}}{\geq} r\big(s, \pi^\star(s)\big) - b\big(s, \pi^\star(s); \widehat{V}^\star_{\mathsf{pe}}\big) + \gamma \widehat{P}_{s,\pi^\star(s)} \widehat{V}^\star_{\mathsf{pe}} - \frac{1}{N} \\
&= r\big(s, \pi^\star(s)\big) - b\big(s, \pi^\star(s); \widehat{V}^\star_{\mathsf{pe}}\big) + \gamma \widehat{P}_{s,\pi^\star(s)} \widehat{V} - \frac{1}{N} - \gamma \widehat{P}_{s,\pi^\star(s)} \big(\widehat{V} - \widehat{V}^\star_{\mathsf{pe}}\big) \\
&\overset{\text{(iii)}}{\geq} r\big(s, \pi^\star(s)\big) - b\big(s, \pi^\star(s); \widehat{V}^\star_{\mathsf{pe}}\big) + \gamma \widehat{P}_{s,\pi^\star(s)} \widehat{V} - \frac{2}{N} \\
&\geq r\big(s, \pi^\star(s)\big) - b\big(s, \pi^\star(s); \widehat{V}\big) + \gamma P_{s,\pi^\star(s)} \widehat{V} - \frac{2}{N} - \gamma \Big| \big(\widehat{P}_{s,\pi^\star(s)} - P_{s,\pi^\star(s)}\big) \widehat{V} \Big| \\
&\qquad - \Big| b\big(s, \pi^\star(s); \widehat{V}^\star_{\mathsf{pe}}\big) - b\big(s, \pi^\star(s); \widehat{V}\big) \Big| \\
&\overset{\text{(iv)}}{\geq} r\big(s, \pi^\star(s)\big) - 2b\big(s, \pi^\star(s); \widehat{V}\big) + \gamma P_{s,\pi^\star(s)} \widehat{V}.
\end{aligned} \tag{5.87}$$

Here, (i) results from the assumption (5.70), (ii) relies on (5.80), (iii) is valid since $\widehat{P}_{s,\pi^\star(s)}\big(\widehat{V} - \widehat{V}^\star_{\mathsf{pe}}\big) \leq \big\|\widehat{P}_{s,\pi^\star(s)}\big\|_1 \big\|\widehat{V} - \widehat{V}^\star_{\mathsf{pe}}\big\|_\infty \leq 1/N$, whereas (iv) holds by virtue of (5.76) and (5.79). Armed with the results in (5.86) and (5.87), we can readily show that

$$\begin{aligned}
\big\langle \rho, V^\star - \widehat{V} \big\rangle &= \sum_{s \in \mathcal{S}} \rho(s) \big(V^\star(s) - \widehat{V}(s)\big) \\
&\leq \sum_{s \in \mathcal{S}} \rho(s) \Big\{ r\big(s, \pi^\star(s)\big) + \gamma P_{s,\pi^\star(s)} V^\star - \Big(r\big(s, \pi^\star(s)\big) - 2b\big(s, \pi^\star(s); \widehat{V}\big) + \gamma P_{s,\pi^\star(s)} \widehat{V} \Big) \Big\} \\
&\leq \gamma \sum_{s \in \mathcal{S}} \rho(s) P_{s,\pi^\star(s)} \big(V^\star - \widehat{V}\big) + 2 \sum_{s \in \mathcal{S}} \rho(s) b\big(s, \pi^\star(s); \widehat{V}\big).
\end{aligned} \tag{5.88}$$

For notational convenience, let us introduce a matrix $P^\star \in \mathbb{R}^{S \times S}$ and a vector $b^\star \in \mathbb{R}^{S \times 1}$ whose s-th row are given respectively by

$$\big[P^\star\big]_{s,\cdot} := P_{s,\pi^\star(s)} \qquad \text{and} \qquad b^\star(s) := b\big(s, \pi^\star(s); \widehat{V}\big) \qquad \text{for all } s \in \mathcal{S}. \tag{5.89}$$

This allows us to rewrite (5.88) in the following matrix/vector form:

$$\rho^\top \big(V^\star - \widehat{V}\big) \leq \gamma \rho^\top P^\star \big(V^\star - \widehat{V}\big) + 2\rho^\top b^\star. \tag{5.90}$$

Note that this relation holds for any arbitrary ρ. Apply it recursively to arrive at

$$\begin{aligned}
\rho^\top\big(V^\star - \widehat{V}\big) &\le \big(\gamma\rho^\top P^\star\big)\big(V^\star - \widehat{V}\big) + 2\rho^\top b^\star \\
&\le \gamma\big(\gamma\rho^\top P^\star\big)P^\star\big(V^\star - \widehat{V}\big) + 2\big(\gamma\rho^\top P^\star\big)b^\star + 2\rho^\top b^\star \\
&= \gamma^2\rho^\top\big(P^\star\big)^2\big(V^\star - \widehat{V}\big) + 2\gamma\rho^\top P^\star b^\star + 2\rho^\top b^\star \\
&\le \cdots \le \left\{\lim_{i\to\infty}\gamma^i\rho^\top\big(P^\star\big)^i\big(V^\star - \widehat{V}\big)\right\} + 2\rho^\top\left\{\sum_{i=0}^{\infty}\gamma^i\big(P^\star\big)^i\right\}b^\star \\
&\overset{\text{(i)}}{=} 2\rho^\top\left\{\sum_{i=0}^{\infty}\gamma^i\big(P^\star\big)^i\right\}b^\star = 2\rho^\top\big(I - \gamma P^\star\big)^{-1}b^\star \\
&= \frac{2}{1-\gamma}\langle d^\star, b^\star\rangle,
\end{aligned} \tag{5.91}$$

where (i) holds since $\lim_{i\to\infty}\gamma^i\rho^\top\big(P^\star\big)^i\big(V^\star - \widehat{V}\big) = 0$ (given that $\lim_{i\to\infty}\gamma^i = 0$ and $\|\rho^\top\big(P^\star\big)^i\|_1 = 1$ for any $i \ge 0$), and the last equality results from the definition of $d^\star$ (see (5.21)) expressed in the following matrix/vector form:

$$\big(d^\star\big)^\top = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\rho^\top\big(P^\star\big)^t = (1-\gamma)\rho^\top\big(I - \gamma P^\star\big)^{-1}. \tag{5.92}$$

Combine the above inequality with (5.85) to reach

$$\big\langle\rho, V^\star - V^{\widehat{\pi}}\big\rangle \le \big\langle\rho, V^\star - \widehat{V}\big\rangle \le \frac{2\big\langle d^\star, b^\star\big\rangle}{1-\gamma}. \tag{5.93}$$

Step 4: using concentrability to control $\big\langle d^\star, b^\star\big\rangle$. We shall control $\big\langle d^\star, b^\star\big\rangle$ by dividing the state set $\mathcal{S}$ into the following two disjoint subsets:

$$\mathcal{S}^{\mathsf{small}} \coloneqq \left\{s \in \mathcal{S} \mid Nd^{\mathsf{b}}\big(s, \pi^\star(s)\big) \le 8\log\frac{NS}{(1-\gamma)\delta}\right\}; \tag{5.94a}$$

$$\mathcal{S}^{\mathsf{large}} \coloneqq \left\{s \in \mathcal{S} \mid Nd^{\mathsf{b}}\big(s, \pi^\star(s)\big) > 8\log\frac{NS}{(1-\gamma)\delta}\right\}. \tag{5.94b}$$

- To begin with, consider any state $s \in \mathcal{S}^{\mathsf{small}}$. Applying Definition 4 and the definition of $\mathcal{S}^{\mathsf{small}}$ yields

$$\min\left\{d^\star(s), \frac{1}{S}\right\} \le C^\star_{\mathsf{rob}}d^{\mathsf{b}}\big(s, \pi^\star(s)\big) \le \frac{8C^\star_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{N} < \frac{1}{S}, \tag{5.95}$$

provided that $N > 8SC^\star_{\mathsf{rob}} \log \frac{NS}{(1-\gamma)\delta}$ (see (5.68)). This inequality necessarily implies that

$$d^\star(s) \leq \frac{8C^\star_{\mathsf{rob}} \log \frac{NS}{(1-\gamma)\delta}}{N} < \frac{1}{S}. \tag{5.96}$$

Combining the preceding inequality with the following fact (see the definition (5.33))

$$b^\star(s) \coloneqq b\big(s, \pi^\star(s); \widehat{V}\big) \leq \frac{1}{1-\gamma} + \frac{5}{N}, \tag{5.97}$$

we arrive at

$$\sum_{s \in \mathcal{S}^{\mathsf{small}}} d^\star(s) b^\star(s) \leq \sum_{s \in \mathcal{S}^{\mathsf{small}}} \left(\frac{8C^\star_{\mathsf{rob}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + d^\star(s) \frac{5}{N} \right) \leq \frac{8SC^\star_{\mathsf{rob}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + \frac{5}{N}. \tag{5.98}$$

- Next, we turn to any state $s \in \mathcal{S}^{\mathsf{large}}$. Using the definition (5.33) of $b(s,a;V)$, we obtain

$$\begin{aligned} b^\star(s) = b\big(s, \pi^\star(s); \widehat{V}\big) &\leq \sqrt{\frac{c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N\big(s, \pi^\star(s)\big)} \mathsf{Var}_{\widehat{P}_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{2c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N\big(s, \pi^\star(s)\big)}} + \frac{5}{N} \\ &\overset{\text{(i)}}{\leq} \sqrt{\frac{c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N\big(s, \pi^\star(s)\big)} \left(2\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{41 \log \frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2 N\big(s, \pi^\star(s)\big)} \right)} + \frac{2c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N\big(s, \pi^\star(s)\big)} + \frac{5}{N} \\ &\overset{\text{(ii)}}{\leq} \sqrt{\frac{2c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N\big(s, \pi^\star(s)\big)} \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{4c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N\big(s, \pi^\star(s)\big)}}, \end{aligned} \tag{5.99}$$

where (i) arises from Lemma 20 and (5.71), (ii) applies the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y \geq 0$ and the fact $N \geq N(s,a)$, in addition to assuming that c_{b} is large enough. To continue, we observe that

$$\frac{1}{N\big(s, \pi^\star(s)\big)} \overset{\text{(i)}}{\leq} \frac{12}{Nd^{\mathsf{b}}\big(s, \pi^\star(s)\big)} \overset{\text{(ii)}}{\leq} \frac{12C^\star_{\mathsf{rob}}}{N \min\big\{d^\star(s), \frac{1}{S}\big\}} \leq \frac{12C^\star_{\mathsf{rob}}}{N} \left(\frac{1}{d^\star(s)} + S \right), \tag{5.100}$$

where (i) follows from the assumption (5.77) and the definition of $\mathcal{S}^{\mathsf{large}}$, and (ii) results from Assumption 4. Substitution into (5.99) yields

$$b^\star(s) \leq \underbrace{\sqrt{\frac{24c_{\mathrm{b}}C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{N} \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} \left(\frac{1}{\sqrt{d^\star(s)}} + \sqrt{S} \right)}_{=: \alpha_1(s)} + \underbrace{\frac{48c_{\mathrm{b}}C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N} \left(\frac{1}{d^\star(s)} + S \right)}_{=: \alpha_2(s)}, \tag{5.101}$$

where the last line comes from the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y \geq 0$. To proceed, observe that the sum of the first terms in (5.101) satisfies

$$\begin{aligned}
&\sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_1(s) \\
&= \sqrt{\frac{24 c_{\mathrm{b}} C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{N}} \left(\sum_{s \in \mathcal{S}^{\mathsf{large}}} \sqrt{d^\star(s) \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \sum_{s \in \mathcal{S}^{\mathsf{large}}} \sqrt{d^\star(s)} \sqrt{S d^\star(s) \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} \right) \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{24 c_{\mathrm{b}} C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{N}} \left(\sqrt{S} \cdot \sqrt{\sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s) \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \sqrt{\sum_{s \in \mathcal{S}^{\mathsf{large}}} S d^\star(s) \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} \right) \\
&= \sqrt{\frac{96 c_{\mathrm{b}} S C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{N}} \sqrt{\sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s) \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)}, \qquad (5.102)
\end{aligned}$$

where (i) arises from the Cauchy-Schwarz inequality and the fact $\sum_s d^\star(s) = 1$. In addition, it is easily verified that the sum of the second terms in (5.101) obeys

$$\sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_2(s) \leq \frac{96 c_{\mathrm{b}} S C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}, \qquad (5.103)$$

which also makes use of the identity $\sum_s d^\star(s) = 1$. Combining (5.102) and (5.103) with (5.101) gives

$$\begin{aligned}
\sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s) b^\star\big(s, \pi^\star(s)\big) &\leq \sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_1(s) + \sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_2(s) \\
&\leq \sqrt{\frac{96 c_{\mathrm{b}} S C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{N}} \sqrt{\sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s) \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \frac{96 c_{\mathrm{b}} S C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N}. \qquad (5.104)
\end{aligned}$$

The above results (5.98) and (5.104) taken collectively give

$$\begin{aligned}
\big\langle d^\star, b^\star \big\rangle &= \sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s) b^\star(s) + \sum_{s \in \mathcal{S}^{\mathsf{small}}} d^\star(s) b^\star(s) \\
&\leq \sqrt{\frac{96 c_{\mathrm{b}} S C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{N}} \sqrt{\sum_{s \in \mathcal{S}^{\mathsf{large}}} d^\star(s) \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \frac{96 c_{\mathrm{b}} S C^\star_{\mathsf{rob}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N} \\
&\quad + \frac{8 S C^\star_{\mathsf{rob}} \log \frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + \frac{5}{N}
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{N}}\sqrt{\sum_{s\in\mathcal{S}} d^{\star}(s)\mathsf{Var}_{P_{s,\pi^{\star}(s)}}\big(\widehat{V}\big)} + \frac{98c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} \\
&\overset{\text{(ii)}}{\leq} \frac{2}{\gamma}\sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}\big\langle d^{\star}, b^{\star}\big\rangle} + \frac{1}{\gamma}\sqrt{\frac{192c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{98c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}, \\
&\overset{\text{(iii)}}{\leq} 4\sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}\big\langle d^{\star}, b^{\star}\big\rangle} + 2\sqrt{\frac{192c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{98c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}, \\
&\overset{\text{(iv)}}{\leq} \frac{1}{2}\big\langle d^{\star}, b^{\star}\big\rangle + \frac{768c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N} + \sqrt{\frac{768c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{98c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}.
\end{aligned}
$$

Here, (i) follows when c_{b} is sufficiently large and $C^{\star}_{\mathsf{rob}} \geq 1/S$ (see (5.26)), (ii) would hold as long as the following inequality could be established:

$$
\sum_{s\in\mathcal{S}} d^{\star}(s)\mathsf{Var}_{P_{s,\pi^{\star}(s)}}\big(\widehat{V}\big) \leq \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)}\big\langle d^{\star}, b^{\star}\big\rangle; \tag{5.105}
$$

(iii) is valid since $\gamma \in [\frac{1}{2}, 1)$, and (iv) follows from the elementary inequality $2xy \leq x^2 + y^2$. Rearranging terms, we are left with

$$
\big\langle d^{\star}, b^{\star}\big\rangle \leq \sqrt{\frac{3072c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}} + \frac{1732c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)N}, \tag{5.106}
$$

which combined with (5.93) yields

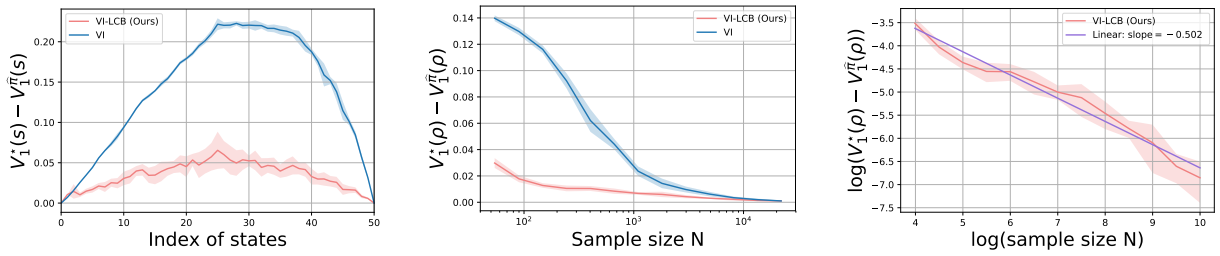
$$
\big\langle \rho, V^{\star} - V^{\widehat{\pi}}\big\rangle \leq \frac{2\big\langle d^{\star}, b^{\star}\big\rangle}{1-\gamma} \leq 120\sqrt{\frac{c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^3N}} + \frac{3464c_{\mathrm{b}}SC^{\star}_{\mathsf{rob}}\log\frac{NS}{(1-\gamma)\delta}}{(1-\gamma)^2N}. \tag{5.107}
$$

This concludes the proof, as long as the inequality (5.105) can be established.

Proof of inequality (5.105). To begin with, we make the observation that

$$
\begin{aligned}
\big(\widehat{V}\circ\widehat{V}\big) - \big(\gamma P^{\star}\widehat{V}\big)\circ\big(\gamma P^{\star}\widehat{V}\big) &= \big(\widehat{V} - \gamma P^{\star}\widehat{V}\big)\circ\big(\widehat{V} + \gamma P^{\star}\widehat{V}\big) \\
&\overset{\text{(i)}}{\leq} \big(\widehat{V} - \gamma P^{\star}\widehat{V} + 2b^{\star}\big)\circ\big(\widehat{V} + \gamma P^{\star}\widehat{V}\big) \\
&\overset{\text{(ii)}}{\leq} \frac{2}{1-\gamma}\big(\widehat{V} - \gamma P^{\star}\widehat{V} + 2b^{\star}\big),
\end{aligned} \tag{5.108}
$$

where (i) holds since $b^{\star} \geq 0$ and $\widehat{V} + \gamma P^{\star}\widehat{V} \geq 0$, (ii) follows from the basic property $\|\widehat{V} + \gamma P^{\star}\widehat{V}\|_{\infty} \leq 2\|\widehat{V}\|_{\infty} \leq \frac{2}{1-\gamma}$ and the fact $\widehat{V} - \gamma P^{\star}\widehat{V} + 2b^{\star} \geq 0$, the latter of which has been verified in (5.87).



(a) Value gap versus states (b) Value gap versus sample size (c) Value gap dependency w.r.t. N

Figure 5.1: The performances of the proposed method VI-LCB and the baseline value iteration (VI) in the gambler's problem. It shows that VI-LCB outperforms VI by taking advantage of the pessimism principle and achieves approximately $1/\sqrt{N}$ sample complexity dependency w.r.t. the sample size N.

Armed with this fact, one can deduce that

$$
\begin{aligned}
\sum_s d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) &\overset{\text{(i)}}{=} \Big\langle d^\star, P^\star\big(\widehat{V}\circ\widehat{V}\big) - \big(P^\star\widehat{V}\big)\circ\big(P^\star\widehat{V}\big)\Big\rangle \\
&\overset{\text{(ii)}}{\le} \left\langle d^\star, P^\star\big(\widehat{V}\circ\widehat{V}\big) - \frac{1}{\gamma^2}\widehat{V}\circ\widehat{V} + \frac{2}{\gamma^2(1-\gamma)}\big(\widehat{V} - \gamma P^\star\widehat{V} + 2b^\star\big)\right\rangle \\
&\overset{\text{(iii)}}{\le} \left\langle d^\star, P^\star\big(\widehat{V}\circ\widehat{V}\big) - \frac{1}{\gamma}\widehat{V}\circ\widehat{V} + \frac{2}{\gamma^2(1-\gamma)}(I-\gamma P^\star)\widehat{V} + \frac{4}{\gamma^2(1-\gamma)}b^\star\right\rangle \\
&= \left\langle d^\star, \frac{1}{\gamma}\big(\gamma P^\star - I\big)\big(\widehat{V}\circ\widehat{V}\big) + \frac{2}{\gamma^2(1-\gamma)}(I-\gamma P^\star)\widehat{V} + \frac{4}{\gamma^2(1-\gamma)}b^\star\right\rangle \\
&= d^{\star\top}\big(I-\gamma P^\star\big)\left\{-\frac{1}{\gamma}\widehat{V}\circ\widehat{V} + \frac{2}{\gamma^2(1-\gamma)}\widehat{V}\right\} + \frac{4}{\gamma^2(1-\gamma)}\langle d^\star, b^\star\rangle \\
&\overset{\text{(iv)}}{\le} (1-\gamma)\rho^\top\left\{-\frac{1}{\gamma}\widehat{V}\circ\widehat{V} + \frac{2}{\gamma^2(1-\gamma)}\widehat{V}\right\} + \frac{4}{\gamma^2(1-\gamma)}\langle d^\star, b^\star\rangle \\
&\le \frac{2}{\gamma^2}\rho^\top\widehat{V} + \frac{4}{\gamma^2(1-\gamma)}\langle d^\star, b^\star\rangle \\
&\overset{\text{(v)}}{\le} \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)}\langle d^\star, b^\star\rangle\,.
\end{aligned}
$$

Here, (i) follows by invoking the definition (1.7), (ii) holds due to (5.108), (iii) is valid since $\gamma < 1$, (iv) is a direct consequence of (5.92), while (v) comes from the basic facts $\|\rho^\top\|_1 = 1$ and $\|\widehat{V}\|_\infty \le \frac{1}{1-\gamma}$.

5.5 Numerical experiments

To confirm the practical applicability of the proposed VI-LCB algorithm, we evaluate its performance in the gambler's problem (Panaganti and Kalathil, 2022; Shi and Chi, 2022; Sutton and Barto, 2018; Zhou et al., 2021). The code can be accessed at:

https://github.com/Laixishi/Model-based-VI-LCB.

Gambler's problem. We start by introducing the formulation of the gambler's problem and its underlying MDP. An agent plays a gambling game in which she bets on a sequence of random coin flips, winning when the coins are heads and losing when they are tails. To bet on each random clip, the agent's policy chooses an integer number of dollars based on an initial balance. If the number of bets hits the maximum length H, or if the agent reaches 50 dollars (win) or 0 dollars (lose), the game ends. Without loss of generality, the problem can be formulated as an episodic finite-horizon MDP. Here, $\mathcal{S}$ is the state space $\{0, 1, \cdots, 50\}$ and the associated accessible actions obey $a \in \big\{0, 1, \cdots, \min\{s, 50 - s\}\big\}$, $H = 100$ is the horizon length, the reward is set to 0 for all other states unless $s = 50$. For the transition kernel, we fix the probability of heads as $p_{\text{head}} = 0.45$ at all steps $h \in [H]$ in the episode. Moreover, the initial state/balance distribution of the agent ρ is taken as a uniform distribution over $\mathcal{S}$. The offline historical dataset is constructed by collecting N independent samples drawn randomly over each state-action pair and time step.

Evaluation results. First, we evaluate the performance of our proposed method VI-LCB (cf. Algorithm 9) with comparisons to the well-known value iteration (VI) method without the pessimism principle. To begin with, Figure 5.1(a) shows the average and standard derivations of the performance gap $V_1^\star(s) - V_1^{\widehat{\pi}}(s)$ over all states $s \in \mathcal{S}$, over 10 independent experiments with a fixed sample size $N = 50$. The results indicate that the proposed VI-LCB method outperforms the baseline VI method uniformly over the entire state space, showing that pessimism brings significant advantages in this sample-scarce regime. Secondly, we evaluate the performance gap $V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho)$ with varying sample size $N \in \big\{54, 90, 148, \cdots, 22026\big\} \approx \{e^4, e^{4.5}, e^5, \cdots, e^{10}\}$, over 10 independent trials. Note that throughout the experiments, we fix the parameter $c_{\text{b}} = 0.05$, which determines the level of the pessimism penalty of VI-LCB (cf. (5.9)). Figure 5.1(b) shows the average and standard derivations of the performance gap $V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho)$ with respect to the sample size N. Clearly, as the sample size increases, both our method VI-LCB and the baseline VI method perform better. Moreover, our VI-LCB method consistently outperforms the baseline VI method over the entire range of the sample size N, especially in the sample-starved regime. In addition, to corroborate the scaling of the sample size on the performance gap, we plot the sub-optimality performance gap of VI-LCB w.r.t. the sample size on a log-log scale in Figure 5.1(c). Fitting using linear regression leads to a slope estimate of -0.502, with the corresponding fitted line plotted in Figure 5.1(c) as well. This nicely matches the finding of Theorem 4, which says the performance gap of VI-LCB scales as $N^{-1/2}$.

5.6 Discussions

Our primary contribution has been to pin down the sample complexity of model-based offline RL for the tabular settings, by establishing its (near) minimax optimality for both infinite- and

finite-horizon MDPs. While reliable estimation of the transition kernel is often infeasible in the sample-starved regime, it does not preclude the success of this "plug-in" approach in learning the optimal policy. Encouragingly, the sample complexity characterization we have derived holds for the entire range of target accuracy level ε, thus revealing that sample optimality comes into effect without incurring any burn-in cost. This is in stark contrast to all prior results, which either suffered from sample sub-optimality or required a large burn-in sample size in order to yield optimal efficiency. We have demonstrated that sophisticated techniques like variance reduction are not necessary, as long as Bernstein-style lower confidence bounds are carefully employed to capture the variance of the estimates in each iteration.

Turning to future directions, we note that the two-fold subsampling adopted in Algorithm 10 is likely unnecessary; it would be of interest to develop sharp analysis for the VI-LCB algorithm without sample splitting, which would call for more refined analysis in order to handle the complicated statistical dependency between different time steps. Notably, while avoiding sample splitting cannot improve the sample complexity in an order-wise sense, the potential gain in terms of the pre-constants as well as the algoritmic simplicity might be of practical interest.

Chapter 6

Model-Based Robust RL with a Generative Model

6.1 Problem formulation

In this chapter, recalling the distributionally robust Markov decision processes (RMDPs) in the discounted infinite-horizon setting in Chapter 2.2.2, we introduce the sampling mechanism, and describe our goal.

Specification of the divergence ρ. Recall that the uncertainty set $\mathcal{U}_\rho^\sigma(\cdot)$ defined in (2.23), we shall first specify the divergence measure function $\rho(\cdot)$. We consider two popular choices of the uncertainty set measured in terms of two different f-divergence metric: the total variation distance and the χ^2 divergence, given respectively by (Tsybakov, 2009)

$$\rho_{\mathsf{TV}}\left(P_{s,a}, P_{s,a}^0\right) := \frac{1}{2}\left\|P_{s,a} - P_{s,a}^0\right\|_1 = \frac{1}{2}\sum_{s'\in\mathcal{S}} P^0(s' \,|\, s,a)\left|1 - \frac{P(s' \,|\, s,a)}{P^0(s' \,|\, s,a)}\right|, \tag{6.1}$$

$$\rho_{\chi^2}\left(P_{s,a}, P_{s,a}^0\right) := \sum_{s'\in\mathcal{S}} P^0(s' \,|\, s,a)\left(1 - \frac{P(s' \,|\, s,a)}{P^0(s' \,|\, s,a)}\right)^2. \tag{6.2}$$

Note that $\rho_{\mathsf{TV}}\left(P_{s,a}, P_{s,a}^0\right) \in [0,1]$ and $\rho_{\chi^2}\left(P_{s,a}, P_{s,a}^0\right) \in [0,\infty)$ in general. As we shall see shortly, these two choices of divergence metrics result in drastically different messages when it comes to sample complexities.

Sampling mechanism: a generative model. Following Panaganti and Kalathil (2022); Zhou et al. (2021), we assume access to a generative model or a simulator (Kearns and Singh, 1999), which allows us to collect N independent samples for each state-action pair generated based on the *nominal* kernel P^0:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \qquad s_{i,s,a} \overset{i.i.d}{\sim} P^0(\cdot \,|\, s,a), \qquad i = 1,2,\cdots,N. \tag{6.3}$$

The total sample size is, therefore, NSA.

Goal. Given the collected samples, the task is to learn the robust optimal policy for the RMDP — w.r.t. some prescribed uncertainty set $\mathcal{U}^\sigma(P^0)$ around the nominal kernel — using as few samples as possible. Specifically, given some target accuracy level $\varepsilon > 0$, the goal is to seek an ε-optimal robust policy $\widehat{\pi}$ obeying

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq \varepsilon. \tag{6.4}$$

6.2 Distributionally robust value iteration

We consider a model-based approach tailored to RMDPs, which first constructs an empirical nominal transition kernel based on the collected samples, and then applies distributionally robust value iteration (DRVI) to compute an optimal robust policy.

Empirical nominal kernel. The empirical nominal transition kernel $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$ can be constructed on the basis of the empirical frequency of state transitions, i.e.,

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widehat{P}^0(s' \,|\, s,a) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\big\{s_{i,s,a} = s'\big\}, \tag{6.5}$$

which leads to an empirical RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}} = \{\mathcal{S},\mathcal{A},\gamma,\mathcal{U}_\rho^\sigma(\widehat{P}^0), r\}$. Analogously, we can define the corresponding robust value function (resp. robust Q-function) of policy π in $\widehat{\mathcal{M}}_{\mathsf{rob}}$ as $\widehat{V}^{\pi,\sigma}$ (resp. $\widehat{Q}^{\pi,\sigma}$) (cf. (2.26)). In addition, we denote the corresponding *optimal robust policy* as $\widehat{\pi}^\star$ and the *optimal robust value function* (resp. *optimal robust Q-function*) as $\widehat{V}^{\star,\sigma}$ (resp. $\widehat{Q}^{\star,\sigma}$) (cf. (2.27)), which satisfies the robust Bellman optimality equation:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widehat{Q}^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}_\rho^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{6.6}$$

Equipped with $\widehat{P}^0$, we can define the empirical robust Bellman operator $\widehat{\mathcal{T}}^\sigma$ as

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widehat{\mathcal{T}}^\sigma(Q)(s,a) \coloneqq r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}_\rho^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V, \quad \text{with} \quad V(s) \coloneqq \max_a Q(s,a). \tag{6.7}$$

DRVI: distributionally robust value iteration. To compute the fixed point of $\widehat{\mathcal{T}}^\sigma$, we introduce distributionally robust value iteration (DRVI), which is summarized in Algorithm 12. From an initialization $\widehat{Q}_0 = 0$, the update rule at the t-th ($t \geq 1$) iteration can be formulated as:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad \widehat{Q}_t(s,a) = \widehat{\mathcal{T}}^\sigma\big(\widehat{Q}_{t-1}\big)(s,a) = r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}_\rho^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}_{t-1}, \tag{6.8}$$

Algorithm 12: Distributionally robust value iteration (DRVI) for infinite-horizon RMDPs.

input: empirical nominal transition kernel $\widehat{P}^0$; reward function r; uncertainty level σ; number of iterations T.
initialization: $\widehat{Q}_0(s,a)=0$, $\widehat{V}_0(s)=0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.
for $t=1,2,\cdots,T$ **do**
  for $s\in\mathcal{S}, a\in\mathcal{A}$ **do**
    Set $\widehat{Q}_t(s,a)$ according to (6.8);
  for $s\in\mathcal{S}$ **do**
    Set $\widehat{V}_t(s)=\max_a \widehat{Q}_t(s,a)$;
output: $\widehat{Q}_T$, $\widehat{V}_T$ and $\widehat{\pi}$ obeying $\widehat{\pi}(s)\coloneqq \arg\max_a \widehat{Q}_T(s,a)$.

where $\widehat{V}_{t-1}(s)=\max_a \widehat{Q}_{t-1}(s,a)$ for all $s\in\mathcal{S}$. However, directly solving (6.8) is computationally expensive since it involves optimization over an S-dimensional probability simplex at each iteration, especially when the dimension of the state space $\mathcal{S}$ is large. Fortunately, in view of strong duality (Iyengar, 2005), (6.8) can be equivalently solved using its dual problem, which concerns optimizing a *scalar* dual variable and thus can be solved efficiently. The specific form of the dual problem depends on the choice of the divergence ρ, which we shall discuss separately in Appendix D.1.2. To complete the description, we output the greedy policy of the final Q-estimate $\widehat{Q}_T$ as the final policy $\widehat{\pi}$, namely,

$$\forall s\in\mathcal{S}: \quad \widehat{\pi}(s)=\arg\max_a \widehat{Q}_T(s,a). \tag{6.9}$$

Encouragingly, the iterates $\{\widehat{Q}_t\}_{t\geq 0}$ of DRVI converge linearly to the fixed point $\widehat{Q}^{\star,\sigma}$, owing to the appealing γ-contraction property of $\widehat{\mathcal{T}}^{\sigma}$.

6.3 Theoretical guarantees: sample complexity analyses

We now present our main results, which concern the sample complexities of learning RMDPs when the uncertainty set is specified using the TV distance or the χ^2 divergence. Somewhat surprisingly, different choices of the uncertainty set can lead to dramatically different consequences in the sample size requirement.

6.3.1 The case of TV distance: RMDPs are easier to learn than standard MDPs

We start with the case where the uncertainty set is measured via the TV distance. The following theorem develops an upper bound on the sample complexity of DRVI in order to return an ε-optimal robust policy. The key challenge of the analysis lies in careful control of the robust value function

$V^{\pi,\sigma}$ as a function of the uncertainty level σ.

Theorem 10 (Upper bound under TV distance). *Let the uncertainty set be $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\mathsf{TV}}^\sigma(\cdot)$, as specified by the TV distance (6.1). Consider any discount factor $\gamma \in \left[\frac{1}{4}, 1\right)$, uncertainty level $\sigma \in (0,1)$, and $\delta \in (0,1)$. Let $\widehat{\pi}$ be the output policy of Algorithm 12 after $T = C_1 \log\left(\frac{N}{1-\gamma}\right)$ iterations. Then with probability at least $1-\delta$, one has*

$$\forall s \in \mathcal{S}: \quad V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \le \varepsilon \tag{6.10}$$

for any $\varepsilon \in \left(0, \sqrt{1/\max\{1-\gamma, \sigma\}}\right]$, as long as the total number of samples obeys

$$NSA \ge \frac{C_2 SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2} \log\left(\frac{SAN}{(1-\gamma)\delta}\right). \tag{6.11}$$

Here, $C_1, C_2 > 0$ are some large enough universal constants.

Remark 4. Note that Theorem 10 is not only valid when invoking Algorithm 12. In fact, the theorem holds for any oracle planning algorithm (designed based on the empirical transitions $\widehat{P}^0$) whose output policy $\widehat{\pi}$ obeys

$$\left\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty \le O\left(\frac{(1-\gamma)^2}{N} \log\left(\frac{SAN}{(1-\gamma)\delta}\right)\right). \tag{6.12}$$

Before discussing the implications of Theorem 10, we present a matching minimax lower bound that confirms the tightness and optimality of the upper bound, which in turn pins down the sample complexity requirement for learning RMDPs with TV distance. The proof is based on constructing new hard instances inspired by the asymmetric structure of RMDPs.

Theorem 11 (Lower bound under TV distance). *Consider any tuple $(S, A, \gamma, \sigma, \varepsilon)$ obeying $\sigma \in (0, 1-c_0]$ with $0 < c_0 \le \frac{1}{8}$ being any small enough positive constant, $\gamma \in \left[\frac{1}{2}, 1\right)$, and $\varepsilon \in \left(0, \frac{c_0}{256(1-\gamma)}\right]$. We can construct two infinite-horizon RMDPs $\mathcal{M}_0, \mathcal{M}_1$ defined by the uncertainty set $\mathcal{U}_\rho^\sigma(\cdot) = \mathcal{U}_{\mathsf{TV}}^\sigma(\cdot)$, an initial state distribution φ, and a dataset with N independent samples for each state-action pair over the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that*

$$\inf_{\widehat{\pi}} \max\left\{\mathbb{P}_0\left(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\right), \mathbb{P}_1\left(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\right)\right\} \ge \frac{1}{8},$$

provided that

$$NSA \le \frac{c_0 SA \log 2}{8192(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}.$$

Here, the infimum is taken over all estimators $\widehat{\pi}$, and $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) denotes the probability when the RMDP is $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

Below, we interpret the above theorems and highlight several key implications about the sample complexity requirements for learning RMDPs for the case w.r.t. the TV distance.

Near minimax-optimal sample complexity. Theorem 10 shows that the total number of samples required for DRVI (or any oracle planning algorithm claimed in Remark 4) to yield ε-accuracy is

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}\right). \tag{6.13}$$

Taken together with the minimax lower bound asserted by Theorem 11, this confirms the near optimality of the sample complexity (up to some logarithmic factor) almost over the full range of the uncertainty level σ. Importantly, this sample complexityscales linearly with the size of the state-action space, and is inversely proportional to σ in the regime where $\sigma \gtrsim 1-\gamma$.

RMDPs is easier than standard MDPs with TV distance. Recall that the sample complexity requirement for learning standard MDPs with a generative model is (Agarwal et al., 2020a; Azar et al., 2013; Li et al., 2023c)

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right) \tag{6.14}$$

in order to yield ε accuracy. Comparing this with the sample complexity requirement in (6.13) for RMDPs under the TV distance, we confirm that the latter is at least as easy as — if not easier than — standard MDPs. In particular, when $\sigma \lesssim 1-\gamma$ is small, the sample complexity of RMDPs is the same as that of standard MDPs as in (6.14), which is as anticipated since the RMDP reduces to the standard MDP when $\sigma = 0$. On the other hand, when $1-\gamma \lesssim \sigma < 1$, the sample complexity of RMDPs simplifies to

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}\right), \tag{6.15}$$

which is smaller than that of standard MDPs by a factor of $\sigma/(1-\gamma)$.

Comparison with state-of-the-art bounds. While the state-of-the-art sample complexity upper bound derived in Clavier et al. (2023) is tight when σ is small (i.e., $\sigma \lesssim 1-\gamma$), the sample complexity bound therein scales as $\widetilde{O}(\frac{SA}{(1-\gamma)^4\varepsilon^2})$ in the regime where $1-\gamma \lesssim \sigma < 1$. Consequently, this is worse than our result by a factor of

$$\frac{\sigma}{(1-\gamma)^2} \in \left(\frac{1}{1-\gamma}, \frac{1}{(1-\gamma)^2}\right).$$

Turning to the lower bound side, Yang et al. (2022) developed a lower bound for RMDPs under the TV distance, which scales as

$$\widetilde{O}\left(\frac{SA(1-\gamma)}{\varepsilon^2}\min\left\{\frac{1}{(1-\gamma)^4},\frac{1}{\sigma^4}\right\}\right).$$

Clearly, this is worse than ours by a factor of $\frac{\sigma^3}{(1-\gamma)^3}\in\left(1,\frac{1}{(1-\gamma)^3}\right)$ in the regime where $1-\gamma\lesssim\sigma<1$.

6.3.2 The case of χ^2 divergence: RMDPs can be harder than standard MDPs

We now switch attention to the case when the uncertainty set is measured via the χ^2 divergence. The theorem below presents an upper bound on the sample complexity for this case.

Theorem 12 (Upper bound under χ^2 divergence)**.** *Let the uncertainty set be $\mathcal{U}_\rho^\sigma(\cdot)=\mathcal{U}_{\chi^2}^\sigma(\cdot)$, as specified using the χ^2 divergence (6.2). Consider any uncertainty level $\sigma\in(0,\infty)$, $\gamma\in[1/4,1)$ and $\delta\in(0,1)$. With probability at least $1-\delta$, the output policy $\widehat{\pi}$ from Algorithm 12 with at most $T=c_1\log\left(\frac{N}{1-\gamma}\right)$ iterations yields*

$$\forall s\in\mathcal{S}:\quad V^{\star,\sigma}(s)-V^{\widehat{\pi},\sigma}(s)\le\varepsilon \tag{6.16}$$

for any $\varepsilon\in\left(0,\frac{1}{1-\gamma}\right]$, as long as the total number of samples obeying

$$NSA\ge\frac{c_2SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\log\left(\frac{SAN}{\delta}\right). \tag{6.17}$$

Here, $c_1,c_2>0$ are some large enough universal constants.

Remark 5. Akin to Remark 4, the sample complexity derived in Theorem 12 continues to hold for any oracle planning algorithm that outputs a policy $\widehat{\pi}$ obeying $\left\|\widehat{V}^{\star,\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty\le O\left(\frac{\log(\frac{SAN}{(1-\gamma)\delta})}{N^2}\right)$.

In addition, in order to gauge the tightness of Theorem 12 and understand the minimal sample complexity requirement under the χ^2 divergence, we further develop a minimax lower bound as follows.

Theorem 13 (Lower bound under χ^2 divergence)**.** *Consider any $(S,A,\gamma,\sigma,\varepsilon)$ obeying $\gamma\in[\frac{3}{4},1)$, $\sigma\in(0,\infty)$, and*

$$\varepsilon\le c_3\begin{cases}\frac{1}{1-\gamma} & \text{if } \sigma\in\left(0,\frac{1-\gamma}{4}\right)\\ \max\left\{\frac{1}{(1+\sigma)(1-\gamma)},1\right\} & \text{if } \sigma\in\left[\frac{1-\gamma}{4},\infty\right)\end{cases} \tag{6.18}$$

for some small universal constant $c_3>0$. Then we can construct two infinite-horizon RMDPs $\mathcal{M}_0,\mathcal{M}_1$ defined by the uncertainty set $\mathcal{U}_\rho^\sigma(\cdot)=\mathcal{U}_{\chi^2}^\sigma(\cdot)$, an initial state distribution φ, and a dataset

with N independent samples per (s,a) pair over the nominal transition kernel (for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively), such that

$$\inf_{\widehat{\pi}} \max\left\{\mathbb{P}_0\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big),\, \mathbb{P}_1\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big)\right\} \geq \frac{1}{8}, \tag{6.19}$$

provided that the total number of samples

$$NSA \leq c_4 \begin{cases} \frac{SA}{(1-\gamma)^3\varepsilon^2} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \frac{\sigma SA}{\min\{1,(1-\gamma)^4(1+\sigma)^4\}\varepsilon^2} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases} \tag{6.20}$$

for some universal constant $c_4 > 0$.

We are now positioned to single out several key implications of the above theorems.

Nearly tight sample complexity. In order to achieve ε-accuracy for RMDPs under the χ^2 divergence, Theorem 12 asserts that a total number of samples on the order of

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right). \tag{6.21}$$

is sufficient for DRVI (or any other oracle planning algorithm as discussed in Remark 5). Taking this together with the minimax lower bound in Theorem 13 confirms that the sample complexity is near-optimal — up to a polynomial factor of the effective horizon $\frac{1}{1-\gamma}$ — over the entire range of the uncertainty level σ. In particular,

- when $\sigma \asymp 1$, our sample complexity $\widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right)$ is sharp and matches the minimax lower bound;
- when $\sigma \gtrsim \frac{1}{(1-\gamma)^3}$, our sample complexity correctly predicts the linear dependency with σ, suggesting that more samples are needed when one wishes to account for a larger χ^2-based uncertainty sets.

RMDPs can be much harder to learn than standard MDPs with χ^2 divergence. The minimax lower bound developed in Theorem 13 exhibits a curious non-monotonic behavior of the sample size requirement over the entire range of the uncertainty level $\sigma \in (0,\infty)$ when the uncertainty set is measured via the χ^2 divergence. When $\sigma \lesssim 1-\gamma$, the lower bound reduces to

$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right),$$

which matches with that of standard MDPs, as $\sigma = 0$ corresponds to standard MDP. However, two additional regimes are worth calling out:

$$1-\gamma \lesssim \sigma \lesssim \frac{1}{(1-\gamma)^{1/3}}: \qquad \widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\min\left\{\sigma, \frac{1}{\sigma^3}\right\}\right),$$
$$\sigma \gtrsim \frac{1}{(1-\gamma)^3}: \qquad \widetilde{O}\left(\frac{SA\sigma}{\varepsilon^2}\right),$$

both of which are *greater* than that of standard MDPs, indicating learning RMDPs under the χ^2 divergence can be much harder.

Comparison with state-of-the-art bounds. Our upper bound significantly improves over the prior art $\widetilde{O}\left(\frac{S^2A(1+\sigma)}{(1-\gamma)^4\varepsilon^2}\right)$ of Panaganti and Kalathil (2022) by a factor of S, and provides the *first* finite-sample complexity that scales *linearly* with respect to S for discounted infinite-horizon RMDPs, which typically exhibit more complicated statistical dependencies than the finite-horizon counterpart. On the other hand, Yang et al. (2022) established a lower bound on the order of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^2\sigma\varepsilon^2}\right)$ when $\sigma \gtrsim 1-\gamma$, which is always smaller than the requirement of standard MDPs, and diminishes when σ grows. Consequently, Yang et al. (2022) does not lead to the rigorous justification that RMDPs can be much harder than standard MDPs, nor the correct linear scaling of the sample size as σ grows.

6.4 Discussions

In this chapter,we have developed improved sample complexity bounds for learning RMDPs when the uncertainty set is measured via the TV distance or the χ^2 divergence, assuming availability of a generative model. Our results have not only strengthened the prior art in both the upper and lower bounds, but have also unlocked curious insights into how the quest for distributional robustness impacts the sample complexity. As a key takeaway of this chapter, RMDPs are not necessarily harder nor easier to learn than standard MDPs, as the answer depends — in a rather subtle maner — on the specific choice of the uncertainty set. For the case w.r.t. the TV distance, we have settled the minimax sample complexity for RMDPs, which is never larger than that required to learn standard MDPs. Regarding the case w.r.t. the χ^2 divergence, we have uncovered that learning RMDPs can oftentimes be provably harder than the standard MDP counterpart. All in all, our findings help raise awareness that the choice of the uncertainty set not only represents a preference in robustness, but also exerts fundamental influences upon the intrinsic statistical complexity.

Chapter 7

Model-Based Robust Offline RL

7.1 Algorithm and theory: episodic finite-horizon RMDPs

In this chapter, we present a model-based algorithm — namely DRVI-LCB — for robust offline RL, along with its performance guarantees.

7.1.1 Problem formulation and assumptions

Specification of the divergence ρ. Recall that the uncertainty set $\mathcal{U}_\rho^\sigma(\cdot)$ defined in (2.23), in this work, we consider a popular choice of the uncertainty set measured in terms of f-divergence metric: Kullback-Leibler (KL) divergence, given by (Tsybakov, 2009)

$$\rho_{\mathsf{KL}}\left(\mathcal{P}, \mathcal{Q}\right) := \sum_{s' \in \mathcal{S}} \mathcal{P}(s') \log \frac{\mathcal{P}(s')}{\mathcal{Q}(s')} \in [0, \infty), \tag{7.1}$$

where $\mathcal{P}$ and $\mathcal{Q}$ are any distribution obeying $\mathcal{P} \in \Delta(\mathcal{S}), \mathcal{Q} \in \Delta(\mathcal{S})$. It directly leads to the corresponding uncertainty set:

$$\mathcal{U}_{\mathsf{KL}}^\sigma(P^0) := \mathcal{U}^\sigma(P^0) := \otimes \mathcal{U}^\sigma(P^0_{h,s,a}), \qquad \mathcal{U}^\sigma(P^0_{h,s,a}) := \left\{P_{h,s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{h,s,a} \,\|\, P^0_{h,s,a}\right) \le \sigma\right\}. \tag{7.2}$$

In words, the KL divergence between the true transition probability vector and the nominal one at each state-action pair is at most σ; moreover, the RMDP reduces to the standard MDP when $\sigma = 0$.

Sampling mechanism: batch data. Let $\mathcal{D}$ be a history/batch dataset, which consists of a collection of K *independent* episodes generated based on executing a behavior policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$ in some nominal MDP $\mathcal{M}^0 = \left(\mathcal{S}, \mathcal{A}, H, P^0 := \{P_h^0\}_{h=1}^H, \{r_h\}_{h=1}^H\right)$. More specifically, for $1 \le k \le K$, the k-th episode $\left(s_1^k, a_1^k, \ldots, s_H^k, a_H^k, s_{H+1}^k\right)$ is generated according to

$$s_1^k \sim \varphi^{\mathsf{b}}, \qquad a_h^k \sim \pi_h^{\mathsf{b}}(\cdot \,|\, s_h^k) \qquad \text{and} \qquad s_{h+1}^k \sim P_h^0(\cdot \,|\, s_h^k, a_h^k), \qquad 1 \le h \le H. \tag{7.3}$$

Throughout the chapter, φ^{b} represents for some initial distribution associated with the history dataset. Then, recalling the notations about occupancy distribution in Chapter 2.2.1, we introduce

the following short-hand notation for the occupancy distribution w.r.t. π^{b}:

$$\forall (h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}: \qquad d_h^{\mathsf{b},P^0}(s) \coloneqq d_h^{\pi^{\mathsf{b}},P^0}(s), \qquad d_h^{\mathsf{b},P^0}(s,a) \coloneqq d_h^{\pi^{\mathsf{b}},P^0}(s,a). \tag{7.4}$$

Robust single-policy clipped concentrability. To quantify the quality of the history dataset, it is desirable to capture the distribution mismatch between the history dataset and the desired ones, inspired by the *single-policy clipped concentrability* assumption in Definition 2, we introduce a tailored assumption for robust MDPs as follows.

Definition 5 (Robust single-policy clipped concentrability)**.** The behavior policy of the history dataset $\mathcal{D}$ satisfies

$$\max_{(s,a,h,P) \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^0}(s,a)} \le C^\star_{\mathsf{rob}} \tag{7.5}$$

for some quantity $C^\star_{\mathsf{rob}} \in \left[\frac{1}{S}, \infty\right]$. Here, we take $C^\star_{\mathsf{rob}}$ to be the smallest quantity satisfying (7.5), and refer to it as the robust single-policy clipped concentrability coefficient. In addition, we follow the convention $0/0 = 0$.

In words, $C^\star_{\mathsf{rob}}$ measures the worst-case discrepancy — between the optimal robust policy $\pi^\star$ in any model $P \in \mathcal{U}^\sigma(P^0)$ within the uncertainty set and the behavior policy π^{b} in the nominal model P^0 — in terms of the maximum density ratio of the state-action occupancy distributions.

- *Distribution shift.* When the uncertainty level $\sigma = 0$, Assumption 5 reduces back to the single-policy clipped concentrability in Definition 2 for standard offline RL, a weaker notion that can be S times smaller than the single-policy concentrability adopted in (Rashidinejad et al., 2021; Shi et al., 2022; Xie et al., 2021b). On the other end, whenever $\sigma > 0$, the proposed robust single-policy clipped concentrability accounts for the distribution shift not only due to the policies in use ($\pi^\star$ versus π^{b}), but also the underlying environments ($P \in \mathcal{U}^\sigma(P^0)$ versus P^0).

- *Partial coverage.* As long as $C^\star_{\mathsf{rob}}$ is finite, i.e., $C^\star_{\mathsf{rob}} < \infty$, it admits the scenarios when the history dataset only provides *partial coverage* over the entire state-action space, as long as the behavior policy π^{b} visits the state-action pairs that are visited by the optimal robust policy $\pi^\star$ under at least one model in the uncertainty set.

Remark 6. To facilitate comparison with prior works assuming full coverage, we can bound $C^\star_{\mathsf{rob}}$ when the batch dataset is generated using a simulator (Panaganti and Kalathil, 2022; Yang et al., 2022); namely, we can generate sample state transitions based on the transition kernel of the nominal MDP for all state-action pairs at all time steps. In this case, it amounts to that $d_h^{\mathsf{b},P^0}(s,a) = \frac{1}{SA}$

for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, which directly leads to the bound

$$C^{\star}_{\mathsf{rob}} = \max_{(s,a,h,P) \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{U}^{\sigma}(P^0)} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^0}(s,a)} \leq \frac{1/S}{1/(SA)} = A.$$

Goal. With the history dataset $\mathcal{D}$ in hand, our goal is to find a near-optimal robust policy $\widehat{\pi}$, which satisfies

$$V_1^{\widehat{\pi},\sigma}(\varphi) \geq V_1^{\star,\sigma}(\varphi) - \varepsilon \tag{7.6}$$

using as few samples as possible, where ε is the target accuracy level, and

$$V_1^{\pi,\sigma}(\varphi) \coloneqq \mathop{\mathbb{E}}_{s_1\sim\varphi}\left[V_1^{\pi,\sigma}(s_1)\right] \qquad \text{and} \qquad V_1^{\star,\sigma}(\varphi) \coloneqq \mathop{\mathbb{E}}_{s_1\sim\varphi}\left[V_1^{\star,\sigma}(s_1)\right] \tag{7.7}$$

are evaluated when the initial state s_1 is drawn from a given distribution φ.

7.1.2 Proposed algorithm: a pessimistic variant of robust value iteration

Building an empirical nominal MDP. For a moment, imagine we have access to N *independent* sample transitions $\mathcal{D}_0 := \{(h_i, s_i, a_i, s_i')\}_{i=1}^N$ drawn from the transition kernel P^0 of the nominal MDP $\mathcal{M}^0$, where each sample (h_i, s_i, a_i, s_i') indicates the transition from state s_i to state s_i' when action a_i is taken at step h_i, drawn according to $s_i' \sim P_{h_i}^0(\cdot \,|\, s_i, a_i)$. It is then natural to build an empirical estimate $\widehat{P}^0 = \{\widehat{P}_h^0\}_{h=1}^H$ of P^0 based on the empirical frequencies of state transitions, where

$$\widehat{P}_h^0(s' \,|\, s,a) \coloneqq \begin{cases} \frac{1}{N_h(s,a)} \sum_{i=1}^N \mathbb{1}\big\{(h_i,s_i,a_i,s_i') = (h,s,a,s')\big\}, & \text{if } N_h(s,a) > 0 \\ 0, & \text{else} \end{cases} \tag{7.8}$$

for any $(h,s,a,s') \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Here, $N_h(s,a)$ denotes the total number of sample transitions from (s,a) at step h as

$$N_h(s,a) \coloneqq \sum_{i=1}^N \mathbb{1}\big\{(h_i,s_i,a_i) = (h,s,a)\big\}. \tag{7.9}$$

While it is possible to directly break down the history dataset $\mathcal{D}$ into sample transitions, unfortunately, the sample transitions from the same episode are not independent, significantly hurdling the analysis. To alleviate this, Chapter 5.1.3 proposed a simple two-fold subsampling scheme to preprocess the history dataset $\mathcal{D}$ and decouple the statistical dependency, resulting into a distributionally equivalent dataset $\mathcal{D}_0$ with independent samples. We have the following lemma paraphrased from Lemma 13.

Lemma 21. *With probability at least* $1-8\delta$*, the output dataset from the two-fold subsampling scheme in Li et al. (2022a) is distributionally equivalent to* $\mathcal{D}_0$*, where* $\{N_h(s,a)\}$ *are independent and obey*

$$N_h(s,a) \geq \frac{K d_h^{\mathsf{b},P^0}(s,a)}{8} - 5\sqrt{K d_h^{\mathsf{b},P^0}(s,a)\log\frac{KH}{\delta}}. \tag{7.10}$$

for all $(h,s,a) \in [H]\times\mathcal{S}\times\mathcal{A}$.

By invoking the two-fold sampling trick from Li et al. (2022a), it is sufficient to treat the dataset $\mathcal{D}_0$ with independent samples onwards with Lemma 21 in place. Armed with the estimate $\widehat{P}^0$ of the nominal transition kernel P^0, we are positioned to introduce our algorithm DRVI-LCB, summarized in Algorithm 13.

Distributionally robust value iteration. Before proceeding, let us recall the update rule of the classical distributionally robust value iteration (DRVI), which serves as the basis of our algorithmic development. Given an estimate of the nominal MDP $\widehat{P}^0$ and the radius σ of the uncertainty set, DRVI updates the robust value functions according to

$$\widehat{Q}_h(s,a) = r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1}, \qquad \text{and} \qquad \widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a), \tag{7.11}$$

which works backwards from $h=H$ to $h=1$, with the terminal condition $\widehat{Q}_{H+1}=0$. Due to strong duality (Hu and Hong, 2013), the update rule of the robust Q-functions in (7.11) can be equivalently reformulated in its dual form as

$$\widehat{Q}_h(s,a) = r_h(s,a) + \sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-\widehat{V}_{h+1}}{\lambda}\right)\right) - \lambda\sigma\right\}, \tag{7.12}$$

which can be solved efficiently (Iyengar, 2005; Panaganti and Kalathil, 2022; Yang et al., 2022).

Our algorithm DRVI-LCB. Motivated by the principle of pessimism in standard offline RL (Jin et al., 2021; Li et al., 2022a; Rashidinejad et al., 2021; Xie et al., 2021b), we propose to perform a pessimistic variant of DRVI, where the update rule of DRVI-LCB at step h is modified as

$$\widehat{Q}_h(s,a) = \max\left\{r_h(s,a) + \sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-\widehat{V}_{h+1}}{\lambda}\right)\right) - \lambda\sigma\right\} - b_h\big(s,a\big),\, 0\right\}. \tag{7.13}$$

Here, the robust Q-function estimate is adjusted by subtracting a carefully designed data-driven penalty term $b_h(s,a)$ that measures the uncertainty of the value estimates. Specifically, for some

Algorithm 13: Robust value iteration with LCB (DRVI-LCB) for robust offline RL.

input: a dataset $\mathcal{D}_0$; reward function r; uncertainty level σ.
initialization: $\widehat{Q}_{H+1} = 0$, $\widehat{V}_{H+1} = 0$.
for $h = H, \cdots, 1$ **do**
  Compute the empirical nominal transition kernel $\widehat{P}_h^0$ according to (7.8);
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    Compute the penalty term $b_h(s,a)$ according to (7.14);
    Set $\widehat{Q}_h(s,a)$ according to (7.13);
  for $s \in \mathcal{S}$ **do**
    Set $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) = \arg\max_a \widehat{Q}_h(s,a)$;
output: $\widehat{\pi} = \{\widehat{\pi}_h\}_{1\le h\le H}$.

$\delta \in (0,1)$ and any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, the penalty term $b_h(s,a)$ is defined as

$$b_h(s,a) = \begin{cases} \min\left\{ c_{\mathrm{b}} \frac{H}{\sigma} \sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},h}(s,a) N_h(s,a)}} \ , \ H \right\} & \text{if } N_h(s,a) > 0, \\ H & \text{otherwise,} \end{cases} \tag{7.14}$$

where c_{b} is some universal constant, and

$$\widehat{P}_{\mathsf{min},h}(s,a) \coloneqq \min_{s'} \left\{ \widehat{P}_h^0(s' \,|\, s,a) : \ \widehat{P}_h^0(s' \,|\, s,a) > 0 \right\}. \tag{7.15}$$

The penalty term is novel and different from the one used in standard (no-robust) offline RL (Jin et al., 2021; Li et al., 2022a; Rashidinejad et al., 2021; Shi et al., 2022; Xie et al., 2021b), by taking into consideration the unique problem structure pertaining to robust MDPs. In particular, it tightly upper bounds the statistical uncertainty which carries a non-linear and implicit dependency w.r.t. the estimated nominal transition kernel induced by the uncertainty set $\mathcal{U}(P^0)$, addressing unique challenges not present for the standard MDP case.

7.1.3 Theoretical guarantees

Before stating the main theorems, let us first introduce several important metrics.

- $P^\star_{\mathsf{min}}$, which only depends on the state-action pairs covered by the optimal robust policy $\pi^\star$ under the nominal model P^0:

$$P^\star_{\mathsf{min}} \coloneqq \min_{h,s,s'} \left\{ P_h^0\left(s' | s, \pi_h^\star(s)\right) : \ P_h^0\left(s' | s, \pi_h^\star(s)\right) > 0 \right\}. \tag{7.16}$$

  In words, $P^\star_{\mathsf{min}}$ is the smallest positive state transition probability of the optimal robust policy $\pi^\star$ under the nominal kernel P^0.

- Similarly, we introduce $P^{\mathsf{b}}_{\mathsf{min}}$ which only depends on the state-action pairs covered by the behavior policy π^{b} under the nominal model P^0:

$$P^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{h,s,a,s'} \left\{ P^0_h\left(s'|s,a\right) : \; d^{\mathsf{b},P^0}_h(s,a) > 0,\; P^0_h\left(s' \,|\, s,a\right) > 0 \right\}. \tag{7.17}$$

  In words, $P^{\mathsf{b}}_{\mathsf{min}}$ is the smallest positive state transition probability of the behavior policy π^{b} under the nominal kernel P^0.

- Finally, let $d^{\mathsf{b}}_{\mathsf{min}}$ denote the smallest positive state-action occupancy distribution of the behavior policy π^{b} under the nominal model P^0:

$$d^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{h,s,a} \left\{ d^{\mathsf{b},P^0}_h(s,a) : \; d^{\mathsf{b},P^0}_h(s,a) > 0 \right\}. \tag{7.18}$$

We are now positioned to present the performance guarantees of DRVI-LCB for robust offline RL.

Theorem 14. *Given an uncertainty level $\sigma > 0$, suppose that the penalty terms in Algorithm 13 are chosen as (7.14) for sufficiently large c_{b}. With probability at least $1-\delta$, the output $\widehat{\pi}$ of Algorithm 13 obeys*

$$V_1^{\star,\sigma}(\rho) - V_1^{\widehat{\pi},\sigma}(\rho) \le c_0 \frac{H^2}{\sigma} \sqrt{\frac{SC^{\star}_{\mathsf{rob}} \log^2(KHS/\delta)}{P^{\star}_{\mathsf{min}} K}}, \tag{7.19}$$

as long as the number of episodes K satisfies

$$K \ge \frac{c_1 \log(KHS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}} P^{\mathsf{b}}_{\mathsf{min}}}, \tag{7.20}$$

where c_0 and c_1 are some sufficiently large universal constants.

Our theorem is the first to characterize the sample complexities of robust offline RL under *partial coverage*, to the best of our knowledge (cf. Table 1.6). Theorem 14 shows that DRVI-LCB finds an ε-optimal robust policy as soon as the sample size $T = KH$ is above the order of

$$\underbrace{\frac{SC^{\star}_{\mathsf{rob}} H^5}{P^{\star}_{\mathsf{min}} \sigma^2 \varepsilon^2}}_{\epsilon\text{-dependent}} + \underbrace{\frac{H}{d^{\mathsf{b}}_{\mathsf{min}} P^{\mathsf{b}}_{\mathsf{min}}}}_{\text{burn-in cost}}, \tag{7.21}$$

up to some logarithmic factor, where the burn-in cost is independent of the accuracy level ε. For sufficiently small accuracy level ε, this results in a sample complexity of

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{rob}} H^5}{P^{\star}_{\mathsf{min}} \sigma^2 \varepsilon^2}\right). \tag{7.22}$$

Our theorem suggests that the sample efficiency of robust offline RL critically depends on the problem structure of the given RMDP (i.e. coverage of the optimal robust policy $\pi^\star$ as measured by $P^\star_{\mathsf{min}}$) as well as the quality of the history dataset (as measured by $C^\star_{\mathsf{rob}}$). Given that $C^\star_{\mathsf{rob}}$ can be as small as on the order of $1/S$, the sample complexity requirement can exhibit a much weaker dependency with the size of the state space S.

On the flip side, to assess the optimality of Theorem 14, we develop an information-theoretic lower bound for robust offline RL as provided in the following theorem.

Theorem 15. *For any $(H, S, C, P^\star_{\mathsf{min}}, \sigma, \varepsilon)$ obeying $H \geq 2e^8$, $C \geq 4/S$, $P^\star_{\mathsf{min}} \in (0, \frac{1}{H}]$, $\log(1/P^\star_{\mathsf{min}}) - 6 \leq \sigma \leq \log(1/P^\star_{\mathsf{min}}) - 5$, and $\varepsilon \leq \frac{H}{384e^6 \log(1/P^\star_{\mathsf{min}})}$, we can construct two finite-horizon RMDPs $\mathcal{M}_0, \mathcal{M}_1$, an initial state distribution ρ, and a batch dataset with K independent sample trajectories each with length H satisfying $2C \leq C^\star_{\mathsf{rob}} \leq 4C$, such that*

$$\inf_{\widehat{\pi}} \max\left\{\mathbb{P}_0\big(V_1^{\star,\sigma}(\rho) - V_1^{\widehat{\pi},\sigma}(\rho) > \varepsilon\big), \mathbb{P}_1\big(V_1^{\star,\sigma}(\rho) - V_1^{\widehat{\pi},\sigma}(\rho) > \varepsilon\big)\right\} \geq \frac{1}{8},$$

provided that

$$T = KH \leq \frac{c_1 S C^\star_{\mathsf{rob}} H^3}{P^\star_{\mathsf{min}} \sigma^2 \varepsilon^2}.$$

Here, $c_1 > 0$ is some universal constant, the infimum is taken over all estimators $\widehat{\pi}$, and $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) denotes the probability when the RMDP is $\mathcal{M}_0$ (resp. $\mathcal{M}_1$).

Theorem 15 shows that no algorithm can succeed in finding an ε-optimal robust policy when the sample complexity falls below the order of

$$\Omega\left(\frac{S C^\star_{\mathsf{rob}} H^3}{P^\star_{\mathsf{min}} \sigma^2 \varepsilon^2}\right),$$

which confirms the near-optimality of DRVI-LCB up to a factor of H^2 ignoring logarithmic factors. Therefore, DRVI-LCB is the first provable algorithm for robust offline RL with a near-optimal sample complexity without requiring the stringent full coverage assumption.

7.2 Algorithm and theory: discounted infinite-horizon RMDPs

Now, we turn to the studies of robust offline RL for discounted infinite-horizon MDPs.

7.2.1 Problem formulation and assumptions

Similar to the finite-horizon setting, using the distance measured in terms of the KL divergence introduced in (7.1), given an uncertainty level $\sigma > 0$, the uncertainty set around P^0 is specified as

$$\mathcal{U}^\sigma_{\mathsf{KL}}(P^0) \coloneqq \mathcal{U}^\sigma(P^0) \coloneqq \otimes \mathcal{U}^\sigma(P^0_{s,a}), \qquad \mathcal{U}^\sigma(P^0_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{s,a} \,\|\, P^0_{s,a}\right) \leq \sigma\right\}, \quad (7.23)$$

where we recall that a vector of the transition kernel P or P^0 at (s,a) is denoted respectively in (2.24).

Data sampling mechanism: batch data. Suppose that we observe a batch/history dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{1\le i\le N}$ consisting of N sample transitions. These transitions are independently generated, where the state-action pair is drawn from some behavior distribution $d^{\mathsf{b}} \in \Delta(\mathcal{S}\times\mathcal{A})$, followed by a next state drawn over the nominal transition kernel P^0, i.e.,

$$(s_i, a_i) \overset{\text{i.i.d.}}{\sim} d^{\mathsf{b}} \quad \text{and} \quad s_i' \overset{\text{i.i.d.}}{\sim} P^0(\cdot \,|\, s_i, a_i), \qquad 1 \le i \le N. \tag{7.24}$$

Similar to Definition 5, we design the following *robust single-policy clipped concentrability* assumption tailored for infinite-horizon RMDPs to characterize the quality of the history dataset.

Definition 6 (Robust single-policy clipped concentrability for infinite-horizon MDPs)**.** The behavior policy of the history dataset $\mathcal{D}$ satisfies

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b},P^0}(s,a)} \le C^\star_{\mathsf{rob}} \tag{7.25}$$

for some finite quantity $C^\star_{\mathsf{rob}} \in \left[\frac{1}{S}, \infty\right)$. Following the convention $0/0 = 0$, we denote $C^\star_{\mathsf{rob}}$ to be the smallest quantity satisfying (7.25), and refer to it as the robust single-policy clipped concentrability coefficient.

Remark 7. Similar to Remark 6, we can bound $C^\star_{\mathsf{rob}} \le A$ when the batch dataset is generated using a simulator (Panaganti and Kalathil, 2022; Yang et al., 2022). By combining this bound of $C^\star_{\mathsf{rob}}$ with the theoretical guarantees developed momentarily in Theorem 16, we obtain the comparison in Table 1.6.

Armed with these, we are ready to introduce the goal in the infinite-horizon setting. Given the history dataset $\mathcal{D}$, for some target accuracy $\varepsilon > 0$, we aim to find a near-optimal robust policy $\widehat{\pi}$, which satisfies

$$V^{\widehat{\pi},\sigma}(\varphi) \ge V^{\star,\sigma}(\varphi) - \varepsilon \tag{7.26}$$

in a sample-efficient manner for some initial state distribution φ.

7.2.2 DRVI-LCB for discounted infinite-horizon RMDPs

Building an empirical nominal MDP Recalling that we have N independent samples in the dataset $\mathcal{D} = \{(s_i, a_i, s_i')\}_{1\le i\le N}$. First, we denote $N(s,a)$ as the total number of sample transitions

Algorithm 14: Robust value iteration with LCB (DRVI-LCB) for infinite-horizon RMDPs.

input: a dataset $\mathcal{D}$; reward function r; uncertainty level σ; number of iterations M.
initialization: $\widehat{Q}_0(s,a) = 0$, $\widehat{V}_0(s) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
Compute the empirical nominal transition kernel $\widehat{P}^0$ according to (7.28);
Compute the penalty term $b(s,a)$ according to (7.32);
for $m = 1, 2, \cdots, M$ **do**
    for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
        Set $\widehat{Q}_m(s,a)$ according to (7.35);
    for $s \in \mathcal{S}$ **do**
        Set $\widehat{V}_m(s) = \max_a \widehat{Q}_m(s,a)$;
output: $\widehat{\pi}$ s.t. $\widehat{\pi}(s) = \arg\max_a \widehat{Q}_M(s,a)$ for all $s \in \mathcal{S}$.

from any state-action pair (s,a) as

$$N(s,a) \coloneqq \sum_{i=1}^{N} \mathbb{1}\big\{(s_i,a_i) = (s,a)\big\}. \tag{7.27}$$

Armed with $N(s,a)$, we construct the empirical estimate $\widehat{P}^0$ of the nominal kernel P^0 by the visiting frequencies of state-action pairs as follows:

$$\widehat{P}^0(s' \,|\, s,a) \coloneqq \begin{cases} \frac{1}{N(s,a)} \sum_{i=1}^{N} \mathbb{1}\big\{(s_i,a_i,s_i') = (s,a,s')\big\}, & \text{if } N(s,a) > 0 \\ 0, & \text{else} \end{cases} \tag{7.28}$$

for any $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

With the estimate $\widehat{P}^0$ of the nominal transition kernel P^0 in hand, we are positioned to introduce our algorithm DRVI-LCB for infinite-horizon RMDPs, which bears some similarity with the finite-horizon version (cf. Algorithm 13), by taking the uncertainties of the value estimates into consideration throughout the value iterations. The procedure is summarized in Algorithm 14.

The pessimistic robust Bellman operator. At the core of DRVI-LCB is a pessimistic variant of the classical robust Bellman operator in the infinite-horizon setting (Iyengar, 2005; Nilim and El Ghaoui, 2005; Zhou et al., 2021), denoted as $\mathcal{T}^\sigma(\cdot) : \mathbb{R}^{SA} \to \mathbb{R}^{SA}$, which we recall as follows:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \mathcal{T}^\sigma(Q)(s,a) \coloneqq r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}V, \quad \text{with} \quad V(s) \coloneqq \max_a Q(s,a). \tag{7.29}$$

Encouragingly, the robust Bellman operator shares the nice γ-contraction property of the standard Bellman operator, ensuring fast convergence of robust value iteration by applying the robust Bellman

operator (7.29) recursively. In the robust offline setting, instead of recursing using the population robust Bellman operator, we need to construct a pessimistic variant of the robust Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ w.r.t. the empirical nominal kernel $\widehat{P}^0$ as follows:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q)(s,a) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big),\, 0 \right\}, \tag{7.30}$$

where $b(s,a)$ denotes the penalty term that measures the data-dependent uncertainty of the value estimates.

To specify the tailored penalty term $b(s,a)$ in (7.30), we first introduce an additional term

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{P}_{\mathsf{min}}(s,a) \coloneqq \min_{s'} \Big\{ \widehat{P}^0(s' \,|\, s,a):\ \widehat{P}^0(s' \,|\, s,a) > 0 \Big\}, \tag{7.31}$$

which in words represents the smallest positive transition probability of the estimated nominal kernel $\widehat{P}^0(s' \,|\, s,a)$. Then for some $\delta \in (0,1)$, some universal constant $c_{\mathrm{b}} > 0$, $b(s,a)$ is defined as

$$b(s,a) = \begin{cases} \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3 S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a) N(s,a)}} + \frac{4}{\sigma N(1-\gamma)}\ ,\ \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N} & \text{if } N(s,a) > 0, \\ \frac{1}{1-\gamma} + \frac{2}{\sigma N} & \text{otherwise.} \end{cases} \tag{7.32}$$

As shall be illuminated, our proposed pessimistic robust Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (7.30)) plays an important role in DRVI-LCB. Encouragingly, despite the additional data-driven penalty term $b(s,a)$, it still enjoys the celebrated γ-contractive property, which greatly facilitates the analysis. Before continuing, we summarize the γ-contraction property below, whose proof is postponed to Appendix E.3.1.

Lemma 22 (γ-Contraction)**.** *For any $\gamma \in [0,1)$, the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (7.30)) is a γ-contraction w.r.t. $\|\cdot\|_{\infty}$. Namely, for any $Q_1, Q_2 \in \mathbb{R}^{SA}$ s.t. $Q_1(s,a), Q_2(s,a) \in \big[0, \frac{1}{1-\gamma}\big]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, one has*

$$\left\| \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q_1) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q_2) \right\|_{\infty} \le \gamma \left\| Q_1 - Q_2 \right\|_{\infty}. \tag{7.33}$$

Additionally, there exists a unique fixed point $\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}$ of the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ obeying $0 \le \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}(s,a) \le \frac{1}{1-\gamma}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Our algorithm DRVI-LCB for infinite-horizon robust offline RL. Armed with the γ-contraction property of the pessimistic robust Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$, we are positioned to introduce DRVI-LCB for infinite-horizon RMDPs, summarized in Algorithm 14. Specifically, DRVI-LCB can be seen as a value iteration algorithm w.r.t. $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (7.30)), whose update rule at the m-th iteration can be

formulated as

$$\widehat{Q}_m(s,a) = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{m-1})(s,a) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}_{m-1} - b\big(s,a\big),\, 0 \right\}, \tag{7.34}$$

and $\widehat{V}_m(s) = \max_a \widehat{Q}_m(s,a)$ for all $m = 1, 2, \cdots, M$. In view of strong duality (Hu and Hong, 2013), the above convex problem can be translated into a dual formulation, leading to the following equivalent update rule:

$$\widehat{Q}_m(s,a) = \max\left\{ r(s,a) + \sup_{\lambda\geq 0}\left\{ -\lambda \log\left(\widehat{P}^0_{s,a} \cdot \exp\left(\frac{-\widehat{V}_{m-1}}{\lambda} \right) \right) - \lambda\sigma \right\} - b\big(s,a\big),\, 0 \right\}, \tag{7.35}$$

which can be solved efficiently (Iyengar, 2005; Panaganti and Kalathil, 2022; Yang et al., 2022) as a one-dimensional optimization problem.

To finish the description, we initialize the estimates of Q-function ($\widehat{Q}_0$) and value function ($\widehat{V}_0$) to be zero and output the greedy policy of the final Q-estimates ($\widehat{Q}_M$) as the final policy $\widehat{\pi}$, namely,

$$\widehat{\pi}(s) = \arg\max_a \widehat{Q}_M(s,a) \quad \text{for all } s \in \mathcal{S}. \tag{7.36}$$

It turns out that the iterates $\{\widehat{Q}_m\}_{m\geq 0}$ of DRVI-LCB converge linearly to the fixed point $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ owing to the nice γ-contraction property. This fact is summarized in the following lemma.

Lemma 23. *Let $\widehat{Q}_0 = 0$. The iterates of Algorithm 14 obey*

$$\forall m \geq 0: \quad \widehat{Q}_m \leq \widehat{Q}^{\star,\sigma}_{\mathsf{pe}} \quad \textit{and} \quad \big\|\widehat{Q}_m - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\big\|_\infty \leq \frac{\gamma^m}{1-\gamma}. \tag{7.37}$$

7.2.3 Theoretical guarantees

Before introducing the main theorems, we first define several essential metrics.

- $d^{\mathsf{b}}_{\mathsf{min}}$: the smallest positive entry of the distribution d^{b,P^0}, i.e.,

$$d^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{s,a}\left\{ d^{\mathsf{b},P^0}(s,a):\ d^{\mathsf{b},P^0}(s,a) > 0 \right\}. \tag{7.38}$$

- $P^{\mathsf{b}}_{\mathsf{min}}$: the smallest positive state transition probability under the nominal kernel P^0 in the region covered by dataset $\mathcal{D}$, i.e.,

$$P^{\mathsf{b}}_{\mathsf{min}} \coloneqq \min_{s,a,s'}\left\{ P^0\left(s' \,|\, s,a\right):\ d^{\mathsf{b},P^0}(s,a) > 0,\ P^0\left(s' \,|\, s,a\right) > 0 \right\}. \tag{7.39}$$

  Note that $P^{\mathsf{b}}_{\mathsf{min}}$ is determined only by the state-action pairs covered by the batch dataset $\mathcal{D}$.

- $P^{\star}_{\mathsf{min}}$: the smallest positive state transition probability of the optimal robust policy $\pi^{\star}$ under the nominal kernel P^0, namely

$$P^{\star}_{\mathsf{min}} \coloneqq \min_{s,s'} \Big\{ P^0\big(s' \,|\, s, \pi^{\star}(s)\big) : \; P^0\big(s' \,|\, s, \pi^{\star}(s)\big) > 0 \Big\}. \tag{7.40}$$

  We also note that $P^{\star}_{\mathsf{min}}$ is determined only by the state-action pairs covered by the optimal robust policy $\pi^{\star}$ under the nominal model P^0.

We are now positioned to introduce the sample complexity upper bound of DRVI-LCB, together with the minimax lower bound, for solving infinite-horizon RMDPs. First, we present the performance guarantees of DRVI-LCB for robust offline RL in the infinite-horizon case.

Theorem 16. *Let c_0 and c_1 be some sufficiently large universal constants. Given an uncertainty level $\sigma > 0$, suppose that the penalty terms in Algorithm 14 are chosen as (7.32) for sufficiently large c_{b}. With probability at least $1-\delta$, the output $\widehat{\pi}$ of Algorithm 14 obeys*

$$V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) \le \frac{c_0}{\sigma(1-\gamma)^2} \sqrt{\frac{SC^{\star}_{\mathsf{rob}} \log^2\left(\frac{(1+\sigma)N^3 S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}} N}}, \tag{7.41}$$

as long as the number of samples N satisfies

$$N \ge \frac{c_1 \log(NS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}} P^{\mathsf{b}}_{\mathsf{min}}}. \tag{7.42}$$

The result directly indicates that DRVI-LCB can finds an ε-optimal policy as long as the sample size in dataset $\mathcal{D}$ exceeds the order of (ignoring logarithmic factors)

$$\underbrace{\frac{SC^{\star}_{\mathsf{rob}}}{P^{\star}_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}}_{\varepsilon\text{-dependent}} + \underbrace{\frac{1}{d^{\mathsf{b}}_{\mathsf{min}} P^{\mathsf{b}}_{\mathsf{min}}}}_{\text{burn-in cost}}. \tag{7.43}$$

Note that the burn-in cost is independent with the accuracy level ε, which tells us that the sample complexity is no more than

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{rob}}}{P^{\star}_{\mathsf{min}}(1-\gamma)^4\sigma^2\varepsilon^2}\right) \tag{7.44}$$

as long as ε is small enough. The sample complexity of DRVI-LCB still dramatically outperforms prior works under full coverage, which has been compared in detail in Table 1.6. In particular, our sample complexity produces an exponential improvement over Panaganti and Kalathil (2022); Zhou et al. (2021) in terms of the dependency with the effective horizon $\frac{1}{1-\gamma}$, which is especially significant for long-horizon problems. Compared with Yang et al. (2022), our sample complexity is better by

at least a factor of $S/P^\star_{\mathsf{min}}$. To achieve the claimed bound, we resort to a delicate technique called the leave-one-out analysis (Agarwal et al., 2020b; Li et al., 2022a, 2023c), by carefully designing an auxiliary set of RMDPs to decouple the statistical dependency introduced across the iterates of pessimistic robust value iteration. This is the first time that the leave-one-out analysis is applied to understanding the sample efficiency of model-based robust RL algorithms, which is of potential independent interest to tighten the sample complexity of other robust RL problems.

To complement the upper bound, we develop an information-theoretic lower bound for robust offline RL as provided in the following theorem.

Theorem 17. *For any* $(S, P^\star_{\mathsf{min}}, C^\star_{\mathsf{rob}}, \gamma, \sigma, \varepsilon)$ *obeying* $\frac{1}{1-\gamma} \geq 2e^8$, $P^\star_{\mathsf{min}} \in \big(0, 1-\gamma\big]$, $S \geq \log\big(1/P^\star_{\mathsf{min}}\big)$, $C^\star_{\mathsf{rob}} \geq 8/S$, $\varepsilon \leq \frac{1}{384e^6(1-\gamma)\log\left(1/P^\star_{\mathsf{min}}\right)}$, *and* $\log\big(1/P^\star_{\mathsf{min}}\big) - 6 \leq \sigma \leq \log\big(1/P^\star_{\mathsf{min}}\big) - 5$, *we can construct two infinite-horizon RMDPs* $\mathcal{M}_0, \mathcal{M}_1$, *an initial state distribution* φ, *and a batch dataset with* N *independent samples, such that*

$$\inf_{\widehat{\pi}} \max\Big\{\mathbb{P}_0\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big),\, \mathbb{P}_1\big(V^{\star,\sigma}(\varphi) - V^{\widehat{\pi},\sigma}(\varphi) > \varepsilon\big)\Big\} \geq \frac{1}{8},$$

provided that

$$N \leq \frac{c_1 S C^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2}.$$

Here, $c_1 > 0$ *is some universal constant, the infimum is taken over all estimators* $\widehat{\pi}$, *and* $\mathbb{P}_0$ *(resp.* $\mathbb{P}_1$*) denotes the probability when the RMDP is* $\mathcal{M}_0$ *(resp.* $\mathcal{M}_1$*).*

The above theorem suggests that there exists some RMDP such that no algorithm can find an ε-optimal policy if the sample complexity is below the order of

$$\Omega\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\varepsilon^2}\right),$$

which directly confirms that DRVI-LCB is near-optimal up to a polynomial factor of the effective horizon length $\frac{1}{1-\gamma}$ (cf. (7.43)). To the best of our knowledge, DRVI-LCB is the first provable algorithm with near-optimal sample complexity for infinite-horizon robust offline RL. Moreover, the requirement imposed on the history dataset is also much weaker than prior literature on robust offline RL (Yang et al., 2022; Zhou et al., 2021), without the need of full coverage of the state-action space.

7.3 Numerical experiments

We conduct experiments on the gambler's problem (Sutton and Barto, 2018; Zhou et al., 2021) to evaluate the performance of the proposed algorithm DRVI-LCB, with comparisons to both the robust value iteration algorithm DRVI without pessimism (Panaganti and Kalathil, 2022). Our code

can be accessed at:

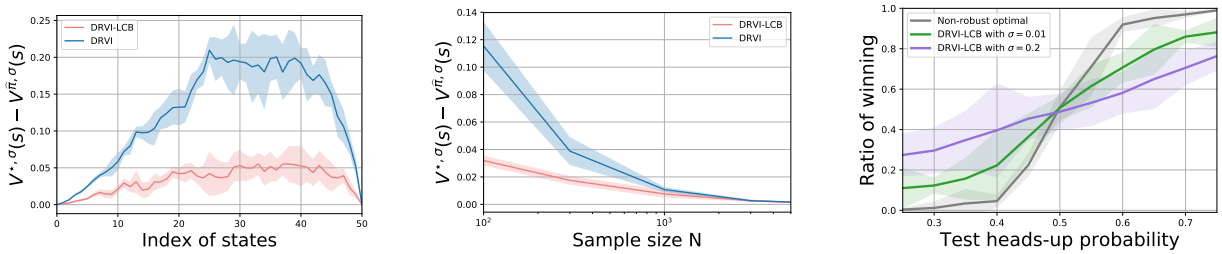
https://github.com/Laixishi/Robust-RL-with-KL-divergence.

Gambler's problem. In the gambler's game (Sutton and Barto, 2018; Zhou et al., 2021), a gambler bets on a sequence of coin flips, winning the stake with heads and losing with tails. Starting from some initial balance, the game ends when the gambler's balance either reaches 50 or 0, or the total number of bets H is hit. This problem can be formulated as an episodic finite-horizon MDP, with a state space $\mathcal{S} = \{0, 1, \cdots, 50\}$ and the associated possible actions $a \in \big\{0, 1, \cdots, \min\{s, 50 - s\}\big\}$ at state s. Here, we set the horizon length $H = 100$. Moreover, the parameter of the transition kernel, which is the probability of heads for the coin flip, is fixed as p_{head} and remains the same in all time steps $h \in [H]$. The reward is set as 1 when the state reaches $s = 50$ and 0 for all other cases. In addition, suppose the initial state (i.e., the gambler's initial balance) distribution ρ is taken uniformly at random within $\mathcal{S}$. Throughout the experiments, we utilize a history dataset with N samples per state-action pair and time step, which is generated from a nominal MDP with $p^0_{\mathsf{head}} = 0.6$.

Results and discussions. First, we evaluate the performance of the learned policy $\widehat{\pi}$ using our proposed method DRVI-LCB with comparison to robust value iteration (DRVI) without pessimism, where we fix the uncertainty level $\sigma = 0.1$ for learning the robust optimal policy. The experiments are repeated 10 times with the average and standard deviations reported. To begin with, Figure 7.1(a) plots the sub-optimality value gap $V_1^{\star,\sigma}(s) - V_1^{\widehat{\pi},\sigma}(s)$ for every $s \in \mathcal{S}$, when a sample size $N = 100$ is used to learn the robust policies. It is shown that DRVI-LCB outperform the baseline DRVI uniformly over the state space when the sample size is small, corroborating the benefit of pessimism in the sample-starved regime. Furthermore, Figure 7.1(b) shows the sub-optimality gap $V_1^{\star,\sigma}(\rho) - V_1^{\widehat{\pi},\sigma}(\rho)$ with varying sample sizes $n = 100, 300, 1000, 3000, 5000$, where the initial test distribution ρ is generated randomly.[1] While the performance of DRVI-LCB and DRVI both improves with the increase of the sample size, the proposed algorithm DRVI-LCB achieves much better performance with fewer samples.

Finally, to corroborate the benefit of distributional robustness, we evaluate the performance of the policy learned from $N = 1000$ samples using DRVI-LCB on perturbed environments with varying model parameters $p_{\mathsf{head}} \in [0.25, 0.75]$. We measure the practical performance based on the ratio of winning (i.e., reaching the state $s = 50$) calculated from 3000 episodes. Figure 7.1(c) illustrates the ratio of winning against the test probability of heads for the policies learned from DRVI-LCB with $\sigma = 0.01$ and $\sigma = 0.2$, which are benchmarked against the non-robust optimal policy of the nominal MDP using the exact model. It can be seen that the policies learned from DRVI-LCB deviate from

[1] The probability distribution vector $\rho \in \Delta(\mathcal{S})$ is generated as $\rho(s) = u_s / \sum_{s \in \mathcal{S}} u_s$, where u_s is drawn independently from a uniform distribution.



(a) Value gap versus states (b) Value gap versus sample size (c) winning rate versus p_{head}

Figure 7.1: The performance evaluation of the proposed algorithm DRVI-LCB, where it shows better sample efficiency than the baseline algorithm DRVI without pessimism, as well as better robustness in the learned policy compare to its non-robust counterpart.

the non-robust optimal policy as σ increases, which achieves better worst-case rates of winning across a wide range of perturbed environments. On the other end, while the non-robust policy maximizes the performance when the test environment is close to the history one used for training, its performance degenerates to be much worse than the robust policies when the probability of heads is mismatched significantly, especially when p_{head} drops below, say around, 0.5.

7.4 Discussions

To accommodate both model robustness and sample efficiency, in this chapter, we propose a distributionally robust model-based algorithm for offline RL with the principle of pessimism. We study the finite-sample complexity of the proposed algorithm DRVI-LCB, and establishes its near-optimality with a matching information-theoretic lower bound. Numerical experiments are provided to demonstrate the efficacy of the proposed algorithm. To the best our knowledge, this provides the first provably near-optimal robust offline RL algorithm that learns under model perturbation and partial coverage.

Chapter 8

Conclusion

This thesis breaks down the sample barriers of various RL problems, taking into consideration additional facets of scalability and robustness. Specifically, for online RL, which permits adaptive interactions with the environment, this thesis presents the first regret-optimal model-free RL algorithm with a small burn-in cost — an initial sampling burden needed for the algorithm to exhibit the desired performance — while maintaining its memory efficiency for scalability. In the context of offline RL, which relies solely on historical datasets, this thesis puts forward the first provable near-optimal model-free offline RL algorithm that doesn't require model estimation. In addition, it settles the sample complexity by establishing the minimax optimality of model-based offline RL algorithms without burn-in cost. Lastly, for a robust variant of standard RL — distributionally robust RL, this thesis reveals a surprising fact: the introduction of additional distributional robustness into the learned policy doesn't inherently increase or decrease the sample requirements compared to standard RL; it largely depends on the defined uncertainty set. This thesis closes by providing the first provable near-optimal algorithm for offline robust RL that can learn under simultaneous model uncertainty and limited historical datasets.

The findings of this thesis naturally suggest numerous potential extensions and future research directions. The thesis concludes by outlining a selection of these possibilities.

- *Improved analysis.* Some results established in this thesis can be further improved. For instance, for online RL, while the proposed algorithm in Chapter 3 provably enables minimal burn-in cost in terms of the dependency on S and A, our current theory falls short of delivering optimal horizon dependency of the burn-in cost. More specifically, even though our burn-in cost improves upon the state-of-the-art theory for sample-optimal model-free algorithms by a factor of at least $S^5A^3H^{18}$ (see Zhang et al. (2020c)), the way we cope with the dependency on H remains inadequate. This calls for more refined analysis tools to optimize the horizon dependency. For offline RL, the ε-range for LCB-Q-Advantage in Chapter 4 to attain sample optimality remains somewhat limited (i.e., $\varepsilon \in (0, 1/H]$)). Further investigation into whether model-free algorithms can accommodate a broader ε-range without compromising sample efficiency is called for.

- *Further investigation in tabular settings.* For robust RL, it is likely that our current analysis framework in Chapter 6 can be extended to tackle finite-horizon RMDPs with a generative model, which would help complete our understanding for the tabular cases. Moreover, Our

work in Chapter 6 raises an interesting question concerning how the geometry of the uncertainty sets intervenes the sample complexity. Hence, characterizing the tight sample complexity for RMDPs under a more general family of uncertainty sets — such as using ℓ_p distance or f-divergence, as well as s-rectangular sets — would be highly desirable. In addition, in light of the results in Chapter 7, it is also promising to design provably efficient model-free algorithms for robust offline RL with partial coverage.

- *Extensions to function approximation settings.* Admittedly, even though we are now able to settle the sample size dependency on the state-action space, the size of SA might remain prohibitively large in many modern RL applications. As a result, parsimonious function representation/approximation of the underlying MDP is needed in order to further reduce the sample complexity. Moving beyond tabular settings, it would be of great interest to extend our analysis to accommodate value-based RL in more general scenarios; examples include MDPs with low-complexity linear representations, realizable MDPs, and RL involving multiple agents (Jin et al., 2020; Li et al., 2021; Nguyen-Tang et al., 2021).

- *Investigation of RL in real-world applications.* In addition to the theoretical investigations of RL problems included in this thesis, my Ph.D. research also focused on seeking practical solutions for diverse applications in real-world settings (Chen et al., 2021b; Ding et al., 2023; Huang et al., 2022; Low et al., 2022; Sang et al., 2018; Shi and Chi, 2021; Shi et al., 2023a, 2021a,b, 2019, 2020; Wang et al., 2023b). Looking ahead, there is great interest in exploring promising application scenarios of RL in both fundamental science and daily life including but not limited to facilitating protein discovery in biology, enhancing the simulation of fluid dynamics, and revolutionizing recommendation systems for social media.

Appendix A

Proofs for Chapter 3

A.1 Freedman's inequality

A.1.1 A user-friendly version of Freedman's inequality

Due to the Markovian structure of the problem, our analysis relies heavily on the celebrated Freedman's inequality (Freedman, 1975; Tropp, 2011), which extends the Bernstein's inequality to accommodate martingales. For ease of reference, we state below a user-friendly version of Freedman's inequality as provided in Li et al. (2023a, Section C).

Theorem 18 (Freedman's inequality)**.** *Consider a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, and let $\mathbb{E}_k$ stand for the expectation conditioned on $\mathcal{F}_k$. Suppose that $Y_n = \sum_{k=1}^{n} X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \le R \qquad \text{and} \qquad \mathbb{E}_{k-1}\big[X_k\big] = 0 \qquad \text{for all } k \ge 1$$

for some quantity $R < \infty$. We also define

$$W_n := \sum_{k=1}^{n} \mathbb{E}_{k-1}\left[X_k^2\right].$$

In addition, suppose that $W_n \le \sigma^2$ holds deterministically for some given quantity $\sigma^2 < \infty$. Then for any positive integer $m \ge 1$, with probability at least $1-\delta$ one has

$$|Y_n| \le \sqrt{8 \max\left\{W_n, \frac{\sigma^2}{2^m}\right\} \log\frac{2m}{\delta}} + \frac{4}{3} R \log\frac{2m}{\delta}. \tag{A.1}$$

A.1.2 Application of Freedman's inequality

We now develop several immediate consequences of Freedman's inequality, which lend themselves well to our context. Before proceeding, we recall that $N_h^i(s,a)$ denotes the number of times that the state-action pair (s,a) has been visited at step h by the end of the i-th episode, and $k_h^n(s,a)$ stands for the episode index when (s,a) is visited at step h for the n-th time (see Appendix 3.3.2).

Our first result is concerned with a martingale concentration bound as follows.

Lemma 24. *Let* $\{W_h^i \in \mathbb{R}^S \mid 1 \le i \le K, 1 \le h \le H+1\}$ *and* $\{u_h^i(s,a,N) \in \mathbb{R} \mid 1 \le i \le K, 1 \le h \le H+1\}$ *be a collections of vectors and scalars, respectively, and suppose that they obey the following properties:*

- W_h^i *is fully determined by the samples collected up to the end of the* $(h-1)$*-th step of the* i*-th episode;*
- $\|W_h^i\|_\infty \le C_{\mathrm{w}}$*;*
- $u_h^i(s,a,N)$ *is fully determined by the samples collected up to the end of the* $(h-1)$*-th step of the* i*-th episode, and a given positive integer* $N \in [K]$*;*
- $0 \le u_h^i(s,a,N) \le C_{\mathrm{u}}$*;*
- $0 \le \sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N) \le 2$.

In addition, consider the following sequence

$$X_i(s,a,h,N) \coloneqq u_h^i(s,a,N)\big(P_h^i - P_{h,s,a}\big)W_{h+1}^i \mathbb{1}\left\{(s_h^i, a_h^i) = (s,a)\right\}, \qquad 1 \le i \le K, \tag{A.2}$$

with P_h^i *defined in* (3.15)*. Consider any* $\delta \in (0,1)$*. Then with probability at least* $1-\delta$*,*

$$\begin{aligned}&\left|\sum_{i=1}^{k} X_i(s,a,h,N)\right| \\ &\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N)\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta}\end{aligned} \tag{A.3}$$

holds simultaneously for all $(k,h,s,a,N) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A} \times [K]$.

Proof. For the sake of notational convenience, we shall abbreviate $X_i(s,a,h,N)$ as X_i throughout the proof of this lemma, as long as it is clear from the context. The plan is to apply Freedman's inequality (cf. Theorem 18) to control the term $\sum_{i=1}^k X_i$ of interest.

Consider any given $(k,h,s,a,N) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A} \times [K]$. It can be easily verified that

$$\mathbb{E}_{i-1}[X_i] = 0,$$

where $\mathbb{E}_{i-1}$ denotes the expectation conditioned on everything happening up to the end of the $(h-1)$-th step of the i-th episode. Additionally, we make note of the following crude bound:

$$|X_i| \le u_h^i(s,a,N)\left|\big(P_h^i - P_{h,s,a}\big)W_{h+1}^i\right|$$

$$\leq u_h^i(s,a,N)\Big(\big\|P_h^i\big\|_1+\big\|P_{h,s,a}\big\|_1\Big)\big\|W_{h+1}^i\big\|_\infty\leq 2C_{\mathrm{w}}C_{\mathrm{u}}, \tag{A.4}$$

which results from the assumptions $\|W_{h+1}^i\|_\infty\leq C_{\mathrm{w}}$, $0\leq u_h^i(s,a,N)\leq C_{\mathrm{u}}$ as well as the basic facts $\big\|P_h^i\big\|_1=\big\|P_{h,s,a}\big\|_1=1$. To continue, recalling the definition of the variance parameter in (B.51), we obtain

$$\begin{aligned}
\sum_{i=1}^k\mathbb{E}_{i-1}\left[\left|X_i\right|^2\right]&=\sum_{i=1}^k\big(u_h^i(s,a,N)\big)^2\mathbb{1}\left\{(s_h^i,a_h^i)=(s,a)\right\}\mathbb{E}_{i-1}\left[\left|(P_h^i-P_{h,s,a})W_{h+1}^i\right|^2\right]\\
&=\sum_{n=1}^{N_h^k(s,a)}\big(u_h^{k_h^n(s,a)}(s,a,N)\big)^2\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big)\\
&\leq C_{\mathrm{u}}\left(\sum_{n=1}^{N_h^k(s,a)}u_h^{k_h^n(s,a)}(s,a,N)\right)\big\|W_{h+1}^{k_h^n(s,a)}\big\|_\infty^2\\
&\leq 2C_{\mathrm{u}}C_{\mathrm{w}}^2,
\end{aligned}\tag{A.5}$$

where the inequalities hold true due to the assumptions $\|W_h^i\|_\infty\leq C_{\mathrm{w}}$, $0\leq u_h^i(s,a,N)\leq C_{\mathrm{u}}$, and $0\leq\sum_{n=1}^{N_h^k(s,a)}u_h^{k_h^n(s,a)}(s,a,N)\leq 1$.

With (A.4) and (A.5) in place, we can invoke Theorem 18 (with $m=\lceil\log_2 N\rceil$) and take the union bound over all $(k,h,s,a,N)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$ to show that: with probability at least $1-\delta$,

$$\begin{aligned}
\left|\sum_{i=1}^k X_i\right|&\lesssim\sqrt{\max\left\{C_{\mathrm{u}}\sum_{n=1}^{N_h^k(s,a)}u_h^{k_h^n(s,a)}(s,a,N)\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big),\frac{C_{\mathrm{u}}C_{\mathrm{w}}^2}{N}\right\}\log\frac{SAT^2\log N}{\delta}}\\
&\quad+C_{\mathrm{u}}C_{\mathrm{w}}\log\frac{SAT^2\log N_h^k}{\delta}\\
&\lesssim\sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)}u_h^{k_h^n(s,a)}(s,a,N)\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big)}+\left(C_{\mathrm{u}}C_{\mathrm{w}}+\sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta}
\end{aligned}$$

holds simultaneously for all $(k,h,s,a,N)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$. □

The next result is concerned with martingale concentration bounds for another type of sequences of interest.

Lemma 25. *Let $\big\{N(s,a,h)\in[K]\mid(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]\big\}$ be a collection of positive integers, and let $\{c_h:0\leq c_h\leq e,h\in[H]\}$ be a collection of fixed and bounded universal constants. Moreover, let $\big\{W_h^i\in\mathbb{R}^S\mid 1\leq i\leq K,1\leq h\leq H+1\big\}$ and $\big\{u_h^i(s_h^i,a_h^i)\in\mathbb{R}\mid 1\leq i\leq K,1\leq h\leq H+1\big\}$ represent respectively a collection of random vectors and scalars, which obey the following properties.*

- W_h^i *is fully determined by the samples collected up to the end of the* $(h-1)$*-th step of the* i*-th episode;*
- $\|W_h^i\|_\infty \le C_{\mathrm{w}}$ *and* $W_h^i \ge 0$*;*
- $u_h^i(s_h^i, a_h^i)$ *is fully determined by the integer* $N(s_h^i, a_h^i, h)$ *and all samples collected up to the end of the* $(h-1)$*-th step of the* i*-th episode;*
- $0 \le u_h^i(s_h^i, a_h^i) \le C_{\mathrm{u}}$.

Consider any $\delta \in (0,1)$*, and introduce the following sequences*

$$X_{i,h} \coloneqq u_h^i(s_h^i, a_h^i)\big(P_h^i - P_{h,s_h^i,a_h^i}\big)W_{h+1}^i, \qquad 1 \le i \le K, 1 \le h \le H+1, \tag{A.6}$$

$$Y_{i,h} \coloneqq c_h\big(P_h^i - P_{h,s_h^i,a_h^i}\big)W_{h+1}^i, \qquad 1 \le i \le K, 1 \le h \le H+1. \tag{A.7}$$

Then with probability at least $1-\delta$*,*

$$\begin{aligned}
\left|\sum_{h=1}^{H}\sum_{i=1}^{K} X_{i,h}\right| &\lesssim \sqrt{C_{\mathrm{u}}^2 \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i})W_{h+1}^i\right|^2\right] \log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}\log\frac{T^{HSA}}{\delta} \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 C_{\mathrm{w}} \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right] \log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}\log\frac{T^{HSA}}{\delta} \\
\left|\sum_{h=1}^{H}\sum_{i=1}^{K} Y_{i,h}\right| &\lesssim \sqrt{TC_{\mathrm{w}}^2 \log\frac{1}{\delta}} + C_{\mathrm{w}}\log\frac{1}{\delta}
\end{aligned}$$

holds simultaneously for all possible collections $\{N(s,a,h) \in [K] \mid (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]\}$.

Proof. This lemma can be proved by Freedman's inequality (cf. Theorem 18).

- We start by controlling the first term of interest $\sum_{h=1}^{H}\sum_{i=1}^{K} X_{i,h}$. As can be easily seen, $a_h^i = \arg\max Q_h^i(s_h^i, a)$ is fully determined by what happens before step h of the i-th episode. Consider any given $\big\{N(s,a,h) \in [K] \mid (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]\big\}$. It is readily seen that

$$\mathbb{E}_{i,h-1}\left[X_i\right] = \mathbb{E}_{i,h-1}\left[u_h^i(s_h^i, a_h^i)\big(P_h^i - P_{h,s_h^i,a_h^i}\big)W_{h+1}^i\right] = 0,$$

  where $\mathbb{E}_{i,h-1}$ denotes the expectation conditioned on everything happening before step h of the i-th episode. In addition, we make note of the following crude bound:

$$\begin{aligned}
\left|X_{i,h}\right| &\le u_h^i(s_h^i, a_h^i)\left|\big(P_h^i - P_{h,s_h^i,a_h^i}\big)W_{h+1}^i\right| \\
&\le u_h^i(s_h^i, a_h^i)\Big(\big\|P_h^i\big\|_1 + \big\|P_{h,s_h^i,a_h^i}\big\|_1\Big)\big\|W_{h+1}^i\big\|_\infty \le 2C_{\mathrm{w}}C_{\mathrm{u}},
\end{aligned} \tag{A.8}$$

which arises from the assumptions $\|W_{h+1}^i\|_\infty \le C_{\mathrm{w}}$, $0 \le u_h^i(s,a,N) \le C_{\mathrm{u}}$ together with the basic facts $\left\|P_h^i\right\|_1 = \left\|P_{h,s_h^i,a_h^i}\right\|_1 = 1$. Additionally, we can calculate that

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[\left|X_{i,h}\right|^2\right] &= \sum_{h=1}^{H}\sum_{i=1}^{K}\left(u_h^i(s_h^i,a_h^i)\right)^2\mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i})W_{h+1}^i\right|^2\right] \\
&\overset{\text{(i)}}{\le} C_{\mathrm{u}}^2\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i})W_{h+1}^i\right|^2\right] \qquad \text{(A.9)} \\
&\le C_{\mathrm{u}}^2\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[\left|P_h^i W_{h+1}^i\right|^2\right] \\
&\overset{\text{(ii)}}{=} C_{\mathrm{u}}^2\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[P_h^i\left(W_{h+1}^i\right)^2\right] \\
&\overset{\text{(iii)}}{\le} C_{\mathrm{u}}^2\sum_{h=1}^{H}\sum_{i=1}^{K}\left\|W_{h+1}^i\right\|_\infty\mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right] \\
&\overset{\text{(iv)}}{\le} C_{\mathrm{u}}^2 C_{\mathrm{w}}\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right] \qquad \text{(A.10)} \\
&\le C_{\mathrm{u}}^2 C_{\mathrm{w}}\sum_{h=1}^{H}\sum_{i=1}^{K}\left\|W_{h+1}^i\right\|_\infty \overset{\text{(v)}}{\le} HKC_{\mathrm{u}}^2C_{\mathrm{w}}^2 = TC_{\mathrm{u}}^2C_{\mathrm{w}}^2. \qquad \text{(A.11)}
\end{aligned}$$

Here, (i) holds true due to the assumption $0 \le u_h^i(s_h^i,a_h^i) \le C_{\mathrm{u}}$, (ii) is valid since P_h^i only has one non-zero entry (cf. (3.15)), (iii) relies on the assumptions that W_h^i is non-negative, whereas (iv) and (v) follow since $\|W_h^i\|_\infty \le C_{\mathrm{w}}$,

With (A.8), (A.10) and (A.11) in mind, we can invoke Theorem 18 (with $m = \lceil \log_2 T \rceil$) and take the union bound over all possible collections $\left\{N(s,a,h) \in [K] \mid (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]\right\}$ — which has at most K^{HSA} possibilities — to show that: with probability at least $1-\delta$,

$$\begin{aligned}
\left|\sum_{h=1}^{H}\sum_{i=1}^{k}X_{i,h}\right| &\lesssim \sqrt{\max\left\{C_{\mathrm{u}}^2\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i})W_{h+1}^i\right|^2\right], \frac{TC_{\mathrm{u}}^2C_{\mathrm{w}}^2}{2^m}\right\}\log\frac{K^{HSA}\log T}{\delta}} \\
&\quad + C_{\mathrm{u}}C_{\mathrm{w}}\log\frac{K^{HSA}\log T}{\delta} \\
&\lesssim \sqrt{C_{\mathrm{u}}^2\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i})W_{h+1}^i\right|^2\right]\log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}\log\frac{T^{HSA}}{\delta} \\
&\lesssim \sqrt{C_{\mathrm{u}}^2C_{\mathrm{w}}\sum_{h=1}^{H}\sum_{i=1}^{K}\mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right]\log\frac{T^{HSA}}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}\log\frac{T^{HSA}}{\delta}
\end{aligned}$$

holds simultaneously for all $\big\{N(s,a,h) \in [K] \mid (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\big\}$.

- Then we turn to control the second term $\left|\sum_{h=1}^{H}\sum_{i=1}^{K} Y_{i,h}\right|$ of interest. Similar to $\left|\sum_{h=1}^{H}\sum_{i=1}^{K} X_{i,h}\right|$, we have

$$|Y_{i,h}| \le 2eC_{\mathrm{w}},$$
$$\sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[|Y_{i,h}|^2\right] \le e^2 T C_{\mathrm{w}}^2.$$

Invoke Theorem 18 (with $m = 1$) to arrive at

$$\left|\sum_{h=1}^{H}\sum_{i=1}^{K} Y_{i,h}\right| \lesssim \sqrt{TC_{\mathrm{w}}^2 \log\frac{1}{\delta}} + C_{\mathrm{w}} \log\frac{1}{\delta} \tag{A.12}$$

with probability at least $1-\delta$.

□

A.2 Proof of Lemma 1

First of all, the properties in (3.14b) follow directly from Jin et al. (2018, Lemma 4.1). Therefore, it suffices to establish the property in (3.14a), which forms the remainder of this sub-chapter.

When $N = 1$, the statement holds trivially since

$$\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \eta_1^1 = 1 \in [1,2].$$

Now suppose that $N \ge 2$. Making use of the basic relation $\eta_n^N = (1-\eta_N)\eta_n^{N-1}$ for all $n = 1, \cdots, N-1$, we observe the following identity:

$$\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \frac{\eta_N}{N^a} + (1-\eta_N)\sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a}. \tag{A.13}$$

We now prove the property in (3.14a) by induction. Suppose for the moment that the property holds for $N-1$, namely,

$$\frac{1}{(N-1)^a} \le \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a} \le \frac{2}{(N-1)^a}. \tag{A.14}$$

Then it is readily seen from (A.13) that

$$\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \frac{\eta_N}{N^a} + (1-\eta_N) \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a} \geq \frac{\eta_N}{N^a} + \frac{1-\eta_N}{(N-1)^a} \geq \frac{\eta_N}{N^a} + \frac{1-\eta_N}{N^a} = \frac{1}{N^a}, \tag{A.15}$$

where the first inequality comes from (A.14). Similarly, one can upper bound

$$\begin{aligned}\sum_{n=1}^{N} \frac{\eta_n^N}{n^a} = \frac{\eta_N}{N^a} + (1-\eta_N) \sum_{n=1}^{N-1} \frac{\eta_n^{N-1}}{n^a} &\overset{\text{(i)}}{\leq} \frac{\eta_N}{N^a} + \frac{2(1-\eta_N)}{(N-1)^a} \overset{\text{(ii)}}{=} \frac{H+1}{N^a(H+N)} + \frac{2(N-1)^{1-a}}{H+N} \\ &\overset{\text{(iii)}}{\leq} \frac{H+1}{N^a(H+N)} + \frac{2N^{1-a}}{H+N} = \frac{1}{N^a}\left(\frac{H+1}{H+N} + \frac{2N}{H+N}\right) \overset{\text{(iv)}}{\leq} \frac{2}{N^a},\end{aligned}$$

where (i) arises from (A.14), (ii) follows from the choice $\eta_N = \frac{H+1}{H+N}$, (iii) holds since $a \leq 1$, and (iv) follows since $H \geq 1$. Consequently, we can immediately establish the advertised property (3.14a) by induction.

A.3 Proof of key lemmas in Chapter 3.3.3

A.3.1 Proof of Lemma 2

To begin with, suppose that we can prove

$$Q_h^k(s,a) \geq Q_h^\star(s,a) \qquad \text{for all } (k,h,s,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}. \tag{A.16}$$

Then this property would immediately lead to the claim w.r.t. V_h^k, namely,

$$V_h^k(s) \geq Q_h^k\big(s, \pi_h^\star(s)\big) \geq Q_h^\star\big(s, \pi_h^\star(s)\big) = V_h^\star(s) \qquad \text{for all } (k,h,s) \in [K] \times [H] \times \mathcal{S}. \tag{A.17}$$

As a result, it suffices to focus on justifying the claim (A.16), which we shall accomplish by induction.

- *Base case.* Given that the initialization obeys $Q_h^1(s,a) = H \geq Q_h^\star(s,a)$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the claim (A.16) holds trivially when $k = 1$.
- *Induction.* Suppose that the claim (A.16) holds all the way up to the k-th episode, and we wish to establish it for the $(k+1)$-th episode as well. To complete the induction argument, it suffices to justify

  $$\min\left\{Q_h^{\mathsf{UCB},k+1}(s,a), Q_h^{\mathsf{R},k+1}(s,a)\right\} \geq Q_h^\star(s,a)$$

  according to line 12 of Algorithm 3. Recognizing that $Q_h^{\mathsf{UCB},k+1}$ is computed via the standard UCB-Q update rule (see line 2 of Algorithm 6), we can readily invoke the argument in Jin

et al. (2018, Lemma 4.3) to show that with probability at least $1-\delta$,

$$Q_h^{\mathsf{UCB},k+1}(s,a) \geq Q_h^\star(s,a)$$

holds simultaneously for all $(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$. Therefore, it is sufficient to prove that

$$Q_h^{\mathsf{R},k+1}(s,a) \geq Q_h^\star(s,a). \tag{A.18}$$

The remainder of the proof is thus devoted to justifying (A.18), assuming that the claim (A.16) holds all the way up to k.

Since $Q_h^{\mathsf{R},k}(s_h^k,a_h^k)$ is updated in the k-th episode while other entries of $Q_h^{\mathsf{R},k}$ remain fixed, it suffices to verify

$$Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) \geq Q_h^\star(s_h^k,a_h^k).$$

We remind the readers of two important short-hand notation that shall be used when it is clear from the context:

- $N_h^k = N_h^k(s_h^k,a_h^k)$ denotes the number of times that the state-action pair (s_h^k,a_h^k) has been visited at step h by the end of the k-th episode;
- $k^n = k_h^n(s_h^k,a_h^k)$ denotes the index of the episode in which the state-action pair (s_h^k,a_h^k) is visited for the n-th time at step h.

Step 1: decomposing $Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)$. To begin with, the above definition of N_h^k and k^n allows us to write

$$Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) = Q_h^{\mathsf{R},k^{N_h^k}+1}(s_h^k,a_h^k), \tag{A.19}$$

since $k^{N_h^k} = k^{N_h^k(s_h^k,a_h^k)} = k$. According to the update rule (i.e., line 11 in Algorithm 3 and line 7 in Algorithm 6), we obtain

$$\begin{aligned}
Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) = Q_h^{\mathsf{R},k^{N_h^k}+1}(s_h^k,a_h^k) &= (1-\eta_{N_h^k})Q_h^{\mathsf{R},k^{N_h^k}}(s_h^k,a_h^k) \\
&\quad + \eta_{N_h^k}\Big\{r_h(s_h^k,a_h^k) + V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - V_{h+1}^{\mathsf{R},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) + \mu_h^{\mathsf{ref},k^{N_h^k}+1}(s_h^k,a_h^k) + b_h^{\mathsf{R},k^{N_h^k}+1}\Big\} \\
&= (1-\eta_{N_h^k})Q_h^{\mathsf{R},k^{N_h^k-1}+1}(s_h^k,a_h^k) \\
&\quad + \eta_{N_h^k}\Big\{r_h(s_h^k,a_h^k) + V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - V_{h+1}^{\mathsf{R},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) + \mu_h^{\mathsf{ref},k^{N_h^k}+1}(s_h^k,a_h^k) + b_h^{\mathsf{R},k^{N_h^k}+1}\Big\},
\end{aligned}$$

where the last identity again follows from our argument for justifying (A.19). Applying this relation recursively and invoking the definitions of η_0^N and η_n^N in (4.16), we are left with

$$Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) = \eta_0^{N_h^k} Q_h^{\mathsf{R},1}(s_h^k,a_h^k) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big\{ r_h(s_h^k,a_h^k) + V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \mu_h^{\mathsf{ref},k^n+1}(s_h^k,a_h^k) + b_h^{\mathsf{R},k^n+1} \Big\}. \tag{A.20}$$

Additionally, the basic relation $\eta_0^{N_h^k} + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see (4.16) and (4.17)) tells us that

$$Q_h^\star(s_h^k,a_h^k) = \eta_0^{N_h^k} Q_h^\star(s_h^k,a_h^k) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} Q_h^\star(s_h^k,a_h^k), \tag{A.21}$$

which combined with (B.65) leads to

$$\begin{aligned} Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) &= \eta_0^{N_h^k} \big(Q_h^{\mathsf{R},1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)\big) \\ &+ \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big\{ r_h(s_h^k,a_h^k) + V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \mu_h^{\mathsf{ref},k^n+1}(s_h^k,a_h^k) + b_h^{\mathsf{R},k^n+1} - Q_h^\star(s_h^k,a_h^k) \Big\}. \end{aligned} \tag{A.22}$$

To continue, invoking the Bellman optimality equation

$$Q_h^\star(s_h^k,a_h^k) = r_h(s_h^k,a_h^k) + P_{h,s_h^k,a_h^k} V_{h+1}^\star \tag{A.23}$$

and using the construction of μ_h^{ref} in line 9 of Algorithm 6 (which is the running mean of V_{h+1}^{R}), we reach

$$\begin{aligned} &r_h(s_h^k,a_h^k) + V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \mu_h^{\mathsf{ref},k^n+1}(s_h^k,a_h^k) + b_h^{\mathsf{R},k^n+1} - Q_h^\star(s_h^k,a_h^k) \\ &= V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \frac{\sum_{i=1}^n V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i})}{n} - P_{h,s_h^k,a_h^k} V_{h+1}^\star + b_h^{\mathsf{R},k^n+1} \qquad \text{(A.24)} \\ &= P_{h,s_h^k,a_h^k} \Big\{ V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n} \Big\} + \frac{\sum_{i=1}^n P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},k^i}\big)}{n} - P_{h,s_h^k,a_h^k} V_{h+1}^\star + b_h^{\mathsf{R},k^n+1} + \xi_h^{k^n}, \\ &= P_{h,s_h^k,a_h^k} \Big\{ V_{h+1}^{k^n} - V_{h+1}^\star + \frac{\sum_{i=1}^n \big(V_{h+1}^{\mathsf{R},k^i} - V_{h+1}^{\mathsf{R},k^n}\big)}{n} \Big\} + b_h^{\mathsf{R},k^n+1} + \xi_h^{k^n}. \qquad \text{(A.25)} \end{aligned}$$

Here, we have introduced the following quantity

$$\xi_h^{k^n} \coloneqq \big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) + \frac{1}{n}\sum_{i=1}^{n}\big(P_h^{k^i} - P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathsf{R},k^i}, \tag{A.26}$$

with the notation P_h^k defined in (3.15). Putting (A.25) and (B.67) together leads to the following decomposition

$$\begin{aligned} Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) &= \eta_0^{N_h^k}\Big(Q_h^{\mathsf{R},1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)\Big) \\ &\quad + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left\{P_{h,s_h^k,a_h^k}\left(V_{h+1}^{k^n} - V_{h+1}^\star + \frac{\sum_{i=1}^n\big(V_{h+1}^{\mathsf{R},k^i} - V_{h+1}^{\mathsf{R},k^n}\big)}{n}\right) + b_h^{\mathsf{R},k^n+1} + \xi_h^{k^n}\right\}. \end{aligned} \tag{A.27}$$

Step 2: two key quantities for lower bounding $Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)$. In order to develop a lower bound on $Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)$ based on the decomposition (B.71), we make note of several simple facts as follows.

(i) The initialization satisfies $Q_h^{\mathsf{R},1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \geq 0$.

(ii) For any $1 \leq k^n \leq k$, one has

$$V_{h+1}^{k^n} \geq V_{h+1}^\star, \tag{A.28}$$

owing to the induction hypotheses (A.16) and (A.17) that hold up to k.

(iii) For all $0 \leq i \leq n$ and any $s \in \mathcal{S}$, one has

$$V_{h+1}^{\mathsf{R},k^i}(s) - V_{h+1}^{\mathsf{R},k^n}(s) \geq 0, \tag{A.29}$$

which holds since the reference value $V_h^{\mathsf{R}}(s)$ is monotonically non-increasing in view of the monotonicity of $V_h(s)$ in (3.17b) and the update rule in line 16 of Algorithm 3.

The above three facts taken collectively with (B.71) allow one to drop several terms and yield

$$Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \geq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(b_h^{\mathsf{R},k^n+1} + \xi_h^{k^n}\big). \tag{A.30}$$

In the sequel, we aim to establish $Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k) \geq Q_h^\star(s_h^k,a_h^k)$ based on this inequality (A.30).

As it turns out, if one could show that

$$\left|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\xi_h^{k^n}\right| \leq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}b_h^{\mathsf{R},k^n+1}, \tag{A.31}$$

then taking this together with (A.30) and the triangle inequality would immediately lead to the desired result

$$Q_h^{\mathsf{R},k+1}(s_h^k,a_h^k)-Q_h^\star(s_h^k,a_h^k)\geq\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}b_h^{\mathsf{R},k^n+1}-\left|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\xi_h^{k^n}\right|\geq 0. \tag{A.32}$$

As a result, the remaining steps come down to justifying the claim (B.73). In order to do so, we need to control the following two quantities (in view of (B.69))

$$I_1 \coloneqq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n}-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n}-V_{h+1}^{\mathsf{R},k^n}\big), \tag{A.33a}$$

$$I_2 \coloneqq \sum_{n=1}^{N_h^k}\frac{1}{n}\eta_n^{N_h^k}\sum_{i=1}^{n}\big(P_h^{k^i}-P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathsf{R},k^i} \tag{A.33b}$$

separately, which constitutes the next two steps. As will be seen momentarily, these two terms can be controlled in a similar fashion using Freedman's inequality.

Step 3: controlling I_1. In the following, we intend to invoke Lemma 24 to control the term I_1 defined in (A.33a). To begin with, consider any $(N,h)\in[K]\times[H]$, and introduce

$$W_{h+1}^i \coloneqq V_{h+1}^i-V_{h+1}^{\mathsf{R},i} \qquad \text{and} \qquad u_h^i(s,a,N)\coloneqq\eta_{N_h^i(s,a)}^N\geq 0. \tag{A.34}$$

Accordingly, we can derive and define

$$\|W_{h+1}^i\|_\infty\leq\|V_{h+1}^{\mathsf{R},i}\|_\infty+\|V_{h+1}^i\|_\infty\leq 2H=:C_{\mathrm{w}}, \tag{A.35}$$

and

$$\max_{N,h,s,a\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}}\eta_{N_h^i(s,a)}^N\leq\frac{2H}{N}=:C_{\mathrm{u}}, \tag{A.36}$$

where the last inequality follows since (according to Lemma 1 and the definition in (4.16))

$$\begin{aligned}&\eta_{N_h^i(s,a)}^N\leq\frac{2H}{N}, && \text{if } 1\leq N_h^i(s,a)\leq N;\\ &\eta_{N_h^i(s,a)}^N=0, && \text{if } N_h^i(s,a)>N.\end{aligned}$$

Moreover, observed from (4.17), we have

$$0\leq\sum_{n=1}^{N}u_h^{k_h^n(s,a)}(s,a,N)=\sum_{n=1}^{N}\eta_n^N\leq 1 \tag{A.37}$$

holds for all $(N,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$. Therefore, choosing $(N,s,a) = (N_h^k, s_h^k, a_h^k)$ and applying Lemma 24 with the quantities (A.34) implies that, with probability at least $1-\delta$,

$$
\begin{aligned}
|I_1| &= \left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right| = \left| \sum_{i=1}^{k} X_i(s_h^k, a_h^k, h, N_h^k) \right| \\
&\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s_h^k, a_h^k, N_h^k) \mathsf{Var}_{h,s_h^k,a_h^k}\big(W_{h+1}^{k^n}\big)} + \left(C_{\mathrm{u}} C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}} C_{\mathrm{w}} \right) \log^2 \frac{SAT}{\delta} \\
&\asymp \sqrt{\frac{H}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)} + \frac{H^2 \log^2 \frac{SAT}{\delta}}{N_h^k} \qquad \text{(A.38)} \\
&\lesssim \sqrt{\frac{H}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k, a_h^k) - \big(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k, a_h^k)\big)^2} + \frac{H^2 \log^2 \frac{SAT}{\delta}}{(N_h^k)^{3/4}}, \qquad \text{(A.39)}
\end{aligned}
$$

where the proof of the last inequality (B.151) needs additional explanation and is postponed to Appendix A.3.1.1 to streamline the presentation.

Step 4: controlling I_2. Next, we turn attention to the quantity I_2 defined in (A.33b). Rearranging terms in the definition (A.33b), we are left with

$$
I_2 = \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \frac{\sum_{i=1}^{n} \big(P_h^{k^i} - P_{h,s_h^k,a_h^k}\big) V_{h+1}^{\mathsf{R},k^i}}{n} = \sum_{i=1}^{N_h^k} \left(\sum_{n=i}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} \right) \big(P_h^{k^i} - P_{h,s_h^k,a_h^k}\big) V_{h+1}^{\mathsf{R},k^i},
$$

which can again be controlled by invoking Lemma 24. To do so, we abuse the notation by taking

$$
W_{h+1}^i \coloneqq V_{h+1}^{\mathsf{R},i} \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \sum_{n=N_h^i(s,a)}^{N} \frac{\eta_n^N}{n} \geq 0. \qquad \text{(A.40)}
$$

These quantities satisfy

$$
\big\|W_{h+1}^i\big\|_\infty \leq \big\|V_{h+1}^{\mathsf{R},i}\big\|_\infty \leq H =: C_{\mathrm{w}} \qquad \text{(A.41)}
$$

and, according to Lemma 1,

$$
\max_{N,h,s,a \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}} \sum_{n=N_h^i(s,a)}^{N} \frac{\eta_n^N}{n} \leq \sum_{n=1}^{N} \frac{\eta_n^N}{n} \leq \frac{2}{N} =: C_{\mathrm{u}}. \qquad \text{(A.42)}
$$

Then it is readily seen from (A.42) that

$$0 \le \sum_{n=1}^{N} u_h^{k_h^n(s,a)}(s,a,N) \le \sum_{n=1}^{N} \frac{2}{N} \le 2. \tag{A.43}$$

holds for all $(N,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$.

With the above relations in mind, Taking $(N,s,a) = (N_h^k, s_h^k, a_h^k)$ and applying Lemma 24 w.r.t. the quantities (A.40) reveals that

$$|I_2| = \left| \sum_{i=1}^{N_h^k} \sum_{n=i}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} \big(P_h^{k^i} - P_{h,s_h^k,a_h^k}\big) V_{h+1}^{\mathsf{R},k^i} \right| = \left| \sum_{i=1}^{k} X_i(s_h^k, a_h^k, h, N_h^k) \right| \tag{A.44}$$

$$\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s_h^k, a_h^k, N_h^k) \mathsf{Var}_{h,s_h^k,a_h^k}\big(W_{h+1}^{k^n}\big) + \left(C_{\mathrm{u}} C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}} C_{\mathrm{w}}\right) \log^2 \frac{SAT}{\delta}}$$

$$\lesssim \sqrt{\frac{1}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\frac{1}{N_h^k} \sum_{n=1}^{N_h^k} \mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},k^n}\big) + \frac{H}{N_h^k} \log^2 \frac{SAT}{\delta}} \tag{A.45}$$

$$\lesssim \sqrt{\frac{1}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k, a_h^k) - \big(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k, a_h^k)\big)^2 + \frac{H}{(N_h^k)^{3/4}} \log^2 \frac{SAT}{\delta}} \tag{A.46}$$

with probability exceeding $1-\delta$, where the proof of the last inequality (B.153) is deferred to Appendix B.3.4.3 in order to streamline presentation.

Step 5: combining the above bounds. Summing up the results in (B.151) and (B.153), we arrive at an upper bound on $\left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n}\right|$ as follows:

$$\begin{aligned}
\left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n} \right| &\le |I_1| + |I_2| \\
&\lesssim \sqrt{\frac{H}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k, a_h^k) - \big(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k, a_h^k)\big)^2} \\
&\quad + \sqrt{\frac{1}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k, a_h^k) - \big(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k, a_h^k)\big)^2} + \frac{H^2 \log^2 \frac{SAT}{\delta}}{(N_h^k)^{3/4}} \\
&\le B_h^{\mathsf{R},k^{N_h^k+1}}(s_h^k, a_h^k) + c_{\mathrm{b}} \frac{H^2 \log^2 \frac{SAT}{\delta}}{(N_h^k)^{3/4}}
\end{aligned} \tag{A.47}$$

for some sufficiently large constant $c_{\mathrm{b}} > 0$, where the last line follows from the definition of $B_h^{\mathsf{R},k^{N_h^k+1}}(s_h^k, a_h^k)$ in line 14 of Algorithm 6.

In order to establish the desired bound (B.73), we still need to control the sum $\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{\mathsf{R},k^n+1}$. Towards this end, the definition of b_h^{R,k^n+1} (resp. δ_h^{R}) in line 6 (resp. line 15) of Algorithm 6 yields

$$b_h^{\mathsf{R},k^n+1} = \Big(1-\frac{1}{\eta_n}\Big) B_h^{\mathsf{R},k^n}(s_h^k,a_h^k) + \frac{1}{\eta_n} B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k) + \frac{c_{\mathrm{b}}}{n^{3/4}} H^2 \log^2 \frac{SAT}{\delta}. \tag{A.48}$$

This taken collectively with the definition (4.16) of η_n^N allows us to expand

$$\begin{aligned}
&\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{\mathsf{R},k^n+1} \\
&= \sum_{n=1}^{N_h^k} \eta_n \prod_{i=n+1}^{N_h^k} (1-\eta_i) \left(\Big(1-\frac{1}{\eta_n}\Big) B_h^{\mathsf{R},k^n}(s_h^k,a_h^k) + \frac{1}{\eta_n} B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k) \right) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^2 \log^2 \frac{SAT}{\delta} \\
&= \sum_{n=1}^{N_h^k} \prod_{i=n+1}^{N_h^k} (1-\eta_i) \Big(-(1-\eta_n) B_h^{\mathsf{R},k^n}(s_h^k,a_h^k) + B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k) \Big) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^2 \log^2 \frac{SAT}{\delta} \\
&= \sum_{n=1}^{N_h^k} \left(\prod_{i=n+1}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k) - \prod_{i=n}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n}(s_h^k,a_h^k) \right) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^2 \log^2 \frac{SAT}{\delta} \\
&\overset{\text{(i)}}{=} \sum_{n=1}^{N_h^k} \prod_{i=n+1}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k) - \sum_{n=2}^{N_h^k} \prod_{i=n}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n}(s_h^k,a_h^k) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^2 \log^2 \frac{SAT}{\delta} \\
&\overset{\text{(ii)}}{=} \sum_{n=1}^{N_h^k} \prod_{i=n+1}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k) - \sum_{n=1}^{N_h^k-1} \prod_{i=n+1}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^2 \log^2 \frac{SAT}{\delta} \\
&= B_h^{\mathsf{R},k^{N_h^k}+1}(s_h^k,a_h^k) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^2 \log^2 \frac{SAT}{\delta}.
\end{aligned} \tag{A.49}$$

Here, (i) is valid due to the fact that $B_h^{\mathsf{R},k^1}(s_h^k,a_h^k) = 0$; (ii) follows from the fact that

$$\begin{aligned}
\sum_{n=2}^{N_h^k} \prod_{i=n}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n}(s_h^k,a_h^k) &= \sum_{n=1}^{N_h^k-1} \prod_{i=n+1}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^{n+1}}(s_h^k,a_h^k) \\
&= \sum_{n=1}^{N_h^k-1} \prod_{i=n+1}^{N_h^k} (1-\eta_i) B_h^{\mathsf{R},k^n+1}(s_h^k,a_h^k),
\end{aligned}$$

where the first relation can be seen by replacing n with $n+1$, and the last relation holds true since the state-action pair (s_h^k,a_h^k) has not been visited at step h between the (k^n+1)-th episode and the $(k^{n+1}-1)$-th episode. Combining the above identity (A.49) with the following property (see

Lemma 1)

$$\frac{1}{(N_h^k)^{3/4}} \leq \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} \leq \frac{2}{(N_h^k)^{3/4}},$$

we can immediately demonstrate that

$$B_h^{\mathsf{R},k^{N_h^k+1}}(s_h^k,a_h^k)+c_{\mathrm{b}}\frac{H^2\log^2\frac{SAT}{\delta}}{(N_h^k)^{3/4}} \leq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}b_h^{\mathsf{R},k^n+1} \leq B_h^{\mathsf{R},k^{N_h^k+1}}(s_h^k,a_h^k)+2c_{\mathrm{b}}\frac{H^2\log^2\frac{SAT}{\delta}}{(N_h^k)^{3/4}}. \quad \text{(A.50)}$$

Taking (B.154) and (B.62) collectively demonstrates that

$$\left|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\xi_h^{k^n}\right| \leq B_h^{\mathsf{R},k^{N_h^k+1}}(s_h^k,a_h^k)+c_{\mathrm{b}}\frac{H^2\log^2\frac{SAT}{\delta}}{(N_h^k)^{3/4}} \leq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}b_h^{\mathsf{R},k^n+1} \quad \text{(A.51)}$$

as claimed in (B.73). We have thus concluded the proof of Lemma 2 based on the argument in Step 2.

A.3.1.1 Proof of the inequality (B.151)

In order to establish the inequality (B.151), it suffices to look at the following term

$$I_3 := \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathsf{R},k^n}\big)-\sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k)+\big(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2, \quad \text{(A.52)}$$

which forms the main content of this sub-chapter.

First of all, the update rules of $\mu_h^{\mathsf{adv},k^{n+1}}$ and $\sigma_h^{\mathsf{adv},k^{n+1}}$ in lines 11-12 of Algorithm 6 tell us that

$$\begin{aligned}
\mu_h^{\mathsf{adv},k^{n+1}}(s_h^k,a_h^k) &= \mu_h^{\mathsf{adv},k^n+1}(s_h^k,a_h^k) = (1-\eta_n)\mu_h^{\mathsf{adv},k^n}(s_h^k,a_h^k)+\eta_n\big(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big),\\
\sigma_h^{\mathsf{adv},k^{n+1}}(s_h^k,a_h^k) &= \sigma_h^{\mathsf{adv},k^n+1}(s_h^k,a_h^k) = (1-\eta_n)\sigma_h^{\mathsf{adv},k^n}(s_h^k,a_h^k)+\eta_n\big(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2.
\end{aligned}$$

Applying this relation recursively and invoking the definitions of η_n^N (resp. P_h^k) in (4.16) (resp. (3.15)) give

$$\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k) \overset{\text{(i)}}{=} \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big) = \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathsf{R},k^n}\big), \quad \text{(A.53a)}$$

$$\sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k) \overset{\text{(ii)}}{=} \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2 = \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n}-V_{h+1}^{\mathsf{R},k^n}\big)^2. \quad \text{(A.53b)}$$

Recognizing that $\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see (4.17)), we can immediately apply Jensen's inequality to the expressions (i) and (ii) to yield

$$\sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k, a_h^k) \geq \Big(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k, a_h^k)\Big)^2. \tag{A.54}$$

Further, in view of the definition (B.51), we have

$$\mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) = P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2 - \Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)\Big)^2,$$

which allows one to decompose and bound I_3 as follows

$$\begin{aligned}
I_3 &= \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2 - \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2 \\
&\qquad + \bigg(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)\bigg)^2 - \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)\Big)^2 \\
&\leq \underbrace{\bigg|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2\bigg|}_{=:I_{3,1}} \\
&\qquad + \underbrace{\bigg(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)\bigg)^2 - \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)\Big)^2}_{=:I_{3,2}}.
\end{aligned} \tag{A.55}$$

It then boils down to controlling the above two terms in (B.156) separately.

Step 1: bounding $I_{3,1}$. To upper bound the term $I_{3,1}$ in (B.156), we resort to Lemma 24 by setting

$$W_{h+1}^i \coloneqq \big(V_{h+1}^i - V_{h+1}^{\mathsf{R},i}\big)^2 \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \eta_{N_h^i(s,a)}^N. \tag{A.56}$$

It is easily seen that

$$\|W_{h+1}^i\|_\infty \leq \Big(\big\|V_{h+1}^{\mathsf{R},i}\big\|_\infty + \big\|V_{h+1}^i\big\|_\infty\Big)^2 \leq 4H^2 =: C_{\mathrm{w}}, \tag{A.57}$$

and it follows from (A.36) that

$$\max_{N,h,s,a\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}} \eta_{N_h^i(s,a)}^N \leq \frac{2H}{N} =: C_{\mathrm{u}}. \tag{A.58}$$

Armed with the properties (A.57) and (A.58) and recalling (B.149), we can invoke Lemma 24 w.r.t. (A.56) and set $(N, s, a) = (N_h^k, s_h^k, a_h^k)$ to yield

$$
\begin{aligned}
I_{3,1} &= \left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2 \right| = \left| \sum_{i=1}^{k} X_i(s_h^k, a_h^k, h, N_h^k) \right| \\
&\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s_h^k, a_h^k, N_h^k) \mathsf{Var}_{h,s_h^k,a_h^k}\big(W_{h+1}^{k^n}\big)} + \left(C_{\mathrm{u}} C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}} C_{\mathrm{w}} \right) \log^2 \frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{H}{N_h^k} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathsf{Var}_{h,s_h^k,a_h^k}\Big(\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2\Big)} + \frac{H^3 \log^2 \frac{SAT}{\delta}}{N_h^k} \\
&\lesssim \sqrt{\frac{H^5}{N_h^k} \log^2 \frac{SAT}{\delta}} + \frac{H^3}{N_h^k} \log^2 \frac{SAT}{\delta}
\end{aligned} \tag{A.59}
$$

with probability at least $1-\delta$. Here, the last inequality results from the fact $\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \le 1$ (see (4.17)) and the following trivial result:

$$
\mathsf{Var}_{h,s_h^k,a_h^k}\Big(\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2\Big) \le \big\| \big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^4 \big\|_\infty \le 16H^4. \tag{A.60}
$$

Step 2: bounding $I_{3,2}$. Jensen's inequality tells us that

$$
\begin{aligned}
\left(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right)^2 &= \left(\sum_{n=1}^{N_h^k} \big(\eta_n^{N_h^k}\big)^{1/2} \cdot \big(\eta_n^{N_h^k}\big)^{1/2} P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right)^2 \\
&\le \left\{ \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \right\} \left\{ \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \Big)^2 \right\} \\
&\le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \Big)^2,
\end{aligned}
$$

where the last line arises from (4.17). Substitution into $I_{3,2}$ (cf. (B.156)) gives

$$
\begin{aligned}
I_{3,2} &\le \left(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right)^2 - \left(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right)^2 \\
&= \left\{ \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right\} \left\{ \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \big(P_h^{k^n} + P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right\}.
\end{aligned} \tag{A.61}
$$

In what follows, we would like to use this relation to show that

$$I_{3,2} \leq C_{32}\left\{\sqrt{\frac{H^5}{N_h^k}\log^2\frac{SAT}{\delta}} + \frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}\right\} \tag{A.62}$$

for some universal constant $C_{32} > 0$.

If $I_{3,2} \leq 0$, then (B.161) holds true trivially. Consequently, it is sufficient to study the case where $I_{3,2} > 0$. To this end, we first note that the term in the first pair of curly brakets of (B.159) is exactly I_1 (see (A.33a)), which can be bounded by recalling (B.150):

$$\begin{aligned}
|I_1| &\lesssim \sqrt{\frac{H}{N_h^k}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)} + \frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k} \\
&\lesssim \sqrt{\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}} + \frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k} \\
&\lesssim \sqrt{\frac{H^3}{N_h^k}\log^2\frac{SAT}{\delta}} + \frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta},
\end{aligned} \tag{A.63}$$

with probability at least $1-\delta$. Here, the second inequality arises from the following property

$$\mathsf{Var}_{h,s_h^k,a_h^k}\Big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\Big) \leq \big\|\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)^2\big\|_\infty \leq 4H^2, \tag{A.64}$$

whereas the last inequality (A.63) holds as a result of the fact $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k} \leq 1$ (see (4.17)).

Moreover, the term in the second pair of curly brakets of (B.159) can be bounded straightforwardly as follows

$$\begin{aligned}
&\Big|\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\big(P_h^{k^n} + P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big)\Big| \\
&\qquad \leq \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(\big\|P_h^{k^n}\big\|_1 + \big\|P_{h,s_h^k,a_h^k}\big\|_1\Big)\big\|V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big\|_\infty \leq 2H,
\end{aligned} \tag{A.65}$$

where we have made use of the property (4.17), as well as the elementary facts $\big\|V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big\|_\infty \leq H$ and $\big\|P_h^{k^n}\big\|_1 = \big\|P_{h,s_h^k,a_h^k}\big\|_1 = 1$. Substituting the above two results (A.63) and (A.65) back into (B.159), we arrive at the bound (B.161) as long as $I_{3,2} > 0$. Putting all cases together, we have established the claim (B.161).

Step 3: putting all this together. To finish up, plugging the bounds (A.59) and (B.161) into (B.156), we can conclude that

$$I_3 \le I_{3,1} + I_{3,2} \le C_3 \left\{ \sqrt{\frac{H^5}{N_h^k} \log^2 \frac{SAT}{\delta}} + \frac{H^3}{N_h^k} \log^2 \frac{SAT}{\delta} \right\}$$

for some constant $C_3 > 0$. This together with the definition (A.52) of I_3 results in

$$\begin{aligned}
&\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \\
&\qquad \le \Big\{ \sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k) - \big(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2 \Big\} + C_3 \left(\sqrt{\frac{H^5}{N_h^k} \log^2 \frac{SAT}{\delta}} + \frac{H^3}{N_h^k} \log^2 \frac{SAT}{\delta} \right),
\end{aligned}$$

which combined with the elementary inequality $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for any $u, v \ge 0$ and (A.54) yields

$$\begin{aligned}
&\left\{ \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{k^n} - V_{h+1}^{\mathsf{R},k^n}\big) \right\}^{1/2} \\
&\qquad \lesssim \Big\{ \sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k) - \big(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2 \Big\}^{1/2} + \frac{H^{5/4}}{\big(N_h^k\big)^{1/4}} \log^{1/2} \frac{SAT}{\delta} + \frac{H^{3/2}}{\big(N_h^k\big)^{1/2}} \log \frac{SAT}{\delta}.
\end{aligned}$$

Substitution into (B.150) establishes the desired result (B.151).

A.3.1.2 Proof of the inequality (B.153)

In order to prove the inequality (B.153), it suffices to look at the following term

$$I_4 := \frac{1}{N_h^k} \sum_{n=1}^{N_h^k} \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathsf{R},k^n}) - \Big(\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k) - \big(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k)\big)^2 \Big). \tag{A.66}$$

In view of the update rules of $\mu_h^{\mathsf{ref},k^{n+1}}$ and $\sigma_h^{\mathsf{ref},k^{n+1}}$ in lines 9-10 of Algorithm 6, we have

$$\begin{aligned}
\mu_h^{\mathsf{ref},k^{n+1}}(s_h^k,a_h^k) &= \mu_h^{\mathsf{ref},k^n+1}(s_h^k,a_h^k) = \left(1 - \frac{1}{n}\right) \mu_h^{\mathsf{ref},k^n}(s_h^k,a_h^k) + \frac{1}{n} V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}), \\
\sigma_h^{\mathsf{ref},k^{n+1}}(s_h^k,a_h^k) &= \sigma_h^{\mathsf{ref},k^n+1}(s_h^k,a_h^k) = \left(1 - \frac{1}{n}\right) \sigma_h^{\mathsf{ref},k^n}(s_h^k,a_h^k) + \frac{1}{n} \big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2,
\end{aligned}$$

Through simple recursion, these identities together with the definition (3.15) of P_h^k lead to

$$\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k) \overset{\text{(i)}}{=} \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) = \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n} V_{h+1}^{\mathsf{R},k^n}, \tag{A.67a}$$

$$\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k) \overset{\text{(ii)}}{=} \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} \big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2 = \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}\big(V_{h+1}^{\mathsf{R},k^n}\big)^2, \tag{A.67b}$$

The expressions (i) and (ii) combined with Jensen's inequality give

$$\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k) \geq \Big(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k)\Big)^2. \tag{A.68}$$

Taking these together with the definition

$$\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathsf{R},k^n}) = P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},k^n}\big)^2 - \big(P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k^n}\big)^2,$$

we obtain

$$\begin{aligned} I_4 &= \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\Big(P_{h,s_h^k,a_h^k}(V_{h+1}^{\mathsf{R},k^n})^2 - \big(P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k^n}\big)^2\Big) - \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}\big(V_{h+1}^{\mathsf{R},k^n}\big)^2 + \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}V_{h+1}^{\mathsf{R},k^n}\right)^2 \\ &= \underbrace{\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\big(P_{h,s_h^k,a_h^k} - P_h^{k^n}\big)\big(V_{h+1}^{\mathsf{R},k^n}\big)^2}_{=:I_{4,1}} + \underbrace{\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}V_{h+1}^{\mathsf{R},k^n}\right)^2 - \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\Big(P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k^n}\Big)^2}_{=:I_{4,2}}. \end{aligned} \tag{A.69}$$

In what follows, we shall bound the terms $I_{4,1}$ and $I_{4,2}$ in (B.166) separately.

Step 1: bounding $I_{4,1}$. The first term $I_{4,1}$ in (B.166) can be bounded by means of Lemma 24 in an almost identical fashion as $I_{3,1}$ in (A.59). Specifically, let us set

$$W_{h+1}^i \coloneqq (V_{h+1}^{\mathsf{R},i})^2 \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \frac{1}{N},$$

which clearly obey

$$|u_h^i(s,a,N)| = \frac{1}{N} =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \leq H^2 =: C_{\mathrm{w}}.$$

It is easily verified that

$$\sum_{n=1}^{N} u_h^{k^n(s,a)}(s,a,N) = \sum_{n=1}^{N}\frac{1}{N} = 1$$

holds for all $(N,s,a)\in[K]\times\mathcal{S}\times\mathcal{A}$. Hence we can take $(N,s,a)=(N_h^k,s_h^k,a_h^k)$ and apply Lemma 24 to yield

$$\begin{aligned}
|I_{4,1}| &= \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\big(P_h^{k^n}-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathsf{R},k^n}\big)^2\right| = \left|\sum_{i=1}^{k}X_i(s_h^k,a_h^k,h,N_h^k)\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k}u_h^{k^n}(s_h^k,a_h^k,N_h^k)\mathsf{Var}_{h,s_h^k,a_h^k}\big(W_{h+1}^{k^n}\big)}+\left(C_{\mathrm{u}}C_{\mathrm{w}}+\sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{H^4\log^2\frac{SAT}{\delta}}{N_h^k}}+\frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k} \qquad \text{(A.70)}
\end{aligned}$$

with probability at least $1-\delta$, where the last inequality results from the fact that

$$\mathsf{Var}_{h,s_h^k,a_h^k}\big(W_{h+1}^{k^n}\big)\le\big\|W_{h+1}^{k^n}\big\|_\infty^2\le C_{\mathrm{w}}^2=H^4.$$

Step 2: bounding $I_{4,2}$. We now turn to the other term $I_{4,2}$ defined in (B.166). Towards this, we first make the observation that

$$\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k^n}\right)^2\le\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k^n}\right)^2, \qquad \text{(A.71)}$$

which follows from Jensen's inequality. Equipped with this relation, we can upper bound $I_{4,2}$ as follows

$$\begin{aligned}
I_{4,2} &\le \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}P_h^{k^n}V_{h+1}^{\mathsf{R},k^n}\right)^2-\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k^n}\right)^2 \\
&= \left\{\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\big(P_h^{k^n}-P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathsf{R},k^n}\right\}\left\{\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\big(P_h^{k^n}+P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathsf{R},k^n}\right\}. \qquad \text{(A.72)}
\end{aligned}$$

In the following, we would like to apply this relation to prove

$$I_{4,2}\le C_{42}\left(\sqrt{\frac{H^4}{N_h^k}\log^2\frac{SAT}{\delta}}+\frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta}\right) \qquad \text{(A.73)}$$

for some constant $C_{42}>0$.

When $I_{4,2}\le 0$, the claim (B.169) holds trivially. As a result, we shall focus on the case where $I_{4,2}>0$. Let us begin with the term in the first pair of curly brackets of (B.168). Towards this, let

us abuse the notation and set

$$W_{h+1}^i \coloneqq V_{h+1}^{\mathsf{R},i} \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \frac{1}{N},$$

which satisfy

$$|u_h^i(s,a,N)| = \frac{1}{N} =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}}.$$

Akin to our argument for bounding $I_{4,1}$, invoking Lemma 24 and setting $(N,s,a) = (N_h^k, s_h^k, a_h^k)$ imply that

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}(P_h^{k^n} - P_{h,s_h^k,a_h^k})V_{h+1}^{\mathsf{R},k^n}\right| \lesssim \sqrt{\frac{H^2\log^2\frac{SAT}{\delta}}{N_h^k}} + \frac{H\log^2\frac{SAT}{\delta}}{N_h^k}$$

with probability at least $1-\delta$. In addition, the term in the second pair of curly brackets of (B.168) can be bounded straightforwardly by

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}(P_h^{k^n} + P_{h,s_h^k,a_h^k})V_{h+1}^{\mathsf{R},k^n}\right| \le \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(\left\|P_h^{k^n}\right\|_1 + \left\|P_{h,s_h^k,a_h^k}\right\|_1\right)\left\|V_{h+1}^{\mathsf{R},k^n}\right\|_\infty \le 2H,$$

where we have used $\left\|V_{h+1}^{\mathsf{R},k^n}\right\|_\infty \le H$ and $\left\|P_h^{k^n}\right\|_1 = \left\|P_{h,s_h^k,a_h^k}\right\|_1 = 1$. Substituting the preceding facts into (B.168) validates the bound (B.169) as long as $I_{4,2} > 0$. We have thus finished the proof of the claim (B.169).

Step 3: putting all pieces together. Combining the results (A.70) and (B.169) with (B.166) yields

$$I_4 \le |I_{4,1}| + I_{4,2} \le C_4\left\{\sqrt{\frac{H^4}{N_h^k}\log^2\frac{SAT}{\delta}} + \frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta}\right\}$$

for some constant $C_4 > 0$. This bound taken together with the definition (A.66) of I_4 gives

$$\begin{aligned}\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathsf{R},k^n}) \le{}& \left\{\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k) - \left(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k)\right)^2\right\} \\ &+ C_4\left\{\sqrt{\frac{H^4}{N_h^k}\log^2\frac{SAT}{\delta}} + \frac{H^2}{N_h^k}\log^2\frac{SAT}{\delta}\right\}.\end{aligned}$$

Invoke the elementary inequality $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for any $u, v \ge 0$ and use the property (A.68) to obtain

$$\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\mathsf{R},k^n})\right)^{1/2} \lesssim \left\{\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k) - \left(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s_h^k,a_h^k)\right)^2\right\}^{1/2} + \frac{H}{(N_h^k)^{1/4}}\log^{1/2}\frac{SAT}{\delta} + \frac{H}{(N_h^k)^{1/2}}\log\frac{SAT}{\delta}.$$

Substitution into (A.45) directly establishes the desired result (B.153).

A.3.2 Proof of Lemma 3

A.3.2.1 Proof of the inequalities (3.21)

Suppose that we can verify the following inequality:

$$Q_h^{\mathsf{LCB},k}(s,a) \le Q_h^\star(s,a) \qquad \text{for all } (s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]. \tag{A.74}$$

which in turn yields

$$\max_a Q_h^{\mathsf{LCB},k}(s,a) \le \max_a Q_h^\star(s,a) = V_h^\star(s) \qquad \text{for all } (k,h,s) \in [K]\times[H]\times\mathcal{S}. \tag{A.75}$$

In addition, the construction of $V_h^{\mathsf{LCB},k}$ (see line 14 of Algorithm 3) allows us to show that

$$V_h^{\mathsf{LCB},k+1}(s) \le \max\left\{\max_{j:j\le k+1}\max_a Q_h^{\mathsf{LCB},j}(s,a),\ \max_{j:j\le k} V_h^{\mathsf{LCB},j}(s)\right\}.$$

This taken together with the initialization $V_h^{\mathsf{LCB},1} = 0$ and a simple induction argument yields

$$V_h^{\mathsf{LCB},k}(s) \le V_h^\star(s) \qquad \text{for all } (k,h,s) \in [K]\times[H]\times\mathcal{S}. \tag{A.76}$$

As a consequence, everything comes down to proving the claim (A.74), which we shall accomplish by induction.

Base case. Given our initialization, we have

$$Q_h^{\mathsf{LCB},1}(s,a) - Q_h^\star(s,a) = 0 - Q_h^\star(s,a) \le 0,$$

and hence the claim (A.74) holds trivially when $k = 1$.

Induction step. Suppose now that the claim (A.74) holds all the way up to k for all (s,a,h), and we would like to validate it for the $(k+1)$-th episode as well. Towards this end, recall that the

state-action pair (s_h^k, a_h^k) is visited in the k-th episode at time step h; this means that $Q_h^{\mathsf{LCB}}(s_h^k, a_h^k)$ is updated once we collect samples in the k-th episode, with all other entries Q_h^{LCB} frozen. It thus suffices to verify that

$$Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) \le Q_h^\star(s_h^k, a_h^k).$$

In what follows, we shall adopt the short-hand notation (see also Chapter 3.3.2)

$$N_h^k = N_h^k(s_h^k, a_h^k) \qquad \text{and} \qquad k^n = k_h^n(s_h^k, a_h^k)$$

which will be used throughout this subchapter as long as it is clear from the context.

The update rule of $Q_h^{\mathsf{LCB},k}$ (cf. line 2 of Algorithm 6) and the Bellman optimality equation in (A.23) tell us the following identities:

$$\begin{aligned}
Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) &= Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s_h^k, a_h^k) \\
&= (1-\eta_{N_h^k}) Q_h^{\mathsf{LCB},k^{N_h^k}}(s_h^k, a_h^k) + \eta_{N_h^k}\Big(r_h(s_h^k, a_h^k) + V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - b_h^{k^{N_h^k}}\Big), \\
Q_h^\star(s_h^k, a_h^k) &= (1-\eta_{N_h^k}) Q_h^\star(s_h^k, a_h^k) + \eta_{N_h^k} Q_h^\star(s_h^k, a_h^k) \\
&= (1-\eta_{N_h^k}) Q_h^\star(s_h^k, a_h^k) + \eta_{N_h^k}\Big(r(s_h^k, a_h^k) + P_{h,s_h^k,a_h^k} V_{h+1}^\star\Big),
\end{aligned}$$

which taken collectively lead to the following identity

$$\begin{aligned}
&Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) = Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \\
&\quad = (1-\eta_{N_h^k})\Big(Q_h^{\mathsf{LCB},k^{N_h^k}}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k)\Big) + \eta_{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - P_{h,s_h^k,a_h^k} V_{h+1}^\star - b_h^{k^{N_h^k}}\Big) \\
&\quad = (1-\eta_{N_h^k})\Big(Q_h^{\mathsf{LCB},k^{N_h^k-1}+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k)\Big) + \eta_{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - P_{h,s_h^k,a_h^k} V_{h+1}^\star - b_h^{k^{N_h^k}}\Big).
\end{aligned}$$

Recall the definitions of η_0^N and η_n^N in (4.16). Applying the above relation recursively and making use of the decomposition of $Q_h^\star(s_h^k, a_h^k)$ in (B.75) result in

$$\begin{aligned}
&Q_h^{\mathsf{LCB},k+1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k) \\
&\quad = \eta_0^{N_h^k}\Big(Q_h^{\mathsf{LCB},1}(s_h^k, a_h^k) - Q_h^\star(s_h^k, a_h^k)\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n}) - P_{h,s_h^k,a_h^k} V_{h+1}^\star - b_h^{k^n}\Big) \\
&\quad \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big(V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n}) - V_{h+1}^\star(s_{h+1}^{k^n}) + \big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big) V_{h+1}^\star - b_h^{k^n}\Big),
\end{aligned} \tag{A.77}$$

where the inequality follows from the initialization $Q_h^{\mathsf{LCB},1}(s_h^k, a_h^k) = 0 \le Q_h^\star(s_h^k, a_h^k)$ and the definition of P_h^k in (3.15). To continue, we invoke a result established in Jin et al. (2018, proof of Lemma 4.3),

which guarantees that with probability at least $1-\delta$,

$$\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right) V_{h+1}^\star \lesssim \sqrt{\frac{H^3 \log(\frac{SAT}{\delta})}{N_h^k}} \leq \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} b_h^{k^n},$$

provided that c_{b} is some sufficiently large constant. Substituting the above relation into (A.77) implies that

$$Q_h^{\mathsf{LCB},k+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \leq \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n}) - V_{h+1}^\star(s_{h+1}^{k^n})\right) \leq 0, \tag{A.78}$$

where the last inequality follows from the induction hypothesis

$$V_{h+1}^{\mathsf{LCB},j}(s) \leq V_{h+1}^\star(s) \qquad \text{for all } s\in\mathcal{S} \text{ and } j\leq k.$$

The proof is thus completed by induction.

A.3.2.2 Proof of the inequality (3.22)

The proof of (3.22) essentially follows the same arguments of Yang et al. (2021, Lemma 4.2) (see also Jin et al. (2018, Lemma C.7)), an algebraic result leveraging certain relations w.r.t. the Q-value estimates. Accounting for the difference between our algorithm and the one in Yang et al. (2021), we paraphrase Yang et al. (2021, Lemma 4.2) into the following form that is convenient for our purpose.

Lemma 26 (paraphrased from Lemma 4.2 in Yang et al. (2021))**.** *Assume there exists a constant $c_{\mathrm{b}}>0$ such that for all $(s,a,k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$, it holds that*

$$\begin{aligned} 0 &\leq Q_h^{k+1}(s,a) - Q_h^{\mathsf{LCB},k+1}(s,a) \\ &\leq \eta_0^{N_h^k(s,a)} H + \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n})\right) + 4c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k(s,a)}}. \end{aligned} \tag{A.79}$$

Consider any $\varepsilon\in(0,H]$. Then for all $\beta=1,\ldots,\lceil\log_2\frac{H}{\varepsilon}\rceil$, one has

$$\left|\sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{1}\left(Q_h^k(s_h^k,a_h^k) - Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k) \in \left[2^{\beta-1}\varepsilon, 2^\beta\varepsilon\right)\right)\right| \lesssim \frac{H^6SA\log\frac{SAT}{\delta}}{4^\beta\varepsilon^2}. \tag{A.80}$$

We first show how to justify (3.22) if the inequality (A.80) holds. As can be seen, the fact

(A.80) immediately leads to

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{1}\left(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)>\varepsilon\right)\lesssim\sum_{\beta=1}^{\lceil\log_2\frac{H}{\varepsilon}\rceil}\frac{H^6SA\log\frac{SAT}{\delta}}{4^\beta\varepsilon^2}\le\frac{H^6SA\log\frac{SAT}{\delta}}{2\varepsilon^2}\tag{A.81}$$

as desired.

We now return to justify the claim (A.80), towards which it suffices to demonstrate that (A.79) holds. Lemma 2 and Lemma 3 directly verify the left-hand side of (A.79) since

$$Q_h^k(s,a)\ge Q_h^\star(s,a)\ge Q_h^{\mathsf{LCB},k}(s,a)\qquad\text{for all }(s,a,k,h)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H].\tag{A.82}$$

The remainder of the proof is thus devoted to justifying the upper bound on $Q_h^{k+1}(s,a)-Q_h^{\mathsf{LCB},k+1}(s,a)$ in (A.79). In view of the update rule in line 12 of Algorithm 3, we have the following basic fact

$$Q_h^{k+1}(s,a)\le Q_h^{\mathsf{UCB},k+1}(s,a).$$

This enables us to obtain

$$Q_h^{k+1}(s,a)-Q_h^{\mathsf{LCB},k+1}(s,a)\le Q_h^{\mathsf{UCB},k+1}(s,a)-Q_h^{\mathsf{LCB},k+1}(s,a)=Q_h^{\mathsf{UCB},k^{N_h^k}+1}(s,a)-Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s,a),\tag{A.83}$$

where we abbreviate

$$N_h^k=N_h^k(s,a)$$

throughout this subchapter as long as it is clear from the context. Making use of the update rules of $Q_h^{\mathsf{UCB},k}$ and $Q_h^{\mathsf{LCB},k}$ in line 2 and line 2 of Algorithm 6, we reach

$$\begin{aligned}
&Q_h^{\mathsf{UCB},k^{N_h^k}+1}(s,a)-Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s,a)\\
&=(1-\eta_{N_h^k})Q_h^{\mathsf{UCB},k^{N_h^k}}(s,a)+\eta_{N_h^k}\left(r_h(s,a)+V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})+c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right)\\
&\quad-(1-\eta_{N_h^k})Q_h^{\mathsf{LCB},k^{N_h^k}}(s,a)-\eta_{N_h^k}\left(r_h(s,a)+V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})-c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right)\\
&=(1-\eta_{N_h^k})\left(Q_h^{\mathsf{UCB},k^{N_h^k}}(s,a)-Q_h^{\mathsf{LCB},k^{N_h^k}}(s,a)\right)\\
&\quad+\eta_{N_h^k}\left(V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})-V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})+2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right)\\
&=(1-\eta_{N_h^k})\left(Q_h^{\mathsf{UCB},k^{N_h^k-1}+1}(s,a)-Q_h^{\mathsf{LCB},k^{N_h^k}}(s,a)\right)
\end{aligned}$$

$$+\eta_{N_h^k}\left(V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})-V_{h+1}^{\mathsf{LCB},k^{N_h^k}}(s_{h+1}^{k^{N_h^k}})+2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\right).$$

Applying this relation recursively leads to the desired result

$$\begin{aligned}
&Q_h^{\mathsf{UCB},k^{N_h^k}+1}(s,a)-Q_h^{\mathsf{LCB},k^{N_h^k}+1}(s,a)\\
&\quad=\eta_0^{N_h^k}\left(Q_h^{\mathsf{UCB},1}(s,a)-Q_h^{\mathsf{LCB},1}(s,a)\right)+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n})+2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{n}}\right)\\
&\quad\le\eta_0^{N_h^k}H+\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathsf{LCB},k^n}(s_{h+1}^{k^n})\right)+4c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}.
\end{aligned}$$

Here, the last line is valid due to the property $0\le Q_h^{\mathsf{LCB},1}(s,a)\le Q_h^{\mathsf{UCB},1}(s,a)\le H$ and the following fact

$$\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}}\le 2c_{\mathrm{b}}\sqrt{\frac{H^3\log\frac{SAT}{\delta}}{N_h^k}},$$

which is an immediate consequence of the elementary property $\sum_{n=1}^{N}\frac{\eta_n^N}{\sqrt{n}}\le\frac{2}{\sqrt{N}}$ (see Lemma 1). This combined with (A.83) establishes the condition (A.79), thus concluding the proof of the inequality (3.22).

A.3.3 Proof of Lemma 4

A.3.3.1 Proof of the inequality (3.25)

Consider any state s that has been visited at least once during the K episodes. Throughout this proof, we shall adopt the shorthand notation

$$k^i=k_h^i(s),$$

which denotes the index of the episode in which state s is visited for the i-th time at step h. Given that $V_h(s)$ and $V_h^{\mathsf{R}}(s)$ are only updated during the episodes with indices coming from $\{i\mid 1\le k^i\le K\}$, it suffices to show that for any s and the corresponding $1\le k^i\le K$, the claim (3.25) holds in the sense that

$$\left|V_h^{k^i+1}(s)-V_h^{\mathsf{R},k^i+1}(s)\right|\le 2. \tag{A.84}$$

Towards this end, we look at three scenarios separately.

Case 1. Suppose that k^i obeys

$$V_h^{k^i+1}(s) - V_h^{\mathsf{LCB},k^i+1}(s) > 1 \tag{A.85}$$

or

$$V_h^{k^i+1}(s) - V_h^{\mathsf{LCB},k^i+1}(s) \le 1 \qquad \text{and} \qquad u_{\mathrm{ref}}^{k^i}(s) = \mathsf{True} \tag{A.86}$$

The above conditions correspond to the ones in line 15 and line 17 of Algorithm 3 (meaning that V_h^{R} is updated during the k^i-th episode), thus resulting in

$$V_h^{k^i+1}(s) = V_h^{\mathsf{R},k^i+1}(s).$$

This clearly satisfies (A.84).

Case 2. Suppose that k^{i_0} is the first time such that (A.85) and (A.86) are violated, namely,

$$i_0 \coloneqq \min\left\{ j \mid V_h^{k^j+1}(s) - V_h^{\mathsf{LCB},k^j+1}(s) \le 1 \ \text{ and } \ u_{\mathrm{ref}}^{k^j}(s) = \mathsf{False} \right\}. \tag{A.87}$$

We make three observations.

- The definition (A.87) taken together with the update rules (lines 15-18 of Algorithm 3) reveals that V_h^{R} has been updated in the k^{i_0-1}-th episode, thus indicating that

$$V_h^{\mathsf{R},k^{i_0}}(s) = V_h^{\mathsf{R},k^{i_0-1}+1}(s) = V_h^{k^{i_0-1}+1}(s) = V_h^{k^{i_0}}(s). \tag{A.88}$$

- Additionally, note that under the definition (A.87), $V_h^{\mathsf{R}}(s)$ is not updated during the k^{i_0}-th episode, namely,

$$V_h^{\mathsf{R},k^{i_0}+1}(s) = V_h^{\mathsf{R},k^{i_0}}(s). \tag{A.89}$$

- The definition of k^{i_0} indicates that either (A.85) or (A.86) is satisfied in the previous episode $k^i = k^{i_0-1}$ in which s was visited. If (A.85) is satisfied, then lines 15-16 in Algorithm 3 tell us that

$$\mathsf{True} = u_{\mathrm{ref}}^{k^{i_0-1}+1}(s) = u_{\mathrm{ref}}^{k^{i_0}}(s), \tag{A.90}$$

  which, however, contradicts the assumption $u_{\mathrm{ref}}^{k^{i_0}}(s) = \mathsf{False}$ in (A.87). Therefore, in the k^{i_0-1}-th episode, (A.86) is satisfied, thus leading to

$$V_h^{k^{i_0}}(s) - V_h^{\mathsf{LCB},k^{i_0}}(s) = V_h^{k^{i_0-1}+1}(s) - V_h^{\mathsf{LCB},k^{i_0-1}+1}(s) \le 1. \tag{A.91}$$

We see from (A.88), (A.89) and (A.91) that

$$V_h^{\mathsf{R},k^{i_0}+1}(s) - V_h^{k^{i_0}+1}(s) = V_h^{\mathsf{R},k^{i_0}}(s) - V_h^{k^{i_0}+1}(s) = V_h^{k^{i_0}}(s) - V_h^{k^{i_0}+1}(s) \tag{A.92}$$

$$\overset{\text{(i)}}{\leq} V_h^{k^{i_0}}(s) - V_h^{\mathsf{LCB},k^{i_0}}(s) \overset{\text{(ii)}}{\leq} 1, \tag{A.93}$$

where (i) holds since $V_h^{k^{i_0}+1}(s) \geq V_h^{\star}(s) \geq V_h^{\mathsf{LCB},k^{i_0}}(s)$, and (ii) follows from (A.91). In addition, we make note of the fact that

$$V_h^{\mathsf{R},k^{i_0}+1}(s) - V_h^{k^{i_0}+1}(s) = V_h^{k^{i_0}}(s) - V_h^{k^{i_0}+1}(s) \geq 0, \tag{A.94}$$

which follows from (A.92) and the monotonicity of $V_h^k(s)$ in k. With the above results in place, we arrive at the advertised bound (A.84) when $i = i_0$.

Case 3. Consider any $i > i_0$. It is easily verified that

$$V_h^{k^i+1}(s) - V_h^{\mathsf{LCB},k^i+1}(s) \leq 1 \qquad \text{and} \qquad u_{\text{ref}}^{k^i}(s) = \mathsf{False}. \tag{A.95}$$

It then follows that

$$\begin{aligned} V_h^{\mathsf{R},k^i+1}(s) &\overset{\text{(i)}}{\leq} V_h^{\mathsf{R},k^{i_0}+1}(s) \overset{\text{(ii)}}{\leq} V_h^{k^{i_0}+1}(s) + 1 \overset{\text{(iii)}}{\leq} V_h^{\mathsf{LCB},k^{i_0}+1}(s) + 2 \\ &\overset{\text{(iv)}}{\leq} V_h^{\star}(s) + 2 \overset{\text{(v)}}{\leq} V_h^{k^i+1}(s) + 2. \end{aligned} \tag{A.96}$$

Here, (i) holds due to the monotonicity of V_h^{R} and V_h^k (see line 14 of Algorithm 3), (ii) is a consequence of (A.93), (iii) comes from the definition (A.87) of i_0, (iv) arises since V_h^{LCB} is a lower bound on $V_h^{\star}$ (see Lemma 3), whereas (v) is valid since $V_h^{k^i+1}(s) \geq V_h^{\star}(s)$ (see Lemma 2). In addition, in view of the monotonicity of V_h^k (see line 14 of Algorithm 3) and the update rule in line 16 of Algorithm 3, we know that

$$V_h^{\mathsf{R},k^i+1}(s) \geq V_h^{k^i+1}(s).$$

The preceding two bounds taken collectively demonstrate that

$$0 \leq V_h^{\mathsf{R},k^i+1}(s) - V_h^{k^i+1}(s) \leq 2,$$

thus justifying (A.84) for this case.

Therefore, we have established (A.84)—and hence (3.25)—for all cases.

A.3.3.2 Proof of the inequality (3.26)

Suppose that

$$V_h^{\mathsf{R},k}(s_h^k) - V_h^{\mathsf{R},K}(s_h^k) \neq 0 \tag{A.97}$$

holds for some $k < K$. Then there are two possible scenarios to look at:

(a) *Case 1: the condition in line 15 and line 17 of Algorithm 3 are violated at step h of the k-th episode.* This means that we have

$$V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) \leq 1 \quad \text{and} \quad u_{\text{ref}}^k(s_h^k) = \mathsf{False} \tag{A.98}$$

in this case. Then for any $k' > k$, one necessarily has

$$\begin{cases} V_h^{k'}(s_h^k) - V_h^{\mathsf{LCB},k'}(s_h^k) \leq V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) \leq 1, \\ u_{\text{ref}}^{k'}(s_h^k) = u_{\text{ref}}^k(s_h^k) = \mathsf{False}, \end{cases} \tag{A.99}$$

where the first property makes use of the monotonicity of V_h^k and $V_h^{\mathsf{LCB},k}$ (see (3.17b) and line 14 of Algorithm 3). In turn, Condition (A.99) implies that V_h^{R} will no longer be updated after the k-th episode (see line 15 of Algorithm 3), thus indicating that

$$V_h^{\mathsf{R},k}(s_h^k) = V_h^{\mathsf{R},k+1}(s_h^k) = \cdots = V_h^{\mathsf{R},K}(s_h^k). \tag{A.100}$$

This, however, contradicts the assumption (A.97).

(b) *Case 2: the condition in either line 15 or line 17 of Algorithm 3 is satisfied at step h of the k-th episode.* If this occurs, then the update rule in line 15 of Algorithm 3 implies that

$$V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) > 1, \tag{A.101}$$

or

$$V_h^{k+1}(s_h^k) - V_h^{\mathsf{LCB},k+1}(s_h^k) \leq 1 \quad \text{and} \quad u_{\text{ref}}^k(s_h^k) = \mathsf{True}. \tag{A.102}$$

To summarize, the above argument demonstrates that (A.97) can only occur if either (A.101) or (A.102) holds.

With the above observation in place, we can proceed with the following decomposition:

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathsf{R},k}(s_h^k) - V_h^{\mathsf{R},K}(s_h^k)\right) = \sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathsf{R},k}(s_h^k) - V_h^{\mathsf{R},K}(s_h^k)\right)\mathbb{1}\left(V_h^{\mathsf{R},k}(s_h^k) - V_h^{\mathsf{R},K}(s_h^k) \neq 0\right)$$

$$\leq \sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathsf{R},k}(s_h^k)-V_h^{\mathsf{R},K}(s_h^k)\right)\mathbb{1}\left(V_h^{k+1}(s_h^k)-V_h^{\mathsf{LCB},k+1}(s_h^k)\leq 1 \text{ and } u_{\mathrm{ref}}^k(s_h^k)=\mathsf{True}\right)$$
$$+\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{k}(s_h^k)-V_h^{\mathsf{LCB},k}(s_h^k)\right)\mathbb{1}\left(V_h^{k}(s_h^k)-V_h^{\mathsf{LCB},k}(s_h^k)>1\right)}_{=:\omega}. \tag{A.103}$$

Regarding the first term in (A.103), it is readily seen that for all $s\in\mathcal{S}$,

$$\sum_{k=1}^{K}\mathbb{1}\left(V_h^{k+1}(s)-V_h^{\mathsf{LCB},k+1}(s)\leq 1 \text{ and } u_{\mathrm{ref}}^k(s)=\mathsf{True}\right)\leq 1, \tag{A.104}$$

which arises since, for each $s\in\mathcal{S}$, the above condition is satisfied in at most one episode, owing to the monotonicity property of V_h, V_h^{LCB} and the update rule for u_{ref} in (17). As a result, one has

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathsf{R},k}(s_h^k)-V_h^{\mathsf{R},K}(s_h^k)\right)\mathbb{1}\left(V_h^{k+1}(s_h^k)-V_h^{\mathsf{LCB},k+1}(s_h^k)\leq 1 \text{ and } u_{\mathrm{ref}}^k(s_h^k)=\mathsf{True}\right)\\
&\quad\leq H\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{1}\left(V_h^{k+1}(s_h^k)-V_h^{\mathsf{LCB},k+1}(s_h^k)\leq 1 \text{ and } u_{\mathrm{ref}}^k(s_h^k)=\mathsf{True}\right)\\
&\quad= H\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{1}\left(V_h^{k+1}(s)-V_h^{\mathsf{LCB},k+1}(s)\leq 1 \text{ and } u_{\mathrm{ref}}^k(s)=\mathsf{True}\right)\\
&\quad\leq H\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}1=H^2S,
\end{aligned}$$

where the first inequality holds since $\|V_h^{\mathsf{R},k}-V_h^{\mathsf{R},K}\|_\infty\leq H$. Substitution into (A.103) yields

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathsf{R},k}(s_h^k)-V_h^{\mathsf{R},K}(s_h^k)\right)\leq H^2S+\omega. \tag{A.105}$$

To complete the proof, it boils down to bounding the term ω defined in (A.103). To begin with, note that

$$V_h^{\mathsf{R},K}(s_h^k)\geq V_h^{\star}(s_h^k)\geq V_h^{\mathsf{LCB},k}(s_h^k),$$

where we make use of the optimism of $V_h^{\mathsf{R},K}(s_h^k)$ stated in Lemma 2 (cf. (3.19)) and the pessimism of V_h^{LCB} in Lemma 3 (see (3.21)). As a result, we can obtain

$$\omega=\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{k}(s_h^k)-V_h^{\mathsf{LCB},k}(s_h^k)\right)\mathbb{1}\left(V_h^{k}(s_h^k)-V_h^{\mathsf{LCB},k}(s_h^k)>1\right)$$

$$\leq \sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)\right)\mathbb{1}\left(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)>1\right), \tag{A.106}$$

where the second line arises from the properties $V_h^k(s_h^k)=Q_h^k(s_h^k,a_h^k)$ (given that $a_h^k=\arg\max_a Q_h^k(s_h^k,a)$) as well as the following fact (see line 14 of Algorithm 3)

$$V_h^{\mathsf{LCB},k}(s_h^k)\geq \max_a Q_h^{\mathsf{LCB},k}(s_h^k,a)\geq Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k).$$

Further, let us make note of the following elementary identity

$$Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)=\int_0^\infty \mathbb{1}\Big(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)>t\Big)\mathrm{d}t.$$

This allows us to obtain

$$\begin{aligned}
\omega &\leq \sum_{h=1}^{H}\sum_{k=1}^{K}\left\{\int_0^\infty \mathbb{1}\big(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)>t\big)\mathrm{d}t\right\}\mathbb{1}\left(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)>1\right)\\
&=\int_1^H \sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{1}\Big(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)>t\Big)\mathrm{d}t\\
&\lesssim \int_1^H \frac{H^6SA\log\frac{SAT}{\delta}}{t^2}\mathrm{d}t\lesssim H^6SA\log\frac{SAT}{\delta},
\end{aligned} \tag{A.107}$$

where the last line follows from the property (3.22) in Lemma 3. Combining the above bounds (A.106) and (A.107) with (A.105) establishes

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\left(V_h^{\mathsf{R},k}(s_h^k)-V_h^{\mathsf{R},K}(s_h^k)\right)\\
&\leq H^2S+\sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)\right)\mathbb{1}\left(Q_h^k(s_h^k,a_h^k)-Q_h^{\mathsf{LCB},k}(s_h^k,a_h^k)>1\right)\\
&\leq H^6SA\log\frac{SAT}{\delta}
\end{aligned}$$

as claimed.

A.4 Proof of Lemma 5

For notational simplicity, we shall adopt the short-hand notation

$$k^n=k_h^n(s_h^k,a_h^k)$$

throughout this chapter. A starting point for proving this lemma is the upper bound already derived in (3.30), and we intend to further bound the first term on the right-hand side of (3.30). Recalling the expression of $Q_h^{\mathsf{R},k+1}(s_h^k, a_h^k)$ in (B.67) and (A.24), we can derive

$$
\begin{aligned}
Q_h^{\mathsf{R},k}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) &= Q_h^{\mathsf{R},k^{N_h^{k-1}(s_h^k,a_h^k)}+1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k) \qquad \text{(A.108)}\\
&= \eta_0^{N_h^{k-1}(s_h^k,a_h^k)}\Big(Q_h^{\mathsf{R},1}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)\Big) + \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)} b_h^{\mathsf{R},k^n+1}\\
&\quad + \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\bigg(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \frac{1}{n}\sum_{i=1}^n V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\bigg)\\
&\le \eta_0^{N_h^{k-1}(s_h^k,a_h^k)} H + B_h^{\mathsf{R},k}(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\big(N_h^{k-1}(s_h^k,a_h^k)\big)^{3/4}}\log\frac{SAT}{\delta}\\
&\quad + \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\bigg(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \frac{1}{n}\sum_{i=1}^n V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\bigg),
\end{aligned}
$$

where the last line follows from (B.62) with $B_h^{\mathsf{R},k^{N_h^{k-1}}+1} = B_h^{\mathsf{R},k}$ and the initialization $Q_h^{\mathsf{R},1}(s_h^k,a_h^k) = H$. Summing over all $1 \le k \le K$ gives

$$
\begin{aligned}
&\sum_{k=1}^K \Big(Q_h^{\mathsf{R},k}(s_h^k,a_h^k) - Q_h^\star(s_h^k,a_h^k)\Big)\\
&\le \sum_{k=1}^K \bigg(H\eta_0^{N_h^{k-1}(s_h^k,a_h^k)} + B_h^{\mathsf{R},k}(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\big(N_h^{k-1}(s_h^k,a_h^k)\big)^{3/4}}\log\frac{SAT}{\delta}\bigg)\\
&\quad + \sum_{k=1}^K \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\bigg(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \frac{\sum_{i=1}^n V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i})}{n} - P_{h,s_h^k,a_h^k}V_{h+1}^\star\bigg)\\
&\le \sum_{k=1}^K \bigg(H\eta_0^{N_h^{k-1}(s_h^k,a_h^k)} + B_h^{\mathsf{R},k}(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\big(N_h^{k-1}(s_h^k,a_h^k)\big)^{3/4}}\log\frac{SAT}{\delta}\bigg)\\
&\quad + \sum_{k=1}^K \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\Big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^\star(s_{h+1}^{k^n})\Big)\\
&\quad + \sum_{k=1}^K \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\bigg(V_{h+1}^\star(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \frac{1}{n}\sum_{i=1}^n V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\bigg).
\end{aligned} \qquad \text{(A.109)}
$$

Next, we control each term in (A.109) separately.

- Regarding the first term of (A.109), we make two observations. To begin with,

$$\sum_{k=1}^{K} \eta_0^{N_h^{k-1}(s_h^k,a_h^k)} \le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{n=0}^{N_h^{K-1}(s,a)} \eta_0^n \le SA, \tag{A.110}$$

where the last inequality follows since $\eta_0^n = 0$ for all $n > 0$ (see (4.16)). Next, it is also observed that

$$\begin{aligned}\sum_{k=1}^{K} \frac{1}{\big(N_h^{k-1}(s_h^k,a_h^k)\big)^{3/4}} &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{n=1}^{N_h^{K-1}(s,a)} \frac{1}{n^{3/4}} \\ &\le \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} 4\big(N_h^{K-1}(s,a)\big)^{1/4} \le 4(SA)^{3/4}K^{1/4}, \end{aligned}\tag{A.111}$$

where the last inequality comes from Holder's inequality

$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big(N_h^{K-1}(s,a)\big)^{1/4} \le \left[\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} 1\right]^{3/4} \left[\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_h^{K-1}(s,a)\right]^{1/4} \le (SA)^{3/4}K^{1/4}.$$

Combine the above bounds to yield

$$\begin{aligned}&\sum_{k=1}^{K}\left(H\eta_0^{N_h^{k-1}(s_h^k,a_h^k)} + B_h^{\mathsf{R},k}(s_h^k,a_h^k) + \frac{2c_{\mathrm{b}}H^2}{\big(N_h^{k-1}(s_h^k,a_h^k)\big)^{3/4}}\log\frac{SAT}{\delta}\right) \\ &\qquad \le HSA + \sum_{k=1}^{K} B_h^{\mathsf{R},k}(s_h^k,a_h^k) + 8c_{\mathrm{b}}(SA)^{3/4}K^{1/4}H^2\log\frac{SAT}{\delta}. \end{aligned}\tag{A.112}$$

- We now turn to the second term of (A.109). A little algebra gives

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n})\big) \\ &= \sum_{l=1}^{K}\sum_{N=N_h^l(s_h^l,a_h^l)}^{N_h^{K-1}(s_h^l,a_h^l)} \eta_{N_h^l(s_h^l,a_h^l)}^{N}\big(V_{h+1}^{l}(s_{h+1}^{l}) - V_{h+1}^{\star}(s_{h+1}^{l})\big) \\ &\le \left(1+\frac{1}{H}\right)\sum_{l=1}^{K}\big(V_{h+1}^{l}(s_{h+1}^{l}) - V_{h+1}^{\star}(s_{h+1}^{l})\big) \\ &= \left(1+\frac{1}{H}\right)\left[\sum_{k=1}^{K}\big(V_{h+1}^{k}(s_{h+1}^{k}) - V_{h+1}^{\pi^k}(s_{h+1}^{k})\big) - \sum_{k=1}^{K}\big(V_{h+1}^{\star}(s_{h+1}^{k}) - V_{h+1}^{\pi^k}(s_{h+1}^{k})\big)\right]. \end{aligned}\tag{A.113}$$

Here, the second line replaces k^n (resp. n) with l (resp. $N_h^l(s_h^l, a_h^l)$), the third line is due to the property $\sum_{N=n}^{\infty} \eta_n^N \le 1 + 1/H$ (see Lemma 1), while the last relation replaces l with k again.

- When it comes to the last term of (A.109), we can derive

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)} \left(V_{h+1}^\star(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) + \frac{1}{n}\sum_{i=1}^{n} V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k} V_{h+1}^\star \right) \\
&= \sum_{k=1}^{K} \sum_{n=1}^{N_h^{k-1}(s_h^k,a_h^k)} \eta_n^{N_h^{k-1}(s_h^k,a_h^k)} \left(\big(P_h^{k^n} - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^\star - V_{h+1}^{\mathsf{R},k^n}\big) + \frac{1}{n}\sum_{i=1}^{n} \big(V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k} V_{h+1}^{\mathsf{R},k^n}\big) \right) \\
&= \sum_{k=1}^{K} \sum_{N=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \eta_{N_h^k(s_h^k,a_h^k)}^{N} \left(\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^\star - V_{h+1}^{\mathsf{R},k}\big) + \frac{\sum_{i=1}^{N_h^k(s_h^k,a_h^k)} \left(V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k} V_{h+1}^{\mathsf{R},k} \right)}{N_h^k(s_h^k,a_h^k)} \right).
\end{aligned}$$

  Here, the first equality holds since $V_{h+1}^\star(s_{h+1}^{k^n}) - V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) = P_h^{k^n}\big(V_{h+1}^\star - V_{h+1}^{\mathsf{R},k^n}\big)$ (in view of the definition of P_h^k in (3.15)), the second equality can be seen via simple rearrangement of the terms, while in the last line we replace k^n (resp. n) with k (resp. $N_h^k(s_h^k, a_h^k)$).

Taking the above bounds together with (A.109) and (3.30), we can rearrange terms to reach

$$\begin{aligned}
&\sum_{k=1}^{K} \big(V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)\big) \\
&\le \left(1 + \frac{1}{H}\right) \sum_{k=1}^{K} \big(V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big) + \sum_{k=1}^{K} B_h^{\mathsf{R},k}(s_h^k, a_h^k) \\
&\quad + HSA + 8c_{\mathrm{b}} H^2 (SA)^{3/4} K^{1/4} \log\frac{SAT}{\delta} + \sum_{k=1}^{K} \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) \\
&\quad + \sum_{k=1}^{K} \sum_{N=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \eta_{N_h^k(s_h^k,a_h^k)}^{N} \left[\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^\star - V_{h+1}^{\mathsf{R},k}\big) + \frac{\sum_{i=1}^{N_h^k(s_h^k,a_h^k)} \left(V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i}) - P_{h,s_h^k,a_h^k} V_{h+1}^{\mathsf{R},k} \right)}{N_h^k(s_h^k,a_h^k)} \right],
\end{aligned} \tag{A.114}$$

where we have dropped the term $-\frac{1}{H}\sum_k \big(V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big)$ owing to the fact that $V_{h+1}^\star \ge V_{h+1}^{\pi^k}$.

Thus far, we have established a crucial connection between $\sum_{k=1}^{K} \big(V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k)\big)$ at step h and $\sum_{k=1}^{K} \big(V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)\big)$ at step $h+1$. Clearly, the term $V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k)$ can be further bounded in the same manner. As a result, by recursively applying the above relation (A.114) over the time steps $h = 1, 2, \cdots, H$ and using the terminal condition $V_{H+1}^k = V_{H+1}^{\pi^k} = 0$, we can immediately arrive at the advertised bound in Lemma 5.

A.5 Proof of Lemma 6

A.5.1 Bounding the term $\mathcal{R}_1$

First of all, let us look at the first two terms of $\mathcal{R}_1$ in (3.32a). Recognizing the following elementary inequality

$$\left(1+\frac{1}{H}\right)^{h-1} \le \left(1+\frac{1}{H}\right)^{H} \le e \quad \text{for all } h = 1, 2, \cdots, H+1, \tag{A.115}$$

we are allowed to upper bound the first two terms in (3.32a) as follows:

$$\begin{aligned}\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left\{HSA + 8c_{\mathrm{b}}H^2(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta}\right\} &\lesssim H^2SA + H^3(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta} \\ &\lesssim H^{4.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK} = H^{4.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^2SAT},\end{aligned} \tag{A.116}$$

where the last inequality can be shown using the AM-GM inequality as follows:

$$H^3(SA)^{3/4}K^{1/4}\log\frac{SAT}{\delta} = \left(H^{9/4}\sqrt{SA}\log\frac{SAT}{\delta}\right)\cdot(H^3SAK)^{1/4} \le H^{4.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK}.$$

We are now left with the last term of $\mathcal{R}_1$ in (3.32a). Towards this, we resort to Lemma 25 by setting

$$W_{h+1}^i \coloneqq V_{h+1}^\star - V_{h+1}^{\pi^k} \qquad \text{and} \qquad c_h \coloneqq \left(1+\frac{1}{H}\right)^{h-1}.$$

In view of (B.44) and the property $H \ge V^\star(s) \ge V^\pi(s) \ge 0$, we see that

$$0 \le c_h \le e, \qquad W_{h+1}^i \ge 0, \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}}.$$

Therefore, applying Lemma 25 yields

$$\begin{aligned}\left|\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\left(P_{h,s_h^k,a_h^k} - P_h^k\right)\left(V_{h+1}^\star - V_{h+1}^{\pi^k}\right)\right| &= \left|\sum_{h=1}^{H}\sum_{k=1}^{K}Y_{k,h}\right| \\ &\lesssim \sqrt{TC_{\mathrm{w}}^2\log\frac{1}{\delta}} + C_{\mathrm{w}}\log\frac{1}{\delta} = \sqrt{H^2T\log\frac{1}{\delta}} + H\log\frac{1}{\delta}\end{aligned} \tag{A.117}$$

with probability exceeding $1-\delta$.

Combining (A.116) and (A.117) with the definition (3.32a) of $\mathcal{R}_1$ immediately leads to the claimed bound.

A.5.2 Bounding the term $\mathcal{R}_2$

In view of the definition of $B_h^{\mathsf{R},k}(s_h^k, a_h^k)$ in line 14 of Algorithm 6, we can decompose $\mathcal{R}_2$ (cf. (3.32b)) as follows:

$$\begin{aligned}
\mathcal{R}_2 &= \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} c_{\mathrm{b}}\sqrt{H\log\frac{SAT}{\delta}}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathsf{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \\
&\quad + \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} c_{\mathrm{b}}\sqrt{\log\frac{SAT}{\delta}}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathsf{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \\
&\lesssim \sqrt{H\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathsf{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \\
&\quad + \sqrt{\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathsf{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}},
\end{aligned} \tag{A.118}$$

where the last relation holds due to (B.44). In what follows, we intend to bound these two terms separately.

Step 1: upper bounding the first term in (B.103). Towards this, we make the observation that

$$\begin{aligned}
\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathsf{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} &\le \sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s_h^k,a_h^k)}{N_h^k(s_h^k,a_h^k)}} \\
&= \sum_{k=1}^{K}\sqrt{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\eta_n^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{k^n}(s_{h+1}^{k^n})-V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}},
\end{aligned} \tag{A.119}$$

where the second line follows from the update rule of $\sigma_h^{\mathsf{adv},k}$ in (B.54). Combining the relation $|V_{h+1}^k(s_h^k)-V_{h+1}^{\mathsf{R},k}(s_h^k)|\le 2$ (cf. (3.25)) and the property $\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\eta_n^{N_h^k(s_h^k,a_h^k)}\le 1$ (cf. (4.17)) with (B.104) yields

$$\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathsf{adv},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \le \sum_{k=1}^{K}\sqrt{\frac{4}{N_h^k(s_h^k,a_h^k)}} \le 2\sqrt{SAK}. \tag{A.120}$$

Here, the last inequality holds due to the following fact:

$$\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s_h^k,a_h^k)}}=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\sqrt{\frac{1}{n}}\le 2\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_h^K(s,a)}$$
$$\le 2\sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}1}\cdot\sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}N_h^K(s,a)}=2\sqrt{SAK}, \tag{A.121}$$

where the last line arises from Cauchy-Schwarz and the basic fact that $\sum_{(s,a)}N_h^K(s,a)=K$.

Step 2: upper bounding the second term in (B.103). Recalling the update rules of $\mu_h^{\mathsf{ref},k}$ and $\sigma_h^{\mathsf{ref},k}$ in (B.55), we have

$$\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s_h^k,a_h^k)-\big(\mu_h^{\mathsf{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}}$$
$$=\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s_h^k,a_h^k)}}\underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}-\left(\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})}{N_h^k(s_h^k,a_h^k)}\right)^2}}_{=:J_h^k}. \tag{A.122}$$

Additionally, the quantity J_h^k defined in (B.107) obeys

$$(J_h^k)^2\le\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2-\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}+\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}-\left(\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}V_{h+1}^{\star}(s_{h+1}^{k^n})}{N_h^k(s_h^k,a_h^k)}\right)^2$$
$$\le\underbrace{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}2H\big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})-V_{h+1}^{\star}(s_{h+1}^{k^n})\big)}{N_h^k(s_h^k,a_h^k)}}_{=:J_1}+\underbrace{\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2}{N_h^k(s_h^k,a_h^k)}-\left(\frac{\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}V_{h+1}^{\star}(s_{h+1}^{k^n})}{N_h^k(s_h^k,a_h^k)}\right)^2}_{=:J_2}, \tag{A.123}$$

which arises from the fact that $H\ge V_{h+1}^{\mathsf{R},k^n}\ge V_{h+1}^{\star}\ge 0$ for all $k^n\le K$ and hence

$$\big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})\big)^2-\big(V_{h+1}^{\star}(s_{h+1}^{k^n})\big)^2=\big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})+V_{h+1}^{\star}(s_{h+1}^{k^n})\big)\big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})-V_{h+1}^{\star}(s_{h+1}^{k^n})\big)$$
$$\le 2H\big(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n})-V_{h+1}^{\star}(s_{h+1}^{k^n})\big).$$

With (A.123) in mind, we shall proceed to bound each term in (A.123) separately.

- The first term J_1 can be straightforwardly bounded as follows

$$J_1 = \frac{2H}{N_h^k(s_h^k,a_h^k)}\left(\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\left(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n})\right)\mathbb{1}\left(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n}) \le 3\right) + \Phi_h^k(s_h^k,a_h^k)\right)$$
$$\le 6H + \frac{2H}{N_h^k(s_h^k,a_h^k)}\Phi_h^k(s_h^k,a_h^k), \tag{A.124}$$

where $\Phi_h^k(s_h^k,a_h^k)$ is defined as

$$\Phi_h^k(s_h^k,a_h^k) \coloneqq \sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\left(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n})\right)\mathbb{1}\left(V_{h+1}^{\mathsf{R},k^n}(s_{h+1}^{k^n}) - V_{h+1}^{\star}(s_{h+1}^{k^n}) > 3\right). \tag{A.125}$$

- When it comes to the second term J_2, we claim that

$$J_2 \lesssim \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star}) + H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k(s_h^k,a_h^k)}}, \tag{A.126}$$

which will be justified in Appendix A.5.2.1.

Plugging (A.124) and (B.141) into (A.123) and (B.107) allows one to demonstrate that

$$\begin{aligned}
&\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s_h^k,a_h^k) - \left(\mu_h^{\mathsf{ref},k}(s_h^k,a_h^k)\right)^2}{N_h^k(s_h^k,a_h^k)}} \\
&\lesssim \sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s_h^k,a_h^k)}}\sqrt{H + \frac{H\Phi_h^k(s_h^k,a_h^k)}{N_h^k(s_h^k,a_h^k)} + \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star}) + H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k(s_h^k,a_h^k)}}} \\
&\le \sum_{k=1}^{K}\left(\sqrt{\frac{H}{N_h^k(s_h^k,a_h^k)}} + \frac{\sqrt{H\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} + \sqrt{\frac{\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})}{N_h^k(s_h^k,a_h^k)}} + \frac{H\log^{1/4}\frac{SAT}{\delta}}{\left(N_h^k(s_h^k,a_h^k)\right)^{3/4}}\right) \\
&\lesssim \sqrt{HSAK} + \sum_{k=1}^{K}\frac{\sqrt{H\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} + \sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\star})}{N_h^k(s_h^k,a_h^k)}} + H(SA)^{3/4}\left(K\log\frac{SAT}{\delta}\right)^{1/4},
\end{aligned} \tag{A.127}$$

where the last line follows from (A.121) and (A.111).

Step 3: putting together the preceding results. Finally, the above results in (A.120) and (A.127) taken collectively with (B.103) lead to

$$\begin{aligned}
\mathcal{R}_2 &\lesssim \sqrt{H^3SAK\log\frac{SAT}{\delta}} + \sum_{h=1}^{H}\sqrt{\log\frac{SAT}{\delta}}\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s_h^k,a_h^k) - \big(\mu_h^{\mathsf{ref},k}(s_h^k,a_h^k)\big)^2}{N_h^k(s_h^k,a_h^k)}} \\
&\lesssim \sqrt{H^3SAK\log\frac{SAT}{\delta}} + H^2(SA)^{3/4}K^{1/4}\log^{5/4}\frac{SAT}{\delta} + \sqrt{\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} \\
&\quad + \sqrt{H\log\frac{SAT}{\delta}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\sqrt{\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} \\
&\overset{\text{(i)}}{\lesssim} \sqrt{H^3SAK\log\frac{SAT}{\delta}} + H^2(SA)^{3/4}K^{1/4}\log^{5/4}\frac{SAT}{\delta} + H^4SA\log^2\frac{SAT}{\delta} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{H^3SAK\log\frac{SAT}{\delta}} + H^4SA\log^2\frac{SAT}{\delta} = \sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^4SA\log^2\frac{SAT}{\delta}.
\end{aligned}$$

Here, (i) holds due to the following two claimed inequalities

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} \lesssim \sqrt{H^2SAT\log\frac{SAT}{\delta}} + H^4SA\log\frac{SAT}{\delta}, \tag{A.128}$$

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\sqrt{\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} \lesssim H^{7/2}SA\log^{3/2}\frac{SAT}{\delta}, \tag{A.129}$$

whose proofs are postponed to to Appendix A.5.2.2 and Appendix A.5.2.3, respectively. Additionally, the inequality (ii) above is valid since

$$\begin{aligned}
H^2(SA)^{3/4}K^{1/4}\log^{5/4}\frac{SAT}{\delta} &= \left(H^{5/4}(SA)^{1/2}\log\frac{SAT}{\delta}\right)\cdot\left(H^3SAK\log\frac{SAT}{\delta}\right)^{1/4} \\
&\lesssim H^{2.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK\log\frac{SAT}{\delta}} = H^{2.5}SA\log^2\frac{SAT}{\delta} + \sqrt{H^2SAT\log\frac{SAT}{\delta}}
\end{aligned}$$

due to the Cauchy-Schwarz inequality. This concludes the proof of the advertised upper bound on $\mathcal{R}_2$.

A.5.2.1 Proof of the inequality (B.141)

Akin to the proof of I_4^1 in (A.70), let

$$W_{h+1}^i \coloneqq (V_{h+1}^\star)^2 \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \frac{1}{N}.$$

By observing and setting

$$C_{\mathrm{u}} := \frac{1}{N}, \qquad \|W_{h+1}^i\|_\infty \le H^2 =: C_{\mathrm{w}},$$

we can apply Lemma 24 to yield

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(V_{h+1}^\star(s_{h+1}^{k^n})\right)^2 - P_{h,s_h^k,a_h^k}(V_{h+1}^\star)^2\right| = \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^\star\right)^2\right| \lesssim H^2\sqrt{\frac{\log^2\frac{SAT}{\delta}}{N_h^k}}$$

with probability at least $1-\delta$. Similarly, by applying the trivial bound $\|V_{h+1}^\star\|_\infty \le H$ and Lemma 24, we can obtain

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right| = \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s_h^k,a_h^k}\right)V_{h+1}^\star\right| \lesssim H\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k}}$$

with probability at least $1-\delta$.

Recalling from (B.51) the definition

$$\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star) = P_{h,s_h^k,a_h^k}(V_{h+1}^\star)^2 - \left(P_{h,s_h^k,a_h^k}V_{h+1}^\star\right)^2,$$

we can use the preceding two bounds and the triangle inequality to show that:

$$\begin{aligned}
&\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})^2 - \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})\right)^2 - \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)\right| \\
&\quad\le \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})^2 - P_{h,s_h^k,a_h^k}(V_{h+1}^\star)^2\right| + \left|\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n})\right)^2 - (P_{h,s_h^k,a_h^k}V_{h+1}^\star)^2\right| \\
&\quad\lesssim H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k}} + \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n}) - P_{h,s_h^k,a_h^k}V_{h+1}^\star\right|\cdot\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}V_{h+1}^\star(s_{h+1}^{k^n}) + P_{h,s_h^k,a_h^k}V_{h+1}^\star\right| \\
&\quad\lesssim H^2\sqrt{\frac{\log\frac{SAT}{\delta}}{N_h^k}}
\end{aligned}$$

with probability at least $1-\delta$, where the last line also makes use of the fact that $\|V_{h+1}^\star\|_\infty \le H$.

A.5.2.2 Proof of the inequality (B.144)

To begin with, we make the observation that

$$\sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{N_h^K(s,a)}\sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{n}} \le 2\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_h^K(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star)},$$

which relies on the fact that $\sum_{n=1}^{N} 1/\sqrt{n} \leq 2\sqrt{N}$. It then follows that

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} &\leq 2\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{N_h^K(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star)} \\
&\leq 2\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} 1}\cdot\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_h^K(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star)} \\
&= 2\sqrt{HSA}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}\,, \qquad \text{(A.130)}
\end{aligned}$$

where the second inequality invokes the Cauchy-Schwarz inequality.

The rest of the proof is then dedicated to bounding (A.130). Towards this end, we first decompose

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{k=1}^{K}\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star) &\leq \sum_{h=1}^{H}\sum_{k=1}^{K}\mathsf{Var}_{h,s_h^k,a_h^k}\big(V_{h+1}^{\pi^k}\big) + \sum_{h=1}^{H}\sum_{k=1}^{K}\left|\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star) - \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})\right| \\
&\overset{\text{(ii)}}{\lesssim} HT + H^3\log\frac{SAT}{\delta} + \sum_{h=1}^{H}\sum_{k=1}^{K}\left|\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star) - \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})\right|, \qquad \text{(A.131)}
\end{aligned}$$

where (ii) follows directly from Jin et al. (2018, Lemma C.5). The second term on the right-hand side of (A.131) can be bounded as follows

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{k=1}^{K}\left|\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star) - \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})\right| \\
&= \sum_{h=1}^{H}\sum_{k=1}^{K}\left|P_{h,s_h^k,a_h^k}(V_{h+1}^\star)^2 - \big(P_{h,s_h^k,a_h^k}V_{h+1}^\star\big)^2 - P_{h,s_h^k,a_h^k}(V_{h+1}^{\pi^k})^2 + \big(P_{h,s_h^k,a_h^k}V_{h+1}^{\pi^k}\big)^2\right| \\
&\leq \sum_{h=1}^{H}\sum_{k=1}^{K}\left\{\left|P_{h,s_h^k,a_h^k}\Big(\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big)\big(V_{h+1}^\star + V_{h+1}^{\pi^k}\big)\Big)\right| + \left|\big(P_{h,s_h^k,a_h^k}V_{h+1}^\star\big)^2 - \big(P_{h,s_h^k,a_h^k}V_{h+1}^{\pi^k}\big)^2\right|\right\} \\
&\overset{\text{(i)}}{\leq} 4H\sum_{h=1}^{H}\sum_{k=1}^{K}P_{h,s_h^k,a_h^k}\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big) \\
&= 4H\sum_{h=1}^{H}\sum_{k=1}^{K}\Big\{V_{h+1}^\star(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k) + \big(P_{h,s_h^k,a_h^k} - P_h^k\big)\big(V_{h+1}^\star - V_{h+1}^{\pi^k}\big)\Big\} \\
&\overset{\text{(ii)}}{\leq} 4H\sum_{h=1}^{H}\sum_{k=1}^{K}\big(\phi_{h+1}^k + \delta_{h+1}^k\big) \overset{\text{(iii)}}{\lesssim} H^2\sqrt{T\log\frac{SAT}{\delta}} + H^4\sqrt{SAT\log\frac{SAT}{\delta}} + H^4SA
\end{aligned}$$

$$\asymp H^4\sqrt{SAT\log\frac{SAT}{\delta}}+H^4SA, \tag{A.132}$$

where we define

$$\delta_{h+1}^k \coloneqq V_{h+1}^{\mathsf{UCB},k}(s_{h+1}^k)-V_{h+1}^{\pi^k}(s_{h+1}^k), \qquad \phi_{h+1}^k \coloneqq \big(P_{h,s_h^k,a_h^k}-P_h^k\big)\big(V_{h+1}^\star-V_{h+1}^{\pi^k}\big). \tag{A.133}$$

We shall take a moment to explain how we derive (A.132). The inequality (i) holds by observing that $V_{h+1}^\star - V_{h+1}^{\pi^k}\geq 0$ and

$$\begin{aligned}
\Big|P_{h,s_h^k,a_h^k}\Big(\big(V_{h+1}^\star-V_{h+1}^{\pi^k}\big)\big(V_{h+1}^\star+V_{h+1}^{\pi^k}\big)\Big)\Big| &\leq P_{h,s_h^k,a_h^k}\big(V_{h+1}^\star-V_{h+1}^{\pi^k}\big)\Big(\big\|V_{h+1}^\star\big\|_\infty+\big\|V_{h+1}^{\pi^k}\big\|_\infty\Big)\\
&\leq 2HP_{h,s_h^k,a_h^k}\big(V_{h+1}^\star-V_{h+1}^{\pi^k}\big),\\
\Big|\big(P_{h,s_h^k,a_h^k}V_{h+1}^\star\big)^2-\big(P_{h,s_h^k,a_h^k}V_{h+1}^{\pi^k}\big)^2\Big| &\leq \Big|P_{h,s_h^k,a_h^k}\big(V_{h+1}^\star-V_{h+1}^{\pi^k}\big)\Big|\cdot\Big|P_{h,s_h^k,a_h^k}\big(V_{h+1}^\star+V_{h+1}^{\pi^k}\big)\Big|\\
&\leq 2HP_{h,s_h^k,a_h^k}\big(V_{h+1}^\star-V_{h+1}^{\pi^k}\big);
\end{aligned}$$

(ii) is valid since $V_{h+1}^{\mathsf{UCB}}\geq V_{h+1}^\star$; and (iii) results from the following two bounds:

$$\sum_{h=1}^H\sum_{k=1}^K\delta_{h+1}^k\lesssim H^3\sqrt{SAT\log\frac{SAT}{\delta}}+H^3SA, \tag{A.134a}$$

$$\sum_{h=1}^H\sum_{k=1}^K\phi_{h+1}^k\lesssim H\sqrt{T\log\frac{SAT}{\delta}}, \tag{A.134b}$$

which come respectively from Jin et al. (2018, Eqn. (C.13)) and the argument for Jin et al. (2018, Eqn. (C.12)).[1]

As a consequence, substituting (A.131) and (A.132) into (A.130), we reach

$$\begin{aligned}
\sum_{h=1}^H\sum_{k=1}^K\sqrt{\frac{\mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)}{N_h^k(s_h^k,a_h^k)}} &\lesssim \sqrt{HSA}\sqrt{HT+H^4\sqrt{SAT\log\frac{SAT}{\delta}}+H^4SA}\\
&\lesssim \sqrt{H^2SAT}+H^{5/2}(SA)^{3/4}\Big(T\log\frac{SAT}{\delta}\Big)^{1/4}+H^{2.5}SA\\
&= \sqrt{H^2SAT}+\Big(H^2SAT\log\frac{SAT}{\delta}\Big)^{1/4}\big(H^4SA\big)^{1/2}+H^{2.5}SA\\
&\lesssim \sqrt{H^2SAT\log\frac{SAT}{\delta}}+H^4SA\log\frac{SAT}{\delta},
\end{aligned}$$

where we have applied the basic inequality $2ab\leq a^2+b^2$ for any $a,b\geq 0$.

[1] Note that the notation δ_h^k used in Jin et al. (2018, Section C.2) and the one in the proof of Jin et al. (2018, Theorem 1) are different; here, we need to adopt the notation used in the proof of Jin et al. (2018, Theorem 1).

A.5.2.3 Proof of the inequality (A.129)

First, it is observed that

$$\sum_{k=1}^{K} \frac{\sqrt{\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{n=1}^{N_h^K(s,a)} \frac{\sqrt{\Phi_h^{k^n(s,a)}(s,a)}}{n}$$
$$\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{\Phi_h^{N_h^K(s,a)}(s,a)}\log T \leq \sqrt{SA\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Phi_h^{N_h^K(s,a)}(s,a)}\log T. \quad \text{(A.135)}$$

Here, the first inequality holds by the monotonicity property of $\Phi_h^k(s_h,a_h)$ with respect to k (see its definition in (A.125)) due to the same property of $V_{h+1}^{\mathsf{R},k}$, while the second inequality comes from Cauchy-Schwarz.

To continue, note that

$$\sum_{h=1}^{H}\sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Phi_h^{N_h^K(s,a)}(s,a)} = \sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}\left(V_{h+1}^{\mathsf{R},k}(s_{h+1}^k) - V_{h+1}^{\star}(s_{h+1}^k)\right)\mathbb{1}\left(V_{h+1}^{\mathsf{R},k}(s_{h+1}^k) - V_{h+1}^{\star}(s_{h+1}^k) > 3\right)}$$
$$\leq \sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}\left(V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right)\mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 3\right)}$$
$$= \sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}\left(V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right)\mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right)}$$
$$\leq \sum_{h=1}^{H}\sqrt{\sum_{k=1}^{K}3\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right)\mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right)}$$
$$\leq \sqrt{H}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}3\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)\right)\mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right)}, \quad \text{(A.136)}$$

where the first inequality follows from Lemma 4 (cf. (3.25)) and Lemma 3 (so that $V_{h+1}^{\mathsf{R},k}(s_{h+1}^k) - V_{h+1}^{\star}(s_{h+1}^k) \leq V_{h+1}^{k}(s_{h+1}^k) + 2 - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)$), the penultimate inequality holds since $1 \leq V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k)$ when $\mathbb{1}\left(V_{h+1}^{k}(s_{h+1}^k) - V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^k) > 1\right) \neq 0$, and the last inequality is a consequence of the Cauchy-Schwarz inequality.

Combining the above relation with (A.106) and applying the triangle inequality, we can demonstrate that

$$\sum_{h=1}^{H}\sqrt{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Phi_h^{N_h^K(s,a)}(s,a)}$$

$$\lesssim \sqrt{H}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\left(Q_{h+1}^{k}(s_{h+1}^{k},a_{h+1}^{k})-Q_{h+1}^{\mathsf{LCB},k}(s_{h+1}^{k},a_{h+1}^{k})\right)\mathbb{1}\left(Q_{h+1}^{k}(s_{h+1}^{k},a_{h+1}^{k})-Q_{h+1}^{\mathsf{LCB},k}(s_{h+1}^{k},a_{h+1}^{k})>1\right)}$$
$$\lesssim \sqrt{H^7SA\log\frac{SAT}{\delta}},$$

where the second inequality follows directly from (3.26), and the first inequality is valid since

$$V_{h+1}^{k}(s_{h+1}^{k})-V_{h+1}^{\mathsf{LCB},k}(s_{h+1}^{k})\le Q_{h+1}^{k}(s_{h+1}^{k},a_{h+1}^{k})-Q_{h+1}^{\mathsf{LCB},k}(s_{h+1}^{k},a_{h+1}^{k}).$$

Substitution into (A.135) gives

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\sqrt{\Phi_h^k(s_h^k,a_h^k)}}{N_h^k(s_h^k,a_h^k)}\lesssim\left(\sqrt{SA}\log T\right)\cdot\sqrt{H^7SA\log\frac{SAT}{\delta}}\asymp H^{7/2}SA\log^{3/2}\frac{SAT}{\delta},$$

thus concluding the proof.

A.5.3 Bounding the term $\mathcal{R}_3$

For notational convenience, we shall use the short-hand notation

$$k^i\coloneqq k_h^i(s_h^k,a_h^k)$$

whenever it is clear from the context. This allows us to decompose the expression of $\mathcal{R}_3$ in (3.32c) as follows

$$\mathcal{R}_3\coloneqq\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_h^k\left(P_h^k-P_{h,s_h^k,a_h^k}\right)\left(V_{h+1}^{\star}-V_{h+1}^{\mathsf{R},k}\right)}_{=:\mathcal{R}_3^1}+\underbrace{\sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_h^k\frac{\sum_{i\le N_h^k(s_h^k,a_h^k)}\left(V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i})-P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},k}\right)}{N_h^k(s_h^k,a_h^k)}}_{=:\mathcal{R}_3^2}$$

with

$$\lambda_h^k\coloneqq\left(1+\frac{1}{H}\right)^{h-1}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\eta^n_{N_h^k(s_h^k,a_h^k)}\le\left(1+\frac{1}{H}\right)^{h}\le\left(1+\frac{1}{H}\right)^{H}\le e. \tag{A.137}$$

Here, the first inequality in (A.137) follows from the property $\sum_{N=n}^{\infty}\eta_n^N\le 1+1/H$ in Lemma 1, while the last inequality in (A.137) results from (B.44). In the sequel, we shall control each of these two terms separately.

Step 1: upper bounding $\mathcal{R}_3^1$. We plan to control this term by means of Lemma 25. For notational simplicity, let us define

$$N(s,a,h) \coloneqq N_h^{K-1}(s,a)$$

and set

$$W_{h+1}^i \coloneqq V_{h+1}^{\mathsf{R},k} - V_{h+1}^\star \qquad \text{and} \qquad u_h^i(s_h^i,a_h^i) \coloneqq \lambda_h^i = \left(1+\frac{1}{H}\right)^{h-1} \sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)} \eta_{N_h^i(s_h^i,a_h^i)}^n.$$

Given the fact that $V_{h+1}^{\mathsf{R},k}(s), V_{h+1}^\star(s) \in [0,H]$ and the condition (A.137), it is readily seen that

$$\left|u_h^i(s_h^i,a_h^i)\right| \le e =: C_{\mathrm{u}} \qquad \text{and} \qquad \left\|W_{h+1}^i\right\|_\infty \le H =: C_{\mathrm{w}}.$$

Apply Lemma 25 to yield

$$\begin{aligned}
\left|\sum_{h=1}^H\sum_{k=1}^K \lambda_h^k\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^\star - V_{h+1}^{\mathsf{R},k}\big)\right| &= \left|\sum_{h=1}^H\sum_{k=1}^K X_{k,h}\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 C_{\mathrm{w}} HSA \sum_{h=1}^H\sum_{i=1}^K \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right]\log\frac{K}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}HSA\log\frac{K}{\delta} \\
&\lesssim \sqrt{H^2SA\sum_{h=1}^H\sum_{k=1}^K \mathbb{E}_{i,h-1}\left[P_h^k\big(V_{h+1}^{\mathsf{R},k} - V_{h+1}^\star\big)\right]\log\frac{T}{\delta}} + H^2SA\log\frac{T}{\delta} \\
&\asymp \sqrt{H^2SA\left\{\sum_{h=1}^H\sum_{k=1}^K P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},k} - V_{h+1}^\star\big)\right\}\log\frac{T}{\delta}} + H^2SA\log\frac{T}{\delta}
\end{aligned} \tag{A.138}$$

with probability at least $1-\delta/2$.

It then comes down to controlling the sum $\sum_{h=1}^H\sum_{k=1}^K P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},k} - V_{h+1}^\star\big)$. Towards this end, we first single out the following useful fact:

$$\begin{aligned}
\sum_{h=1}^H\sum_{k=1}^K P_h^k\big(V_{h+1}^{\mathsf{R},k} - V_{h+1}^\star\big) &\overset{\text{(i)}}{\le} \sum_{h=1}^H\sum_{k=1}^K P_h^k\big(V_{h+1}^k + 2 - V_{h+1}^\star\big) \\
&\le 2HK + \sum_{h=1}^H\sum_{k=1}^K\Big(V_{h+1}^k(s_{h+1}^k) - V_{h+1}^\star(s_{h+1}^k)\Big) \overset{\text{(ii)}}{\lesssim} \sqrt{H^7SAK\log\frac{SAT}{\delta}} + H^3SA + HK
\end{aligned} \tag{A.139}$$

with probability at least $1-\delta/4$, where (i) holds according to (3.25), and (ii) is valid since

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\Big(V_{h+1}^{k}(s_{h+1}^{k})-V_{h+1}^{\star}(s_{h+1}^{k})\Big)\le\sum_{h=1}^{H}\sum_{k=1}^{K}\Big(V_{h+1}^{\mathsf{UCB},k}(s_{h+1}^{k})-V_{h+1}^{\pi^k}(s_{h+1}^{k})\Big)$$
$$\lesssim\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA,$$

where the first inequality follows since $V_{h+1}^{\mathsf{UCB},k}\ge V_{h+1}^{k}$ and $V_{h+1}^{\star}\ge V_{h+1}^{\pi^k}$, and the second inequality comes from (A.134a). Additionally, invoking Freedman's inequality (see Lemma 25) with $c_h=1$ and $\widetilde{W}_h^i=V_{h+1}^{\mathsf{R},k}-V_{h+1}^{\star}$ (so that $0\le\widetilde{W}_h^i(s)\le H$) directly leads to

$$\Big|\sum_{h=1}^{H}\sum_{k=1}^{K}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathsf{R},k}-V_{h+1}^{\star}\big)\Big|\lesssim\sqrt{TH^2\log\frac{1}{\delta}}+H\log\frac{1}{\delta}\asymp\sqrt{H^3K\log\frac{1}{\delta}}$$

with probability at least $1-\delta/4$, which taken collectively with (A.139) reveals that

$$\sum_{h=1}^{H}\sum_{k=1}^{K}P_{s_h^k,a_h^k,h}\big(V_{h+1}^{\mathsf{R},k}-V_{h+1}^{\star}\big)\le\sum_{h=1}^{H}\sum_{k=1}^{K}P_h^k\big(V_{h+1}^{\mathsf{R},k}-V_{h+1}^{\star}\big)+\Big|\sum_{h=1}^{H}\sum_{k=1}^{K}\big(P_h^k-P_{s_h^k,a_h^k,h}\big)\big(V_{h+1}^{\mathsf{R},k}-V_{h+1}^{\star}\big)\Big|$$
$$\lesssim\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA+HK \tag{A.140}$$

with probability at least $1-\delta/2$. Substitution into (A.138) then gives

$$\Big|\sum_{h=1}^{H}\sum_{k=1}^{K}\lambda_h^k\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\star}-V_{h+1}^{\mathsf{R},k}\big)\Big|$$
$$\lesssim\sqrt{H^2SA\sum_{h=1}^{H}\sum_{k=1}^{K}P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},k}-V_{h+1}^{\star}\big)\log\frac{T}{\delta}}+H^2SA\log\frac{T}{\delta}$$
$$\lesssim\sqrt{H^2SA\left(\sqrt{H^7SAK\log\frac{SAT}{\delta}}+H^3SA+HK\right)\log\frac{T}{\delta}}+H^2SA\log\frac{T}{\delta}$$
$$\asymp\sqrt{H^2SA\left(H^6SA\log\frac{SAT}{\delta}+H^3SA+HK\right)\log\frac{T}{\delta}}+H^2SA\log\frac{T}{\delta}$$
$$\lesssim\sqrt{H^3SAK\log\frac{SAT}{\delta}}+H^4SA\log\frac{SAT}{\delta}$$
$$=\sqrt{H^2SAT\log\frac{SAT}{\delta}}+H^4SA\log\frac{SAT}{\delta} \tag{A.141}$$

with probability exceeding $1-\delta$, where the third line holds since (due to Cauchy-Schwarz)

$$\sqrt{H^7SAK\log\frac{SAT}{\delta}}=\sqrt{H^6SA\log\frac{SAT}{\delta}}\sqrt{HK}\lesssim H^6SA\log\frac{SAT}{\delta}+HK.$$

Step 2: upper bounding $\mathcal{R}_3^2$. We start by making the following observation:

$$\begin{aligned}
\mathcal{R}_3^2 &\le \sum_{h=1}^{H}\sum_{k=1}^{K}\frac{\lambda_h^k}{N_h^k(s_h^k,a_h^k)}\sum_{i\le N_h^k(s_h^k,a_h^k)}\big(V_{h+1}^{\mathsf{R},k^i}(s_{h+1}^{k^i})-P_{h,s_h^k,a_h^k}V_{h+1}^{\mathsf{R},K}\big)\\
&=\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\Big(V_{h+1}^{\mathsf{R},k}(s_{h+1}^k)-V_{h+1}^{\mathsf{R},K}(s_{h+1}^k)+\big(P_h^k-P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\mathsf{R},K}\Big)\\
&\le(e\log T)\sum_{h=1}^{H}\sum_{k=1}^{K}\big(V_{h+1}^{\mathsf{R},k}(s_{h+1}^k)-V_{h+1}^{\mathsf{R},K}(s_{h+1}^k)\big)+\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)V_{h+1}^{\star}\\
&\quad+\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathsf{R},K}-V_{h+1}^{\star}\big),
\end{aligned}\tag{A.142}$$

where the first inequality comes from the monotonicity property $V_{h+1}^{\mathsf{R},k}\ge V_{h+1}^{\mathsf{R},k+1}\ge\cdots\ge V_{h+1}^{\mathsf{R},K}$, and the last line follows from the facts that $\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{1}{n}\le\log T$ and $\lambda_h^k\le e$ (cf. (A.137)). In what follows, we shall control the three terms in (A.142) separately.

- The first term in (A.142) can be controlled by Lemma 4 (cf. (3.26)) as follows:

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\big(V_{h+1}^{\mathsf{R},k}(s_{h+1}^k)-V_{h+1}^{\mathsf{R},K}(s_{h+1}^k)\big)\lesssim H^6SA\log\frac{SAT}{\delta}\tag{A.143}$$

  with probability at least $1-\delta/3$.

- To control the second term in (A.142), we abuse the notation by setting

$$N(s,a,h)\coloneqq N_h^{K-1}(s,a)$$

  and

$$W_{h+1}^i\coloneqq V_{h+1}^{\star},\qquad\text{and}\qquad u_h^i(s_h^i,a_h^i)\coloneqq\sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)}\frac{\lambda_h^i}{n},$$

which clearly satisfy

$$\left|u_h^i(s_h^i,a_h^i)\right| \le e\sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)} \frac{1}{n} \le e\log T =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}}.$$

Here, we have made use of the properties $\sum_{n=N_h^i(s_h^i,a_h^i)}^{N_h^{K-1}(s_h^i,a_h^i)} \frac{1}{n} \le \log T$ and $\lambda_h^k \le e$ (cf. (A.137)). With these in place, applying Lemma 25 reveals that

$$\begin{aligned}
&\left|\sum_{h=1}^H\sum_{k=1}^K \sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)V_{h+1}^\star\right| = \left|\sum_{h=1}^H\sum_{k=1}^K X_{k,h}\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 HSA\sum_{h=1}^H\sum_{i=1}^K \mathbb{E}_{i,h-1}\left[\left|(P_h^i - P_{h,s_h^i,a_h^i})W_{h+1}^i\right|^2\right]\log\frac{T}{\delta}} + C_{\mathrm{u}}C_{\mathrm{w}}HSA\log\frac{T}{\delta} \\
&\overset{\text{(i)}}{\lesssim} \sqrt{\sum_{h=1}^H\sum_{k=1}^K \mathsf{Var}_{h,s_h^k,a_h^k}(V_{h+1}^\star)\cdot HSA\log^3\frac{T}{\delta}} + H^2SA\log^2\frac{T}{\delta} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{HSA\big(HT + H^4\sqrt{SAT}\big)\log^4\frac{SAT}{\delta}} + H^2SA\log^2\frac{T}{\delta} \\
&\lesssim \sqrt{HSA\big(HT + H^7SA\big)\log^4\frac{SAT}{\delta}} + H^2SA\log^2\frac{T}{\delta} \\
&\overset{\text{(iii)}}{\lesssim} \sqrt{H^2SAT\log^4\frac{SAT}{\delta}} + H^4SA\log^2\frac{SAT}{\delta}
\end{aligned} \tag{A.144}$$

with probability at least $1-\delta/3$. Here, (i) comes from the definition in (B.51), (ii) holds due to (A.131) and (A.132), whereas (iii) is valid since

$$HT + H^4\sqrt{SAT} = HT + \sqrt{H^7SA}\cdot\sqrt{HT} \lesssim HT + H^7SA$$

due to the Cauchy-Schwarz inequality.

- Turning attention the third term of (A.142), we need to properly cope with the dependency between P_h^k and $V_{h+1}^{\mathsf{R},K}$. Towards this, we shall resort to the standard epsilon-net argument (see, e.g., (Tao, 2012)), which will be presented in Appendix A.5.3.1. The final bound reads like

$$\left|\sum_{h=1}^H\sum_{k=1}^K \sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathsf{R},K} - V_{h+1}^\star\big)\right| \lesssim H^4SA\log^2\frac{SAT}{\delta} + \sqrt{H^3SAK\log^3\frac{SAT}{\delta}}. \tag{A.145}$$

- Combining (A.143), (A.144) and (A.145) with (A.142), we can use the union bound to demonstrate that

$$\mathcal{R}_3^2 \le C_{3,2}\left\{H^6SA\log^3\frac{SAT}{\delta}+\sqrt{H^2SAT\log^4\frac{SAT}{\delta}}\right\} \tag{A.146}$$

with probability at least $1-\delta$, where $C_{3,2}>0$ is some constant.

Step 3: final bound of $\mathcal{R}_3$. Putting the above results (A.141) and (A.146) together, we immediately arrive at

$$\mathcal{R}_3 \le \left|\mathcal{R}_3^1\right| + \mathcal{R}_3^2 \le C_{\mathrm{r},3}\left\{H^6SA\log^3\frac{SAT}{\delta}+\sqrt{H^2SAT\log^4\frac{SAT}{\delta}}\right\} \tag{A.147}$$

with probability at least $1-2\delta$, where $C_{\mathrm{r},3}>0$ is some constant. This immediately concludes the proof.

A.5.3.1 Proof of (A.145)

Step 1: concentration bounds for a fixed group of vectors. Consider a fixed group of vectors $\{V_{h+1}^{\mathrm{d}}\in\mathbb{R}^S \mid 1\le h\le H\}$ obeying the following properties:

$$V_{h+1}^\star \le V_{h+1}^{\mathrm{d}} \le H \qquad \text{for } 1\le h\le H. \tag{A.148}$$

We intend to control the following sum

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{d}}-V_{h+1}^\star\big).$$

To do so, we shall resort to Lemma 25. For the moment, let us take $N(s,a,h)\coloneqq N_h^{K-1}(s,a)$ and

$$W_{h+1}^i \coloneqq V_{h+1}^{\mathrm{d}}-V_{h+1}^\star, \qquad u_h^i(s_h^i,a_h^i)\coloneqq\sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)}\frac{\lambda_h^i}{n}.$$

It is easily seen that

$$\left|u_h^i(s_h^i,a_h^i)\right| \le e\sum_{n=N_h^i(s_h^i,a_h^i)}^{N(s_h^i,a_h^i,h)}\frac{1}{n}\le e\log T =: C_{\mathrm{u}} \qquad \text{and} \qquad \|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}},$$

which hold due to the facts $\sum_{n=N_h^i(s_h^i,a_h^i)}^{N_h^K(s_h^i,a_h^i)} \frac{1}{n} \le \log T$ and $\lambda_h^k \le e$ (cf. (A.137)) as well as the property that $V_{h+1}^{\mathrm{d}}(s), V_{h+1}^{\star}(s) \in [0,H]$. Thus, invoking Lemma 25 yields

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{d}} - V_{h+1}^{\star}\big)\right| = \left|\sum_{h=1}^{H}\sum_{k=1}^{K} X_{k,h}\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}^2 C_{\mathrm{w}} \sum_{h=1}^{H}\sum_{i=1}^{K} \mathbb{E}_{i,h-1}\left[P_h^i W_{h+1}^i\right] \log\frac{K^{HSA}}{\delta_0}} + C_{\mathrm{u}} C_{\mathrm{w}} \log\frac{K^{HSA}}{\delta_0} \\
&\lesssim \sqrt{H\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{d}} - V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{K^{HSA}}{\delta_0}} + H\big(\log T\big)\log\frac{K^{HSA}}{\delta_0}
\end{aligned} \tag{A.149}$$

with probability at least $1-\delta_0$, where the choice of δ_0 will be revealed momentarily.

Step 2: constructing and controlling an epsilon net. Our argument in Step 1 is only applicable to a fixed group of vectors. The next step is then to construct an epsilon net that allows one to cover the set of interest. Specifically, let us construct an epsilon net $\mathcal{N}_{h+1,\alpha}$ (the value of α will be specified shortly) for each $h \in [H]$ such that:

a) for any $V_{h+1} \in [0,H]^S$, one can find a point $V_{h+1}^{\mathrm{net}} \in \mathcal{N}_{h+1,\alpha}$ obeying

$$0 \le V_{h+1}(s) - V_{h+1}^{\mathrm{net}}(s) \le \alpha \qquad \text{for all } s \in \mathcal{S};$$

b) its cardinality obeys

$$\left|\mathcal{N}_{h+1,\alpha}\right| \le \Big(\frac{H}{\alpha}\Big)^S. \tag{A.150}$$

Clearly, this also means that

$$\left|\mathcal{N}_{2,\alpha} \times \mathcal{N}_{3,\alpha} \times \cdots \times \mathcal{N}_{H+1,\alpha}\right| \le \Big(\frac{H}{\alpha}\Big)^{SH}.$$

Set $\delta_0 = \frac{1}{6}\delta/\big(\frac{H}{\alpha}\big)^{SH}$. Taking (A.149) together the union bound implies that: with probability at least $1 - \delta_0\big(\frac{H}{\alpha}\big)^{SH} = 1 - \delta/6$, one has

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{net}} - V_{h+1}^{\star}\big)\right| \\
&\lesssim \sqrt{H\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{net}} - V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{K^{HSA}}{\delta_0}} + H(\log T)\log\frac{K^{HSA}}{\delta_0}
\end{aligned}$$

$$\lesssim \sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{net}}-V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{SAT}{\delta\alpha}} + H^2SA\log^2\frac{SAT}{\delta\alpha} \tag{A.151}$$

simultaneously for all $\{V_{h+1}^{\mathrm{net}} \mid 1\le h\le H\}$ obeying $V_{h+1}^{\mathrm{d}}\in\mathcal{N}_{h+1,\alpha}$ $(h\in[H])$.

Step 3: obtaining uniform bounds. We are now positioned to establish a uniform bound over the entire set of interest. Consider an arbitrary group of vectors $\{V_{h+1}^{\mathrm{u}}\in\mathbb{R}^S \mid 1\le h\le H\}$ obeying (A.148). By construction, one can find a group of points $\left\{V_{h+1}^{\mathrm{net}}\in\mathcal{N}_{h+1,\alpha} \mid h\in[H]\right\}$ such that

$$0\le V_{h+1}^{\mathrm{u}}(s)-V_{h+1}^{\mathrm{net}}(s)\le\alpha \qquad \text{for all } (h,s)\in\mathcal{S}\times[H]. \tag{A.152}$$

It is readily seen that

$$\begin{aligned}
&\left|\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\mathrm{net}}\big)\right| \\
&\le\left|\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\Big(\big\|P_h^k\big\|_1+\big\|P_{h,s_h^k,a_h^k}\big\|_1\Big)\big\|V_{h+1}^{\mathrm{u}}-V_{h+1}^{\mathrm{net}}\big\|_\infty\right| \\
&\le 2eK\alpha\log T,
\end{aligned} \tag{A.153}$$

where the last inequality follows from $\sum_{n=N_h^i(s_h^i,a_h^i)}^{N_h^{K-1}(s_h^i,a_h^i)}\frac{1}{n}\le\log T$ and $\lambda_h^k\le e$ (cf. (A.137)). Consequently, by taking $\alpha=1/(SAT)$, we can deduce that

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\star}\big)\right| \\
&\quad\le\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{net}}-V_{h+1}^{\star}\big)\right| \\
&\qquad+\sum_{h=1}^{H}\left|\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\mathrm{net}}\big)\right| \\
&\quad\lesssim\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)}\frac{\lambda_h^k}{n}\big(P_h^k-P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathrm{net}}-V_{h+1}^{\star}\big)\right|+HK\alpha\log T \\
&\quad\lesssim\sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K}P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{net}}-V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{SAT}{\delta\alpha}}+H^2SA\log^2\frac{SAT}{\delta\alpha}+HK\alpha\log T
\end{aligned}$$

$$\asymp \sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathrm{u}}-V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{SAT}{\delta}} + H^2SA\log^2\frac{SAT}{\delta}, \tag{A.154}$$

where the last line holds due to the condition (A.152) and our choice of α. To summarize, with probability exceeding $1-\delta/6$, the property (A.154) holds simultaneously for all $\{V_{h+1}^{\mathrm{u}} \in \mathbb{R}^S \mid 1 \le h \le H\}$ obeying (A.148).

Step 4: controlling the original term of interest. With the above union bound in hand, we are ready to control the original term of interest

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathsf{R},K} - V_{h+1}^{\star}\big). \tag{A.155}$$

To begin with, it can be easily verified using (3.20) that

$$V_{h+1}^{\star} \le V_{h+1}^{\mathsf{R},K} \le H \qquad \text{for all } 1 \le h \le H. \tag{A.156}$$

Moreover, we make the observation that

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K} P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},K} - V_{h+1}^{\star}\big) &\overset{\text{(i)}}{\le} \sum_{h=1}^{H}\sum_{k=1}^{K} P_{h,s_h^k,a_h^k}\big(V_{h+1}^{\mathsf{R},k} - V_{h+1}^{\star}\big) \\ &\overset{\text{(ii)}}{\le} \sqrt{H^7SAK\log\frac{SAT}{\delta}} + H^3SA + HK \end{aligned}\tag{A.157}$$

with probability exceeding $1-\delta/6$, where (i) holds because V_{h+1}^{R} is monotonically non-increasing (in view of the monotonicity of $V_h(s)$ in (3.17b) and the update rule in line 16 of Algorithm 3), and (ii) follows from (A.140). Substitution into (A.154) yields

$$\begin{aligned}&\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{n=N_h^k(s_h^k,a_h^k)}^{N_h^{K-1}(s_h^k,a_h^k)} \frac{\lambda_h^k}{n}\big(P_h^k - P_{h,s_h^k,a_h^k}\big)\big(V_{h+1}^{\mathsf{R},K} - V_{h+1}^{\star}\big)\right| \\ &\lesssim \sqrt{H^2SA\sum_{h=1}^{H}\sum_{i=1}^{K} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\mathsf{R},K}-V_{h+1}^{\star}\right)\left(\log^2 T\right)\log\frac{SAT}{\delta}} + H^2SA\log^2\frac{SAT}{\delta} \\ &\lesssim \sqrt{H^2SA\left\{\sqrt{H^7SAK\log\frac{SAT}{\delta}} + H^3SA + HK\right\}\left(\log^2 T\right)\log\frac{SAT}{\delta}} + H^2SA\log^2\frac{SAT}{\delta} \\ &\lesssim \sqrt{H^2SA\left\{H^6SA\log\frac{SAT}{\delta} + H^3SA + HK\right\}\log^3\frac{SAT}{\delta}} + H^2SA\log^2\frac{SAT}{\delta}\end{aligned}$$

$$\lesssim H^4 SA \log^2 \frac{SAT}{\delta} + \sqrt{H^3 SAK \log^3 \frac{SAT}{\delta}}, \tag{A.158}$$

where the penultimate line holds since

$$\sqrt{H^7 SAK \log \frac{SAT}{\delta}} = \sqrt{H^6 SA \log \frac{SAT}{\delta}} \sqrt{HK} \lesssim H^6 SA \log \frac{SAT}{\delta} + HK.$$

Appendix B

Proofs for Chapter 4

B.1 Technical lemmas

B.1.1 Preliminary facts

Our results rely heavily on proper choices of the learning rates, leading to several useful properties which have been established in Lemma 1.

In addition, we gather a few elementary properties about the Binomial distribution, which will be useful throughout the proof. The lemma below is adapted from Xie et al. (2021b, Lemma A.1).

Lemma 27. *Suppose $N \sim \mathsf{Binomial}(n, p)$, where $n \geq 1$ and $p \in [0, 1]$. For any $\delta \in (0, 1)$, we have*

$$\frac{p}{N \vee 1} \leq \frac{8 \log\left(\frac{1}{\delta}\right)}{n}, \tag{B.1}$$

and

$$N \geq \frac{np}{8 \log\left(\frac{1}{\delta}\right)} \qquad \text{if } np \geq 8 \log\left(\frac{1}{\delta}\right), \tag{B.2a}$$

$$N \leq \begin{cases} e^2 np & \text{if } np \geq \log\left(\frac{1}{\delta}\right), \\ 2e^2 \log\left(\frac{1}{\delta}\right) & \text{if } np \leq 2 \log\left(\frac{1}{\delta}\right). \end{cases} \tag{B.2b}$$

with probability at least $1 - 4\delta$.

Proof. To begin with, we directly invoke Xie et al. (2021b, Lemma A.1) which yields the results in (B.1) and (B.2a). Regarding (B.2b), invoking the Chernoff bound (Vershynin, 2018, Theorem 2.3.1) with $\mathbb{E}[N] = np$, when $np \geq \log\left(\frac{1}{\delta}\right)$, it satisfies

$$\mathbb{P}(N \geq e^2 np) \leq e^{-np} \left(\frac{enp}{e^2 np}\right)^{e^2 np} \leq e^{-np} \leq \delta.$$

Similarly, when $np \leq 2 \log\left(\frac{1}{\delta}\right)$, we have

$$\mathbb{P}\left(N \geq 2e^2 \log\left(\frac{1}{\delta}\right)\right) \overset{\text{(i)}}{\leq} e^{-np} \left(\frac{enp}{2e^2 \log\left(\frac{1}{\delta}\right)}\right)^{2e^2 \log\left(\frac{1}{\delta}\right)}$$

$$\overset{\text{(ii)}}{\leq} e^{-np}\left(\frac{enp}{e^2 np}\right)^{2e^2\log\left(\frac{1}{\delta}\right)} \leq e^{-2e^2\log\left(\frac{1}{\delta}\right)} \leq \delta,$$

where (i) results from Vershynin (2018, Theorem 2.3.1), and (ii) follows from the basic fact $e^2\log\left(\frac{1}{\delta}\right) \geq 2\log\left(\frac{1}{\delta}\right) \geq np$. Taking the union bound thus completes the proof. □

B.1.2 Application of Freedman's inequality

Both the samples collected within each episode and the algorithms analyzed herein exhibit certain Markovian structure. As a result, concentration inequalities tailored to martingales become particularly effective for our analysis. Besides Freedman's inequality (cf. Theorem 18) and its consequence established in Lemma 24, we shall make note of an immediate consequence of Lemma 24 tailored to our problem. Recall that $N_h^i(s,a)$ denotes the number of times that (s,a) has been visited at step h before the beginning of the i-th episode, and $k^n(s,a)$ stands for the index of the episode in which (s,a) is visited for the n-th time.

Lemma 28. *Let $\left\{W_h^i \in \mathbb{R}^S \mid 1 \leq i \leq K, 1 \leq h \leq H+1\right\}$ be a collection of vectors satisfying the following properties:*

- *W_h^i is fully determined by the samples collected up to the end of the $(h-1)$-th step of the i-th episode;*
- *$\|W_h^i\|_\infty \leq C_{\mathrm{w}}$.*

For any positive $N \geq H$, we consider the following sequence

$$X_i(s,a,h,N) \coloneqq \eta^N_{N_h^i(s,a)}\big(P_h^i - P_{h,s,a}\big)W_{h+1}^i \mathbb{1}\left\{(s_h^i, a_h^i) = (s,a)\right\}, \qquad 1 \leq i \leq K, \tag{B.3}$$

with P_h^i defined in (4.18). Consider any $\delta \in (0,1)$. With probability at least $1-\delta$,

$$\left|\sum_{i=1}^{k} X_i(s,a,h,N)\right| \lesssim \sqrt{\frac{H}{N}C_{\mathrm{w}}^2}\log^2\frac{SAT}{\delta} \tag{B.4}$$

holds simultaneously for all $(k,h,s,a,N) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$.

Proof. Taking $u_h^i(s,a,N) = \eta^N_{N_h^i(s,a)}$, one can see from (3.14b) in Lemma 1 that

$$\left|u_h^i(s,a,N)\right| \leq \frac{2H}{N} =: C_{\mathrm{u}}.$$

Recognizing the trivial bound $\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big) \leq C_{\mathrm{w}}^2$, we can invoke Lemma 24 to obtain that,

with probability at least $1-\delta$,

$$\left|\sum_{i=1}^{k} X_i(s,a,h,N)\right| \lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)}\eta_n^N C_{\mathrm{w}}^2} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta}$$
$$\lesssim \sqrt{\frac{H}{N}\log^2\frac{SAT}{\delta}}\cdot C_{\mathrm{w}} + \frac{HC_{\mathrm{w}}}{N}\log^2\frac{SAT}{\delta} \lesssim \sqrt{\frac{HC_{\mathrm{w}}^2}{N}}\log^2\frac{SAT}{\delta}$$

holds simultaneously for all $(k,h,s,a,N)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}\times[K]$, where the last line applies (3.14b) in Lemma 1 once again. □

Finally, we introduce another lemma by invoking Freedman's inequality in Theorem 18.

Lemma 29. *Let $\left\{W_h^k(s,a)\in\mathbb{R}^S \mid (s,a)\in\mathcal{S}\times\mathcal{A}, 1\le k\le K, 1\le h\le H+1\right\}$ be a collection of vectors satisfying the following properties:*

- *$W_h^k(s,a)$ is fully determined by the given state-action pair (s,a) and the samples collected up to the end of the $(k-1)$-th episode;*
- *$\|W_h^k(s,a)\|_\infty \le C_{\mathrm{w}}$.*

For any positive $C_{\mathrm{d}}\ge 0$, we consider the following sequences

$$X_{h,k} \coloneqq C_{\mathrm{d}}\left[\frac{d_h^{\pi_\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}P_{h,s_h^k,a_h^k}W_{h+1}^k(s_h^k,a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}W_{h+1}^k(s,a)\right], \qquad 1\le k\le K, \tag{B.5}$$

$$\overline{X}_{h,k} \coloneqq C_{\mathrm{d}}\left[\frac{d_h^{\pi_\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}P_h^k W_{h+1}^k(s_h^k,a_h^k) - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}W_{h+1}^k(s,a)\right], \qquad 1\le k\le K. \tag{B.6}$$

Consider any $\delta\in(0,1)$. Then with probability at least $1-\delta$,

$$\left|\sum_{k=1}^{K} X_{h,k}\right| \le \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[P_{h,s,a}W_{h+1}^k(s,a)\right]^2\log\frac{2H}{\delta}} + 2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \tag{B.7}$$

$$\left|\sum_{k=1}^{K} \overline{X}_{h,k}\right| \le \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left[W_{h+1}^k(s,a)\right]^2\log\frac{2H}{\delta}} + 2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \tag{B.8}$$

hold simultaneously for all $h\in[H]$.

Proof. We intend to apply Freedman's inequality (cf. Theorem 18) to control $\sum_{k=1}^{K} X_{h,k}$. Considering any given time step h, it is easily verified that

$$\mathbb{E}_{k-1}[X_{h,k}] = 0, \qquad \mathbb{E}_{k-1}[\overline{X}_{h,k}] = 0,$$

where $\mathbb{E}_{k-1}$ denotes the expectation conditioned on everything happening up to the end of the $(k-1)$-th episode. To continue, we observe that

$$|X_{h,k}| \le C_{\mathrm{d}} \left(\frac{d_h^{\pi_\star}(s_h^k, a_h^k)}{d_h^{\mu}(s_h^k, a_h^k)} + 1 \right) \left\| W_{h+1}^k(s,a) \right\|_\infty \le 2C_{\mathrm{d}} C^\star C_{\mathrm{w}}, \tag{B.9}$$

$$|\overline{X}_{h,k}| \le C_{\mathrm{d}} \left(\frac{d_h^{\pi_\star}(s_h^k, a_h^k)}{d_h^{\mu}(s_h^k, a_h^k)} + 1 \right) \left\| W_{h+1}^k(s,a) \right\|_\infty \le 2C_{\mathrm{d}} C^\star C_{\mathrm{w}}, \tag{B.10}$$

where we use the assumptions $\frac{d_h^{\pi_\star}(s,a)}{d_h^{\mu}(s,a)} \le C^\star$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ (cf. Assumption 1) and $\left\| W_{h+1}^k(s_h^k, a_h^k) \right\|_\infty \le C_{\mathrm{w}}$.

Recall that $\Delta(\mathcal{S} \times \mathcal{A})$ is the probability simplex over the set $\mathcal{S} \times \mathcal{A}$ of all state-action pairs, and we denote by $d_h^\mu \in \Delta(\mathcal{S} \times \mathcal{A})$ the state-action visitation distribution induced by the behavior policy μ at time step $h \in [H]$. With this in hand, we obtain

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}_{k-1}[|X_{h,k}|^2] &\le \sum_{k=1}^{K} C_{\mathrm{d}}^2 \mathbb{E}_{k-1} \left[\frac{d_h^{\pi_\star}(s_h^k, a_h^k)}{d_h^{\mu}(s_h^k, a_h^k)} P_{h,s_h^k,a_h^k} W_{h+1}^k(s_h^k, a_h^k) - \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\pi_\star}(s,a) P_{h,s,a} W_{h+1}^k(s,a) \right]^2 \\
&\le \sum_{k=1}^{K} C_{\mathrm{d}}^2 \mathbb{E}_{(s_h^k, a_h^k) \sim d_h^\mu} \left[\frac{d_h^{\pi_\star}(s_h^k, a_h^k)}{d_h^{\mu}(s_h^k, a_h^k)} P_{h,s_h^k,a_h^k} W_{h+1}^k(s_h^k, a_h^k) \right]^2 \\
&= \sum_{k=1}^{K} C_{\mathrm{d}}^2 \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{d_h^{\pi_\star}(s,a)}{d_h^{\mu}(s,a)} d_h^{\pi_\star}(s,a) \left[P_{h,s,a} W_{h+1}^k(s,a) \right]^2 \\
&\overset{\text{(i)}}{\le} \sum_{k=1}^{K} C_{\mathrm{d}}^2 C^\star \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\pi_\star}(s,a) \left[P_{h,s,a} W_{h+1}^k(s,a) \right]^2 \qquad \text{(B.11)} \\
&\le \sum_{k=1}^{K} C_{\mathrm{d}}^2 \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} C^\star d_h^{\pi_\star}(s,a) \left\| W_{h+1}^k(s_h^k, a_h^k) \right\|_\infty^2 \le C_{\mathrm{d}}^2 C^\star C_{\mathrm{w}}^2 K, \qquad \text{(B.12)}
\end{aligned}$$

where (i) follows from $\frac{d_h^{\pi_\star}(s,a)}{d_h^{\mu}(s,a)} \le C^\star$ (see Assumption 1) and the assumption $\left\| W_{h+1}^k(s_h^k, a_h^k) \right\|_\infty \le C_{\mathrm{w}}$.

Similarly, we can derive

$$\sum_{k=1}^{K} \mathbb{E}_{k-1}[|\overline{X}_{h,k}|^2] \le \sum_{k=1}^{K} C_{\mathrm{d}}^2 \mathbb{E}_{k-1} \left[\frac{d_h^{\pi_\star}(s_h^k, a_h^k)}{d_h^{\mu}(s_h^k, a_h^k)} P_h^k W_{h+1}^k(s_h^k, a_h^k) - \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\pi_\star}(s,a) P_{h,s,a} W_{h+1}^k(s,a) \right]^2$$

$$
\begin{aligned}
&\leq \sum_{k=1}^{K} C_{\mathrm{d}}^2 \mathbb{E}_{(s_h^k,a_h^k)\sim d_h^\mu}\left[\mathbb{E}_{P_h^k\sim P_{h,s_h^k,a_h^k}}\left[\frac{d_h^{\pi_\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k W_{h+1}^k(s_h^k,a_h^k)\right]^2\right] \\
&= \sum_{k=1}^{K} C_{\mathrm{d}}^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{d_h^{\pi_\star}(s,a)}{d_h^\mu(s,a)} d_h^{\pi_\star}(s,a)\mathbb{E}_{P_h^k\sim P_{h,s,a}}\left[P_h^k W_{h+1}^k(s,a)\right]^2 \\
&\overset{\text{(i)}}{\leq} \sum_{k=1}^{K} C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\mathbb{E}_{P_h^k\sim P_{h,s,a}}\left[P_h^k W_{h+1}^k(s,a)\right]^2 \qquad \text{(B.13)} \\
&= \sum_{k=1}^{K} C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a) P_{h,s,a}\left[W_{h+1}^k(s,a)\right]^2 \qquad \text{(B.14)} \\
&\leq \sum_{k=1}^{K} C_{\mathrm{d}}^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} C^\star d_h^{\pi_\star}(s,a)\left\|W_{h+1}^k(s,a)\right\|_\infty^2 \leq C_{\mathrm{d}}^2 C^\star C_{\mathrm{w}}^2 K, \qquad \text{(B.15)}
\end{aligned}
$$

where (i) follows from $\frac{d_h^{\pi_\star}(s,a)}{d_h^\mu(s,a)} \leq C^\star$ (see Assumption 1) and the assumption $\left\|W_{h+1}^k(s_h^k,a_h^k)\right\|_\infty \leq C_{\mathrm{w}}$.

Plugging in the results in (B.9) and (B.11) (resps. (B.10) and (B.14)) to control $\sum_{k=1}^K |X_{h,k}|$ (resps. $\sum_{k=1}^K |\overline{X}_{h,k}|$), we invoke Theorem 18 with $m = \lceil \log_2 K \rceil$ and take the union bound over $h \in [H]$ to show that with probability at least $1-\delta$,

$$
\begin{aligned}
\left|\sum_{k=1}^{K} X_{h,k}\right| &\leq \sqrt{8\max\left\{\sum_{k=1}^{K} C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[P_{h,s,a}W_{h+1}^k(s,a)\right]^2, \frac{C_{\mathrm{d}}^2 C^\star C_{\mathrm{w}}^2 K}{2^m}\right\}\log\frac{2H}{\delta}} \\
&\quad + \frac{8}{3}C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \\
&\leq \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[P_{h,s,a}W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 6C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta}
\end{aligned}
$$

and

$$
\begin{aligned}
\left|\sum_{k=1}^{K} \overline{X}_{h,k}\right| &\leq \sqrt{8\max\left\{\sum_{k=1}^{K} C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left[W_{h+1}^k(s,a)\right]^2, \frac{C_{\mathrm{d}}^2 C^\star C_{\mathrm{w}}^2 K}{2^m}\right\}\log\frac{2H}{\delta}} \\
&\quad + \frac{8}{3}C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \\
&\leq \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left[W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 6C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta}
\end{aligned}
$$

holds simultaneously for all $h \in [H]$. □

B.2 Proof of main lemmas for LCB-Q (Theorem 2)

B.2.1 Proof of Lemma 7

B.2.1.1 Proof of inequality (4.22)

To begin with, we shall control $\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \big(P_{h,s,a} - P_h^{k^n(s,a)}\big) V_{h+1}^{k^n(s,a)}$ by invoking Lemma 28. Let

$$W_{h+1}^i \coloneqq V_{h+1}^i,$$

which satisfies

$$\|W_{h+1}^i\|_\infty \le H =: C_{\mathrm{w}}.$$

Applying Lemma 28 with $N = N_h^k(s,a)$ reveals that, with probability at least $1-\delta$,

$$\left| \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \Big(P_{h,s,a} - P_h^{k^n(s,a)}\Big) V_{h+1}^{k^n(s,a)} \right| = \left| \sum_{i=1}^{k} X_i\big(s,a,h,N_h^k(s,a)\big) \right| \le c_{\mathrm{b}} \sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}} \tag{B.16a}$$

holds simultaneously for all $(s,a,k,h) \in \mathcal{S}\times\mathcal{A}\times[K]\times[H]$, provided that the constant $c_{\mathrm{b}} > 0$ is large enough and that $N_h^k(s,a) > 0$. If $N_h^k(s,a) = 0$, then we have the trivial bound

$$\left| \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \Big(P_{h,s,a} - P_h^{k^n(s,a)}\Big) V_{h+1}^{k^n(s,a)} \right| = 0. \tag{B.16b}$$

Additionally, from the definition $b_n = c_{\mathrm{b}}\sqrt{\frac{H^3\iota^2}{n}}$, we observe that

$$\begin{cases} \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n \in \Big[c_{\mathrm{b}}\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}}, 2c_{\mathrm{b}}\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)}}\Big], & \text{if } N_h^k(s,a) > 0 \\ \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n = 0, & \text{if } N_h^k(s,a) = 0 \end{cases} \tag{B.17}$$

holds simultaneously for all $s,a,h,k \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, which follows directly from the property (3.14a) in Lemma 1.

Combining the above bounds (B.16) and (B.17), we arrive at the advertised result

$$\left| \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \Big(P_{h,s,a} - P_h^{k^n(s,a)}\Big) V_{h+1}^{k^n(s,a)} \right| \le \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n.$$

B.2.1.2 Proof of inequality (4.23)

Note that the second inequality of (4.23) holds straightforwardly as

$$V_h^{\pi}(s) \leq V^{\star}(s)$$

holds for any policy π. As a consequence, it suffices to establish the first inequality of (4.23), namely,

$$V_h^k(s) \leq V_h^{\pi^k}(s) \qquad \text{for all } (s,h,k) \in \mathcal{S} \times [H] \times [K]. \tag{B.18}$$

Before proceeding, let us introduce the following auxiliary index

$$k_o(h,k,s) \coloneqq \max\left\{l : l < k \text{ and } V_h^l(s) = \max_a Q_h^l(s,a)\right\} \tag{B.19}$$

for any $(h,k,s) \in [H] \times [K] \times \mathcal{S}$, which denotes the index of the latest episode — before the end of the $(k-1)$-th episode — in which $V_h(s)$ has been updated. In what follows, we shall often abbreviate $k_o(h,k,s)$ as $k_o(h)$ whenever it is clear from the context.

Towards establishing the relation (B.18), we proceed by means of an inductive argument. In what follows, we shall first justify the desired inequality for the base case when $h+1 = H+1$ for all episodes $k \in [K]$, and then use induction to complete the argument for other cases. More specifically, consider any step $h \in [H]$ in any episode $k \in [K]$, and suppose that the first inequality of (4.23) is satisfied for all previous episodes as well as all steps $h' \geq h+1$ in the current episode, namely,

$$V_{h'}^{k'}(s) \leq V_{h'}^{\pi^{k'}}(s) \qquad \text{for all } (k',h',s) \in [k-1] \times [H+1] \times \mathcal{S}, \tag{B.20a}$$

$$V_{h'}^{k}(s) \leq V_{h'}^{\pi^{k}}(s) \qquad \text{for all } h' \geq h+1 \text{ and } s \in \mathcal{S}. \tag{B.20b}$$

We intend to justify that the following is valid

$$V_h^k(s) \leq V_h^{\pi^k}(s) \qquad \text{for all } s \in \mathcal{S}, \tag{B.21}$$

assuming that the induction hypothesis (B.20) holds.

Step 1: base case. Let us begin with the base case when $h+1 = H+1$ for all episodes $k \in [K]$. Recognizing the fact that $V_{H+1}^{\pi} = V_{H+1}^k = 0$ for any π and any $k \in [K]$, we directly arrive at

$$V_{H+1}^k(s) \leq V_{H+1}^{\pi^k}(s) \qquad \text{for all } (k,s) \in [K] \times \mathcal{S}. \tag{B.22}$$

Step 2: induction. To justify (B.21) under the induction hypothesis (B.20), we decompose the difference term to obtain

$$\begin{aligned} V_h^{\pi^k}(s) - V_h^k(s) &= V_h^{\pi^k}(s) - \max\big\{\max_a Q_h^k(s,a), V_h^{k-1}(s)\big\} \\ &= Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - \max\big\{\max_a Q_h^k(s,a), V_h^{k_o(h)}(s)\big\}, \end{aligned} \tag{B.23}$$

where the last line holds since $V_h(s)$ has not been updated during episodes $k_o(h), k_o(h)+1, \cdots, k-1$ (in view of the definition of $k_o(h)$ in (B.19)). We shall prove that the right-hand side of (B.23) is non-negative by discussing the following two cases separately.

- Consider the case where $V_h^k(s) = \max_a Q_h^k(s,a)$. Before continuing, it is easily observed from the update rule in line 13 and line 12 of Algorithm 4 that: $V_h(s)$ and $\pi_h(s)$ are updated hand-in-hand for every h. Thus, it implies that

$$\pi_h^k(s) = \arg\max_a Q_h^k(s,a), \qquad \text{when } V_h^k(s) = \max_a Q_h^k(s,a) \tag{B.24}$$

holds for all $(k,h) \in [K] \times [H]$. As a result, we express the term of interest as follows:

$$V_h^{\pi^k}(s) - V_h^k(s) = Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - \max_a Q_h^k(s,a) = Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - Q_h^k\big(s, \pi_h^k(s)\big). \tag{B.25}$$

To continue, we turn to controlling a more general term $Q_h^{\pi^k}(s,a) - Q_h^k(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Invoking the fact $\eta_0^{N_h^k} + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see (4.16) and (4.17)) leads to

$$Q_h^{\pi^k}(s,a) = \eta_0^{N_h^k} Q_h^{\pi^k}(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} Q_h^{\pi^k}(s,a).$$

This relation combined with (4.20) allows us to express the difference between $Q_h^{\pi^k}$ and Q_h^k as follows

$$\begin{aligned} Q_h^{\pi^k}(s,a) - Q_h^k(s,a) &= \eta_0^{N_h^k}\Big(Q_h^{\pi^k}(s,a) - Q_h^1(s,a)\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[Q_h^{\pi^k}(s,a) - r_h(s,a) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + b_n\Big] \\ &\overset{\text{(i)}}{=} \eta_0^{N_h^k}\Big(Q_h^{\pi^k}(s,a) - Q_h^1(s,a)\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[P_{h,s,a}V_{h+1}^{\pi^k} - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + b_n\Big] \\ &\overset{\text{(ii)}}{\geq} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[P_{h,s,a}V_{h+1}^{\pi^k} - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + b_n\Big] \\ &\overset{\text{(iii)}}{=} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\Big(V_{h+1}^{\pi^k} - V_{h+1}^{k^n}\Big) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big[\Big(P_{h,s,a} - P_h^{k^n}\Big)V_{h+1}^{k^n} + b_n\Big] \end{aligned}$$

$$\overset{\text{(iv)}}{\geq} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left[\left(P_{h,s,a} - P_h^{k^n} \right) V_{h+1}^{k^n} + b_n \right]. \tag{B.26}$$

Here, (i) invokes the Bellman equation $Q_h^{\pi^k}(s,a) = r_h(s,a) + P_{h,s,a} V_{h+1}^{\pi^k}$; (ii) holds since $Q_h^{\pi^k}(s,a) \geq 0 = Q_h^1(s,a)$; (iii) relies on the notaion (4.18); and (iv) comes from the fact

$$V_{h+1}^{\pi^k} \geq V_{h+1}^k \geq V_{h+1}^{k^n},$$

owing to the induction hypothesis in (B.20) as well as the monotonicity of V_{h+1} in (4.21). Consequently, it follows from (B.26) that

$$\begin{aligned} Q_h^{\pi^k}(s,a) - Q_h^k(s,a) &\geq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(P_{h,s,a} - P_h^{k^n(s,a)} \right) V_{h+1}^{k^n(s,a)} + \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n \\ &\geq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} b_n - \left| \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(P_{h,s,a} - P_h^{k^n(s,a)} \right) V_{h+1}^{k^n(s,a)} \right| \geq 0 \end{aligned} \tag{B.27}$$

for all state-action pair (s,a), where the last inequality holds due to the bound (4.22) in Lemma 7. Plugging the above result into (B.25) directly establishes that

$$V_h^{\pi^k}(s) - V_h^k(s) = Q_h^{\pi^k}\big(s, \pi^k(s)\big) - Q_h^k\big(s, \pi^k(s)\big) \geq 0. \tag{B.28}$$

- When $V_h^k(s) = V_h^{k_o(h)}(s)$, it indicates that

$$V_h^{k_o(h)}(s) = \max_a Q_h^{k_o(h)}(s,a), \qquad \pi_h^{k_o(h)}(s) = \arg\max_a Q_h^{k_o(h)}(s,a), \tag{B.29}$$

which follows from the definition of $k_o(h)$ in (B.19) and the corresponding fact in (B.24). We also make note of the fact that

$$\pi_h^k(s) = \pi_h^{k_o(h)}(s), \tag{B.30}$$

which holds since $V_h(s)$ (and hence $\pi_h(s)$) has not been updated during episodes $k_o(h), k_o(h)+1, \cdots, k-1$ (in view of the definition (B.19)). Combining the above two results, we can show that

$$\begin{aligned} V_h^{\pi^k}(s) - V_h^k(s) &= Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - V_h^{k_o(h)}(s) = Q_h^{\pi^k}\big(s, \pi_h^k(s)\big) - \max_a Q_h^{k_o(h)}(s,a) \\ &= Q_h^{\pi^k}\big(s, \pi_h^{k_o(h)}(s)\big) - Q_h^{k_o(h)}\big(s, \pi_h^{k_o(h)}(s)\big) \\ &\geq 0, \end{aligned} \tag{B.31}$$

where the final line can be verified using exactly the same argument as in the previous case to show (B.26) and then (B.28). Here, we omit the proof of this step for brevity.

To conclude, substituting the relations (B.28) and (B.31) in the above two cases back into (B.23), we arrive at

$$V_h^{\pi^k}(s) - V_h^k(s) \geq 0$$

as desired in (B.21). This immediately completes the induction argument.

B.2.2 Proof of Lemma 8

We make the observation that Lemma 8 would follow immediately if we could establish the following relation:

$$\begin{aligned} A_h &:= \sum_{k=1}^{K} \underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a) P_{h,s,a} \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \left(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\right)}_{=:A_{h,k}} \\ &\leq \sum_{k=1}^{K} \underbrace{\left(1+\frac{1}{H}\right) \sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s) \left(V_{h+1}^\star(s) - V_{h+1}^k(s)\right)}_{=:B_{h,k}} + 24\sqrt{H^2 C^\star K \log\frac{2H}{\delta}} + 12HC^\star \log\frac{2H}{\delta}. \end{aligned} \tag{B.32}$$

The remainder of the proof is thus dedicated to proving (B.32).

To continue, let us first consider two auxiliary sequences $\{Y_{h,k}\}_{k=1}^K$ and $\{Z_{h,k}\}_{k=1}^K$ which are the empirical estimates of $A_{h,k}$ and $B_{h,k}$, respectively. For any time step h in episode k, $Y_{h,k}$ and $Z_{h,k}$ are defined as follows

$$\begin{aligned} Y_{h,k} &:= \frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)} \left(V_{h+1}^\star - V_{h+1}^{k^n(s_h^k,a_h^k)}\right), \\ Z_{h,k} &:= \left(1+\frac{1}{H}\right) \frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \left(V_{h+1}^\star - V_{h+1}^k\right). \end{aligned}$$

To begin with, let us establish the relationship between $\{Y_{h,k}\}_{k=1}^K$ and $\{Z_{h,k}\}_{k=1}^K$:

$$\begin{aligned} \sum_{k=1}^{K} Y_{h,k} &= \sum_{k=1}^{K} \frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)} \left(V_{h+1}^\star - V_{h+1}^{k^n(s_h^k,a_h^k)}\right) \\ &\overset{\text{(i)}}{=} \sum_{l=1}^{K} \frac{d_h^{\pi^\star}(s_h^l,a_h^l)}{d_h^\mu(s_h^l,a_h^l)} P_{h,s_h^l,a_h^l} \left\{ \sum_{N=N_h^l(s_h^l,a_h^l)}^{N_h^K(s_h^l,a_h^l)} \eta_{N_h^l(s_h^l,a_h^l)}^N \right\} \left(V_{h+1}^\star - V_{h+1}^l\right) \end{aligned} \tag{B.33}$$

$$\leq \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}P_{h,s_h^k,a_h^k}\left(V_{h+1}^\star - V_{h+1}^k\right) = \sum_{k=1}^{K} Z_{h,k}. \tag{B.34}$$

Here, (i) holds by replacing $k^n(s_h^k,a_h^k)$ with l and gathering all terms that involve $V_{h+1}^\star - V_{h+1}^l$; in the last line, we have invoked the property $\sum_{N=n}^{N_h^K(s,a)}\eta_n^N \leq \sum_{N=n}^{\infty}\eta_n^N = 1+1/H$ (see (3.14b)) together with the fact $V_{h+1}^\star - V_{h+1}^l \geq 0$ (see Lemma 7), and have further replaced l with k.

With the above relation in hand, in order to verify (B.32), we further decompose A_h into several terms

$$\begin{aligned} A_h &= \sum_{k=1}^{K} A_{h,k} = \sum_{k=1}^{K} Y_{h,k} + \sum_{k=1}^{K}(A_{h,k}-Y_{h,k}) \overset{\text{(i)}}{\leq} \sum_{k=1}^{K} Z_{h,k} + \sum_{k=1}^{K}(A_{h,k}-Y_{h,k}) \\ &= \sum_{k=1}^{K} B_{h,k} + \sum_{k=1}^{K}(Z_{h,k}-B_{h,k}) + \sum_{k=1}^{K}(A_{h,k}-Y_{h,k}) \end{aligned} \tag{B.35}$$

where (i) follows from (B.34).

As a result, it remains to control $\sum_{k=1}^{K}(Z_{h,k}-B_{h,k})$ and $\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})$ separately in the following.

Step 1: controlling $\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})$. We shall first control this term by means of Lemma 29. Specifically, consider

$$W_{h+1}^k(s,a) \coloneqq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\left(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\right), \qquad C_{\mathrm{d}} \coloneqq 1 \tag{B.36}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_\infty \leq \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\left(\left\|V_{h+1}^\star\right\|_\infty + \left\|V_{h+1}^{k^n(s,a)}\right\|_\infty\right) \leq 2H =: C_{\mathrm{w}}. \tag{B.37}$$

Here we use the fact that $\eta_0^{N_h^k} + \sum_{n=1}^{N_h^k}\eta_n^{N_h^k} = 1$ (see (4.16) and (4.17)). Then, applying Lemma 29 with (B.36), we have with probability at least $1-\delta$, the following inequality holds true

$$\begin{aligned} \left|\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})\right| &=: \left|\sum_{k=1}^{K} X_{h,k}\right| \\ &\leq \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left[P_{h,s,a}W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \end{aligned}$$

$$\overset{\text{(i)}}{\leq} \sqrt{8C^\star \sum_{k=1}^{K} \left\|W_{h+1}^k(s,a)\right\|_\infty^2 \log\frac{2H}{\delta}} + 4HC^\star \log\frac{2H}{\delta}$$
$$\leq 8\sqrt{H^2 C^\star K \log\frac{2H}{\delta}} + 4HC^\star \log\frac{2H}{\delta}, \tag{B.38}$$

where (i) holds since $\left|P_{h,s,a} W_{h+1}^k(s,a)\right| \leq \left\|P_{h,s,a}\right\|_1 \|W_{h+1}^k(s,a)\|_\infty = \|W_{h+1}^k(s,a)\|_\infty$.

Step 2: controlling $\sum_{k=1}^{K}(Z_{h,k} - B_{h,k})$. Similarly, we shall control $\sum_{k=1}^{K}(Z_{h,k} - B_{h,k})$ by invoking Lemma 29.

Recall that

$$Z_{h,k} - B_{h,k} = \left(1+\frac{1}{H}\right)\frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k}\left(V_{h+1}^\star - V_{h+1}^k\right) - \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s) - V_{h+1}^k(s)\right), \tag{B.39}$$

and let us consider

$$W_{h+1}^k(s,a) \coloneqq V_{h+1}^\star - V_{h+1}^k, \qquad C_{\mathrm{d}} \coloneqq \left(1+\frac{1}{H}\right) \leq 2 \tag{B.40}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_\infty \leq \left\|V_{h+1}^\star\right\|_\infty + \left\|V_{h+1}^k\right\|_\infty \leq 2H =: C_{\mathrm{w}}. \tag{B.41}$$

Again, in view of Lemma 29, we can show that with probability at least $1-\delta$,

$$\left|\sum_{k=1}^{K}(B_{h,k} - Z_{h,k})\right| = \left|\sum_{k=1}^{K} X_{h,k}\right|$$
$$\leq \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left[P_{h,s,a} W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}} C^\star C_{\mathrm{w}} \log\frac{2H}{\delta}$$
$$\overset{\text{(i)}}{\leq} \sqrt{32C^\star \sum_{k=1}^{K} \left\|W_{h+1}^k(s,a)\right\|_\infty^2 \log\frac{2H}{\delta}} + 8HC^\star \log\frac{2H}{\delta}$$
$$\leq 16\sqrt{H^2 C^\star K \log\frac{2H}{\delta}} + 8HC^\star \log\frac{2H}{\delta}, \tag{B.42}$$

where (i) holds due to the fact $\left\|P_{h,s,a}\right\|_1 = 1$.

Step 3: putting all this together. Substitution results in (B.38) and (B.42) back into (B.35) completes the proof of (B.32) as follows

$$\begin{aligned} A_h &\le \sum_{k=1}^{K} B_{h,k} + \left|\sum_{k=1}^{K}(Z_{h,k} - B_{h,k})\right| + \left|\sum_{k=1}^{K}(A_{h,k} - Y_{h,k})\right| \\ &\le \sum_{k=1}^{K} B_{h,k} + 24\sqrt{H^2C^\star K \log\frac{2H}{\delta}} + 12HC^\star \log\frac{2H}{\delta}. \end{aligned}$$

This in turn concludes the proof of Lemma 8.

B.2.3 Proof of Lemma 9

Recall that the term of interest in (4.33) is given by

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}\right) + \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} I_h. \tag{B.43}$$

First, it is easily seen that

$$\left(1+\frac{1}{H}\right)^{h-1} \le \left(1+\frac{1}{H}\right)^{H} \le e \qquad \text{for every } h = 1, \cdots, H, \tag{B.44}$$

which taken collectively with the expression of the first term in (B.43) yields

$$\begin{aligned} \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(24\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 12HC^\star\log\frac{2H}{\delta}\right) &\le 24e\sum_{h=1}^{H}\left(\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + HC^\star\log\frac{2H}{\delta}\right) \\ &\lesssim \sqrt{H^4C^\star K\log\frac{H}{\delta}} + H^2C^\star\log\frac{H}{\delta}. \end{aligned} \tag{B.45}$$

As a result, it remains to control the second term in (B.43). Plugging the expression of I_h (cf. (4.30)) and invoking the fact (B.44) give

$$\begin{aligned} \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} I_h = \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\eta_0^{N_h^k(s,a)} H \\ + 2\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)} b_n \end{aligned}$$

$$\leq \underbrace{e\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\eta_0^{N_h^k(s,a)}H}_{=:A} + \underbrace{2e\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}b_n}_{=:B}. \tag{B.46}$$

Step 1: controlling the quantities A and B in (B.46). We first develop an upper bound on the quantity A in (B.46). Recognizing the fact that $\eta_0^N = 0$ for any $N > 0$ (see (4.16)), we have

$$\begin{aligned}
A &= e\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\eta_0^{N_h^k(s,a)}H \\
&\leq eH\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\mathbb{1}\big(N_h^k(s,a)<1\big) \\
&\leq eH\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\frac{8\iota}{d_h^\mu(s,a)} + eH\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=\lceil\frac{8\iota}{d_h^\mu(s,a)}\rceil}^{K}\mathbb{1}\big(N_h^k(s,a)<1\big) \\
&= eH\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi^\star(s)\big)\frac{8\iota}{d_h^\mu\big(s,\pi^\star(s)\big)} + eH\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi^\star(s)\big)\sum_{k=\lceil\frac{8\iota}{d_h^\mu(s,\pi^\star(s))}\rceil}^{K}\mathbb{1}\big(N_h^k\big(s,\pi^\star(s)\big)<1\big),
\end{aligned}$$

where the last equality holds since $\pi^\star$ is a deterministic policy (so that $d_h^{\pi^\star}(s,a)\neq 0$ only when $a=\pi^\star(s)$). Recalling $\frac{d_h^{\pi^\star}(s,a)}{d_h^\mu(s,a)}\leq C^\star$ under Assumption 1, we can further bound A by

$$\begin{aligned}
A &\leq 8eH^2SC^\star\iota + eH\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi^\star(s)\big)\sum_{k=\lceil\frac{8\iota}{d_h^\mu(s,\pi^\star(s))}\rceil}^{K}\mathbb{1}\big(N_h^k\big(s,\pi^\star(s)\big)<1\big) \\
&= 8eH^2SC^\star\iota,
\end{aligned} \tag{B.47}$$

where the last inequality follows since when $k\geq\frac{8\iota}{d_h^\mu(s,a)}$, one has — with probability at least $1-\delta$ — that

$$N_h^k(s,a)\geq\frac{kd_h^\mu(s,a)}{8\iota}\geq 1,$$

holds simultaneously for all $(s,a,h,k)\in\mathcal{S}\times\mathcal{A}\times[K]\times[H]$ (as implied by (B.2a)).

Turning to the quantity B in (B.46), one can deduce that

$$B = 2e\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}b_n$$

$$\lesssim \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sqrt{\frac{H^3\iota^2}{N_h^k(s,a)\vee 1}} = \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi^\star(s)\big)\sqrt{\frac{H^3\iota^2}{N_h^k\big(s,\pi^\star(s)\big)\vee 1}}, \tag{B.48}$$

where the inequality follows from inequality (B.17), and the last equality is valid since $\pi^\star$ is a deterministic policy.

To further control the right hand side above, Lemma 27 provides an upper bound for $\sqrt{1/\big(N_h^k\big(s,\pi^\star(s)\big)\vee 1\big)}$ which in turn leads to

$$\begin{aligned}
B &\lesssim \sqrt{H^3\iota^3}\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi^\star(s)\big)\sqrt{\frac{1}{k d_h^{\mu}\big(s,\pi^\star(s)\big)}} \\
&\lesssim \sqrt{H^3C^\star\iota^3}\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} \sqrt{d_h^{\pi^\star}\big(s,\pi^\star(s)\big)}\sqrt{\frac{1}{k}} \\
&\lesssim \sqrt{H^5C^\star K\iota^3}\max_h \sum_{s\in\mathcal{S}} \sqrt{d_h^{\pi^\star}\big(s,\pi^\star(s)\big)} \\
&\lesssim \sqrt{H^5C^\star K\iota^3}\cdot\left(\sqrt{S}\cdot\sqrt{\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\big(s,\pi^\star(s)\big)}\right) \asymp \sqrt{H^5SC^\star K\iota^3},
\end{aligned} \tag{B.49}$$

where the second inequality follows from the fact $\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)} \le C^\star$ under Assumption 1, and the last line invokes the Cauchy-Schwarz inequality.

Taking the upper bounds on both A and B collectively establishes

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} I_h \le A + B \lesssim H^2SC^\star\iota + \sqrt{H^5SC^\star K\iota^3}. \tag{B.50}$$

Step 2: putting everything together. Combining (B.45) and (B.50) allows us to establish that

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(I_h + 16\sqrt{H^2C^\star K\log\frac{2H}{\delta}} + 8HC^\star\log\frac{2H}{\delta}\right) \lesssim H^2SC^\star\iota + \sqrt{H^5SC^\star K\iota^3},$$

as advertised.

B.3 Proof of lemmas for LCB-Q-Advantage (Theorem 3)

Additional notation for LCB-Q-Advantage. Let us also introduce, and remind the reader of, several notation of interest in Algorithm 8 as follows.

- $N_h^k(s,a)$ (resp. $N_h^{(m,t)}(s,a)$) denotes the value of $N_h(s,a)$ — the number of episodes that has visited (s,a) at step h at the *beginning* of the k-th episode (resp. the *beginning* of t-th episode of the m-th epoch); for the sake of conciseness, we shall often abbreviate $N_h^k = N_h^k(s,a)$ (resp. $N_h^{(m,t)} = N_h^{(m,t)}(s,a)$) when it is clear from context.

- $L_m = 2^m$: the total number of in-epoch episodes in the m-th epoch.

- $k_h^n(s,a)$: the index of the episode in which (s,a) is visited for the n-th time at time step h; $(m_h^n(s,a), t_h^n(s,a))$ denote respectively the index of the epoch and that of the in-epoch episode in which (s,a) is visited for the n-th time at step h; for the sake of conciseness, we shall often use the shorthand $k^n = k_h^n(s,a)$, $(m^n, k^n) = (m_h^n(s,a), k_h^n(s,a))$ whenever it is clear from context.

- $Q_h^k(s,a)$, $Q_h^{\mathsf{LCB},k}(s,a)$, $\overline{Q}_h^k(s,a)$ and $V_h^k(s)$ are used to denote $Q_h(s,a)$, $Q_h^{\mathsf{LCB}}(s,a)$, $\overline{Q}_h(s,a)$, and $V_h(s)$ at the *beginning* of the k-th episode, respectively.

- $\overline{V}_h^k(s), \overline{V}_h^{\mathsf{next},k}(s), \overline{\mu}_h^k(s,a), \overline{\mu}_h^{\mathsf{next},k}(s,a)$ denote the values of $\overline{V}_h(s), \overline{V}_h^{\mathsf{next}}(s), \overline{\mu}_h(s,a)$ and $\overline{\mu}_h^{\mathsf{next}}(s,a)$ at the *beginning* of the k-th episode, respectively.

- $\widehat{N}_h^{(m,t)}(s,a)$ represents $\widehat{N}_h(s,a)$ at the *beginning* of the t-th in-epoch episode in the m-th epoch.

- $\widehat{N}_h^{\mathsf{epo},m}(s,a)$ denotes $\widehat{N}_h^{(m,L_m+1)}(s,a)$, representing the number of visits to (s,a) in the entire duration of the m-th epoch.

- $[\mu_h^{\mathsf{ref},k}, \sigma_h^{\mathsf{ref},k}, \mu_h^{\mathsf{adv},k}, \sigma_h^{\mathsf{adv},k}, \overline{\delta}_h^k, \overline{B}_h^k, \overline{b}_h^k]$: the values of $[\mu_h^{\mathsf{ref}}, \sigma_h^{\mathsf{ref}}, \mu_h^{\mathsf{adv}}, \sigma_h^{\mathsf{adv}}, \overline{\delta}_h, \overline{B}_h, \overline{b}_h]$ at the *beginning* of the k-th episode, respectively.

In addition, for a fixed vector $V \in \mathbb{R}^{|\mathcal{S}|}$, let us define a variance parameter with respect to $P_{h,s,a}$ as follows

$$\mathsf{Var}_{h,s,a}(V) \coloneqq \mathop{\mathbb{E}}_{s' \sim P_{h,s,a}} \left[\big(V(s') - P_{h,s,a}V\big)^2 \right] = P_{h,s,a}(V^2) - (P_{h,s,a}V)^2. \tag{B.51}$$

This notation will be useful in the subsequent proof. We remind the reader that there exists a one-to-one mapping between the index of the episode k and the index pair (m,t) (i.e., the epoch m and in-epoch episode t), as specified in (4.36).

In the following, for any episode k, we recall the expressions of $\overline{V}_{h+1}$ and $\overline{\mu}_h$ (which is the running mean of $\overline{V}_{h+1}$).

- Recalling the update rule of $\overline{V}_h$ and $\overline{V}_h^{\mathsf{next}}$ in line 26 and line 27 of Algorithm 8, we observe that both the reference values for the current epoch $\overline{V}_h$ and for the next epoch $\overline{V}_h^{\mathsf{next}}$ remain unchanged within each epoch. Additionally, for any epoch m, $\overline{V}_h$ takes the value of $\overline{V}_h^{\mathsf{next}}$ in

the previous $(m-1)$-th epoch; namely, for any episode k happening in the m-th epoch, we have

$$\overline{V}_h^k = \overline{V}_h^{\mathsf{next},k'} \tag{B.52}$$

for all episode k' within the $(m-1)$-th epoch.

- $\overline{\mu}_h^k$ serves as the estimate of $P_{h,s,a}\overline{V}_{h+1}^k$ constructed by the samples in the previous $(m-1)$-th epoch (collected by updating $\overline{\mu}_h^{\mathsf{next}}$). Recall the update rule of $\overline{\mu}_h$ in line 26 and line 24 of Algorithm 8: for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, we can write $\overline{\mu}_h^k$ as

$$\begin{aligned}\overline{\mu}_h^k(s,a) &= \overline{\mu}_h^{(m,1)}(s,a) = \overline{\mu}_h^{\mathsf{next},(m,1)}(s,a) = \overline{\mu}_h^{\mathsf{next},(m-1,L_{m-1})}(s,a)\\ &= \frac{\sum_{i=N_h^{(m-1,1)}+1}^{N_h^{(m,1)}} \overline{V}_{h+1}^{\mathsf{next},k^i}(s_{h+1}^{k^i})}{\widehat{N}_h^{\mathsf{epo},m-1}(s,a)\vee 1} = \frac{\sum_{i=N_h^{(m-1,1)}+1}^{N_h^{(m,1)}} \overline{V}_{h+1}^{k}(s_{h+1}^{k^i})}{\widehat{N}_h^{\mathsf{epo},m-1}(s,a)\vee 1},\end{aligned} \tag{B.53}$$

where the last equality follows from (B.52) using the fact that the indices of episodes in which (s,a) is visited within the $(m-1)$-th epoch are $\{i : i = N_h^{(m-1,1)}+1, N_h^{(m-1,1)}+2, \cdots, N_h^{(m,1)}\}$.

Finally, according to the update rules of $\mu_h^{\mathsf{adv},k^{n+1}}(s_h^k,a_h^k)$ and $\sigma_h^{\mathsf{adv},k^{n+1}}(s_h^k,a_h^k)$ in lines 11-12 of Algorithm 6, we have

$$\begin{aligned}\mu_h^{\mathsf{adv},k^{n+1}}(s_h^k,a_h^k) &= \mu_h^{\mathsf{adv},k^n+1}(s_h^k,a_h^k) = (1-\eta_n)\mu_h^{\mathsf{adv},k^n}(s_h^k,a_h^k) + \eta_n\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n})\big),\\ \sigma_h^{\mathsf{adv},k^{n+1}}(s_h^k,a_h^k) &= \sigma_h^{\mathsf{adv},k^n+1}(s_h^k,a_h^k) = (1-\eta_n)\sigma_h^{\mathsf{adv},k^n}(s_h^k,a_h^k) + \eta_n\big(V_{h+1}^{k^n}(s_{h+1}^{k^n}) - \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n})\big)^2.\end{aligned}$$

Applying this relation recursively and invoking the definitions of $\eta_n^{N_h^k}$ in (4.16) give

$$\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s,a) = \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\big),\quad \sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s,a) = \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_h^{k^n}\big(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\big)^2. \tag{B.54}$$

Similarly, according to the update rules of $\mu_h^{\mathsf{ref},k^{n+1}}(s,a)$ and $\sigma_h^{\mathsf{ref},k^{n+1}}(s,a)$ in lines 9-10 of Algorithm 6, we obtain

$$\begin{aligned}\mu_h^{\mathsf{ref},k^{n+1}}(s,a) &= \mu_h^{\mathsf{ref},k^n+1}(s,a) = \left(1-\frac{1}{n}\right)\mu_h^{\mathsf{ref},k^n}(s,a) + \frac{1}{n}\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n}),\\ \sigma_h^{\mathsf{ref},k^{n+1}}(s,a) &= \sigma_h^{\mathsf{ref},k^n+1}(s,a) = \left(1-\frac{1}{n}\right)\sigma_h^{\mathsf{ref},k^n}(s,a) + \frac{1}{n}\left(\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)^2.\end{aligned}$$

Simple recursion leads to

$$\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s,a) = \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}\overline{V}_{h+1}^{\mathsf{next},k^n}, \quad \sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s,a) = \frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}\big(\overline{V}_{h+1}^{\mathsf{next},k^n}\big)^2. \tag{B.55}$$

B.3.1 Proof of Lemma 10

Akin to the proof of Lemma 7, the second inequality of (4.38) holds trivially since

$$V_h^{\pi}(s) \le V_h^{\star}(s)$$

holds for any policy π. Thus, it suffices to focus on justifying the first inequality of (4.38), namely,

$$V_h^k(s) \le V_h^{\pi^k}(s) \qquad \forall (k,h,s) \in [K]\times[H]\times\mathcal{S}, \tag{B.56}$$

which we shall prove by induction.

Step 1: introducing the induction hypothesis. For notational simplicity, let us define

$$k_o(h,k,s) \coloneqq \max\Big\{l : l<k \text{ and } V_h^l(s) = \max_a \max\Big\{Q_h^{\mathsf{LCB},l}(s,a), \overline{Q}_h^l(s,a)\Big\}\Big\} \tag{B.57}$$

for any $(h,k,s)\in[H]\times[K]\times\mathcal{S}$. Here, $k_o(h,k,s)$ denotes the index of the latest episode — right at the end of the $(k-1)$-th episode — in which $V_h(s)$ has been updated, which shall be abbreviated as $k_o(h)$ whenever it is clear from context.

In what follows, we shall first justify the advertised inequality for the base case where $h=H+1$ for all episodes $k\in[K]$, followed by an induction argument. Regarding the induction part, let us consider any $k\in[K]$ and any $h\in[H]$, and suppose that

$$V_{h'}^{k'}(s) \le V_{h'}^{\pi^{k'}}(s) \qquad \text{for all } (k',h',s)\in[k-1]\times[H+1]\times\mathcal{S}, \tag{B.58a}$$

$$V_{h'}^{k}(s) \le V_{h'}^{\pi^{k}}(s) \qquad \text{for all } h'\ge h+1 \text{ and } s\in\mathcal{S}. \tag{B.58b}$$

We intend to justify

$$V_h^k(s) \le V_h^{\pi^k}(s) \qquad \forall s\in\mathcal{S}, \tag{B.59}$$

assuming that the induction hypotheses (B.58) hold.

Step 2: controlling the confident bound $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{b}_h^{k^n+1}$. Before proceeding, we first introduce an auxiliary result on bounding $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\overline{b}_h^{k^n+1}$, which plays a crucial role. For any (s,a), it

is easily seen that

$$N_h^k(s,a)=0 \qquad \Longrightarrow \qquad \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \overline{b}_h^{k^n(s,a)+1} = 0. \tag{B.60}$$

When $N_h^k(s,a)>0$, expanding the definitions of $\overline{b}_h^{k^n+1}$ (cf. line 6 of Algorithm 6) and $\overline{\delta}_h^{k+1}$ (cf. line 15 of Algorithm 6) leads to

$$\begin{aligned}
&\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1} \\
&= \sum_{n=1}^{N_h^k} \eta_n \prod_{i=n+1}^{N_h^k} (1-\eta_i) \cdot \left(\left(1-\frac{1}{\eta_n}\right) \overline{B}_h^{k^n}(s,a) + \frac{1}{\eta_n} \overline{B}_h^{k^n+1}(s,a) \right) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^{7/4}\iota + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} H^2\iota \\
&= \sum_{n=1}^{N_h^k} \left(\prod_{i=n+1}^{N_h^k} (1-\eta_i) \overline{B}_h^{k^n+1}(s,a) - \prod_{i=n}^{N_h^k} (1-\eta_i) \overline{B}_h^{k^n}(s,a) \right) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^{7/4}\iota + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} H^2\iota \\
&\overset{\text{(i)}}{=} \sum_{n=1}^{N_h^k} \prod_{i=n+1}^{N_h^k} (1-\eta_i) \overline{B}_h^{k^n+1}(s,a) - \sum_{n=2}^{N_h^k} \prod_{i=n}^{N_h^k} (1-\eta_i) \overline{B}_h^{k^n}(s,a) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^{7/4}\iota + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} H^2\iota \\
&\overset{\text{(ii)}}{=} \sum_{n=1}^{N_h^k} \prod_{i=n+1}^{N_h^k} (1-\eta_i) \overline{B}_h^{k^n+1}(s,a) - \sum_{n=1}^{N_h^k-1} \prod_{i=n+1}^{N_h^k} (1-\eta_i) \overline{B}_h^{k^n+1}(s,a) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^{7/4}\iota + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} H^2\iota \\
&= \overline{B}_h^{k^{N_h^k}+1}(s,a) + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} H^{7/4}\iota + c_{\mathrm{b}} \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} H^2\iota,
\end{aligned} \tag{B.61}$$

where we abuse the notation to let $\prod_{i=j+1}^{j}(1-\eta_i)=1$. Here, (i) holds since $\overline{B}^{k^1}(s,a)=0$, (ii) follows from the fact that $\overline{B}^{k^n+1}(s,a)=\overline{B}^{k^{n+1}}(s,a)$, since (s,a) has not been visited at step h during the episodes between the indices k^n+1 and $k^{n+1}-1$. Combining the above result in (B.61) with the properties $\frac{1}{(N_h^k)^{3/4}} \le \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n^{3/4}} \le \frac{2}{(N_h^k)^{3/4}}$ and $\frac{1}{N_h^k} \le \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{n} \le \frac{2}{N_h^k}$ (see Lemma 1), we arrive at

$$\overline{B}_h^{k^{N_h^k}+1}(s,a) + c_{\mathrm{b}} \frac{H^{7/4}\iota}{(N_h^k)^{3/4}} + c_{\mathrm{b}} \frac{H^2\iota}{N_h^k} \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1} \le \overline{B}_h^{k^{N_h^k}+1}(s,a) + 2c_{\mathrm{b}} \frac{H^{7/4}\iota}{(N_h^k)^{3/4}} + 2c_{\mathrm{b}} \frac{H^2\iota}{N_h^k} \tag{B.62}$$

as long as $N_h^k(s,a)>0$.

Step 3: base case. Let us look at the base case with $h = H + 1$ for any $k \in [K]$. Recalling the facts that $V_{H+1}^{\pi} = V_{H+1}^{k} = 0$ for any π and any $k \in [K]$, we reach

$$V_{H+1}^{k}(s) \le V_{H+1}^{\pi^k}(s) \qquad \text{for all } (k,s) \in [K] \times \mathcal{S}. \tag{B.63}$$

Step 4: induction arguments. We now turn to the induction arguments. Suppose that (B.58) holds for a pair $(k,h) \in [K] \times [H]$. Everything comes down to justifying (B.59) for time step h in the episode k.

First, we recall the update rule of $V_h(s)$ in lines 21-22 of Algorithm 8:

$$V_h^k(s) = \max_a Q_h^k(s,a) = Q_h^k\big(s, \pi_h^k(s)\big) = \max\left\{Q_h^{\mathsf{LCB},k}\left(s, \pi_h^k(s)\right), \overline{Q}_h^k\left(s, \pi_H^k(s)\right), Q_h^{k-1}\left(s, \pi_h^k(s)\right)\right\}.$$

Then we shall verify (B.59) in three different cases.

- When $V_h^k(s) = Q_h^{\mathsf{LCB},k}\left(s, \pi_h^k(s)\right)$, the term of interest can be controlled by

$$V_h^{\pi^k}(s) - V_h^k(s) \overset{\text{(i)}}{=} Q_h^{\pi^k}\left(s, \pi_h^k(s)\right) - Q_h^{\mathsf{LCB},k}\left(s, \pi_h^k(s)\right) \ge 0,$$

  where (i) holds since π^k is set to be the greedy policy such that $V_h^{\pi^k}(s) = Q_h^{\pi^k}(s, \pi_h^k(s))$, and the last inequality follows directly from the analysis for LCB-Q (see (B.27)).

- When $V_h^k(s) = \overline{Q}_h^k\left(s, \pi_h^k(s)\right)$, we obtain

$$V_h^{\pi^k}(s) - V_h^k(s) = Q_h^{\pi^k}\left(s, \pi_h^k(s)\right) - \overline{Q}_h^k\left(s, \pi_h^k(s)\right). \tag{B.64}$$

  To prove the term on the right-hand side of (B.64) is non-negative, we proceed by developing a more general lower bound on $Q_h^{\pi^k}(s,a) - \overline{Q}_h^k(s,a)$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$. Towards this, recalling the definition of N_h^k and k^n, we can express

$$\overline{Q}_h^k(s,a) = \overline{Q}_h^{k^{N_h^k+1}}(s,a).$$

  Thus, according to the update rule (cf. line 7 in Algorithm 6), we arrive at

$$\begin{aligned}\overline{Q}_h^k(s,a) &= \overline{Q}_h^{k^{N_h^k+1}}(s,a) \\ &= (1-\eta_{N_h^k})\overline{Q}_h^{k^{N_h^k}}(s,a) + \eta_{N_h^k}\left\{r_h(s,a) + V_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) - \overline{V}_{h+1}^{k^{N_h^k}}(s_{h+1}^{k^{N_h^k}}) + \overline{\mu}_h^{k^{N_h^k}}(s,a) - \overline{b}_h^{k^{N_h^k+1}}\right\}.\end{aligned}$$

Applying this relation recursively and invoking the definitions of $\eta_0^{N_h^k}$ and $\eta_n^{N_h^k}$ in (4.16) give

$$\overline{Q}_h^k(s,a) = \eta_0^{N_h^k}\overline{Q}_h^1(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left\{r_h(s,a) + V_{h+1}^{k^n}(s_{h+1}^{k^n}) - \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n}) + \overline{\mu}_h^{k^n}(s,a) - \overline{b}_h^{k^n+1}\right\}. \tag{B.65}$$

Additionally, for any policy π^k, the basic relation $\eta_0^{N_h^k} + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see (4.17) and (4.16)) gives

$$Q_h^{\pi^k}(s,a) = \eta_0^{N_h^k} Q_h^{\pi^k}(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} Q_h^{\pi^k}(s,a). \tag{B.66}$$

Combing (B.65) and (B.66) leads to

$$\begin{aligned} Q_h^{\pi^k}(s,a) - \overline{Q}_h^k(s,a) &= \eta_0^{N_h^k}\left(Q_h^{\pi^k}(s,a) - \overline{Q}_h^1(s,a)\right) \\ &\quad + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left\{Q_h^{\pi^k}(s,a) - r_h(s,a) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n}) - \overline{\mu}_h^{k^n}(s,a) + \overline{b}_h^{k^n+1}\right\}. \end{aligned} \tag{B.67}$$

Plugging in the construction of $\overline{\mu}_h$ in (B.53) and invoking the Bellman equation

$$Q_h^{\pi^k}(s,a) = r_h(s,a) + P_{h,s,a}V_{h+1}^{\pi^k}, \tag{B.68}$$

we arrive at

$$\begin{aligned} &Q_h^{\pi^k}(s,a) - r_h(s,a) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n}) - \overline{\mu}_h^{k^n}(s,a) + \overline{b}_h^{k^n+1} \\ &= P_{h,s,a}V_{h+1}^{\pi^k} + \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n}) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^i})}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} + \overline{b}_h^{k^n+1} \\ &= P_{h,s,a}V_{h+1}^{\pi^k} - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + \left(P_h^{k^n} - P_{h,s,a}\right)\overline{V}_{h+1}^{k^n} + \left(P_{h,s,a} - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1}\right)\overline{V}_{h+1}^{k^n} + \overline{b}_h^{k^n+1} \\ &= P_{h,s,a}\left(V_{h+1}^{\pi^k} - V_{h+1}^{k^n}\right) + \overline{b}_h^{k^n+1} + \xi_h^{k^n}, \end{aligned}$$

where

$$\xi_h^{k^n} := \left(P_h^{k^n} - P_{h,s,a}\right)\left(\overline{V}_{h+1}^{k^n} - V_{h+1}^{k^n}\right) + \left(P_{h,s,a} - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1}\right)\overline{V}_{h+1}^{k^n}. \tag{B.69}$$

Inserting the above result into (B.67) leads to the following decomposition

$$Q_h^{\pi^k}(s,a) - \overline{Q}_h^k(s,a) = \eta_0^{N_h^k}\left(Q_h^{\pi^k}(s,a) - \overline{Q}_h^1(s,a)\right) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\left\{P_{h,s,a}\left(V_{h+1}^{\pi^k} - V_{h+1}^{k^n}\right) + \overline{b}_h^{k^n+1} + \xi_h^{k^n}\right\} \tag{B.70}$$

$$\geq \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}(\overline{b}_h^{k^n+1} + \xi_h^{k^n}), \tag{B.71}$$

which holds by virtue of the following facts:

(i) The initialization $\overline{Q}_h^1(s,a) = 0$ and the non-negativity of $Q_h^\pi(s,a)$ for any policy π and $(s,a) \in \mathcal{S} \times \mathcal{A}$ lead to $Q_h^{\pi^k}(s,a) - \overline{Q}_h^1(s,a) = Q_h^{\pi^k}(s,a) \geq 0$.

(ii) For any episode k^n appearing before k, making use of the induction hypothesis $V_{h+1}^{\pi^k}(s) \geq V_{h+1}^k(s)$ in (B.58b) and the monotonicity of $V_h(s)$ in (4.37), we obtain

$$V_{h+1}^{\pi^k}(s) - V_{h+1}^{k^n}(s) \geq V_{h+1}^k(s) - V_{h+1}^{k^n}(s) \geq 0. \tag{B.72}$$

The following lemma ensures that the right-hand side of (B.71) is non-negative. We postpone the proof of Lemma 30 to Appendix B.3.4 to streamline our discussion.

Lemma 30. *For any $\delta \in (0,1)$, there exists some sufficiently large constant $c_{\mathrm{b}} > 0$, such that with probability at least $1-\delta$,*

$$\left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n}\right| \leq \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1}, \qquad \forall k \in [K]. \tag{B.73}$$

Taking this lemma together with the inequalities (B.64) and (B.71) yields

$$V_h^{\pi^k}(s) - V_h^k(s) = Q_h^{\pi^k}(s,a) - \overline{Q}_h^k(s,a) \geq \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1} - \left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n}\right| \geq 0.$$

- Next, consider the case where $V_h^k(s) = Q_h^{k-1}\left(s, \pi_h^k(s)\right)$. In view of the definition of $k_o(h)$ in (B.57), one has

$$V_h^k(s) = Q_h^{k-1}\left(s, \pi_h^k(s)\right) = Q_h^{k_o(h)}\left(s, \pi_h^k(s)\right) = \max\left\{Q_h^{\mathsf{LCB},k_o(h)}\left(s, \pi_h^k(s)\right), \overline{Q}_h^{k_o(h)}\left(s, \pi_h^k(s)\right)\right\},$$

since $Q_h\left(s, \pi_h^k(s)\right)$ has not been updated during the episode $k_o(h)$ and remains unchanged in the episodes $k_o(h)+1, k_o(h)+2, \cdots, k-1$. With this equality in hand, the term of interest

in (B.59) can be controlled by

$$V_h^{\pi^k}(s) - V_h^k(s) = Q_h^{\pi^k}(s, \pi_h^k(s)) - \max\left\{Q_h^{\mathsf{LCB},k_o(h)}\left(s, \pi_h^k(s)\right), \overline{Q}_h^{k_o(h)}\left(s, \pi_h^k(s)\right)\right\} \geq 0,$$

where the last inequality follows from the facts

$$Q_h^{\pi^k}(s, \pi_h^k(s)) - Q_h^{\mathsf{LCB},k_o(h)}(s, \pi_h^k(s)) \overset{\text{(i)}}{\geq} 0,$$
$$Q_h^{\pi^k}(s, \pi_h^k(s)) - \overline{Q}_h^{k_o(h)}(s, \pi_h^k(s)) \overset{\text{(ii)}}{\geq} 0.$$

Here, (i) follows from the same analysis framework for showing (B.26) and (B.28); (ii) holds due to the following fact

$$Q_h^{\pi^k}(s,a) - \overline{Q}_h^{k_o(h)}(s,a) \geq \sum_{n=1}^{N_h^{k_o(h)}} \eta_n^{N_h^{k_o(h)}} (\bar{b}_h^{k^n+1} + \xi_h^{k^n}) \geq 0,$$

which is obtained directly by adapting (B.71) and then invoking (B.73) for $k = k_o(h)$; since the analysis follows verbatim, we omit their proofs here.

Combining the above three cases verifies the induction hypothesis in (B.59), provided that (B.58) is satisfied.

Step 5: putting everything together. Combining the base case in Step 3 and induction arguments in Step 4, we can readily verify the induction hypothesis in Step 1, which in turn establishes Lemma 10.

B.3.2 Proof of Lemma 11

For every $h \in [H]$, we can decompose

$$\begin{aligned}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s) - V_h^k(s)\right) &\overset{\text{(i)}}{\leq} \sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}\left(s, \pi_h^\star(s)\right)\left(Q_h^\star\left(s, \pi_h^\star(s)\right) - \overline{Q}_h^k\left(s, \pi_h^\star(s)\right)\right) \\ &= \sum_{k=1}^{K}\sum_{s,a\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left(Q_h^\star(s,a) - \overline{Q}_h^k(s,a)\right), \end{aligned} \tag{B.74}$$

where (i) follows from the fact $V_h^k(s) = \max_a Q_h^k(s,a) \geq \max_a \overline{Q}_h^k(s,a) \geq \overline{Q}_h^k(s, \pi_h^\star(s))$ (see lines 21-22 in Algorithm 8). Here, the last equality is due to (4.26).

Step 1: bounding $Q_h^\star(s,a) - \overline{Q}_h^k(s,a)$. The basic relation $\eta_0^{N_h^k} + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see (4.17) and (4.16)) gives

$$Q_h^\star(s,a) = \eta_0^{N_h^k} Q_h^\star(s,a) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} Q_h^\star(s,a), \tag{B.75}$$

which combined with (B.65) leads to

$$\begin{aligned} Q_h^\star(s,a) - \overline{Q}_h^k(s,a) &= \eta_0^{N_h^k}\left(Q_h^\star(s,a) - \overline{Q}_h^1(s,a)\right) \\ &\quad + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left\{ Q_h^\star(s,a) - r_h(s,a) - V_{h+1}^{k^n}(s_{h+1}^{k^n}) + \overline{V}_{h+1}^{k^n}(s_{h+1}^{k^n}) - \overline{\mu}_h^{k^n}(s,a) + \overline{b}_h^{k^n+1} \right\}. \end{aligned} \tag{B.76}$$

Invoking the Bellman optimality equation

$$Q_h^\star(s,a) = r_h(s,a) + P_{h,s,a} V_{h+1}^\star, \tag{B.77}$$

we can decompose $Q_h^\star(s,a) - \overline{Q}_h^k(s,a)$ similar to (B.70) by inserting (B.69) as follows:

$$\begin{aligned} Q_h^\star(s,a) - \overline{Q}_h^k(s,a) &= \eta_0^{N_h^k}\left(Q_h^\star(s,a) - \overline{Q}_h^1(s,a)\right) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left\{ P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^{k^n}\right) + \overline{b}_h^{k^n+1} + \xi_h^{k^n} \right\} \\ &\overset{\text{(i)}}{\le} \eta_0^{N_h^k} H + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(\overline{b}_h^{k^n+1} + \xi_h^{k^n}\right) + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^{k^n}\right) \\ &\overset{\text{(ii)}}{\le} \eta_0^{N_h^k} H + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^{k^n}\right) + 2\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1} \\ &\overset{\text{(iii)}}{\le} \eta_0^{N_h^k} H + \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^{k^n}\right) + 2\left(\overline{B}_h^k(s,a) + 2c_{\mathrm{b}} \frac{H^{7/4}\iota}{\left(N_h^k \vee 1\right)^{3/4}} + 2c_{\mathrm{b}} \frac{H^2\iota}{N_h^k \vee 1}\right), \end{aligned} \tag{B.78}$$

where (i) follows from the initialization $\overline{Q}_h^1(s,a) = 0$ and the trivial upper bound $Q_h^\pi(s,a) \le H$ for any policy π, (ii) holds owing to the fact (see (B.73))

$$\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(\overline{b}_h^{k^n+1} + \xi_h^{k^n}\right) \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1} + \left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n} \right| \le 2\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1}, \tag{B.79}$$

and (iii) comes from (B.62) with the fact $\overline{B}_h^{k^{N_h^k+1}}(s,a) = \overline{B}_h^k(s,a)$.

Step 2: decomposing the error in (B.74). Plugging (B.78) into (B.74) and rearranging terms yield

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right) \tag{B.80}$$

$$\begin{aligned}
&\le \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left[\eta_0^{N_h^k(s,a)}H + 2\overline{B}_h^k(s,a) + \frac{4c_{\mathrm{b}}H^{7/4}\iota}{\left(N_h^k(s,a)\vee 1\right)^{3/4}} + \frac{4c_{\mathrm{b}}H^2\iota}{N_h^k(s,a)\vee 1}\right] \\
&\quad + \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\right) \\
&\le \underbrace{\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left[\eta_0^{N_h^k(s,a)}H + \frac{4c_{\mathrm{b}}H^{7/4}\iota}{\left(N_h^k(s,a)\vee 1\right)^{3/4}} + \frac{4c_{\mathrm{b}}H^2\iota}{N_h^k(s,a)\vee 1}\right]}_{=:J_h^1} + \underbrace{2\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\overline{B}_h^k(s,a)}_{=:J_h^2} \\
&\quad + \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\right).
\end{aligned} \tag{B.81}$$

Step 3: controlling the last term in (B.81). If we could verify the following result

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\right) \\
&\le \underbrace{\left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right) + 48\sqrt{HC^\star K\log\frac{2H}{\delta}} + 28c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2}_{=:J_h^3},
\end{aligned} \tag{B.82}$$

then combining this result with inequality (B.81) would immediately establish Lemma 11. As a result, it suffices to verify the inequality (B.82), which shall be accomplished as follows.

Proof of inequality (B.82). We first make the observation that the left-hand side of inequality (B.82) is the same as what Lemma 8 shows. Therefore, we shall establish this inequality following the same framework as in Appendix B.2.2. To begin with, let us recall several definitions in Appendix B.2.2:

$$A_h := \underbrace{\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\right)}_{=:A_{h,k}},$$

$$
\begin{aligned}
B_{h,k} &:= \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi_\star}(s)\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right), \\
Y_{h,k} &= \frac{d_h^{\pi_\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k} \sum_{n=1}^{N_h^k(s_h^k,a_h^k)} \eta_n^{N_h^k(s_h^k,a_h^k)} \left(V_{h+1}^{\star}-V_{h+1}^{k^n(s_h^k,a_h^k)}\right), \\
Z_{h,k} &= \left(1+\frac{1}{H}\right)\frac{d_h^{\pi_\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)} P_{h,s_h^k,a_h^k}\left(V_{h+1}^{\star}-V_{h+1}^{k}\right),
\end{aligned} \tag{B.83}
$$

and we also remind the reader of the relation in (B.35) as follows

$$
A_h \le \sum_{k=1}^{K} B_{h,k} + \sum_{k=1}^{K}\left(Z_{h,k}-B_{h,k}\right) + \sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right). \tag{B.84}
$$

Equipped with these relations, we aim to control $\sum_{k=1}^{K}\left(Z_{h,k}-B_{h,k}\right)$ and $\sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right)$ respectively as follows.

- We first bound $\sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right)$, which is similar to (B.38) (as controlled by Lemma 29). Repeating the argument and tightening the bound from the second line of (B.38), we have for all $(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$,

$$
\begin{aligned}
\left|\sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right)\right| &\le \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[P_{h,s,a}W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \\
&\le \sqrt{8C^\star\log\frac{2H}{\delta}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_{h,s,a}\left(V_{h+1}^{\star}-V_{h+1}^{k^n(s,a)}\right)\right]^2} + 4HC^\star\log\frac{2H}{\delta} \\
&\overset{\text{(i)}}{\le} \sqrt{8C^\star\log\frac{2H}{\delta}\left(36HK+3c_{\mathrm{a}}^2H^6SC^\star\iota\right)} + 4HC^\star\log\frac{2H}{\delta} \\
&\le 32\sqrt{HC^\star K\log\frac{2H}{\delta}} + 12c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2.
\end{aligned} \tag{B.85}
$$

  Here, (i) holds by virtue of the following fact

$$
\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_{h,s,a}\left(V_{h+1}^{\star}-V_{h+1}^{k^n(s,a)}\right)\right]^2 \le 36HK+3c_{\mathrm{a}}^2H^6SC^\star\iota, \tag{B.86}
$$

  whose proof is postponed to Appendix B.3.2.1.

- Next, we turn to $\sum_{k=1}^{K}\left(Z_{h,k}-B_{h,k}\right)$, which can be bounded similar to (B.42) (as controlled via Lemma 29). Repeating the argument and tightening the bound from the second line of

(B.42) yield

$$\left|\sum_{k=1}^{K}(B_{h,k}-Z_{h,k})\right| \leq \sqrt{\sum_{k=1}^{K} 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[P_{h,s,a}W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta}$$
$$\leq 8\sqrt{C^\star \log\frac{2H}{\delta}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^k\right)\right]^2} + 8HC^\star\log\frac{2H}{\delta}. \tag{B.87}$$

To further control (B.87), we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^k\right)\right]^2 &\overset{\text{(i)}}{\leq} \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^k\right)^2 \\ &\overset{\text{(ii)}}{\leq} H\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^k\right) \\ &\overset{\text{(iii)}}{\leq} 2HK + c_{\mathrm{a}}^2 H^6 SC^\star\iota. \end{aligned}\tag{B.88}$$

Here, (i) holds due to the non-negativity of the variance

$$\mathsf{Var}_{h,s,a}(V_{h+1}^\star - \overline{V}_{h+1}^k) = P_{h,s,a}(V_{h+1}^\star - V_{h+1}^k)^2 - \left(P_{h,s,a}(V_{h+1}^\star - V_{h+1}^k)\right)^2 \geq 0; \tag{B.89}$$

(ii) follows from the basic property $\left\|V_{h+1}^\star - V_{h+1}^k\right\|_\infty \leq H$; to see why (iii) holds, we refer the reader to (B.96), which will be proven in Appendix B.3.2.1 as well. Inserting (B.88) back into (B.87) yields

$$\begin{aligned}\left|\sum_{k=1}^{K}(B_{h,k}-Z_{h,k})\right| &\leq 8\sqrt{C^\star\log\frac{2H}{\delta}\left(2KH + c_{\mathrm{a}}^2H^6SC^\star\iota\right)} + 8HC^\star\log\frac{2H}{\delta} \\ &\leq 16\sqrt{HC^\star K\log\frac{2H}{\delta}} + 16c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota. \end{aligned}\tag{B.90}$$

Substituting the inequalities (B.85) and (B.90) into (B.84), and using the definitions in (B.83), we arrive at

$$\begin{aligned}A_h &= \sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^\star - V_{h+1}^{k^n(s,a)}\right) \\ &\leq \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s) - V_{h+1}^k(s)\right) + \sum_{k=1}^{K}(Z_{h,k}-B_{h,k}) + \sum_{k=1}^{K}(A_{h,k}-Y_{h,k})\end{aligned}$$

$$\leq \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right)+32\sqrt{HC^\star K\log\frac{2H}{\delta}}+12c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2$$
$$+16\sqrt{HC^\star K\log\frac{2H}{\delta}}+16c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota$$
$$\leq \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right)+48\sqrt{HC^\star K\log\frac{2H}{\delta}}+28c_{\mathrm{a}}H^3C^\star\sqrt{S}\iota^2, \quad \text{(B.91)}$$

which directly verifies (B.82) and completes the proof.

B.3.2.1 Proof of inequality (B.86)

Step 1: rewriting the term of interest. We first invoke Jensen's inequality to obtain

$$\Big(\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\left(V_{h+1}^\star-V_{h+1}^{k^n}\right)\Big)^2\leq\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\Big(P_{h,s,a}\left(V_{h+1}^\star-V_{h+1}^{k^n}\right)\Big)^2\leq\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}P_{h,s,a}\left(V_{h+1}^\star-V_{h+1}^{k^n}\right)^2,$$

where the first inequality follows from $\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}=1$ (see (4.17) and (4.16)), and the last inequality holds by the non-negativity of the variance $\mathsf{Var}_{h,s,a}[V_{h+1}^\star-V_{h+1}^{k^n}]$. This allows one to derive

$$\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_{h,s,a}\left(V_{h+1}^\star-V_{h+1}^{k^n}\right)\right]^2$$
$$\leq\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^\star-V_{h+1}^{k^n}\right)^2$$
$$\overset{\text{(i)}}{\leq}\left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s)-V_{h+1}^k(s)\right)^2+32\sqrt{H^4C^\star K\log\frac{2H}{\delta}}+32H^2C^\star\log\frac{2H}{\delta}, \quad \text{(B.92)}$$

where (i) can be verified in a way similar to the proof of Lemma 8 in Appendix B.2.2. We omit the details for conciseness.

Step 2: controlling the first term in (B.92). Let us introduce the following short-hand notation

$$k_{\mathsf{stop}}:=c_{\mathrm{a}}^2H^5SC^\star\iota,$$

and decompose the term in (B.92) as follows

$$\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{k=1}^{K}\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right)^2 \overset{\text{(i)}}{\le} H\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right)$$
$$= H\sum_{k=1}^{k_{\mathsf{stop}}}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\big(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\big) + H\sum_{k=k_{\mathsf{stop}}+1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\big(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\big). \quad \text{(B.93)}$$

Here, (i) holds since $0 \le V_{h+1}^{\star}(s) - V_{h+1}^{k}(s) \le H$. The first term in (B.93) satisfies

$$H\sum_{k=1}^{k_{\mathsf{stop}}}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\big(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\big) \le H\left(c_{\mathrm{a}}\sqrt{H^5SC^\star\iota k_{\mathsf{stop}}} + c_{\mathrm{a}}H^2SC^\star\iota\right) \le c_{\mathrm{a}}^2H^6SC^\star\iota, \quad \text{(B.94)}$$

where the first inequality holds by applying the results of LCB-Q in (4.35) with $K = k_{\mathsf{stop}}$. The second term in (B.93) can be controlled as follows:

$$\begin{aligned} H\sum_{k=k_{\mathsf{stop}}+1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\big(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\big) &\le HK\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}(s)-V_{h+1}^{k_{\mathsf{stop}}}(s)\right) \\ &\le HK\frac{1}{k_{\mathsf{stop}}}\sum_{k=1}^{k_{\mathsf{stop}}}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right) \\ &\le HK\left(c_{\mathrm{a}}\sqrt{\frac{H^5SC^\star\iota}{k_{\mathsf{stop}}}} + \frac{c_{\mathrm{a}}H^2SC^\star\iota}{k_{\mathsf{stop}}}\right) \le 2HK, \quad \text{(B.95)} \end{aligned}$$

where the first and the second inequalities hold by the monotonicity property $V_{h+1}^{k+1} \ge V_{h+1}^{k}$ introduced in (4.37), and the final inequality follows from applying (4.35).

Inserting the results in (B.94) and (B.95) into (B.93) yields

$$\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{k=1}^{K}\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right)^2 \le H\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right) \le 2HK + c_{\mathrm{a}}^2H^6SC^\star\iota. \quad \text{(B.96)}$$

Step 3: combining the above results. Inserting the above result (B.96) back into (B.92), we reach:

$$\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)\left[\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_{h,s,a}\left(V_{h+1}^{\star}-V_{h+1}^{k^n}\right)\right]^2$$

$$
\begin{aligned}
&\leq \left(1+\frac{1}{H}\right)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}-V_{h+1}^{k}\right)^2 + 32\sqrt{H^4C^\star K\log\frac{2H}{\delta}} + 32H^2C^\star\log\frac{2H}{\delta} \\
&\overset{\text{(i)}}{\leq} 4HK + 2c_{\mathrm{a}}^2H^6SC^\star\iota + 32\sqrt{H^4C^\star K\log\frac{2H}{\delta}} + 32H^2C^\star\log\frac{2H}{\delta} \\
&\overset{\text{(ii)}}{\leq} 36HK + 3c_{\mathrm{a}}^2H^6SC^\star\iota,
\end{aligned} \tag{B.97}
$$

where (i) holds due to (B.96) and $1+\frac{1}{H}\leq 2$, and (ii) results from the Cauchy-Schwarz inequality.

B.3.3 Proof of Lemma 12

We shall verify the three inequalities in (4.45) separately.

B.3.3.1 Proof of inequality (4.45a)

We start by rewriting the term of interest using the expression of J_h^1 in (4.42) as

$$
\begin{aligned}
&\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}J_h^1 \\
&=\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left[\eta_0^{N_h^k(s,a)}H + \frac{4c_{\mathrm{b}}H^{7/4}\iota}{\left(N_h^k(s,a)\vee 1\right)^{3/4}} + \frac{4c_{\mathrm{b}}H^2\iota}{N_h^k(s,a)\vee 1}\right] \\
&=\underbrace{\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\eta_0^{N_h^k(s,a)}H}_{=:\mathcal{J}_1^1} + \underbrace{\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\frac{4c_{\mathrm{b}}H^{7/4}\iota}{(N_h^k(s,a)\vee 1)^{3/4}}}_{=:\mathcal{J}_1^2} \\
&\quad+\underbrace{\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\frac{4c_{\mathrm{b}}H^2\iota}{N_h^k(s,a)\vee 1}}_{=:\mathcal{J}_1^3}.
\end{aligned} \tag{B.98}
$$

Invoking (B.47) and (B.44) yields

$$
\mathcal{J}_1^1 \lesssim H^2SC^\star\iota. \tag{B.99}
$$

In terms of $\mathcal{J}_1^2$, one has

$$
\begin{aligned}
\mathcal{J}_1^2 &= \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\frac{4c_{\mathrm{b}}H^{7/4}\iota}{(N_h^k(s,a)\vee 1)^{\frac{3}{4}}} \\
&\overset{\text{(i)}}{\lesssim} H^{7/4}\iota^2\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\frac{1}{(kd_h^{\mu}(s,a))^{\frac{3}{4}}}
\end{aligned}
$$

$$\overset{\text{(ii)}}{\lesssim} H^{7/4}\iota^2(C^\star)^{\frac{3}{4}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{1}{k^{\frac{3}{4}}}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(d_h^{\pi^\star}(s,a)\right)^{\frac{1}{4}}$$
$$= H^{7/4}\iota^2(C^\star)^{\frac{3}{4}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{1}{k^{\frac{3}{4}}}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{1}\left(a=\pi_h^\star(s)\right)\left(d_h^{\pi^\star}(s,a)\right)^{\frac{1}{4}},$$

where (i) holds due to (B.44) and $\frac{1}{N_h^k(s,a)\vee 1}\le\frac{8\iota}{kd_h^\mu(s,a)}$ from Lemma 27, and (ii) follows from the definition of $C^\star$ in Assumption 1. A direct application of Hölder's inequality leads to

$$\mathcal{J}_1^2 \le H^{7/4}\iota^2(C^\star)^{\frac{3}{4}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{1}{k^{\frac{3}{4}}}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{1}(a=\pi_h^\star(s))\right)^{3/4}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^\star}(s,a)\right)^{1/4}$$
$$\overset{\text{(iii)}}{\le} H^{7/4}\iota^2(SC^\star)^{\frac{3}{4}}\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{1}{k^{\frac{3}{4}}}\lesssim H^{2.75}(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2, \tag{B.100}$$

where (iii) follows since $\pi^\star$ is assumed to be a deterministic policy.

Similarly, we can derive an upper bound on $\mathcal{J}_1^3$ as follows:

$$\mathcal{J}_1^3 = \sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^\star}(s,a)\frac{4c_{\mathrm{b}}H^2\iota}{N_h^k(s,a)\vee 1}$$
$$\overset{\text{(i)}}{\lesssim} H^2\iota^2\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^\star}(s,a)}{kd_h^\mu(s,a)}\lesssim H^3SC^\star\iota^3, \tag{B.101}$$

where (i) follows from the result in (B.44) and the fact $\frac{1}{N_h^k(s,a)\vee 1}\le\frac{8\iota}{kd_h^\mu(s,a)}$ (cf. Lemma 27), and the last relation results from the definition of $C^\star$ (cf. Assumption 1) and the assumption that $\pi^\star$ is a deterministic policy.

Putting the preceding results (B.99), (B.100) and (B.101) together, we conclude that

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}J_h^1\lesssim H^{2.75}(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2+H^3SC^\star\iota^3. \tag{B.102}$$

B.3.3.2 Proof of inequality (4.45b)

Making use of the definition of $\overline{B}_h^k(s,a)$ (cf. (14)) in the expression of J_h^2 (cf. (4.42)), we obtain

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}J_h^2 = 2\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^\star}(s,a)\overline{B}_h^k(s,a)$$

$$
\begin{aligned}
&= 2\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} c_{\mathrm{b}}\sqrt{H\iota}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s,a)-\left(\mu_h^{\mathsf{adv},k}(s,a)\right)^2}{N_h^k(s,a)\vee 1}} \\
&\quad + 2\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1} c_{\mathrm{b}}\sqrt{\iota}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s,a)-\left(\mu_h^{\mathsf{ref},k}(s,a)\right)^2}{N_h^k(s,a)\vee 1}} \\
&\lesssim \underbrace{\sqrt{H\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s,a)-\left(\mu_h^{\mathsf{adv},k}(s,a)\right)^2}{N_h^k(s,a)\vee 1}}}_{=:\mathcal{J}_2^1} \\
&\quad + \underbrace{\sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s,a)-\left(\mu_h^{\mathsf{ref},k}(s,a)\right)^2}{N_h^k(s,a)\vee 1}}}_{=:\mathcal{J}_2^2}, \qquad \text{(B.103)}
\end{aligned}
$$

where the last inequality follows from (B.44). In the following, we shall look at the two terms in (B.103) separately.

Step 1: controlling $\mathcal{J}_2^1$. Recalling the expressions of $\sigma_h^{\mathsf{adv},k}(s,a)=\sigma_h^{\mathsf{adv},k^{N_h^k}+1}(s,a)$ in (B.54), we observe that the main part of $\mathcal{J}_2^1$ in (B.103) satisfies

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{adv},k}(s,a)-\left(\mu_h^{\mathsf{adv},k}(s,a)\right)^2}{N_h^k(s,a)\vee 1}} \le \sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{d_h^{\pi^\star}(s,a)\frac{d_h^{\pi^\star}(s,a)\cdot\sigma_h^{\mathsf{adv},k}(s,a)}{kd_h^{\mu}(s,a)}} \\
&= \sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{d_h^{\pi^\star}(s,a)\frac{d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2}{kd_h^{\mu}(s,a)}} \\
&\overset{\text{(i)}}{\le} \sqrt{C^\star\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{\frac{1}{k}\mathbb{1}\left(a=\pi_h^\star(s)\right)\cdot d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2} \\
&\overset{\text{(ii)}}{\le} \sqrt{C^\star\iota}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{1}\left(a=\pi_h^\star(s)\right)\frac{1}{k}} \\
&\lesssim \sqrt{HSC^\star\iota^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2}, \qquad \text{(B.104)}
\end{aligned}
$$

where the first inequality is due to the fact $\frac{1}{N_h^k(s,a)\vee 1}\le\frac{8\iota}{kd_h^{\mu}(s,a)}$ from Lemma 27, (i) follows from the definition of $C^\star$ in Assumption 1 and (4.26), and (ii) follows from the Cauchy-Schwarz inequality.

To continue, we claim the following bound holds, which will be proven in Appendix B.3.3.4:

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2$$
$$\lesssim H^2\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)+K+H^5\sqrt{S}C^\star\iota^2. \tag{B.105}$$

Combining the above inequality with (B.104), we arrive at

$$\mathcal{J}_2^1 \lesssim \sqrt{H^2SC^\star\iota^3}\sqrt{H^2\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)+K+H^5\sqrt{S}C^\star\iota^2}$$
$$\lesssim \sqrt{H^4SC^\star\iota^3\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)}+\sqrt{H^2SC^\star K\iota^3}+H^{3.5}SC^\star\iota^{2.5}. \tag{B.106}$$

Step 2: controlling $\mathcal{J}_2^2$. Recalling the expressions of $\mu_h^{\mathsf{ref},k+1}(s,a)=\mu_h^{\mathsf{ref},k^{N_h^k}+1}(s,a)$ and $\sigma_h^{\mathsf{ref},k+1}(s,a)=\sigma_h^{\mathsf{ref},k^{N_h^k}+1}(s,a)$ in (B.55) to $\mathcal{J}_2^2$ in (B.103), we can deduce that

$$\mathcal{J}_2^2=\sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\sigma_h^{\mathsf{ref},k}(s,a)-\left(\mu_h^{\mathsf{ref},k}(s,a)\right)^2}{N_h^k(s,a)\vee 1}}$$
$$\le\sqrt{\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}\underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\left(\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}}_{=:F_{h,k}}. \tag{B.107}$$

We further decompose and bound $F_{h,k}$ as follows:

$$F_{h,k}\overset{\text{(i)}}{\le}\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^\star(s_{h+1}^{k^n})\right)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}$$
$$=\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^\star(s_{h+1}^{k^n})\right)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2+\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}$$

$$\overset{\text{(ii)}}{\leq} \underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^\star(s_{h+1}^{k^n})\right)^2}{N_h^k(s,a)\vee 1}-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2}}_{G_{h,k}}+\underbrace{\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}2H\left(V_{h+1}^\star(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)}{N_h^k(s,a)\vee 1}}}_{=:L_{h,k}}, \tag{B.108}$$

where (i) follows from the fact that for some $k'\in[K]$, $\overline{V}_{h+1}^{\mathsf{next},k^n}=V_{h+1}^{k'}\leq V_{h+1}^\star$ (see the update rule of $\overline{V}^{\mathsf{next}}$ in line 27 and the fact in (4.38)), and (ii) holds due to the fact that

$$\left(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2-\left(\frac{\sum_{n=1}^{N_h^k(s,a)}\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\right)^2\leq 2H\frac{\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^\star(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)}{N_h^k(s,a)\vee 1}.$$

Inserting (B.108) back into (B.107), we arrive at

$$\begin{aligned}\mathcal{J}_2^2&\leq\sqrt{\iota}\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^\star}(s,a)\sum_{k=1}^K\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}\left(G_{h,k}+L_{h,k}\right)\\&\overset{\text{(i)}}{\lesssim}\sqrt{\iota}\left(\sqrt{H^3SC^\star K\iota^4}+H^4SC^\star\iota^3+\sqrt{H^3SC^\star K\iota^2}+H^{2.5}SC^\star\iota^3\right)\lesssim\sqrt{H^3SC^\star K\iota^5}+H^4SC^\star\iota^4,\end{aligned} \tag{B.109}$$

where (i) follows from the following facts

$$\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^\star}(s,a)\sum_{k=1}^K\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}L_{h,k}\lesssim\sqrt{H^3SC^\star K\iota^4}+H^4SC^\star\iota^3, \tag{B.110}$$

$$\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi^\star}(s,a)\sum_{k=1}^K\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}G_{h,k}\lesssim\sqrt{H^3SC^\star K\iota^2}+H^{2.5}SC^\star\iota^3. \tag{B.111}$$

We postpone the proofs of (B.110) and (B.111) to Appendix B.3.3.5 and Appendix B.3.3.6, respectively.

Putting the bounds together. Substitute (B.106) and (B.109) back into (B.103) to yield

$$\begin{aligned}\sum_{h=1}^H\left(1+\frac{1}{H}\right)^{h-1}J_h^2&\lesssim\sqrt{H^4SC^\star\iota^3\max_{h\in[H]}\sum_{k=1}^K\sum_{s\in\mathcal{S}}d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)}+\sqrt{H^2SC^\star K\iota^3}+H^{3.5}SC^\star\iota^{2.5}\\&\quad+\sqrt{H^3SC^\star K\iota^5}+H^4SC^\star\iota^4\\&\lesssim\sqrt{H^4SC^\star\iota^3\max_{h\in[H]}\sum_{k=1}^K\sum_{s\in\mathcal{S}}d_h^{\pi^\star}(s)\left(V_h^\star(s)-V_h^k(s)\right)}+\sqrt{H^3SC^\star K\iota^5}+H^4SC^\star\iota^4.\end{aligned}$$

B.3.3.3 Proof of inequality (4.45c)

Invoking inequality (B.44) directly leads to

$$\sum_{h=1}^{H}\left(1+\frac{1}{H}\right)^{h-1}\left(48\sqrt{HC^{\star}K\log\frac{2H}{\delta}}+28c_{\mathrm{a}}H^{3}C^{\star}\sqrt{S}\iota^{2}\right)\lesssim\sqrt{H^{3}C^{\star}K\log\frac{2H}{\delta}}+H^{4}C^{\star}\sqrt{S}\iota^{2}$$

as claimed.

B.3.3.4 Proof of inequality (B.105)

We shall control the term in (B.105) in a way similar to the proof of Lemma 8 in Appendix B.2.2.

Step 1: decomposing the terms of interest. Akin to Appendix B.2.2, let us introduce the terms of interest and definitions as follows:

$$\begin{aligned}
A_h &\coloneqq \sum_{k=1}^{K}\underbrace{\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^{\star}}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2}_{=:A_{h,k}},\\
B_{h,k} &\coloneqq \left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^{\star}}(s)\left(V_{h+1}^{k}(s)-\overline{V}_{h+1}^{k}(s)\right)^2,\\
Y_{h,k} &= \frac{d_h^{\pi^{\star}}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}\sum_{n=1}^{N_h^k(s_h^k,a_h^k)}\eta_n^{N_h^k(s_h^k,a_h^k)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2,\\
Z_{h,k} &= \left(1+\frac{1}{H}\right)\frac{d_h^{\pi^{\star}}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}P_h^k\left(V_{h+1}^{k}-\overline{V}_{h+1}^{k}\right)^2.
\end{aligned}\tag{B.112}$$

With these definitions in place, we directly adapt the argument in (B.35) to arrive at

$$A_h\le\sum_{k=1}^{K}B_{h,k}+\sum_{k=1}^{K}\left(Z_{h,k}-B_{h,k}\right)+\sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right).\tag{B.113}$$

As a consequence, it remains to control $\sum_{k=1}^{K}\left(Z_{h,k}-B_{h,k}\right)$ and $\sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right)$ separately.

Step 2: controlling $\sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right)$. To control $\sum_{k=1}^{K}\left(A_{h,k}-Y_{h,k}\right)$, we resort to Lemma 29 by setting

$$W_{h+1}^{k}(s,a)\coloneqq\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2,\qquad C_{\mathrm{d}}\coloneqq 1,\tag{B.114}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_\infty \le 4H^2 =: C_{\mathrm{w}}.$$

Applying Lemma 29 with (B.114) yields that: with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{k=1}^K (A_{h,k}-Y_{h,k})\right| = \left|\sum_{k=1}^K \overline{X}_{h,k}\right| \\
&\le \sqrt{\sum_{k=1}^K 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left[W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \\
&\lesssim \sqrt{C^\star \log\frac{2H}{\delta}\sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left[\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right]} + C^\star H^2\log\frac{2H}{\delta}.
\end{aligned} \tag{B.115}$$

To further control the first term in (B.115), it follows from Jensen's inequality that

$$P_{h,s,a}\left[\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right] \le P_{h,s,a}\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^4, \tag{B.116}$$

which yields

$$\begin{aligned}
&\sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\left[\sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2\right] \\
&\le \sum_{k=1}^K \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi_\star}(s,a)P_{h,s,a}\sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^4 \\
&\le \left(1+\frac{1}{H}\right)\sum_{k=1}^K\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^k(s)-\overline{V}_{h+1}^k(s)\right)^4 + 32\sqrt{H^8C^\star K\log\frac{2H}{\delta}} + 32H^4C^\star\log\frac{2H}{\delta}.
\end{aligned} \tag{B.117}$$

This can be verified similar to the proof for Lemma 8 in Appendix B.2.2. We omit the details for conciseness. To continue, it follows that

$$\sum_{k=1}^K\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^k(s)-\overline{V}_{h+1}^k(s)\right)^4$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \sum_{m=1}^{M}\sum_{t=1}^{L_m}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}(s)-\overline{V}_{h+1}^{(m,t)}(s)\right)^4\\
&\overset{\text{(ii)}}{=} \sum_{m=1}^{M}\sum_{t=1}^{L_m}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}(s)-V_{h+1}^{((m-1)\vee 1,1)}(s)\right)^4\\
&\overset{\text{(iii)}}{=} \sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{m=1}^{M} 2^m\left(V_{h+1}^{\star}(s)-V_{h+1}^{((m-1)\vee 1,1)}(s)\right)^4\\
&= 4\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{m-2=-1}^{M-2} 2^{m-2}\left(V_{h+1}^{\star}(s)-V_{h+1}^{((m-1)\vee 1,1)}(s)\right)^4\\
&= 4\sum_{m-2=-1}^{0} 2^{m-2}\left(V_{h+1}^{\star}(s)-V_{h+1}^{(1,1)}(s)\right)^4 + 4\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{m-2=1}^{M-2} 2^{m-2}\left(V_{h+1}^{\star}(s)-V_{h+1}^{(m-1,1)}(s)\right)^4.
\end{aligned}
$$

Here, (i) holds by using the pessimistic property $V^\star \geq V^k \geq \overline{V}^k$ for all $k\in[K]$ (see (4.38)) and by regrouping the summands; (ii) follows from the fact (see updating rules in line 26 and line 27) that for any $(m,s,h)\in[M]\times\mathcal{S}\times[H+1]$,

$$
\overline{V}_h^{(m,t)}(s) = V_h^{((m-1)\vee 1,1)}(s), \qquad t=1,2,\cdots,L_m; \tag{B.118}
$$

and (iii) results from the choice of the parameter $L_m = 2^m$. In addition, we can further control

$$
\begin{aligned}
\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{k}(s)-\overline{V}_{h+1}^{k}(s)\right)^4 &\overset{\text{(iv)}}{\leq} 8H^4 + 4\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{m=1}^{M-2}\sum_{t=1}^{L_m}\left(V_{h+1}^{\star}(s)-V_{h+1}^{(m+1,1)}(s)\right)^4\\
&\overset{\text{(v)}}{\leq} 8H^4 + 4\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{m=1}^{M-2}\sum_{t=1}^{L_m}\left(V_{h+1}^{\star}(s)-V_{h+1}^{(m,t)}(s)\right)^4\\
&\leq 8H^4 + 4\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{k=1}^{K}\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right)^4 \qquad \text{(B.119)}\\
&\leq 8H^4 + 4H^3\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\sum_{k=1}^{K}\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right)\\
&\overset{\text{(vi)}}{\lesssim} H^3K + H^8SC^\star\iota. \qquad \text{(B.120)}
\end{aligned}
$$

Here, (iv) follows from the fact $0\leq V_{h+1}^\star(s)-V_{h+1}^{(1,1)}(s)\leq H-0=H$; (v) holds since $V_{h+1}^\star \geq V_{h+1}^{(m+1,1)} = V_{h+1}^{(m,L_m)} \geq V_{h+1}^{(m,t)}$ for all $t\in[L_m]$ (using the monotonic increasing property of V_{h+1} introduced in (4.37)); and (vi) follows from (B.96). Putting (B.120) and (B.117) together with

(B.115), we arrive at

$$\begin{aligned}
&\left|\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})\right| \\
&\lesssim \sqrt{C^\star \log\frac{2H}{\delta}\left(H^3K+H^8SC^\star\iota+\sqrt{H^8C^\star K\log\frac{2H}{\delta}}+H^4C^\star\log\frac{2H}{\delta}\right)}+C^\star H^2\log\frac{2H}{\delta} \\
&\lesssim \sqrt{H^3C^\star K\iota}+H^4\sqrt{S}C^\star\iota^2. \qquad \text{(B.121)}
\end{aligned}$$

Step 3: controlling $\sum_{k=1}^{K}(Z_{h,k}-B_{h,k})$. Similarly, we also invoke Lemma 29. Let's set

$$W_{h+1}^k(s,a) \coloneqq \left(V_{h+1}^k-\overline{V}_{h+1}^k\right)^2, \qquad C_{\mathrm{d}} \coloneqq \left(1+\frac{1}{H}\right)\le 2, \qquad \text{(B.122)}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_\infty \le 4H^2 =: C_{\mathrm{w}}.$$

Applying Lemma 29 with (B.122) yields that: with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{k=1}^{K}(B_{h,k}-Z_{h,k})\right| = \left|\sum_{k=1}^{K}\overline{X}_{h,k}\right| \\
&\le \sqrt{\sum_{k=1}^{K}8C_{\mathrm{d}}^2C^\star\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi_\star}(s,a)P_{h,s,a}\left[W_{h+1}^k(s,a)\right]^2\log\frac{2H}{\delta}}+2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \\
&\lesssim \sqrt{C^\star\log\frac{2H}{\delta}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi_\star}(s,a)P_{h,s,a}\left[V_{h+1}^k-\overline{V}_{h+1}^k\right]^4}+C^\star H^2\log\frac{2H}{\delta} \\
&\overset{\text{(i)}}{\lesssim} \sqrt{C^\star\log\frac{2H}{\delta}\left(H^3K+H^8SC^\star\iota\right)}+C^\star H^2\log\frac{2H}{\delta} \lesssim \sqrt{H^3C^\star K\iota}+H^4\sqrt{S}C^\star\iota^2, \qquad \text{(B.123)}
\end{aligned}$$

where (i) follows from (B.119) and (B.120).

Step 4: combining the results. Inserting (B.123) and (B.121) back into (B.113), we can conclude that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}d_h^{\pi_\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^{k^n}-\overline{V}_{h+1}^{k^n}\right)^2=\sum_{h=1}^{H}A_h \\
&\le \sum_{h=1}^{H}\sum_{k=1}^{K}B_{h,k}+\sum_{h=1}^{H}\sum_{k=1}^{K}(Z_{h,k}-B_{h,k})+\sum_{h=1}^{H}\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{h=1}^{H}\sum_{k=1}^{K}\left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{k}(s)-\overline{V}_{h+1}^{k}(s)\right)^2+\sum_{h=1}^{H}\left|\sum_{k=1}^{K}(Z_{h,k}-B_{h,k})\right|+\sum_{h=1}^{H}\left|\sum_{k=1}^{K}(A_{h,k}-Y_{h,k})\right| \\
&\leq H\sum_{h=1}^{H}\sum_{k=1}^{K}\left(1+\frac{1}{H}\right)\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{k}(s)-\overline{V}_{h+1}^{k}(s)\right)+\sqrt{H^5C^\star K\iota}+H^5\sqrt{S}C^\star\iota^2 \\
&\overset{\text{(i)}}{\lesssim} H\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^{\star}(s)-V_{h+1}^{k}(s)\right)+K+H^5\sqrt{S}C^\star\iota^2 \\
&\lesssim H^2\max_{h\in[H]}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h}^{\pi^\star}(s)\left(V_{h}^{\star}(s)-V_{h}^{k}(s)\right)+K+H^5\sqrt{S}C^\star\iota^2, \qquad \text{(B.124)}
\end{aligned}
$$

where (i) follows from the same routine to obtain (B.119) and the Cauchy-Schwarz inequality.

B.3.3.5 Proof of inequality (B.110)

Step 1: decomposing the error in (B.110). The term in (B.110) obeys

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}L_{h,k} \\
&=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}\sqrt{\frac{\sum_{n=1}^{N_h^k(s,a)}2H\left(V_{h+1}^{\star}(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)}{N_h^k(s,a)\vee 1}} \\
&\overset{\text{(i)}}{\lesssim}\sqrt{H\iota}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{\frac{d_h^{\pi^\star}(s,a)}{kd_h^{\mu}(s,a)}}\sqrt{\frac{d_h^{\pi^\star}(s,a)\iota\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^{\star}(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)}{kd_h^{\mu}(s,a)}} \\
&\overset{\text{(ii)}}{\lesssim}\sqrt{HC^\star\iota^2}\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\sqrt{\frac{\mathbb{1}(a=\pi^\star(s))}{k}}\sqrt{\frac{d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^{\star}(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)}{kd_h^{\mu}(s,a)}} \\
&\overset{\text{(iii)}}{\lesssim}\sqrt{HC^\star\iota^2}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s,a)\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^{\star}(s_{h+1}^{k^n})-\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n})\right)}{kd_h^{\mu}(s,a)}}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^{K}\frac{\mathbb{1}(a=\pi^\star(s))}{k}} \\
&\lesssim\sqrt{H^2SC^\star\iota^3}\sqrt{\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{d_h^{\pi^\star}(s,a)}{d_h^{\mu}(s,a)}\sum_{k=1}^{K}\frac{1}{k}\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^{\star}(s_{h+1}^{k^n(s,a)})-\overline{V}_{h+1}^{\mathsf{next},k^n}(s_{h+1}^{k^n(s,a)})\right)} \\
&\overset{\text{(iv)}}{=}\sqrt{H^2SC^\star\iota^3}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^{\mu}(s_h^k,a_h^k)}P_h^k\sum_{k'=k}^{K}\frac{1}{k'}(V_{h+1}^{\star}-\overline{V}_{h+1}^{\mathsf{next},k})} \\
&\lesssim\sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s_h^k,a_h^k)P_h^k}{d_h^{\mu}(s_h^k,a_h^k)}(V_{h+1}^{\star}-\overline{V}_{h+1}^{\mathsf{next},k})}. \qquad \text{(B.125)}
\end{aligned}
$$

Here, (i) follows from the fact $\frac{1}{N_h^k(s,a)\vee 1} \le \frac{8\iota}{k d_h^\mu(s,a)}$ (cf. Lemma 27); (ii) follows from the definition of $C^\star$ in Assumption 1; (iii) invokes the Cauchy-Schwarz inequality; (iv) can be obtained by regrouping the terms (the terms involving $(V_{h+1}^\star - \overline{V}_{h+1}^{\mathsf{next},k})$ associated with index k will only been added during episodes $k' = k, k+1, \cdots, K$).

With this upper bound in hand, we further decompose

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}L_{h,k} &\lesssim \sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s_h^k,a_h^k)P_h^k}{d_h^\mu(s_h^k,a_h^k)}(V_{h+1}^\star - \overline{V}_{h+1}^{\mathsf{next},k})} \\
&\overset{\text{(i)}}{\lesssim} \sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k\left(V_{h+1}^\star - \overline{V}_{h+1}^{k}\right)} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{H^2SC^\star\iota^4}\sqrt{\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\left(V_{h+1}^\star - \overline{V}_{h+1}^{k}\right)} \\
&\quad + \sqrt{H^2SC^\star\iota^4}\sqrt{\left|\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a} - \frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k\right)\left(V_{h+1}^\star - \overline{V}_{h+1}^{k}\right)\right|}.
\end{aligned} \tag{B.126}$$

Here (i) holds due to the following observation: denoting by m the index of the epoch in which episode k occurs, we have

$$\overline{V}_{h+1}^{\mathsf{next},k} = V_{h+1}^{(m,1)} \ge V_{h+1}^{((m-1\vee 1),1)} = \overline{V}_{h+1}^{k}, \tag{B.127}$$

which invokes the monotonicity of V_{h+1}^k in (4.37). In addition, (ii) arises from the Cauchy-Schwarz inequality.

Step 2: controlling the first term in (B.126). The first term in (B.126) satisfies

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\left(V_{h+1}^\star - \overline{V}_{h+1}^{k}\right) &= \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\left\langle P_h(\cdot\,|\,s,a), V_{h+1}^\star - \overline{V}_{h+1}^{k}\right\rangle \\
&\overset{\text{(i)}}{=} \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s'\in\mathcal{S}} d_{h+1}^{\pi^\star}(s')\left(V_{h+1}^\star(s') - \overline{V}_{h+1}^{k}(s')\right) \\
&\overset{\text{(ii)}}{\lesssim} H^2 + \sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}} d_{h+1}^{\pi^\star}(s)\left(V_{h+1}^\star(s) - V_{h+1}^{k}(s)\right) \\
&\overset{\text{(iii)}}{\lesssim} HK + H^6SC^\star\iota,
\end{aligned} \tag{B.128}$$

where (i) holds due to the fact $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_h(\cdot\,|\,s,a) = d_{h+1}^{\pi^\star}(\cdot)$, (ii) comes from the same argument employed to establish (B.119), and (iii) follows from (B.96).

Step 3: controlling the second term in (B.126). We shall invoke Lemma 29 for this purpose. To proceed, let

$$W_{h+1}^k(s,a) \coloneqq V_{h+1}^\star - \overline{V}_{h+1}^k, \qquad C_{\mathrm{d}} =: 1, \tag{B.129}$$

which satisfies

$$\left\|W_{h+1}^k(s,a)\right\|_\infty \le H =: C_{\mathrm{w}}.$$

Applying Lemma 29 with (B.129) yields, for all $h \in [H]$, with probability at least $1-\delta$

$$\begin{aligned}
&\left|\sum_{k=1}^K \left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a} - \frac{d_h^{\pi^\star}(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)}P_h^k\right)\left(V_{h+1}^\star - \overline{V}_{h+1}^k\right)\right| = \left|\sum_{k=1}^K \overline{X}_{h,k}\right| \\
&\le \sqrt{\sum_{k=1}^K 8C_{\mathrm{d}}^2 C^\star \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\left[W_{h+1}^k(s,a)\right]^2 \log\frac{2H}{\delta}} + 2C_{\mathrm{d}}C^\star C_{\mathrm{w}}\log\frac{2H}{\delta} \\
&\lesssim \sqrt{C^\star\log\frac{2H}{\delta}\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\left(V_{h+1}^\star - \overline{V}_{h+1}^k\right)^2} + HC^\star\log\frac{2H}{\delta} \\
&\overset{\text{(i)}}{\lesssim} \sqrt{C^\star\log\frac{2H}{\delta}\left(H^2 + \sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a}\left(V_{h+1}^\star - V_{h+1}^k\right)^2\right)} + HC^\star\log\frac{2H}{\delta} \\
&\overset{\text{(ii)}}{\lesssim} \sqrt{C^\star\log\frac{2H}{\delta}\left(HK + H^6SC^\star\iota\right)} + HC^\star\log\frac{2H}{\delta} \\
&\lesssim \sqrt{HC^\star K\iota} + H^3\sqrt{S}C^\star\iota.
\end{aligned} \tag{B.130}$$

Here (i) follows from the same routine to arrive at (B.119), and (ii) comes from (B.96). As a result, the second term in (B.126) satisfies, with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\sum_{h=1}^H\sum_{k=1}^K \left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a} - \frac{d_h^{\pi^\star}(s_h^k,a_h^k)P_h^k}{d_h^\mu(s_h^k,a_h^k)}\right)\left(V_{h+1}^\star - \overline{V}_{h+1}^k\right)\right| \\
&\le \sum_{h=1}^H\left|\sum_{k=1}^K \left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)P_{h,s,a} - \frac{d_h^{\pi^\star}(s_h^k,a_h^k)P_h^k}{d_h^\mu(s_h^k,a_h^k)}\right)\left(V_{h+1}^\star - \overline{V}_{h+1}^k\right)\right| \lesssim \sqrt{H^3C^\star K\iota} + H^4\sqrt{S}C^\star\iota.
\end{aligned} \tag{B.131}$$

Step 4: combining the results. Finally, inserting (B.128) and (B.131) into (B.126), we arrive at

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)}}L_{h,k}\\
&\lesssim \sqrt{H^2SC^\star\iota^4}\sqrt{HK+H^6SC^\star\iota}+\sqrt{H^2SC^\star\iota^4}\sqrt{\sqrt{H^3C^\star K\iota}+H^4\sqrt{S}C^\star\iota}\\
&\lesssim \sqrt{H^3SC^\star K\iota^4}+H^4SC^\star\iota^3+\sqrt{H^2SC^\star\iota^4}\sqrt{HK+H^4\sqrt{S}C^\star\iota}\lesssim \sqrt{H^3SC^\star K\iota^4}+H^4SC^\star\iota^3,
\end{aligned} \tag{B.132}$$

where the last two inequalities follow from the Cauchy-Schwarz inequality.

B.3.3.6 Proof of inequality (B.111)

Recall the expression of $G_{h,k}$ in (B.108) as

$$\begin{aligned}
G_{h,k}^2 &= \frac{\sum_{n=1}^{N_h^k(s,a)}\left(V_{h+1}^\star(s_{h+1}^{k^n})\right)^2}{N_h^k(s,a)\vee 1}-\Big(\frac{\sum_{n=1}^{N_h^k(s,a)}V_{h+1}^\star(s_{h+1}^{k^n})}{N_h^k(s,a)\vee 1}\Big)^2\\
&= \frac{\sum_{n=1}^{N_h^k(s,a)}P_h^{k^n}\left(V_{h+1}^\star\right)^2}{N_h^k(s,a)\vee 1}-\Big(\frac{\sum_{n=1}^{N_h^k(s,a)}P_h^{k^n}V_{h+1}^\star}{N_h^k(s,a)\vee 1}\Big)^2.
\end{aligned} \tag{B.133}$$

To continue, we make the following observation

$$\begin{aligned}
G_{h,k} &\le \left\{\left|G_{h,k}^2-\mathsf{Var}_{h,s,a}(V_{h+1}^\star)\right|+\mathsf{Var}_{h,s,a}(V_{h+1}^\star)\right\}^{1/2}\\
&\le \left|G_{h,k}^2-\mathsf{Var}_{h,s,a}(V_{h+1}^\star)\right|^{1/2}+\sqrt{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}
\end{aligned} \tag{B.134}$$

due to the elementary inequality $\sqrt{a^2+b^2}\le a+b$ for any $a,b\ge 0$. Here, we remind the reader that $\mathsf{Var}_{h,s,a}(V_{h+1}^\star)=P_{h,s,a}(V_{h+1}^\star)^2-(P_{h,s,a}V_{h+1}^\star)^2$ (cf. (B.51)). This allows us to rewrite

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}G_{h,k}\\
&\le \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\left|G_{h,k}^2-\mathsf{Var}_{h,s,a}(V_{h+1}^\star)\right|}{N_h^k(s,a)\vee 1}}+\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{N_h^k(s,a)\vee 1}},
\end{aligned} \tag{B.135}$$

leaving us with two terms to cope with.

Step 1: controlling the first term of (B.135). By definition, we have

$$\begin{aligned}
\left|G_{h,k}^2 - \mathsf{Var}_{h,s,a}(V_{h+1}^\star)\right| &= \left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n}\left(V_{h+1}^\star\right)^2}{N_h^k(s,a)\vee 1} - \Big(\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1}\Big)^2 - P_{h,s,a}(V_{h+1}^\star)^2 + \big(P_{h,s,a}V_{h+1}^\star\big)^2\right| \\
&\le \left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n}\left(V_{h+1}^\star\right)^2}{N_h^k(s,a)\vee 1} - P_{h,s,a}(V_{h+1}^\star)^2\right| + \left|\Big(\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1}\Big)^2 - \big(P_{h,s,a}V_{h+1}^\star\big)^2\right| \\
&\le \left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n}\left(V_{h+1}^\star\right)^2}{N_h^k(s,a)\vee 1} - P_{h,s,a}(V_{h+1}^\star)^2\right| + 2H\left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} - P_{h,s,a}V_{h+1}^\star\right|, \qquad \text{(B.136)}
\end{aligned}$$

where the last inequality holds due to

$$\begin{aligned}
\left|\Big(\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1}\Big)^2 - \big(P_{h,s,a}V_{h+1}^\star\big)^2\right| &= \left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} - P_{h,s,a}V_{h+1}^\star\right| \cdot \left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} + P_{h,s,a}V_{h+1}^\star\right| \\
&\le 2H\left|\frac{\sum_{n=1}^{N_h^k(s,a)} P_h^{k^n} V_{h+1}^\star}{N_h^k(s,a)\vee 1} - P_{h,s,a}V_{h+1}^\star\right|.
\end{aligned}$$

We now control the two terms in (B.136) separately by invoking Lemma 24. For the first term in (B.136), let us set

$$W_{h+1}^i \coloneqq \left(V_{h+1}^\star\right)^2, \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \frac{1}{N\vee 1} \coloneqq C_{\mathrm{u}}, \qquad \text{(B.137)}$$

which indicates that

$$\|W_{h+1}^i\|_\infty \le H^2 =: C_{\mathrm{w}}, \qquad \text{(B.138)}$$

Applying Lemma 24 with (B.137) and $N = N_h^k = N_h^k(s,a)$, with probability at least $1-\frac{\delta}{2}$, we arrive at

$$\begin{aligned}
&\left|\frac{1}{N_h^k(s,a)\vee 1}\sum_{n=1}^{N_h^k}(P_h^{k^n} - P_{h,s,a})(V_{h+1}^\star)^2\right| = \left|\sum_{i=1}^{k} X_i\left(s,a,h,N_h^k\right)\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s,a,N_h^k)\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k^n}\big)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N_h^k\vee 1}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\asymp \sqrt{\frac{\iota^2}{N_h^k\vee 1}}\sqrt{\sum_{n=1}^{N_h^k}\frac{1}{N_h^k\vee 1}\|W_{h+1}^{k^n}\|_\infty^2} + \frac{H^2\iota^2}{N_h^k\vee 1} \lesssim H^2\iota^2\sqrt{\frac{1}{N_h^k\vee 1}}. \qquad \text{(B.139)}
\end{aligned}$$

Similarly, for the second term in (B.136), with $W_{h+1}^i \coloneqq V_{h+1}^\star$, we have with probability at least $1-\frac{\delta}{2}$,

$$\frac{1}{N_h^k(s,a)\vee 1}\sum_{n=1}^{N_h^k}\left(P_h^{k^n}-P_{h,s,a}\right)V_{h+1}^\star \lesssim H\iota^2\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}. \tag{B.140}$$

Inserting (B.139) and (B.140) back into (B.136) yields

$$\left|G_{h,k}^2-\mathsf{Var}_{h,s,a}(V_{h+1}^\star)\right| \lesssim H^2\iota^2\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}. \tag{B.141}$$

Consequently, the first term in (B.135) can be controlled as

$$\begin{aligned}\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^K\sqrt{\frac{\left|G_{h,k}^2-\mathsf{Var}_{h,s,a}(V_{h+1}^\star)\right|}{N_h^k(s,a)\vee 1}} &\lesssim H\iota\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^K\frac{1}{\left(N_h^k(s,a)\right)^{\frac{3}{4}}\vee 1}\\ &\lesssim H^2(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2,\end{aligned} \tag{B.142}$$

where the last inequality holds due to (B.100).

Step 2: controlling the second term of (B.135). The second term can be decomposed as

$$\begin{aligned}&\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^K\sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{N_h^k(s,a)\vee 1}}\\ &\overset{\text{(i)}}{\lesssim}\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^K\sqrt{\frac{C^\star\iota d_h^{\pi^\star}(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{k}\mathbb{1}\left(a=\pi_h^\star(s)\right)}\\ &\overset{\text{(ii)}}{\lesssim}\sqrt{C^\star\iota}\sqrt{\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^K\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}\sqrt{\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{k=1}^K\frac{1}{k}\mathbb{1}\left(a=\pi_h^\star(s)\right)}\\ &\lesssim\sqrt{HSC^\star K\iota^2}\sqrt{\sum_{h=1}^H\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star)},\end{aligned} \tag{B.143}$$

where (i) follows from the facts $\frac{1}{N_h^k(s,a)\vee 1}\le\frac{8\iota}{kd_h^\mu(s,a)}$ by Lemma 27 and the definition of $C^\star$ in Assumption 1, (ii) holds by the Cauchy-Schwarz inequality, and the final inequality comes from the fact that $\pi^\star$ is deterministic.

We are then left with bounding $\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star)$. Note that

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star) = \mathbb{E}_{s_1\sim\rho, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\sum_{h=1}^{H}\mathsf{Var}_{h,s_h,\pi_h^\star(s_h)}(V_{h+1}^\star)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{s_1\sim\rho, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\sum_{h=1}^{H}\left(r_h\left(s_h,\pi_h^\star(s_h)\right)+V_{h+1}^\star(s_{h+1})-V_h^\star(s_h)\right)^2\right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_{s_1\sim\rho, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\sum_{h=1}^{H}\left(r_h(s_h,\pi_h^\star(s_h))+V_{h+1}^\star(s_{h+1})-V_h^\star(s_h)\right)\right]^2 \\
&\overset{\text{(iii)}}{=} \mathbb{E}_{s_1\sim\rho, s_{h+1}\sim P_{h,s_h,\pi_h^\star(s_h)}}\left[\left(\sum_{h=1}^{H} r_h(s_h,\pi_h^\star(s_h))\right)-V_1^\star(s_1)\right]^2 \overset{\text{(iv)}}{\le} H^2,
\end{aligned} \tag{B.144}$$

where (i) follows from Bellman's optimality equation, (ii) follows from the Markov property, (iii) holds due to the fact that $V_{H+1}^\star(s)=0$ for all $s\in\mathcal{S}$, and (iv) arises from the fact $r_h(s,a)\le 1$ for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$. Substituting (B.144) back into (B.143), we get

$$\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{N_h^k(s,a)\vee 1}} \lesssim \sqrt{H^3SC^\star K\iota^2}. \tag{B.145}$$

Step 4: combing the results. Combining (B.142) and (B.145) with (B.135) yields

$$\begin{aligned}
\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^\star}(s,a)\sum_{k=1}^{K}\sqrt{\frac{1}{N_h^k(s,a)\vee 1}}G_{h,k} &\lesssim H^2(SC^\star)^{\frac{3}{4}}K^{\frac{1}{4}}\iota^2+\sqrt{H^3SC^\star K\iota^2} \\
&\lesssim \sqrt{H^3SC^\star K\iota^2}+H^{2.5}SC^\star\iota^3.
\end{aligned} \tag{B.146}$$

B.3.4 Proof of Lemma 30

In view of (B.69), we can decompose the term of interest into

$$\left|\sum_{n=1}^{N_h^k(s,a)}\eta_n^{N_h^k(s,a)}\xi_h^{k^n}\right| \le |U_1|+|U_2|,$$

where

$$U_1 := \sum_{n=1}^{N_h^k}\eta_n^{N_h^k}\left(P_h^{k^n}-P_{h,s,a}\right)\left(\overline{V}_{h+1}^{k^n}-V_{h+1}^{k^n}\right), \tag{B.147a}$$

$$U_2 \coloneqq \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_{h,s,a} - \frac{\sum_{i=N_h^{(m^n-1,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \right) \overline{V}_{h+1}^{k^n}. \tag{B.147b}$$

Next, we turn to controlling these two terms separately with the assistance of Lemma 24.

Step 1: controlling U_1. In the following, we invoke Lemma 24 to control U_1 in (B.147a). Let us set

$$W_{h+1}^i \coloneqq \overline{V}_{h+1}^i - V_{h+1}^i, \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \eta_{N_h^i(s,a)}^N \geq 0,$$

which indicates that

$$\|W_{h+1}^i\|_\infty \leq \|\overline{V}_{h+1}^i\|_\infty + \|V_{h+1}^i\|_\infty \leq 2H =: C_{\mathrm{w}},$$

and

$$\max_{N,h,s,a \in \left(\{0\}\cup[K]\right)\times[H]\times\mathcal{S}\times\mathcal{A}} \eta_{N_h^i(s,a)}^N \leq \frac{2H}{N \vee 1} =: C_{\mathrm{u}}. \tag{B.148}$$

Here, the last inequality follows since (according to Lemma 1 and the definition in (4.16))

$$\begin{aligned} \eta_{N_h^i(s,a)}^N &\leq \frac{2H}{N \vee 1}, \qquad \text{if } 0 \leq N_h^i(s,a) \leq N; \\ \eta_{N_h^i(s,a)}^N &= 0, \qquad\quad \text{if } N_h^i(s,a) > N. \end{aligned}$$

To continue, it can be seen from (4.17) that

$$0 \leq \sum_{n=1}^{N} u_h^{k_h^n(s,a)}(s,a,N) = \sum_{n=1}^{N} \eta_n^N \leq 1 \tag{B.149}$$

holds for all $(N,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}$. Therefore, choosing $N = N_h^k(s,a) = N_h^k$ for any (s,a) and applying Lemma 24 with the above quantities, we arrive at

$$\begin{aligned} |U_1| &= \left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_h^{k^n} - P_{h,s,a}\right)\left(\overline{V}_{h+1}^{k^n} - V_{h+1}^{k^n}\right) \right| = \left| \sum_{i=1}^{k} X_i \left(s,a,h,N_h^k\right) \right| \\ &\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s,a,N_h^k) \mathsf{Var}_{h,s,a}\left(W_{h+1}^{k^n}\right)} + \left(C_{\mathrm{u}} C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N \vee 1}} C_{\mathrm{w}} \right) \log^2 \frac{SAT}{\delta} \end{aligned}$$

$$\asymp \sqrt{\frac{H\iota^2}{N_h^k \vee 1}} \sqrt{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathsf{Var}_{h,s,a}\big(\overline{V}_{h+1}^{k^n} - V_{h+1}^{k^n}\big) + \frac{H^2\iota^2}{N_h^k \vee 1}} \tag{B.150}$$

$$\lesssim \sqrt{\frac{H\iota^2}{N_h^k \vee 1}} \sqrt{\sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s,a) - \big(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s,a)\big)^2} + \frac{H^{7/4}\iota^2}{(N_h^k \vee 1)^{3/4}} + \frac{H^2\iota^2}{N_h^k \vee 1}. \tag{B.151}$$

with probability at least $1-\delta$. Here, the proof of the inequality (B.151) is postponed to Appendix B.3.4.1 in order to streamline the presentation of the analysis.

Step 2: bounding U_2. Making use of the result in (B.53), we arrive at

$$\frac{\sum_{i=N_h^{(m^n,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \overline{V}_{h+1}^{k^n} = \frac{\sum_{i=N_h^{(m^n,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i} \overline{V}_{h+1}^{\mathsf{next},k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1}.$$

To continue, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, we rewrite and rearrange U_2 (cf. (B.147b)) as follows:

$$\begin{aligned}
U_2 &= \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_{h,s,a} - \frac{\sum_{i=N_h^{(m^n,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \right) \overline{V}_{h+1}^{k^n} \\
&= \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_{h,s,a} \overline{V}_{h+1}^{k^n} - \frac{\sum_{i=N_h^{(m^n,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i} \overline{V}_{h+1}^{\mathsf{next},k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \right) \\
&\overset{\text{(i)}}{=} \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(\frac{\sum_{i=N_h^{(m^n,1)}+1}^{N_h^{(m^n,1)}} P_{h,s,a}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \overline{V}_{h+1}^{k^n} - \frac{\sum_{i=N_h^{(m^n,1)}+1}^{N_h^{(m^n,1)}} P_h^{k^i} \overline{V}_{h+1}^{\mathsf{next},k^i}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \right) \\
&= \sum_{n=1}^{N_h^k} \frac{\eta_n^{N_h^k}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \sum_{i=N_h^{(m^n,1)}+1}^{N_h^{(m^n,1)}} \left(P_{h,s,a} - P_h^{k^i} \right) \overline{V}_{h+1}^{\mathsf{next},k^i} \\
&\overset{\text{(ii)}}{=} \sum_{i=1}^{N_h^k} \left(\sum_{n=N_h^{(m^i+1,1)}+1}^{N_h^{(m^i+2,1)} \wedge N_h^k} \frac{\eta_n^{N_h^k}}{\widehat{N}_h^{\mathsf{epo},m^n-1}(s,a) \vee 1} \right) \left(P_{h,s,a} - P_h^{k^i} \right) \overline{V}_{h+1}^{\mathsf{next},k^i} \\
&= \sum_{i=1}^{N_h^k} \left(\sum_{n=N_h^{(m^i+1,1)}+1}^{N_h^{(m^i+2,1)} \wedge N_h^k} \frac{\eta_n^{N_h^k}}{\widehat{N}_h^{\mathsf{epo},m^i} \vee 1} \right) \left(P_{h,s,a} - P_h^{k^i} \right) \overline{V}_{h+1}^{\mathsf{next},k^i},
\end{aligned}$$

where (i) follows from the fact that $N_h^{(m^n,1)} - N_h^{(m^n-1,1)} = \widehat{N}_h^{\mathsf{epo},m^n-1}(s,a)$, and (ii) is obtained by rearranging terms with respect to i (the terms with respect to $\overline{V}_{h+1}^{\mathsf{next},k^i}$ will only be added during the epoch m^i+1), and the last equality holds since $m^n - 1 = m^i$ for all $n = N_h^{(m^i+1,1)}+1, N_h^{(m^i+1,1)}+2, N_h^{(m^i+2,1)}$.

With the above relation in mind, we are ready to invoke Lemma 24 to control U_2. To continue, for any episode $j \le k$, let us denote by $m(j)$ the index of the epoch in which episode j happens (with slight abuse of notation). Let us set

$$W_{h+1}^j \coloneqq \overline{V}_{h+1}^{\mathsf{next},j}, \qquad \text{and} \qquad u_h^j(s,a,N) \coloneqq \sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{\eta_n^N}{\widehat{N}_h^{\mathsf{epo},m(j)}(s,a)\vee 1}.$$

As a result, we see that

$$\|W_{h+1}^j\|_\infty \le \|\overline{V}_{h+1}^{\mathsf{next},j}\|_\infty \le H =: C_{\mathrm{w}}$$

and the following fact (which will be established in Appendix B.3.4.2)

$$0 \le u_h^j(s,a,N) = \sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{\eta_n^N}{\widehat{N}_h^{\mathsf{epo},m(j)}(s,a)\vee 1} \le \frac{64e^2\iota}{N\vee 1} =: C_u \tag{B.152}$$

holds for all $(j,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}$ with probability at least $1-\delta$.

Given that $N = N_h^k(s,a) = N_h^k$, applying Lemma 24 with the above quantities, we can show that for any state-action pair $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
|U_2| &= \left|\sum_{i=1}^{N_h^k}\left(\sum_{n=N_h^{(m^i+1,1)}+1}^{N_h^{(m^i+2,1)}\wedge N_h^k}\frac{\eta_n^{N_h^k}}{\widehat{N}_h^{\mathsf{epo},m^i}\vee 1}\right)\left(P_{h,s,a}-P_h^{k^i}\right)\overline{V}_{h+1}^{\mathsf{next},k^i}\right| = \left|\sum_{j=1}^{k} X_j\left(s,a,h,N_h^k\right)\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{i=1}^{N_h^k(s,a)} u_h^{k_h^i(s,a)}(s,a,N)\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k_h^i(s,a)}\big)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N\vee 1}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{\iota^3}{N_h^k\vee 1}}\sqrt{\frac{1}{N_h^k\vee 1}\sum_{i=1}^{N_h^k}\mathsf{Var}_{h,s,a}\big(\overline{V}_{h+1}^{\mathsf{next},k^i}\big)} + \frac{H\iota^3}{N_h^k\vee 1} \\
&\lesssim \sqrt{\frac{\iota^3}{N_h^k\vee 1}}\sqrt{\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s,a) - \big(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s,a)\big)^2} + \frac{H\iota^3}{(N_h^k\vee 1)^{3/4}}.
\end{aligned} \tag{B.153}$$

To streamline the presentation of the analysis, we shall postpone the proof of (B.153) to Appendix B.3.4.3.

Step 3: summing up. Combining the bounds in (B.151) and (B.153) yields that: for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
\left| \sum_{n=1}^{N_h^k(s,a)} \eta_n^{N_h^k(s,a)} \xi_h^{k^n} \right| &\leq |U_1| + |U_2| \\
&\lesssim \sqrt{\frac{H\iota^2}{N_h^k \vee 1}} \sqrt{\sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s,a) - \left(\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s,a)\right)^2} \\
&\quad + \sqrt{\frac{\iota^3}{N_h^k \vee 1}} \sqrt{\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s,a) - \left(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s,a)\right)^2} + c_{\mathrm{b}} \frac{H^{7/4}\iota^2}{(N_h^k \vee 1)^{3/4}} + c_{\mathrm{b}} \frac{H^2\iota^2}{N_h^k \vee 1} \\
&\leq \overline{B}_h^{k^{N_h^k+1}}(s,a) + c_{\mathrm{b}} \frac{H^{7/4}\iota^2}{(N_h^k \vee 1)^{3/4}} + c_{\mathrm{b}} \frac{H^2\iota^2}{N_h^k \vee 1} \qquad \text{(B.154)}
\end{aligned}$$

holds for some sufficiently large constant $c_{\mathrm{b}} > 0$, where the last line follows from the definition of $\overline{B}_h^{k^{N_h^k+1}}(s,a)$ in line 14 of Algorithm 6. As a consequence of the inequality (B.154), for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, one has

$$\left| \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \xi_h^{k^n} \right| \leq \overline{B}_h^{k^{N_h^k+1}}(s,a) + c_{\mathrm{b}} \frac{H^{7/4}\iota^2}{(N_h^k \vee 1)^{3/4}} + c_{\mathrm{b}} \frac{H^2\iota^2}{N_h^k \vee 1} \leq \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \overline{b}_h^{k^n+1},$$

where the last inequality holds due to (B.62). We have thus concluded the proof of Lemma 30.

B.3.4.1 Proof of inequality (B.151)

To establish the inequality (B.151), it is sufficient to consider the difference

$$W_1 := \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathsf{Var}_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}) - \sigma_h^{\mathsf{adv},k^{N_h^k+1}}(s,a) + (\mu_h^{\mathsf{adv},k^{N_h^k+1}}(s,a))^2.$$

Before continuing, it is easily verified that if $N_h^k = N_h^k(s,a) = 0$, the basic fact $\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 0$ leads to $W_1 = 0$, and therefore, (B.151) holds directly. The remainder of the proof is thus dedicated to controlling W_1 when $N_h^k = N_h^k(s,a) \geq 1$. Recalling the definition in (B.51)

$$\mathsf{Var}_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}) = P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})^2 - \left(P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\right)^2, \qquad \text{(B.155)}$$

we can take this result together with (B.54) to yield

$$W_1 = \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})^2 - \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \left(P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\right)^2$$

$$
\begin{aligned}
&\quad + \Big(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big)^2 - \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})^2 \\
&\le \underbrace{\left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}(P_h^{k^n} - P_{h,s,a})(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})^2\right|}_{=:W_1^1} \\
&\qquad + \underbrace{\Big(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big)^2 - \sum_{n=1}^{N_h^k} \eta_n^{N_h^k}\Big(P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big)^2}_{=:W_1^2}. \qquad \text{(B.156)}
\end{aligned}
$$

It then boils down to control the above two terms in (B.156) separately when $N_h^k = N_h^k(s,a) \ge 1$.

Step 1: controlling W_1^1. To control W_1^1, we shall invoke Lemma 24 by setting

$$
W_{h+1}^i \coloneqq (V_{h+1}^i - \overline{V}_{h+1}^i)^2, \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \eta_{N_h^i(s,a)}^N \ge 0,
$$

which obey

$$
\|W_{h+1}^i\|_\infty \le \|\overline{V}_{h+1}^i\|_\infty^2 + \|V_{h+1}^i\|_\infty^2 \le 2H^2 =: C_{\mathrm{w}}.
$$

Invoking the facts in (B.148) and (B.149), we arrive at

$$
\frac{2H}{N \vee 1} =: C_{\mathrm{u}}
$$

and

$$
0 \le \sum_{n=1}^{N} u_h^{k_h^n(s,a)}(s,a,N) \le 1, \qquad \forall (N,s,a) \in [K] \times \mathcal{S} \times \mathcal{A}.
$$

Therefore, choosing $N = N_h^k(s,a) = N_h^k$ for any (s,a) and applying Lemma 24 with the above quantities, we arrive at, with probability at least $1-\delta$,

$$
\begin{aligned}
|W_1^1| &= \left|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k}(P_h^{k^n} - P_{h,s,a})(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})^2\right| = \left|\sum_{i=1}^{k} X_i\left(s,a,h,N_h^k\right)\right| \\
&\lesssim \sqrt{C_{\mathrm{u}} \log^2 \frac{SAT}{\delta}} \sqrt{\sum_{n=1}^{N_h^k} u_h^{k^n}(s,a,N_h^k) \mathsf{Var}_{h,s,a}\big(W_{h+1}^{k^n}\big)} + \left(C_{\mathrm{u}} C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N \vee 1}} C_{\mathrm{w}}\right) \log^2 \frac{SAT}{\delta}
\end{aligned}
$$

$$\lesssim \sqrt{\frac{H}{N_h^k \vee 1}\iota^2}\sqrt{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \|W_{h+1}^{k^n}\|_\infty^2} + \frac{H^3\iota^2}{N_h^k \vee 1} \lesssim \sqrt{\frac{H^5}{N_h^k \vee 1}\iota^2} + \frac{H^3\iota^2}{N_h^k \vee 1}. \tag{B.157}$$

Step 2: controlling W_1^2. Observe that Jensen's inequality gives

$$\Big(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big)^2 \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big)^2, \tag{B.158}$$

due to the fact $\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} = 1$ (see (4.17) and (4.16)). Plugging the above relation into (B.156) gives

$$\begin{aligned} W_1^2 &\le \Big(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_h^{k^n}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big)^2 - \Big(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} P_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big)^2 \\ &= \Big(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} (P_h^{k^n} - P_{h,s,a})(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big) \cdot \Big(\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} (P_h^{k^n} + P_{h,s,a})(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big). \end{aligned} \tag{B.159}$$

Note that the first term in (B.159) is exactly $|U_1|$ defined in (B.147a), which can be controlled by invoking (B.150) to achieve that, with probability at least $1-\delta$,

$$\begin{aligned} &\Big|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} (P_h^{k^n} - P_{h,s,a})(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})\Big| \\ &\lesssim \sqrt{\frac{H}{N_h^k \vee 1}\iota^2}\sqrt{\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \mathsf{Var}_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n})} + \frac{H^2\iota^2}{N_h^k \vee 1} \lesssim \sqrt{\frac{H^3\iota^2}{N_h^k \vee 1}} + \frac{H^2}{N_h^k \vee 1}\iota^2, \end{aligned} \tag{B.160}$$

where the final inequality holds since $\mathsf{Var}_{h,s,a}(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}) \lesssim H^2$ and the fact in (4.17). In addition, the second term in (B.159) can be controlled straightforwardly by

$$\Big|\sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(P_h^{k^n} + P_{h,s,a}\Big)\Big(V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\Big)\Big| \le \sum_{n=1}^{N_h^k} \eta_n^{N_h^k} \Big(\big\|P_h^{k^n}\big\|_1 + \big\|P_{h,s,a}\big\|_1\Big)\big\|V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\big\|_\infty \le 2H,$$

where we have used the fact in (4.17), $\big\|V_{h+1}^{k^n} - \overline{V}_{h+1}^{k^n}\big\|_\infty \le H$ and $\big\|P_h^{k^n}\big\|_1 = \big\|P_{h,s,a}\big\|_1 = 1$.

Taking the above two facts collectively with (B.159) yields

$$W_1^2 \lesssim \sqrt{\frac{H^5\iota^2}{N_h^k \vee 1}} + \frac{H^3\iota^2}{N_h^k \vee 1}. \tag{B.161}$$

Step 3: summing up. Plugging the results in (B.157) and (B.161) back into (B.156), we have

$$W_1 \le W_1^1 + W_1^2 \lesssim \sqrt{\frac{H^5\iota^2}{N_h^k \vee 1}} + \frac{H^3\iota^2}{N_h^k \vee 1},$$

which leads to the desired result (B.151) directly.

B.3.4.2 Proof of inequality (B.152)

To begin with, let us recall two pieces of notation that shall be used throughout this proof:

1. $m(j)$: the index of the epoch in which the j-th episode occurs.
2. $\widehat{N}_h^{\mathsf{epo},m}(s,a)$: the value of $\widehat{N}_h^{(m,L_m+1)}(s,a)$, representing the number of visits to (s,a) in the entire m-th epoch with length $L_m = 2^m$.

Applying (B.1) and taking the union bound over $(m(j),h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$ yield

$$\widehat{N}_h^{\mathsf{epo},m(j)}(s,a) \vee 1 \ge \frac{2^{m(j)} d_h^\mu(s,a)}{8\log\left(\frac{SAT}{\delta}\right)} \tag{B.162}$$

with probability at least $1-\delta/2$.

For any epoch m, if we denote by $k_{\text{last}}(m)$ the index of the last episode in the m-th epoch, we can immediately see that

$$k_{\text{last}}(m) = \sum_{i=1}^{m} L_i = \sum_{i=1}^{m} 2^i = 2^{m+1} - 2 \le 2^{m+1}. \tag{B.163}$$

Applying (B.1) again and taking the union bound over $(m(j),h,s,a) \in [M]\times[H]\times\mathcal{S}\times\mathcal{A}$, one can guarantee that for every $n \in [N_h^{(m(j)+1,1)}, N_h^{(m(j)+2,1)}]$, with probability at least $1-\delta/2$,

$$\begin{aligned} N_h^{(m(j)+1,1)} \le n \le N_h^{(m(j)+2,1)} &= N_h^{k_{\text{last}}(m(j)+1)} \\ &\le N_h^{2^{m(j)+2}} \le \begin{cases} e^2 2^{m(j)+2} d_h^\mu(s,a) & \text{if } 2^{m(j)+2} d_h^\mu(s,a) \ge \log\left(\frac{SAT}{\delta}\right) \\ 2e^2 \log\left(\frac{SAT}{\delta}\right) & \text{if } 2^{m(j)+2} d_h^\mu(s,a) \le 2\log\left(\frac{SAT}{\delta}\right) \end{cases}. \end{aligned} \tag{B.164}$$

Combine the above results to yield

$$\begin{cases} \widehat{N}_h^{\mathsf{epo},m(j)}(s,a) \vee 1 \overset{\text{(i)}}{\ge} \frac{2^{m(j)} d_h^\mu(s,a)}{8\log\left(\frac{SAT}{\delta}\right)} \overset{\text{(ii)}}{\ge} \frac{1}{32e^2\log\left(\frac{SAT}{\delta}\right)} n, & \text{if } 2^{m(j)+2}\cdot d_h^\mu(s,a) \ge \log\left(\frac{SAT}{\delta}\right), \\ \widehat{N}_h^{\mathsf{epo},m(j)}(s,a) \vee 1 \ge 1 \overset{\text{(iii)}}{\ge} \frac{1}{2e^2\log\left(\frac{SAT}{\delta}\right)} n & \text{if } 2^{m(j)+2}\cdot d_h^\mu(s,a) \le 2\log\left(\frac{SAT}{\delta}\right), \end{cases} \tag{B.165}$$

where (i) follows from (B.162), (ii) and (iii) hold due to (B.164). As a result, we arrive at

$$\sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{\eta_n^N}{\widehat{N}_h^{\mathsf{epo},m(j)}(s,a)\vee 1} \le \sum_{n=N_h^{(m(j)+1,1)}+1}^{N_h^{(m(j)+2,1)}\wedge N} \frac{32e^2\log\left(\frac{SAT}{\delta}\right)\eta_n^N}{n} \le \sum_{n=N_h^{(m(j)+1,1)}+1}^{N} \frac{32e^2\log\left(\frac{SAT}{\delta}\right)\eta_n^N}{n} \le \frac{64e^2\log\left(\frac{SAT}{\delta}\right)}{N\vee 1},$$

where the last inequality holds since $\sum_{i=1}^N \frac{\eta_i^N}{i} \le \frac{2}{N\vee 1}$ (see Lemma 1).

B.3.4.3 Proof of inequality (B.153)

In this subchapter, we intend to control the following term

$$W_2 := \frac{1}{N_h^k\vee 1}\sum_{n=1}^{N_h^k}\mathsf{Var}_{h,s,a}\left(\overline{V}_{h+1}^{\mathsf{next},k^n}\right) - \left(\sigma_h^{\mathsf{ref},k^{N_h^k+1}}(s,a) - \left(\mu_h^{\mathsf{ref},k^{N_h^k+1}}(s,a)\right)^2\right)$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. First, it is easily seen that if $N_h^k=0$, then we have $W_2=0$ and thus (B.153) is satisfied. Therefore, the remainder of the proof is devoted to verifying (B.153) when $N_h^k = N_h^k(s,a)\ge 1$.

Combining the expression (B.55) with the following definition

$$\mathsf{Var}_{h,s,a}\left(\overline{V}_{h+1}^{\mathsf{next},k^n}\right) = P_{h,s,a}\left(\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 - \left(P_{h,s,a}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2,$$

we arrive at

$$\begin{aligned} W_2 &= \frac{1}{N_h^k\vee 1}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}\left(\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 - \left(P_{h,s,a}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2\right) \\ &\quad - \frac{1}{N_h^k\vee 1}\sum_{n=1}^{N_h^k}P_h^{k^n}\left(\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 + \left(\frac{1}{N_h^k\vee 1}\sum_{n=1}^{N_h^k}P_h^{k^n}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 \\ &= \underbrace{\frac{1}{N_h^k\vee 1}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}-P_h^{k^n}\right)\left(\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2}_{=:W_2^1} + \underbrace{\left(\frac{1}{N_h^k\vee 1}\sum_{n=1}^{N_h^k}P_h^{k^n}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 - \frac{1}{N_h^k\vee 1}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2}_{=:W_2^2}. \end{aligned} \tag{B.166}$$

In the sequel, we intend to control the terms in (B.166) separately.

Step 1: controlling W_2^1. The first term W_2^1 can be controlled by invoking Lemma 24 and set

$$W_{h+1}^i \coloneqq \left(\overline{V}_{h+1}^{\mathsf{next},i}\right)^2, \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \frac{1}{N} =: C_u.$$

To proceeding, with the fact

$$\left\|W_{h+1}^i\right\|_\infty \le \left\|\overline{V}_{h+1}^{\mathsf{next},i}\right\|_\infty^2 \le H^2 =: C_{\mathrm{w}}$$

and $N = N_h^k(s,a) = N_h^k$, applying Lemma 24 with the above quantities, we have for all state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
|W_2^1| &= \left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s,a} - P_h^{k^n}\right)\left(\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2\right| = \left|\sum_{i=1}^{k} X_i\left(s,a,h,N_h^k\right)\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N)\mathsf{Var}_{h,s,a}\left(W_{h+1}^{k_h^n(s,a)}\right)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{\iota^2}{N_h^k}}\sqrt{\|W_{h+1}^i\|_\infty^2} + \frac{H^2\iota^2}{N_h^k} \lesssim \sqrt{\frac{H^4\iota^2}{N_h^k}} + \frac{H^2\iota^2}{N_h^k}. \qquad \text{(B.167)}
\end{aligned}$$

Step 2: controlling W_2^2. Towards controlling W_2^2 in (B.166), we observe that by Jensen's inequality,

$$\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_{h,s,a}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 \le \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_{h,s,a}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2.$$

Equipped with this relation, W_2^2 satisfies

$$\begin{aligned}
W_2^2 &\le \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_h^{k^n}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 - \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k} P_{h,s,a}\overline{V}_{h+1}^{\mathsf{next},k^n}\right)^2 \\
&= \left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s,a}\right)\overline{V}_{h+1}^{\mathsf{next},k^n}\right)\cdot\left(\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} + P_{h,s,a}\right)\overline{V}_{h+1}^{\mathsf{next},k^n}\right). \qquad \text{(B.168)}
\end{aligned}$$

As for the first term in (B.168), let us set

$$W_{h+1}^i \coloneqq \overline{V}_{h+1}^{\mathsf{next},i}, \qquad \text{and} \qquad u_h^i(s,a,N) \coloneqq \frac{1}{N} =: C_u,$$

which satisfy

$$\left\|W_{h+1}^i\right\|_\infty \le \left\|\overline{V}_{h+1}^{\mathsf{next},i}\right\|_\infty \le H =: C_{\mathrm{w}}.$$

For any (s,a), Lemma 24 together with the above quantities and $N = N_h^k = N_h^k(s,a)$ gives

$$\begin{aligned}
&\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} - P_{h,s,a}\right)\overline{V}_{h+1}^{\mathsf{next},k^n}\right| \\
&\lesssim \sqrt{C_{\mathrm{u}}\log^2\frac{SAT}{\delta}}\sqrt{\sum_{n=1}^{N_h^k(s,a)} u_h^{k_h^n(s,a)}(s,a,N)\mathsf{Var}_{h,s,a}\big(W_{h+1}^{k_h^n(s,a)}\big)} + \left(C_{\mathrm{u}}C_{\mathrm{w}} + \sqrt{\frac{C_{\mathrm{u}}}{N}}C_{\mathrm{w}}\right)\log^2\frac{SAT}{\delta} \\
&\lesssim \sqrt{\frac{\iota^2}{N_h^k}}\sqrt{\left\|W_{h+1}^{k_h^n(s,a)}\right\|_\infty^2} + \frac{H\iota^2}{N_h^k} \lesssim \sqrt{\frac{H^2\iota^2}{N_h^k}} + \frac{H\iota^2}{N_h^k}
\end{aligned}$$

with probability at least $1-\delta$. In addition, the second term can be bounded straightforwardly by

$$\left|\frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(P_h^{k^n} + P_{h,s,a}\right)\overline{V}_{h+1}^{\mathsf{next},k^n}\right| \le \frac{1}{N_h^k}\sum_{n=1}^{N_h^k}\left(\left\|P_h^{k^n}\right\|_1 + \left\|P_{h,s,a}\right\|_1\right)\left\|\overline{V}_{h+1}^{\mathsf{next},k^n}\right\|_\infty \le 2H,$$

where the last inequality is valid since $\left\|\overline{V}_{h+1}^{\mathsf{next},k^n}\right\|_\infty \le H$ and $\left\|P_h^{k^n}\right\|_1 = \left\|P_{h,s,a}\right\|_1 = 1$. Substitution of the above two observations back into (B.168) yields

$$W_2^2 \lesssim \sqrt{\frac{H^4}{N_h^k \vee 1}\iota^2} + \frac{H^2}{N_h^k \vee 1}\iota^2. \tag{B.169}$$

Step 3: combining the above results. Plugging the results in (B.167) and (B.169) into (B.166), we reach

$$W_2 \le W_2^1 + W_2^2 \lesssim \sqrt{\frac{H^4}{N_h^k \vee 1}\iota^2} + \frac{H^2}{N_h^k \vee 1}\iota^2,$$

thus establishing the desired inequality (B.153).

Appendix C

Proofs for Chapter 5

C.1 Proof of auxiliary lemmas: episodic finite-horizon MDPs

C.1.1 Proof of Lemma 13

(a) Let us begin by proving the claim (5.11a). Recall from our construction that $\mathcal{D}^{\mathsf{aux}}$ is composed of the second half of the sample trajectories, and hence for each $s \in \mathcal{S}$ and $1 \leq h \leq H$,

$$N_h^{\mathsf{aux}}(s) = \sum_{k=K/2+1}^{K} \mathbb{1}\{s_h^k = s\}$$

can be viewed as the sum of $K/2$ independent Bernoulli random variables, each with mean $d_h^{\mathsf{b}}(s)$. According to the union bound and the Bernstein inequality, we obtain

$$\begin{aligned}\mathbb{P}\left\{\exists(s,h) \in \mathcal{S} \times [H] : \left|N_h^{\mathsf{aux}}(s) - \frac{K}{2} d_h^{\mathsf{b}}(s)\right| \geq \tau\right\} &\leq \sum_{s\in\mathcal{S}, h\in[H]} \mathbb{P}\left\{\left|N_h^{\mathsf{aux}}(s) - \frac{K}{2} d_h^{\mathsf{b}}(s)\right| \geq \tau\right\} \\ &\leq 2SH \exp\left(-\frac{\tau^2/2}{v_{s,h} + \tau/3}\right)\end{aligned}$$

for any $\tau \geq 0$, where

$$v_{s,h} \coloneqq \frac{K}{2}\mathsf{Var}\big(\mathbb{1}\{s_h^t = s\}\big) = \frac{K d_h^{\mathsf{b}}(s)\big(1 - d_h^{\mathsf{b}}(s)\big)}{2} \leq \frac{K d_h^{\mathsf{b}}(s)}{2}.$$

A little algebra then yields that with probability at least $1 - 2\delta$, one has

$$\left|N_h^{\mathsf{aux}}(s) - \frac{K}{2} d_h^{\mathsf{b}}(s)\right| \leq \sqrt{4 v_{s,h} \log\frac{HS}{\delta}} + \frac{2}{3}\log\frac{HS}{\delta} \leq \sqrt{2K d_h^{\mathsf{b}}(s) \log\frac{HS}{\delta}} + \log\frac{HS}{\delta} \tag{C.1}$$

simultaneously for all $s \in \mathcal{S}$ and all $1 \leq h \leq H$. The same argument also reveals that with probability exceeding $1 - 2\delta$,

$$\left|N_h^{\mathsf{main}}(s) - \frac{K}{2} d_h^{\mathsf{b}}(s)\right| \leq \sqrt{2K d_h^{\mathsf{b}}(s) \log\frac{HS}{\delta}} + \log\frac{HS}{\delta} \tag{C.2}$$

holds simultaneously for all $s \in \mathcal{S}$ and all $1 \le h \le H$. Combine (C.1) and (C.2) to show that

$$\left|N_h^{\mathsf{main}}(s) - N_h^{\mathsf{aux}}(s)\right| \le 2\sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} + 2\log\frac{HS}{\delta} \tag{C.3}$$

for all $s \in \mathcal{S}$ and all $1 \le h \le H$.

To establish the claimed result (5.11a), we divide into two cases.

- *Case 1:* $N_h^{\mathsf{aux}}(s) \le 100\log\frac{HS}{\delta}$. By construction, it is easily seen that

$$N_h^{\mathsf{trim}}(s) = \max\left\{N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{HS}{\delta}},\, 0\right\} = 0 \le N_h^{\mathsf{main}}(s). \tag{C.4}$$

- *Case 2:* $N_h^{\mathsf{aux}}(s) > 100\log\frac{HS}{\delta}$. In this case, invoking (C.1) reveals that

$$\frac{K}{2}d_h^{\mathsf{b}}(s) + \sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} + \log\frac{HS}{\delta} \ge N_h^{\mathsf{aux}}(s) > 100\log\frac{HS}{\delta},$$

and hence one necessarily has

$$Kd_h^{\mathsf{b}}(s) \ge (9\sqrt{2})^2\log\frac{HS}{\delta} \ge 100\log\frac{HS}{\delta}. \tag{C.5}$$

In turn, this property (C.5) taken collectively with (C.28) ensures that

$$N_h^{\mathsf{aux}}(s) \ge \frac{K}{2}d_h^{\mathsf{b}}(s) - \sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} - \log\frac{HS}{\delta} \ge \frac{K}{4}d_h^{\mathsf{b}}(s). \tag{C.6}$$

Therefore, in the case with $N_h^{\mathsf{aux}}(s) > 100\log\frac{HS}{\delta}$, we can demonstrate that

$$\begin{aligned} N_h^{\mathsf{trim}}(s) &= \max\left\{N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{HS}{\delta}},\, 0\right\} = N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{HS}{\delta}} \\ &\overset{\text{(i)}}{\le} N_h^{\mathsf{aux}}(s) - 5\sqrt{Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} \overset{\text{(ii)}}{\le} N_h^{\mathsf{aux}}(s) - \left\{2\sqrt{2Kd_h^{\mathsf{b}}(s)\log\frac{HS}{\delta}} + 2\log\frac{HS}{\delta}\right\} \\ &\overset{\text{(iii)}}{\le} N_h^{\mathsf{main}}(s), \end{aligned} \tag{C.7}$$

where (i) comes from Condition (C.6), (ii) is valid under the condition (C.5), and (iii) holds true with probability at least $1 - 2\delta$ due to the inequality (C.3).

Putting the above two cases together establishes the claim (5.11a).

(b) We now turn to the second claim (5.11b). Towards this, we first claim that the following bound

holds simultaneously for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ with probability exceeding $1-2\delta$:

$$N_h^{\mathsf{trim}}(s,a) \geq N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) - \sqrt{4N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s)\log\frac{KH}{\delta}} - \log\frac{KH}{\delta}. \tag{C.8}$$

Let us take this claim as given for the moment, and return to establish it towards the end of this subchapter. We shall discuss the following two cases separately.

- If $Kd_h^{\mathsf{b}}(s,a) = Kd_h^{\mathsf{b}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) > 1600\log\frac{KH}{\delta}$, then it follows from (C.6) (with slight modification) that

$$N_h^{\mathsf{aux}}(s) \geq \frac{K}{4}d_h^{\mathsf{b}}(s) \geq 400\log\frac{KH}{\delta}. \tag{C.9}$$

  This property together with the definition of $N_h^{\mathsf{trim}}(s)$ in turn allows us to derive

$$\begin{aligned} N_h^{\mathsf{trim}}(s) &\geq N_h^{\mathsf{aux}}(s) - 10\sqrt{N_h^{\mathsf{aux}}(s)\log\frac{KH}{\delta}} \geq \frac{K}{4}d_h^{\mathsf{b}}(s) - 10\sqrt{\frac{K}{4}d_h^{\mathsf{b}}(s)\log\frac{KH}{\delta}} \\ &\geq \frac{K}{8}d_h^{\mathsf{b}}(s), \end{aligned}$$

  and as a result,

$$N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) \geq \frac{K}{8}d_h^{\mathsf{b}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) = \frac{K}{8}d_h^{\mathsf{b}}(s,a) \geq 200\log\frac{KH}{\delta},$$

  where the last inequality arises from the assumption of this case. Taking this lower bound with (C.8) implies that

$$\begin{aligned} N_h^{\mathsf{trim}}(s,a) &\geq \frac{K}{8}d_h^{\mathsf{b}}(s,a) - \sqrt{\frac{K}{2}d_h^{\mathsf{b}}(s,a)\log\frac{KH}{\delta}} - \log\frac{KH}{\delta} \\ &\geq \frac{K}{8}d_h^{\mathsf{b}}(s,a) - 2\sqrt{Kd_h^{\mathsf{b}}(s,a)\log\frac{KH}{\delta}}. \end{aligned}$$

- If $Kd_h^{\mathsf{b}}(s,a) \leq 1600\log\frac{KH}{\delta}$, then one can easily verify that

$$\frac{K}{8}d_h^{\mathsf{b}}(s,a) - 5\sqrt{Kd_h^{\mathsf{b}}(s,a)\log\frac{KH}{\delta}} \leq 0 \leq N_h^{\mathsf{trim}}(s,a).$$

Putting these two cases together concludes the proof, provided that the claim (C.8) is valid.

Proof of inequality (C.8). Let us look at two cases separately.

- If $N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) \leq 4\log\frac{KH}{\delta}$, then the right-hand side of (C.8) is negative, and hence the claim (C.8) holds trivially.

- We then turn attention to the following set:

$$\mathcal{A}_{\mathsf{large}} \coloneqq \left\{(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] \;\middle|\; N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) > 4\log\frac{KH}{\delta}\right\}. \tag{C.10}$$

  Recognizing that

$$\begin{aligned}\sum_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) &= \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{trim}}(s)\sum_{a\in\mathcal{A}}\pi_h^{\mathsf{b}}(a\,|\,s) = \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{trim}}(s)\\ &\le \sum_{(s,h)\in\mathcal{S}\times[H]} N_h^{\mathsf{aux}}(s) = \frac{KH}{2},\end{aligned}$$

  we can immediately bound the cardinality of $\mathcal{A}_{\mathsf{large}}$ as follows:

$$\left|\mathcal{A}_{\mathsf{large}}\right| < \frac{\sum_{(s,a,h)} N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s)}{4\log\frac{KH}{\delta}} \le KH/2. \tag{C.11}$$

  Additionally, it follows from our construction that: conditional on $N_h^{\mathsf{trim}}(s)$, $N_h^{\mathsf{main}}(s)$ and the high-probability event (5.11a), $N_h^{\mathsf{trim}}(s,a)$ can be viewed as the sum of $\min\left\{N_h^{\mathsf{trim}}(s), N_h^{\mathsf{main}}(s)\right\} = N_h^{\mathsf{trim}}(s)$ independent Bernoulli random variables each with mean $\pi_h(a\,|\,s)$. As a result, repeating the Bernstein-type argument in (C.28) on the event (5.11a) reveals that, with probability at least $1-2\delta/(KH)$,

$$N_h^{\mathsf{trim}}(s,a) \ge N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s) - \sqrt{4N_h^{\mathsf{trim}}(s)\pi_h^{\mathsf{b}}(a\,|\,s)\log\frac{KH}{\delta}} - \log\frac{KH}{\delta} \tag{C.12}$$

  for any fixed triple (s,a,h). Taking the union bound over all $(s,a,h)\in\mathcal{A}_{\mathsf{large}}$ and using the bound (C.11) imply that with probability exceeding $1-\delta$, (C.12) holds simultaneously for all $(s,a,h)\in\mathcal{A}_{\mathsf{large}}$.

Combining the above two cases allows one to conclude that with probability at least $1-\delta$, the advertised property (C.8) holds simultaneously for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$.

C.2 Proof of auxiliary lemmas: infinite-horizon MDPs

C.2.1 Proof of Lemma 14

Before embarking on the proof, we introduce several notation. To make explicit the dependency on V, we shall express the penalty term using the following notation throughout this subchapter:

$$b(s,a;V) = \min\left\{\max\left\{\sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(V)}, \frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\}, \frac{1}{1-\gamma}\right\} + \frac{5}{N} \tag{C.13}$$

For any $Q, Q_1, Q_2 \in \mathbb{R}^{SA}$, we write

$$V(s) \coloneqq \max_a Q(s,a), \qquad V_1(s) \coloneqq \max_a Q_1(s,a) \qquad \text{and} \qquad V_2(s) \coloneqq \max_a Q_2(s,a) \tag{C.14}$$

for all $s \in \mathcal{S}$. Unless otherwise noted, we assume that

$$Q(s,a), Q_1(s,a), Q_2(s,a) \in \left[0, \frac{1}{1-\gamma}\right] \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}$$

throughout this subchapter. In addition, let us define another operator $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ obeying

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a) = r(s,a) - b(s,a;V) + \gamma \widehat{P}_{s,a} V \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A} \tag{C.15}$$

for any $Q \in \mathbb{R}^{SA}$. It is self-evident that

$$\widehat{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a) = \max\left\{\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a), 0\right\} \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{C.16}$$

γ-contraction. The main step of the proof lies in showing the monotonicity of the operator $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ in the sense that

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q) \le \widetilde{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) \qquad \text{for any } Q \le \widetilde{Q}. \tag{C.17}$$

Suppose that this claim is valid for the moment, then one can demonstrate that: for any $Q_1, Q_2 \in \mathbb{R}^{SA}$,

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2) \le \widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 + \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2), \tag{C.18a}$$

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2) \ge \widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 - \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2), \tag{C.18b}$$

with 1 denoting the all-one vector. Additionally, observe that

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) = \mathsf{Var}_{\widehat{P}_{s,a}}(V + c \cdot 1) \qquad \text{and hence} \qquad b(s,a;V) = b(s,a;V + c \cdot 1)$$

for any constant c, which together with the identity $\widehat{P}1 = 1$ immediately leads to

$$\begin{aligned}
\left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 - \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty &\le \gamma \left\|\widehat{P}\big(\|Q_1 - Q_2\|_\infty 1\big)\right\|_\infty = \gamma \|Q_1 - Q_2\|_\infty, \\
\left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}\big(Q_2 + \|Q_1 - Q_2\|_\infty 1\big) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty &\le \gamma \left\|\widehat{P}\big(\|Q_1 - Q_2\|_\infty 1\big)\right\|_\infty = \gamma \|Q_1 - Q_2\|_\infty.
\end{aligned}$$

Taking this together with (C.18) yields

$$\left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty,$$

which combined with the basic property $\left\|\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_1)-\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty \le \left\|\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_1)-\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty$ (as a result of (C.16)) justifies that

$$\left\|\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_1)-\widehat{\mathcal{T}}_{\mathsf{pe}}(Q_2)\right\|_\infty \le \gamma\|Q_1-Q_2\|_\infty. \tag{C.19}$$

The remainder of the proof is thus devoted to establishing the monotonicity property (C.17).

Proof of the monotonicity property (C.17). Consider any point $Q\in\mathbb{R}^{SA}$, and we would like to examine the derivative of $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ at point Q. Towards this end, we consider any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and divide into several cases.

- *Case 1:* $\max\left\{\sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(V)},\ \frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\} > \frac{1}{1-\gamma}$. In this case, the penalty term (C.13) simplifies to
$$b(s,a;V)=\frac{1}{1-\gamma}+\frac{5}{N}.$$
Taking the derivative of $\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)$ w.r.t. the s'-th component of V leads to
$$\frac{\partial\big(\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)\big)}{\partial V(s')}=\frac{\partial\big(r(s,a)-\frac{1}{1-\gamma}+\gamma\widehat{P}_{s,a}V\big)}{\partial V(s')}=\gamma\widehat{P}(s'\,|\,s,a)\ge 0 \tag{C.20}$$
for any $s'\in\mathcal{S}$.

- *Case 2:* $\sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(V)} < \frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} < \frac{1}{1-\gamma}$. The penalty (C.13) in this case reduces to
$$b(s,a;V)=\frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}+\frac{5}{N},$$
an expression that is independent of V. As a result, repeating the argument for Case 1 indicates that (C.20) continues to hold for this case.

- *Case 3:* $\frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} < \sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(V)} < \frac{1}{1-\gamma}$. In this case, the penalty term is given by
$$b(s,a;V)=\sqrt{\frac{c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(V)}+\frac{5}{N}.$$
Note that in this case, we necessarily have
$$\mathsf{Var}_{\widehat{P}_{s,a}}(V)\ge\frac{4c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2N(s,a)},$$

which together with the definition in (1.7) indicates that

$$\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a}V\big)^2 \geq \frac{4c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a)} > 0. \tag{C.21}$$

As a result, for any $s' \in \mathcal{S}$, taking the derivative of $b(s,a;V)$ w.r.t. the s'-th component of V gives

$$\begin{aligned}
\frac{\partial b(s,a;V)}{\partial V(s')} &= \sqrt{\frac{c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)}} \frac{\partial \sqrt{\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a}V\big)^2}}{\partial V(s')} \\
&= \sqrt{\frac{c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)}} \frac{\widehat{P}(s' \,|\, s,a)V(s') - \big(\widehat{P}_{s,a}V\big)\widehat{P}(s' \,|\, s,a)}{\sqrt{\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a}V\big)^2}} \\
&\leq \sqrt{\frac{c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)}} \frac{\widehat{P}(s' \,|\, s,a)V(s')}{\sqrt{\widehat{P}_{s,a}(V \circ V) - \big(\widehat{P}_{s,a}V\big)^2}} \\
&\leq \frac{1}{2}(1-\gamma)\widehat{P}(s' \,|\, s,a)V(s') \leq \gamma \widehat{P}(s' \,|\, s,a),
\end{aligned}$$

where the penultimate inequality relies on (C.21), and the last inequality is valid since $V(s') = \max_a Q(s',a) \leq \frac{1}{1-\gamma}$ and $\gamma \geq 1/2$. In turn, the preceding relation allows one to derive

$$\frac{\partial\big(\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)\big)}{\partial V(s')} = \gamma \widehat{P}(s' \,|\, s,a) - \frac{\partial b(s,a;V)}{\partial V(s')} \geq 0$$

for any $s' \in \mathcal{S}$.

Putting the above cases together reveals that

$$\frac{\partial\big(\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)(s,a)\big)}{\partial V(s')} \geq 0 \qquad \text{for all } (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$$

holds almost everywhere (except for the boundary points of these cases). Recognizing that $\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q)$ is continuous in Q and that V is non-decreasing in Q, one can immediately conclude that

$$\widetilde{\mathcal{T}}_{\mathsf{pe}}(Q) \leq \widetilde{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) \qquad \text{for any } Q \leq \widetilde{Q}. \tag{C.22}$$

Existence and uniqueness of fixed points. To begin with, note that for any $0 \leq Q \leq \frac{1}{1-\gamma} \cdot 1$, one has $0 \leq \widehat{\mathcal{T}}_{\mathsf{pe}}(Q) \leq \frac{1}{1-\gamma} \cdot 1$. If we produce the following sequence recursively:

$$Q^{(0)} = 0 \qquad \text{and} \qquad Q^{(t+1)} = \widehat{\mathcal{T}}_{\mathsf{pe}}(Q^{(t)}) \quad \text{for all } t \geq 0,$$

then the standard proof for the Banach fixed-point theorem (e.g., Agarwal et al. (2001, Theorem 1)) tells us that $Q^{(t)}$ converges to some point $Q^{(\infty)}$ as $t \to \infty$. Clearly, $Q^{(\infty)}$ is a fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}$ obeying $0 \le Q^{(\infty)} \le \frac{1}{1-\gamma} \cdot 1$.

We then turn to justifying the uniqueness of fixed points of $\widehat{\mathcal{T}}_{\mathsf{pe}}$. Suppose that there exists another point $\widetilde{Q}$ obeying $\widetilde{Q} = \widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q})$, which clearly satisfies $\widetilde{Q} \ge 0$. If $\|\widetilde{Q}\|_\infty > \frac{1}{1-\gamma}$, then

$$\big\|\widetilde{Q}\big\|_\infty = \big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q})\big\|_\infty \le \|r\|_\infty + \gamma\|\widehat{P}\|_1\big\|\widetilde{Q}\big\|_\infty \le 1 + \gamma\big\|\widetilde{Q}\big\|_\infty < (1-\gamma)\big\|\widetilde{Q}\big\|_\infty + \gamma\big\|\widetilde{Q}\big\|_\infty = \big\|\widetilde{Q}\big\|_\infty,$$

resulting in contradiction. Consequently, one necessarily has $0 \le \widetilde{Q} \le \frac{1}{1-\gamma} \cdot 1$. Further, the γ-contraction property (C.19) implies that

$$\big\|\widetilde{Q} - Q^{(\infty)}\big\|_\infty = \big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) - \widehat{\mathcal{T}}_{\mathsf{pe}}(Q^{(\infty)})\big\|_\infty \le \gamma\big\|\widetilde{Q} - Q^{(\infty)}\big\|_\infty.$$

Given that $\gamma < 1$, this inequality cannot happen unless $\widetilde{Q} = Q^\infty$, thus confirming the uniqueness of Q^∞.

C.2.2 Proof of Lemma 15

Let us first recall the monotone non-decreasing property (C.17) of the operator $\widetilde{\mathcal{T}}_{\mathsf{pe}}$ defined in (C.15), which taken together with the property (C.16) readily yields

$$\widehat{\mathcal{T}}_{\mathsf{pe}}(Q) \le \widehat{\mathcal{T}}_{\mathsf{pe}}(\widetilde{Q}) \tag{C.23}$$

for any Q and $\widetilde{Q}$ obeying $Q \le \widetilde{Q}$, $0 \le Q \le \frac{1}{1-\gamma} \cdot 1$ and $0 \le \widetilde{Q} \le \frac{1}{1-\gamma} \cdot 1$ (with 1 the all-one vector). Given that $\widehat{Q}_0 = 0 \le \widehat{Q}^\star_{\mathsf{pe}}$, we can apply (C.23) to obtain

$$\widehat{Q}_1 = \widehat{\mathcal{T}}_{\mathsf{pe}}(Q_0) \le \widehat{\mathcal{T}}_{\mathsf{pe}}(\widehat{Q}^\star_{\mathsf{pe}}) = \widehat{Q}^\star_{\mathsf{pe}}.$$

Repeat this argument recursively to arrive at

$$\widehat{Q}_\tau \le \widehat{Q}^\star_{\mathsf{pe}} \qquad \text{for all } \tau \ge 0.$$

In addition, it comes directly from Lemma 14 that

$$\begin{aligned}\big\|\widehat{Q}_\tau - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty &= \big\|\widehat{\mathcal{T}}_{\mathsf{pe}}(\widehat{Q}_{\tau-1}) - \widehat{\mathcal{T}}_{\mathsf{pe}}(\widehat{Q}^\star_{\mathsf{pe}})\big\|_\infty \le \gamma\big\|\widehat{Q}_{\tau-1} - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \\ &\le \cdots \le \gamma^\tau\big\|\widehat{Q}_0 - \widehat{Q}^\star_{\mathsf{pe}}\big\|_\infty \\ &\le \frac{\gamma^\tau}{1-\gamma}\end{aligned} \tag{C.24}$$

for any $\tau \ge 0$, where the last inequality is valid since $\widehat{Q}_0 = 0$ and $\|\widehat{Q}^\star_{\mathsf{pe}}\|_\infty \le \frac{1}{1-\gamma}$ (see Lemma 14).

The other claim (5.38) also follows immediately by taking the right-hand side of (C.24) to be no larger than $1/N$.

C.2.3 Proof of Lemma 19

For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, if $\frac{Nd^{\mathsf{b}}(s,a)}{12} < \frac{2}{3}\log\frac{SN}{\delta}$, then it is self-evident that this pair satisfies (5.69). As a consequence, it suffices to focus attention on the following set of state-action pairs:

$$\mathcal{N}_{\mathsf{large}} := \left\{(s,a) \,\middle|\, d^{\mathsf{b}}(s,a) \geq \frac{8\log\frac{SN}{\delta}}{N}\right\}. \tag{C.25}$$

To bound the cardinality of $\mathcal{N}_{\mathsf{large}}$, we make the observation that

$$\left|\mathcal{N}_{\mathsf{large}}\right| \cdot \frac{8\log\frac{SN}{\delta}}{N} \leq \sum_{(s,a)\in\mathcal{N}_{\mathsf{large}}} d^{\mathsf{b}}(s,a) \leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\mathsf{b}}(s,a) \leq 1,$$

thus leading to the crude bound

$$\left|\mathcal{N}_{\mathsf{large}}\right| \leq \frac{N}{8\log\frac{SN}{\delta}} \leq \frac{N}{8}. \tag{C.26}$$

Let us now look at any $(s,a) \in \mathcal{N}_{\mathsf{large}}$. Given that $N(s,a)$ can be viewed as the sum of N independent Bernoulli random variables each with mean $d^{\mathsf{b}}(s,a)$, we can apply the Bernstein inequality to yield

$$\mathbb{P}\Big\{\Big|N(s,a) - Nd^{\mathsf{b}}(s,a)\Big| \geq \tau\Big\} \leq 2\exp\left(-\frac{\tau^2/2}{v_{s,a}+\tau/3}\right)$$

for any $\tau \geq 0$, where we define

$$v_{s,a} := N\mathsf{Var}\Big(\mathbb{1}\big\{(s_i,a_i)=(s,a)\big\}\Big) \leq Nd^{\mathsf{b}}(s,a).$$

A little algebra then yields that with probability at least $1-\delta$,

$$\Big|N(s,a) - Nd^{\mathsf{b}}(s,a)\Big| \leq \sqrt{4v_{s,a}\log\frac{2}{\delta}} + \frac{2}{3}\log\frac{2}{\delta} \leq \sqrt{4Nd^{\mathsf{b}}(s,a)\log\frac{2}{\delta}} + \log\frac{2}{\delta}. \tag{C.27}$$

Combining this result with the union bound over $(s,a) \in \mathcal{N}_{\mathsf{large}}$ and making use of (C.26) give: with probability at least $1-\delta$,

$$\Big|N(s,a) - Nd^{\mathsf{b}}(s,a)\Big| \leq \sqrt{4Nd^{\mathsf{b}}(s,a)\log\frac{N}{\delta}} + \log\frac{N}{\delta} \tag{C.28}$$

holds simultaneously for all $(s,a) \in \mathcal{N}_{\mathsf{large}}$. Recalling that $Nd^{\mathsf{b}}(s,a) \geq 8\log\frac{NS}{\delta}$ holds for any

$(s,a) \in \mathcal{N}_{\mathsf{large}}$, we can easily verify that

$$N(s,a) \geq Nd^{\mathsf{b}}(s,a) - \left(\sqrt{4Nd^{\mathsf{b}}(s,a)\log\frac{N}{\delta}} + \log\frac{N}{\delta}\right) \geq \frac{Nd^{\mathsf{b}}(s,a)}{12}, \tag{C.29}$$

thereby establishing (5.69) for any $(s,a) \in \mathcal{N}_{\mathsf{large}}$. This concludes the proof.

C.2.4 Proof of Lemma 20

If $N(s,a) = 0$, then the inequalities hold trivially. Hence, it is sufficient to focus on the case where $N(s,a) > 0$. Before proceeding, we make note of a key Bernstein-style result; the proof is deferred to Appendix C.2.4.1.

Lemma 31. *Consider any given pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ with $N(s,a) > 0$. Let $V \in \mathbb{R}^S$ be any vector independent of $\widehat{P}_{s,a}$ obeying $\|V\|_\infty \leq \frac{1}{1-\gamma}$. With probability at least $1-4\delta$, one has*

$$\left|(\widehat{P}_{s,a} - P_{s,a})V\right| \leq \sqrt{\frac{48\mathsf{Var}_{\widehat{P}_{s,a}}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{48\log\frac{N}{\delta}}{(1-\gamma)N(s,a)} \tag{C.30a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) \leq 2\mathsf{Var}_{P_{s,a}}(V) + \frac{5\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)} \tag{C.30b}$$

Remark 8. In words, Lemma 31 develops a Bernstein bound (C.30a) on $\left|(\widehat{P}_{s,a} - P_{s,a})V\right|$ that makes clear the importance of the variance parameter. Lemma 31 (cf. (C.30b)) also ascertains that the variance w.r.t. the empirical distribution $\widehat{P}_{s,a}$ does not deviate much from the variance w.r.t. the true distribution $P_{s,a}$.

Equipped with this result, we are now ready to present the proof of Lemma 20, which is built upon a leave-one-out decoupling argument and consists of the following steps.

Step 1: construction of auxiliary state-absorbing MDPs. Recall that $\widehat{\mathcal{M}}$ is the empirical MDP. For each state $s \in \mathcal{S}$ and each scalar $u \geq 0$, we construct an auxiliary state-absorbing MDP $\widehat{\mathcal{M}}^{s,u}$ in a way that makes it identical to the empirical MDP $\widehat{\mathcal{M}}$ except for state s. More specifically, the transition kernel of the auxiliary MDP $\widehat{\mathcal{M}}^{s,u}$ — denoted by $P^{s,u}$ — is chosen such that

$$\begin{aligned} P^{s,u}(\widetilde{s}\,|\,s,a) &= \mathbb{1}(\widetilde{s} = s) && \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A}, \\ P^{s,u}(\cdot\,|\,s',a) &= \widehat{P}(\cdot\,|\,s',a) && \text{for all } (s',a) \in \mathcal{S}\times\mathcal{A} \text{ and } s' \neq s; \end{aligned}$$

and the reward function of $\widehat{\mathcal{M}}^{s,u}$ — denoted by $r^{s,u}$ — is set to be

$$\begin{aligned} r^{s,u}(s,a) &= u && \text{for all } a \in \mathcal{A}, \\ r^{s,u}(s',a) &= r(s',a) && \text{for all } (s',a) \in \mathcal{S}\times\mathcal{A} \text{ and } s' \neq s. \end{aligned}$$

In words, the probability transition kernel of $\widehat{\mathcal{M}}^{s,u}$ is obtained by dropping all randomness of $\widehat{P}_{s,a}$ ($a \in \mathcal{A}$) that concerns state s and making s an absorbing state. In addition, let us define the pessimistic Bellman operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{s,u}$ based on the auxiliary MDP $\widehat{\mathcal{M}}^{s,u}$ such that

$$\widehat{\mathcal{T}}_{\mathsf{pe}}^{s,u}(Q)(s,a) \coloneqq \max\Big\{r^{s,u}(s,a) + \gamma P_{s,a}^{s,u} V - b^{s,u}(s,a;V),\, 0\Big\} \tag{C.31}$$

for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, where the penalty term is taken to be

$$b^{s,u}(s,a;V) = \min\left\{\max\left\{\sqrt{\frac{c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{N(s,a)} \mathsf{Var}_{P^{s,u}(\cdot\,|\,s,a)}(V)},\ \frac{2c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}\right\},\ \frac{1}{1-\gamma}\right\} + \frac{5}{N}. \tag{C.32}$$

Step 2: the correspondence between the empirical MDP and auxiliary MDP. Taking

$$u^\star = (1-\gamma)\widehat{V}_{\mathsf{pe}}^\star(s) + \min\left\{\frac{2c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)\max_a N(s,a)},\ \frac{1}{1-\gamma}\right\} + \frac{5}{N}, \tag{C.33}$$

we claim that there exists a fixed point $\widehat{Q}_{s,u^\star}^\star$ of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{s,u^\star}$ whose corresponding value function $\widehat{V}_{s,u^\star}^\star$ coincides with $\widehat{V}_{\mathsf{pe}}^\star$. To justify this, it suffices to verify the following properties:

- Consider any $a \in \mathcal{A}$. Given that $P^{s,u}(\cdot\,|\,s,a)$ only has a single non-zero entry (equal to 1), it is easily seen that $\mathsf{Var}_{P^{s,u}(\cdot\,|\,s,a)}(V) = 0$ holds for any V and any u, thus indicating that

$$b^{s,u}(s,a;V) = \min\left\{\frac{2c_{\mathrm{b}} \log \frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},\ \frac{1}{1-\gamma}\right\} + \frac{5}{N}. \tag{C.34}$$

  Consequently, for state s, one has

$$\begin{aligned}
\max_a \Big\{r^{s,u^\star}(s,a) - b^{s,u^\star}\big(s,a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma\big\langle P^{s,u^\star}(\cdot\,|\,s,a),\ \widehat{V}_{\mathsf{pe}}^\star\big\rangle\Big\} &= \max_a \Big\{u^\star - b^{s,u^\star}\big(s,a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma \widehat{V}_{\mathsf{pe}}^\star(s)\Big\} \\
&= u^\star - \min_a b^{s,u^\star}\big(s,a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma \widehat{V}_{\mathsf{pe}}^\star(s) \\
&= (1-\gamma)\widehat{V}_{\mathsf{pe}}^\star(s) + \gamma \widehat{V}_{\mathsf{pe}}^\star(s) \\
&= \widehat{V}_{\mathsf{pe}}^\star(s),
\end{aligned} \tag{C.35}$$

  where the third identity makes use of our choice (C.33) of $u^\star$ and (C.34).

- Next, consider any $s' \neq s$ and any $a \in \mathcal{A}$. We make the observation that

$$\begin{aligned}
&\max\Big\{r^{s,u^\star}(s',a) - b^{s,u^\star}\big(s',a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma\big\langle P^{s,u^\star}(\cdot\,|\,s',a),\ \widehat{V}_{\mathsf{pe}}^\star\big\rangle,\, 0\Big\} \\
&\quad = \max\Big\{r(s',a) - b\big(s',a;\widehat{V}_{\mathsf{pe}}^\star\big) + \gamma\big\langle \widehat{P}(\cdot\,|\,s',a),\ \widehat{V}_{\mathsf{pe}}^\star\big\rangle,\, 0\Big\} = \widehat{Q}_{\mathsf{pe}}^\star(s',a),
\end{aligned} \tag{C.36}$$

where the last relation holds since $\widehat{Q}^\star_{\mathsf{pe}}$ is a fixed point of $\widehat{\mathcal{T}}_{\mathsf{pe}}$.

Armed with (C.35) and (C.36), we see that $\widehat{Q}^\star_{s,u^\star} = \widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}\big(\widehat{Q}^\star_{s,u^\star}\big)$ by taking

$$\begin{aligned}
&\max_{a\in\mathcal{A}} \widehat{Q}^\star_{s,u^\star}(s,a) = \widehat{V}^\star_{\mathsf{pe}}(s), \\
&\quad \widehat{Q}^\star_{s,u^\star}(s',a) = \widehat{Q}^\star_{\mathsf{pe}}(s',a) \qquad\qquad \text{for all } s'\neq s \text{ and } a\in\mathcal{A}.
\end{aligned}$$

This readily confirms the existence of a fixed point of $\widehat{\mathcal{T}}^{s,u^\star}_{\mathsf{pe}}$ whose corresponding value coincides with $\widehat{V}^\star_{\mathsf{pe}}$.

Step 3: building an ϵ-net. Consider any $(s,a)\in\mathcal{S}\times\mathcal{A}$ with $N(s,a)>0$. Construct a set $\mathcal{U}_{\mathrm{cover}}$ as follows

$$\mathcal{U}_{\mathrm{cover}} := \left\{ \frac{i}{N} \mid 1\le i\le N u_{\max} \right\}, \tag{C.37}$$

with $u_{\max} = \min\left\{ \frac{2c_{\mathrm{b}}\log\frac{N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \frac{1}{1-\gamma}\right\} + \frac{5}{N} + 1$. This can be viewed as the ϵ-net (Vershynin, 2018) of the range $[0,u_{\max}]\subseteq\big[0,\frac{2}{1-\gamma}\big]$ with $\epsilon = 1/N$. Let us construct an auxiliary MDP $\widehat{\mathcal{M}}^{s,u}$ as in Step 1 for each $u\in\mathcal{U}_{\mathrm{cover}}$. Repeating the argument in the proof of Lemma 14 (see Chapter C.2.1), we can easily show that there exists a unique fixed point $\widehat{Q}^\star_{s,u}$ of $\widehat{\mathcal{M}}^{s,u}$, which also obeys $0\le\widehat{Q}^\star_{s,u}\le\frac{1}{1-\gamma}\cdot 1$. In what follows, we denote by $\widehat{V}^\star_{s,u}$ the corresponding value function of $\widehat{Q}^\star_{s,u}$.

Recognizing that $\widehat{\mathcal{M}}^{s,u}$ is statistically independent from $\widehat{P}_{s,a}$ for any $u\in\mathcal{U}_{\mathrm{cover}}$ (by construction), we can apply Lemma 31 in conjunction with the union bound (over all $u\in\mathcal{U}_{\mathsf{cover}}$) to show that, with probability exceeding $1-\delta$,

$$\left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}^\star_{s,u}\right| \le \sqrt{\frac{48\log\frac{8N^2}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big)} + \frac{48\log\frac{8N^2}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \tag{C.38a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big) \le 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u}\big) + \frac{5\log\frac{8N^2}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)} \tag{C.38b}$$

hold simultaneously for all $u\in\mathcal{U}_{\mathrm{cover}}$. Clearly, the total number of (s,a) pairs with $N(s,a)>0$ cannot exceed N. Thus, taking the union bound over all these pairs yield that, with probability at least $1-\delta$,

$$\left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}^\star_{s,u}\right| \le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big)} + \frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}, \tag{C.39a}$$

$$\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^\star_{s,u}\big) \le 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^\star_{s,u}\big) + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)} \tag{C.39b}$$

hold simultaneously for all $(s,a,u) \in \mathcal{S} \times \mathcal{A} \times \mathcal{U}_{\text{cover}}$ obeying $N(s,a) > 0$.

Step 4: a covering argument. In this step, we shall work on the high-probability event (C.38) that holds simultaneously for all $u \in \mathcal{U}_{\text{cover}}$. Given that $\widehat{V}^{\star}_{\mathsf{pe}}$ satisfies the trivial bound $0 \le \widehat{V}^{\star}_{\mathsf{pe}}(s) \le \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, one can find some $u_0 \in \mathcal{U}_{\text{cover}}$ such that $|u_0 - u^{\star}| \le 1/N$, where we recall the choice of $u^{\star}$ in (C.33). From the definition of the MDP $\widehat{\mathcal{M}}^{s,u}$ and the operator (C.31), it is readily seen that

$$\big\|\widehat{\mathcal{T}}^{s,u_0}_{\mathsf{pe}}(Q) - \widehat{\mathcal{T}}^{s,u^{\star}}_{\mathsf{pe}}(Q)\big\|_{\infty} \le \big|u_0 - u^{\star}\big| \le \frac{1}{N}$$

holds for any $Q \in \mathbb{R}^{SA}$. Consequently, we can use γ-contraction of the operator to obtain

$$\begin{aligned}
\big\|\widehat{Q}^{\star}_{s,u_0} - \widehat{Q}^{\star}_{s,u^{\star}}\big\|_{\infty} &= \Big\|\widehat{\mathcal{T}}^{s,u_0}_{\mathsf{pe}}\big(\widehat{Q}^{\star}_{s,u_0}\big) - \widehat{\mathcal{T}}^{s,u^{\star}}_{\mathsf{pe}}\big(\widehat{Q}^{\star}_{s,u^{\star}}\big)\Big\|_{\infty} \\
&\le \Big\|\widehat{\mathcal{T}}^{s,u^{\star}}_{\mathsf{pe}}\big(\widehat{Q}^{\star}_{s,u_0}\big) - \widehat{\mathcal{T}}^{s,u^{\star}}_{\mathsf{pe}}\big(\widehat{Q}^{\star}_{s,u^{\star}}\big)\Big\|_{\infty} + \Big\|\widehat{\mathcal{T}}^{s,u_0}_{\mathsf{pe}}\big(\widehat{Q}^{\star}_{s,u_0}\big) - \widehat{\mathcal{T}}^{s,u^{\star}}_{\mathsf{pe}}\big(\widehat{Q}^{\star}_{s,u_0}\big)\Big\|_{\infty} \\
&\le \gamma\big\|\widehat{Q}^{\star}_{s,u_0} - \widehat{Q}^{\star}_{s,u^{\star}}\big\|_{\infty} + \frac{1}{N},
\end{aligned}$$

which implies that

$$\big\|\widehat{Q}^{\star}_{s,u_0} - \widehat{Q}^{\star}_{s,u^{\star}}\big\|_{\infty} \le \frac{1}{(1-\gamma)N}$$

and therefore

$$\big\|\widehat{V}^{\star}_{s,u_0} - \widehat{V}^{\star}_{s,u^{\star}}\big\|_{\infty} \le \big\|\widehat{Q}^{\star}_{s,u_0} - \widehat{Q}^{\star}_{s,u^{\star}}\big\|_{\infty} \le \frac{1}{(1-\gamma)N}.$$

This in turn allows us to demonstrate that

$$\begin{aligned}
&\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big) - \mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u^{\star}}\big) \\
&\quad= P_{s,a}\Big(\big(\widehat{V}^{\star}_{s,u_0} - P_{s,a}\widehat{V}^{\star}_{s,u_0}\big)\circ\big(\widehat{V}^{\star}_{s,u_0} - P_{s,a}\widehat{V}^{\star}_{s,u_0}\big) - \big(\widehat{V}^{\star}_{s,u^{\star}} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}}\big)\circ\big(\widehat{V}^{\star}_{s,u^{\star}} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}}\big)\Big) \\
&\quad\le P_{s,a}\Big(\big(\widehat{V}^{\star}_{s,u_0} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}}\big)\circ\big(\widehat{V}^{\star}_{s,u_0} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}}\big) - \big(\widehat{V}^{\star}_{s,u^{\star}} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}}\big)\circ\big(\widehat{V}^{\star}_{s,u^{\star}} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}}\big)\Big) \\
&\quad\le P_{s,a}\Big(\big(\widehat{V}^{\star}_{s,u_0} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}} + \widehat{V}^{\star}_{s,u^{\star}} - P_{s,a}\widehat{V}^{\star}_{s,u^{\star}}\big)\circ\big(\widehat{V}^{\star}_{s,u_0} - \widehat{V}^{\star}_{s,u^{\star}}\big)\Big) \\
&\quad\le \frac{2}{1-\gamma}\Big|P_{s,a}\big(\widehat{V}^{\star}_{s,u_0} - \widehat{V}^{\star}_{s,u^{\star}}\big)\Big| \le \frac{2}{1-\gamma}\big\|\widehat{V}^{\star}_{s,u_0} - \widehat{V}^{\star}_{s,u^{\star}}\big\|_{\infty} \le \frac{2}{(1-\gamma)^2 N},
\end{aligned}$$

where the third line comes from the fact that $\mathbb{E}[X] = \arg\min_c \mathbb{E}[(X-c)^2]$, and the last line relies on the property $0 \le \widehat{V}^{\star}_{s,u_0}, \widehat{V}^{\star}_{s,u^{\star}} \le \frac{1}{1-\gamma}$. In addition, by swapping $\widehat{V}^{\star}_{s,u_0}$ and $\widehat{V}^{\star}_{s,u^{\star}}$, we can derive

$$\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u^{\star}}\big) - \mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big) \le \frac{2}{(1-\gamma)^2 N},$$

and then

$$\left|\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big)-\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big)\right|\le\frac{2}{(1-\gamma)^2N}. \tag{C.40}$$

Clearly, this bound (C.40) continues to be valid if we replace $P_{s,a}$ with $\widehat{P}_{s,a}$.

With the above perturbation bounds in mind, we can invoke the triangle inequality and (C.39a) to reach

$$\begin{aligned}
\left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}^{\star}_{\mathsf{pe}}\right| &= \left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}^{\star}_{s,u^\star}\right| \le \left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}^{\star}_{s,u_0}\right| + \left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\big(\widehat{V}^{\star}_{s,u^\star}-\widehat{V}^{\star}_{s,u_0}\big)\right| \\
&\le \left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\widehat{V}^{\star}_{s,u_0}\right| + \frac{2}{N(1-\gamma)} \\
&\le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big)} + \frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{2}{N(1-\gamma)} \\
&\le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big)} + \sqrt{\frac{96\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)^2N(s,a)}} + \frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{2}{N(1-\gamma)} \\
&\le \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big)} + \frac{60\log\frac{8N^3}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},
\end{aligned} \tag{C.41}$$

where the second line holds since

$$\left|\big(\widehat{P}_{s,a}-P_{s,a}\big)\big(\widehat{V}^{\star}_{s,u^\star}-\widehat{V}^{\star}_{s,u_0}\big)\right| \le \big(\big\|\widehat{P}_{s,a}\big\|_1+\big\|P_{s,a}\big\|_1\big)\big\|\widehat{V}^{\star}_{s,u^\star}-\widehat{V}^{\star}_{s,u_0}\big\|_\infty \le \frac{2}{N(1-\gamma)},$$

the penultimate line is valid due to (C.40), and the last line holds true under the conditions that $T\ge N(s,a)$ and that T is sufficiently large. Moreover, apply (C.39b) and the triangle inequality to arrive at

$$\begin{aligned}
\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big) = \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big) &\le \mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big) + \left|\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big)-\mathsf{Var}_{\widehat{P}_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big)\right| \\
&\overset{\text{(i)}}{\le} 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big) + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)} + \frac{2}{(1-\gamma)^2N} \\
&\le 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big) + 2\left|\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u^\star}\big)-\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{s,u_0}\big)\right| + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)} + \frac{2}{(1-\gamma)^2N} \\
&\overset{\text{(ii)}}{\le} 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big) + \frac{6}{(1-\gamma)^2N} + \frac{5\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)} \\
&\le 2\mathsf{Var}_{P_{s,a}}\big(\widehat{V}^{\star}_{\mathsf{pe}}\big) + \frac{23\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2N(s,a)},
\end{aligned} \tag{C.42}$$

where (i) arise from (C.39b) and (C.40), (ii) follows from (C.40), and the last line holds true since $N \geq N(s,a)$.

Step 5: extending the bounds to $\widetilde{V}$. Consider any $\widetilde{V}$ obeying $\|\widetilde{V} - \widehat{V}^{\star}_{\mathsf{pe}}\|_{\infty} \leq \frac{1}{N}$ and $\|\widetilde{V}\|_{\infty} \leq \frac{1}{1-\gamma}$. Invoke (C.41) and the triangle inequality to arrive at

$$\begin{aligned}
\left|(\widehat{P}_{s,a} - P_{s,a})\widetilde{V}\right| &\leq \left|(\widehat{P}_{s,a} - P_{s,a})\widehat{V}^{\star}_{\mathsf{pe}}\right| + \left|(\widehat{P}_{s,a} - P_{s,a})(\widehat{V}^{\star}_{\mathsf{pe}} - \widetilde{V})\right| \\
&\leq \sqrt{\frac{48\log\frac{8N^3}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(\widehat{V}^{\star}_{s,u^{\star}})} + \frac{60\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} + \frac{2}{N}, \\
&\leq 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(\widehat{V}^{\star}_{s,u^{\star}})} + \frac{62\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} \\
&= 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(\widehat{V}^{\star}_{\mathsf{pe}})} + \frac{62\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)},
\end{aligned} \tag{C.43}$$

where the penultimate inequality relies on $N \geq N(s,a)$, and the second line holds since

$$\left|(\widehat{P}_{s,a} - P_{s,a})(\widehat{V}^{\star}_{\mathsf{pe}} - \widetilde{V})\right| \leq (\|\widehat{P}_{s,a}\|_1 + \|P_{s,a}\|_1)\|\widehat{V}^{\star}_{\mathsf{pe}} - \widetilde{V}\|_{\infty} \leq \frac{2}{N}.$$

Given that $\left\|\widetilde{V} - \widehat{V}^{\star}_{\mathsf{pe}}\right\|_{\infty} \leq 1/N$, we can repeat the argument for (C.40) allows one to demonstrate that

$$\left|\mathsf{Var}_{\widehat{P}_{s,a}}(\widehat{V}^{\star}_{\mathsf{pe}}) - \mathsf{Var}_{\widehat{P}_{s,a}}(\widetilde{V})\right| \leq \frac{2}{(1-\gamma)^2 N}$$

which taken together with (C.43) and the basic inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ gives

$$\begin{aligned}
\left|(\widehat{P}_{s,a} - P_{s,a})\widetilde{V}\right| &\leq 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(\widetilde{V})} + 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\cdot\frac{2}{(1-\gamma)^2 N}} + \frac{62\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)} \\
&\leq 12\sqrt{\frac{\log\frac{2N}{(1-\gamma)\delta}}{N(s,a)}\mathsf{Var}_{\widehat{P}_{s,a}}(\widetilde{V})} + \frac{74\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)N(s,a)}.
\end{aligned}$$

Additionally, repeating the argument for (C.42) leads to another desired inequality:

$$\begin{aligned}
\mathsf{Var}_{\widehat{P}_{s,a}}(\widetilde{V}) &\leq 2\mathsf{Var}_{P_{s,a}}(\widetilde{V}) + \frac{6}{(1-\gamma)N} + \frac{23\log\frac{8N^3}{(1-\gamma)\delta}}{3(1-\gamma)^2 N(s,a)} \\
&\leq 2\mathsf{Var}_{P_{s,a}}(\widetilde{V}) + \frac{41\log\frac{2N}{(1-\gamma)\delta}}{(1-\gamma)^2 N(s,a)}.
\end{aligned}$$

C.2.4.1 Proof of Lemma 31

In this proof, we shalle often use $\mathsf{Var}_{s,a}$ to abbreviate $\mathsf{Var}_{P_{s,a}}$ for notational simplicity. Before proceeding, let us define the following vector

$$\overline{V} = V - (P_{s,a}V)\mathbf{1}, \tag{C.44}$$

with $\mathbf{1}$ denoting the all-one vector. It is clearly seen that

$$P_{s,a}\big(\overline{V} \circ \overline{V}\big) = P_{s,a}\big(V \circ V\big) - \big(P_{s,a}V\big)^2 = \mathsf{Var}_{s,a}(V). \tag{C.45}$$

In addition, we make note of the following basic facts that will prove useful:

$$\|V\|_\infty \le \frac{1}{1-\gamma}, \qquad \|\overline{V}\|_\infty \le \frac{1}{1-\gamma}, \qquad \|\overline{V} \circ \overline{V}\|_\infty \le \|\overline{V}\|_\infty^2 \le H^2, \tag{C.46a}$$

$$\mathsf{Var}_{s,a}\big(\overline{V} \circ \overline{V}\big) \le P_{s,a}\big(\overline{V} \circ \overline{V} \circ \overline{V} \circ \overline{V}\big) \le \frac{1}{(1-\gamma)^2} P_{s,a}\big(\overline{V} \circ \overline{V}\big) = \frac{1}{(1-\gamma)^2}\mathsf{Var}_{s,a}\big(V\big). \tag{C.46b}$$

Proof of inequality (C.30a). If $0 < N(s,a) < 48\log\frac{N}{\delta}$, then we can immediately see that

$$\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\Big| \le \|V\|_\infty \le \frac{1}{1-\gamma} \le \frac{48\log\frac{N}{\delta}}{(1-\gamma)N(s,a)}, \tag{C.47}$$

and hence the claim (C.30a) is valid. As a result, it suffices to focus on the case where

$$N(s,a) \ge 48\log\frac{N}{\delta}. \tag{C.48}$$

Note that the total number of pairs (s,a) with nonzero $N(s,a)$ cannot exceed N. Akin to (C.28), taking the Bernstein inequality together with (C.46) and invoking the union bound, we can demonstrate that with probability at least $1-4\delta$,

$$\begin{aligned}
\Big|\big(\widehat{P}_{s,a} - P_{s,a}\big)V\Big| &\le \sqrt{\frac{4\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{2\|V\|_\infty \log\frac{N}{\delta}}{3N(s,a)} \\
&\le \sqrt{\frac{4\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)N(s,a)} \qquad \text{(C.49a)} \\
\Big|\big(P_{s,a} - \widehat{P}_{s,a}\big)\big(\overline{V} \circ \overline{V}\big)\Big| &\le \sqrt{\frac{4\mathsf{Var}_{s,a}\big(\overline{V} \circ \overline{V}\big)\log\frac{N}{\delta}}{N(s,a)}} + \frac{2\|\overline{V} \circ \overline{V}\|_\infty \log\frac{N}{\delta}}{3N(s,a)} \\
&\le \sqrt{\frac{4\mathsf{Var}_{s,a}(V)\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}} + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)} \qquad \text{(C.49b)}
\end{aligned}$$

hold simultaneously over all (s,a) with $N(s,a) > 0$. Note, however, that the Bernstein bounds in

(C.49) involve the variance $\mathsf{Var}_{s,a}(V)$; we still need to connect $\mathsf{Var}_{s,a}(V)$ with its empirical estimate $\mathsf{Var}_{\widehat{P}_{s,a}}(V)$.

In the sequel, let us look at two cases separately.

- *Case 1:* $\mathsf{Var}_{s,a}(V) \le \frac{9\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}$. In this case, our bound (C.49a) immediately leads to

$$\left|(\widehat{P}_{s,a} - P_{s,a})V\right| \le \frac{7\log\frac{N}{\delta}}{(1-\gamma)N(s,a)}. \tag{C.50}$$

- *Case 2:* $\mathsf{Var}_{s,a}(V) > \frac{9\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}$. We first single out the following useful identity:

$$\begin{aligned}
\widehat{P}_{s,a}(\overline{V}\circ\overline{V}) - \mathsf{Var}_{\widehat{P}_{s,a}}(V) &= \widehat{P}_{s,a}(\overline{V}\circ\overline{V}) - \left[\widehat{P}_{s,a}(V\circ V) - (\widehat{P}_{s,a}V)^2\right] \\
&= \widehat{P}_{s,a}(V\circ V) - 2(\widehat{P}_{s,a}V)(P_{s,a}V) + (P_{s,a}V)^2 - \left[\widehat{P}_{s,a}(V\circ V) - (\widehat{P}_{s,a}V)^2\right] \\
&= \left|(\widehat{P}_{s,a} - P_{s,a})V\right|^2.
\end{aligned} \tag{C.51}$$

Combining (C.51) with (C.49b) then implies that, with probability exceeding $1-4\delta$,

$$\begin{aligned}
&\mathsf{Var}_{s,a}(V) = P_{s,a}(\overline{V}\circ\overline{V}) = (P_{s,a} - \widehat{P}_{s,a})(\overline{V}\circ\overline{V}) + \widehat{P}_{s,a}(\overline{V}\circ\overline{V}) \\
&= (P_{s,a} - \widehat{P}_{s,a})(\overline{V}\circ\overline{V}) + \left\{\left|(\widehat{P}_{s,a} - P_{s,a})V\right|^2 + \mathsf{Var}_{\widehat{P}_{s,a}}(V)\right\} \qquad \text{(C.52)} \\
&\le \sqrt{\frac{4\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}}\sqrt{\mathsf{Var}_{s,a}(V)} + \left|(\widehat{P}_{s,a} - P_{s,a})V\right|^2 + \mathsf{Var}_{\widehat{P}_{s,a}}(V) + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)} \\
&\le \frac{2}{3}\mathsf{Var}_{s,a}(V) + \left|(\widehat{P}_{s,a} - P_{s,a})V\right|^2 + \mathsf{Var}_{\widehat{P}_{s,a}}(V) + \frac{2\log\frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}, \qquad \text{(C.53)}
\end{aligned}$$

where the second line arises from the identity (C.51), the penultimate inequality results from (C.49b), and the last inequality holds true due to the assumption $\mathsf{Var}_{s,a}(V) > \frac{9\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}$ in this case. Rearranging terms of the above inequality, we are left with

$$\mathsf{Var}_{s,a}(V) \le 3\left|(\widehat{P}_{s,a} - P_{s,a})V\right|^2 + 3\mathsf{Var}_{\widehat{P}_{s,a}}(V) + \frac{2\log\frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}$$

Taking this upper bound on $\mathsf{Var}_{s,a}(V)$ collectively with (C.49a) and using a little algebra lead to

$$\left|(\widehat{P}_{s,a} - P_{s,a})V\right| \le \sqrt{\frac{12\log\frac{N}{\delta}}{N(s,a)}}\left|(\widehat{P}_{s,a} - P_{s,a})V\right| + \sqrt{\frac{12\mathsf{Var}_{\widehat{P}_{s,a}}(V)\log\frac{N}{\delta}}{N(s,a)}} + \frac{5\log\frac{N}{\delta}}{(1-\gamma)N(s,a)} \tag{C.54}$$

with probability at least $1-4\delta$. When $N(s,a) \geq 48 \log \frac{N}{\delta}$ (cf. (C.48)), one has $\sqrt{\frac{12 \log \frac{N}{\delta}}{N(s,a)}} \leq 1/2$. Substituting this into (C.54) and rearranging terms, we arrive at

$$\left|\left(\widehat{P}_{s,a} - P_{s,a}\right)V\right| \leq \sqrt{\frac{48 \mathsf{Var}_{\widehat{P}_{s,a}}(V) \log \frac{N}{\delta}}{N(s,a)}} + \frac{10 \log \frac{N}{\delta}}{(1-\gamma)N(s,a)}$$

with probability at least $1-4\delta$.

Putting the above two cases together establishes the advertised bound (C.30a).

Proof of inequality (C.30b). It follows from (C.52) and (C.49a) that with probability at least $1-4\delta$,

$$\begin{aligned}
\mathsf{Var}_{s,a}(V) &\geq -\left|\left(P_{s,a} - \widehat{P}_{s,a}\right)\left(\overline{V} \circ \overline{V}\right)\right| + \mathsf{Var}_{\widehat{P}_{s,a}}(V) \\
&\geq -\sqrt{\frac{4\mathsf{Var}_{s,a}\left(\overline{V}\right) \log \frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}} - \frac{2 \log \frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)} + \mathsf{Var}_{\widehat{P}_{s,a}}(V),
\end{aligned}$$

or equivalently,

$$\mathsf{Var}_{\widehat{P}_{s,a}}(V) \leq \mathsf{Var}_{s,a}(V) + 2\sqrt{\frac{\mathsf{Var}_{s,a}\left(\overline{V}\right) \log \frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}} + \frac{2 \log \frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}.$$

Invoke the elementary inequality $2xy \leq x^2 + y^2$ to establish the claimed bound:

$$\begin{aligned}
\mathsf{Var}_{\widehat{P}_{s,a}}(V) &\leq \mathsf{Var}_{s,a}(V) + \left(\mathsf{Var}_{s,a}\left(\overline{V}\right) + \frac{\log \frac{N}{\delta}}{(1-\gamma)^2 N(s,a)}\right) + \frac{2 \log \frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)} \\
&= 2\mathsf{Var}_{s,a}(V) + \frac{5 \log \frac{N}{\delta}}{3(1-\gamma)^2 N(s,a)}.
\end{aligned}$$

C.2.5 Proof of Theorem 5

To establish Theorem 5, we shall first generate a collection of hard problem instances (including MDPs and the associated batch datasets), and then conduct sample complexity analyses over these hard instances.

C.2.5.1 Construction of hard problem instances

Construction of the hard MDPs. To begin with, for any integer $H \geq 32$, let us consider a set $\Theta \subseteq \{0,1\}^H$ of H-dimensional vectors, which we shall construct shortly. We then generate a

collection of MDPs

$$\mathsf{MDP}(\Theta) = \Big\{ \mathcal{M}^\theta = \big(\mathcal{S}, \mathcal{A}, P^\theta = \{P_h^{\theta_h}\}_{h=1}^H, \{r_h\}_{h=1}^H, H\big) \mid \theta = [\theta_h]_{1 \le h \le H} \in \Theta \Big\}, \tag{C.55}$$

where

$$\mathcal{S} = \{0, 1, \ldots, S-1\}, \qquad \text{and} \qquad \mathcal{A} = \{0, 1\}.$$

To define the transition kernel of these MDPs, we find it convenient to introduce the following state distribution supported on the state subset $\{0, 1\}$:

$$\mu(s) = \frac{1}{CS} \mathbb{1}\{s = 0\} + \Big(1 - \frac{1}{CS}\Big) \mathbb{1}\{s = 1\}, \tag{C.56}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $C > 0$ is some constant that will determine the concentrability coefficient $C^\star_{\mathsf{clipped}}$ (as we shall detail momentarily). It is assumed that

$$\frac{1}{CS} \le \frac{1}{4}. \tag{C.57}$$

With this distribution in mind, we can specify the transition kernel $P^\theta = \{P_h^{\theta_h}\}_{h=1}^H$ of the MDP $\mathcal{M}^\theta$ as follows:

$$P_h^{\theta_h}(s' \mid s, a) = \begin{cases} p \mathbb{1}\{s' = 0\} + (1-p)\mu(s') & \text{if} \quad (s,a) = (0, \theta_h) \\ q \mathbb{1}\{s' = 0\} + (1-q)\mu(s') & \text{if} \quad (s,a) = (0, 1-\theta_h) \\ \mathbb{1}\{s' = 1\} & \text{if} \quad (s,a) = (1,0) \\ \big(1 - \frac{2c_1}{H}\big) \mathbb{1}\{s' = 1\} + \frac{2c_1}{H} \mu(s') & \text{if} \quad (s,a) = (1,1) \\ \big(1 - \frac{1}{H}\big) \mathbb{1}\{s' = s\} + \frac{1}{H} \mu(s') & \text{if} \quad s > 1 \end{cases} \tag{C.58}$$

for any $(s, a, s', h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H]$, where p and q are set to be

$$p = 1 - \frac{c_1}{H} + \frac{c_2 \varepsilon}{H^2} \qquad \text{and} \qquad q = 1 - \frac{c_1}{H} - \frac{c_2 \varepsilon}{H^2} \tag{C.59}$$

for $c_1 = 1/4$ and $c_2 = 4096$ such that

$$\frac{c_2 \varepsilon}{H^2} \le \frac{c_1}{2H} \le \frac{1}{8}. \tag{C.60}$$

It is readily seen from the above assumption that

$$p > q \ge \frac{1}{2}. \tag{C.61}$$

In view of the transition kernel (C.58), the MDP will never leave the state subset $\{0,1\}$ if its initial state belongs to $\{0,1\}$. The reward function of all these MDPs is chosen to be

$$r_h(s,a) = \begin{cases} 1 & \text{if} \quad s = 0 \\ \frac{1}{2} & \text{if} \quad (s,a) = (1,0) \\ 0 & \text{if} \quad (s,a) = (1,1) \\ 0 & \text{if} \quad s > 1 \end{cases} \tag{C.62}$$

for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

Finally, let us choose the set $\Theta \subseteq \{0,1\}^H$. By virtue of the Gilbert-Varshamov lemma (Gilbert, 1952), one can construct $\Theta \subseteq \{0,1\}^H$ in a way that

$$|\Theta| \geq e^{H/8} \qquad \text{and} \qquad \|\theta - \widetilde{\theta}\|_1 \geq \frac{H}{8} \quad \text{for any } \theta, \widetilde{\theta} \in \Theta \text{ obeying } \theta \neq \widetilde{\theta}. \tag{C.63}$$

In other words, the set Θ we construct contains an exponentially large number of vectors that are sufficiently separated. This property plays an important role in the ensuing analysis.

Value functions and optimal policies. Next, we look at the value functions of the constructed MDPs and identify the optimal policies. For the sake of notational clarity, for the MDP $\mathcal{M}_\theta$, we denote by $\pi^{\star,\theta} = \{\pi_h^{\star,\theta}\}_{h=1}^H$ the optimal policy, and let $V_h^{\pi,\theta}$ (resp. $V_h^{\star,\theta}$) indicate the value function of policy π (resp. $\pi^{\star,\theta}$) at time step h. The following lemma collects a couple of useful properties concerning the value functions and optimal policies; the proof can be found in Appendix E.2.3.5.

Lemma 32. *Consider any $\theta \in \Theta$ and any policy π. Then it holds that*

$$V_h^{\pi,\theta}(0) = 1 + \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big)V_{h+1}^{\pi,\theta}(0) + (1 - x_h^{\pi,\theta})\mu(1)V_{h+1}^{\pi,\theta}(1) \tag{C.64}$$

for any $h \in [H]$, where

$$x_h^{\pi,\theta} = p\pi_h(\theta_h \,|\, 0) + q\pi_h(1 - \theta_h \,|\, 0). \tag{C.65}$$

In addition, for any $h \in [H]$, the optimal policies and the optimal value functions obey

$$\pi_h^{\star,\theta}(\theta_h \,|\, 0) = 1, \qquad V_h^{\star,\theta}(0) \geq \tfrac{2}{3}(H + 1 - h), \tag{C.66a}$$

$$\pi_h^{\star,\theta}(0 \,|\, 1) = 1, \qquad V_h^{\star,\theta}(1) = \tfrac{1}{2}(H + 1 - h), \tag{C.66b}$$

provided that $0 < c_1 \leq 1/2$.

Construction of the batch dataset. A batch dataset is then generated, which consists of K *independent* sample trajectories each of length H. The initial state distribution ρ^{b} and the behavior

policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$ (according to (7.3)) are chosen as follows:

$$\rho^{\mathsf{b}}(s) = \mu(s) \quad \text{and} \quad \pi_h^{\mathsf{b}}(a \,|\, s) = \frac{1}{2}, \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H],$$

where μ has been defined in (E.67). As it turns out, for any MDP $\mathcal{M}_\theta$, the occupancy distributions of the above batch dataset admit the following simple characterization:

$$d_h^{\mathsf{b}}(s) = \mu(s), \qquad d_h^{\mathsf{b}}(s,a) = \frac{1}{2}\mu(s), \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{C.67}$$

Additionally, we shall choose the initial state distribution ρ as follows

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0, \\ 0, & \text{if } s > 0. \end{cases} \tag{C.68}$$

With this choice of ρ, the single-policy clipped concentrability coefficient $C^\star_{\mathsf{clipped}}$ and the quantity C are intimately connected as follows:

$$C^\star_{\mathsf{clipped}} = 2C. \tag{C.69}$$

The proof of the claims (E.69) and (E.80) can be found in Appendix E.2.3.3.

C.2.5.2 Establishing the minimax lower bound

We are now positioned to establish our sample complexity lower bounds. Recalling our choice of ρ in (E.70), our proof seeks to control the quantity

$$\left\langle \rho, V_1^{\star,\theta} - V_1^{\widehat{\pi},\theta} \right\rangle = V_1^{\star,\theta}(0) - V_1^{\widehat{\pi},\theta}(0),$$

where $\widehat{\pi}$ is any policy estimator computed based on the batch dataset.

Step 1: converting $\widehat{\pi}$ into an estimate $\widehat{\theta}$ of θ. Towards this, we first make the following claim: for an arbitrary policy π obeying

$$\sum_{h=1}^{H} \left\| \pi_h(\cdot \,|\, 0) - \pi_h^{\star,\theta}(\cdot \,|\, 0) \right\|_1 \geq \frac{H}{8}, \tag{C.70}$$

one has

$$\left\langle \rho, V_1^{\star,\theta} - V_1^{\pi,\theta} \right\rangle > \varepsilon. \tag{C.71}$$

We shall postpone the proof of this claim to Appendix E.2.3.3. Suppose for the moment that there exists a policy estimate $\widehat{\pi}$ that achieves

$$\mathbb{P}\left\{\left\langle \rho, V_1^{\star,\theta} - V_1^{\widehat{\pi},\theta}\right\rangle \leq \varepsilon\right\} \geq \frac{3}{4}, \tag{C.72}$$

then in view of (E.83), we necessarily have

$$\mathbb{P}\left\{\sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\right\|_1 < H/8\right\} \geq \frac{3}{4}. \tag{C.73}$$

With the above observation in mind, we are motivated to construct the following estimate $\widehat{\theta}$ for $\theta \in \Theta$:

$$\widehat{\theta} = \arg\min_{\overline{\theta}\in\Theta} \sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\overline{\theta}}(\cdot\,|\,0)\right\|_1. \tag{C.74}$$

If $\sum_h \left\|\widehat{\pi}(\cdot\,|\,0) - \pi^{\star,\theta}(\cdot\,|\,0)\right\|_1 < H/8$ holds for some $\theta \in \Theta$, then for any $\widetilde{\theta} \in \Theta$ with $\widetilde{\theta} \neq \theta$ one has

$$\begin{aligned}
\sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\widetilde{\theta}}(\cdot\,|\,0)\right\|_1 &\geq \sum_{h=1}^{H} \left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h^{\star,\widetilde{\theta}}(\cdot\,|\,0)\right\|_1 - \sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\right\|_1 \\
&= 2\left\|\theta - \widetilde{\theta}\right\|_1 - \sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\right\|_1 \\
&> \frac{H}{4} - \frac{H}{8} = \frac{H}{8},
\end{aligned} \tag{C.75}$$

where the first inequality holds by the triangle inequality, the second line arises from the fact $\pi_h^{\star,\theta}(\theta_h\,|\,0) = 1$ for all $1 \leq h \leq H$ (see (E.79)), and the last line comes from the properties (C.63) about Θ. Putting (C.74) and (C.75) together implies that $\widehat{\theta} = \theta$ if

$$\sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\right\|_1 < \frac{H}{8} < \sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\widetilde{\theta}}(\cdot\,|\,0)\right\|_1$$

is valid for all $\widetilde{\theta} \in \Theta$ with $\widetilde{\theta} \neq \theta$. As a consequence,

$$\mathbb{P}\big(\widehat{\theta} = \theta\big) \geq \mathbb{P}\left(\sum_{h=1}^{H} \left\|\widehat{\pi}_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\right\|_1 < \frac{H}{8}\right) \geq \frac{3}{4}. \tag{C.76}$$

In the sequel, we aim to demonstrate that (C.76) cannot possibly happen without enough samples, which would in turn contradict (E.84).

Step 2: probability of error in testing multiple hypotheses. Next, we turn attention to a $|\Theta|$-ary hypothesis testing problem. For any $\theta \in \Theta$, denote by $\mathbb{P}_\theta$ the probability distribution when the MDP is $\mathcal{M}_\theta$. We will then study the minimax probability of error defined as follows:

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max_{\theta \in \Theta} \mathbb{P}_\theta(\psi \neq \theta), \tag{C.77}$$

where the infimum is taken over all possible tests ψ (constructed based on the batch dataset available).

Let $\mu^{\mathsf{b},\theta}$ (resp. $\mu_h^{\mathsf{b},\theta_h}(s_h)$) represent the distribution of a sample trajectory $\{s_1, a_1, s_2, a_2, \cdots, s_H, a_H\}$ (resp. a sample (a_h, s_{h+1}) conditional on s_h) for the MDP $\mathcal{M}_\theta$. Recalling that the K trajectories in the batch dataset are independently generated, one obtains

$$\begin{aligned} p_{\mathrm{e}} &\overset{\text{(i)}}{\geq} 1 - \frac{K \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big) + \log 2}{\log|\Theta|} \\ &\overset{\text{(ii)}}{\geq} 1 - \frac{8K}{H} \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big) - \frac{8\log 2}{H} \\ &\overset{\text{(iii)}}{\geq} \frac{1}{2} - \frac{8K}{H} \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big), \end{aligned} \tag{C.78}$$

where (i) arises from Fano's inequality (cf. (Tsybakov, 2009, Corollary 2.6)) and the additivity property of the KL divergence (cf. Tsybakov (2009, Page 85)), (ii) holds since $|\Theta| \geq e^{H/8}$ (according to our construction (C.63)), and (iii) is valid when $H \geq 16\log 2$. Recalling that the occupancy state distribution d_h^{b} is the same for any MDP $\mathcal{M}_\theta$ with $\theta \in \Theta$ (see (E.69)), one can invoke the chain rule of the KL divergence (Duchi, 2018, Lemma 5.2.8) and the Markovian nature of the sample trajectories to obtain

$$\mathsf{KL}\big(\mu^{\mathsf{b},\theta} \,\|\, \mu^{\mathsf{b},\widetilde{\theta}}\big) = \sum_{h=1}^{H} \mathop{\mathbb{E}}_{s_h\sim d_h^{\mathsf{b}}} \Big[\mathsf{KL}\big(\mu_h^{\mathsf{b},\theta_h}(s_h) \,\|\, \mu_h^{\mathsf{b},\widetilde{\theta}_h}(s_h)\big)\Big] = \frac{1}{2}\mu(0) \sum_{h=1}^{H} \sum_{a\in\{0,1\}} \mathsf{KL}\left(P_h^{\theta_h}(\cdot\,|\,0,a) \,\|\, P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,a)\right),$$

where the last identity holds true since (by construction and (E.69))

$$\begin{aligned} \mathop{\mathbb{E}}_{s_h\sim d_h^{\mathsf{b}}} \Big[\mathsf{KL}\big(\mu_h^{\mathsf{b},\theta_h}(s_h) \,\|\, \mu_h^{\mathsf{b},\widetilde{\theta}_h}(s_h)\big)\Big] &= \sum_s d_h^{\mathsf{b}}(s) \left\{ \sum_{a,s'} \pi_h^{\mathsf{b}}(a\,|\,s) P_h^{\theta_h}(s'\,|\,s,a) \log \frac{\pi_h^{\mathsf{b}}(a\,|\,s) P_h^{\theta_h}(s'\,|\,s,a)}{\pi_h^{\mathsf{b}}(a\,|\,s) P_h^{\widetilde{\theta}_h}(s'\,|\,s,a)} \right\} \\ &= \frac{1}{2}\mu(0) \sum_a \sum_{s'} P_h^{\theta_h}(s'\,|\,0,a) \log \frac{P_h^{\theta_h}(s'\,|\,0,a)}{P_h^{\widetilde{\theta}_h}(s'\,|\,0,a)} \\ &= \frac{1}{2}\mu(0) \sum_a \mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,a) \,\|\, P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,a)\big). \end{aligned}$$

Substitution into (C.78) yields

$$p_{\mathrm{e}} \geq \frac{1}{2} - \frac{4K\mu(0)}{H} \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \sum_{h=1}^{H} \Big[\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,0) \,\|\, P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,0)\big) + \mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,1) \,\|\, P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,1)\big)\Big]. \tag{C.79}$$

It then boils down to bounding the KL divergence terms in (E.88). If $\theta_h = \widetilde{\theta}_h$, then it is self-evident that

$$\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,0) \,\|\, P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,0)\big) + \mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,1) \,\|\, P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,1)\big) = 0. \tag{C.80}$$

Consider now the case that $\theta_h \neq \widetilde{\theta}_h$, and suppose without loss of generality that $\theta_h = 0$ and $\widetilde{\theta}_h = 1$. It is seen that

$$\begin{aligned}
P_h^{\theta_h}(0\,|\,0,0) &= P_h^{\theta_h}(\theta_h\,|\,0,0) = \left(1 - \frac{1}{CS}\right) p + \frac{1}{CS}, \\
P_h^{\widetilde{\theta}_h}(0\,|\,0,0) &= P_h^{\widetilde{\theta}_h}(1 - \widetilde{\theta}_h\,|\,0,0) = \left(1 - \frac{1}{CS}\right) q + \frac{1}{CS}.
\end{aligned}$$

Given that $p \geq q \geq 1/2$ (see (C.61)), we can apply Lemma 60 to arrive at

$$\begin{aligned}
\mathsf{KL}\left(P_h^{\theta_h}(0\,|\,0,0) \,\|\, P_h^{\widetilde{\theta}_h}(0\,|\,0,0)\right) &= \mathsf{KL}\left(\left(1 - \frac{1}{CS}\right) p + \frac{1}{CS} \,\|\, \left(1 - \frac{1}{CS}\right) q + \frac{1}{CS}\right) \\
&\leq \frac{\left(1 - \frac{1}{CS}\right)^2 (p-q)^2}{\left(\left(1 - \frac{1}{CS}\right) p + \frac{1}{CS}\right)\left(1 - p - (1-p)\frac{1}{CS}\right)} \\
&\overset{\text{(i)}}{\leq} \frac{\left(1 - \frac{1}{CS}\right)^2 (p-q)^2}{\left(\left(1 - \frac{1}{CS}\right) p\right)\left((1-p)(1 - \frac{1}{CS})\right)} = \frac{4(c_2)^2\varepsilon^2}{H^4 p(1-p)} \\
&\overset{\text{(ii)}}{=} \frac{4(c_2)^2\varepsilon^2}{H^4\left(1 - \frac{c_1}{H} + \frac{c_2\varepsilon}{H^2}\right)\left(\frac{c_1}{H} - \frac{c_2\varepsilon}{H^2}\right)} \\
&\leq \frac{4(c_2)^2\varepsilon^2}{H^4 \frac{1}{2}\frac{c_1}{2H}} = \frac{16(c_2)^2\varepsilon^2}{c_1 H^3},
\end{aligned} \tag{C.81}$$

where (i) and (ii) make use of the definition (E.61) of (p, q), and the last line follows as long as $\frac{c_2\varepsilon}{H^2} \leq \frac{c_1}{2H} \leq \frac{1}{4}$. Similarly, it can be easily verified that $\mathsf{KL}\big(P_h^{\theta_h}(0\,|\,0,1) \,\|\, P_h^{\widetilde{\theta}_h}(0\,|\,0,1)\big)$ can be upper bounded in the same way. Substituting (E.89) and (C.80) back into (E.88) indicates that: if the sample size obeys

$$N = KH \leq \frac{c_1 CSH^4}{512(c_2)^2\varepsilon^2} = \frac{c_1 C^\star_{\mathsf{clipped}} SH^4}{1024(c_2)^2\varepsilon^2}, \tag{C.82}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{2} - \frac{4K\mu(0)}{H} \max_{\theta,\widetilde{\theta}\in\Theta,\,\theta\neq\widetilde{\theta}} \sum_{h=1}^{H} \Big[\mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,0) \parallel P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,0)\big) + \mathsf{KL}\big(P_h^{\theta_h}(\cdot\,|\,0,1) \parallel P_h^{\widetilde{\theta}_h}(\cdot\,|\,0,1)\big)\Big]$$
$$\geq \frac{1}{2} - \frac{4K\mu(0)}{H} \sum_{h=1}^{H} \frac{32(c_2)^2\varepsilon^2}{c_1 H^3} \geq \frac{1}{4}. \tag{C.83}$$

Step 3: combining the above results. Suppose that there exists an estimator $\widehat{\pi}$ satisfying

$$\max_{\theta\in\Theta} \mathbb{P}_\theta \left\{ \big\langle \rho, V_1^{\star,\theta} - V_1^{\widehat{\pi},\theta} \big\rangle \geq \varepsilon \right\} < \frac{1}{4}, \tag{C.84}$$

where $\mathbb{P}_\theta$ denotes the probability when the MDP is $\mathcal{M}_\theta$. Then in view of the analysis in Step 1, we must have

$$\mathbb{P}_\theta \left(\sum_{h=1}^{H} \big\|\widehat{\pi}(\cdot\,|\,0) - \pi^{\star,\theta}(\cdot\,|\,0)\big\|_1 < \frac{H}{8} \right) \geq \frac{3}{4}, \qquad \text{for all } \theta \in \Theta,$$

and as a consequence of (C.76), the estimator $\widehat{\theta}$ defined in (C.74) must satisfy

$$\mathbb{P}_\theta\big(\widehat{\theta} \neq \theta\big) < \frac{1}{4}, \qquad \text{for all } \theta \in \Theta. \tag{C.85}$$

Nevertheless, this cannot possibly happen under the sample size condition (C.82); otherwise it is contradictory to the result in (C.83). This concludes the proof by inserting $c_1 = 1/4$ and $c_2 = 4096$.

C.2.5.3 Proof of Lemma 63

To start with, for any policy π, it is observed that the value function of state $s = 0$ at step h is

$$\begin{aligned}
V_h^{\pi,\theta}(0) &= \mathop{\mathbb{E}}_{a\sim\pi_h(\cdot\,|\,0)} \left[1 + \sum_{s'} P_h^{\theta_h}(s'\,|\,0,a) V_{h+1}^{\pi,\theta}(s')\right] \\
&= 1 + \pi_h(\theta_h\,|\,0) \Big[\big(p + (1-p)\mu(0)\big) V_{h+1}^{\pi,\theta}(0) + (1-p)\mu(1) V_{h+1}^{\pi,\theta}(1)\Big] \\
&\quad + \pi(1-\theta_h\,|\,0) \Big[\big(q + (1-q)\mu(0)\big) V_{h+1}^{\pi,\theta}(0) + (1-q)\mu(1) V_{h+1}^{\pi,\theta}(1)\Big] \\
&= 1 + \Big[p\pi_h(\theta_h\,|\,0) + q\pi(1-\theta_h\,|\,0) + \mu(0) - p\pi_h(\theta_h\,|\,0)\mu(0) - q\pi(1-\theta_h\,|\,0)\mu(0)\Big] V_{h+1}^{\pi,\theta}(0) \\
&\quad + \mu(1)\Big[1 - p\pi_h(\theta_h\,|\,0) - q\pi(1-\theta_h\,|\,0)\Big] V_{h+1}^{\pi,\theta}(1) \\
&\overset{\text{(i)}}{=} 1 + \Big[x_h^{\pi,\theta} + (1 - x_h^{\pi,\theta})\mu(0) V_{h+1}^{\pi,\theta}(0) + (1 - x_h^{\pi,\theta})\mu(1) V_{h+1}^{\pi,\theta}(1)\Big] \\
&\overset{\text{(ii)}}{=} 1 + \big(\mu(1) x_h^{\pi} + \mu(0)\big) V_{h+1}^{\pi,\theta}(0) + (1 - x_h^{\pi})\mu(1) V_{h+1}^{\pi,\theta}(1),
\end{aligned} \tag{C.86}$$

where (i) is valid due to the choice

$$x_h^{\pi,\theta} = p\pi_h(\theta_h \,|\, 0) + q\pi_h(1-\theta_h \,|\, 0), \tag{C.87}$$

and (ii) holds since $\mu(0) + \mu(1) = 1$.

Additionally, the value function of state 1 at any step h obeys

$$\begin{aligned} V_h^{\pi,\theta}(1) &= \pi_h(0\,|\,1)\left(\frac{1}{2} + V_{h+1}^{\pi,\theta}(1)\right) + \pi_h(1\,|\,1)\left[\left(1 - \frac{2c_1}{HCS}\right)V_{h+1}^{\pi,\theta}(1) + \frac{2c_1}{HCS}V_{h+1}^{\pi,\theta}(0)\right] \qquad \text{(C.88)} \\ &\overset{\text{(i)}}{\le} \pi_h(0\,|\,1)\left(\frac{1}{2} + V_{h+1}^{\pi,\theta}(1)\right) + \pi_h(1\,|\,1)\left[\left(1 - \frac{2c_1}{HCS}\right)V_{h+1}^{\pi,\theta}(1) + \frac{2c_1}{HCS}(H-h)\right] \\ &\overset{\text{(ii)}}{\le} \pi_h(0\,|\,1)\left(\frac{1}{2} + V_{h+1}^{\pi,\theta}(1)\right) + \pi_h(1\,|\,1)\left[\frac{1}{2} + \left(1 - \frac{2c_1}{HCS}\right)V_{h+1}^{\pi,\theta}(1)\right] \\ &= \frac{1}{2} + V_{h+1}^{\pi,\theta}(1) - \frac{2c_1}{HCS}\pi_h(1\,|\,1)V_{h+1}^{\pi,\theta}(1), \qquad \text{(C.89)} \end{aligned}$$

where (i) arises from the basic fact $0 \le V_h^{\pi,\theta}(s) \le H - h + 1$ for any policy π and all $(s,h) \in \mathcal{S} \times [H]$, and (ii) holds since $\frac{2c_1}{HCS}(H-h) \le \frac{1}{2}$ for c_1 small enough. The above results lead to several immediate facts.

- If we choose π such that $\pi_h(0\,|\,1) = 1$ for all $h \in [H]$, then (C.88) tells us that

$$V_h^{\pi,\theta}(1) = \frac{1}{2} + V_{h+1}^{\pi,\theta}(1). \tag{C.90}$$

  A recursive application of this relation reveals that

$$V_h^{\pi,\theta}(1) = \frac{1}{2} + V_{h+1}^{\pi,\theta}(1) = \cdots = \sum_{j=h}^{H}\frac{1}{2} = \frac{1}{2}(H+1-h). \tag{C.91}$$

- For any policy π, applying (C.89) recursively tells us that

$$V_h^{\pi,\theta}(1) \le \frac{1}{2} + V_{h+1}^{\pi,\theta}(1) \le \cdots \le \sum_{j=h}^{H}\frac{1}{2} = \frac{1}{2}(H+1-h). \tag{C.92}$$

The above two facts taken collectively imply that the optimal policy and optimal value function obey

$$\pi_h^{\star,\theta}(0\,|\,1) = 1, \qquad V_h^{\star,\theta}(1) = \frac{1}{2}(H+1-h), \qquad \forall h \in [H]. \tag{C.93}$$

We then return to state 0. By taking π such that $\pi_h(\theta_h \,|\, 0) = 1$ (and hence $x_h^{\pi,\theta} = p$) for all $h \in [H]$,

one can invoke (C.86) to derive

$$\begin{aligned} V_h^{\pi,\theta}(0) &= 1 + \big(\mu(1)p + \mu(0)\big)V_{h+1}^{\pi,\theta}(0) + (1-p)\mu(1)V_{h+1}^{\pi,\theta}(1) \\ &\geq 1 + pV_{h+1}^{\pi,\theta}(0) \geq \sum_{j=0}^{H-h} p^j \geq \sum_{j=0}^{H-h} \Big(1 - \frac{c_1}{H}\Big)^j = \frac{1 - \big(1 - \frac{c_1}{H}\big)^{H-h+1}}{c_1/H} \\ &\geq \frac{2}{3}(H+1-h). \end{aligned} \tag{C.94}$$

To see that why the last inequality holds, it suffices to observe that

$$\Big(1 - \frac{c_1}{H}\Big)^{H-h+1} \leq \exp\Big(-\frac{c_1}{H}(H-h+1)\Big) \leq 1 - \frac{2c_1(H-h+1)}{3H},$$

as long as $c_1 \leq 0.5$, which follows due to the elementary inequalities $1 - x \leq \exp(-x)$ for any $x \geq 0$ and $\exp(-x) \leq 1 - 2x/3$ for any $0 \leq x \leq 1/2$. Combine (C.94) with (C.93) to reach

$$V_h^{\star,\theta}(0) \geq V_h^{\pi,\theta}(0) \geq \frac{2}{3}(H+1-h) > V_h^{\star,\theta}(1). \tag{C.95}$$

Moreover, it follows from (C.86) that

$$\begin{aligned} V_h^{\star,\theta}(0) &= 1 + \big(\mu(1)x_h^{\pi^\star,\theta,\theta} + \mu(0)\big)V_{h+1}^{\star,\theta}(0) + (1 - x_h^{\pi^\star,\theta,\theta})\mu(1)V_{h+1}^{\star,\theta}(1) \\ &= 1 + \mu(0)V_{h+1}^{\star,\theta}(0) + \mu(1)V_{h+1}^{\star,\theta}(1) + \mu(1)\big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\big)x_h^{\pi^\star,\theta,\theta}. \end{aligned} \tag{C.96}$$

Observing that the function

$$\mu(1)\big(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\big)x \tag{C.97}$$

is increasing in x (as a result of (C.95)) and that $x_h^{\pi,\theta}$ is increasing in $\pi_h(\theta_h \,|\, 0)$ (since $p \geq q$), we can readily conclude that the optimal policy in state 0 obeys

$$\pi_h^{\star,\theta}(\theta_h \,|\, 0) = 1, \qquad \text{for all } h \in [H]. \tag{C.98}$$

C.2.5.4 Proof of auxiliary properties

Throughout this subchapter, we shall suppress the dependency on θ in the notation $d_h^\star$ whenever it is clear from the context.

Proof of claim (E.69). For any MDP $\mathcal{M}_\theta$, from the definition of $d_h^{\mathsf{b}}(s,a)$ in (5.1) and the Markov property, it is clearly seen that

$$d_{h+1}^{\mathsf{b}}(s) = d_{h+1}^{\pi^{\mathsf{b}}}(s;\rho^{\mathsf{b}}) = \mathbb{P}(s_{h+1} = s \,|\, s_h \sim d_h^{\mathsf{b}}; \pi^{\mathsf{b}}), \qquad \forall (s,h) \in \mathcal{S} \times [H]. \tag{C.99}$$

Recalling that $d_1^{\mathsf{b}}(s) = \rho^{\mathsf{b}}(s) = \mu(s)$ for all $s \in \mathcal{S}$, one can then show that

$$\begin{aligned}
d_2^{\mathsf{b}}(0) &= \mathbb{P}\{s_2 = 0 \,|\, s_1 \sim d_1^{\mathsf{b}}; \pi^{\mathsf{b}}\} \\
&= \mu(0)\left[\pi_1^{\mathsf{b}}(\theta_1 \,|\, 0) P_1^{\theta_1}(0 \,|\, 0, \theta_1) + \pi_1^{\mathsf{b}}(1-\theta_1 \,|\, 0) P_1^{\theta_1}(0 \,|\, 0, 1-\theta_1)\right] \\
&\quad + \mu(1)\left[\pi_1^{\mathsf{b}}(0 \,|\, 1) P_1^{\theta_1}(0 \,|\, 1, 0) + \pi_1^{\mathsf{b}}(1 \,|\, 1) P_1^{\theta_1}(0 \,|\, 1, 1)\right] \\
&= \frac{\mu(0)}{2}\left[P_1^{\theta_1}(0 \,|\, 0, \theta_1) + P_1^{\theta_1}(0 \,|\, 0, 1-\theta_1)\right] + \frac{\mu(1)}{2}\left[P_1^{\theta_1}(0 \,|\, 1, 0) + P_1^{\theta_1}(0 \,|\, 1, 1)\right] \\
&= \frac{\mu(0)}{2}\left[(p+q) + (2-p-q)\mu(0)\right] + \frac{\mu(1)}{2}\mu(0)\frac{2c_1}{H} \\
&= \frac{\mu(0)}{2}\left[2 - \frac{2c_1}{H} + \frac{2c_1}{H}\mu(0)\right] + \frac{\mu(1)}{2}\mu(0)\frac{2c_1}{H} = \mu(0),
\end{aligned}$$

where the last inequality holds since $\mu(1) + \mu(0) = 1$. Similarly, it can be verified that $d_1^{\mathsf{b}}(1) = \mu(1)$, thereby implying that $d_2^{\mathsf{b}} = \mu$. Repeating this argument recursively for steps $h = 2, \ldots, H$ confirms that

$$d_h^{\mathsf{b}}(s) = \mu(s), \qquad \forall (s,h) \in \mathcal{S} \times [H]. \tag{C.100}$$

This further allows one to demonstrate that

$$d_h^{\mathsf{b}}(s,a) = d_h^{\mathsf{b}}(s)\pi_h^{\mathsf{b}}(a \,|\, s) = \mu(s)\pi_h^{\mathsf{b}}(a \,|\, s) = \mu(s)/2, \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{C.101}$$

Proof of claim (E.80)**.** Consider any MDP $\mathcal{M}_\theta$, for which we have shown in Lemma 63 that $\pi_h^{\star,\theta}(\theta_h \,|\, 0) = 1$ for all $h \in [H]$. It is observed that

$$\begin{aligned}
d_h^\star(0, \theta_h) &= d_h^\star(0)\pi_h^{\star,\theta}(\theta_h \,|\, 0) = d_h^\star(0) = \mathbb{P}\big\{s_h = 0 \,|\, s_{h-1} \sim d_{h-1}^\star; \pi^{\star,\theta}\big\} \\
&\geq d_{h-1}^\star(0)\pi_{h-1}^{\star,\theta}(\theta_{h-1} \,|\, 0) P_{h-1}^{\theta_{h-1}}(0 \,|\, 0, \theta_{h-1}) = d_{h-1}^\star(0) P_{h-1}^{\theta_{h-1}}(0 \,|\, 0, \theta_{h-1}) \\
&\geq \cdots \geq d_1^\star(0) \prod_{j=0}^{h-1} P_j^{\theta_j}(0 \,|\, 0, \theta_j) = \rho(0) \prod_{j=0}^{h-1} P_j^{\theta_j}(0 \,|\, 0, \theta_j) \\
&\geq \rho(0) \prod_{j=0}^{h-1} p \geq \left(1 - \frac{c_1}{H}\right)^H > \frac{1}{2},
\end{aligned} \tag{C.102}$$

where the last line makes use of the properties $p \geq 1 - c_1/H$, $\rho(0) = 1$, and

$$\left(1 - \frac{c_1}{H}\right)^H \geq \left(1 - \frac{1}{2H}\right)^H > \frac{1}{2}$$

provided that $0 < c_1 < 1/2$. Combining this with (E.69), we arrive at

$$\begin{aligned}
\max_{h\in[H]} \frac{\min\left\{d_h^\star(0,\theta_h), \frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,\theta_h)} &= \frac{2}{S\mu(0)} = 2C, \\
\max_{h\in[H]} \frac{\min\left\{d_h^\star(0,1-\theta_h), \frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,1-\theta_h)} &= \max_{h\in[H]} \frac{\min\left\{d_h^\star(0)\pi_h^\star(1-\theta_h\,|\,0), \frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,1-\theta_h)} = 0, \\
\max_{a\in\{0,1\},h\in[H]} \frac{\min\left\{d_h^\star(1,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b}}(1,a)} &\overset{\text{(i)}}{\le} \frac{1/S}{\mu(1)/2} \overset{\text{(ii)}}{=} \frac{2}{S\left(1-\frac{1}{SC}\right)} \le \frac{4}{S} \le 2C,
\end{aligned}$$

where (i) arises from (E.69), (ii) relies on the definition in (E.67), and the final two inequalities come from the assumption in (C.57). Taking this together with the straightforward condition $d_h^\star(s) = 0$ $(s > 1)$ yields

$$C_{\mathsf{clipped}}^\star = \max_{h\in[H]} \frac{\min\left\{d_h^\star(0,\theta_h), \frac{1}{S}\right\}}{d_h^{\mathsf{b}}(0,\theta_h)} = 2C. \tag{C.103}$$

Proof of inequality (E.83). By virtue of (E.77) and (E.79), we see that $x_h^{\pi^\star,\theta,\theta} = p$ for all $h \in [H]$, which combined with (E.78) gives

$$\begin{aligned}
\left\langle \rho, V_h^{\star,\theta} - V_h^{\pi,\theta} \right\rangle &= V_h^{\star,\theta}(0) - V_h^{\pi,\theta}(0) \\
&= \big(\mu(1)p + \mu(0)\big)V_{h+1}^{\star,\theta}(0) + (1-p)\mu(1)V_{h+1}^{\star,\theta}(1) \\
&\quad - \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big)V_{h+1}^{\pi,\theta}(0) - (1-x_h^{\pi,\theta})\mu(1)V_{h+1}^{\pi,\theta}(1) \\
&\overset{\text{(i)}}{\ge} \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big)\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\right) + \mu(1)\big(p - x_h^{\pi,\theta}\big)V_{h+1}^{\star,\theta}(0) \\
&\quad + (1-p)\mu(1)V_{h+1}^{\star,\theta}(1) - \big(1-x_h^{\pi,\theta}\big)\mu(1)V_{h+1}^{\star,\theta}(1) \\
&= \big(\mu(1)x_h^{\pi,\theta} + \mu(0)\big)\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\right) + \big(p - x_h^{\pi,\theta}\big)\mu(1)\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\right) \\
&\overset{\text{(ii)}}{\ge} q\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\right) + \big(p - x_h^{\pi,\theta}\big)\mu(1)\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\right) \\
&\overset{\text{(iii)}}{\ge} q\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\right) + \frac{3}{8}(p-q)\left\|\pi_h^{\star,\theta}(0) - \pi_h(0)\right\|_1\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\star,\theta}(1)\right) \\
&\overset{\text{(iv)}}{\ge} q\left(V_{h+1}^{\star,\theta}(0) - V_{h+1}^{\pi,\theta}(0)\right) + \frac{c_2\varepsilon}{8H^2}(H+1-h)\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1,
\end{aligned} \tag{C.104}$$

where (i) holds since $V_{h+1}^{\pi,\theta}(1) \le V_{h+1}^{\star,\theta}(1)$, (ii) follows from the fact that $x_h^\pi \ge q$ for any π and $h \in [H]$, and (iv) arises from the facts (E.79) and the choice (E.61) of (p, q). To see why (iii) is valid, it suffices to note that $\mu(1) = 1 - \frac{1}{CS} \ge \frac{3}{4}$ (as a consequence of (E.67) and (C.57)) and

$$p - x_h^{\pi,\theta} = (p-q)\big(1-\pi_h(\theta_h\,|\,0)\big) = \frac{1}{2}(p-q)\big(1-\pi_h(\theta_h\,|\,0)+\pi_h(1-\theta_h\,|\,0)\big) = \frac{1}{2}(p-q)\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1.$$

To continue, under the condition

$$\sum_{h=1}^{H} \left\|\pi_h(\cdot\,|\,0) - \pi_h^{\star,\theta}(\cdot\,|\,0)\right\|_1 \geq \frac{H}{8}, \tag{C.105}$$

applying the relation in (C.104) recursively yields

$$\begin{aligned}
V_1^{\star,\theta}(0) - V_1^{\pi,\theta}(0) &\geq \sum_{h=1}^{H} q^{h-1} \frac{c_2\varepsilon}{8H^2}(H+1-h)\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1 \\
&= \sum_{h=1}^{H} \left(1 - \frac{c_1}{H} - \frac{c_2\varepsilon}{H^2}\right)^{h-1} \frac{c_2\varepsilon}{8H^2}(H+1-h)\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1 \\
&\overset{\text{(i)}}{>} \frac{c_2\varepsilon}{16H^2} \sum_{h=1}^{H} (H+1-h)\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1 \\
&= \frac{c_2\varepsilon}{16H^2} \sum_{h=1}^{H} h\left\|\pi_{H+1-h}^{\star,\theta}(\cdot\,|\,0) - \pi_{H+1-h}(\cdot\,|\,0)\right\|_1 \\
&\overset{\text{(ii)}}{\geq} \frac{c_2\varepsilon}{16H^2} \sum_{h=1}^{\lfloor H/16\rfloor} 2h = \frac{c_2\varepsilon}{8H^2}\left\lfloor\frac{H}{16}\right\rfloor\left(\left\lfloor\frac{H}{16}\right\rfloor + 1\right).
\end{aligned} \tag{C.106}$$

Here, (i) follows since

$$\left(1 - \frac{c_1}{H} - \frac{c_2\varepsilon}{H^2}\right)^{h-1} \geq \left(1 - \frac{2c_1}{H}\right)^{H} > \frac{1}{2}, \qquad \text{for all } h \in [H]$$

holds as long as $0 < c_1 \leq 1/4$ and $c_2\varepsilon/H \leq c_1$. To see why (ii) is valid, we note that for any $0 \leq x_1, \cdots, x_H \leq x_{\max}$ obeying $\sum_{i=1}^{H} x_i \geq x_{\text{sum}}$, the following elementary inequality holds:

$$\sum_{i=1}^{H} x_i a_i \geq \sum_{i=1}^{\lfloor x_{\text{sum}}/x_{\max}\rfloor} x_{\max} a_i;$$

this together with $\left\|\pi_h^{\star,\theta}(\cdot\,|\,0) - \pi_h(\cdot\,|\,0)\right\|_1 \leq 2$ and (C.105) reveals that (by taking $a_h = h$ and $x_h = \left\|\pi_{H+1-h}^{\star,\theta}(\cdot\,|\,0) - \pi_{H+1-h}(\cdot\,|\,0)\right\|_1$)

$$\sum_{h=1}^{H} h\left\|\pi_{H+1-h}^{\star,\theta}(\cdot\,|\,0) - \pi_{H+1-h}(\cdot\,|\,0)\right\|_1 \geq \sum_{h=1}^{\lfloor H/16\rfloor} 2h,$$

thus validating inequality (ii). As a result, we can continue the derivation to obtain

$$\text{(C.106)} \geq \frac{c_2\varepsilon}{8H^2} \frac{\frac{H}{16}\left(\frac{H}{16}+1\right)}{2} > \varepsilon, \tag{C.107}$$

provided that $c_2 \geq 4096$.

C.3 Proof of minimax lower bounds

C.3.1 Preliminary facts

For any two distributions P and Q, we denote by $\mathsf{KL}(P \parallel Q)$ the Kullback-Leibler (KL) divergence of P and Q. Letting $\mathsf{Ber}(p)$ be the Bernoulli distribution with mean p, we also introduce

$$\mathsf{KL}(p \parallel q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \qquad \text{and} \qquad \chi^2(p \parallel q) := \frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q}, \tag{C.108}$$

which represent respectively the KL divergence and the chi-square divergence of $\mathsf{Ber}(p)$ from $\mathsf{Ber}(q)$ (Tsybakov, 2009). We make note of the following useful properties about the KL divergence.

Lemma 33. *For any $p, q \in \left[\frac{1}{2}, 1\right)$ and $p > q$, it holds that*

$$\mathsf{KL}(p \parallel q) \leq \mathsf{KL}(q \parallel p) \leq \chi^2(q \parallel p) = \frac{(p-q)^2}{p(1-p)}. \tag{C.109}$$

Proof. The second inequality in (E.11) is a well-known relation between KL divergence and chi-square divergence; see Tsybakov (2009, Lemma 2.7). As a result, it suffices to justify the first inequality. Towards this end, let us introduce $a = \frac{p+q}{2} \in \left[\frac{1}{2}, 1\right]$ and $b = \frac{p-q}{2} \in \left[0, \frac{1}{4}\right]$, which allow us to re-parameterize (p, q) as $p = a + b$ and $q = a - b$. The definition (E.10) together with a little algebra gives

$$\begin{aligned} \mathsf{KL}(p \parallel q) - \mathsf{KL}(q \parallel p) &= (p+q) \log \frac{p}{q} + (2-p-q) \log \frac{1-p}{1-q} \\ &= 2a \log \left(\frac{a+b}{a-b} \right) + 2(1-a) \log \frac{1-a-b}{1-a+b} =: g(a, b). \end{aligned}$$

Taking the derivative w.r.t. b yields

$$\frac{\partial g(a,b)}{\partial b} = 2a \left\{ \frac{1}{a+b} + \frac{1}{a-b} \right\} - 2(1-a) \left\{ \frac{1}{1-a+b} + \frac{1}{1-a-b} \right\} = f(a) - f(1-a) \leq 0,$$

with $f(x) := \frac{2x}{x+b} + \frac{2x}{x-b}$ (for $x > b$). Here, the last inequality follows since $f(\cdot)$ is a decreasing function and that $a \geq 1 - a$. This implies that $g(a, b)$ is non-increasing in $b \geq 0$ for any given a, which in turn leads to

$$\mathsf{KL}(p \parallel q) - \mathsf{KL}(q \parallel p) = g(a, b) \leq g(a, 0) = 0$$

as claimed. □

C.3.2 Proof of Theorem 7

We now construct some hard problem instances and use them to establish the minimax lower bounds claimed in Theorem 7. It is assumed throughout this subchapter that

$$\frac{2}{3} \leq \gamma < 1 \qquad \text{and} \qquad \frac{14(1-\gamma)\varepsilon}{\gamma} \leq \frac{1}{2}. \tag{C.110}$$

C.3.2.1 Construction of hard problem instances

Construction of the hard MDPs. Let us introduce two MDPs $\{\mathcal{M}_\theta = (\mathcal{S}, \mathcal{A}, P_\theta, r, \gamma) \mid \theta \in \{0,1\}\}$ parameterized by θ, which involve S states and 2 actions as follows:

$$\mathcal{S} = \{0, 1, \ldots, S-1\} \qquad \text{and} \qquad \mathcal{A} = \{0, 1\}.$$

We single out a crucial state distribution (supported on the state subset $\{0,1\}$) as follows:

$$\mu(s) = \frac{1}{CS}\mathbb{1}\{s=0\} + \Big(1 - \frac{1}{CS}\Big)\mathbb{1}\{s=1\} \tag{C.111}$$

for some quantity $C > 0$ obeying

$$\frac{1}{CS} \leq \frac{1}{4\gamma}. \tag{C.112}$$

We shall make clear the relation between C and the concentrability coefficient $C^\star_{\mathsf{clipped}}$ shortly (see (C.122)). Armed with this distribution, we are ready to define the transition kernel P_θ of the MDP $\mathcal{M}_\theta$ as follows:

$$P_\theta(s' \,|\, s, a) = \begin{cases} p\mathbb{1}\{s'=0\} + (1-p)\mu(s') & \text{for} \quad (s,a) = (0,\theta), \\ q\mathbb{1}\{s'=0\} + (1-q)\mu(s') & \text{for} \quad (s,a) = (0,1-\theta), \\ \mathbb{1}\{s'=1\} & \text{for} \quad (s,a) = (1,0), \\ (2\gamma-1)\mathbb{1}\{s'=1\} + 2(1-\gamma)\mu(s') & \text{for} \quad (s,a) = (1,1), \\ \gamma\mathbb{1}(s'=s) + (1-\gamma)\mu(s') & \text{for} \quad s > 1, \end{cases} \tag{C.113}$$

where the parameters p and q are chosen to be

$$p = \gamma + \frac{14(1-\gamma)^2\varepsilon}{\gamma}, \qquad q = \gamma - \frac{14(1-\gamma)^2\varepsilon}{\gamma}. \tag{C.114}$$

In view of the assumptions (C.110), one has

$$p > q \geq \gamma - \frac{1-\gamma}{2} \geq \frac{1}{2}. \tag{C.115}$$

As can be clearly seen from the construction, if the MDP is initialized to either state 0 or state 1, then it will never leave the state subset $\{0, 1\}$. In addition, the reward function for any MDP $\mathcal{M}_\theta$ is chosen to be

$$r(s,a) = \begin{cases} 1 & \text{for} \quad s = 0, \\ \frac{1}{2} & \text{for} \quad (s,a) = (1,0), \\ 0 & \text{for} \quad (s,a) = (1,1), \\ 0 & \text{for} \quad s > 1, \end{cases} \tag{C.116}$$

where the reward gained in state 0 is clearly higher than that in other states.

Value functions and optimal policies. Next, let us take a moment to compute the value functions of the constructed MDPs and identify the optimal policies. For notational clarity, for the MDP $\mathcal{M}_\theta$ with $\theta \in \{0, 1\}$, we denote by $\pi_\theta^\star$ the optimal policy, and let V_θ^π (resp. $V_\theta^\star$) represent the value function of policy π (resp. $\pi_\theta^\star$). The lemma below collects several useful properties about the value functions and the optimal policies; the proof is deferred to Appendix C.3.2.3.

Lemma 34. *Consider any $\theta \in \{0, 1\}$ and any policy π. One has*

$$V_\theta^\pi(0) = \frac{1 + \gamma(1 - x_{\pi,\theta})\mu(1)V_\theta^\pi(1)}{1 - \gamma\big(\mu(1)x_{\pi,\theta} + \mu(0)\big)} = V_\theta^\pi(1) + \frac{1 - (1-\gamma)V_\theta^\pi(1)}{1 - \gamma\mu(0) - \gamma\mu(1)x_{\pi,\theta}}, \tag{C.117}$$

where we define

$$x_{\pi,\theta} \coloneqq p\pi(\theta \,|\, 0) + q\pi(1 - \theta \,|\, 0). \tag{C.118}$$

In addition, the optimal policy $\pi_\theta^\star$ and the optimal value function obey

$$\pi_\theta^\star(\theta \,|\, 0) = 1, \qquad \pi_\theta^\star(0 \,|\, 1) = 1, \qquad \text{and} \qquad V_\theta^\star(1) = \frac{1}{2(1-\gamma)}. \tag{C.119}$$

Construction of the batch dataset. Given any constructed MDP $\mathcal{M}_\theta$, we generate a dataset containing N i.i.d. samples $\{(s_i, a_i, s_i')\}_{1 \le i \le N}$ according to (5.22), where the initial state distribution ρ^{b} and behavior policy π^{b} are chosen to be:

$$\rho^{\mathsf{b}}(s) = \mu(s) \qquad \text{and} \qquad \pi^{\mathsf{b}}(a \,|\, s) = 1/2, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

with μ denoting the distribution defined in (E.194). Interestingly, the occupancy state distribution of this dataset coincides with μ, in the sense that

$$d^{\mathsf{b}}(s) = \mu(s) \qquad \text{and} \qquad d^{\mathsf{b}}(s,a) = \mu(s)/2, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{C.120}$$

Moreover, letting us choose the test distribution ρ in a way that

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{if } s > 0. \end{cases} \tag{C.121}$$

we can also characterize the single-policy clipped concentrability coefficient $C^\star_{\mathsf{clipped}}$ of the dataset w.r.t. the constructed MDP $\mathcal{M}_\theta$ as follows

$$C^\star_{\mathsf{clipped}} = 2C. \tag{C.122}$$

The proof of the claims (C.120) and (C.122) can be found in Appendix C.3.2.4.

C.3.2.2 Establishing the minimax lower bound

Equipped with the above construction, we are ready to develop our lower bounds. We remind the reader of the test distribution ρ chosen in (C.121), and hence we need to control $\langle \rho, V_\theta^\star - V_\theta^{\widehat{\pi}} \rangle = V_\theta^\star(0) - V_\theta^{\widehat{\pi}}(0)$ with $\widehat{\pi}$ representing a policy estimate (computed based on the batch dataset).

Step 1: converting $\widehat{\pi}$ into an estimate $\widehat{\theta}$ of θ. Consider first an arbitrary policy π. By combining the definition (E.205) with the properties (C.119), we see that $x_{\pi^\star_\theta, \theta} = p$, which together with (C.117) gives

$$\begin{aligned} \langle \rho, V_\theta^\star - V_\theta^\pi \rangle = V_\theta^\star(0) - V_\theta^\pi(0) &= \frac{1 + \gamma(1-p)\mu(1)V_\theta^\star(1)}{1 - \gamma\big(\mu(1)p + \mu(0)\big)} - \frac{1 + \gamma(1 - x_{\pi,\theta})\mu(1)V_\theta^\pi(1)}{1 - \gamma\big(\mu(1)x_{\pi,\theta} + \mu(0)\big)} \\ &\geq \frac{1 + \gamma(1-p)\mu(1)V_\theta^\star(1)}{1 - \gamma\big(\mu(1)p + \mu(0)\big)} - \frac{1 + \gamma(1 - x_{\pi,\theta})\mu(1)V_\theta^\star(1)}{1 - \gamma\big(\mu(1)x_{\pi,\theta} + \mu(0)\big)} \\ &\geq \frac{21\varepsilon}{8}\big(1 - \pi(\theta \,|\, 0)\big). \end{aligned} \tag{C.123}$$

Here, the second line holds since $V_\theta^\pi \leq V_\theta^\star$, and the last inequality will be established in Appendix C.3.2.4.

Denoting by $\mathbb{P}_\theta$ the probability distribution when the MDP is $\mathcal{M}_\theta$, suppose for the moment that the policy estimate $\widehat{\pi}$ achieves

$$\mathbb{P}_\theta\big\{\big\langle \rho, V_\theta^\star - V_\theta^{\widehat{\pi}} \big\rangle \leq \varepsilon\big\} \geq \frac{7}{8},$$

then in view of (C.123), one necessarily has $\widehat{\pi}(\theta \,|\, 0) \geq \frac{13}{21}$ with probability at least $7/8$. If this were true, then we could then construct the following estimate $\widehat{\theta}$ for θ:

$$\widehat{\theta} = \arg\max_a \widehat{\pi}(a \,|\, 0), \tag{C.124}$$

which would necessarily satisfy

$$\mathbb{P}_\theta\big(\widehat{\theta} = \theta\big) \geq \mathbb{P}_\theta\big\{\widehat{\pi}(\theta \,|\, 0) > 1/2\big\} \geq \mathbb{P}_\theta\bigg\{\widehat{\pi}(\theta \,|\, 0) \geq \frac{13}{21}\bigg\} \geq \frac{7}{8}. \tag{C.125}$$

In what follows, we would like to show that (E.86) cannot happen — i.e., one cannot possibly find such a good estimator for θ — without a sufficient number of samples.

Step 2: probability of error in testing two hypotheses. The next step lies in studying the feasibility of differentiating two hypotheses $\theta = 0$ and $\theta = 1$. Define the minimax probability of error as follows

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max\big\{\mathbb{P}_0(\psi \neq 0),\, \mathbb{P}_1(\psi \neq 1)\big\}, \tag{C.126}$$

where the infimum is taken over all possible tests ψ (based on the batch dataset in hand). Letting μ_θ^{b} denote the distribution of a sample (s_i, a_i, s_i') under the MDP $\mathcal{M}_\theta$ and recalling that the samples are independently generated, one can demonstrate that

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{4}\exp\Big(-N\mathsf{KL}\big(\mu_0^{\mathsf{b}} \,\|\, \mu_1^{\mathsf{b}}\big)\Big) \\ &= \frac{1}{4}\exp\bigg\{-\frac{1}{2}N\mu(0)\Big(\mathsf{KL}\big(P_0(\cdot \,|\, 0,0) \,\|\, P_1(\cdot \,|\, 0,0)\big) + \mathsf{KL}\big(P_0(\cdot \,|\, 0,1) \,\|\, P_1(\cdot \,|\, 0,1)\big)\Big)\bigg\}. \end{aligned} \tag{C.127}$$

Here, the first inequality results from Tsybakov (2009, Theorem 2.2) and the additivity property of the KL divergence (cf. Tsybakov (2009, Page 85)), and the second line holds true since

$$\begin{aligned} \mathsf{KL}(\mu_0^{\mathsf{b}} \,\|\, \mu_1^{\mathsf{b}}) &= \sum_{s,a,s'} \mu(s)\pi^{\mathsf{b}}(a \,|\, s)P_0(s' \,|\, s,a) \log \frac{\mu(s)\pi^{\mathsf{b}}(a \,|\, s)P_0(s' \,|\, s,a)}{\mu(s)\pi^{\mathsf{b}}(a \,|\, s)P_1(s' \,|\, s,a)} \\ &= \frac{1}{2}\mu(0)\sum_a \sum_{s'} P_0(s' \,|\, 0,a) \log \frac{P_0(s' \,|\, 0,a)}{P_1(s' \,|\, 0,a)} \\ &= \frac{1}{2}\mu(0)\sum_a \mathsf{KL}\big(P_0(\cdot \,|\, 0,a) \,\|\, P_1(\cdot \,|\, 0,a)\big), \end{aligned}$$

where the second line is valid since $P_0(\cdot \,|\, s,a)$ and $P_1(\cdot \,|\, s,a)$ differ only when $s = 0$.

Next, we turn attention to the KL divergence of interest. Recall that

$$P_0(0 \,|\, 0,0) = \bigg(1 - \frac{1}{CS}\bigg)p + \frac{1}{CS}, \qquad P_1(0 \,|\, 0,0) = \bigg(1 - \frac{1}{CS}\bigg)q + \frac{1}{CS}.$$

Given that $p \geq q \geq 1/2$ (see (C.115)), we can apply Lemma 60 to arrive at

$$\mathsf{KL}\big(P_0(\cdot \,|\, 0,0) \,\|\, P_1(\cdot \,|\, 0,0)\big) = \mathsf{KL}\left(\bigg(1 - \frac{1}{CS}\bigg)p + \frac{1}{CS} \,\|\, \bigg(1 - \frac{1}{CS}\bigg)q + \frac{1}{CS}\right)$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \frac{\big(1-\frac{1}{CS}\big)^2 (p-q)^2}{\big(\big(1-\frac{1}{CS}\big)p+\frac{1}{CS}\big)\big(1-p-(1-p)\frac{1}{CS}\big)} \\
&\leq \frac{\big(1-\frac{1}{CS}\big)^2 (p-q)^2}{p\big((1-p)\big(1-\frac{1}{CS}\big)\big)} \\
&\overset{\text{(ii)}}{=} \frac{784(1-\gamma)^4\varepsilon^2}{\gamma^2\Big(\gamma+\frac{14(1-\gamma)^2\varepsilon}{\gamma}\Big)\Big(1-\gamma-\frac{14(1-\gamma)^2\varepsilon}{\gamma}\Big)} \\
&\overset{\text{(iii)}}{\leq} \frac{1568(1-\gamma)^4\varepsilon^2}{\gamma^3(1-\gamma)} \overset{\text{(iv)}}{\leq} 12544(1-\gamma)^3\varepsilon^2,
\end{aligned}
$$

where (i) arises from Lemma 60, (ii) follows from the definitions of p and q (C.114), (iii) holds true as long as $\frac{14(1-\gamma)^2\varepsilon}{\gamma} \leq \frac{1-\gamma}{2}$, and (iv) results from the assumption $\gamma \in [\frac{1}{2}, 1)$. Evidently, the same upper bound holds for $\mathsf{KL}\big(P_0(\cdot\,|\,0,1) \,\|\, P_1(\cdot\,|\,0,1)\big)$ as well. Substitution back into (C.127) reveals that: if the sample size does not exceed

$$
N \leq \frac{CS\log 2}{12544(1-\gamma)^3\varepsilon^2} = \frac{C^\star_{\mathsf{clipped}} S\log 2}{25088(1-\gamma)^3\varepsilon^2}, \tag{C.128}
$$

then one necessarily has

$$
p_{\mathrm{e}} \geq \frac{1}{4}\exp\Big(-12544N\mu(0)(1-\gamma)^3\varepsilon^2\Big) = \frac{1}{4}\exp\Big(-\frac{12544N(1-\gamma)^3\varepsilon^2}{CS}\Big) \geq \frac{1}{8}. \tag{C.129}
$$

Step 3: putting all this together. To finish up, suppose that there exists an estimator $\widehat{\pi}$ such that

$$
\mathbb{P}_0\big\{\big\langle \rho, V_0^\star - V_0^{\widehat{\pi}}\big\rangle > \varepsilon\big\} < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big\{\big\langle \rho, V_0^\star - V_0^{\widehat{\pi}}\big\rangle > \varepsilon\big\} < \frac{1}{8}.
$$

Then in view of our arguments in Step 1, the estimator $\widehat{\theta}$ defined in (E.85) must satisfy

$$
\mathbb{P}_0\big(\widehat{\theta} \neq \theta\big) < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big(\widehat{\theta} \neq \theta\big) < \frac{1}{8}.
$$

This, however, cannot possibly happen under the sample size condition (C.128); otherwise it contradicts the lower bound (E.91).

C.3.2.3 Proof of Lemma 34

To begin with, for any policy π, the value function of state 0 obeys

$$
\begin{aligned}
V_\theta^\pi(0) &= \mathop{\mathbb{E}}_{a\sim\pi(\cdot\,|\,0)}\Big[r(0,a) + \gamma\sum_{s'} P_\theta(s'\,|\,0,a)V_\theta^\pi(s')\Big] \\
&= 1 + \gamma\pi(\theta\,|\,0)\Big[\big(p+(1-p)\mu(0)\big)V_\theta^\pi(0) + (1-p)\mu(1)V_\theta^\pi(1)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \gamma\pi(1-\theta\,|\,0)\Big[\big(q+(1-q)\mu(0)\big)V_\theta^\pi(0) + (1-q)\mu(1)V_\theta^\pi(1)\Big] \\
&= 1 + \gamma\Big[p\pi(\theta\,|\,0) + q\pi(1-\theta\,|\,0) + \mu(0) - p\pi(\theta\,|\,0)\mu(0) - q\pi(1-\theta\,|\,0)\mu(0)\Big]V_\theta^\pi(0) \\
&\quad + \gamma\mu(1)\Big[1 - p\pi(\theta\,|\,0) - q\pi(1-\theta\,|\,0)\Big]V_\theta^\pi(1) \\
&\overset{\text{(i)}}{=} 1 + \gamma\Big[x_{\pi,\theta} + (1-x_{\pi,\theta})\mu(0)V_\theta^\pi(0) + (1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)\Big] \\
&\overset{\text{(ii)}}{=} 1 + \gamma\Big[\big(\mu(1)x_{\pi,\theta} + \mu(0)\big)V_\theta^\pi(0) + (1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)\Big], \qquad \text{(C.130)}
\end{aligned}
$$

where in (i) we have defined the following quantity

$$
x_{\pi,\theta} = p\pi(\theta\,|\,0) + q\pi(1-\theta\,|\,0) = q + (p-q)\pi(\theta\,|\,0), \qquad \text{(C.131)}
$$

and (ii) relies on the fact that $\mu(0)+\mu(1)=1$. Rearranging terms in (C.130), we are left with

$$
V_\theta^\pi(0) = \frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\pi(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)} = V_\theta^\pi(1) + \frac{1-(1-\gamma)V_\theta^\pi(1)}{1-\gamma\mu(0)-\gamma\mu(1)x_{\pi,\theta}}. \qquad \text{(C.132)}
$$

Additionally, the value function of state 1 can be calculated as

$$
\begin{aligned}
V_\theta^\pi(1) &= \pi(0\,|\,1)\left(\frac{1}{2} + \gamma V_\theta^\pi(1)\right) + \pi(1\,|\,1)\gamma\Big[\big((2\gamma-1) + 2(1-\gamma)\mu(1)\big)V_\theta^\pi(1) + 2(1-\gamma)\mu(0)V_\theta^\pi(0)\Big] \\
&= \pi(0\,|\,1)\left(\frac{1}{2} + \gamma V_\theta^\pi(1)\right) + \pi(1\,|\,1)\gamma\left[\left(1 - \frac{2(1-\gamma)}{CS}\right)V_\theta^\pi(1) + \frac{2(1-\gamma)}{CS}V_\theta^\pi(0)\right] \qquad \text{(C.133)} \\
&\overset{\text{(i)}}{\le} \pi(0\,|\,1)\left(\frac{1}{2} + \gamma V_\theta^\pi(1)\right) + \pi(1\,|\,1)\gamma\left[\left(1 - \frac{2(1-\gamma)}{CS}\right)V_\theta^\pi(1) + \frac{2(1-\gamma)}{CS}\frac{1}{1-\gamma}\right] \\
&\overset{\text{(ii)}}{\le} \pi(0\,|\,1)\left(\frac{1}{2} + \gamma V_\theta^\pi(1)\right) + \pi(1\,|\,1)\left[\frac{1}{2} + \gamma\left(1 - \frac{2(1-\gamma)}{CS}\right)V_\theta^\pi(1)\right] \\
&= \frac{1}{2} + \gamma V_\theta^\pi(1) - \frac{2\gamma(1-\gamma)}{CS}V_\theta^\pi(1)\pi(1\,|\,1), \qquad \text{(C.134)}
\end{aligned}
$$

where (i) arises from the elementary property $0 \le V_\theta^\pi(s) \le \frac{1}{1-\gamma}$ for any π and $s \in \mathcal{S}$, and (ii) comes from the assumption (C.112). The above observation reveals several facts:

- If we take $\pi(0\,|\,1) = 1$, then (C.133) tells us that

$$
V_\theta^\pi(1) = \frac{1}{2} + \gamma V_\theta^\pi(1) \qquad \Longrightarrow \qquad V_\theta^\pi(1) = \frac{1}{2(1-\gamma)}. \qquad \text{(C.135)}
$$

- It also follows from (C.134) that for any policy π, one has

$$
V_\theta^\pi(1) \le \frac{1}{2} + \gamma V_\theta^\pi(1) \qquad \Longrightarrow \qquad V_\theta^\pi(1) \le \frac{1}{2(1-\gamma)}. \qquad \text{(C.136)}
$$

These two facts taken collectively imply that the optimal policy and the optimal value function obey

$$\pi_\theta^\star(0\,|\,1) = 1 \qquad \text{and} \qquad V_\theta^\star(1) = \frac{1}{2(1-\gamma)}. \tag{C.137}$$

Next, we have learned from (C.132) that

$$V_\theta^\star(0) = V_\theta^\star(1) + \frac{1-(1-\gamma)V_\theta^\star(1)}{1-\gamma\mu(0)-\gamma\mu(1)x_{\pi_\theta^\star,\theta}}.$$

Note that $1-(1-\gamma)V_\theta^\star(1) \geq 1-(1-\gamma)\frac{1}{1-\gamma} = 0$. Since the function

$$g(x) = V_\theta^\star(1) + \frac{1-(1-\gamma)V_\theta^\star(1)}{1-\gamma\mu(0)-\gamma\mu(1)x}$$

is increasing in x and that $x_{\pi,\theta}$ (cf. (C.131)) is increasing in $\pi(\theta\,|\,0)$ (given that $p \geq q$), one can easily see that the optimal policy obeys

$$\pi_\theta^\star(\theta\,|\,0) = 1. \tag{C.138}$$

C.3.2.4 Proof of auxiliary properties

Proof of claim (C.120). We begin by proving the property (C.120). Towards this, let us abuse the notation by considering a MDP trajectory denoted by $\{(s_t, a_t)\}_{t\geq 0}$, and suppose that it starts from $s_0 \sim \rho^{\mathsf{b}} = \mu$. It can be straightforwardly calculated that

$$\begin{aligned}
\mathbb{P}\{s_1 = 0\} &= \sum\nolimits_s \mu(s)\Big\{\pi^{\mathsf{b}}(0\,|\,s)\mathbb{P}\{s_1 = 0 \mid s_0 = s, a_0 = 0\} + \pi^{\mathsf{b}}(1\,|\,s)\mathbb{P}\{s_1 = 0 \mid s_0 = s, a_0 = 1\}\Big\} \\
&= \mu(0)\left\{\frac{1}{2}P_\theta(0\,|\,0,0) + \frac{1}{2}P_\theta(0\,|\,0,1)\right\} + \mu(1)\left\{\frac{1}{2}P_\theta(0\,|\,1,0) + \frac{1}{2}P_\theta(0\,|\,1,1)\right\} \\
&= \mu(0)\{\gamma + (1-\gamma)\mu(0)\} + \mu(1)\{(1-\gamma)\mu(0)\} = \mu(0),
\end{aligned}$$

where the last identity holds since $\mu(0) + \mu(1) = 1$. Similarly, one can derive $\mathbb{P}\{s_1 = 1\} = \mu(1)$, thus indicating that $s_1 \sim \mu$. Repeating this analysis reveals that $s_t \sim \mu$ for any $t \geq 0$. Consequently, one has

$$d^{\mathsf{b}}(s) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\big(s_t = s \mid s_0 \sim \rho^{\mathsf{b}}; \pi^{\mathsf{b}}\big)\right] = \mu(s), \qquad \forall s \in \mathcal{S}.$$

Additionally, it it observed that

$$d^{\mathsf{b}}(s,a) = d^{\mathsf{b}}(s)\pi^{\mathsf{b}}(a\,|\,s) = \mu(s)/2. \tag{C.139}$$

Proof of claim (C.122). Consider the MDP $\mathcal{M}_\theta$, whose optimal policy $\pi_\theta^\star$ satisfies $\pi_\theta^\star(\theta\,|\,0) = 1$ (see Lemma 34). Let us generate a MDP trajectory denoted by $\{(s_t, a_t)\}_{t\geq 0}$ with $a_t \sim \pi_\theta^\star(\cdot\,|\,s_t)$,

where we have again abused notation as long as it is clear from the context. In this case, we can deduce that

$$d^\star(0,\theta) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\big(s_t=0 \mid s_0\sim\rho;\pi_\theta^\star\big)\pi_\theta^\star(\theta\,|\,0)\right] = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\big(s_t=0 \mid s_0\sim\rho;\pi_\theta^\star\big)\right]$$
$$\overset{\text{(i)}}{\geq}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\rho(0)\big[\mathbb{P}_\theta(0\,|\,0,\theta)\big]^t \overset{\text{(ii)}}{\geq}(1-\gamma)\sum_{t=0}^{\infty}\rho(0)\gamma^{2t} = \frac{1-\gamma}{1-\gamma^2} = \frac{1}{1+\gamma} \geq \frac{1}{2},$$

where in (i) we compute, for each t, the probability of a special trajectory with $s_1=\cdots=s_t=0$ and $a_0=\cdots=a_{t-1}=\theta$, and (ii) holds true since $P_\theta(0\,|\,0,\theta)\geq p\geq\gamma$. Taking this together with (C.139) yields

$$\begin{aligned}\frac{\min\big\{d^\star(0,\theta),\frac{1}{S}\big\}}{d^{\mathsf{b}}(0,\theta)} &= \frac{2}{S\mu(0)} = 2C,\\ \frac{\min\big\{d^\star(0,1-\theta),\frac{1}{S}\big\}}{d^{\mathsf{b}}(0,1-\theta)} &= \frac{\min\big\{d^\star(0,1-\theta),\frac{1}{S}\big\}}{d^{\mathsf{b}}(0,\theta)} \leq \frac{\min\big\{d^\star(0,\theta),\frac{1}{S}\big\}}{d^{\mathsf{b}}(0,\theta)} = 2C.\end{aligned} \tag{C.140}$$

In addition, it is easily seen that $d^\star(s,a)=0$ for any $s>1$, and that

$$\frac{\min\big\{d^\star(1,a),\frac{1}{S}\big\}}{d^{\mathsf{b}}(1,a)} \leq \frac{1/S}{\mu(1)/2} = \frac{2}{S(1-1/CS)} \leq \frac{4}{S} \leq 2C,$$

where the first inequality comes from (C.139), the first identity uses the definition (E.194), and the last two inequalities result from an immediate consequence of (C.112) and $\gamma\geq 1/2$, i.e.,

$$\frac{1}{CS} \leq \frac{1}{4\gamma} \leq \frac{1}{2}. \tag{C.141}$$

As a result, putting the above relations together leads to

$$C^\star_{\mathsf{clipped}} = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\frac{\min\big\{d^\star(s,a),\frac{1}{S}\big\}}{d^{\mathsf{b}}(s,a)} = \frac{\min\big\{d^\star(0,\theta),\frac{1}{S}\big\}}{d^{\mathsf{b}}(0,\theta)} = 2C. \tag{C.142}$$

Proof of inequality (C.123)**.** Observing the basic identity (using $\mu(0)+\mu(1)=1$)

$$\frac{1+\gamma(1-x)\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)x+\mu(0)\big)} = V_\theta^\star(1) + \frac{1-(1-\gamma)V_\theta^\star(1)}{1-\gamma\mu(0)-\gamma\mu(1)x},$$

we can obtain

$$\frac{1+\gamma(1-p)\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)p+\mu(0)\big)} - \frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)} = \frac{1-(1-\gamma)V_\theta^\star(1)}{1-\gamma\mu(0)-\gamma\mu(1)p} - \frac{1-(1-\gamma)V_\theta^\star(1)}{1-\gamma\mu(0)-\gamma\mu(1)x_{\pi,\theta}}$$

$$= \big(1-(1-\gamma)V_\theta^\star(1)\big)\underbrace{\frac{\gamma\mu(1)(p-x_{\pi,\theta})}{\big[1-\gamma\big(\mu(1)p+\mu(0)\big)\big]\big[1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)\big]}}_{=:\,\alpha}$$

$$= \frac{\gamma\mu(1)(p-x_{\pi,\theta})}{2\alpha}, \tag{C.143}$$

where the last relation arises from the fact (C.119).

The remainder of the proof boils down to controlling α. Making use of the definition of p (cf. (C.114)), $\mu(s)$ (cf. (E.194)) and x_π (cf. (E.205)), we can demonstrate that

$$\begin{aligned}\alpha &= \left[1-\gamma\left(\left(1-\frac{1}{CS}\right)p+\frac{1}{CS}\right)\right]\left[1-\gamma\left(\left(1-\frac{1}{CS}\right)x_{\pi,\theta}+\frac{1}{CS}\right)\right] \le (1-\gamma p)(1-\gamma x_{\pi,\theta}) \\ &\overset{\text{(i)}}{\le} (1-\gamma p)(1-\gamma q) \overset{\text{(ii)}}{\le} \left(1-\gamma\frac{p+q}{2}\right)^2 \\ &= (1-\gamma^2)^2 = (1-\gamma)^2(1+\gamma)^2 \le 4(1-\gamma)^2, \end{aligned} \tag{C.144}$$

where (i) holds true owing to the trivial fact that $x_{\pi,\theta} \ge q$ for any policy π (as long as $p \ge q$), and (ii) is a consequence of the AM-GM inequality. Substituting it into (C.143) and using the definition (E.205) give

$$\begin{aligned}\frac{1+\gamma(1-p)\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)p+\mu(0)\big)} - \frac{1+\gamma(1-x_{\pi,\theta})\mu(1)V_\theta^\star(1)}{1-\gamma\big(\mu(1)x_{\pi,\theta}+\mu(0)\big)} &= \frac{\gamma\mu(1)(p-x_{\pi,\theta})}{2\alpha} \ge \frac{\gamma\mu(1)(p-x_{\pi,\theta})}{8(1-\gamma)^2} \\ &= \frac{\gamma\mu(1)}{8(1-\gamma)^2}(p-q)\pi(1-\theta\,|\,0) \\ &\ge \frac{3\gamma}{32(1-\gamma)^2}\frac{28(1-\gamma)^2\varepsilon}{\gamma}\pi(1-\theta\,|\,0) = \frac{21\varepsilon}{8}\big(1-\pi(\theta\,|\,0)\big).\end{aligned}$$

C.4 Discounted infinite-horizon MDPs with Markovian data

In this subchapter, we extend the i.i.d. sampling model (5.22) for discounted infinite-horizon MDPs to accommodate Markovian data.

C.4.1 Sampling models and assumptions

A single Markovian trajectory. Suppose that we have access to a historical dataset $\mathcal{D}$, which comprises a single trajectory of samples with length T generated by a behavior policy π^{b}. More precisely, the sample trajectory starts from an arbitrary state s_0 and takes the form

$$\{s_0, a_0, s_1, a_1, \cdots, s_T, a_T\}, \tag{C.145}$$

which is generated by the MDP $\mathcal{M}$ (cf. Chapter 5.2.1) in the following manner

$$a_t \sim \pi^{\mathsf{b}}(\cdot \,|\, s_t), \qquad s_t' = s_{t+1} \sim P(\cdot \,|\, s_t, a_t), \qquad 0 \le t \le T. \tag{C.146}$$

Note that the dataset $\mathcal{D}$ here consists of a single trajectory, thus resulting in substantially more complicated statistical dependency in comparison to the independent sampling model. This model has also been investigated in our companion paper Yan et al. (2022a), although the focus therein is on model-free algorithms. With regards to the Markov chain generating the above sample trajectory, we shall assume it to be uniformly ergodic (see, e.g., Paulin (2015, Definition 1.1)). We denote by μ^{b} the stationary state-action distribution of this chain, and let t_{mix} indicate its mixing time as follows (Paulin, 2015)

$$t_{\mathsf{mix}}(\delta) \coloneqq \min \left\{ t \;\middle|\; \max_{s_0 \in \mathcal{S}} d_{\mathsf{TV}}\big(\mu(s_t, a_t \,|\, s_0), \mu^{\mathsf{b}}\big) \le \delta \right\}, \tag{C.147a}$$

$$t_{\mathsf{mix}} \coloneqq t_{\mathsf{mix}}(1/4), \tag{C.147b}$$

where $\mu(s_t, a_t \,|\, s_0)$ denotes the probability distribution of (s_t, a_t) given the initial state s_0, and $d_{\mathsf{TV}}(\mu, \nu)$ indicates the total-variation distance between two distributions μ and ν.

Similar to Definition 4, we introduce the following concentrability to capture the distribution shift.

Definition 7 (Single-policy clipped concentrability for discounted infinite-horizon MDPs)**.** The single-policy clipped concentrability coefficient of the dataset $\mathcal{D}$ (cf. (C.145)) is given by

$$C^{\star}_{\mathsf{clipped}} \coloneqq \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\min\left\{d^{\star}(s,a), \frac{1}{S}\right\}}{\mu^{\mathsf{b}}(s,a)}. \tag{C.148}$$

In comparison to Definition 4, the denominator in (C.148) uses the stationary distribution μ^{b} of the underlying Markov chain. In particular, if the initial state s_0 is also drawn from this distribution μ^{b}, then it is self-evident that $d^{\pi^{\mathsf{b}}} = \mu^{\mathsf{b}}$; in such a case, Condition (C.148) can be alternatively expressed as

$$C^{\star}_{\mathsf{clipped}} = \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{\min\left\{d^{\star}(s,a), \frac{1}{S}\right\}}{d^{\pi^{\mathsf{b}}}(s,a)}, \tag{C.149}$$

which is largely dictated by the closeness between the resulting occupancy of the target policy $\pi^{\star}$ and that of the behavior policy π^{b}.

C.4.2 A subsampling trick

Before continuing, we single out a crucial property that allows us to reuse the algorithmic idea developed for the i.i.d. sampling model. Imagine that the Markovian trajectory runs until $t = \infty$,

although only the first T sample transitions are revealed to us. For any given (s,a), denote by $t_i(s,a)$ the time stamp when the trajectory visits this state-action pair (s,a) for the i-th time. As has been rigorized in Li et al. (2021, Section B.1), for any given $(s,a) \in \mathcal{S} \times \mathcal{A}$, the following sample transitions

$$\left\{(s, a, s_{t_i(s,a)+1}) \mid 1 \le i \le T\right\} \tag{C.150}$$

are statistically independent, resembling the i.i.d. sampling model in some sense.

On a high level, our algorithm is built upon a two-fold subsampling trick to decouple the statistical dependence across the sample rollout. Roughly speaking, the main steps are as follows, with a detailed description provided in Algorithm 15.

i) Split data into two parts: $\mathcal{D}^{\mathsf{main}}$ and $\mathcal{D}^{\mathsf{aux}}$, each containing half of the sample trajectory. Let $N^{\mathsf{main}}(s,a)$ denote the number of sample transitions in $\mathcal{D}^{\mathsf{main}}$ from state s when action a is taken.

ii) For each $(s,a) \in \mathcal{S} \times \mathcal{A}$, use $\mathcal{D}^{\mathsf{aux}}$ to compute lower bounds $\{N^{\mathsf{trim}}(s,a)\}$ on $\{N^{\mathsf{main}}(s,a)\}$. Subsample the first $N^{\mathsf{trim}}(s,a)$ sample transitions from $\mathcal{D}^{\mathsf{main}}$ to construct data subset $\mathcal{D}^{\mathsf{trim}}$.

iii) Run VI-LCB (i.e., Algorithm 11) on the subsampled dataset $\mathcal{D}^{\mathsf{trim}}$.

The lemma below states several properties about $N_k^{\mathsf{trim}}(s,a)$ that resemble the properties in Lemma 13.

Lemma 35. *With probability at least $1-2\delta$, the quantities constructed in (C.151) obey*

$$N_k^{\mathsf{trim}}(s,a) \le N^{\mathsf{main}}(s,a), \qquad \text{for all } k \ge 665 t_{\mathsf{mix}} \tag{C.153a}$$

$$\max\left\{N_k^{\mathsf{trim}}(s,a), 222 t_{\mathsf{mix}} \log\frac{SA}{\delta}\right\} \ge \frac{T\mu^{\mathsf{b}}(s,a)}{12}, \qquad \text{for all } k \le 665 t_{\mathsf{mix}} \tag{C.153b}$$

simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Lemma 35 tells us the following important properties:

- When $N^{\mathsf{trim}}(s,a) = N_{665 t_{\mathsf{mix}}}^{\mathsf{trim}}(s,a)$, then it follows directly from Lemma 35 that

$$\max\left\{N^{\mathsf{trim}}(s,a), 222 t_{\mathsf{mix}} \log\frac{SA}{\delta}\right\} \ge \frac{T\mu^{\mathsf{b}}(s,a)}{12}. \tag{C.154a}$$

- When $N^{\mathsf{trim}}(s,a) = N_{\widehat{k}(s,a)}^{\mathsf{trim}}(s,a)$ (cf. (C.152)), it is easy to show that $\widehat{k}(s,a) \le 665 t_{\mathsf{mix}}$ and

$$\max\left\{N^{\mathsf{trim}}(s,a), 222 t_{\mathsf{mix}} \log\frac{SA}{\delta}\right\} \ge \frac{T\mu^{\mathsf{b}}(s,a)}{12}. \tag{C.154b}$$

Algorithm 15: Subsampled VI-LCB for discounted infinite MDPs with Markovian data

input: Markovian dataset $\mathcal{D}$ (cf. (C.145)); reward function r.

subsampling: run the following procedure to generate the subsampled dataset $\mathcal{D}^{\mathsf{trim}}$.

1) *Data splitting.* Split $\mathcal{D}$ into two halves: $\mathcal{D}^{\mathsf{main}} = \{s_0, a_0, s_1, a_1, \cdots, s_{T/2}\}$ (which contains the first $T/2$ transitions), and $\mathcal{D}^{\mathsf{aux}} = \{s_{T/2}, a_{T/2}, s_{T/2+1}, a_{T/2+1}, \cdots, s_T\}$ (which contains the remaining $T/2$ transitions); we let $N^{\mathsf{main}}(s,a)$ (resp. $N^{\mathsf{aux}}(s,a)$) denote the number of sample transitions in $\mathcal{D}^{\mathsf{main}}$ (resp. $\mathcal{D}^{\mathsf{aux}}$) that transition from state s with action a taken.

2) *Lower bounding* $\{N^{\mathsf{main}}(s,a)\}$ *using* $\mathcal{D}^{\mathsf{aux}}$. For each $(s,a) \in \mathcal{S} \times \mathcal{A}$, let

$$N_k^{\mathsf{trim}}(s,a) := \frac{1}{3} N^{\mathsf{aux}}(s,a) \mathbb{1}\Big(N^{\mathsf{aux}}(s,a) > k \log \frac{SA}{\delta}\Big), \qquad k \in \mathbb{N}. \tag{C.151}$$

If we know t_{mix}, set $N^{\mathsf{trim}}(s,a) = N^{\mathsf{trim}}_{665 t_{\mathsf{mix}}}(s,a)$; otherwise, set $N^{\mathsf{trim}}(s,a) = N^{\mathsf{trim}}_{\widehat{k}(s,a)}(s,a)$ with

$$\widehat{k}(s,a) = \min\big\{k : N_k^{\mathsf{trim}}(s,a) \le N^{\mathsf{main}}(s,a)\big\}. \tag{C.152}$$

3) *Subsampling.* Let $\mathcal{D}^{\mathsf{main}'}$ be the set of all sample transitions (i.e., the tuples taking the form (s,a,s')) from $\mathcal{D}^{\mathsf{main}}$. Subsample $\mathcal{D}^{\mathsf{main}'}$ to obtain $\mathcal{D}^{\mathsf{trim}}$, such that for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\mathcal{D}^{\mathsf{trim}}$ contains the first $N^{\mathsf{trim}}(s,a)$ sample transitions from $\mathcal{D}^{\mathsf{main}'}$.

run VI-LCB: set $\mathcal{D}_0 = \mathcal{D}^{\mathsf{trim}}$, and run Algorithm 11 to compute a policy $\widehat{\pi}$.

To see why this is true, it suffices to combine the fact (C.153a) with the choice (C.152).

Proof of Lemma 35. Since $N_k^{\mathsf{trim}}(s,a)$ is non-increasing as k grows, it is sufficient to prove (C.153) for $k = 665 t_{\mathsf{mix}}$. Repeating similar arguments as for Li et al. (2021, Lemma 8) — which concerns the concentration of measure for the empirical distribution of a uniformly ergodic Markov chain — implies that: with probability at least $1-\delta$,

$$\left| N^{\mathsf{aux}}(s,a) - \frac{T\mu^{\mathsf{b}}(s,a)}{2} \right| \le \frac{T\mu^{\mathsf{b}}(s,a)}{4} \tag{C.155a}$$

holds for all (s,a) obeying $\frac{T\mu^{\mathsf{b}}(s,a)}{2} \ge 443 t_{\mathsf{mix}} \log \frac{SA}{\delta}$, and

$$N^{\mathsf{aux}}(s,a) \le 665 t_{\mathsf{mix}} \log \frac{SA}{\delta} \tag{C.155b}$$

holds for all (s,a) obeying $\frac{T\mu^{\mathsf{b}}(s,a)}{2} < 443 t_{\mathsf{mix}} \log \frac{SA}{\delta}$; we omit the proof for brevity. Recalling the definition (C.151), we can readily see from (C.155) that when $k = 665 t_{\mathsf{mix}}$: $N_k^{\mathsf{trim}}(s,a) = 0$ if

$\frac{T\mu^{\mathsf{b}}(s,a)}{2} < 443 t_{\mathsf{mix}} \log \frac{SA}{\delta}$, and

$$\frac{T\mu^{\mathsf{b}}(s,a)}{12} \le N_k^{\mathsf{trim}}(s,a) \le \frac{T\mu^{\mathsf{b}}(s,a)}{4} \tag{C.156}$$

if $\frac{T\mu^{\mathsf{b}}(s,a)}{2} > 1330 t_{\mathsf{mix}} \log \frac{SA}{\delta}$. Therefore, we have established (C.153b).

We then prove relation (C.153a). Similar to (C.155a), with probability at least $1-\delta$ we have

$$\left| N^{\mathsf{main}}(s,a) - \frac{T\mu^{\mathsf{b}}(s,a)}{2} \right| \le \frac{T\mu^{\mathsf{b}}(s,a)}{4} \tag{C.157}$$

for all (s,a) with $\frac{T\mu^{\mathsf{b}}(s,a)}{2} \ge 443 t_{\mathsf{mix}} \log \frac{SA}{\delta}$. Putting (C.156) and (C.157) together establishes relation (C.153a). □

C.4.3 Performance guarantees

We are now ready to present our theoretical guarantees for Algorithm 15 in the presence of Markovian data.

Theorem 19. *Consider any $\gamma \in [\frac{1}{2}, 1)$, $0 < \delta < 1$ and any $\varepsilon \in \left(0, \frac{1}{1-\gamma}\right]$. Suppose that the number of iterations exceeds $\tau_{\max} > \frac{1}{1-\gamma} \log \frac{T}{1-\gamma}$. With probability at least $1-3\delta$, the policy $\widehat{\pi}$ returned by Algorithm 15 obeys*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \le \varepsilon, \tag{C.158}$$

as long as the penalty terms are chosen according to the Bernstein-style quantity (5.33) for any constant $c_{\mathrm{b}} \ge 144$, and the total number of samples exceeds

$$T \ge \frac{c_1 S C^{\star}_{\mathsf{clipped}} \log \frac{665 S t_{\mathsf{mix}} T}{(1-\gamma)\delta}}{(1-\gamma)^3 \varepsilon^2} + \frac{c_1 t_{\mathsf{mix}} S C^{\star}_{\mathsf{clipped}} \log \frac{665 S t_{\mathsf{mix}} T}{(1-\gamma)\delta}}{(1-\gamma)^2 \varepsilon} \tag{C.159}$$

for some large enough numerical constant $c_1 > 0$ (e.g., $c_1 = 22000$).

Theorem 19 provides a sample complexity bound that is very similar to the i.i.d. sampling case (i.e., Theorem 6), except that the length T of the sample trajectory also needs to exceed some linear scaling in the mixing time. This is unavoidable though (see also other related works Li et al. (2021); Yan et al. (2022a)); unless the sample trajectory mixes well, in general one would not be able to obtain enough information associated with all states given only this sample trajectory. In comparison to the variance-reduced offline model-free algorithm proposed in Yan et al. (2022a), our sample complexity (C.159) is strictly better (more precisely, the sample complexity in Yan et al. (2022a) has two additional terms $\frac{SC^{\star}}{(1-\gamma)^4 \varepsilon} + \frac{t_{\mathsf{mix}} C^{\star}}{(1-\gamma)^3 \varepsilon}$ and hence incurs a longer burn-in phase than the one derived herein).

Proof of Theorem 19. The proof follows very similar arguments as for Theorem 9, although we need to replace Lemma 19 by Lemma 35. For brevity, we shall only point out the parts of the analysis that need modification.

To begin with, we shall suppose for the moment that

$$N(s,a) = N^{\mathsf{trim}}(s,a) = N_k^{\mathsf{trim}}(s,a), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

for a fixed integer $k \leq 665 t_{\mathsf{mix}}$, and assume that the two parts $\mathcal{D}^{\mathsf{aux}}$ and $\mathcal{D}^{\mathsf{main}}$ are statistically independent. We will come back to remove these two assumptions towards the end of this proof.

This analysis mainly differs from that of Theorem 9 in its Step 4 when controlling $\langle d^\star, b^\star \rangle$, which we shall detail now. Let us first divide $\mathcal{S}$ into the following two disjoint state subsets:

$$\mathcal{S}^{\mathsf{small}} := \left\{ s \in \mathcal{S} \mid T\mu^{\mathsf{b}}\big(s, \pi^\star(s)\big) \leq 2660 t_{\mathsf{mix}} \log \frac{SA}{\delta} \right\}; \tag{C.160a}$$

$$\mathcal{S}^{\mathsf{large}} := \left\{ s \in \mathcal{S} \mid T\mu^{\mathsf{b}}\big(s, \pi^\star(s)\big) > 2660 t_{\mathsf{mix}} \log \frac{SA}{\delta} \right\}. \tag{C.160b}$$

- Firstly, for any state $s \in \mathcal{S}^{\mathsf{small}}$, combine Definition 7 with the definition of $\mathcal{S}^{\mathsf{small}}$ to give

$$\min\left\{ d^\star(s), \frac{1}{S} \right\} \leq C^\star_{\mathsf{clipped}} \mu^{\mathsf{b}}\big(s, \pi^\star(s)\big) \leq \frac{2660 C^\star_{\mathsf{clipped}} t_{\mathsf{mix}} \log \frac{T}{(1-\gamma)\delta}}{T} < \frac{1}{S}, \tag{C.161}$$

  provided that $T > 2660 S C^\star_{\mathsf{clipped}} t_{\mathsf{mix}} \log \frac{T}{(1-\gamma)\delta}$. An immediate consequence of this result is that

$$d^\star(s) \leq \frac{2660 C^\star_{\mathsf{clipped}} t_{\mathsf{mix}} \log \frac{T}{(1-\gamma)\delta}}{T} < \frac{1}{S}. \tag{C.162}$$

  Taking this inequality together with the fact below (see the definition (5.33))

$$b^\star(s) := b\big(s, \pi^\star(s); \widehat{V}\big) \leq \frac{1}{1-\gamma} + \frac{5}{T}, \tag{C.163}$$

  we can demonstrate that

$$\begin{aligned} \sum_{s \in \mathcal{S}^{\mathsf{small}}} d^\star(s) b^\star(s) &\leq \sum_{s \in \mathcal{S}^{\mathsf{small}}} \left(\frac{2660 C^\star_{\mathsf{clipped}} t_{\mathsf{mix}} \log \frac{T}{(1-\gamma)\delta}}{(1-\gamma)T} + d^\star(s) \frac{5}{T} \right) \\ &\leq \frac{2660 S C^\star_{\mathsf{clipped}} t_{\mathsf{mix}} \log \frac{T}{(1-\gamma)\delta}}{(1-\gamma)T} + \frac{5}{T}. \end{aligned} \tag{C.164}$$

- Secondly, we look at any state $s \in \mathcal{S}^{\mathsf{large}}$. From the definition (5.33) of $b(s,a;V)$, one obtains

$$\begin{aligned} b^\star(s) &\leq \sqrt{\frac{c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{N\big(s,\pi^\star(s)\big)} \mathsf{Var}_{\widehat{P}_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{2c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{(1-\gamma)N\big(s,\pi^\star(s)\big)}} + \frac{5}{T} \\ &\overset{\text{(i)}}{\leq} \sqrt{\frac{c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{N\big(s,\pi^\star(s)\big)} \left(2\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{41 \log \frac{2T}{(1-\gamma)\delta}}{(1-\gamma)^2 N\big(s,\pi^\star(s)\big)}\right) + \frac{2c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{(1-\gamma)N\big(s,\pi^\star(s)\big)}} + \frac{5}{T} \\ &\overset{\text{(ii)}}{\leq} \sqrt{\frac{2c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{N\big(s,\pi^\star(s)\big)} \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{4c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{(1-\gamma)N\big(s,\pi^\star(s)\big)}}, \end{aligned} \tag{C.165}$$

where (i) comes from Lemma 20 and inequality (5.71), (ii) follows since $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y \geq 0$ and $T \geq N(s,a)$ for c_{b} large enough. Moreover, it is seen that

$$\frac{1}{N\big(s,\pi^\star(s)\big)} \overset{\text{(i)}}{\leq} \frac{12}{T\mu^{\mathsf{b}}\big(s,\pi^\star(s)\big)} \overset{\text{(ii)}}{\leq} \frac{12C^\star_{\mathsf{clipped}}}{T\min\big\{d^\star(s), \frac{1}{S}\big\}} \leq \frac{12C^\star_{\mathsf{clipped}}}{T}\left(\frac{1}{d^\star(s)} + S\right), \tag{C.166}$$

where (i) follows from (C.154), and (ii) invokes Definition 7. Plugging this inequality into (C.165) gives

$$\begin{aligned} b^\star(s) &\leq \sqrt{\frac{2c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{N\big(s,\pi^\star(s)\big)} \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) + \frac{4c_{\mathrm{b}} \log \frac{T}{(1-\gamma)\delta}}{(1-\gamma)N\big(s,\pi^\star(s)\big)}} \\ &\leq \underbrace{\sqrt{\frac{24c_{\mathrm{b}}C^\star_{\mathsf{clipped}} \log \frac{T}{(1-\gamma)\delta}}{T} \mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)}\left(\frac{1}{\sqrt{d^\star(s)}} + \sqrt{S}\right)}_{=:\alpha_1(s)} + \underbrace{\frac{48c_{\mathrm{b}}C^\star_{\mathsf{clipped}} \log \frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}\left(\frac{1}{d^\star(s)} + S\right)}_{=:\alpha_2(s)}, \end{aligned} \tag{C.167}$$

where the last line also applies the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for any $x, y \geq 0$.

To continue, observe that summing the first terms in (C.167) over $s \in \mathcal{S}^{\mathsf{large}}$ satisfies

$$\begin{aligned} &\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\alpha_1(s) \\ &= \sqrt{\frac{24c_{\mathrm{b}}C^\star_{\mathsf{clipped}} \log \frac{T}{(1-\gamma)\delta}}{T}}\left(\sum_{s\in\mathcal{S}^{\mathsf{large}}} \sqrt{d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \sum_{s\in\mathcal{S}^{\mathsf{large}}} \sqrt{d^\star(s)}\sqrt{Sd^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)}\right) \\ &\overset{\text{(i)}}{\leq} \sqrt{\frac{24c_{\mathrm{b}}C^\star_{\mathsf{clipped}} \log \frac{T}{(1-\gamma)\delta}}{T}}\left(\sqrt{S}\sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)} + \sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^\star(s)S\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big)}\right) \end{aligned}$$

$$= \sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{T}}\sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)\mathsf{Var}_{P_{s,\pi^{\star}(s)}}\big(\widehat{V}\big)}, \tag{C.168}$$

where (i) comes from Cauchy-Schwarz and the fact $\sum_s d^{\star}(s)=1$. In addition, the sum of the second terms in (C.167) over $s\in\mathcal{S}^{\mathsf{large}}$ obeys

$$\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)\alpha_2(s) \le \frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}, \tag{C.169}$$

which follows since $\sum_s d^{\star}(s)=1$. Combine (C.168) and (C.169) with (C.167) to yield

$$\begin{aligned}\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)b^{\star}\big(s,\pi^{\star}(s)\big) &\le \sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)\alpha_1(s) + \sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)\alpha_2(s)\\ &\le \sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{T}}\sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)\mathsf{Var}_{P_{s,\pi^{\star}(s)}}\big(\widehat{V}\big)} + \frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}.\end{aligned} \tag{C.170}$$

The above results (C.164) and (C.170) taken collectively give

$$\begin{aligned}\big\langle d^{\star}, b^{\star}\big\rangle &= \sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)b^{\star}(s) + \sum_{s\in\mathcal{S}^{\mathsf{small}}} d^{\star}(s)b^{\star}(s)\\
&\le \sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{T}}\sqrt{\sum_{s\in\mathcal{S}^{\mathsf{large}}} d^{\star}(s)\mathsf{Var}_{P_{s,\pi^{\star}(s)}}\big(\widehat{V}\big)} + \frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}\\
&\quad + \frac{2660SC^{\star}_{\mathsf{clipped}}t_{\mathsf{mix}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T} + \frac{5}{T}\\
&\overset{\text{(i)}}{\le} \sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{T}}\sqrt{\sum_{s\in\mathcal{S}} d^{\star}(s)\mathsf{Var}_{P_{s,\pi^{\star}(s)}}\big(\widehat{V}\big)} + \frac{192c_{\mathrm{b}}t_{\mathsf{mix}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}\\
&\overset{\text{(ii)}}{\le} \frac{2}{\gamma}\sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}\big\langle d^{\star}, b^{\star}\big\rangle} + \frac{1}{\gamma}\sqrt{\frac{192c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}} + \frac{192c_{\mathrm{b}}t_{\mathsf{mix}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T},\\
&\overset{\text{(iii)}}{\le} 4\sqrt{\frac{96c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}\big\langle d^{\star}, b^{\star}\big\rangle} + 2\sqrt{\frac{192c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}} + \frac{192c_{\mathrm{b}}t_{\mathsf{mix}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T},\\
&\overset{\text{(iv)}}{\le} \frac{1}{2}\big\langle d^{\star}, b^{\star}\big\rangle + \frac{768c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T} + \sqrt{\frac{768c_{\mathrm{b}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}} + \frac{192c_{\mathrm{b}}t_{\mathsf{mix}}SC^{\star}_{\mathsf{clipped}}\log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}.\end{aligned}$$

Here, (i) follows when c_{b} is sufficiently large, $t_{\mathsf{mix}}\ge 1$ and $C^{\star}_{\mathsf{clipped}}\ge 1/S$ (see (5.26)), (ii) holds by

recalling (5.105) that

$$\sum_{s\in\mathcal{S}} d^\star(s)\mathsf{Var}_{P_{s,\pi^\star(s)}}\big(\widehat{V}\big) \le \frac{2}{\gamma^2(1-\gamma)} + \frac{4}{\gamma^2(1-\gamma)}\big\langle d^\star, b^\star\big\rangle;$$

(iii) is valid since $\gamma \in [\frac{1}{2}, 1)$, and (iv) follows since $2xy \le x^2 + y^2$. Rearrange terms to yield

$$\big\langle d^\star, b^\star\big\rangle \le \sqrt{\frac{3072 c_{\mathrm{b}} SC^\star_{\mathsf{clipped}} \log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T}} + \frac{1920 c_{\mathrm{b}} t_{\mathsf{mix}} SC^\star_{\mathsf{clipped}} \log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)T},$$

which in turn leads to

$$\big\langle \rho, V^\star - V^{\widehat{\pi}}\big\rangle \le \frac{2\big\langle d^\star, b^\star\big\rangle}{1-\gamma} \le 120\sqrt{\frac{c_{\mathrm{b}} SC^\star_{\mathsf{clipped}} \log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)^3 T}} + \frac{3840 c_{\mathrm{b}} t_{\mathsf{mix}} SC^\star_{\mathsf{clipped}} \log\frac{T}{(1-\gamma)\delta}}{(1-\gamma)^2 T}, \quad \text{(C.171)}$$

where the first inequality follows from relation (5.91).

Finally, we explain how to relax the two strong assumptions imposed at the beginning of this proof.

- Note that when t_{mix} is not known *a priori*, the choice $\widehat{k}(s,a)$ is not a fixed integer independent from the data. Fortunately, we have already demonstrated right after Lemma 35 that $\widehat{k}(s,a) \le 665 t_{\mathsf{mix}}$. Taking the union bound over all integers $1 \le k \le 665 t_{\mathsf{mix}}$ suffices to handle this statistical dependency.

- In general, $\mathcal{D}^{\mathsf{aux}}$ and $\mathcal{D}^{\mathsf{main}}$ are statistically dependent as they are two parts of the same trajectory. To resolve this issue, consider an additional collection of S *independent* Markovian trajectories each of length $T/2$ (denoted by $\mathcal{D}^{\mathsf{aux}}(1), \mathcal{D}^{\mathsf{aux}}(2), \cdots, \mathcal{D}^{\mathsf{aux}}(S)$), where the $\mathcal{D}^{\mathsf{aux}}(s)$ starts from state s (note that these trajectories are introduced merely to assist in the proof). Taking the union bound over all $s \in \mathcal{S}$, we can establish the desired result, for every $s \in \mathcal{S}$, if the second dataset is $\mathcal{D}^{\mathsf{aux}}(s)$. Additionally, due to the Markovian property, we know that the true $\mathcal{D}^{\mathsf{aux}}$ must be statistically equivalent to one of these trajectories $\mathcal{D}^{\mathsf{aux}}(1), \cdots, \mathcal{D}^{\mathsf{aux}}(S)$; this in turn establishes the claimed results for the true dataset.

Putting everything together ensures the desired sample complexity as stated in Theorem 19.

□

Appendix D

Proofs for Chapter 6

D.1 Preliminaries

For any vector x, we overload the notation by letting $x^{\circ 2} = \big[x(s,a)^2\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ (resp. $x^{\circ 2} = \big[x(s)^2\big]_{s\in\mathcal{S}}$). With slight abuse of notation, we denote 0 (resp. 1) as the all-zero (resp. all-one) vector, and drop the subscript ρ to write $\mathcal{U}^\sigma(\cdot) = \mathcal{U}^\sigma_\rho(\cdot)$ whenever the argument holds for all divergence ρ.

Matrix notation. To continue, we recall or introduce some additional matrix notation that is useful throughout the analysis.

- $P^0 \in \mathbb{R}^{SA\times S}$: the matrix of the nominal transition kernel with $P^0_{s,a}$ as the (s,a)-th row.
- $\widehat{P}^0 \in \mathbb{R}^{SA\times S}$: the matrix of the estimated nomimal transition kernel with $\widehat{P}^0_{s,a}$ as the (s,a)-th row.
- $r \in \mathbb{R}^{SA}$: a vector representing the reward function r (so that $r_{(s,a)} = r(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$).
- $\Pi^\pi \in \{0,1\}^{S\times SA}$: a projection matrix associated with a given deterministic policy π taking the following form

$$\Pi^\pi = \begin{pmatrix} e_{\pi(1)}^\top & 0^\top & \cdots & 0^\top \\ 0^\top & e_{\pi(2)}^\top & \cdots & 0^\top \\ \vdots & \vdots & \ddots & \vdots \\ 0^\top & 0^\top & \cdots & e_{\pi(S)}^\top \end{pmatrix}, \tag{D.1}$$

  where $e_{\pi(1)}^\top, e_{\pi(2)}^\top, \ldots, e_{\pi(S)}^\top \in \mathbb{R}^A$ are standard basis vectors.
- $r_\pi \in \mathbb{R}^S$: a reward vector restricted to the actions chosen by the policy π, namely, $r_\pi(s) = r(s,\pi(s))$ for all $s \in \mathcal{S}$ (or simply, $r_\pi = \Pi^\pi r$).
- $\mathrm{Var}_P(V) \in \mathbb{R}^{SA}$: for any transition kernel $P \in \mathbb{R}^{SA\times S}$ and vector $V \in \mathbb{R}^S$, we denote the (s,a)-th row of $\mathrm{Var}_P(V)$ as

$$\mathsf{Var}_P(s,a) \coloneqq \mathrm{Var}_{P_{s,a}}(V). \tag{D.2}$$

- $P^V \in \mathbb{R}^{SA\times S}$, $\widehat{P}^V \in \mathbb{R}^{SA\times S}$: the matrices representing the probability transition kernel in the uncertainty set that leads to the worst-case value for any vector $V \in \mathbb{R}^S$. We denote $P^V_{s,a}$ (resp. $\widehat{P}^V_{s,a}$) as the (s,a)-th row of the transition matrix P^V (resp. $\widehat{P}^V$). In truth, the (s,a)-th rows of these transition matrices are defined as

$$P^V_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}V, \qquad \text{and} \qquad \widehat{P}^V_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}V. \tag{D.3a}$$

  Furthermore, we make use of the following short-hand notation:

$$P^{\pi,V}_{s,a} := P^{V^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}V^{\pi,\sigma}, \qquad P^{\pi,\widehat{V}}_{s,a} := P^{\widehat{V}^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{D.3b}$$

$$\widehat{P}^{\pi,V}_{s,a} := \widehat{P}^{V^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{P\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} PV^{\pi,\sigma}, \qquad \widehat{P}^{\pi,\widehat{V}}_{s,a} := \widehat{P}^{\widehat{V}^{\pi,\sigma}}_{s,a} = \operatorname{argmin}_{P\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} P\widehat{V}^{\pi,\sigma}. \tag{D.3c}$$

  The corresponding probability transition matrices are denoted by $P^{\pi,V} \in \mathbb{R}^{SA\times S}$, $P^{\pi,\widehat{V}} \in \mathbb{R}^{SA\times S}$, $\widehat{P}^{\pi,V} \in \mathbb{R}^{SA\times S}$ and $\widehat{P}^{\pi,\widehat{V}} \in \mathbb{R}^{SA\times S}$, respectively.

- $P^\pi \in \mathbb{R}^{S\times S}$, $\widehat{P}^\pi \in \mathbb{R}^{S\times S}$, $\underline{P}^{\pi,V} \in \mathbb{R}^{S\times S}$, $\underline{P}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$, $\underline{\widehat{P}}^{\pi,V} \in \mathbb{R}^{S\times S}$ and $\underline{\widehat{P}}^{\pi,\widehat{V}} \in \mathbb{R}^{S\times S}$: six *square* probability transition matrices w.r.t. policy π over the states, namely

$$\begin{aligned} &P^\pi := \Pi^\pi P^0, \qquad \widehat{P}^\pi := \Pi^\pi \widehat{P}^0, \qquad \underline{P}^{\pi,V} := \Pi^\pi P^{\pi,V}, \qquad \underline{P}^{\pi,\widehat{V}} := \Pi^\pi P^{\pi,\widehat{V}}, \\ &\underline{\widehat{P}}^{\pi,V} := \Pi^\pi \widehat{P}^{\pi,V}, \qquad \text{and} \qquad \underline{\widehat{P}}^{\pi,\widehat{V}} := \Pi^\pi \widehat{P}^{\pi,\widehat{V}}. \end{aligned} \tag{D.4}$$

  We denote P^π_s as the s-th row of the transition matrix P^π; similar quantities can be defined for the other matrices as well.

D.1.1 Basic facts

Kullback-Leibler (KL) divergence. First, for any two distributions P and Q, we denote by $\mathsf{KL}(P \parallel Q)$ the Kullback-Leibler (KL) divergence of P and Q. Letting $\mathsf{Ber}(p)$ be the Bernoulli distribution with mean p, we also introduce

$$\mathsf{KL}(p \parallel q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \qquad \text{and} \qquad \chi^2(p \parallel q) := \frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q} = \frac{(p-q)^2}{q(1-q)}, \tag{D.5}$$

which represent respectively the KL divergence and the χ^2 divergence of $\mathsf{Ber}(p)$ from $\mathsf{Ber}(q)$ (Tsybakov, 2009). We make note of the following useful property about the KL divergence in Tsybakov (2009, Lemma 2.7).

Lemma 36. *For any $p, q \in (0,1)$, it holds that*

$$\mathsf{KL}(p \parallel q) \le \frac{(p-q)^2}{q(1-q)}. \tag{D.6}$$

Variance. For any probability vector $P \in \mathbb{R}^{1\times S}$ and vector $V \in \mathbb{R}^S$, we denote the variance

$$\mathrm{Var}_P(V) \coloneqq P(V \circ V) - (PV) \circ (PV). \tag{D.7}$$

The following lemma bounds the Lipschitz constant of the variance function.

Lemma 37. *Consider any $0 \le V_1, V_2 \le \frac{1}{1-\gamma}$ obeying $\|V_1 - V_2\|_\infty \le x$ and any probability vector $P \in \Delta(S)$, one has*

$$|\mathrm{Var}_P(V_1) - \mathrm{Var}_P(V_2)| \le \frac{2x}{(1-\gamma)}. \tag{D.8}$$

Proof. It is immediate to check that

$$\begin{aligned} |\mathrm{Var}_P(V_1) - \mathrm{Var}_P(V_2)| &= |P(V_1 \circ V_1) - (PV_1) \circ (PV_1) - P(V_2 \circ V_2) + (PV_2) \circ (PV_2)| \\ &\le \big|P\big(V_1 \circ V_1 - V_2 \circ V_2\big)\big| + |(PV_1 + PV_2)P(V_1 - V_2)| \\ &\le 2\|V_1 + V_2\|_\infty \|V_1 - V_2\|_\infty \le \frac{2x}{(1-\gamma)}. \end{aligned} \tag{D.9}$$

where the penultimate inequality holds by the triangle inequality. □

D.1.2 Properties of the robust Bellman operator

γ-contraction of the robust Bellman operator. It is worth noting that the robust Bellman operator (cf. (2.29)) shares the nice γ-contraction property of the standard Bellman operator, stated as below.

Lemma 38 (γ-Contraction)**.** *(Iyengar, 2005, Theorem 3.2) For any $\gamma \in [0,1)$, the robust Bellman operator $\mathcal{T}^\sigma(\cdot)$ (cf. (2.29)) is a γ-contraction w.r.t. $\|\cdot\|_\infty$. Namely, for any $Q_1, Q_2 \in \mathbb{R}^{SA}$ s.t. $Q_1(s,a), Q_2(s,a) \in \big[0, \frac{1}{1-\gamma}\big]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, one has*

$$\left\|\mathcal{T}^\sigma(Q_1) - \mathcal{T}^\sigma(Q_2)\right\|_\infty \le \gamma \left\|Q_1 - Q_2\right\|_\infty. \tag{D.10}$$

Additionally, $Q^{\star,\sigma}$ is the unique fixed point of $\mathcal{T}^\sigma(\cdot)$ obeying $0 \le Q^{\star,\sigma}(s,a) \le \frac{1}{1-\gamma}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Dual equivalence of the robust Bellman operator. Fortunately, the robust Bellman operator can be evaluated efficiently by resorting to its dual formulation (Iyengar, 2005). In what follows, we

shall illustrate this for the two choices of the divergence ρ of interest. Before continuing, for any $V \in \mathbb{R}^S$, we denote $[V]_\alpha$ as its clipped version by some non-negative value α, namely,

$$[V]_\alpha(s) \coloneqq \begin{cases} \alpha, & \text{if } V(s) > \alpha, \\ V(s), & \text{otherwise.} \end{cases} \tag{D.11}$$

- TV distance, where the uncertainty set is $\mathcal{U}^\sigma_\rho(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\mathsf{TV}}(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\rho_{\mathsf{TV}}}(\widehat{P}^0_{s,a})$ w.r.t. the TV distance $\rho = \rho_{\mathsf{TV}}$ defined in (6.1). In particular, we have the following lemma due to strong duality, which is a direct consequence of Iyengar (2005, Lemma 4.3).

  Lemma 39 (Strong duality for TV)**.** *Consider any probability vector $P \in \Delta(\mathcal{S})$, any fixed uncertainty level σ and the uncertainty set $\mathcal{U}^\sigma(P) \coloneqq \mathcal{U}^\sigma_{\mathsf{TV}}(P)$. For any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, recalling the definition of $[V]_\alpha$ in (D.11), one has*

$$\inf_{\mathcal{P} \in \mathcal{U}^\sigma(P)} \mathcal{P}V = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P\,[V]_\alpha - \sigma \left(\alpha - \min_{s'} [V]_\alpha\,(s') \right) \right\}. \tag{D.12}$$

  In view of the above lemma, the following dual update rule is equivalent to (6.8) in DRVI:

$$\widehat{Q}_t(s,a) = r(s,a) + \gamma \max_{\alpha \in \left[\min_s \widehat{V}_{t-1}(s), \max_s \widehat{V}_{t-1}(s)\right]} \left\{ \widehat{P}^0_{s,a} \left[\widehat{V}_{t-1}\right]_\alpha - \sigma \left(\alpha - \min_{s'} \left[\widehat{V}_{t-1}\right]_\alpha (s') \right) \right\}. \tag{D.13}$$

- χ^2 divergence, where the uncertainty set is $\mathcal{U}^\sigma_\rho(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\chi^2}(\widehat{P}^0_{s,a}) \coloneqq \mathcal{U}^\sigma_{\rho_{\chi^2}}(\widehat{P}^0_{s,a})$ w.r.t. the χ^2 divergence $\rho = \rho_{\chi^2}$ defined in (6.2). We introduce the following lemma which directly follows from (Iyengar, 2005, Lemma 4.2).

  Lemma 40 (Strong duality for χ^2)**.** *Consider any probability vector $P \in \Delta(\mathcal{S})$, any fixed uncertainty level σ and the uncertainty set $\mathcal{U}^\sigma(P) \coloneqq \mathcal{U}^\sigma_{\chi^2}(P)$. For any vector $V \in \mathbb{R}^S$ obeying $V \geq 0$, one has*

$$\inf_{\mathcal{P} \in \mathcal{U}^\sigma(P)} \mathcal{P}V = \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left\{ P[V]_\alpha - \sqrt{\sigma \mathsf{Var}_P\left([V]_\alpha\right)} \right\}, \tag{D.14}$$

  where $\mathsf{Var}_P(\cdot)$ is defined as (D.7).

  In view of the above lemma, the update rule (6.8) in DRVI can be equivalently written as:

$$\widehat{Q}_t(s,a) = r(s,a) + \gamma \max_{\alpha \in \left[\min_s \widehat{V}_{t-1}(s), \max_s \widehat{V}_{t-1}(s)\right]} \left\{ \widehat{P}^0_{s,a} \left[\widehat{V}_{t-1}\right]_\alpha - \sqrt{\sigma \mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}_{t-1}\right]_\alpha\right)} \right\}. \tag{D.15}$$

The proofs of Lemma 39 and Lemma 40 are provided as follows.

Proof of Lemma 39. To begin with, applying (Iyengar, 2005, Lemma 4.3), the term of interest obeys

$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P)} \mathcal{P}V = \max_{\mu\in\mathbb{R}^S,\mu\geq 0} \left\{ P\left(V-\mu\right) - \sigma\left(\max_{s'}\left\{V(s')-\mu(s')\right\} - \min_{s'}\left\{V(s')-\mu(s')\right\}\right)\right\}, \quad \text{(D.16)}$$

where $\mu(s')$ represents the s'-th entry of $\mu\in\mathbb{R}^S$. Denoting $\mu^\star$ as the optimal dual solution, taking $\alpha = \max_{s'}\left\{V(s')-\mu^\star(s')\right\}$, it is easily verified that $\mu^\star$ obeys

$$\mu^\star(s) = \begin{cases} V(s)-\alpha, & \text{if } V(s)>\alpha \\ 0, & \text{otherwise.} \end{cases} \quad \text{(D.17)}$$

Therefore, (D.16) can be solved by optimizing α as below (Iyengar, 2005, Lemma 4.3):

$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P)} \mathcal{P}V = \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left\{ P\left[V\right]_\alpha - \sigma\left(\alpha - \min_{s'}\left[V\right]_\alpha(s')\right)\right\}. \quad \text{(D.18)}$$

□

Proof of Lemma 40. Due to strong duality (Iyengar, 2005, Lemma 4.2), it holds that

$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P)} \mathcal{P}V = \max_{\mu\in\mathbb{R}^S,\mu\geq 0} \left\{ P\left(V-\mu\right) - \sqrt{\sigma \mathsf{Var}_P\left(V-\mu\right)}\right\}, \quad \text{(D.19)}$$

and the optimal $\mu^\star$ obeys

$$\mu^\star(s) = \begin{cases} V(s)-\alpha, & \text{if } V(s)>\alpha \\ 0, & \text{otherwise.} \end{cases} \quad \text{(D.20)}$$

for some $\alpha\in[\min_s V(s),\max_s V(s)]$. As a result, solving (D.19) is equivalent to optimizing the scalar α as below:

$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P)} \mathcal{P}V = \max_{\alpha\in[\min_s V(s),\max_s V(s)]} \left\{ P[V]_\alpha - \sqrt{\sigma \mathsf{Var}_P\left([V]_\alpha\right)}\right\}. \quad \text{(D.21)}$$

□

D.1.3 Additional facts of the empirical robust MDP

Bellman equations of the empirical robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$. To begin with, recall that the empirical robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}} = \{\mathcal{S},\mathcal{A},\gamma,\mathcal{U}^{\sigma}(\widehat{P}^0),r\}$ based on the estimated nominal distribution $\widehat{P}^0$ constructed in (6.5) and its corresponding robust value function (resp. robust Q-function) $\widehat{V}^{\pi,\sigma}$ (resp. $\widehat{Q}^{\pi,\sigma}$).

Note that $\widehat{Q}^{\star,\sigma}$ is the unique fixed point of $\widehat{\mathcal{T}}^{\sigma}(\cdot)$ (see Lemma 38), the empirical robust Bellman operator constructed using $\widehat{P}^0$. Moreover, similar to (2.27), for $\widehat{\mathcal{M}}_{\mathsf{rob}}$, the Bellman's optimality principle gives the following *robust Bellman consistency equation* (resp. *robust Bellman optimality equation*):

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{Q}^{\pi,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{D.22a}$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{Q}^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{D.22b}$$

With these in mind, combined with the matrix notation, for any policy π, we can write the robust Bellman consistency equations as

$$Q^{\pi,\sigma} = r + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P^0)} \mathcal{P}V^{\pi,\sigma} \quad \text{and} \quad \widehat{Q}^{\pi,\sigma} = r + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0)} \mathcal{P}\widehat{V}^{\pi,\sigma}, \tag{D.23}$$

which leads to

$$\begin{aligned} V^{\pi,\sigma} &= r_\pi + \gamma \Pi^\pi \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P^0)} \mathcal{P}V^{\pi,\sigma} \overset{\text{(i)}}{=} r_\pi + \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}, \\ \widehat{V}^{\pi,\sigma} &= r_\pi + \gamma \Pi^\pi \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0)} \mathcal{P}\widehat{V}^{\pi,\sigma} \overset{\text{(ii)}}{=} r_\pi + \gamma \underline{\widehat{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}, \end{aligned} \tag{D.24}$$

where (i) and (ii) holds by the definitions in (D.1), (D.3) and (D.4).

Encouragingly, the above property of the robust Bellman operator ensures the fast convergence of DRVI. We collect this consequence in the following lemma.

Lemma 41. *Let $\widehat{Q}_0 = 0$. The iterates $\{\widehat{Q}_t\}, \{\widehat{V}_t\}$ of* DRVI *obey*

$$\forall t \geq 0: \quad \big\|\widehat{Q}_t - \widehat{Q}^{\star,\sigma}\big\|_\infty \leq \frac{\gamma^t}{1-\gamma} \qquad \text{and} \qquad \big\|\widehat{V}_t - \widehat{V}^{\star,\sigma}\big\|_\infty \leq \frac{\gamma^t}{1-\gamma}. \tag{D.25}$$

Furthermore, the output policy $\widehat{\pi}$ obeys

$$\big\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\big\|_\infty \leq \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}, \qquad \text{where} \quad \big\|\widehat{V}^{\star,\sigma} - \widehat{V}_{T-1}\big\|_\infty =: \varepsilon_{\mathsf{opt}}. \tag{D.26}$$

Proof of Lemma 41. Applying the γ-contraction property in Lemma 38 directly yields that for any $t \geq 0$,

$$\begin{aligned} \|\widehat{Q}_t - \widehat{Q}^{\star,\sigma}\|_\infty = \big\|\widehat{\mathcal{T}}^{\sigma}(\widehat{Q}_{t-1}) - \widehat{\mathcal{T}}^{\sigma}(\widehat{Q}^{\star,\sigma})\big\|_\infty &\leq \gamma \big\|\widehat{Q}_{t-1} - \widehat{Q}^{\star,\sigma}\big\|_\infty \\ &\leq \cdots \leq \gamma^t \big\|\widehat{Q}_0 - \widehat{Q}^{\star,\sigma}\big\|_\infty = \gamma^t \big\|\widehat{Q}^{\star,\sigma}\big\|_\infty \leq \frac{\gamma^t}{1-\gamma}, \end{aligned}$$

where the last inequality holds by the fact $\|\widehat{Q}^{\star,\sigma}\|_\infty \le \frac{1}{1-\gamma}$ (see Lemma 38). In addition,

$$\left\|\widehat{V}_t - \widehat{V}^{\star,\sigma}\right\|_\infty = \max_{s\in\mathcal{S}} \left| \max_{a\in\mathcal{A}} \widehat{Q}_t(s,a) - \max_{a\in\mathcal{A}} \widehat{Q}^{\star,\sigma}(s,a)\right| _\infty \le \left\|\widehat{Q}_t - \widehat{Q}^{\star,\sigma}\right\|_\infty \le \frac{\gamma^t}{1-\gamma},$$

where the penultimate inequality holds by the maximum operator is 1-Lipschitz. This completes the proof of (D.25).

We now move to establish (D.26). Note that there exists at least one state $s_0 \in \mathcal{S}$ that is associated with the maximum of the value gap, i.e.,

$$\left\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty = \widehat{V}^{\star,\sigma}(s_0) - \widehat{V}^{\widehat{\pi},\sigma}(s_0) \ge \widehat{V}^{\star,\sigma}(s) - \widehat{V}^{\widehat{\pi},\sigma}(s), \qquad \forall s \in \mathcal{S}.$$

Recall $\widehat{\pi}^\star$ is the optimal robust policy for the empirical RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$. For convenience, we denote $a_1 = \widehat{\pi}^\star(s_0)$ and $a_2 = \widehat{\pi}(s_0)$. Then, since $\widehat{\pi}$ is the greedy policy w.r.t. $\widehat{Q}_T$, one has

$$r(s_0,a_1) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}_{T-1} = \widehat{Q}_T(s_0,a_1) \le \widehat{Q}_T(s_0,a_2) = r(s_0,a_2) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}_{T-1}. \tag{D.27}$$

Recalling the notation in (D.3), the above fact and (D.26) altogether yield

$$\begin{aligned}
r(s_0,a_1) + \gamma \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\left(\widehat{V}^{\star,\sigma} - \varepsilon_{\mathsf{opt}}1\right) &\le r(s_0,a_1) + \gamma \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\widehat{V}_{T-1} \\
&\le r(s_0,a_2) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}_{T-1} \\
&\overset{\text{(i)}}{\le} r(s_0,a_2) + \gamma \widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2}\widehat{V}_{T-1} \\
&\le r(s_0,a_2) + \gamma \widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2}\left(\widehat{V}^{\star,\sigma} + \varepsilon_{\mathsf{opt}}1\right),
\end{aligned} \tag{D.28}$$

where (i) follows from the optimality criteria. The term of interest can be controlled as

$$\begin{aligned}
\left\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty &= \widehat{V}^{\star,\sigma}(s_0) - \widehat{V}^{\widehat{\pi},\sigma}(s_0) \\
&= r(s_0,a_1) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \left(r(s_0,a_2) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma}\right) \\
&= r(s_0,a_1) - r(s_0,a_2) + \gamma\left(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma}\right) \\
&\overset{\text{(i)}}{\le} 2\gamma\varepsilon_{\mathsf{opt}} + \gamma\left(\widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2}\widehat{V}^{\star,\sigma} - \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\widehat{V}^{\star,\sigma} + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma}\right) \\
&= 2\gamma\varepsilon_{\mathsf{opt}} + \gamma\left(\widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2}\widehat{V}^{\star,\sigma} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_2})} \mathcal{P}\widehat{V}^{\widehat{\pi},\sigma}\right) + \gamma\left(\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})} \mathcal{P}\widehat{V}^{\star,\sigma} - \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\widehat{V}^{\star,\sigma}\right)
\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} 2\gamma\varepsilon_{\mathsf{opt}} + \gamma \widehat{P}^{\widehat{V}^{\widehat{\pi},\sigma}}_{s_0,a_2}\left(\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right) + \gamma\left(\widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\widehat{V}^{\star,\sigma} - \widehat{P}^{\widehat{V}_{T-1}}_{s_0,a_1}\widehat{V}^{\star,\sigma}\right)$$
$$\leq 2\gamma\varepsilon_{\mathsf{opt}} + \gamma\left\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty, \tag{D.29}$$

where (i) holds by plugging in (D.28), and (ii) follows from $\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})}\mathcal{P}\widehat{V}^{\star,\sigma} \leq \mathcal{P}\widehat{V}^{\star,\sigma}$ for any $\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s_0,a_1})$. Rearranging (D.29) leads to

$$\left\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty \leq \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}.$$

□

D.2 Proof of the upper bound with TV distance: Theorem 10

Throughout this subchapter, for any transition kernel P, the uncertainty set is taken as (see (6.1))

$$\mathcal{U}^\sigma(P) \coloneqq \mathcal{U}^\sigma_{\mathsf{TV}}(P) = \otimes\,\mathcal{U}^\sigma_{\mathsf{TV}}(P_{s,a}), \qquad \mathcal{U}^\sigma_{\mathsf{TV}}(P_{s,a}) \coloneqq \left\{P'_{s,a}\in\Delta(\mathcal{S}) : \frac{1}{2}\left\|P'_{s,a} - P_{s,a}\right\|_1 \leq \sigma\right\}. \tag{D.30}$$

D.2.1 Technical lemmas

We begin with a key lemma concerning the dynamic range of the robust value function $V^{\pi,\sigma}$ (cf. (2.25)), which produces tighter control when σ is large; the proof is deferred to Appendix D.2.3.1.

Lemma 42. *For any nominal transition kernel $P\in\mathbb{R}^{SA\times S}$, any fixed uncertainty level σ, and any policy π, its corresponding robust value function $V^{\pi,\sigma}$ (cf. (2.25)) satisfies*

$$\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) - \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq \frac{1}{\gamma\max\{1-\gamma,\sigma\}}.$$

Next, we introduce the following lemma, whose proof is postponed in Appendix D.2.3.2.

Lemma 43. *Consider an MDP with transition kernel matrix P and reward function $0\leq r\leq 1$. For any policy π and its associated state transition matrix $P_\pi \coloneqq \Pi^\pi P$ and value function $0\leq V^{\pi,P}\leq\frac{1}{1-\gamma}$ (cf. (2.20)), one has*

$$(I-\gamma P_\pi)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \leq \sqrt{\frac{8(\max_s V^{\pi,P}(s) - \min_s V^{\pi,P}(s))}{\gamma^2(1-\gamma)^2}}1.$$

D.2.2 Proof of Theorem 10

The main proof idea of Theorem 10 is similar to that of Agarwal et al. (2020a) and Li et al. (2023c) while the argument needs essential adjustments in order to adapt to the robustness setting. Before

proceeding, applying Lemma 41 yields that for any $\varepsilon_{\mathsf{opt}} > 0$, as long as $T \geq \log(\frac{1}{(1-\gamma)\varepsilon_{\mathsf{opt}}})$, one has

$$\left\|\widehat{V}^{\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty \leq \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}, \tag{D.31}$$

allowing us to justify the more general statement in Remark 4. To control the performance gap $\left\|V^{\star,\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$, the proof is divided into several key steps.

Step 1: decomposing the error. Recall the optimal robust policy $\pi^\star$ w.r.t. $\mathcal{M}_{\mathsf{rob}}$ and the optimal robust policy $\widehat{\pi}^\star$, the optimal robust value function $\widehat{V}^{\star,\sigma}$ (resp. robust value function $\widehat{Q}^{\pi,\sigma}$) w.r.t. $\widehat{\mathcal{M}}_{\mathsf{rob}}$. The term of interest $V^{\star,\sigma} - V^{\widehat{\pi},\sigma}$ can be decomposed as

$$\begin{aligned}
V^{\star,\sigma} - V^{\widehat{\pi},\sigma} &= \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right) + \left(\widehat{V}^{\pi^\star,\sigma} - \widehat{V}^{\widehat{\pi}^\star,\sigma}\right) + \left(\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right) + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right) \\
&\overset{\text{(i)}}{\leq} \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right) + \left(\widehat{V}^{\widehat{\pi}^\star,\sigma} - \widehat{V}^{\widehat{\pi},\sigma}\right) + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right) \\
&\overset{\text{(ii)}}{\leq} \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\right) + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1 + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right)
\end{aligned} \tag{D.32}$$

where (i) holds by $\widehat{V}^{\pi^\star,\sigma} - \widehat{V}^{\widehat{\pi}^\star,\sigma} \leq 0$ since $\widehat{\pi}^\star$ is the robust optimal policy for $\widehat{\mathcal{M}}_{\mathsf{rob}}$, and (ii) comes from the fact in (D.31).

To control the two important terms in (D.32), we first consider a more general term $\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}$ for any policy π. Towards this, plugging in (D.24) yields

$$\begin{aligned}
\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} &= r_\pi + \gamma\underline{\widehat{P}}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} - \left(r_\pi + \gamma\underline{P}^{\pi,V}V^{\pi,\sigma}\right) \\
&= \left(\gamma\underline{\widehat{P}}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} - \gamma\underline{P}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma}\right) + \left(\gamma\underline{P}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} - \gamma\underline{P}^{\pi,V}V^{\pi,\sigma}\right) \\
&\overset{\text{(i)}}{\leq} \gamma\left(\underline{P}^{\pi,V}\widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,V}V^{\pi,\sigma}\right) + \left(\gamma\underline{\widehat{P}}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} - \gamma\underline{P}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma}\right),
\end{aligned}$$

where (i) holds by observing

$$\underline{P}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} \leq \underline{P}^{\pi,V}\widehat{V}^{\pi,\sigma}$$

due to the optimality of $\underline{P}^{\pi,\widehat{V}}$ (cf. (D.3)). Rearranging terms leads to

$$\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} \leq \gamma\left(I - \gamma\underline{P}^{\pi,V}\right)^{-1}\left(\underline{\widehat{P}}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma}\right). \tag{D.33}$$

Similarly, we can also deduce

$$\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma} = r_\pi + \gamma\underline{\widehat{P}}^{\pi,\widehat{V}}\widehat{V}^{\pi,\sigma} - \left(r_\pi + \gamma\underline{P}^{\pi,V}V^{\pi,\sigma}\right)$$

$$
\begin{aligned}
&= \left(\gamma \widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) + \left(\gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \\
&\geq \gamma \left(\underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} V^{\pi,\sigma}\right) + \left(\gamma \widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \gamma \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) \\
&\geq \gamma \left(I - \gamma \underline{P}^{\pi,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right). \qquad \text{(D.34)}
\end{aligned}
$$

Combining (D.33) and (D.34), we arrive at

$$
\begin{aligned}
\left\|\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \underline{P}^{\pi,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \underline{P}^{\pi,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma} - \underline{P}^{\pi,\widehat{V}} \widehat{V}^{\pi,\sigma}\right) \right\|_\infty \Big\}. \qquad \text{(D.35)}
\end{aligned}
$$

By decomposing the error in a symmetric way, we can similarly obtain

$$
\begin{aligned}
\left\|\widehat{V}^{\pi,\sigma} - V^{\pi,\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right) \right\|_\infty \Big\}. \qquad \text{(D.36)}
\end{aligned}
$$

With the above facts in mind, we are ready to control the two terms $\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty$ and $\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty$ in (D.32) separately. More specifically, taking $\pi = \pi^\star$, applying (D.36) leads to

$$
\begin{aligned}
\left\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi^\star,V}\right)^{-1} \left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \widehat{\underline{P}}^{\pi^\star,\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right) \right\|_\infty \Big\}. \qquad \text{(D.37)}
\end{aligned}
$$

Similarly, taking $\pi = \widehat{\pi}$, applying (D.35) leads to

$$
\begin{aligned}
\left\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\right\|_\infty \leq \gamma \max\Big\{ &\left\| \left(I - \gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1} \left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right) \right\|_\infty, \\
&\left\| \left(I - \gamma \underline{P}^{\widehat{\pi},V}\right)^{-1} \left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right) \right\|_\infty \Big\}. \qquad \text{(D.38)}
\end{aligned}
$$

Step 2: controlling $\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\|_\infty$: bounding the first term in (D.37). To control the two terms in (D.37), we first introduce the following lemma whose proof is postponed to Appendix D.2.3.3.

Lemma 44. *Consider any $\delta \in (0,1)$. Setting $N \geq \log(\frac{18SAN}{\delta})$, with probability at least $1-\delta$, one has*

$$
\left|\widehat{\underline{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\right| \leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} 1
$$

$$\leq 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}1, \tag{D.39}$$

where $\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})$ *is defined in* (D.2).

Armed with the above lemma, now we control the first term on the right hand side of (D.37) as follows:

$$\begin{aligned}
&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right)\\
&\overset{\text{(i)}}{\leq}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left\|\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right\|_\infty\\
&\overset{\text{(ii)}}{\leq}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}+\frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}1\right)\\
&\leq\frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}1+\underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})}}_{=:\mathcal{C}_1}\\
&\quad+\underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})-\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|}}_{=:\mathcal{C}_2}\\
&\quad+\underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}-\sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)}_{=:\mathcal{C}_3},
\end{aligned}\tag{D.40}$$

where (i) holds by $\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\geq 0$, (ii) follows from Lemma 44, and the last inequality arise from

$$\begin{aligned}
&\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}=\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}-\sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)+\sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\\
&\leq\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}-\sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)+\sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})-\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})\right|}+\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})}
\end{aligned}$$

by applying the triangle inequality.

To continue, observing that each row of $\underline{\widehat{P}}^{\pi^\star,V}$ is a probability distribution obeying that the sum is 1, we arrive at

$$\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}1=\left(I+\sum_{t=1}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\right)1=\frac{1}{1-\gamma}1. \tag{D.41}$$

Armed with this fact, we shall control the other three terms $\mathcal{C}_1,\mathcal{C}_2,\mathcal{C}_3$ in (D.40) separately.

- Consider $\mathcal{C}_1$. We first introduce the following lemma, whose proof is postponed to Appendix D.2.3.4.

  Lemma 45. *Consider any $\delta \in (0,1)$. With probability at least $1-\delta$, one has*

$$\left(I - \gamma \underline{\widehat{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star, V}}(V^{\star,\sigma})} \leq 4\sqrt{\frac{\left(1 + \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right)}{\gamma^3 (1-\gamma)^2 \max\{1-\gamma, \sigma\}}} 1 \leq 4\sqrt{\frac{\left(1 + \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right)}{\gamma^3 (1-\gamma)^3}} 1.$$

  Applying Lemma 45 and inserting back to (D.40) leads to

$$\begin{aligned}
\mathcal{C}_1 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \left(I - \gamma \underline{\widehat{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star, V}}(V^{\star,\sigma})} \\
&\leq 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3 (1-\gamma)^2 \max\{1-\gamma, \sigma\} N} \left(1 + \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right)} 1. \qquad \text{(D.42)}
\end{aligned}$$

- Consider $\mathcal{C}_2$. First, denote $V' := V^{\star,\sigma} - \min_{s' \in \mathcal{S}} V^{\star,\sigma}(s') 1$, by Lemma 42, it follows that

$$0 \leq V' \leq \frac{1}{\gamma \max\{1-\gamma, \sigma\}} 1. \qquad \text{(D.43)}$$

  Then, we have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $P_{s,a} \in \Delta(\mathcal{S})$, and $\widetilde{P}_{s,a} \in \mathcal{U}^\sigma(P_{s,a})$:

$$\begin{aligned}
\left|\mathsf{Var}_{\widetilde{P}_{s,a}}(V^{\star,\sigma}) - \mathsf{Var}_{P_{s,a}}(V^{\star,\sigma})\right| &= \left|\mathsf{Var}_{\widetilde{P}_{s,a}}(V') - \mathsf{Var}_{P_{s,a}}(V')\right| \\
&\leq \left\|\widetilde{P}_{s,a} - P_{s,a}\right\|_1 \left\|V'\right\|_\infty^2 \\
&\leq \frac{2\sigma}{\gamma^2 (\max\{1-\gamma, \sigma\})^2} 1 \leq \frac{2}{\gamma^2 \max\{1-\gamma, \sigma\}} 1. \qquad \text{(D.44)}
\end{aligned}$$

  Applying the above relation we obtain

$$\begin{aligned}
\mathcal{C}_2 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \left(I - \gamma \underline{\widehat{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\left|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star, V}}(V^{\star,\sigma})\right|} \\
&= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \left(I - \gamma \underline{\widehat{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\left|\Pi^{\pi^\star}\left(\mathrm{Var}_{\widehat{P}^0}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{P}^{\pi^\star, V}}(V^{\star,\sigma})\right)\right|} \\
&\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \left(I - \gamma \underline{\widehat{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\left\|\mathrm{Var}_{\widehat{P}^0}(V^{\star,\sigma}) - \mathrm{Var}_{\widehat{P}^{\pi^\star, V}}(V^{\star,\sigma})\right\|_\infty 1} \\
&\leq 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \left(I - \gamma \underline{\widehat{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\frac{2}{\gamma^2 \max\{1-\gamma, \sigma\}}} 1 = 2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2 (1-\gamma)^2 \max\{1-\gamma, \sigma\} N}} 1, \qquad \text{(D.45)}
\end{aligned}$$

where the last equality uses $\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}1=\frac{1}{1-\gamma}$ (cf. (D.41)).

- Consider $\mathcal{C}_3$. The following lemma plays an important role.

Lemma 46. *(Panaganti and Kalathil, 2022, Lemma 6) Consider any $\delta\in(0,1)$. For any fixed policy π and fixed value vector $V\in\mathbb{R}^S$, one has with probability at least $1-\delta$,*

$$\left|\sqrt{\mathrm{Var}_{\widehat{P}^\pi}(V)}-\sqrt{\mathrm{Var}_{P^\pi}(V)}\right|\le\sqrt{\frac{2\|V\|_\infty^2\log(\frac{2SA}{\delta})}{N}}1.$$

Applying Lemma 46 with $\pi=\pi^\star$ and $V=V^{\star,\sigma}$ leads to

$$\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}-\sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\le\sqrt{\frac{2\|V^{\star,\sigma}\|_\infty^2\log(\frac{2SA}{\delta})}{N}}1,$$

which can be plugged in (D.40) to verify

$$\begin{aligned}\mathcal{C}_3&=2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}-\sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)\\&\le\frac{4}{(1-\gamma)}\frac{\log(\frac{SAN}{\delta})\|V^{\star,\sigma}\|_\infty}{N}1\le\frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1,\end{aligned}\tag{D.46}$$

where the last line uses $\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}1=\frac{1}{1-\gamma}$ (cf. (D.41)).

Finally, inserting the results of $\mathcal{C}_1$ in (D.42), $\mathcal{C}_2$ in (D.45), $\mathcal{C}_3$ in (D.46), and (D.41) back into (D.40) gives

$$\begin{aligned}&\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right)\le 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}\left(1+\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\right)}1\\&\quad+2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1+\frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)^2}1\\&\le 10\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}\left(1+\sqrt{\frac{\log(\frac{SAN}{\delta})}{(1-\gamma)^2N}}\right)}1+\frac{5\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1\\&\le 160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\frac{5\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1,\end{aligned}\tag{D.47}$$

where the last inequality holds by the fact $\gamma\ge\frac{1}{4}$ and letting $N\ge\frac{\log(\frac{SAN}{\delta})}{(1-\gamma)^2}$.

Step 3: controlling $\|\widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma}\|_\infty$: bounding the second term in (D.37). To proceed, applying Lemma 44 on the second term of the right hand side of (D.37) leads to

$$
\begin{aligned}
&\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\Big(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}\Big) \\
&\le 2\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\left(\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}1\right) \\
&\le \frac{2\log(\frac{18SAN}{\delta})}{N(1-\gamma)}\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}1 + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(\widehat{V}^{\pi^\star,\sigma})}}_{=:\mathcal{C}_4} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\left(\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma})}\right)}_{=:\mathcal{C}_5} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\left(\sqrt{\Big|\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma}) - \mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\star,\sigma})\Big|}\right)}_{=:\mathcal{C}_6} \\
&\quad + \underbrace{2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\left(\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})} - \sqrt{\mathrm{Var}_{\widehat{P}^{\pi^\star}}(V^{\star,\sigma})}\right)}_{=:\mathcal{C}_7},
\end{aligned}
\tag{D.48}
$$

where the last term $\widetilde{\mathcal{C}}_3$ can be controlled the same as $\mathcal{C}_3$ in (D.46). We now bound the above terms separately.

- Applying Lemma 43 with $P = \widehat{P}^{\pi^\star,\widehat{V}}$, $\pi = \pi^\star$ and taking $V = \widehat{V}^{\pi^\star,\sigma}$ which obeys $\widehat{V}^{\pi^\star,\sigma} = r_{\pi^\star} + \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\widehat{V}^{\pi^\star,\sigma}$, and in view of (D.41), the term $\mathcal{C}_4$ in (D.48) can be controlled as follows:

$$
\begin{aligned}
\mathcal{C}_4 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\Big(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(\widehat{V}^{\pi^\star,\sigma})} \\
&\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\frac{8(\max_s \widehat{V}^{\pi^\star,\sigma}(s) - \min_s \widehat{V}^{\pi^\star,\sigma}(s))}{\gamma^2(1-\gamma)^2}}1 \\
&\le 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1,
\end{aligned}
\tag{D.49}
$$

  where the last inequality holds by applying Lemma 42.

- To continue, considering $\mathcal{C}_5$, we directly observe that (in view of (D.41))

$$\begin{aligned}\mathcal{C}_5 &= 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\Big(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,\widehat{V}}}(V^{\pi^\star,\sigma}-\widehat{V}^{\pi^\star,\sigma})}\\ &\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1. \end{aligned} \tag{D.50}$$

- Then, it is easily verified that $\mathcal{C}_6$ can be controlled similarly as (D.45) as follows:

$$\mathcal{C}_6 \le 2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \tag{D.51}$$

- Similarly, $\mathcal{C}_7$ can be controlled the same as (D.46) shown below:

$$\mathcal{C}_7 \le \frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1. \tag{D.52}$$

Combining the results in (D.49), (D.50), (D.51), and (D.52) and inserting back to (D.48) leads to

$$\begin{aligned}\Big(I-\gamma\underline{\widehat{P}}^{\pi^\star,\widehat{V}}\Big)^{-1}\Big(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\Big) &\le 8\sqrt{\frac{\log(\frac{18SAN}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1\\ &\quad+2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1+2\sqrt{\frac{2\log(\frac{18SAN}{\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1\\ &\le 80\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty 1+\frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}1, \end{aligned} \tag{D.53}$$

where the last inequality follows from the assumption $\gamma \ge \frac{1}{4}$.

Finally, inserting (D.47) and (D.53) back to (D.37) yields

$$\begin{aligned}\left\|\widehat{V}^{\pi^\star,\sigma}-V^{\pi^\star,\sigma}\right\|_\infty &\le \max\Bigg\{160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{5\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N},\\ &\qquad 80\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}\left\|V^{\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\right\|_\infty+\frac{4\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}\Bigg\}\\ &\le 160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{8\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}, \end{aligned} \tag{D.54}$$

where the last inequality holds by taking $N \ge \frac{16\log(\frac{SAN}{\delta})}{(1-\gamma)^2}$.

Step 4: controlling $\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\|_\infty$: bounding the first term in (D.38). Unlike the earlier term, we now need to deal with the complicated statistical dependency between $\widehat{\pi}$ and the empirical RMDP. To begin with, we introduce the following lemma which controls the main term on the right hand side of (D.38), which is proved in Appendix D.2.3.5.

Lemma 47. *Consider any $\delta \in (0,1)$. Taking $N \geq \log\left(\frac{54SAN^2}{(1-\gamma)\delta}\right)$, with probability at least $1-\delta$, one has*

$$\begin{aligned}\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| &\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})}1 + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}1 + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1 \\ &\leq 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2 N}}1 + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1. \end{aligned} \tag{D.55}$$

With Lemma 47 in hand, we have

$$\begin{aligned}&\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right) \\ &\overset{\text{(i)}}{\leq} \left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma} - \underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| \\ &\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})} + \left(I - \gamma P_Q^{\widehat{\pi},V^{\widehat{\pi}}}\right)^{-1}\left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1 \\ &\overset{\text{(ii)}}{\leq} \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1 + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}}_{=:\mathcal{D}_1} \\ &\quad + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})\right|}}_{=:\mathcal{D}_2} \\ &\quad + \underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})\right|}}_{=:\mathcal{D}_3}, \end{aligned} \tag{D.56}$$

where (i) and (ii) hold by the fact that each row of $(1-\gamma)\left(I - \gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}$ is a probability vector that falls into $\Delta(\mathcal{S})$.

The remainder of the proof will focus on controlling the three terms in (D.56) separately.

- For $\mathcal{D}_1$, we introduce the following lemma, whose proof is postponed to D.2.3.6.

  Lemma 48. *Consider any $\delta \in (0,1)$. Taking $N \geq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2}$ and $\varepsilon_{\mathsf{opt}} \leq \frac{1-\gamma}{\gamma}$, one has with*

probability at least $1-\delta$,

$$\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \le 6\sqrt{\frac{1}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1 \le 6\sqrt{\frac{1}{(1-\gamma)^3\gamma^2}}1.$$

Applying Lemma 48 and (D.41) to (D.56) leads to

$$\begin{aligned}\mathcal{D}_1 &= 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \\ &\le 12\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \end{aligned} \tag{D.57}$$

- Applying Lemma 37 with $\|\widehat{V}^{\star,\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\|_\infty \le \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}$ and (D.41), $\mathcal{D}_2$ can be controlled as

$$\begin{aligned}\mathcal{D}_2 &= 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})\right|} \\ &\le 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\frac{\sqrt{\gamma\varepsilon_{\mathsf{opt}}}}{1-\gamma} \le 4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4N}}1. \end{aligned} \tag{D.58}$$

- $\mathcal{D}_3$ can be controlled similar to $\mathcal{C}_2$ in (D.45) as follows:

$$\begin{aligned}\mathcal{D}_3 &= 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\star,\sigma})\right|} \\ &\le 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\frac{1}{\gamma^2\max\{1-\gamma,\sigma\}}}1 \le 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1 \end{aligned} \tag{D.59}$$

Finally, summing up the results in (D.57), (D.58), and (D.59) and inserting them back to (D.56) yields: taking $N \ge \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2}$ and $\varepsilon_{\mathsf{opt}} \le \frac{1-\gamma}{\gamma}$, with probability at least $1-\delta$,

$$\begin{aligned}\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right) &\le \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1 \\ &+12\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4N}}1+4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1\end{aligned}$$

$$\leq 16\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}1, \tag{D.60}$$

where the last inequality holds by taking $\varepsilon_{\mathsf{opt}} \leq \min\left\{\frac{1-\gamma}{\gamma}, \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}\right\} = \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$.

Step 5: controlling $\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\|_\infty$: bounding the second term in (D.38). Towards this, applying Lemma 47 leads to

$$\begin{aligned}
&\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right) \leq \left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| \\
&\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\sqrt{\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})}+\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1 \\
&\leq \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1+\underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(V^{\widehat{\pi},\sigma})}}_{=:\mathcal{D}_4} \\
&\quad+\underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma})}}_{=:\mathcal{D}_5} \\
&\quad+\underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma})\right|}}_{=:\mathcal{D}_6} \\
&\quad+\underbrace{2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\left|\mathrm{Var}_{P^{\widehat{\pi}}}(\widehat{V}^{\star,\sigma})-\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\star,\sigma})\right|}}_{=:\mathcal{D}_7}.
\end{aligned} \tag{D.61}$$

We shall bound each of the terms separately.

- Applying Lemma 43 with $P = \underline{P}^{\widehat{\pi},V}$, $\pi = \widehat{\pi}$, and taking $V = V^{\widehat{\pi},\sigma}$ which obeys $V^{\widehat{\pi},\sigma} = r_{\widehat{\pi}} + \gamma\underline{P}^{\widehat{\pi},V}V^{\widehat{\pi},\sigma}$, the term $\mathcal{D}_4$ can be controlled similar to (D.49) as follows:

$$\mathcal{D}_4 \leq 8\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \tag{D.62}$$

- For $\mathcal{D}_5$, it is observed that

$$\mathcal{D}_5 = 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},V}}(\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma})}$$

$$\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2 N}}\left\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty 1. \tag{D.63}$$

- Next, observing that $\mathcal{D}_6$ and $\mathcal{D}_7$ are almost the same as the terms $\mathcal{D}_2$ (controlled in (D.58)) and $\mathcal{D}_3$ (controlled in (D.59)) in (D.56), it is easily verified that they can be controlled as follows

$$\mathcal{D}_6 \leq 4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4 N}}1, \qquad \mathcal{D}_7 \leq 4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1. \tag{D.64}$$

Then inserting the results in (D.62), (D.63), and (D.64) back to (D.61) leads to

$$\begin{aligned}
&\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right) \leq \left(\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}\right)1+8\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1\\
&\quad+2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2 N}}\left\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty 1+4\sqrt{\frac{\gamma\varepsilon_{\mathsf{opt}}\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^4 N}}1+4\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^2(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1\\
&\leq 12\sqrt{\frac{2\log(\frac{8SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}1+\frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}1+2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2 N}}\left\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty 1,
\end{aligned} \tag{D.65}$$

where the last inequality holds by letting $\varepsilon_{\mathsf{opt}} \leq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$, which directly satisfies $\varepsilon_{\mathsf{opt}} \leq \frac{1-\gamma}{\gamma}$ by letting $N \geq \frac{\log(\frac{54SAN^2}{\delta})}{1-\gamma}$.

Finally, inserting (D.60) and (D.65) back to (D.38) yields: taking $\varepsilon_{\mathsf{opt}} \leq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$ and $N \geq \frac{16\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2}$, with probability at least $1-\delta$, one has

$$\begin{aligned}
\left\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\right\|_\infty &\leq \max\Bigg\{16\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2},\\
&\qquad 12\sqrt{\frac{2\log(\frac{8SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{14\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}+2\sqrt{\frac{\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2 N}}\left\|V^{\widehat{\pi},\sigma}-\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty\Bigg\}\\
&\leq 24\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}}+\frac{28\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2}.
\end{aligned} \tag{D.66}$$

Step 6: summing up the results. Summing up the results in (D.54) and (D.66) and inserting back to (D.32) complete the proof as follows: taking $\varepsilon_{\mathsf{opt}} \leq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma N}$ and $N \geq \frac{16\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2}$, with probability at least $1-\delta$,

$$\begin{aligned}
\big\|V^{\star,\sigma} - V^{\widehat{\pi},\sigma}\big\|_\infty &\leq \big\|V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma}\big\|_\infty + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + \big\|\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma}\big\|_\infty \\
&\leq \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} + 160\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{8\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N} \\
&\quad + 24\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{28\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} \\
&\leq 184\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}N}} + \frac{36\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)^2} \\
&\leq 1508\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2\max\{1-\gamma,\sigma\}N}},
\end{aligned} \tag{D.67}$$

where the last inequality holds by $\gamma \geq \frac{1}{4}$ and $N \geq \frac{16\log(\frac{54SAN^2}{\delta})}{(1-\gamma)^2}$.

D.2.3 Proof of the auxiliary lemmas

D.2.3.1 Proof of Lemma 42

To begin, note that there at leasts exist one state s_0 for any $V^{\pi,\sigma}$ such that $V^{\pi,\sigma}(s_0) = \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s)$. With this in mind, for any policy π, one has by the definition in (2.25) and the Bellman's equation (2.27a),

$$\begin{aligned}
\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) &= \max_{s\in\mathcal{S}} \mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\Big[r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma}\Big] \\
&\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big(1 + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma}\Big),
\end{aligned}$$

where the second line holds since the reward function $r(s,a) \in [0,1]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. To continue, note that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, there exists some $\widetilde{P}_{s,a} \in \mathbb{R}^S$ constructed by reducing the values of some elements of $P_{s,a}$ to obey $P_{s,a} \geq \widetilde{P}_{s,a} \geq 0$ and $\sum_{s'}(P_{s,a}(s') - \widetilde{P}_{s,a}(s')) = \sigma$. This implies $\widetilde{P}_{s,a} + \sigma e_{s_0}^\top \in \mathcal{U}^\sigma(P_{s,a})$, where e_{s_0} is the standard basis vector supported on s_0, since $\frac{1}{2}\|\widetilde{P}_{s,a} + \sigma e_{s_0}^\top - P_{s,a}\|_1 \leq \frac{1}{2}\|\widetilde{P}_{s,a} - P_{s,a}\|_1 + \frac{\sigma}{2} = \sigma$. Consequently,

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{s,a})} \mathcal{P}V^{\pi,\sigma} \leq \Big(\widetilde{P}_{s,a} + \sigma e_{s_0}^\top\Big) V^{\pi,\sigma} \leq \big\|\widetilde{P}_{s,a}\big\|_1 \big\|V^{\pi,\sigma}\big\|_\infty + \sigma V^{\pi,\sigma}(s_0)$$

$$\leq (1-\sigma)\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \sigma \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s), \tag{D.68}$$

where the second inequality holds by $\left\|\widetilde{P}_{s,a}\right\|_1 = \sum_{s'} \widetilde{P}_{s,a}(s') = -\sum_{s'}\left(P_{s,a}(s') - \widetilde{P}_{s,a}(s')\right) + \sum_{s'} P_{s,a}(s') = 1-\sigma$. Plugging this back to the previous relation gives

$$\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq 1 + \gamma(1-\sigma)\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) + \gamma\sigma \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s),$$

which, by rearranging terms, immediately yields

$$\begin{aligned}\max_{s\in\mathcal{S}} V^{\pi,\sigma}(s) &\leq \frac{1+\gamma\sigma\min_{s\in\mathcal{S}} V^{\pi,\sigma}(s)}{1-\gamma(1-\sigma)} \\ &\leq \frac{1}{(1-\gamma)+\gamma\sigma} + \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s) \leq \frac{1}{\gamma\max\{1-\gamma,\sigma\}} + \min_{s\in\mathcal{S}} V^{\pi,\sigma}(s).\end{aligned}$$

D.2.3.2 Proof of Lemma 43

Observing that each row of P_π belongs to $\Delta(S)$, it can be directly verified that each row of $(1-\gamma)(I-\gamma P_\pi)^{-1}$ falls into $\Delta(S)$. As a result,

$$\begin{aligned}(I-\gamma P_\pi)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} &= \frac{1}{1-\gamma}(1-\gamma)(I-\gamma P_\pi)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \\ &\overset{\text{(i)}}{\leq} \frac{1}{1-\gamma}\sqrt{(1-\gamma)(I-\gamma P_\pi)^{-1}\mathrm{Var}_{P_\pi}(V^{\pi,P})} \\ &= \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t (P_\pi)^t \mathrm{Var}_{P_\pi}(V^{\pi,P})},\end{aligned} \tag{D.69}$$

where (i) holds by Jensen's inequality.

To continue, denoting the minimum value of V as $V_{\min} = \min_{s\in\mathcal{S}} V^{\pi,P}(s)$ and $V' \coloneqq V^{\pi,P} - V_{\min}1$. We control $\mathrm{Var}_{P_\pi}(V^{\pi,P})$ as follows:

$$\begin{aligned}&\mathrm{Var}_{P_\pi}(V^{\pi,P}) \\ &\overset{\text{(i)}}{=} \mathrm{Var}_{P_\pi}(V') = P_\pi\left(V'\circ V'\right) - \left(P_\pi V'\right)\circ\left(P_\pi V'\right) \\ &\overset{\text{(ii)}}{=} P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma^2}\left(V' - r_\pi + (1-\gamma)V_{\min}1\right)\circ\left(V' - r_\pi + (1-\gamma)V_{\min}1\right) \\ &= P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma^2}V'\circ V' + \frac{2}{\gamma^2}V'\circ(r_\pi - (1-\gamma)V_{\min}1) - \frac{1}{\gamma^2}(r_\pi - (1-\gamma)V_{\min}1)\circ(r_\pi - (1-\gamma)V_{\min}1) \\ &\leq P_\pi\left(V'\circ V'\right) - \frac{1}{\gamma}V'\circ V' + \frac{2}{\gamma^2}\|V'\|_\infty 1,\end{aligned} \tag{D.70}$$

where (i) holds by the fact that $\mathrm{Var}_{P_\pi}(V^{\pi,P} - b1) = \mathrm{Var}_{P_\pi}(V^{\pi,P})$ for any scalar b and $V^{\pi,P}\in\mathbb{R}^S$, (ii) follows from $V' = r_\pi + \gamma P_\pi V^{\pi,P} - V_{\min}1 = r_\pi - (1-\gamma)V_{\min}1 + \gamma P_\pi V'$, and the last line arises

from $\frac{1}{\gamma^2}V'\circ V' \geq \frac{1}{\gamma}V'\circ V'$ and $\|r_\pi - (1-\gamma)V_{\min}1\|_\infty \leq 1$. Plugging (D.70) back to (D.69) leads to

$$
\begin{aligned}
&(I-\gamma P_\pi)^{-1}\sqrt{\mathrm{Var}_{P_\pi}(V^{\pi,P})} \leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t (P_\pi)^t\left(P_\pi\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'+\frac{2}{\gamma^2}\|V'\|_\infty 1\right)} \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t (P_\pi)^t\left(P_\pi\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'\right)\right|}+\sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t (P_\pi)^t\frac{2}{\gamma^2}\|V'\|_\infty 1} \\
&\leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\left(\sum_{t=0}^{\infty}\gamma^t (P_\pi)^{t+1}-\sum_{t=0}^{\infty}\gamma^{t-1}(P_\pi)^t\right)\left(V'\circ V'\right)\right|}+\sqrt{\frac{2\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2 1}{\gamma(1-\gamma)}}+\sqrt{\frac{2\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}} \\
&\leq \sqrt{\frac{8\|V'\|_\infty 1}{\gamma^2(1-\gamma)^2}},
\end{aligned}
\tag{D.71}
$$

where (i) holds by the triangle inequality, (ii) holds by following recursion, and the last inequality holds by $\|V'\|_\infty \leq \frac{1}{1-\gamma}$.

D.2.3.3 Proof of Lemma 44

Step 1: controlling the point-wise concentration. We first consider a more general term w.r.t. any fixed (independent from $\widehat{P}^0$) value vector V obeying $0 \leq V \leq \frac{1}{1-\gamma}1$ and any policy π. Invoking Lemma 39 leads to that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$
\begin{aligned}
\left|\widehat{P}_{s,a}^{\pi,V}V - P_{s,a}^{\pi,V}V\right| &\leq \Big|\max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{\widehat{P}_{s,a}^0[V]_\alpha-\sigma\left(\alpha-\min_{s'}[V]_\alpha(s')\right)\right\} \\
&\qquad -\max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{P_{s,a}^0[V]_\alpha-w\sigma\left(\alpha-\min_{s'}[V]_\alpha(s')\right)\right\}\Big| \\
&\leq \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\underbrace{\left|\left(P_{s,a}^0-\widehat{P}_{s,a}^0\right)[V]_\alpha\right|}_{=:g_{s,a}(\alpha,V)},
\end{aligned}
\tag{D.72}
$$

where the last inequality holds by that the maximum operator is 1-Lipschitz.

Then for a fixed α and any vector V that is independent with $\widehat{P}^0$, using the Bernstein's inequality, one has with probability at least $1-\delta$,

$$
\begin{aligned}
g_{s,a}(\alpha,V) = \left|\left(P_{s,a}^0-\widehat{P}_{s,a}^0\right)[V]_\alpha\right| &\leq \sqrt{\frac{2\log(\frac{2}{\delta})}{N}}\sqrt{\mathrm{Var}_{P_{s,a}^0}([V]_\alpha)}+\frac{2\log(\frac{2}{\delta})}{3N(1-\gamma)} \\
&\leq \sqrt{\frac{2\log(\frac{2}{\delta})}{N}}\sqrt{\mathrm{Var}_{P_{s,a}^0}(V)}+\frac{2\log(\frac{2}{\delta})}{3N(1-\gamma)}.
\end{aligned}
\tag{D.73}
$$

Step 2: deriving the uniform concentration. To obtain the union bound, we first notice that $g_{s,a}(\alpha, V)$ is 1-Lipschitz w.r.t. α for any V obeying $\|V\|_\infty \le \frac{1}{1-\gamma}$. In addition, we can construct an ε_1-net N_{ε_1} over $[0, \frac{1}{1-\gamma}]$ whose size satisfies $|N_{\varepsilon_1}| \le \frac{3}{\varepsilon_1(1-\gamma)}$ (Vershynin, 2018). By the union bound and (D.73), it holds with probability at least $1 - \frac{\delta}{SA}$ that for all $\alpha \in N_{\varepsilon_1}$,

$$g_{s,a}(\alpha, V) \le \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)}. \tag{D.74}$$

Combined with (D.72), it yields that,

$$\begin{aligned}
\left|\widehat{P}^{\pi,V}_{s,a} V - P^{\pi,V}_{s,a} V\right| &\le \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| \\
&\overset{\text{(i)}}{\le} \varepsilon_1 + \sup_{\alpha \in N_{\varepsilon_1}} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| \\
&\overset{\text{(ii)}}{\le} \varepsilon_1 + \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)} \qquad \text{(D.75)} \\
&\overset{\text{(iii)}}{\le} \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{N(1-\gamma)} \\
&\overset{\text{(iv)}}{\le} 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V)} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} \qquad \text{(D.76)} \\
&\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \|V\|_\infty + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} \\
&\le 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}} \qquad \text{(D.77)}
\end{aligned}$$

where (i) follows from that the optimal $\alpha^\star$ falls into the ε_1-ball centered around some point inside N_{ε_1} and $g_{s,a}(\alpha, V)$ is 1-Lipschitz, (ii) holds by (D.74), (iii) arises from taking $\varepsilon_1 = \frac{\log(\frac{2SA|N_{\varepsilon_1}|}{\delta})}{3N(1-\gamma)}$, (iv) is verified by $|N_{\varepsilon_1}| \le \frac{3}{\varepsilon_1(1-\gamma)} \le 9N$, and the last inequality is due to the fact $\|V^{\star,\sigma}\|_\infty \le \frac{1}{1-\gamma}$ and letting $N \ge \log(\frac{18SAN}{\delta})$.

To continue, applying (D.76) and (D.77) with $\pi = \pi^\star$ and $V = V^{\star,\sigma}$ (independent with $\widehat{P}^0$) and taking the union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$ gives that with probability at least $1-\delta$, it holds simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned}
\left|\widehat{P}^{\pi^\star,V}_{s,a} V^{\star,\sigma} - P^{\pi^\star,V}_{s,a} V^{\star,\sigma}\right| &\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(V^{\star,\sigma})} + \frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)} \\
&\le 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}. \qquad \text{(D.78)}
\end{aligned}$$

By converting (D.78) to the matrix form, one has with probability at least $1-\delta$,

$$\begin{aligned}\left|\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right| &\le 2\sqrt{\frac{\log(\frac{18SAN}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^{\pi^\star}}(V^{\star,\sigma})}+\frac{\log(\frac{18SAN}{\delta})}{N(1-\gamma)}1\\ &\le 3\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}1. \end{aligned} \tag{D.79}$$

D.2.3.4 Proof of Lemma 45

Following the same argument as (D.69), it follows

$$\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})}=\sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{\widehat{P}}^{\pi^\star,V}\right)^t\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})}. \tag{D.80}$$

To continue, we first focus on controlling $\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})$. Towards this, denoting the minimum value of $V^{\star,\sigma}$ as $V_{\min}:=\min_{s\in\mathcal{S}}V^{\star,\sigma}(s)$ and $V':=V^{\star,\sigma}-V_{\min}1$, we arrive at (see the robust Bellman's consistency equation in (D.24))

$$\begin{aligned}V'=V^{\star,\sigma}-V_{\min}1&=r_{\pi^\star}+\gamma\underline{P}^{\pi^\star,V}V^{\star,\sigma}-V_{\min}1\\ &=r_{\pi^\star}+\gamma\underline{\widehat{P}}^{\pi^\star,V}V^{\star,\sigma}+\gamma\left(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\right)V^{\star,\sigma}-V_{\min}1\\ &=r_{\pi^\star}-(1-\gamma)V_{\min}1+\gamma\underline{\widehat{P}}^{\pi^\star,V}V'+\gamma\left(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\right)V^{\star,\sigma}\\ &=r'_{\pi^\star}+\gamma\underline{\widehat{P}}^{\pi^\star,V}V'+\gamma\left(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\right)V^{\star,\sigma}, \end{aligned} \tag{D.81}$$

where the last line holds by letting $r'_{\pi^\star}:=r_{\pi^\star}-(1-\gamma)V_{\min}1\le r_{\pi^\star}$. With the above fact in hand, we control $\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})$ as follows:

$$\begin{aligned}\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V^{\star,\sigma})&\overset{\text{(i)}}{=}\mathrm{Var}_{\underline{\widehat{P}}^{\pi^\star,V}}(V')=\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\left(\underline{\widehat{P}}^{\pi^\star,V}V'\right)\circ\left(\underline{\widehat{P}}^{\pi^\star,V}V'\right)\\ &\overset{\text{(ii)}}{=}\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma^2}\left(V'-r'_{\pi^\star}-\gamma\left(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\right)V^{\star,\sigma}\right)^{\circ 2}\\ &=\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma^2}V'\circ V'+\frac{2}{\gamma^2}V'\circ\left(r'_{\pi^\star}+\gamma\left(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\right)V^{\star,\sigma}\right)\\ &\quad-\frac{1}{\gamma^2}\left(r'_{\pi^\star}+\gamma\left(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\right)V^{\star,\sigma}\right)^{\circ 2}\\ &\overset{\text{(iii)}}{\le}\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'+\frac{2}{\gamma^2}\|V'\|_\infty 1+\frac{2}{\gamma}\|V'\|_\infty\left|\left(\underline{P}^{\pi^\star,V}-\underline{\widehat{P}}^{\pi^\star,V}\right)V^{\star,\sigma}\right| \end{aligned} \tag{D.82}$$

$$\le\underline{\widehat{P}}^{\pi^\star,V}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'+\frac{2}{\gamma^2}\|V'\|_\infty 1+\frac{6}{\gamma}\|V'\|_\infty\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2N}}1, \tag{D.83}$$

where (i) holds by the fact that $\mathrm{Var}_{P_\pi}(V - b1) = \mathrm{Var}_{P_\pi}(V)$ for any scalar b and $V \in \mathbb{R}^S$, (ii) follows from (D.81), (iii) arises from $\frac{1}{\gamma^2} V' \circ V' \geq \frac{1}{\gamma} V' \circ V'$ and $-1 \leq r_{\pi^\star} - (1-\gamma) V_{\min} 1 = r'_{\pi^\star} \leq r_{\pi^\star} \leq 1$, and the last inequality holds by Lemma 44.

Plugging (D.83) into (D.80) leads to

$$
\begin{aligned}
&\left(I - \gamma \widehat{\underline{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star, V}}(V^{\star,\sigma})} \\
&\leq \sqrt{\frac{1}{1-\gamma}} \sqrt{\sum_{t=0}^{\infty} \gamma^t \left(\widehat{\underline{P}}^{\pi^\star, V}\right)^t \left(\widehat{\underline{P}}^{\pi^\star, V} (V' \circ V') - \frac{1}{\gamma} V' \circ V' + \frac{2}{\gamma^2} \|V'\|_\infty 1 + \frac{6}{\gamma} \|V'\|_\infty \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}} 1\right)} \\
&\overset{\text{(i)}}{\leq} \sqrt{\frac{1}{1-\gamma}} \sqrt{\left| \sum_{t=0}^{\infty} \gamma^t \left(\widehat{\underline{P}}^{\pi^\star, V}\right)^t \left(\widehat{\underline{P}}^{\pi^\star, V} (V' \circ V') - \frac{1}{\gamma} V' \circ V'\right) \right|} \\
&\quad + \sqrt{\frac{1}{1-\gamma}} \sqrt{\sum_{t=0}^{\infty} \gamma^t \left(\widehat{\underline{P}}^{\pi^\star, V}\right)^t \left(\frac{2}{\gamma^2} \|V'\|_\infty 1 + \frac{6}{\gamma} \|V'\|_\infty \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}} 1\right)} \\
&\leq \sqrt{\frac{1}{1-\gamma}} \sqrt{\left| \sum_{t=0}^{\infty} \gamma^t \left(\widehat{\underline{P}}^{\pi^\star, V}\right)^t \left[\widehat{\underline{P}}^{\pi^\star, V} (V' \circ V') - \frac{1}{\gamma} V' \circ V'\right] \right|} + \sqrt{\frac{\left(2 + 6\sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right) \|V'\|_\infty}{(1-\gamma)^2 \gamma^2}} 1,
\end{aligned}
\tag{D.84}
$$

where (i) holds by the triangle inequality. Therefore, the remainder of the proof shall focus on the first term, which follows

$$
\begin{aligned}
&\left| \sum_{t=0}^{\infty} \gamma^t \left(\widehat{\underline{P}}^{\pi^\star, V}\right)^t \left(\widehat{\underline{P}}^{\pi^\star, V} (V' \circ V') - \frac{1}{\gamma} V' \circ V'\right) \right| \\
&= \left| \left(\sum_{t=0}^{\infty} \gamma^t \left(\widehat{\underline{P}}^{\pi^\star, V}\right)^{t+1} - \sum_{t=0}^{\infty} \gamma^{t-1} \left(\widehat{\underline{P}}^{\pi^\star, V}\right)^t \right) (V' \circ V') \right| \leq \frac{1}{\gamma} \|V'\|_\infty^2 1
\end{aligned}
\tag{D.85}
$$

by recursion. Inserting (D.85) back to (D.84) leads to

$$
\begin{aligned}
&\left(I - \gamma \widehat{\underline{P}}^{\pi^\star, V}\right)^{-1} \sqrt{\mathrm{Var}_{\widehat{\underline{P}}^{\pi^\star, V}}(V^{\star,\sigma})} \\
&\leq \sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}} 1 + 3\sqrt{\frac{\left(1 + \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right) \|V'\|_\infty}{(1-\gamma)^2 \gamma^2}} 1 \\
&\leq 4\sqrt{\frac{\left(1 + \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right) \|V'\|_\infty}{(1-\gamma)^2 \gamma^2}} 1 \leq 4\sqrt{\frac{\left(1 + \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right)}{\gamma^3 (1-\gamma)^2 \max\{1-\gamma, \sigma\}}} 1 \leq 4\sqrt{\frac{\left(1 + \sqrt{\frac{\log(\frac{18SAN}{\delta})}{(1-\gamma)^2 N}}\right)}{\gamma^3 (1-\gamma)^3}} 1,
\end{aligned}
\tag{D.86}
$$

where the penultimate inequality follows from applying Lemma 42 with $P = P^0$ and $\pi = \pi^\star$:

$$\|V'\|_\infty = \max_{s\in\mathcal{S}} V^{\star,\sigma}(s) - \min_{s\in\mathcal{S}} V^{\star,\sigma}(s) \le \frac{1}{\gamma \max\{1-\gamma, \sigma\}}.$$

D.2.3.5 Proof of Lemma 47

To begin with, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, invoking the results in (D.72), we have

$$\begin{aligned}
\left|\widehat{P}^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma} - P^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}\right| &\le \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right| \\
&\overset{\text{(i)}}{\le} \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left(\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha - \left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)\right|\right) \\
&\le \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left(\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + \left\|P^0_{s,a} - \widehat{P}^0_{s,a}\right\|_1 \left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha - \left[\widehat{V}^{\star,\sigma}\right]_\alpha\right\|_\infty\right) \\
&\overset{\text{(ii)}}{\le} \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + 2\left\|\widehat{V}^{\widehat{\pi},\sigma} - \widehat{V}^{\star,\sigma}\right\|_\infty \\
&\overset{\text{(iii)}}{\le} \max_{\alpha\in[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right| + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma},
\end{aligned} \tag{D.87}$$

where (i) holds by the triangle inequality, and (ii) follows from $\left\|P^0_{s,a} - \widehat{P}^0_{s,a}\right\|_1 \le 2$ and $\left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha - \left[\widehat{V}^{\star,\sigma}\right]_\alpha\right\|_\infty \le \left\|\widehat{V}^{\widehat{\pi},\sigma} - \widehat{V}^{\star,\sigma}\right\|_\infty$, and (iii) follows from (D.31).

To control $\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right|$ in (D.87) for any given $\alpha \in \left[0, \frac{1}{1-\gamma}\right]$, and tame the dependency between $\widehat{V}^{\star,\sigma}$ and $\widehat{P}^0$, we resort to the following leave-one-out argument motivated by (Agarwal et al., 2020a; Li et al., 2022a; Shi and Chi, 2022). Specifically, we first construct a set of auxiliary RMDPs which simultaneously have the desired statistical independence between robust value functions and the estimated nominal transition kernel, and are minimally different from the original RMDPs under consideration. Then we control the term of interest associated with these auxiliary RMDPs and show the value is close to the target quantity for the desired RMDP. The process is divided into several steps as below.

Step 1: construction of auxiliary RMDPs with deterministic empirical nominal transitions. Recall that we target the empirical infinite-horizon robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}$ with the nominal transition kernel $\widehat{P}^0$. Towards this, we can construct an auxiliary robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ for each state s and any non-negative scalar $u \ge 0$, so that it is the same as $\widehat{\mathcal{M}}_{\mathsf{rob}}$ except for the transition properties in state s. In particular, we define the nominal transition kernel and reward function of $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ as $P^{s,u}$ and $r^{s,u}$, which are expressed as follows

$$\begin{cases}
P^{s,u}(s' \,|\, s,a) = \mathbb{1}(s'=s) & \text{for all } (s',a) \in \mathcal{S}\times\mathcal{A}, \\
P^{s,u}(\cdot \,|\, \widetilde{s},a) = \widehat{P}^0(\cdot \,|\, \widetilde{s},a) & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s} \ne s,
\end{cases} \tag{D.88}$$

and

$$\begin{cases} r^{s,u}(s,a) = u & \text{for all } a \in \mathcal{A}, \\ r^{s,u}(\widetilde{s},a) = r(\widetilde{s},a) & \text{for all } (\widetilde{s},a) \in \mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s} \neq s. \end{cases} \tag{D.89}$$

It is evident that the nominal transition probability at state s of the auxiliary $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$, i.e. it never leaves state s once entered. This useful property removes the randomness of $\widehat{P}_{s,a}^0$ for all $a \in \mathcal{A}$ in state s, which will be leveraged later.

Correspondingly, the robust Bellman operator $\widehat{\mathcal{T}}_{s,u}^{\sigma}(\cdot)$ associated with the RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ is defined as

$$\forall(\widetilde{s},a) \in \mathcal{S}\times\mathcal{A}: \quad \widehat{\mathcal{T}}_{s,u}^{\sigma}(Q)(\widetilde{s},a) = r^{s,u}(\widetilde{s},a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P_{\widetilde{s},a}^{s,u})} \mathcal{P}V, \qquad \text{with } V(\widetilde{s}) = \max_a Q(\widetilde{s},a). \tag{D.90}$$

Step 2: fixed-point equivalence between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and the auxiliary RMDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$. Recall that $\widehat{Q}^{\star,\sigma}$ is the unique fixed point of $\widehat{\mathcal{T}}^{\sigma}(\cdot)$ with the corresponding robust value $\widehat{V}^{\star,\sigma}$. We assert that the corresponding robust value function $\widehat{V}_{s,u^\star}^{\star,\sigma}$ obtained from the fixed point of $\widehat{\mathcal{T}}_{s,u}^{\sigma}(\cdot)$ aligns with the robust value function $\widehat{V}^{\star,\sigma}$ derived from $\widehat{\mathcal{T}}^{\sigma}(\cdot)$, as long as we choose u in the following manner:

$$u^\star := u^\star(s) = \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{D.91}$$

where e_s is the s-th standard basis vector in $\mathbb{R}^S$. Towards verifying this, we shall break our arguments in two different cases.

- **For state s**: One has for any $a \in \mathcal{A}$:

$$\begin{aligned} r^{s,u^\star}(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P_{s,a}^{s,u^\star})} \mathcal{P}\widehat{V}^{\star,\sigma} &= u^\star + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} \\ &= \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma} = \widehat{V}^{\star,\sigma}(s), \end{aligned} \tag{D.92}$$

  where the first equality follows from the definition of $P_{s,a}^{s,u^\star}$ in (E.164), and the second equality follows from plugging in the definition of $u^\star$ in (E.169).

- **For state $s' \neq s$**: It is easily verified that for all $a \in \mathcal{A}$,

$$r^{s,u^\star}(s',a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P_{s',a}^{s,u^\star})} \mathcal{P}\widehat{V}^{\star,\sigma} = r(s',a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}_{s',a}^0)} \mathcal{P}\widehat{V}^{\star,\sigma}$$

$$= \widehat{\mathcal{T}}^{\sigma}(\widehat{Q}^{\star,\sigma})(s',a) = \widehat{Q}^{\star,\sigma}(s',a), \tag{D.93}$$

where the first equality follows from the definitions in (E.165) and (E.164), and the last line arises from the definition of the robust Bellman operator in (6.7), and that $\widehat{Q}^{\star,\sigma}$ is the fixed point of $\widehat{\mathcal{T}}^{\sigma}(\cdot)$ (see Lemma 38).

Combining the facts in the above two cases, we establish that there exists a fixed point $\widehat{Q}^{\star,\sigma}_{s,u^\star}$ of the operator $\widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\cdot)$ by taking

$$\begin{cases} \widehat{Q}^{\star,\sigma}_{s,u^\star}(s,a) = \widehat{V}^{\star,\sigma}(s) & \text{for all } a \in \mathcal{A}, \\ \widehat{Q}^{\star,\sigma}_{s,u^\star}(s',a) = \widehat{Q}^{\star,\sigma}(s',a) & \text{for all } s' \neq s \text{ and } a \in \mathcal{A}. \end{cases} \tag{D.94}$$

Consequently, we confirm the existence of a fixed point of the operator $\widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\cdot)$. In addition, its corresponding value function $\widehat{V}^{\star,\sigma}_{s,u^\star}$ also coincides with $\widehat{V}^{\star,\sigma}$. Note that the corresponding facts between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ in Step 1 and step 2 holds in fact for any uncertainty set.

Step 3: building an ε-net for all reward values u. It is easily verified that

$$0 \leq u^\star \leq \widehat{V}^{\star,\sigma}(s) \leq \frac{1}{1-\gamma}. \tag{D.95}$$

We can construct a N_{ε_2}-net over the interval $\left[0, \frac{1}{1-\gamma}\right]$, where the size is bounded by $|N_{\varepsilon_2}| \leq \frac{3}{\varepsilon_2(1-\gamma)}$ (Vershynin, 2018). Following the same arguments in the proof of Lemma 38, we can demonstrate that for each $u \in N_{\varepsilon_2}$, there exists a unique fixed point $\widehat{Q}^{\star,\sigma}_{s,u}$ of the operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$, which satisfies $0 \leq \widehat{Q}^{\star,\sigma}_{s,u} \leq \frac{1}{1-\gamma} \cdot 1$. Consequently, the corresponding robust value function also satisfies $\left\|\widehat{V}^{\star,\sigma}_{s,u}\right\|_\infty \leq \frac{1}{1-\gamma}$.

By the definitions in (E.164) and (E.165), we observe that for all $u \in N_{\varepsilon_2}$, $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is statistically independent from $\widehat{P}^0_{s,a}$. This independence indicates that $[\widehat{V}^{\star,\sigma}_{s,u}]_\alpha$ and $\widehat{P}^0_{s,a}$ are independent for a fixed α. With this in mind, invoking the fact in (D.76) and (D.77) and taking the union bound over all $(s,a,\alpha) \in \mathcal{S} \times \mathcal{A} \times N_{\varepsilon_1}$, $u \in N_{\varepsilon_2}$ yields that, with probability at least $1-\delta$, it holds for all $(s,a,u) \in \mathcal{S} \times \mathcal{A} \times N_{\varepsilon_2}$ that

$$\begin{aligned} \max_{\alpha \in [0,1/(1-\gamma)]} \left| \left(P^0_{s,a} - \widehat{P}^0_{s,a}\right) [\widehat{V}^{\star,\sigma}_{s,u}]_\alpha \right| &\leq \varepsilon_2 + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}} \sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,u})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\ &\leq \varepsilon_2 + 3\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}}, \end{aligned} \tag{D.96}$$

where the last inequality holds by the fact $\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,u}) \leq \|\widehat{V}^{\star,\sigma}_{s,u}\|_\infty \leq \frac{1}{1-\gamma}$ and letting $N \geq \log\left(\frac{18SAN|N_{\varepsilon_2}|}{\delta}\right)$.

Step 4: uniform concentration. Recalling that $u^\star \in \left[0, \frac{1}{1-\gamma}\right]$ (see (D.95)), we can always find some $\overline{u} \in N_{\varepsilon_2}$ such that $|\overline{u} - u^\star| \le \varepsilon_2$. Consequently, plugging in the operator $\widehat{\mathcal{T}}^\sigma_{s,u}(\cdot)$ in (D.90) yields

$$\forall Q \in \mathbb{R}^{SA}: \quad \left\|\widehat{\mathcal{T}}^\sigma_{s,\overline{u}}(Q) - \widehat{\mathcal{T}}^\sigma_{s,u^\star}(Q)\right\|_\infty = |\overline{u} - u^\star| \le \varepsilon_2$$

With this in mind, we observe that the fixed points of $\widehat{\mathcal{T}}^\sigma_{s,\overline{u}}(\cdot)$ and $\widehat{\mathcal{T}}^\sigma_{s,u^\star}(\cdot)$ obey

$$\begin{aligned}
\left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty &= \left\|\widehat{\mathcal{T}}^\sigma_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,\overline{u}}) - \widehat{\mathcal{T}}^\sigma_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\
&\le \left\|\widehat{\mathcal{T}}^\sigma_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,\overline{u}}) - \widehat{\mathcal{T}}^\sigma_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty + \left\|\widehat{\mathcal{T}}^\sigma_{s,\overline{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star}) - \widehat{\mathcal{T}}^\sigma_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty \\
&\le \gamma \left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty + \varepsilon_2,
\end{aligned}$$

where the last inequality holds by the fact that $\widehat{\mathcal{T}}^\sigma_{s,u}(\cdot)$ is a γ-contraction. It directly indicates that

$$\left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{\varepsilon_2}{(1-\gamma)} \quad \text{and} \quad \left\|\widehat{V}^{\star,\sigma}_{s,\overline{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \left\|\widehat{Q}^{\star,\sigma}_{s,\overline{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty \le \frac{\varepsilon_2}{(1-\gamma)}. \tag{D.97}$$

Armed with the above facts, to control the first term in (D.87), invoking the identity $\widehat{V}^{\star,\sigma} = \widehat{V}^{\star,\sigma}_{s,u^\star}$ established in Step 2 gives that: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
&\max_{\alpha \in [\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right| \\
&\le \max_{\alpha \in [0,1/(1-\gamma)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right| = \max_{\alpha \in [0,1/(1-\gamma)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right| \\
&\overset{\text{(i)}}{\le} \max_{\alpha \in [0,1/(1-\gamma)]} \left\{\left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha\right| + \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)\left([\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha - [\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right)\right|\right\} \\
&\overset{\text{(ii)}}{\le} \max_{\alpha \in [0,1/(1-\gamma)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha\right| + \frac{2\varepsilon_2}{(1-\gamma)} \\
&\overset{\text{(iii)}}{\le} \frac{2\varepsilon_2}{(1-\gamma)} + \varepsilon_2 + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,u})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\
&\le \frac{3\varepsilon_2}{(1-\gamma)} + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} \\
&\quad + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\left|\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}) - \mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma}_{s,\overline{u}})\right|} \\
&\overset{\text{(iv)}}{\le} \frac{3\varepsilon_2}{(1-\gamma)} + 2\sqrt{\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{3N(1-\gamma)} + 2\sqrt{\frac{2\varepsilon_2\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N(1-\gamma)^2}} \\
&\le 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})} + \frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}
\end{aligned} \tag{D.98}$$

$$\leq 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2 N}}, \tag{D.99}$$

where (i) holds by the triangle inequality, (ii) arises from (the last inequality holds by (E.181))

$$\begin{aligned}\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)\left([\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha-[\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right)\right| &\leq \left\|P^0_{s,a}-\widehat{P}^0_{s,a}\right\|_1\left\|[\widehat{V}^{\star,\sigma}_{s,\overline{u}}]_\alpha-[\widehat{V}^{\star,\sigma}_{s,u^\star}]_\alpha\right\|_\infty \\ &\leq 2\left\|\widehat{V}^{\star,\sigma}_{s,\overline{u}}-\widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_\infty \leq \frac{2\varepsilon_2}{(1-\gamma)},\end{aligned} \tag{D.100}$$

(iii) follows from (D.96), (iv) can be verified by applying Lemma 37 with (E.181). Here, the penultimate inequality holds by letting $\varepsilon_2=\frac{\log(\frac{18SAN|N_{\varepsilon_2}|}{\delta})}{N}$, which leads to $|N_{\varepsilon_2}|\leq\frac{3}{\varepsilon_2(1-\gamma)}\leq\frac{3N}{1-\gamma}$, and the last inequality holds by the fact $\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})\leq\|\widehat{V}^{\star,\sigma}\|_\infty\leq\frac{1}{1-\gamma}$ and letting $N\geq\log\left(\frac{54SAN^2}{(1-\gamma)\delta}\right)$.

Step 5: finishing up. Inserting (D.98) and (D.99) back into (D.87) and combining with (D.99) give that with probability at least $1-\delta$,

$$\begin{aligned}\left|\widehat{P}^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}-P^{\widehat{\pi},\widehat{V}}_{s,a}\widehat{V}^{\widehat{\pi},\sigma}\right| &\leq \max_{\alpha\in[\min_s\widehat{V}^{\widehat{\pi},\sigma}(s),\max_s\widehat{V}^{\widehat{\pi},\sigma}(s)]}\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right|+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} \\ &\leq \max_{\alpha\in[0,1/(1-\gamma)]}\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)[\widehat{V}^{\star,\sigma}]_\alpha\right|+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} \\ &\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})}+\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma} \\ &\leq 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\end{aligned} \tag{D.101}$$

holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

Finally, we complete the proof by compiling everything into the matrix form as follows:

$$\begin{aligned}\left|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right| &\leq 2\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N}}\sqrt{\mathrm{Var}_{P^0_{s,a}}(\widehat{V}^{\star,\sigma})}1+\frac{8\log(\frac{54SAN^2}{(1-\gamma)\delta})}{N(1-\gamma)}1+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1 \\ &\leq 10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}1+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}1.\end{aligned} \tag{D.102}$$

D.2.3.6 Proof of Lemma 48

The proof can be achieved by directly applying the same routine as Appendix D.2.3.4. Towards this, similar to (D.80), we arrive at

$$\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} \leq \sqrt{\frac{1}{1-\gamma}}\sqrt{\sum_{t=0}^{\infty}\gamma^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\right)^t \mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}. \tag{D.103}$$

To control $\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})$, we denote the minimum value of $\widehat{V}^{\widehat{\pi},\sigma}$ as $V_{\min}=\min_{s\in\mathcal{S}}\widehat{V}^{\widehat{\pi},\sigma}(s)$ and $V' \coloneqq \widehat{V}^{\widehat{\pi},\sigma}-V_{\min}1$. By the same argument as (D.82), we arrive at

$$\begin{aligned}
\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma}) &\leq \underline{P}^{\widehat{\pi},\widehat{V}}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'+\frac{2}{\gamma^2}\|V'\|_\infty 1+\frac{2}{\gamma}\|V'\|_\infty\left|\left(\widehat{\underline{P}}^{\widehat{\pi},\widehat{V}}-\underline{P}^{\widehat{\pi},\widehat{V}}\right)\widehat{V}^{\widehat{\pi},\sigma}\right| \\
&\leq \underline{P}^{\widehat{\pi},\widehat{V}}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'+\frac{2}{\gamma^2}\|V'\|_\infty 1+\frac{2}{\gamma}\|V'\|_\infty\left(10\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)1,
\end{aligned} \tag{D.104}$$

where the last inequality makes use of Lemma 47. Plugging (D.104) back into (D.103) leads to

$$\begin{aligned}
\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})} &\overset{\text{(i)}}{\leq} \sqrt{\frac{1}{1-\gamma}}\sqrt{\left|\sum_{t=0}^{\infty}\gamma^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\right)^t\left(\underline{P}^{\widehat{\pi},\widehat{V}}\left(V'\circ V'\right)-\frac{1}{\gamma}V'\circ V'\right)\right|} \\
&\quad+\sqrt{\frac{1}{(1-\gamma)^2\gamma^2}\left(2+20\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)\|V'\|_\infty 1} \\
&\overset{\text{(ii)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}1}+\sqrt{\frac{\left(2+20\sqrt{\frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}\right)\|V'\|_\infty}{(1-\gamma)^2\gamma^2}1} \\
&\overset{\text{(iii)}}{\leq} \sqrt{\frac{\|V'\|_\infty^2}{\gamma(1-\gamma)}1}+\sqrt{\frac{24\|V'\|_\infty}{(1-\gamma)^2\gamma^2}1} \leq 6\sqrt{\frac{\|V'\|_\infty}{(1-\gamma)^2\gamma^2}1},
\end{aligned} \tag{D.105}$$

where (i) arises from following the routine of (D.84), (ii) holds by repeating the argument of (D.85), (iii) follows by taking $N \geq \frac{\log(\frac{54SAN^2}{(1-\gamma)\delta})}{(1-\gamma)^2}$ and $\varepsilon_{\mathsf{opt}} \leq \frac{1-\gamma}{\gamma}$, and the last inequality holds by $\|V'\|_\infty \leq \|V^{\star,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$.

Finally, applying Lemma 42 with $P=\widehat{P}^0$ and $\pi=\widehat{\pi}$ yields

$$\|V'\|_\infty \leq \max_{s\in\mathcal{S}}\widehat{V}^{\widehat{\pi},\sigma}(s)-\min_{s\in\mathcal{S}}\widehat{V}^{\widehat{\pi},\sigma}(s) \leq \frac{1}{\gamma\max\{1-\gamma,\sigma\}},$$

which can be inserted into (D.105) and gives

$$\left(I-\gamma \underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\sqrt{\mathrm{Var}_{\underline{P}^{\widehat{\pi},\widehat{V}}}(\widehat{V}^{\widehat{\pi},\sigma})}\leq 6\sqrt{\frac{1}{\gamma^3(1-\gamma)^2\max\{1-\gamma,\sigma\}}}1\leq 6\sqrt{\frac{1}{(1-\gamma)^3\gamma^2}}1.$$

D.3 Proof of the lower bound with TV distance: Theorem 11

To prove Theorem 11, we shall first construct some hard instances and then characterize the sample complexity requirements over these instances. Note that the hard instances for robust MDPs are different from those for standard MDPs, due to the asymmetric structure induced by the robust RL problem formulation to consider the worst-case performance. By constructing a new class of hard instances inspired by the asymmetric structure of the RMDP, we develop a new lower bound in Theorem 11 that is tighter than prior art (Yang et al., 2022).

D.3.1 Construction of the hard problem instances

Construction of two hard MDPs. Suppose there are two standard MDPs defined as below:

$$\left\{\mathcal{M}_\phi=\left(\mathcal{S},\mathcal{A},P^\phi,r,\gamma\right)\mid\phi=\{0,1\}\right\}.$$

Here, γ is the discount parameter, $\mathcal{S}=\{0,1,\ldots,S-1\}$ is the state space. Given any state $s\in\{2,3,\cdots,S-1\}$, the corresponding action space are $\mathcal{A}=\{0,1,2,\cdots,A-1\}$. While for states $s=0$ or $s=1$, the action space is only $\mathcal{A}'=\{0,1\}$. For any $\phi\in\{0,1\}$, the transition kernel P^ϕ of the constructed MDP $\mathcal{M}_\phi$ is defined as

$$P^\phi(s'\mid s,a)=\begin{cases}p\mathbb{1}(s'=1)+(1-p)\mathbb{1}(s'=0) & \text{if}\quad (s,a)=(0,\phi)\\ q\mathbb{1}(s'=1)+(1-q)\mathbb{1}(s'=0) & \text{if}\quad (s,a)=(0,1-\phi)\\ \mathbb{1}(s'=1) & \text{if}\quad s\geq 1\end{cases},\tag{D.106}$$

where p and q are set to satisfy

$$0\leq p\leq 1\quad\text{ and }\quad 0\leq q=p-\Delta\tag{D.107}$$

for some p and $\Delta>0$ that shall be introduced later. The above transition kernel P^ϕ implies that state 1 is an absorbing state, namely, the MDP will always stay after it arrives at 1.

Then, we define the reward function as

$$r(s,a)=\begin{cases}1 & \text{if } s=1\\ 0 & \text{otherwise}\end{cases}.\tag{D.108}$$

Additionally, we choose the following initial state distribution:

$$\varphi(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{D.109}$$

Here, the constructed two instances are set with different probability transition from state 0 with reward 0 but not state 1 with reward 1 (which were used in standard MDPs (Li et al., 2022a)), yielding a larger gap between the value functions of the two instances.

Uncertainty set of the transition kernels. Recalling the uncertainty set assumed throughout this subchapter is defined as $\mathcal{U}^\sigma(P^\phi)$ with TV distance:

$$\mathcal{U}^\sigma(P) \coloneqq \mathcal{U}^\sigma_{\mathsf{TV}}(P) = \otimes\, \mathcal{U}^\sigma_{\mathsf{TV}}(P_{s,a}), \qquad \mathcal{U}^\sigma_{\mathsf{TV}}(P_{s,a}) \coloneqq \Big\{ P'_{s,a} \in \Delta(\mathcal{S}) : \frac{1}{2} \left\| P'_{s,a} - P_{s,a} \right\|_1 \le \sigma \Big\}, \tag{D.110}$$

where $P^\phi_{s,a} \coloneqq P^\phi(\cdot \,|\, s, a)$ is defined similar to (2.24). In addition, without loss of generality, we recall the radius $\sigma \in (0, 1 - c_0]$ with $0 < c_0 < 1$. With the uncertainty level in hand, taking $c_1 \coloneqq \frac{c_0}{2}$, p and Δ which determines the instances obey

$$p = (1 + c_1) \max\{1 - \gamma, \sigma\} \qquad \text{and} \qquad \Delta \le c_1 \max\{1 - \gamma, \sigma\}, \tag{D.111}$$

which ensure $0 \le p \le 1$ as follows:

$$(1 + c_1)\, \sigma \le 1 - c_0 + c_1 \sigma \le 1 - \frac{c_0}{2} < 1, \qquad (1 + c_1)\,(1 - \gamma) \le \frac{3}{2}(1 - \gamma) \le \frac{3}{4} < 1. \tag{D.112}$$

Consequently, applying (E.190) directly leads to

$$p \ge q \ge \max\{1 - \gamma, \sigma\}. \tag{D.113}$$

To continue, for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we denote the infimum probability of moving to the next state s' associated with any perturbed transition kernel $P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})$ as

$$\underline{P}^\phi(s' \,|\, s, a) \coloneqq \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})} P(s' \,|\, s, a) = \max\{P(s' \,|\, s, a) - \sigma, 0\}, \tag{D.114}$$

where the last equation can be easily verified by the definition of $\mathcal{U}^\sigma(P^\phi)$ in (D.110). As shall be seen, the transition from state 0 to state 1 plays an important role in the analysis, for convenience, we denote

$$\underline{p} \coloneqq \underline{P}^\phi(1 \,|\, 0, \phi) = p - \sigma, \qquad \underline{q} \coloneqq \underline{P}^\phi(1 \,|\, 0, 1 - \phi) = q - \sigma, \tag{D.115}$$

which follows from the fact that $p \geq q \geq \sigma$ in (E.192).

Robust value functions and robust optimal policies. To proceed, we are ready to derive the corresponding robust value functions, identify the optimal policies, and characterize the optimal values. For any MDP $\mathcal{M}_\phi$ with the above uncertainty set, we denote $\pi_\phi^\star$ as the optimal policy, and the robust value function of any policy π (resp. the optimal policy $\pi_\phi^\star$) as $V_\phi^{\pi,\sigma}$ (resp. $V_\phi^{\star,\sigma}$). Then, we introduce the following lemma which describes some important properties of the robust (optimal) value functions and optimal policies. The proof is postponed to Appendix D.3.3.1.

Lemma 49. *For any $\phi = \{0, 1\}$ and any policy π, the robust value function obeys*

$$V_\phi^{\pi,\sigma}(0) = \frac{\gamma\big(z_\phi^\pi - \sigma\big)}{(1-\gamma)\left(1 + \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma(1-\sigma)}\right)(1 - \gamma\,(1-\sigma))}, \tag{D.116}$$

where z_ϕ^π is defined as

$$z_\phi^\pi \coloneqq p\pi(\phi\,|\,0) + q\pi(1-\phi\,|\,0). \tag{D.117}$$

In addition, the robust optimal value functions and the robust optimal policies satisfy

$$V_\phi^{\star,\sigma}(0) = \frac{\gamma\,(p-\sigma)}{(1-\gamma)\left(1 + \frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)(1 - \gamma\,(1-\sigma))}, \tag{D.118a}$$

$$\pi_\phi^\star(\phi\,|\,s) = 1, \qquad \text{for } s \in \mathcal{S}. \tag{D.118b}$$

D.3.2 Establishing the minimax lower bound

Note that our goal is to control the quantity w.r.t. any policy estimator $\widehat{\pi}$ based on the chosen initial distribution φ in (E.197) and the dataset consisting of N samples over each state-action pair generated from the nominal transition kernel P^ϕ, which gives

$$\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \right\rangle = V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0).$$

Step 1: converting the goal to estimate ϕ. We make the following useful claim which shall be verified in Appendix D.3.3.2: With $\varepsilon \leq \frac{c_1}{32(1-\gamma)}$, letting

$$\Delta = 32(1-\gamma)\max\{1-\gamma, \sigma\}\varepsilon \leq c_1 \max\{1-\gamma, \sigma\} \tag{D.119}$$

which satisfies (D.111), it leads to that for any policy $\widehat{\pi}$,

$$\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \right\rangle \geq 2\varepsilon\big(1 - \widehat{\pi}(\phi\,|\,0)\big). \tag{D.120}$$

With this connection established between the policy $\widehat{\pi}$ and its sub-optimality gap as depicted in (D.120), we can now proceed to build an estimate for ϕ. Here, we denote $\mathbb{P}_\phi$ as the probability distribution when the MDP is $\mathcal{M}_\phi$, where ϕ can take on values in the set $\{0, 1\}$.

Let's assume momentarily that an estimated policy $\widehat{\pi}$ achieves

$$\mathbb{P}_\phi\big\{\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma}\big\rangle \le \varepsilon\big\} \ge \frac{7}{8}, \tag{D.121}$$

then in view of (D.120), we necessarily have $\widehat{\pi}(\phi \,|\, 0) \ge \frac{1}{2}$ with probability at least $\frac{7}{8}$. With this in mind, we are motivated to construct the following estimate $\widehat{\phi}$ for $\phi \in \{0, 1\}$:

$$\widehat{\phi} = \arg\max_{a \in \{0,1\}} \widehat{\pi}(a \,|\, 0), \tag{D.122}$$

which obeys

$$\mathbb{P}_\phi\big\{\widehat{\phi} = \phi\big\} \ge \mathbb{P}_\phi\big\{\widehat{\pi}(\phi \,|\, 0) > 1/2\big\} \ge \frac{7}{8}. \tag{D.123}$$

Subsequently, our aim is to demonstrate that (E.86) cannot occur without an adequate number of samples, which would in turn contradict (D.120).

Step 2: probability of error in testing two hypotheses. Equipped with the aforementioned groundwork, we can now delve into differentiating between the two hypotheses $\phi \in \{0, 1\}$. To achieve this, we consider the concept of minimax probability of error, defined as follows:

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max\big\{\mathbb{P}_0(\psi \neq 0),\, \mathbb{P}_1(\psi \neq 1)\big\}. \tag{D.124}$$

Here, the infimum is taken over all possible tests ψ constructed from the samples generated from the nominal transition kernel P^ϕ.

Moving forward, let us denote μ_ϕ (resp. $\mu_\phi(s)$) as the distribution of a sample tuple (s_i, a_i, s_i') under the nominal transition kernel P^ϕ associated with $\mathcal{M}_\phi$ and the samples are generated independently. Applying standard results from Tsybakov (2009, Theorem 2.2) and the additivity of the KL divergence (cf. Tsybakov (2009, Page 85)), we obtain

$$\begin{aligned} p_{\mathrm{e}} &\ge \frac{1}{4}\exp\Big(-NSA\mathsf{KL}\big(\mu_0 \,\|\, \mu_1\big)\Big) \\ &= \frac{1}{4}\exp\Big\{-N\Big(\mathsf{KL}\big(P^0(\cdot \,|\, 0, 0) \,\|\, P^1(\cdot \,|\, 0, 0)\big) + \mathsf{KL}\big(P^0(\cdot \,|\, 0, 1) \,\|\, P^1(\cdot \,|\, 0, 1)\big)\Big)\Big\}, \end{aligned} \tag{D.125}$$

where the last inequality holds by observing that

$$\mathsf{KL}\big(\mu_0 \,\|\, \mu_1\big) = \frac{1}{SA}\sum_{s,a,s'} \mathsf{KL}\big(P^0(s' \,|\, s, a) \,\|\, P^1(s' \,|\, s, a)\big)$$

$$= \frac{1}{SA} \sum_{a\in\{0,1\}} \mathsf{KL}\big(P^0(\cdot\,|\,0,a) \,\|\, P^1(\cdot\,|\,0,a)\big),$$

Here, the last equality holds by the fact that $P^0(\cdot\,|\,s,a)$ and $P^1(\cdot\,|\,s,a)$ only differ when $s=0$.

Now, our focus shifts towards bounding the terms involving the KL divergence in (E.88). Given $p \geq q \geq \max\{1-\gamma,\sigma\}$ (cf. (E.192)), applying Lemma 60 (cf. (E.11)) gives

$$\begin{aligned}\mathsf{KL}\big(P^0(\cdot\,|\,0,1) \,\|\, P^1(\cdot\,|\,0,1)\big) = \mathsf{KL}\left(p \,\|\, q\right) &\leq \frac{(p-q)^2}{(1-p)p} \overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{=} \frac{1024(1-\gamma)^2\max\{1-\gamma,\sigma\}^2\varepsilon^2}{p(1-p)} \\ &\leq \frac{1024(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}{1-p} \leq \frac{4096}{c_1}(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2,\end{aligned} \tag{D.126}$$

where (i) stems from the definition in (E.190), (ii) follows by the expression of Δ in (E.82), and the last inequality arises from $1-q \geq 1-p \geq \frac{c_0}{4}$ (see (D.112)).

Note that it can be shown that $\mathsf{KL}\big(P^0(\cdot\,|\,0,0) \,\|\, P^1(\cdot\,|\,0,0)\big)$ can be upper bounded in a same manner. Substituting (E.89) back into (E.88) demonstrates that: if the sample size is selected as

$$N \leq \frac{c_1 \log 2}{8192(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2}, \tag{D.127}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{-N\frac{8192}{c_1}(1-\gamma)^2\max\{1-\gamma,\sigma\}\varepsilon^2\right\} \geq \frac{1}{8}, \tag{D.128}$$

Step 3: putting the results together. Lastly, suppose that there exists an estimator $\widehat{\pi}$ such that

$$\mathbb{P}_0\big\{\big\langle\varphi, V_0^{\star,\sigma} - V_0^{\widehat{\pi},\sigma}\big\rangle > \varepsilon\big\} < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big\{\big\langle\varphi, V_1^{\star,\sigma} - V_1^{\widehat{\pi},\sigma}\big\rangle > \varepsilon\big\} < \frac{1}{8}.$$

According to Step 1, the estimator $\widehat{\phi}$ defined in (E.85) must satisfy

$$\mathbb{P}_0\big(\widehat{\phi} \neq 0\big) < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big(\widehat{\phi} \neq 1\big) < \frac{1}{8}.$$

However, this cannot occur under the sample size condition (E.90) to avoid contradiction with (E.91). Thus, we have completed the proof.

D.3.3 Proof of the auxiliary facts

D.3.3.1 Proof of Lemma 49

Deriving the robust value function over different states. For any $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, we first characterize the robust value function of any policy π over different states. Before proceeding, we denote the minimum of the robust value function over states as below:

$$V_{\phi,\min}^{\pi,\sigma} \coloneqq \min_{s\in\mathcal{S}} V_\phi^{\pi,\sigma}(s). \tag{D.129}$$

Clearly, there exists at least one state $s_{\phi,\min}^\pi$ that satisfies $V_\phi^{\pi,\sigma}(s_{\phi,\min}^\pi) = V_{\phi,\min}^{\pi,\sigma}$.

With this in mind, it is easily observed that for any policy π, the robust value function at state $s = 1$ obeys

$$\begin{aligned} V_\phi^{\pi,\sigma}(1) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,1)}\left[r(1,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{1,a}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\ &\overset{\text{(i)}}{=} 1 + \gamma\mathbb{E}_{a\sim\pi(\cdot\,|\,1)}\left[\underline{P}^\phi(1\,|\,1,a)V_\phi^{\pi,\sigma}(1)\right] + \gamma\sigma V_{\phi,\min}^{\pi,\sigma} \overset{\text{(ii)}}{=} 1 + \gamma(1-\sigma)V_\phi^{\pi,\sigma}(1) + \gamma\sigma V_{\phi,\min}^{\pi,\sigma}, \end{aligned} \tag{D.130}$$

where (i) holds by $r(1,a) = 1$ for all $a \in \mathcal{A}'$ and (E.201), and (ii) follows from $P^\phi(1\,|\,1,a) = 1$ for all $a \in \mathcal{A}'$.

Similarly, for any $s \in \{2,3,\cdots,S-1\}$, we have

$$\begin{aligned} V_\phi^{\pi,\sigma}(s) &= 0 + \gamma\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[\underline{P}^\phi(1\,|\,s,a)V_\phi^{\pi,\sigma}(1)\right] + \gamma\sigma V_{\phi,\min}^{\pi,\sigma} \\ &= \gamma\,(1-\sigma)\,V_\phi^{\pi,\sigma}(1) + \gamma\sigma V_{\phi,\min}^{\pi,\sigma}, \end{aligned} \tag{D.131}$$

since $r(s,a) = 0$ for all $s \in \{2,3,\cdots,S-1\}$ and the definition in (E.201).

Finally, we move onto compute $V_\phi^{\pi,\sigma}(0)$, the robust value function at state 0 associated with any policy π. First, it obeys

$$\begin{aligned} V_\phi^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,0)}\left[r(0,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,a}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\ &= 0 + \gamma\pi(\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma} + \gamma\pi(1-\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,1-\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma}. \end{aligned} \tag{D.132}$$

Recall the transition kernel defined in (E.189) and the fact about the uncertainty set over state 0 in (E.202), it is easily verified that the following probability vector $P_1 \in \Delta(\mathcal{S})$ obeys $P_1 \in \mathcal{U}^\sigma(P_{0,\phi}^\phi)$, which is defined as

$$P_1(0) = 1 - p + \sigma\,\mathbb{1}\left(0 = s_{\phi,\min}^\pi\right), \qquad P_1(1) = \underline{p} = p - \sigma,$$

$$P_1(s) = \sigma \mathbb{1}\left(s = s^{\pi}_{\phi,\min}\right), \qquad \forall s \in \{2, 3, \cdots, S-1\}, \tag{D.133}$$

where $\underline{p} = p - \sigma$ due to (E.202). Similarly, the following probability vector $P_2 \in \Delta(\mathcal{S})$ also falls into the uncertainty set $\mathcal{U}^{\sigma}(P^{\phi}_{0,1-\phi})$:

$$\begin{aligned} P_2(0) = 1 - q + \sigma \mathbb{1}\left(0 = s^{\pi}_{\phi,\min}\right), \qquad P_2(1) = \underline{q} = q - \sigma, \\ P_2(s) = \sigma \mathbb{1}\left(0 = s^{\pi}_{\phi,\min}\right) \qquad \forall s \in \{2, 3, \cdots, S-1\}. \end{aligned} \tag{D.134}$$

It is noticed that P_0 and P_1 defined above are the worst-case perturbations, since the probability mass at state 1 will be moved to the state with the least value. Plugging the above facts about $P_1 \in \mathcal{U}^{\sigma}(P^{\phi}_{0,\phi})$ and $P_2 \in \mathcal{U}^{\sigma}(P^{\phi}_{0,1-\phi})$ into (D.132), we arrive at

$$\begin{aligned} V^{\pi,\sigma}_{\phi}(0) &\leq \gamma\pi(\phi \,|\, 0) P_1 V^{\pi,\sigma}_{\phi} + \gamma\pi(1-\phi \,|\, 0) P_2 V^{\pi,\sigma}_{\phi} \\ &= \gamma\pi(\phi \,|\, 0)\Big[(p-\sigma) V^{\pi,\sigma}_{\phi}(1) + (1-p) V^{\pi,\sigma}_{\phi}(0) + \sigma V^{\pi,\sigma}_{\phi,\min}\Big] \\ &\quad + \gamma\pi(1-\phi \,|\, 0)\Big[(q-\sigma) V^{\pi,\sigma}_{\phi}(1) + (1-q) V^{\pi,\sigma}_{\phi}(0) + \sigma V^{\pi,\sigma}_{\phi,\min}\Big] \\ &\overset{\text{(i)}}{=} \gamma\left(z^{\pi}_{\phi} - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1 - z^{\pi}_{\phi}) V^{\pi,\sigma}_{\phi}(0), \end{aligned} \tag{D.135}$$

where the last equality holds by the definition of z^{π}_{ϕ} in (E.205). To continue, recursively applying (D.135) yields

$$\begin{aligned} V^{\pi,\sigma}_{\phi}(0) &\leq \gamma\left(z^{\pi}_{\phi} - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1 - z^{\pi}_{\phi})\Big[\gamma\left(z^{\pi}_{\phi} - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1 - z^{\pi}_{\phi}) V^{\pi,\sigma}_{\phi}(0)\Big] \\ &\overset{\text{(i)}}{\leq} \gamma\left(z^{\pi}_{\phi} - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1 - z^{\pi}_{\phi})\Big[\gamma z^{\pi}_{\phi} V^{\pi,\sigma}_{\phi}(1) + \gamma(1 - z^{\pi}_{\phi}) V^{\pi,\sigma}_{\phi}(0)\Big] \\ &\leq \ldots \\ &\leq \gamma\left(z^{\pi}_{\phi} - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma z^{\pi}_{\phi} \sum_{t=1}^{\infty} \gamma^t (1 - z^{\pi}_{\phi})^t V^{\pi,\sigma}_{\phi}(1) + \lim_{t\to\infty} \gamma^t (1 - z^{\pi}_{\phi})^t V^{\pi,\sigma}_{\phi}(0) \\ &\overset{\text{(ii)}}{\leq} \gamma\left(z^{\pi}_{\phi} - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1 - z^{\pi}_{\phi}) \frac{\gamma z^{\pi}_{\phi}}{1 - \gamma(1 - z^{\pi}_{\phi})} V^{\pi,\sigma}_{\phi}(1) + 0 \\ &< \gamma\left(z^{\pi}_{\phi} - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min} + \gamma(1 - z^{\pi}_{\phi}) V^{\pi,\sigma}_{\phi}(1) \\ &= \gamma\left(1 - \sigma\right) V^{\pi,\sigma}_{\phi}(1) + \gamma\sigma V^{\pi,\sigma}_{\phi,\min}, \end{aligned} \tag{D.136}$$

where (i) uses $V^{\pi,\sigma}_{\phi,\min} \leq V^{\pi,\sigma}_{\phi}(1)$, (ii) follows from $\gamma(1 - z^{\pi}_{\phi}) < 1$, and the penultimate line follows from the trivial fact that $\frac{\gamma z^{\pi}_{\phi}}{1-\gamma(1-z^{\pi}_{\phi})} < 1$.

Combining (D.130), (D.131), and (D.136), we have that for any policy π,

$$V^{\pi,\sigma}_{\phi}(0) = V^{\pi,\sigma}_{\phi,\min}, \tag{D.137}$$

which directly leads to

$$V_\phi^{\pi,\sigma}(1) = 1 + \gamma\,(1-\sigma)\,V_\phi^{\pi,\sigma}(1) + \gamma\sigma V_{\phi,\min}^{\pi,\sigma} = \frac{1+\gamma\sigma V_\phi^{\pi,\sigma}(0)}{1-\gamma\,(1-\sigma)}. \tag{D.138}$$

Let's now return to the characterization of $V_\phi^{\pi,\sigma}(0)$. In view of (D.137), the equality in (D.135) holds, and we have

$$\begin{aligned}
V_\phi^{\pi,\sigma}(0) &= \gamma\left(z_\phi^\pi - \sigma\right) V_\phi^{\pi,\sigma}(1) + \gamma\left(1 - z_\phi^\pi + \sigma\right) V_\phi^{\pi,\sigma}(0) \\
&\overset{\text{(i)}}{=} \gamma\left(z_\phi^\pi - \sigma\right) \frac{1+\gamma\sigma V_\phi^{\pi,\sigma}(0)}{1-\gamma\,(1-\sigma)} + \gamma\left(1 - z_\phi^\pi + \sigma\right) V_\phi^{\pi,\sigma}(0) \\
&= \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma\,(1-\sigma)} + \gamma\left(1 + \left(z_\phi^\pi - \sigma\right) \frac{\gamma\sigma - (1-\gamma\,(1-\sigma))}{1-\gamma\,(1-\sigma)}\right) V_\phi^{\pi,\sigma}(0) \\
&= \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma\,(1-\sigma)} + \gamma\left(1 - \frac{(1-\gamma)\left(z_\phi^\pi - \sigma\right)}{1-\gamma\,(1-\sigma)}\right) V_\phi^{\pi,\sigma}(0),
\end{aligned}$$

where (i) arises from (D.138). Solving this relation gives

$$V_\phi^{\pi,\sigma}(0) = \frac{\frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma(1-\sigma)}\right)}. \tag{D.139}$$

The optimal robust policy and optimal robust value function. We move on to characterize the robust optimal policy and its corresponding robust value function. To begin with, denoting

$$z \coloneqq \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma\,(1-\sigma)}, \tag{D.140}$$

we rewrite (D.139) as

$$V_\phi^{\pi,\sigma}(0) = \frac{z}{(1-\gamma)(1+z)} =: f(z).$$

Plugging in the fact that $z_\phi^\pi \geq q \geq \sigma > 0$ in (E.192), it follows that $z > 0$. So for any $z > 0$, the derivative of $f(z)$ w.r.t. z obeys

$$\frac{(1-\gamma)(1+z) - (1-\gamma)z}{(1-\gamma)^2(1+z)^2} = \frac{1}{(1-\gamma)(1+z)^2} > 0. \tag{D.141}$$

Observing that $f(z)$ is increasing in z, z is increasing in z_ϕ^π, and z_ϕ^π is also increasing in $\pi(\phi\,|\,0)$ (see the fact $p \geq q$ in (E.192)), the optimal policy in state 0 thus obeys

$$\pi_\phi^\star(\phi\,|\,0) = 1. \tag{D.142}$$

Considering that the action does not influence the state transition for all states $s > 0$, without loss of generality, we choose the robust optimal policy to obey

$$\forall s > 0: \quad \pi_\phi^\star(\phi\,|\,s) = 1. \tag{D.143}$$

Taking $\pi = \pi_\phi^\star$, we complete the proof by showing that the corresponding robust optimal robust value function at state 0 as follows:

$$V_\phi^{\star,\sigma}(0) = \frac{\frac{\gamma\left(z_\phi^{\pi^\star} - \sigma\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma\left(z_\phi^{\pi^\star} - \sigma\right)}{1-\gamma(1-\sigma)}\right)} = \frac{\frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)}. \tag{D.144}$$

D.3.3.2 Proof of the claim (D.120)

Plugging in the definition of φ, we arrive at that for any policy π,

$$\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \right\rangle = V_\phi^{\star,\sigma}(0) - V_\phi^{\pi,\sigma}(0) = \frac{\frac{\gamma\left(p - z_\phi^\pi\right)}{1-\gamma(1-\sigma)}}{(1-\gamma)\left(1 + \frac{\gamma(p-\sigma)}{1-\gamma(1-\sigma)}\right)\left(1 + \frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1-\gamma(1-\sigma)}\right)}, \tag{D.145}$$

which follows from applying (E.204) and basic calculus. Then, we proceed to control the above term in two cases separately in terms of the uncertainty level σ.

- When $\sigma \in (0, 1-\gamma]$. Then regarding the important terms in (D.145), we observe that

  $$1 - \gamma < 1 - \gamma\,(1-\sigma) \leq 1 - \gamma\,(1 - (1-\gamma)) = (1-\gamma)(1+\gamma) \leq 2(1-\gamma), \tag{D.146}$$

  which directly leads to

  $$\frac{\gamma\left(z_\phi^\pi - \sigma\right)}{1 - \gamma\,(1-\sigma)} \overset{\text{(i)}}{\leq} \frac{\gamma\,(p-\sigma)}{1-\gamma\,(1-\sigma)} \leq \frac{\gamma c_1 (1-\gamma)}{1-\gamma\,(1-\sigma)} \overset{\text{(ii)}}{<} c_1 \gamma, \tag{D.147}$$

  where (i) holds by $z_\phi^\pi < p$, and (ii) is due to (D.146). Inserting (D.146) and (D.147) back into (D.145), we arrive at

  $$\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \right\rangle \geq \frac{\frac{\gamma\left(p - z_\phi^\pi\right)}{2(1-\gamma)}}{(1-\gamma)(1+c_1\gamma)^2} \geq \frac{\gamma\left(p - z_\phi^\pi\right)}{8(1-\gamma)^2}$$

$$= \frac{\gamma\,(p-q)\,\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)^2} = \frac{\gamma\Delta\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)^2} \geq 2\varepsilon\big(1-\pi(\phi\,|\,0)\big), \quad \text{(D.148)}$$

where the last inequality holds by setting ($\gamma \geq 1/2$)

$$\Delta = 32(1-\gamma)^2\varepsilon. \quad \text{(D.149)}$$

Finally, it is easily verified that

$$\varepsilon \leq \frac{c_1}{32(1-\gamma)} \quad \Longrightarrow \quad \Delta \leq c_1(1-\gamma).$$

- When $\sigma \in (1-\gamma, 1-c_1]$. Regarding (D.145), we observe that

$$\gamma\sigma < 1-\gamma\,(1-\sigma) = 1-\gamma+\gamma\sigma \leq (1+\gamma)\sigma \leq 2\sigma, \quad \text{(D.150)}$$

which directly leads to

$$\frac{\gamma\big(z_\phi^\pi-\sigma\big)}{1-\gamma\,(1-\sigma)} \leq \frac{\gamma\,(p-\sigma)}{1-\gamma\,(1-\sigma)} \leq \frac{\gamma c_1\sigma}{1-\gamma\,(1-\sigma)} \overset{\text{(i)}}{<} c_1, \quad \text{(D.151)}$$

where (i) holds by (D.150). Inserting (D.150) and (D.151) back into (D.145), we arrive at

$$\begin{aligned}\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma}\big\rangle &\geq \frac{\frac{\gamma(p-z_\phi^\pi)}{2\sigma}}{(1-\gamma)(1+c_1)^2} \geq \frac{\gamma\left(p-z_\phi^\pi\right)}{8(1-\gamma)\sigma} = \frac{\gamma\,(p-q)\,\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)\sigma} \\ &= \frac{\gamma\Delta\big(1-\pi(\phi\,|\,0)\big)}{8(1-\gamma)\sigma} \geq 2\varepsilon\big(1-\pi(\phi\,|\,0)\big), \end{aligned} \quad \text{(D.152)}$$

where the last inequality holds by letting ($\gamma \geq 1/2$)

$$\Delta = 32(1-\gamma)\sigma\varepsilon. \quad \text{(D.153)}$$

Finally, it is easily verified that

$$\varepsilon \leq \frac{c_1}{32(1-\gamma)} \quad \Longrightarrow \quad \Delta \leq c_1\sigma. \quad \text{(D.154)}$$

D.4 Proof of the upper bound with χ^2 divergence: Theorem 12

The proof of Theorem 12 mainly follows the structure of the proof of Theorem 10 in Appendix D.2. Throughout this subchapter, for any nominal transition kernel P, the uncertainty set is taken as

(see (6.2))

$$\mathcal{U}^{\sigma}(P) = \mathcal{U}^{\sigma}_{\chi^2}(P) := \otimes \mathcal{U}^{\sigma}_{\chi^2}(P_{s,a}), \qquad \mathcal{U}^{\sigma}_{\chi^2}(P_{s,a}) := \left\{ P'_{s,a} \in \Delta(\mathcal{S}) : \sum_{s' \in \mathcal{S}} \frac{(P'(s' \,|\, s,a) - P(s' \,|\, s,a))^2}{P(s' \,|\, s,a)} \le \sigma \right\}. \tag{D.155}$$

D.4.1 Proof of Theorem 12

In order to control the performance gap $\left\| V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \right\|_{\infty}$, recall the error decomposition in (D.32):

$$V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \le \left(V^{\pi^\star,\sigma} - \widehat{V}^{\pi^\star,\sigma} \right) + \frac{2\gamma \varepsilon_{\mathsf{opt}}}{1-\gamma} 1 + \left(\widehat{V}^{\widehat{\pi},\sigma} - V^{\widehat{\pi},\sigma} \right), \tag{D.156}$$

where $\varepsilon_{\mathsf{opt}}$ (cf. (D.31)) shall be specified later (which justifies Remark 5). To further control (D.156), we bound the remaining two terms separately.

Step 1: controlling $\left\| \widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma} \right\|_{\infty}$. Towards this, recall the bound in (D.37) which holds for any uncertainty set:

$$\begin{aligned} \left\| \widehat{V}^{\pi^\star,\sigma} - V^{\pi^\star,\sigma} \right\|_{\infty} \le \gamma \max \Big\{ & \left\| \left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}} \right)^{-1} \left(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma} \right) \right\|_{\infty}, \\ & \left\| \left(I - \gamma \underline{\widehat{P}}^{\pi^\star,V} \right)^{-1} \left(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma} \right) \right\|_{\infty} \Big\}. \end{aligned} \tag{D.157}$$

To control the main term $\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma}$ in (D.157), we first introduce an important lemma whose proof is postponed to Appendix D.4.2.1.

Lemma 50. *Consider any $\sigma > 0$ and the uncertainty set $\mathcal{U}^{\sigma}(\cdot) := \mathcal{U}^{\sigma}_{\chi^2}(\cdot)$. For any $\delta \in (0,1)$ and any fixed policy π, one has with probability at least $1-\delta$,*

$$\left\| \underline{\widehat{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma} \right\|_{\infty} \le 4 \sqrt{\frac{2(1+\sigma) \log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}.$$

Applying Lemma 50 by taking $\pi = \pi^\star$ gives

$$\left\| \underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma} \right\|_{\infty} \le 4 \sqrt{\frac{2(1+\sigma) \log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}, \tag{D.158}$$

which directly leads to

$$\begin{aligned} & \left\| \left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}} \right)^{-1} \left(\underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma} \right) \right\|_{\infty} \\ & \le \left\| \underline{\widehat{P}}^{\pi^\star,V} V^{\pi^\star,\sigma} - \underline{P}^{\pi^\star,V} V^{\pi^\star,\sigma} \right\|_{\infty} \cdot \left\| \left(I - \gamma \underline{\widehat{P}}^{\pi^\star,\widehat{V}} \right)^{-1} 1 \right\|_{\infty} \le 4 \sqrt{\frac{2(1+\sigma) \log(\frac{24SAN}{\delta})}{(1-\gamma)^4 N}}. \end{aligned} \tag{D.159}$$

Similarly, we have

$$\left\|\left(I-\gamma\underline{\widehat{P}}^{\pi^\star,V}\right)^{-1}\left(\underline{\widehat{P}}^{\pi^\star,V}V^{\pi^\star,\sigma}-\underline{P}^{\pi^\star,V}V^{\pi^\star,\sigma}\right)\right\|_\infty \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^4N}}. \tag{D.160}$$

Inserting (D.159) and (D.160) back to (D.157) yields

$$\|\widehat{V}^{\pi^\star,\sigma}-V^{\pi^\star,\sigma}\|_\infty \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^4N}}. \tag{D.161}$$

Step 2: controlling $\left\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\right\|_\infty$. Recall the bound in (D.38) which holds for any uncertainty set:

$$\begin{aligned}\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\|_\infty \le \gamma\max\Big\{&\left\|\left(I-\gamma\underline{P}^{\widehat{\pi},V}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right)\right\|_\infty,\\ &\left\|\left(I-\gamma\underline{P}^{\widehat{\pi},\widehat{V}}\right)^{-1}\left(\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right)\right\|_\infty\Big\}.\end{aligned} \tag{D.162}$$

We introduce the following lemma which controls $\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}$ in (D.162); the proof is deferred to Appendix D.4.2.2.

Lemma 51. *Consider the uncertainty set $\mathcal{U}^\sigma(\cdot)\coloneqq\mathcal{U}^\sigma_{\chi^2}(\cdot)$ and any $\delta\in(0,1)$. With probability at least $1-\delta$, one has*

$$\left\|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty \le 12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}+4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}}. \tag{D.163}$$

Repeating the arguments from (D.158) to (D.161) yields

$$\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\|_\infty \le 12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^4N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}+4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^4}}. \tag{D.164}$$

Finally, inserting (D.161) and (D.164) back to (D.156) complete the proof

$$\begin{aligned}\|V^{\star,\sigma}-V^{\widehat{\pi},\sigma}\|_\infty &\le \|V^{\pi^\star,\sigma}-\widehat{V}^{\pi^\star,\sigma}\|_\infty+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}+\|\widehat{V}^{\widehat{\pi},\sigma}-V^{\widehat{\pi},\sigma}\|_\infty\\ &\le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^4N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}+12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^4N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}+4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^4}}\\ &\le 24\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^4N}},\end{aligned} \tag{D.165}$$

where the last line holds by taking $\varepsilon_{\mathsf{opt}} \le \min\left\{\sqrt{\frac{32(1+\sigma)\log(\frac{36SAN^2}{\delta})}{N}}, \frac{4\log(\frac{36SAN^2}{\delta})}{N}\right\}$.

D.4.2 Proof of the auxiliary lemmas

D.4.2.1 Proof of Lemma 50

Step 1: controlling the point-wise concentration. Consider any fixed policy π and the corresponding robust value vector $V \coloneqq V^{\pi,\sigma}$ (independent from $\widehat{P}^0$). Invoking Lemma 40 leads to that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
\left|\widehat{P}^{\pi,V}_{s,a}V^{\pi,\sigma} - P^{\pi,V}_{s,a}V^{\pi,\sigma}\right| &= \Bigg|\max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{P^0_{s,a}[V]_\alpha - \sqrt{\sigma\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right\} \\
&\qquad - \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left\{\widehat{P}^0_{s,a}[V]_\alpha - \sqrt{\sigma\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)}\right\}\Bigg| \\
&\le \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)[V]_\alpha + \sqrt{\sigma\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\sigma\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\le \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)[V]_\alpha\right| + \\
&\qquad + \max_{\alpha\in[\min_s V(s),\max_s V(s)]}\sqrt{\sigma}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right|,
\end{aligned} \tag{D.166}$$

where the first inequality follows by that the maximum operator is 1-Lipschitz, and the second inequality follows from the triangle inequality. Observing that the first term in (D.166) is exactly the same as (D.72), recalling the fact in (D.77) directly leads to: with probability at least $1-\delta$,

$$\max_{\alpha\in[\min_s V(s),\max_s V(s)]}\left|\left(P^0_{s,a}-\widehat{P}^0_{s,a}\right)[V]_\alpha\right| \le 2\sqrt{\frac{\log(\frac{2SAN}{\delta})}{(1-\gamma)^2 N}} \tag{D.167}$$

holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Then the remainder of the proof focuses on controlling the second term in (D.166).

Step 2: controlling the second term in (D.166). For any given $(s,a)\in\mathcal{S}\times\mathcal{A}$ and fixed $\alpha\in[0,\frac{1}{1-\gamma}]$, applying the concentration inequality (Panaganti and Kalathil, 2022, Lemma 6) with $\|[V]_\alpha\|_\infty \le \frac{1}{1-\gamma}$, we arrive at

$$\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \le \sqrt{\frac{2\log(\frac{2}{\delta})}{(1-\gamma)^2 N}} \tag{D.168}$$

holds with probability at least $1-\delta$. To obtain a uniform bound, we first observe the follow lemma proven in Appendix D.4.2.3.

Lemma 52. *For any V obeying $\|V\|_\infty \le \frac{1}{1-\gamma}$, the function $J_{s,a}(\alpha, V) := \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right|$ w.r.t. α obeys*

$$|J_{s,a}(\alpha_1, V) - J_{s,a}(\alpha_2, V)| \le 4\sqrt{\frac{|\alpha_1 - \alpha_2|}{1-\gamma}}.$$

In addition, we can construct an ε_3-net N_{ε_3} over $[0, \frac{1}{1-\gamma}]$ whose size is $|N_{\varepsilon_3}| \le \frac{3}{\varepsilon_3(1-\gamma)}$ (Vershynin, 2018). Armed with the above, we can derive the uniform bound over $\alpha \in [\min_s V(s), \max_s V(s)] \subset [0, 1/(1-\gamma)]$: with probability at least $1 - \frac{\delta}{SA}$, it holds that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
&\max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\le \max_{\alpha \in [0, 1/(1-\gamma)]} \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\overset{\text{(i)}}{\le} 4\sqrt{\frac{\varepsilon_3}{1-\gamma}} + \sup_{\alpha \in N_{\varepsilon_3}} \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\overset{\text{(ii)}}{\le} 4\sqrt{\frac{\varepsilon_3}{1-\gamma}} + \sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_3}|}{\delta})}{(1-\gamma)^2 N}} \\
&\overset{\text{(iii)}}{\le} 2\sqrt{\frac{2\log(\frac{2SA|N_{\varepsilon_3}|}{\delta})}{(1-\gamma)^2 N}} \le 2\sqrt{\frac{2\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}},
\end{aligned} \tag{D.169}$$

where (i) holds by the property of N_{ε_3}, (ii) follows from (D.168), (iii) arises from taking $\varepsilon_3 = \frac{\log(\frac{2SA|N_{\varepsilon_3}|}{\delta})}{8N(1-\gamma)}$, and the last inequality is verified by $|N_{\varepsilon_3}| \le \frac{3}{\varepsilon_3(1-\gamma)} \le 24N$.

Inserting (D.167) and (D.169) back to (D.166) and taking the union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$, we arrive at that for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta$,

$$\begin{aligned}
\left|\widehat{P}^{\pi,V}_{s,a} V - P^{\pi,V}_{s,a} V\right| &\le \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\left(P^0_{s,a} - \widehat{P}^0_{s,a}\right)[V]_\alpha\right| + \\
&\quad + \max_{\alpha \in [\min_s V(s), \max_s V(s)]} \left|\sqrt{\sigma \mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_\alpha)} - \sqrt{\sigma \mathsf{Var}_{P^0_{s,a}}([V]_\alpha)}\right| \\
&\le \sqrt{\frac{2\log(\frac{2SAN}{\delta})}{(1-\gamma)^2 N}} + 2\sqrt{\frac{2\sigma\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}} \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}.
\end{aligned}$$

Finally, we complete the proof by recalling the matrix form as below:

$$\left\|\underline{\widehat{P}}^{\pi,V} V^{\pi,\sigma} - \underline{P}^{\pi,V} V^{\pi,\sigma}\right\|_\infty \le \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left|\widehat{P}^{\pi,V}_{s,a} V - P^{\pi,V}_{s,a} V\right| \le 4\sqrt{\frac{2(1+\sigma)\log(\frac{24SAN}{\delta})}{(1-\gamma)^2 N}}.$$

D.4.2.2 Proof of Lemma 51

Step 1: decomposing the term of interest. The proof follows the routine of the proof of Lemma 47 in Appendix D.2.3.5. To begin with, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, following the same arguments of (D.166) yields

$$\left|\widehat{P}_{s,a}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma} - P_{s,a}^{\widehat{\pi},\widehat{V}} \widehat{V}^{\widehat{\pi},\sigma}\right| \le \max_{\alpha \in \left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \left|\left(P_{s,a}^0 - \widehat{P}_{s,a}^0\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right| + \\ + \max_{\alpha \in \left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \sqrt{\sigma} \left|\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}^0}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{P_{s,a}^0}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}\right|. \tag{D.170}$$

Invoking the fact in (D.101) (for proving Lemma 47), the first term in (D.170) obeys

$$\max_{\alpha \in \left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s), \max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \left|\left(P_{s,a}^0 - \widehat{P}_{s,a}^0\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right| \le \max_{\alpha \in [0,1/(1-\gamma)]} \left|\left(P_{s,a}^0 - \widehat{P}_{s,a}^0\right)\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right| \\ \le 4\sqrt{\frac{\log(\frac{3SAN^{3/2}}{(1-\gamma)\delta})}{(1-\gamma)^2 N}} + \frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}. \tag{D.171}$$

The remainder of the proof will focus on controlling the second term of (D.170).

Step 2: controlling the second term of (D.170). Towards this, we recall the auxiliary robust MDP $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ defined in Appendix D.2.3.5. Taking the uncertainty set $\mathcal{U}^\sigma(\cdot) \coloneqq \mathcal{U}_{\chi^2}^\sigma(\cdot)$ for both $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ and $\widehat{\mathcal{M}}_{\mathsf{rob}}$, we recall the corresponding robust Bellman operator $\widehat{\mathcal{T}}_{s,u}^\sigma(\cdot)$ in (D.90) and the following definition in (E.169)

$$u^\star \coloneqq \widehat{V}^{\star,\sigma}(s) - \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma(e_s)} \mathcal{P}\widehat{V}^{\star,\sigma}. \tag{D.172}$$

Following the arguments in Appendix D.2.3.5, it can be verified that there exists a unique fixed point $\widehat{Q}_{s,u}^{\star,\sigma}$ of the operator $\widehat{\mathcal{T}}_{s,u}^\sigma(\cdot)$, which satisfies $0 \le \widehat{Q}_{s,u}^{\star,\sigma} \le \frac{1}{1-\gamma}1$. In addition, the corresponding robust value function coincides with that of the operator $\widehat{\mathcal{T}}^\sigma(\cdot)$, i.e., $\widehat{V}_{s,u}^{\star,\sigma} = \widehat{V}^{\star,\sigma}$.

We recall the N_{ε_2}-net over $\left[0, \frac{1}{1-\gamma}\right]$ whose size obeying $|N_{\varepsilon_2}| \le \frac{3}{\varepsilon_2(1-\gamma)}$ (Vershynin, 2018). Then for all $u \in N_{\varepsilon_2}$ and a fixed α, $\widehat{\mathcal{M}}_{\mathsf{rob}}^{s,u}$ is statistically independent from $\widehat{P}_{s,a}^0$, which indicates the independence between $[\widehat{V}_{s,u}^{\star,\sigma}]_\alpha$ and $\widehat{P}_{s,a}^0$. With this in mind, invoking the fact in (D.169) and taking the union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $u \in N_{\varepsilon_2}$ yields that, with probability at least $1-\delta$,

$$\max_{\alpha \in [0,1/(1-\gamma)]} \left|\sqrt{\mathsf{Var}_{\widehat{P}_{s,a}^0}\left(\left[\widehat{V}_{s,u}^{\star,\sigma}\right]_\alpha\right)} - \sqrt{\mathsf{Var}_{P_{s,a}^0}\left(\left[\widehat{V}_{s,u}^{\star,\sigma}\right]_\alpha\right)}\right| \le 2\sqrt{\frac{2\log(\frac{24SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2 N}} \tag{D.173}$$

holds for all $(s,a,u) \in \mathcal{S} \times \mathcal{A} \times N_{\varepsilon_2}$.

To continue, we decompose the term of interest in (D.170) as follows:

$$
\begin{aligned}
&\max_{\alpha\in\left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}\right| \\
&\le \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}\right| \\
&\overset{\text{(i)}}{\le} \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}\right| \\
&\quad+\max_{\alpha\in[0,1/(1-\gamma)]}\left[\sqrt{\left|\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)-\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)\right|}+\sqrt{\left|\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)-\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)\right|}\right] \\
&\overset{\text{(ii)}}{\le} \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}\right|+\max_{\alpha\in[0,1/(1-\gamma)]}2\sqrt{\frac{2}{(1-\gamma)}\left\|\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha-\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right\|_\infty} \\
&\le \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}\right|+4\sqrt{\frac{\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}}, \qquad \text{(D.174)}
\end{aligned}
$$

where (i) holds by the triangle inequality, (ii) arises from applying Lemma 37, and the last inequality holds by (D.31).

Armed with the above facts, invoking the identity $\widehat{V}^{\star,\sigma}=\widehat{V}^{\star,\sigma}_{s,u^\star}$ leads to that for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$,

$$
\begin{aligned}
&\max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}\right]_\alpha\right)}\right| \\
&=\max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_\alpha\right)}\right| \\
&\overset{\text{(i)}}{\le} \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_\alpha\right)}\right| \\
&\quad+\max_{\alpha\in[0,1/(1-\gamma)]}\left[\sqrt{\left|\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_\alpha\right)-\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_\alpha\right)\right|}+\sqrt{\left|\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,u^\star}\right]_\alpha\right)-\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_\alpha\right)\right|}\right] \\
&\overset{\text{(ii)}}{\le} \max_{\alpha\in[0,1/(1-\gamma)]}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\star,\sigma}_{s,\overline{u}}\right]_\alpha\right)}\right|+4\sqrt{\frac{\varepsilon_2}{(1-\gamma)}} \\
&\overset{\text{(iii)}}{\le} 2\sqrt{\frac{2\log(\frac{24SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2N}}+4\sqrt{\frac{\varepsilon_2}{(1-\gamma)}} \\
&\le 6\sqrt{\frac{2\log(\frac{36SAN^2|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2N}}, \qquad \text{(D.175)}
\end{aligned}
$$

where (i) holds by the triangle inequality, (ii) arises from applying Lemma 37 and the fact $\left\|\widehat{V}^{\star,\sigma}_{s,\overline{u}}-\widehat{V}^{\star,\sigma}_{s,u^\star}\right\|_\infty\le\frac{\varepsilon_2}{(1-\gamma)}$ (see (E.181)), (iii) follows from (D.173), and the last inequality holds by letting $\varepsilon_2=\frac{2\log(\frac{24SAN|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)N}$, which leads to $|N_{\varepsilon_2}|\le\frac{3}{\varepsilon_2(1-\gamma)}\le\frac{3N}{2}$.

In summary, inserting (D.175) back to (D.174) and (D.174) leads to with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{\alpha\in\left[\min_s \widehat{V}^{\widehat{\pi},\sigma}(s),\max_s \widehat{V}^{\widehat{\pi},\sigma}(s)\right]} \left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}-\sqrt{\mathsf{Var}_{P^0_{s,a}}\left(\left[\widehat{V}^{\widehat{\pi},\sigma}\right]_\alpha\right)}\right| \\
&\le 6\sqrt{\frac{2\sigma\log(\frac{36SAN^2|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2N}}+4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}} \qquad \text{(D.176)}
\end{aligned}$$

holds for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

Step 4: finishing up. Inserting (D.176) and (D.171) back to (D.170), we complete the proof: with probability at least $1-\delta$,

$$\begin{aligned}
\left\|\underline{\widehat{P}}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}-\underline{P}^{\widehat{\pi},\widehat{V}}\widehat{V}^{\widehat{\pi},\sigma}\right\|_\infty &\le 4\sqrt{\frac{\log(\frac{3SAN^{3/2}}{(1-\gamma)\delta})}{(1-\gamma)^2N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}+6\sqrt{\frac{2\sigma\log(\frac{36SAN^2|N_{\varepsilon_2}|}{\delta})}{(1-\gamma)^2N}}+4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}} \\
&\le 12\sqrt{\frac{2(1+\sigma)\log(\frac{36SAN^2}{\delta})}{(1-\gamma)^2N}}+\frac{2\gamma\varepsilon_{\mathsf{opt}}}{1-\gamma}+4\sqrt{\frac{\sigma\varepsilon_{\mathsf{opt}}}{(1-\gamma)^2}}. \qquad \text{(D.177)}
\end{aligned}$$

D.4.2.3 Proof of Lemma 52

For any $0\le\alpha_1,\alpha_2\le 1/(1-\gamma)$, one has

$$\begin{aligned}
&|J_{s,a}(\alpha_1,V)-J_{s,a}(\alpha_2,V)| \\
&=\left|\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_1})}-\sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_1})}\right|-\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_2})}-\sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_2})}\right|\right| \\
&\overset{\text{(i)}}{\le}\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_1})}-\sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_1})}-\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_2})}+\sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_2})}\right| \\
&\le\left|\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_1})}-\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_2})}\right|+\left|\sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_1})}-\sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_2})}\right| \\
&\overset{\text{(ii)}}{\le}\sqrt{\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_2})-\mathsf{Var}_{\widehat{P}^0_{s,a}}([V]_{\alpha_1})}+\sqrt{\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_2})-\mathsf{Var}_{P^0_{s,a}}([V]_{\alpha_1})} \\
&\overset{\text{(iii)}}{\le}\sqrt{\left|\widehat{P}^0_{s,a}\left[([V]_{\alpha_1})\circ([V]_{\alpha_1})-([V]_{\alpha_2})\circ([V]_{\alpha_2})\right]\right|+\left|\widehat{P}^0_{s,a}\left([V]_{\alpha_1}+[V]_{\alpha_2}\right)\cdot\widehat{P}^0_{s,a}\left([V]_{\alpha_1}-[V]_{\alpha_2}\right)\right|} \\
&\quad+\sqrt{\left|P^0_{s,a}\left[([V]_{\alpha_1})\circ([V]_{\alpha_1})-([V]_{\alpha_2})\circ([V]_{\alpha_2})\right]\right|+\left|P^0_{s,a}\left([V]_{\alpha_1}+[V]_{\alpha_2}\right)\cdot P^0_{s,a}\left([V]_{\alpha_1}-[V]_{\alpha_2}\right)\right|} \\
&\le 2\sqrt{2(\alpha_1+\alpha_2)|\alpha_1-\alpha_2|}\le 4\sqrt{\frac{|\alpha_1-\alpha_2|}{1-\gamma}}. \qquad \text{(D.178)}
\end{aligned}$$

where (i) holds by the fact $||x|-|y||\le|x-y|$ for all $x,y\in\mathbb{R}$, (ii) follows from the fact that $\sqrt{x}-\sqrt{y}\le\sqrt{x-y}$ for any $x\ge y\ge 0$ and $\mathsf{Var}_P([V]_{\alpha_2})\ge\mathsf{Var}_P([V]_{\alpha_1})$ for any transition kernel

$P \in \Delta(\mathcal{S})$, (iii) holds by the definition of $\mathrm{Var}_P(\cdot)$ defined in (D.7), and the last inequality arises from $0 \le \alpha_1, \alpha_2 \le 1/(1-\gamma)$.

D.5 Proof of the lower bound with χ^2 divergence: Theorem 13

To prove Theorem 13, we shall first construct some hard instances and then characterize the sample complexity requirements over these instances. The structure of the hard instances are the same as the ones used in the proof of Theorem 11.

D.5.1 Construction of the hard problem instances

First, note that we shall use the same MDPs defined in Appendix D.3.1 as follows

$$\left\{\mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi, r, \gamma\right) \mid \phi = \{0,1\}\right\}.$$

In particular, we shall keep the structure of the transition kernel in (E.189), reward function in (E.193) and initial state distribution in (E.197), while p and Δ shall be specified differently later.

Uncertainty set of the transition kernels. Recalling the uncertainty set associated with χ^2 divergence in (D.155), for any uncertainty level σ, the uncertainty set throughout this subchapter is defined as $\mathcal{U}^\sigma(P^\phi)$:

$$\mathcal{U}^\sigma(P^\phi) := \otimes \mathcal{U}^\sigma_{\chi^2}(P^\phi_{s,a}), \qquad \mathcal{U}^\sigma_{\chi^2}(P^\phi_{s,a}) := \left\{P_{s,a} \in \Delta(\mathcal{S}) : \sum_{s' \in \mathcal{S}} \frac{\left(P(s' \mid s,a) - P^\phi(s' \mid s,a)\right)^2}{P^\phi(s' \mid s,a)} \le \sigma\right\}. \tag{D.179}$$

Clearly, $\mathcal{U}^\sigma(P^\phi_{s,a}) = P^\phi_{s,a}$ whenever the state transition is deterministic for χ^2 divergence. Here, q and Δ (whose choice will be specified later in more detail) which determine the instances are specified as

$$0 \le q = \begin{cases} 1-\gamma & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \frac{\sigma}{1+\sigma} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases}, \qquad p = q + \Delta, \tag{D.180}$$

and

$$0 < \Delta \le \begin{cases} \frac{1}{4}(1-\gamma) & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(1+\sigma)}\right\} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases}. \tag{D.181}$$

This directly ensures that

$$p = \Delta + q \le \max\left\{\frac{\frac{1}{2}+\sigma}{1+\sigma}, \frac{5}{4}(1-\gamma)\right\} \le 1$$

since $\gamma \in \left[\frac{3}{4}, 1\right)$.

To continue, for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we denote the infimum probability of moving to the next state s' associated with any perturbed transition kernel $P_{s,a} \in \mathcal{U}^{\sigma}(P^{\phi}_{s,a})$ as

$$\underline{P}^{\phi}(s' \,|\, s, a) \coloneqq \inf_{P_{s,a} \in \mathcal{U}^{\sigma}(P^{\phi}_{s,a})} P(s' \,|\, s, a). \tag{D.182}$$

In addition, we denote the transition from state 0 to state 1 as follows, which plays an important role in the analysis,

$$\underline{p} \coloneqq \underline{P}^{\phi}(1 \,|\, 0, \phi), \qquad \underline{q} \coloneqq \underline{P}^{\phi}(1 \,|\, 0, 1-\phi). \tag{D.183}$$

Before continuing, we introduce some facts about $\underline{p}$ and $\underline{q}$ which are summarized as the following lemma; the proof is postponed to Appendix D.5.3.1.

Lemma 53. *Consider any $\sigma \in (0, \infty)$ and any p, q, Δ obeying (D.180) and (D.181), the following properties hold*

$$\begin{cases} \frac{1-\gamma}{2} < \underline{q} < 1-\gamma, \quad \underline{q} + \frac{3}{4}\Delta \le \underline{p} \le \underline{q} + \Delta \le \frac{5(1-\gamma)}{4} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right), \\ \underline{q} = 0, \quad \frac{\sigma+1}{2}\Delta \le \underline{p} \le (3+\sigma)\Delta & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right). \end{cases} \tag{D.184}$$

Value functions and optimal policies. Armed with above facts, we are positioned to derive the corresponding robust value functions, the optimal policies, and its corresponding optimal robust value functions. For any RMDP $\mathcal{M}_{\phi}$ with the uncertainty set defined in (D.179), we denote the robust optimal policy as $\pi^{\star}_{\phi}$, the robust value function of any policy π (resp. the optimal policy $\pi^{\star}_{\phi}$) as $V^{\pi,\sigma}_{\phi}$ (resp. $V^{\star,\sigma}_{\phi}$). The following lemma describes some key properties of the robust (optimal) value functions and optimal policies whose proof is postponed to Appendix D.5.3.2.

Lemma 54. *For any $\phi = \{0, 1\}$ and any policy π, one has*

$$V^{\pi,\sigma}_{\phi}(0) = \frac{\gamma z^{\pi}_{\phi}}{(1-\gamma)\left(1-\gamma\left(1-z^{\pi}_{\phi}\right)\right)}, \tag{D.185}$$

where z^{π}_{ϕ} is defined as

$$z^{\pi}_{\phi} \coloneqq \underline{p}\pi(\phi \,|\, 0) + \underline{q}\pi(1-\phi \,|\, 0). \tag{D.186}$$

In addition, the optimal value functions and the optimal policies obey

$$V_\phi^{\star,\sigma}(0) = \frac{\gamma \underline{p}}{(1-\gamma)\big(1-\gamma\big(1-\underline{p}\big)\big)}, \tag{D.187a}$$

$$\pi_\phi^\star(\phi \,|\, s) = 1, \qquad \text{for } s \in \mathcal{S}. \tag{D.187b}$$

D.5.2 Establishing the minimax lower bound

Our goal is to control the performance gap w.r.t. any policy estimator $\widehat{\pi}$ based on the generated dataset and the chosen initial distribution φ in (E.197), which gives

$$\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \big\rangle = V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0). \tag{D.188}$$

Step 1: converting the goal to estimate ϕ. To achieve the goal, we first introduce the following fact which shall be verified in Appendix D.5.3.3: given

$$\varepsilon \le \begin{cases} \frac{1}{72(1-\gamma)} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right), \\ \frac{1}{256(1+\sigma)(1-\gamma)} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right), \\ \frac{3}{32} & \text{if } \sigma > \frac{1}{3(1-\gamma)}. \end{cases} \tag{D.189}$$

choosing

$$\Delta = \begin{cases} 18(1-\gamma)^2\varepsilon & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right), \\ 64(1+\sigma)(1-\gamma)^2\varepsilon & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right), \\ \frac{16}{3(1+\sigma)}\varepsilon & \text{if } \sigma > \frac{1}{3(1-\gamma)}. \end{cases} \tag{D.190}$$

which satisfies the requirement of Δ in (D.180), it holds that for any policy $\widehat{\pi}$,

$$\big\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\widehat{\pi},\sigma} \big\rangle \ge 2\varepsilon\big(1 - \widehat{\pi}(\phi \,|\, 0)\big). \tag{D.191}$$

Step 2: arriving at the final results. To continue, following the same definitions and argument in Appendix D.3.2, we recall the minimax probability of the error and its property as follows:

$$p_{\mathrm{e}} \ge \frac{1}{4}\exp\left\{ - N\Big(\mathsf{KL}\big(P^0(\cdot \,|\, 0,0) \parallel P^1(\cdot \,|\, 0,0)\big) + \mathsf{KL}\big(P^0(\cdot \,|\, 0,1) \parallel P^1(\cdot \,|\, 0,1)\big)\Big)\right\}, \tag{D.192}$$

then we can complete the proof by showing $p_{\mathrm{e}} \ge \frac{1}{8}$ given the bound for the sample size N. In the following, we shall control the KL divergence terms in (D.192) in three different cases.

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, applying $\gamma \in [\frac{3}{4}, 1)$ yields

$$\begin{aligned} 1-q > 1-p = 1-q-\Delta > \gamma - \frac{1-\gamma}{4} > \frac{3}{4} - \frac{1}{16} > \frac{1}{2}, \\ p \geq q = 1-\gamma. \end{aligned} \tag{D.193}$$

Armed with the above facts, applying Lemma 60 (cf. (E.11)) yields

$$\begin{aligned} \mathsf{KL}\big(P^0(\cdot\,|\,0,1) \,\|\, P^1(\cdot\,|\,0,1)\big) = \mathsf{KL}\left(p \,\|\, q\right) &\leq \frac{(p-q)^2}{(1-p)p} \overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{=} \frac{324(1-\gamma)^4\varepsilon^2}{p(1-p)} \\ &\overset{\text{(iii)}}{\leq} 648(1-\gamma)^3\varepsilon^2, \end{aligned} \tag{D.194}$$

where (i) follows from the definition in (D.180), (ii) holds by plugging in the expression of Δ in (D.190), and (iii) arises from (D.193). The same bound can be established for $\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0) \,\|\, P_1^1(\cdot\,|\,0,0)\big)$. Substituting (D.194) back into (D.192) demonstrates that: if the sample size is chosen as

$$N \leq \frac{\log 2}{1296(1-\gamma)^3\varepsilon^2}, \tag{D.195}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\Big\{-N \cdot 1296(1-\gamma)^3\varepsilon^2\Big\} \geq \frac{1}{8}. \tag{D.196}$$

- Case 2: $\sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right)$. Applying the facts of Δ in (D.181), one has

$$\begin{aligned} 1-q > 1-p = 1-q-\Delta \geq \frac{1}{1+\sigma} - \frac{1}{2(1+\sigma)} = \frac{1}{2(1+\sigma)}, \\ p \geq q = \frac{\sigma}{1+\sigma}. \end{aligned} \tag{D.197}$$

Given (D.197), applying Lemma 60 (cf. (E.11)) yields

$$\begin{aligned} \mathsf{KL}\big(P^0(\cdot\,|\,0,1) \,\|\, P^1(\cdot\,|\,0,1)\big) = \mathsf{KL}\left(p \,\|\, q\right) &\leq \frac{(p-q)^2}{(1-p)p} \overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{=} \frac{4096(1+\sigma)^2(1-\gamma)^4\varepsilon^2}{p(1-p)} \\ &\overset{\text{(iii)}}{\leq} \frac{4096(1+\sigma)^2(1-\gamma)^4\varepsilon^2}{\frac{\sigma}{2(1+\sigma)^2}} \leq \frac{8192(1-\gamma)^4(1+\sigma)^4\varepsilon^2}{\sigma}, \end{aligned} \tag{D.198}$$

where (i) follows from the definition in (D.180), (ii) holds by plugging in the expression of Δ in (D.190), and (iii) arises from (D.197). The same bound can be established for $\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0)\,\|\,P_1^1(\cdot\,|\,0,0)\big)$.

Substituting (D.198) back into (E.88) demonstrates that: if the sample size is chosen as

$$N \leq \frac{\sigma \log 2}{16384(1-\gamma)^4(1+\sigma)^4\varepsilon^2}, \tag{D.199}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{-N\frac{16384(1-\gamma)^4(1+\sigma)^4\varepsilon^2}{\sigma}\right\} \geq \frac{1}{8}. \tag{D.200}$$

- Case 3: $\sigma > \frac{1}{3(1-\gamma)} \geq \frac{1}{3}$. Regarding this, one gives

$$\begin{aligned} &1-q > 1-p = 1-q-\Delta \geq \frac{1}{1+\sigma} - \frac{1}{4(1+\sigma)} \geq \frac{1}{2(1+\sigma)}, \\ &p \geq q \geq \frac{1}{4}. \end{aligned} \tag{D.201}$$

Given $p \geq q \geq 1/2$ and (D.201), applying Lemma 60 (cf. (E.11)) yields

$$\begin{aligned} \mathsf{KL}\big(P^0(\cdot\,|\,0,1)\,\|\,P^1(\cdot\,|\,0,1)\big) = \mathsf{KL}\left(p\,\|\,q\right) &\leq \frac{(p-q)^2}{(1-p)p} \overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{\leq} \frac{\frac{64}{(1+\sigma)^2}\varepsilon^2}{p(1-p)} \\ &\overset{\text{(iii)}}{\leq} \frac{492\varepsilon^2}{\sigma}, \end{aligned} \tag{D.202}$$

where (i) follows from the definition in (D.180), (ii) holds by plugging in the expression of Δ in (D.190), and (iii) arises from (D.201). The same bound can be established for $\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0)\,\|\,P_1^1(\cdot\,|\,0,0)\big)$. Substituting (D.202) back into (E.88) demonstrates that: if the sample size is chosen as

$$N \leq \frac{\sigma \log 2}{984\varepsilon^2}, \tag{D.203}$$

then one necessarily has

$$p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{-N\frac{984\varepsilon^2}{\sigma}\right\} \geq \frac{1}{8}. \tag{D.204}$$

Step 3: putting things together. Finally, summing up the results in (D.195), (D.199), and (D.203), combined with the requirement in (D.189), one has when

$$\varepsilon \le c_1 \begin{cases} \frac{1}{1-\gamma} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \max\left\{\frac{1}{(1+\sigma)(1-\gamma)}, 1\right\} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases}, \tag{D.205}$$

taking

$$N \le c_2 \begin{cases} \frac{1}{(1-\gamma)^3\varepsilon^2} & \text{if } \sigma \in \left(0, \frac{1-\gamma}{4}\right) \\ \frac{\sigma}{\min\{1,(1-\gamma)^4(1+\sigma)^4\}\varepsilon^2} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \infty\right) \end{cases} \tag{D.206}$$

leads to $p_e \ge \frac{1}{8}$, for some universal constants $c_1, c_2 > 0$.

D.5.3 Proof of the auxiliary facts

We begin with some basic facts about the χ^2 divergence defined in (E.10) for any two Bernoulli distributions $\mathsf{Ber}(w)$ and $\mathsf{Ber}(x)$, denoted as

$$f(w,x) \coloneqq \chi^2(x \parallel w) = \frac{(w-x)^2}{w} + \frac{(1-w-(1-x))^2}{1-w} = \frac{(w-x)^2}{w(1-w)}. \tag{D.207}$$

For $x \in [0, w)$, it is easily verified that the partial derivative w.r.t. x obeys $\frac{\partial f(w,x)}{\partial x} = \frac{2(x-w)}{w(1-w)} < 0$, implying that

$$\forall\, x_1 < x_2 \in [0, w), \qquad f(w, x_1) > f(w, x_2). \tag{D.208}$$

In other words, the χ^2 divergence $f(w,x)$ increases as x decreases from w to 0.

Next, we introduce the following function for any fixed $\sigma \in (0,\infty)$ and any $x \in \left[\frac{\sigma}{1+\sigma}, 1\right)$:

$$f_\sigma(x) \coloneqq \inf_{\{y:\chi^2(y\|x)\le\sigma, y\in[0,x]\}} y \overset{\text{(i)}}{=} \max\left\{0, x - \sqrt{\sigma x(1-x)}\right\} = x - \sqrt{\sigma x(1-x)}, \tag{D.209}$$

where (i) has been verified in Yang et al. (2022, Corollary B.2), and the last equality holds since $x \ge \frac{\sigma}{1+\sigma}$. The next lemma summarizes some useful facts about $f_\sigma(\cdot)$, which again has been verified in Yang et al. (2022, Lemma B.12 and Corollary B.2).

Lemma 55. *Consider any $\sigma \in (0,\infty)$. For $x \in [\frac{\sigma}{1+\sigma}, 1)$, $f_\sigma(x)$ is convex and differentiable, which obeys*

$$f_\sigma'(x) = 1 + \frac{\sqrt{\sigma}(2x-1)}{2\sqrt{x(1-x)}}.$$

D.5.3.1 Proof of Lemma 53

Let us control $\underline{q}$ and $\underline{p}$ respectively.

Step 1: controlling $\underline{q}$. We shall control $\underline{q}$ in different cases w.r.t. the uncertainty level σ.

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, recall that $q = 1-\gamma$ defined in (D.180), applying (D.209) with $x = q$ leads to

$$1-\gamma = q > \underline{q} = f_\sigma(q) = 1-\gamma-\sqrt{\sigma\gamma(1-\gamma)} \geq 1-\gamma-\sqrt{\frac{1-\gamma}{4}\gamma(1-\gamma)} > \frac{1-\gamma}{2}. \quad \text{(D.210)}$$

- Case 2: $\sigma \in \left[\frac{1-\gamma}{4}, \infty\right)$. Note that it suffices to treat $P^{\phi}_{0,1-\phi}$ as a Bernoulli distribution $\mathsf{Ber}(q)$ over states 1 and 0, since we do not allow transition to other states. Recalling $q = \frac{\sigma}{1+\sigma}$ in (D.180) and noticing the fact that

$$f(q,0) = \frac{q^2}{q} + \frac{(1-(1-q))^2}{1-q} = \frac{q}{(1-q)} = \sigma, \quad \text{(D.211)}$$

one has the probability $\mathsf{Ber}(0)$ falls into the uncertainty set of $\mathsf{Ber}(q))$ of size σ. As a result, recalling the definition (D.183) leads to

$$\underline{q} = \underline{P}^{\phi}(1 \,|\, 0, 1-\phi) = 0, \quad \text{(D.212)}$$

since $\underline{q} \geq 0$.

Step 2: controlling $\underline{p}$. To characterize the value of $\underline{p}$, we also divide into several cases separately.

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, note that $p > q = 1-\gamma \geq \frac{\sigma}{1+\sigma}$. Therefore, applying that $f_\sigma(\cdot)$ is convex and the form of its derivative in Lemma 55, one has

$$\begin{aligned}\underline{p} &= f_\sigma(p) \geq f_\sigma(q) + f'_\sigma(q)(p-q) \\ &= \underline{q} + \left(1 + \frac{\sqrt{\sigma}(2q-1)}{2\sqrt{q(1-q)}}\right)\Delta \geq \underline{q} + \left(1 - \frac{\sqrt{\frac{1-\gamma}{4}}(1-2(1-\gamma))}{2\sqrt{(1-\gamma)\gamma}}\right)\Delta \geq \underline{q} + \frac{3\Delta}{4}. \end{aligned} \quad \text{(D.213)}$$

Similarly, applying Lemma 55 leads to

$$\begin{aligned}\underline{p} &= f_\sigma(p) \leq f_\sigma(q) + f'_\sigma(p)(p-q) \\ &= \underline{q} + \left(1 - \frac{\sqrt{\sigma}(1-2p)}{2\sqrt{p(1-p)}}\right)\Delta \leq \underline{q} + \Delta, \end{aligned} \quad \text{(D.214)}$$

where the last inequality holds by $1-2p>0$ due to the fact $p=q+\Delta\le\frac{5}{4}(1-\gamma)\le\frac{5}{16}<\frac{1}{2}$ (cf. (D.181) and $\gamma\in[\frac{3}{4},1)$). To sum up, given $\sigma\in\left(0,\frac{1-\gamma}{4}\right)$, combined with (D.210), we arrive at

$$\underline{q}+\frac{3}{4}\Delta\le\underline{p}\le\underline{q}+\Delta\le\frac{5(1-\gamma)}{4}, \tag{D.215}$$

where the last inequality holds by $\Delta\le\frac{1}{4}(1-\gamma)$ (see (D.180)).

- Case 2: $\sigma\in\left[\frac{1-\gamma}{4},\infty\right)$. We recall that $p=q+\Delta>q=\frac{\sigma}{1+\sigma}$ in (D.180). To derive the lower bound for $\underline{p}$ in (D.183), similar to (D.213), one has

$$\begin{aligned}\underline{p}&=f_\sigma(p)\ge f_\sigma(q)+f'_\sigma(q)(p-q)\\&=\underline{q}+\left(1+\frac{\sqrt{\sigma}(2q-1)}{2\sqrt{q(1-q)}}\right)\Delta\\&\overset{\text{(i)}}{=}0+\left(1+\frac{\sqrt{\sigma}\frac{\sigma-1}{1+\sigma}}{2\sqrt{\frac{\sigma}{1+\sigma}\frac{1}{1+\sigma}}}\right)\Delta=\left(1+\frac{\sigma-1}{2}\right)\Delta=\left(\frac{\sigma+1}{2}\right)\Delta,\end{aligned} \tag{D.216}$$

where (i) follows from $q=\frac{\sigma}{1+\sigma}$ and $\underline{q}=0$ (see (D.212)). For the other direction, similar to (D.214), we have

$$\begin{aligned}\underline{p}&=f_\sigma(p)\le f_\sigma(q)+f'_\sigma(p)(p-q)=\underline{q}+\left(1+\frac{\sqrt{\sigma}(2p-1)}{2\sqrt{p(1-p)}}\right)\Delta\\&\overset{\text{(i)}}{=}\left(1+\frac{\sqrt{\sigma}(2p-1)}{2\sqrt{p(1-p)}}\right)\Delta\overset{\text{(ii)}}{=}\left(1+\frac{\sqrt{\sigma}\left(\frac{\sigma-1}{1+\sigma}+2\Delta\right)}{2\sqrt{\left(\frac{\sigma}{1+\sigma}+\Delta\right)\left(\frac{1}{1+\sigma}-\Delta\right)}}\right)\Delta\\&\overset{\text{(iii)}}{\le}\left(1+\frac{\sqrt{\sigma}(1+2\Delta)}{2\sqrt{\frac{\sigma}{1+\sigma}\cdot\frac{1}{2(1+\sigma)}}}\right)\Delta\overset{\text{(iv)}}{\le}\left(1+(1+\sigma)\left(1+\frac{1}{1+\sigma}\right)\right)\Delta=(3+\sigma)\Delta,\end{aligned} \tag{D.217}$$

where (i) holds by $\underline{q}=0$ (see (D.212)), (ii) follows from plugging in $p=q+\Delta=\frac{\sigma}{1+\sigma}+\Delta$, and (iii) and (iv) arises from $\Delta=\min\left\{\frac{1}{4}(1-\gamma),\frac{1}{2(1+\sigma)}\right\}\le 1$ in (D.181). Combining (D.216) and (D.217) yields

$$\frac{\sigma+1}{2}\Delta\le\underline{p}\le(3+\sigma)\Delta. \tag{D.218}$$

Step 3: combining all the results. Finally, summing up the results for both $\underline{q}$ (in (D.210) and (D.212)) and $\underline{p}$ (in (D.215) and (D.218)), we arrive at the advertised bound.

D.5.3.2 Proof of Lemma 54

The robust value function for any policy π. For any $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, we first characterize the robust value function of any policy π over different states.

Towards this, it is easily observed that for any policy π, the robust value functions at state $s=1$ or any $s \in \{2,3,\cdots,S-1\}$ obey

$$V_\phi^{\pi,\sigma}(1) \overset{\text{(i)}}{=} 1+\gamma V_\phi^{\pi,\sigma}(1) = \frac{1}{1-\gamma} \tag{D.219a}$$

and

$$\forall s \in \{2,3,\cdots,S\}: \qquad V_\phi^{\pi,\sigma}(s) \overset{\text{(ii)}}{=} 0+\gamma V_\phi^{\pi,\sigma}(1) = \frac{\gamma}{1-\gamma}, \tag{D.219b}$$

where (i) and (ii) is according to the facts that the transitions defined over states $s \geq 1$ in (E.189) give only one possible next state 1, leading to a non-random transition in the uncertainty set associated with χ^2 divergence, and $r(1,a)=1$ for all $a \in \mathcal{A}'$ and $r(s,a)=0$ holds all $(s,a) \in \{2,3,\cdots,S-1\}\times\mathcal{A}$.

To continue, the robust value function at state 0 with policy π satisfies

$$\begin{aligned} V_\phi^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,0)}\left[r(0,a)+\gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,a}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\ &= 0+\gamma\pi(\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma} + \gamma\pi(1-\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,1-\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma} \qquad \text{(D.220)} \\ &\overset{\text{(i)}}{\leq} \frac{\gamma}{1-\gamma}, \qquad \text{(D.221)} \end{aligned}$$

where (i) holds by that $\|V_\phi^{\pi,\sigma}\|_\infty \leq \frac{1}{1-\gamma}$. Summing up the results in (D.219b) and (D.221) leads to

$$\forall s \in \{2,3,\cdots,S\}, \qquad V_\phi^{\pi,\sigma}(1) > V_\phi^{\pi,\sigma}(s) \geq V_\phi^{\pi,\sigma}(0). \tag{D.222}$$

With the transition kernel in (E.189) over state 0 and the fact in (D.222), (D.220) can be rewritten as

$$\begin{aligned} V_\phi^{\pi,\sigma}(0) &= \gamma\pi(\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma} + \gamma\pi(1-\phi\,|\,0)\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P_{0,1-\phi}^\phi)} \mathcal{P}V_\phi^{\pi,\sigma} \\ &\overset{\text{(i)}}{=} \gamma\pi(\phi\,|\,0)\left[\underline{p}V_\phi^{\pi,\sigma}(1)+\left(1-\underline{p}\right)V_\phi^{\pi,\sigma}(0)\right] + \gamma\pi(1-\phi\,|\,0)\left[\underline{q}V_\phi^{\pi,\sigma}(1)+\left(1-\underline{q}\right)V_\phi^{\pi,\sigma}(0)\right] \\ &\overset{\text{(ii)}}{=} \gamma z_\phi^\pi V_\phi^{\pi,\sigma}(1)+\gamma\left(1-z_\phi^\pi\right)V_\phi^{\pi,\sigma}(0) \\ &= \frac{\gamma z_\phi^\pi}{(1-\gamma)\Big(1-\gamma\big(1-z_\phi^\pi\big)\Big)}, \qquad \text{(D.223)} \end{aligned}$$

where (i) holds by the definition of $\underline{p}$ and $\underline{q}$ in (D.183), (ii) follows from the definition of z_ϕ^π in (D.186), and the last line holds by applying (D.219a) and solving the resulting linear equation for $V_\phi^{\pi,\sigma}(0)$.

Optimal policy and its optimal value function. To continue, observing that $V_\phi^{\pi,\sigma}(0) =: f(z_\phi^\pi)$ is increasing in z_ϕ^π since the derivative of $f(z_\phi^\pi)$ w.r.t. z_ϕ^π obeys

$$f'(z_\phi^\pi) = \frac{\gamma(1-\gamma)\left(1-\gamma\left(1-z_\phi^\pi\right)\right) - \gamma^2 z_\phi^\pi (1-\gamma)}{(1-\gamma)^2\left(1-\gamma\left(1-z_\phi^\pi\right)\right)^2} = \frac{\gamma}{\left(1-\gamma\left(1-z_\phi^\pi\right)\right)^2} > 0,$$

where the last inequality holds by $0 \le z_\phi^\pi \le 1$. Further, z_ϕ^π is also increasing in $\pi(\phi\,|\,0)$ (see the fact $\underline{p} \ge \underline{q}$ in (D.183)), the optimal robust policy in state 0 thus obeys

$$\pi_\phi^\star(\phi\,|\,0) = 1. \tag{D.224}$$

Considering that the action does not influence the state transition for all states $s > 0$, without loss of generality, we choose the optimal robust policy to obey

$$\forall s > 0: \quad \pi_\phi^\star(\phi\,|\,s) = 1. \tag{D.225}$$

Taking $\pi = \pi_\phi^\star$ and $z_\phi^{\pi_\phi^\star} = \underline{p}$ in (D.223), we complete the proof by showing the corresponding optimal robust value function at state 0 as follows:

$$V_\phi^{\star,\sigma}(0) = \frac{\gamma z_\phi^{\pi_\phi^\star}}{(1-\gamma)\left(1-\gamma\left(1-z_\phi^{\pi_\phi^\star}\right)\right)} = \frac{\gamma\underline{p}}{(1-\gamma)\left(1-\gamma\left(1-\underline{p}\right)\right)}.$$

D.5.3.3 Proof of the claim (D.191)

Plugging in the definition of φ, we arrive at that for any policy π,

$$\begin{aligned}
\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \right\rangle &= V_\phi^{\star,\sigma}(0) - V_\phi^{\pi,\sigma}(0) \\
&\overset{\text{(i)}}{=} \frac{\gamma\underline{p}}{(1-\gamma)\left(1-\gamma(1-\underline{p})\right)} - \frac{\gamma z_\phi^\pi}{(1-\gamma)\left(1-\gamma(1-z_\phi^\pi)\right)} \\
&= \frac{\gamma\left(\underline{p}-z_\phi^\pi\right)}{\left(1-\gamma(1-\underline{p})\right)\left(1-\gamma(1-z_\phi^\pi)\right)} \overset{\text{(ii)}}{\ge} \frac{\gamma\left(\underline{p}-z_\phi^\pi\right)}{\left(1-\gamma\left(1-\underline{p}\right)\right)^2} \overset{\text{(iii)}}{=} \frac{\gamma(\underline{p}-\underline{q})\left(1-\pi(\phi\,|\,0)\right)}{\left(1-\gamma\left(1-\underline{p}\right)\right)^2},
\end{aligned} \tag{D.226}$$

where (i) holds by applying Lemma 54, (ii) arises from $z_\phi^\pi \le \underline{p}$ (see the definition of z_ϕ^π in (D.186) and the fact $\underline{p} \ge \underline{q} + \frac{3\Delta}{4}$ in (D.183)), and (iii) follows from the definition of z_ϕ^π in (D.186).

To further control (D.226), we consider it in two cases separately:

- Case 1: $\sigma \in \left(0, \frac{1-\gamma}{4}\right)$. In this case, applying Lemma 53 to (D.226) yields

$$\begin{aligned}\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \right\rangle &\ge \frac{\gamma(\underline{p}-\underline{q})\big(1-\pi(\phi\,|\,0)\big)}{\big(1-\gamma\left(1-\underline{p}\right)\big)^2} \ge \frac{\gamma\frac{3\Delta}{4}\big(1-\pi(\phi\,|\,0)\big)}{\left(1-\gamma\left(1-\frac{5(1-\gamma)}{4}\right)\right)^2} \\ &\ge \frac{\Delta\big(1-\pi(\phi\,|\,0)\big)}{9(1-\gamma)^2} = 2\varepsilon\big(1-\pi(\phi\,|\,0)\big), \end{aligned} \tag{D.227}$$

  where the penultimate inequality follows from $\gamma \ge 3/4$, and the last inequality holds by taking the specification of Δ in (D.190) as follows:

$$\Delta = 18(1-\gamma)^2\varepsilon. \tag{D.228}$$

  It is easily verified that taking $\varepsilon \le \frac{1}{72(1-\gamma)}$ as in (D.189) directly leads to meeting the requirement in (D.181), i.e., $\Delta \le \frac{1}{4}(1-\gamma)$.

- Case 2: $\sigma \in \left[\frac{1-\gamma}{4}, \infty\right)$. Similarly, applying Lemma 53 to (D.226) gives

$$\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \right\rangle \ge \frac{\gamma(\underline{p}-\underline{q})\big(1-\pi(\phi\,|\,0)\big)}{\big(1-\gamma\left(1-\underline{p}\right)\big)^2} \ge \frac{\gamma\frac{\sigma+1}{2}\Delta\big(1-\pi(\phi\,|\,0)\big)}{\min\left\{1, \left(1-\gamma\left(1-(3+\sigma)\Delta\right)\right)^2\right\}} \tag{D.229}$$

  Before continuing, it can be verified that

$$\begin{aligned} 1-\gamma\left(1-(3+\sigma)\Delta\right) &= 1-\gamma+\gamma(3+\sigma)\Delta \overset{\text{(i)}}{\le} 1-\gamma+(3+\sigma)\min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(\sigma+1)}\right\} \\ &\le \min\left\{2(1+\sigma)(1-\gamma), \frac{3}{2}\right\}, \end{aligned} \tag{D.230}$$

  where (i) is obtained by $\Delta \le \min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(1+\sigma)}\right\}$ (see (D.180)). Applying the above fact to (D.229) gives

$$\begin{aligned}\left\langle \varphi, V_\phi^{\star,\sigma} - V_\phi^{\pi,\sigma} \right\rangle &\ge \frac{\gamma\frac{\sigma+1}{2}\Delta\big(1-\pi(\phi\,|\,0)\big)}{\min\left\{1, \left(1-\gamma\left(1-(3+\sigma)\Delta\right)\right)^2\right\}} \overset{\text{(i)}}{\ge} \frac{3(\sigma+1)\Delta\big(1-\pi(\phi\,|\,0)\big)}{8\min\left\{4(1+\sigma)^2(1-\gamma)^2, 1\right\}} \\ &\ge \frac{\Delta\big(1-\pi(\phi\,|\,0)\big)}{\min\left\{32(1+\sigma)(1-\gamma)^2, \frac{8}{3(1+\sigma)}\right\}} = 2\varepsilon\big(1-\pi(\phi\,|\,0)\big), \end{aligned} \tag{D.231}$$

  where (i) holds by $\gamma \ge \frac{3}{4}$ and (D.229), and the last equality holds by the specification in

(D.190):

$$\Delta = \begin{cases} 64(1+\sigma)(1-\gamma)^2\varepsilon & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right), \\ \frac{16}{3(1+\sigma)}\varepsilon & \text{if } \sigma > \frac{1}{3(1-\gamma)}. \end{cases} \tag{D.232}$$

As a result, it is easily verified that the requirement in (D.181)

$$\Delta \leq \min\left\{\frac{1}{4}(1-\gamma), \frac{1}{2(1+\sigma)}\right\} \tag{D.233}$$

is met if we let

$$\varepsilon \leq \begin{cases} \frac{1}{256(1+\sigma)(1-\gamma)} & \text{if } \sigma \in \left[\frac{1-\gamma}{4}, \frac{1}{3(1-\gamma)}\right), \\ \frac{3}{32} & \text{if } \sigma > \frac{1}{3(1-\gamma)}, \end{cases} \tag{D.234}$$

as in (D.189).

The proof is then completed by summing up the results in the above two cases.

Appendix E

Proofs for Chapter 7

E.1 Preliminaries

Before starting, let us introduce some additional notation useful throughout the theoretical analysis. Let $\operatorname{ess\,inf} X$ denote the essential infimum of a function/variable X.

E.1.1 Properties of the robust Bellman operator

To begin with, we introduce the following strong duality lemma which is widely used in distributionally robust optimization when the uncertainty set is defined with respect to the KL divergence.

Lemma 56 ((Hu and Hong, 2013), Theorem 1)**.** *Suppose $f(x)$ has a finite moment generating function in some neighborhood around $x = 0$, then for any $\sigma > 0$ and a nominal distribution P^0, we have*

$$\sup_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathbb{E}_{X\sim\mathcal{P}}[f(X)] = \inf_{\lambda\geq 0}\left\{\lambda\log\mathbb{E}_{X\sim P^0}\left[\exp\left(\frac{f(X)}{\lambda}\right)\right] + \lambda\sigma\right\}. \tag{E.1}$$

Armed with the above lemma, it is easily verified that for any positive constant M and a nominal distribution vector $P^0 \in \mathbb{R}^{1\times S}$ supported over the state space $\mathcal{S}$, if $X(s) \in [0, M]$ for all $s \in \mathcal{S}$, then

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0)} \mathcal{P}X = \sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0\exp\left(-\frac{X}{\lambda}\right)\right) - \lambda\sigma\right\}. \tag{E.2}$$

For convenience, we introduce the following lemma, paraphrased from Zhou et al. (2021, Lemma 4) and its proof, to further characterize several essential properties of the optimal dual value.

Lemma 57 ((Zhou et al., 2021))**.** *Let $X \sim P$ be a bounded random variable with $X \in [0, M]$. Let $\sigma > 0$ be any uncertainty level and the corresponding optimal dual variable be*

$$\lambda^\star \in \arg\max_{\lambda\geq 0}\ f(\lambda, P), \qquad \text{where } f(\lambda, P) \coloneqq \left\{-\lambda\log\mathbb{E}_{X\sim P}\left[\exp\left(\frac{-X}{\lambda}\right)\right] - \lambda\sigma\right\}. \tag{E.3}$$

Then the optimal value $\lambda^\star$ obeys

$$\lambda^\star \in \left[0, \frac{M}{\sigma}\right], \tag{E.4}$$

where $\lambda^\star = 0$ *if and only if*

$$\log\big(\mathbb{P}(X = \operatorname{essinf} X)\big) + \sigma \geq 0. \tag{E.5}$$

Moreover, when $\lambda^\star = 0$*, we have*

$$\lim_{\lambda \to 0} f(\lambda, P) = \lim_{\lambda \to 0} \left\{ -\lambda \log \mathbb{E}_{X \sim P} \left[\exp\left(\frac{-X}{\lambda} \right) \right] - \lambda\sigma \right\} = \operatorname{essinf} X. \tag{E.6}$$

E.1.2 Concentration inequalities

In light of Lemma 57 (cf. E.6), we are interested in comparing the values of $\operatorname{essinf} X$ when X is drawn from the population nominal distribution or its empirical estimate. This is supplied by the following lemma from Zhou et al. (2021).

Lemma 58 ((Zhou et al., 2021))**.** *Let* $X \sim P$ *be a discrete bounded random variable with* $X \in [0, M]$*. Let* P_n *denote the empirical distribution constructed from* n *independent samples* $X_1, X_2, \cdots, X_n$*, and let* $\widehat{X} \sim P_n$*. Denote* $P_{\mathsf{min},X}$ *as the smallest positive probability* $P_{\mathsf{min},X} \coloneqq \min\{\mathbb{P}(X = x) : x \in \mathsf{supp}(X)\}$*, where* $\mathsf{supp}(X)$ *is the support of* X*. Then for any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*, we have*

$$\min_{i \in [n]} X_i = \operatorname{essinf} \widehat{X} = \operatorname{essinf} X, \tag{E.7}$$

as long as

$$n \geq -\frac{\log(2/\delta)}{\log(1 - P_{\mathsf{min},X})}. \tag{E.8}$$

We next gather a elementary fact about the Binomial distribution, which will be useful throughout the proof.

Lemma 59 (Chernoff's inequality)**.** *Suppose* $N \sim \mathsf{Binomial}(n, p)$*, where* $n \geq 1$ *and* $p \in [0, 1)$*. For some universal constant* $c_{\mathsf{f}} > 0$*, we have*

$$\mathbb{P}\left(|N/n - p| \geq pt\right) \leq \exp\left(-c_{\mathsf{f}} n p t^2\right), \qquad \forall t \in [0, 1]. \tag{E.9}$$

E.1.3 Kullback-Leibler (KL) divergence

We next introduce some useful facts about the Kullback-Leibler (KL) divergence for two distributions P and Q, denoted as $\mathsf{KL}(P \parallel Q)$. Denoting $\mathsf{Ber}(p)$(resp. $\mathsf{Ber}(q)$) as the Bernoulli distribution with mean p (resp. q), we introduce

$$\mathsf{KL}\big(\mathsf{Ber}(p) \parallel \mathsf{Ber}(q)\big) \coloneqq p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q}, \tag{E.10}$$

which represents the KL divergence from $\mathsf{Ber}(p)$ to $\mathsf{Ber}(q)$. We now introduce the following lemma.

Lemma 60. *For any $p, q \in \left[\frac{1}{2}, 1\right)$ and $p > q$, it holds that*

$$\mathsf{KL}\big(\mathsf{Ber}(p) \,\|\, \mathsf{Ber}(q)\big) \le \mathsf{KL}\big(\mathsf{Ber}(q) \,\|\, \mathsf{Ber}(p)\big) \le \frac{(p-q)^2}{p(1-p)}. \tag{E.11}$$

Moreover, for any $0 \le x < y < q$, it holds

$$\mathsf{KL}\left(\mathsf{Ber}\left(x\right) \,\|\, \mathsf{Ber}(q)\right) > \mathsf{KL}\left(\mathsf{Ber}\left(y\right) \,\|\, \mathsf{Ber}(q)\right). \tag{E.12}$$

Proof. The first half of this lemma is proven in Li et al. (2022a, Lemma 10). For the latter half, it follows from that the function

$$f(x, q) \coloneqq \mathsf{KL}\left(\mathsf{Ber}\left(x\right) \,\|\, \mathsf{Ber}(q)\right)$$

is monotonically decreasing for all $x \in (0, q]$, since its derivative with respect to x satisfies $\frac{\partial f(x,q)}{\partial x} = \log \frac{x}{q} + \log \frac{1-q}{1-x} < 0$. □

E.2 Analysis: episodic finite-horizon RMDPs

E.2.1 Proof of Theorem 14

Before starting, we introduce several additional notation that will be useful in the analysis. First, we denote the state-action space covered by the behavior policy π^{b} in the nominal model P^0 as

$$\mathcal{C}^{\mathsf{b}} = \left\{(h, s, a) : d_h^{\mathsf{b}, P^0}(s, a) > 0\right\}. \tag{E.13}$$

Moreover, we recall the definition in (7.15) and define a similar one based on the exact nominal model P^0 as

$$P_{\mathsf{min},h}(s, a) \coloneqq \min_{s'} \left\{P_h^0(s' \,|\, s, a) : \; P_h^0(s' \,|\, s, a) > 0\right\}. \tag{E.14}$$

Clearly, by comparing with the definitions (7.16) and (7.17), it holds that

$$P_{\mathsf{min}}^{\star} = \min_{h,s} \; P_{\mathsf{min},h}(s, \pi_h^{\star}(s)), \qquad P_{\mathsf{min}}^{\mathsf{b}} = \min_{(h,s,a) \in \mathcal{C}^{\mathsf{b}}} \; P_{\mathsf{min},h}(s, a). \tag{E.15}$$

For any time step $h \in [H]$, we denote the set of possible state occupancy distributions associated with the optimal policy $\pi^{\star}$ in a model within the uncertainty set $P \in \mathcal{U}^{\sigma}\left(P^0\right)$ as

$$\mathcal{D}_h^{\star} \coloneqq \left\{\left[d_h^{\star, P}(s)\right]_{s \in \mathcal{S}} : P \in \mathcal{U}^{\sigma}\left(P^0\right)\right\} = \left\{\left[d_h^{\star, P}\big(s, \pi_h^{\star}(s)\big)\right]_{s \in \mathcal{S}} : P \in \mathcal{U}^{\sigma}\left(P^0\right)\right\}, \tag{E.16}$$

where the second equality is due to the fact that $\pi^\star$ is chosen to be deterministic.

With these in place, the proof of Theorem 14 is separated into several key steps, as outlined below.

Step 1: establishing the pessimism property. To achieve this claim, we heavily count on the following lemma whose proof can be found in Appendix E.2.2.

Lemma 61. *Instate the assumptions in Theorem 14. Then for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, consider any vector $V \in \mathbb{R}^S$ independent of $\widehat{P}^0_{h,s,a}$ obeying $\|V\|_\infty \le H$. With probability at least $1-\delta$, one has*

$$\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}V - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V \right| \le b_h(s,a) \tag{E.17}$$

with $b_h(s,a)$ given in (7.14). Moreover, for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$, with probability at least $1-\delta$, one has

$$\frac{P_{\mathsf{min},h}(s,a)}{8\log(KHS/\delta)} \le \widehat{P}_{\mathsf{min},h}(s,a) \le e^2 P_{\mathsf{min},h}(s,a). \tag{E.18}$$

Armed with the above lemma, with probability at least $1-\delta$, we shall show the following relation holds

$$\forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H+1]: \qquad \widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi},\sigma}(s,a), \qquad \widehat{V}_h(s) \le V_h^{\widehat{\pi},\sigma}(s), \tag{E.19}$$

which means that $\widehat{Q}_h$ (resp. $\widehat{V}_h$) is a pessimistic estimate of $Q_h^{\widehat{\pi},\sigma}$ (resp. $V_h^{\widehat{\pi},\sigma}$). Towards this, it is easily verified that the latter assertion concerning $V_h^{\widehat{\pi},\sigma}$ is implied by the former, since

$$\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a) \le \max_a Q_h^{\widehat{\pi},\sigma}(s,a) = V_h^{\widehat{\pi},\sigma}(s). \tag{E.20}$$

Therefore, the remainder of this step focuses on verifying the former assertion in (E.19) by induction.

- To begin, the claim (E.19) holds at the base case when $h = H+1$, by invoking the trivial fact $\widehat{Q}_{H+1}(s,a) = Q_{H+1}^{\widehat{\pi},\sigma}(s,a) = 0$.
- Then, suppose that $\widehat{Q}_{h+1}(s,a) \le Q_{h+1}^{\widehat{\pi},\sigma}(s,a)$ holds for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ at some time step $h \in [H]$, it boils down to show $\widehat{Q}_h(s,a) \le Q_h^{\widehat{\pi},\sigma}(s,a)$.

  By the update rule of $\widehat{Q}_h(s,a)$ in Algorithm 13 (cf. line 7), the above relation holds immediately if $\widehat{Q}_h(s,a) = 0$ since $\widehat{Q}_h(s,a) = 0 \le Q_h^{\widehat{\pi},\sigma}(s,a)$. Otherwise, $\widehat{Q}_h(s,a)$ is updated via

$$\begin{aligned}
\widehat{Q}_h(s,a) &= r_h(s,a) + \sup_{\lambda \ge 0} \left\{ -\lambda \log\left(\widehat{P}^0_{h,s,a} \cdot \exp\left(\frac{-\widehat{V}_{h+1}}{\lambda} \right) \right) - \lambda\sigma \right\} - b_h(s,a) \\
&\overset{\text{(i)}}{=} r_h(s,a) + \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} - b_h(s,a)
\end{aligned}$$

$$\leq r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} + \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}\widehat{V}_{h+1} \right| - b_h(s,a)$$

$$\overset{\text{(ii)}}{\leq} r_h(s,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})} \mathcal{P}V^{\widehat{\pi},\sigma}_{h+1} + 0 \overset{\text{(iii)}}{=} Q^{\widehat{\pi},\sigma}_h(s,a), \tag{E.21}$$

where (i) rewrites the update rule back to its primal form (cf. (7.11)), (ii) holds by applying (E.17) with the condition (7.20) satisfied and the induction hypothesis $\widehat{V}_{h+1} \leq V^{\widehat{\pi},\sigma}_{h+1}$, and lastly, (iii) follows by the robust Bellman consistency equation (2.18).

Putting them together, we have verified the claim (E.19) by induction.

Step 2: bounding $V^{\star,\sigma}_h(s) - V^{\widehat{\pi},\sigma}_h(s)$. With the pessimism property (E.19) in place, we observe that the following relation holds

$$0 \leq V^{\star,\sigma}_h(s) - V^{\widehat{\pi},\sigma}_h(s) \leq V^{\star,\sigma}_h(s) - \widehat{V}_h(s) \leq Q^{\star,\sigma}_h\big(s,\pi^\star_h(s)\big) - \widehat{Q}_h\big(s,\pi^\star_h(s)\big), \tag{E.22}$$

where the last inequality follows from $\widehat{Q}_h\big(s,\pi^\star_h(s)\big) \leq \max_a \widehat{Q}_h(s,a) = \widehat{V}_h(s)$. Then, by the robust Bellman optimality equation in (2.19) and the primal version of the update rule (cf. (7.11))

$$Q^{\star,\sigma}_h\big(s,\pi^\star_h(s)\big) = r_h\big(s,\pi^\star_h(s)\big) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1},$$

$$\widehat{Q}_h\big(s,\pi^\star_h(s)\big) = r_h\big(s,\pi^\star_h(s)\big) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} - b_h\left(s,\pi^\star_h(s)\right),$$

we arrive at

$$\begin{aligned}
V^{\star,\sigma}_h(s) - \widehat{V}_h(s) &\leq Q^{\star,\sigma}_h\big(s,\pi^\star_h(s)\big) - \widehat{Q}_h\big(s,\pi^\star_h(s)\big) \\
&= \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} + b_h\big(s,\pi^\star_h(s)\big) \\
&\leq \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \\
&\quad + \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(\widehat{P}^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \right| + b_h\big(s,\pi^\star_h(s)\big) \\
&\overset{\text{(i)}}{\leq} \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} + 2b_h\big(s,\pi^\star_h(s)\big) \\
&\overset{\text{(ii)}}{\leq} \widehat{P}^{\inf}_{h,s,\pi^\star_h(s)}\big(V^{\star,\sigma}_{h+1} - \widehat{V}_{h+1}\big) + 2b_h\big(s,\pi^\star_h(s)\big),
\end{aligned} \tag{E.23}$$

where (i) holds by applying Lemma 2 (cf. (E.17)) since $\widehat{V}_{h+1}$ is independent of $P^0_{h,s,\pi^\star_h(s)}$ by construction, and (ii) arises from introducing the notation

$$\widehat{P}^{\inf}_{h,s,\pi^\star_h(s)} \coloneqq \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} \tag{E.24}$$

and consequently,

$$\inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}V^{\star,\sigma}_{h+1} \le \widehat{P}^{\inf}_{h,s,\pi^\star_h(s)} V^{\star,\sigma}_{h+1}, \qquad \text{and} \qquad \inf_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{h,s,\pi^\star_h(s)}\right)} \mathcal{P}\widehat{V}_{h+1} = \widehat{P}^{\inf}_{h,s,\pi^\star_h(s)} \widehat{V}_{h+1}.$$

To continue, let us introduce some additional notation for convenience. Define a sequence of matrices $\widehat{P}^{\inf}_h \in \mathbb{R}^{S\times S}$ and vectors $b^\star_h \in \mathbb{R}^S$ for $h\in[H]$, where their s-th rows (resp. entries) are given by

$$\left[\widehat{P}^{\inf}_h\right]_{s,\cdot} = \widehat{P}^{\inf}_{h,s,\pi^\star_h(s)}, \qquad \text{and} \qquad b^\star_h(s) = b_h\big(s,\pi^\star_h(s)\big). \tag{E.25}$$

Applying (E.23) recursively over the time steps $h, h+1, \cdots, H$ using the above notation gives

$$\begin{aligned} 0 \le V^{\star,\sigma}_h - \widehat{V}_h &\le \widehat{P}^{\inf}_h\big(V^{\star,\sigma}_{h+1} - \widehat{V}_{h+1}\big) + 2b^\star_h \\ &\le \widehat{P}^{\inf}_h \widehat{P}^{\inf}_{h+1}\big(V^{\star,\sigma}_{h+2} - \widehat{V}_{h+2}\big) + 2\widehat{P}^{\inf}_h b^\star_{h+1} + 2b^\star_h \le \cdots \le 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widehat{P}^{\inf}_j\right) b^\star_i, \end{aligned} \tag{E.26}$$

where we let $\left(\prod_{j=i}^{i-1}\widehat{P}^{\inf}_j\right) = I$ for convenience.

For any $d^\star_h \in \mathcal{D}^\star_h$ (cf. (E.16)), taking inner product with (E.26) leads to

$$\left\langle d^\star_h, V^{\star,\sigma}_h - \widehat{V}_h\right\rangle \le \left\langle d^\star_h, 2\sum_{i=h}^{H}\left(\prod_{j=h}^{i-1}\widehat{P}^{\inf}_j\right) b^\star_i\right\rangle = 2\sum_{i=h}^{H}\left\langle d^\star_i, b^\star_i\right\rangle, \tag{E.27}$$

where

$$d^\star_i \coloneqq \left[\left(d^\star_h\right)^\top\left(\prod_{j=h}^{i-1}\widehat{P}^{\inf}_j\right)\right]^\top \in \mathcal{D}^\star_i \tag{E.28}$$

by the definition of $\mathcal{D}^\star_i$ (cf. (E.16)) for all $i = h+1, \cdots, H$.

Step 3: controlling $\langle d^\star_i, b^\star_i\rangle$ using concentrability. Since $\langle d^\star_i, b^\star_i\rangle = \sum_{s\in\mathcal{S}} d^\star_i(s) b^\star_i(s)$, we shall divide the discussion in two different cases.

- For $s\in S$ where $\max_{P\in\mathcal{U}^\sigma(P^0)} d^{\star,P}_i\big(s,\pi^\star_i(s)\big) = \max_{P\in\mathcal{U}^\sigma(P^0)} d^{\star,P}_i(s) = 0$, it follows from the

definition (cf. (E.16)) that for any $d_i^\star \in \mathcal{D}_i^\star$, it satisfies that

$$d_i^\star(s) = 0. \tag{E.29}$$

- For $s \in S$ where $\max_{P\in\mathcal{U}^\sigma(P^0)} d_i^{\star,P}\big(s,\pi_i^\star(s)\big) = \max_{P\in\mathcal{U}^\sigma(P^0)} d_i^{\star,P}(s) > 0$, by the assumption in (7.5)

$$\max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\big\{d_i^{\star,P}\big(s,\pi_i^\star(s)\big),\frac{1}{S}\big\}}{d_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)} = \max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\big\{d_i^{\star,P}(s),\frac{1}{S}\big\}}{d_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)} \le C^\star_{\mathsf{rob}} < \infty,$$

it implies that

$$d_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big) > 0 \quad \text{and} \quad \big(i,s,\pi_i^\star(s)\big) \in \mathcal{C}^{\mathsf{b}}. \tag{E.30}$$

Lemma 21 tells that with probability at least $1-8\delta$,

$$\begin{aligned} N_i\big(s,\pi_i^\star(s)\big) &\ge \frac{Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)}{8} - 5\sqrt{Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)\log\frac{KH}{\delta}} \overset{\text{(i)}}{\ge} \frac{Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)}{16} \\ &\overset{\text{(ii)}}{\ge} \frac{K\max_{P\in\mathcal{U}^\sigma(P^0)}\min\Big\{d_i^{\star,P}\big(s,\pi_i^\star(s)\big),\frac{1}{S}\Big\}}{16C^\star_{\mathsf{rob}}} \ge \frac{K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}{16C^\star_{\mathsf{rob}}}, \end{aligned} \tag{E.31}$$

where (i) holds due to

$$Kd_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big) \ge c_1\frac{d_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big)\log(KHS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}}P^{\mathsf{b}}_{\mathsf{min}}} \ge \frac{c_1\log\frac{KH}{\delta}}{P^{\mathsf{b}}_{\mathsf{min}}} \ge c_1\log\frac{KH}{\delta} \tag{E.32}$$

for some sufficiently large c_1, where the first inequality follows from Condition (7.20), the second inequality follows from

$$d^{\mathsf{b}}_{\mathsf{min}} = \min_{h,s,a}\Big\{d_h^{\mathsf{b},P^0}(s,a) : d_h^{\mathsf{b},P^0}(s,a) > 0\Big\} \le d_i^{\mathsf{b},P^0}\big(s,\pi_i^\star(s)\big) \tag{E.33}$$

and the last inequality follows from $P^{\mathsf{b}}_{\mathsf{min}} \le 1$. In addition, (ii) follows from Assumption 5.

With this in place, we observe that the pessimistic penalty (see (7.14)) obeys

$$\begin{aligned} b_i^\star(s) &\le c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)N_i\big(s,\pi_i^\star(s)\big)}} \overset{\text{(i)}}{\le} 4c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{\log^2(\frac{KHS}{\delta})}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)N_i\big(s,\pi_i^\star(s)\big)}} \\ &\le 16c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{C^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}}, \end{aligned} \tag{E.34}$$

where (i) holds by applying (E.18) in view of the fact that $\big(i,s,\pi_i^\star(s)\big) \in \mathcal{C}^{\mathsf{b}}$ by (E.30), and

the last inequality holds by (E.31).

Combining the results in the above two cases leads to

$$\begin{aligned}\sum_{s\in\mathcal{S}} d_i^\star(s) b_i^\star(s) &\le \sum_{s\in\mathcal{S}} 16 d_i^\star(s) c_{\mathrm{b}} \frac{H}{\sigma}\sqrt{\frac{C^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}} \\ &\overset{\text{(i)}}{\le} 16c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\sum_{s\in\mathcal{S}} d_i^\star(s)\frac{C^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K\min\big\{d_i^\star(s),\frac{1}{S}\big\}}}\sqrt{\sum_{s\in\mathcal{S}} d_i^\star(s)} \\ &\le 32c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{SC^\star_{\mathsf{rob}}\log^2\frac{KHS}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K}},\end{aligned} \tag{E.35}$$

where (i) follows from the Cauchy-Schwarz inequality and the last inequality hold by the trivial fact

$$\sum_{s\in\mathcal{S}}\frac{d_i^\star(s)}{\min\big\{d_i^\star(s),\frac{1}{S}\big\}} \le \sum_{s\in\mathcal{S}} d_i^\star(s)\left(\frac{1}{d_i^\star(s)}+\frac{1}{1/S}\right) = \sum_{s\in\mathcal{S}} 1+\frac{1}{S}\sum_{s\in\mathcal{S}} d_i^\star(s) \le 2S. \tag{E.36}$$

Step 4: finishing up the proof. Then, inserting (E.35) back into (E.27) with $h=1$ shows

$$\left\langle d_1^\star, V_1^{\star,\sigma}-\widehat{V}_1\right\rangle \le 2\sum_{i=1}^H \langle d_i^\star, b_i^\star\rangle \le \sum_{i=1}^H 64c_{\mathrm{b}}\frac{H}{\sigma}\sqrt{\frac{SC^\star_{\mathsf{rob}}\log^2\frac{KH}{\delta}}{P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)K}} \le c_2\frac{H^2}{\sigma}\sqrt{\frac{SC^\star_{\mathsf{rob}}\log^2\frac{KH}{\delta}}{P^\star_{\mathsf{min}}K}}, \tag{E.37}$$

where the last inequality holds by plugging in the relation $P^\star_{\mathsf{min}} \le P_{\mathsf{min},i}\big(s,\pi_i^\star(s)\big)$ for $i=1,\ldots,H$ by the definition in (7.16) (see also (E.15)), and choosing c_2 to be large enough. The proof is completed.

E.2.2 Proof of Lemma 61

To begin, we shall introduce the following fact that

$$\forall(h,s,a)\in\mathcal{C}^{\mathsf{b}}: \qquad N_h(s,a) \ge \frac{c_1\log\frac{KHS}{\delta}}{16P_{\mathsf{min},h}(s,a)} \ge -\frac{\log\frac{2KHS}{\delta}}{\log(1-P_{\mathsf{min},h}(s,a))}, \tag{E.38}$$

as long as Condition (7.20) holds. The proof is postponed to Appendix E.2.2.3. With this in mind, we shall first establish the simpler bound (E.18) and then move on to show (E.17).

E.2.2.1 Proof of (E.18)

To begin, recall that (E.38) is satisfied for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$. By Lemma 27 and the union bound, it holds that with probability at least $1-\delta$ that for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$:

$$\forall s' \in \mathcal{S}: \qquad P_h^0(s' \,|\, s,a) \geq \frac{\widehat{P}_h^0(s' \,|\, s,a)}{e^2} \geq \frac{P_h^0(s' \,|\, s,a)}{8e^2 \log(\frac{KHS}{\delta})}. \tag{E.39}$$

To characterize the relation between $P_{\mathsf{min},h}(s,a)$ and $\widehat{P}_{\mathsf{min},h}(s,a)$ for any $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$, we suppose—without loss of generality—that $P_{\mathsf{min},h}(s,a) = P_h^0(s_1 \,|\, s,a)$ and $\widehat{P}_{\mathsf{min},h}(s,a) = \widehat{P}_h^0(s_2 \,|\, s,a)$ for some $s_1, s_2 \in \mathcal{S}$. Then, it follows that

$$\begin{aligned} P_{\mathsf{min},h}(s,a) = P_h^0(s_1 \,|\, s,a) &\overset{\text{(i)}}{\geq} \frac{\widehat{P}_h^0(s_1 \,|\, s,a)}{e^2} \geq \frac{\widehat{P}_{\mathsf{min},h}(s,a)}{e^2} = \frac{\widehat{P}_h^0(s_2 \,|\, s,a)}{e^2} \\ &\overset{\text{(ii)}}{\geq} \frac{P_h^0(s_2 \,|\, s,a)}{8e^2 \log(\frac{KHS}{\delta})} \geq \frac{P_{\mathsf{min},h}(s,a)}{8e^2 \log(\frac{KHS}{\delta})}, \end{aligned}$$

where (i) and (ii) follow from (E.39).

E.2.2.2 Proof of (E.17)

The main goal of (E.17) is to control the gap between robust Bellman operations based on the nominal transition kernel $P_{h,s,a}^0$ and the estimated kernel $\widehat{P}_{h,s,a}^0$ by the constructed penalty term. Towards this, first consider $(h,s,a) \notin \mathcal{C}^{\mathsf{b}}$, which corresponds to the state-action pairs (s,a) that haven't been visited at step h by the behavior policy. In other words, $N_h(s,a) = 0$. In this case, (E.17) can be easily verified that

$$\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}_{h,s,a}^0)} \mathcal{P}V - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P_{h,s,a}^0)} \mathcal{P}V \right| \overset{\text{(i)}}{=} \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P_{h,s,a}^0)} \mathcal{P}V \leq \|V\|_\infty \overset{\text{(ii)}}{\leq} H \overset{\text{(iii)}}{=} b_h(s,a), \tag{E.40}$$

where (i) follows from the fact $\widehat{P}_{h,s,a}^0 = 0$ when $N_h(s,a) = 0$ (see (7.8)), (ii) arises from the assumption $\|V\|_\infty \leq H$, and (iii) holds by the definition of $b_h(s,a)$ in (7.14). Therefore, the remainder of the proof will focus on verifying (E.17) for $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$. Rewriting the term of interest via duality (cf. Lemma 56) yields

$$\begin{aligned} &\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma(\widehat{P}_{h,s,a}^0)} \mathcal{P}V - \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P_{h,s,a}^0)} \mathcal{P}V \right| \\ &= \left| \sup_{\lambda \geq 0} \left\{ -\lambda \log\left(\widehat{P}_{h,s,a}^0 \exp\left(\frac{-V}{\lambda} \right) \right) - \lambda\sigma \right\} - \sup_{\lambda \geq 0} \left\{ -\lambda \log\left(P_{h,s,a}^0 \exp\left(\frac{-V}{\lambda} \right) \right) - \lambda\sigma \right\} \right|. \end{aligned} \tag{E.41}$$

Denoting

$$\widehat{\lambda}^{\star}_{h,s,a} \coloneqq \arg\max_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}, \tag{E.42a}$$

$$\lambda^{\star}_{h,s,a} \coloneqq \arg\max_{\lambda\geq 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}, \tag{E.42b}$$

Lemma 57 (cf. (E.4)) then gives that

$$\lambda^{\star}_{h,s,a}\in\left[0,\frac{H}{\sigma}\right], \qquad \widehat{\lambda}^{\star}_{h,s,a}\in\left[0,\frac{H}{\sigma}\right], \tag{E.43}$$

due to $\|V\|_\infty\leq H$. We shall control (E.41) in three different cases separately: (a) $\lambda^{\star}_{h,s,a}=0$ and $\widehat{\lambda}^{\star}_{h,s,a}=0$; (b) $\lambda^{\star}_{h,s,a}>0$ and $\widehat{\lambda}^{\star}_{h,s,a}=0$ or $\lambda^{\star}_{h,s,a}=0$ and $\widehat{\lambda}^{\star}_{h,s,a}>0$; and (c) $\lambda^{\star}_{h,s,a}\neq 0$ or $\widehat{\lambda}^{\star}_{h,s,a}\neq 0$.

Case (a): $\lambda^{\star}_{h,s,a}=0$ and $\widehat{\lambda}^{\star}_{h,s,a}=0$. Applying Lemma 57 and Lemma 58 to (E.41) gives that, with probability at least $1-\frac{\delta}{KH}$,

$$\begin{aligned}\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})}\mathcal{P}V\right| &\overset{\text{(i)}}{=}\left|\operatorname{essinf}_{s\sim\widehat{P}^0_{h,s,a}}V(s)-\operatorname{essinf}_{s\sim P^0_{h,s,a}}V(s)\right|\\ &\overset{\text{(ii)}}{=}\left|\operatorname{essinf}_{s\sim P^0_{h,s,a}}V(s)-\operatorname{essinf}_{s\sim P^0_{h,s,a}}V(s)\right|\\ &=0\leq b_h(s,a).\end{aligned} \tag{E.44}$$

where (i) holds by Lemma 57 (cf. (E.6)) and (ii) arises from Lemma 58 (cf. (E.7)) given (E.38).

Case (b): $\lambda^{\star}_{h,s,a}>0$ and $\widehat{\lambda}^{\star}_{h,s,a}=0$ or $\lambda^{\star}_{h,s,a}=0$ and $\widehat{\lambda}^{\star}_{h,s,a}>0$. Towards this, note that two trivial facts are implied by the definition (E.42):

$$\sup_{\lambda\geq 0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\geq-\widehat{\lambda}^{\star}_{h,s,a}\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^{\star}_{h,s,a}}\right)\right)-\widehat{\lambda}^{\star}_{h,s,a}\sigma, \tag{E.45a}$$

$$\sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\geq-\lambda^{\star}_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^{\star}_{h,s,a}}\right)\right)-\lambda^{\star}_{h,s,a}\sigma. \tag{E.45b}$$

To continue, first, we consider a subcase when $\lambda^{\star}_{h,s,a}=0$ and $\widehat{\lambda}^{\star}_{h,s,a}>0$. With probability at least $1-\frac{\delta}{KH}$, it follows from Lemma 57 (cf. (E.6)) and Lemma 58 (cf. (E.7)) that

$$\sup_{\lambda\geq 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\geq\lim_{\lambda\to 0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}$$

$$= \operatorname{essinf}_{s\sim \widehat{P}^0_{h,s,a}} V(s) = \operatorname{essinf}_{s\sim P^0_{h,s,a}} V(s)$$
$$= \sup_{\lambda\geq 0}\left\{-\lambda \log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\}, \tag{E.46}$$

leading to

$$\begin{aligned}
&\left|\sup_{\lambda\geq 0}\left\{-\lambda \log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\} - \sup_{\lambda\geq 0}\left\{-\lambda \log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right) - \lambda\sigma\right\}\right| \\
&\overset{\text{(i)}}{\leq} \left(-\widehat{\lambda}^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \widehat{\lambda}^\star_{h,s,a}\sigma\right) - \left(-\widehat{\lambda}^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \widehat{\lambda}^\star_{h,s,a}\sigma\right) \\
&\leq \widehat{\lambda}^\star_{h,s,a}\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right) - \log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}\right)\right)\right|,
\end{aligned} \tag{E.47}$$

where (i) follows from the definition of $\widehat{\lambda}^\star_{h,s,a}$ in (E.42) and the fact in (E.45a).

We pause to claim that with probability at least $1-\delta$, the following bound holds

$$\forall (h,s,a)\in\mathcal{C}^{\mathsf{b}},\, V\in\mathbb{R}^S:\quad \frac{\left|\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)} \leq \sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}}N_h(s,a)P_{\mathsf{min},h}(s,a)}} \leq \frac{1}{2}. \tag{E.48}$$

The proof is postponed to Appendix E.2.2.4. With (E.48) in place, we can further bound (E.47) (which is plugged into (E.41)) as

$$\begin{aligned}
\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})}\mathcal{P}V - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})}\mathcal{P}V\right| &\leq \widehat{\lambda}^\star_{h,s,a}\left|\log\left(1+\frac{\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)}\right)\right| \\
&\overset{\text{(i)}}{\leq} 2\widehat{\lambda}^\star_{h,s,a}\frac{\left|\left(\widehat{P}^0_{h,s,a}-P^0_{h,s,a}\right)\cdot\exp\left(\frac{-V}{\lambda}\right)\right|}{P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)} \\
&\overset{\text{(ii)}}{\leq} \frac{2H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}}N_h(s,a)P_{\mathsf{min},h}(s,a)}} \\
&\leq \frac{2eH}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}}N_h(s,a)\widehat{P}_{\mathsf{min},h}(s,a)}} \leq c_{\mathsf{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{\widehat{P}_{\mathsf{min},h}(s,a)N_h(s,a)}},
\end{aligned} \tag{E.49}$$

where (i) follows from $\log(1+x)\leq 2|x|$ for any $|x|\leq\frac{1}{2}$ in view of (E.48), (ii) follows from (E.43) as well as (E.48), and the last line follows from (E.18) and choosing c_{b} to be sufficiently large.

Moreover, note that it can be easily verified that

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})}\mathcal{P}V\right|\le H$$

due to the assumption $\|V\|_\infty\le H$. Plugging in the definition of $b_h(s,a)$ in (7.14), combined with the above bounds, we have that with probability at least $1-\delta$,

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})}\mathcal{P}V\right|\le\min\left\{c_{\mathsf{b}}\frac{H}{\sigma}\sqrt{\frac{\log(\frac{KHS}{\delta})}{N_h(s,a)\widehat{P}_{\mathsf{min},h}(s,a)}}\ ,\ H\right\}=:b_h(s,a).\quad\text{(E.50)}$$

The other subcase when $\lambda^\star_{h,s,a}>0$ and $\widehat{\lambda}^\star_{h,s,a}=0$ follows similarly from the bound

$$\begin{aligned}&\left|\sup_{\lambda\ge0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}-\sup_{\lambda\ge0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\right|\\&\le\lambda^\star_{h,s,a}\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right)-\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda^\star_{h,s,a}}\right)\right)\right|,\end{aligned}\quad\text{(E.51)}$$

and therefore, will be omitted for simplicity.

Case (c): $\lambda^\star_{h,s,a}>0$ and $\widehat{\lambda}^\star_{h,s,a}>0$. It follows that

$$\begin{aligned}&\left|\sup_{\lambda\ge0}\left\{-\lambda\log\left(\widehat{P}^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}-\sup_{\lambda\ge0}\left\{-\lambda\log\left(P^0_{h,s,a}\exp\left(\frac{-V}{\lambda}\right)\right)-\lambda\sigma\right\}\right|\\&\overset{\text{(i)}}{\le}\max\left\{\left(-\widehat{\lambda}^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot e^{\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}}\right)-\widehat{\lambda}^\star_{h,s,a}\sigma\right)-\left(-\widehat{\lambda}^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot e^{\frac{-V}{\widehat{\lambda}^\star_{h,s,a}}}\right)-\widehat{\lambda}^\star_{h,s,a}\sigma\right),\right.\\&\qquad\left.\left(-\lambda^\star_{h,s,a}\log\left(P^0_{h,s,a}\cdot e^{\frac{-V}{\lambda^\star_{h,s,a}}}\right)-\lambda^\star_{h,s,a}\sigma\right)-\left(-\lambda^\star_{h,s,a}\log\left(\widehat{P}^0_{h,s,a}\cdot e^{\frac{-V}{\lambda^\star_{h,s,a}}}\right)-\lambda^\star_{h,s,a}\sigma\right)\right\}\\&\le\max_{\lambda\in\{\lambda^\star_{h,s,a},\widehat{\lambda}^\star_{h,s,a}\}}\lambda\left|\log\left(\widehat{P}^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)\right)-\log\left(P^0_{h,s,a}\cdot\exp\left(\frac{-V}{\lambda}\right)\right)\right|,\end{aligned}\quad\text{(E.52)}$$

where (i) can be verified by applying the facts in (E.45). Hence, the above term (E.52) can be controlled again in a similar manner as (E.47); we omit the details for simplicity.

Summing up. Combining the previous results in different cases by the union bound, with probability at least $1-10\delta$, it is satisfied that for all $(h,s,a)\in\mathcal{C}^{\mathsf{b}}$:

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{h,s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{h,s,a})}\mathcal{P}V\right|\le b_h(s,a),$$

which concludes the proof.

E.2.2.3 Proof of (E.38)

Observe that for all $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$:

$$Kd_h^{\mathsf{b},P^0}(s,a) \overset{\text{(i)}}{\geq} \frac{c_1 d_h^{\mathsf{b},P^0}(s,a)\log(KHS/\delta)}{d_{\mathsf{min}}^{\mathsf{b}} P_{\mathsf{min}}^{\mathsf{b}}} \overset{\text{(ii)}}{\geq} \frac{c_1\log(KHS/\delta)}{P_{\mathsf{min}}^{\mathsf{b}}} \overset{\text{(iii)}}{\geq} \frac{c_1\log(KHS/\delta)}{P_{\mathsf{min},h}(s,a)}, \tag{E.53}$$

where (i) follows from Condition (7.20), (ii) follows from the definition that $d_{\mathsf{min}}^{\mathsf{b}} \leq d_h^{\mathsf{b},P^0}(s,a)$ for $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$, and (iii) comes from (E.15).

Lemma 21 then tells that with probability at least $1-8\delta$,

$$\begin{aligned} N_h(s,a) &\geq \frac{Kd_h^{\mathsf{b},P^0}(s,a)}{8} - 5\sqrt{Kd_h^{\mathsf{b},P^0}(s,a)\log\frac{KH}{\delta}} \\ &\geq \frac{Kd_i^{\mathsf{b},P^0}(s,a)}{16} \geq \frac{c_1\log\frac{KH}{\delta}}{16P_{\mathsf{min},h}(s,a)}, \end{aligned} \tag{E.54}$$

where the second line follows from the above relation as long as c_1 is sufficiently large. The last inequality of (E.38) then follows from

$$\frac{c_1\log\frac{KHS}{\delta}}{16P_{\mathsf{min},h}(s,a)} \geq -\frac{\log\frac{2KHS}{\delta}}{\log(1-P_{\mathsf{min},h}(s,a))}, \tag{E.55}$$

since $x \leq -\log(1-x)$ for all $x \in [0,1]$.

E.2.2.4 Proof of (E.48)

Denoting

$$\mathsf{supp}\big(P_{h,s,a}^0\big) \coloneqq \big\{s' \in \mathcal{S} : P_h^0(s' \,|\, s,a) > 0\big\}$$

as the support of $P_{h,s,a}^0$, we observe that

$$\begin{aligned} \frac{\left|\left(\widehat{P}_{h,s,a}^0 - P_{h,s,a}^0\right)\cdot\exp\left(\frac{-V}{\lambda}\right)\right|}{P_{h,s,a}^0\cdot\exp\left(\frac{-V}{\lambda}\right)} &\leq \frac{\sum_{s'\in\mathsf{supp}\left(P_{h,s,a}^0\right)}\left|\widehat{P}_h^0(s' \,|\, s,a) - P_h^0(s' \,|\, s,a)\right|\exp\left(\frac{-V(s')}{\lambda}\right)}{\sum_{s'\in\mathsf{supp}\left(P_{h,s,a}^0\right)}P_h^0(s' \,|\, s,a)\exp\left(\frac{-V(s')}{\lambda}\right)} \\ &\leq \max_{s'\in\mathsf{supp}\left(P_{h,s,a}^0\right)}\frac{\left|\widehat{P}_h^0(s' \,|\, s,a) - P_h^0(s' \,|\, s,a)\right|}{P_h^0(s' \,|\, s,a)}, \end{aligned} \tag{E.56}$$

where the second line follows from $\sum_i a_i = \sum_i b_i\frac{a_i}{b_i} \leq (\max_i \frac{a_i}{b_i})\sum_i b_i$ for any positive sequences $\{a_i,b_i\}_i$ obeying $a_i,b_i>0$.

To continue, note that for any $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$ and $s' \in \mathsf{supp}\big(P^0_{h,s,a}\big)$, $N_h(s,a)\widehat{P}^0_h(s' \,|\, s,a)$ follows the binomial distribution $\mathsf{Binomial}\left(N_h(s,a), P^0_h(s' \,|\, s,a)\right)$. Thus, applying Lemma 59 with $t = \sqrt{\frac{\log\left(\frac{KHS}{\delta}\right)}{c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a)}}$ yields

$$\mathbb{P}\left(\left|\widehat{P}^0_h(s' \,|\, s,a) - P^0_h(s' \,|\, s,a)\right| \geq P^0_h(s' \,|\, s,a)t\right) \leq \exp\left(-c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a) t^2\right) \leq \frac{\delta}{KHS}, \quad \text{(E.57)}$$

as soon as $t \leq \frac{1}{2}$, which can be verified by the fact (E.38) and $P_{\mathsf{min},h}(s,a) \leq P^0_h(s' \,|\, s,a)$ (cf. (E.14)), namely,

$$N_h(s,a) \geq \frac{c_1 \log \frac{KHS}{\delta}}{16 P_{\mathsf{min},h}(s,a)} \geq \frac{\log\left(\frac{KHS}{\delta}\right)}{4 c_{\mathsf{f}} P_{\mathsf{min},h}(s,a)} \geq \frac{\log\left(\frac{KHS}{\delta}\right)}{4 c_{\mathsf{f}} P^0_h(s' \,|\, s,a)} \quad \text{(E.58)}$$

as long as c_1 is sufficiently large.

Applying (E.57) and taking the union bound over $s \in \mathsf{supp}\big(P^0_{h,s,a}\big)$ lead to that with probability at least $1 - \frac{\delta}{KH}$,

$$\begin{aligned}
\max_{s' \in \mathsf{supp}\left(P^0_{h,s,a}\right)} \frac{\left|\widehat{P}^0_h(s' \,|\, s,a) - P^0_h(s' \,|\, s,a)\right|}{P^0_h(s' \,|\, s,a)} &\leq \max_{s' \in \mathsf{supp}\left(P^0_{h,s,a}\right)} \frac{P^0_h(s' \,|\, s,a)\sqrt{\frac{\log\left(\frac{KHS}{\delta}\right)}{c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a)}}}{P^0_h(s' \,|\, s,a)} \\
&= \max_{s' \in \mathsf{supp}\left(P^0_{h,s,a}\right)} \sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}} N_h(s,a) P^0_h(s' \,|\, s,a)}} \\
&\leq \sqrt{\frac{\log(\frac{KHS}{\delta})}{c_{\mathsf{f}} N_h(s,a) P_{\mathsf{min},h}(s,a)}} \leq \frac{1}{2},
\end{aligned}$$

where the last line uses again (E.58). Plugging this back into (E.56) and applying the union bound over $(h,s,a) \in \mathcal{C}^{\mathsf{b}}$ then completes the proof.

E.2.3 Proof of Theorem 15

The proof of Theorem 15 is inspired by the construction in Li et al. (2022a) for standard MDPs, but is considerably more involved to handle the uncertainty set unique in robust MDPs. We shall first construct some hard instances and then characterize the sample complexity requirements over these instances.

E.2.3.1 Construction of hard problem instances

Construction of a collection of hard MDPs. Let us introduce two MDPs

$$\left\{\mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi = \{P^\phi_h\}_{h=1}^H, \{r_h\}_{h=1}^H, H\right) \,|\, \phi = \{0,1\}\right\}, \quad \text{(E.59)}$$

where the state space is $\mathcal{S} = \{0, 1, \ldots, S-1\}$, and the action space is $\mathcal{A} = \{0, 1\}$. The transition kernel P^ϕ of the constructed MDP $\mathcal{M}_\phi$ is defined as

$$P_1^\phi(s' \,|\, s, a) = \begin{cases} p\mathbb{1}(s'=0) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,\phi) \\ q\mathbb{1}(s'=0) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0,1-\phi) \\ \mathbb{1}(s'=1) & \text{if} \quad s=1 \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s>1 \end{cases} \tag{E.60a}$$

and

$$P_h^\phi(s' \,|\, s, a) = \mathbb{1}(s'=s), \qquad \forall (h,s,a) \in \{2, \ldots, H\} \times \mathcal{S} \times \mathcal{A}. \tag{E.60b}$$

In words, except at step $h=1$, the MDP always stays in the same state. Additionally, the MDP will always stay in the state subset $\{0,1\}$ if the initial distribution is supported only on $\{0,1\}$, in view of (E.60). Here, p and q are set to be

$$p = 1-\alpha \qquad \text{and} \qquad q = 1-\alpha-\Delta \tag{E.61}$$

for some $H \geq 2e^8$, α and Δ obeying

$$0 < \alpha \leq \frac{1}{H} \leq \frac{1}{2e^8} \qquad \text{and} \qquad \Delta \leq \frac{\alpha}{2} \leq \frac{1}{2H} \leq \frac{1}{4e^8}, \tag{E.62}$$

where β is set as

$$\beta := \frac{\log \frac{1}{\alpha+\Delta}}{2} \geq \frac{\log(2H/3)}{2} \geq 4. \tag{E.63}$$

The assumption (E.62) immediately indicates the facts

$$1 > p > q \geq \frac{1}{2}. \tag{E.64}$$

Moreover, for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the reward function is defined as

$$r_h(s,a) = \begin{cases} 1 & \text{if } s=0 \\ 0 & \text{otherwise} \end{cases}. \tag{E.65}$$

Construction of the history/batch dataset. In the nominal environment $\mathcal{M}_\phi$, a batch dataset is generated consisting of K *independent* sample trajectories each of length H, where each trajectory is generated according to (7.3), based on the following initial state distribution ρ^{b} and behavior

policy $\pi^{\mathsf{b}} = \{\pi_h^{\mathsf{b}}\}_{h=1}^H$:

$$\rho^{\mathsf{b}}(s) = \mu(s) \quad \text{and} \quad \pi_h^{\mathsf{b}}(a \,|\, s) = \frac{1}{2}, \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{E.66}$$

Here, $\mu(s)$ is defined as the following state distribution supported on the state subset $\{0,1\}$:

$$\mu(s) = \frac{1}{CS}\mathbb{1}(s=0) + \Big(1 - \frac{1}{CS}\Big)\mathbb{1}(s=1), \tag{E.67}$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $C > 0$ is some constant that will determine the concentrability coefficient $C^\star_{\mathsf{rob}}$ (as we shall detail momentarily) and obeys

$$\frac{1}{CS} \le \frac{1}{4}. \tag{E.68}$$

As it turns out, for any MDP $\mathcal{M}_\phi$, the occupancy distributions of the above batch dataset are the same (due to symmetry) and admit the following simple characterization:

$$d_1^{\mathsf{b},P^\phi}(0,a) = \frac{1}{2}\mu(0), \qquad \forall a \in \mathcal{A}, \tag{E.69a}$$

$$\frac{\mu(s)}{2} \le d_h^{\mathsf{b},P^\phi}(s) \le 2\mu(s), \qquad \frac{\mu(s)}{4} \le d_h^{\mathsf{b},P^\phi}(s,a) \le \mu(s), \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{E.69b}$$

In addition, we choose the following initial state distribution

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{if } s > 0 \end{cases}. \tag{E.70}$$

The proof of the claim (E.69) is postponed to Appendix E.2.3.3.

Uncertainty set of the transition kernels. Denote the transition kernel vector as

$$P^\phi_{h,s,a} \coloneqq P^\phi_h(\cdot \,|\, s,a) \in [0,1]^{1\times S}. \tag{E.71}$$

For any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the perturbation of the transition kernels in $\mathcal{M}_\phi$ is restricted to the following uncertainty set

$$\mathcal{U}^\sigma(P^\phi) \coloneqq \otimes\, \mathcal{U}^\sigma\left(P^\phi_{h,s,a}\right), \qquad \mathcal{U}^\sigma(P^\phi_{h,s,a}) \coloneqq \left\{P_{h,s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{h,s,a} \,\|\, P^\phi_{h,s,a}\right) \le \sigma\right\}, \tag{E.72}$$

where the radius of the uncertainty set σ obeys

$$\left(1 - \frac{3}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \le \sigma \le \left(1 - \frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right). \tag{E.73}$$

Before continuing, we shall introduce some notation for convenience. For any $P_h^\phi(\cdot \,|\, s,a)$ in (E.60), we define the limit of the perturbed kernel transiting to the next state s' from the current state-action pair (s,a) by

$$\underline{P}_h^\phi(s' \,|\, s,a) \coloneqq \inf_{P_{h,s,a} \in \mathcal{U}^\sigma(P^\phi_{h,s,a})} P_h(s' \,|\, s,a), \tag{E.74}$$

and in particular, denote

$$\underline{p} \coloneqq \underline{P}_1^\phi(0 \,|\, 0,\phi), \qquad \underline{q} = \underline{P}_1^\phi(0 \,|\, 0,1-\phi). \tag{E.75}$$

Armed with the above definitions, we introduce the following lemma which implies some useful properties of the uncertainty set.

Lemma 62. *When β satisfies (E.63) and the uncertainty level σ satisfies (E.73), the perturbed transition kernels obey*

$$\underline{p} \geq \underline{q} \geq \frac{1}{\beta}. \tag{E.76}$$

Proof. See Appendix E.2.3.4. □

Value functions and optimal policies. We take a moment to derive the corresponding value functions and identify the optimal policies. With some abuse of notation, for any MDP $\mathcal{M}_\phi$, we denote $\pi^{\star,\phi} = \{\pi_h^{\star,\phi}\}_{h=1}^H$ as the optimal policy, and let $V_h^{\pi,\sigma,\phi}$ (resp. $V_h^{\star,\sigma,\phi}$) represent the robust value function of policy π (resp. $\pi^{\star,\phi}$) at step h with uncertainty radius σ. Armed with these notation, we introduce the following lemma which collects the properties concerning the value functions and optimal policies.

Lemma 63. *For any $\phi = \{0,1\}$ and any policy π, defining*

$$z_\phi^\pi \coloneqq \underline{p}\pi_1(\phi \,|\, 0) + \underline{q}\pi_1(1-\phi \,|\, 0), \tag{E.77}$$

it holds that

$$V_1^{\pi,\sigma,\phi}(0) = 1 + z_\phi^\pi(H-1). \tag{E.78}$$

In addition, the optimal policies and the optimal value functions obey

$$V_1^{\star,\sigma,\phi}(0) = 1 + \underline{p}(H-1), \tag{E.79a}$$

$$\forall h \in [H] \setminus \{1\}: \quad V_h^{\star,\sigma,\phi}(0) = H - h + 1, \tag{E.79b}$$

$$\forall h \in [H]: \quad \pi_h^{\star,\phi}(\phi \,|\, 0) = 1, \quad \pi_h^{\star,\phi}(\phi \,|\, 1) = 1, \qquad V_h^{\star,\sigma,\phi}(1) = 0. \tag{E.79c}$$

The robust single-policy clipped concentrability coefficient $C^\star_{\mathsf{rob}}$ *obeys*

$$2C \le C^\star_{\mathsf{rob}} \le 4C. \tag{E.80}$$

Proof. See Appendix E.2.3.5. □

In view of Lemma 63, we note that the smallest positive state transition probability of the optimal policy $\pi^\star$ under any MDP $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$ thus can be given by

$$P^\star_{\mathsf{min}} \coloneqq \min_{h,s,s'} \left\{ P_h^\phi\left(s'|s,\pi_h^{\star,\phi}(s)\right) :\ P_h^\phi\left(s'|s,\pi_h^{\star,\phi}(s)\right) > 0 \right\} = P_1^\phi\left(1|0,1-\phi\right) = 1-p, \tag{E.81}$$

which obeys

$$\alpha = P^\star_{\mathsf{min}} \in (0, 1/H]$$

according to (E.61) and (E.62).

E.2.3.2 Establishing the minimax lower bound

We are now ready to establish the sample complexity lower bound. With the choice of the initial distribution ρ in (E.70), for any policy estimator $\widehat{\pi}$ computed based on the batch dataset, we plan to control the quantity

$$\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \right\rangle = V_1^{\star,\sigma,\phi}(0) - V_1^{\widehat{\pi},\sigma,\phi}(0).$$

Step 1: converting the goal to estimate ϕ. We make the following claim which shall be verified in Appendix E.2.3.6: given $\varepsilon \le \frac{H}{384e^6\log\left(\frac{1}{\alpha}\right)} \le \frac{H}{384e^6\log\left(\frac{1}{\alpha+\Delta}\right)}$, choosing

$$\Delta = \frac{128e^6\sigma(1-q)\varepsilon}{H} = \frac{128e^6\sigma(\alpha+\Delta)\varepsilon}{H} \le \frac{128e^6(\alpha+\Delta)\varepsilon\log\left(\frac{1}{\alpha+\Delta}\right)}{H} \le \frac{\alpha}{2}, \tag{E.82}$$

which satisfies (E.62) with the aid of (E.73) and (E.61), it holds that for any policy $\widehat{\pi}$,

$$\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \right\rangle \ge 2\varepsilon\big(1 - \widehat{\pi}_1(\phi\,|\,0)\big). \tag{E.83}$$

Armed with this relation between the policy $\widehat{\pi}$ and its sub-optimality gap, we are positioned to construct an estimate of ϕ. We denote $\mathbb{P}_\phi$ as the probability distribution when the MDP is $\mathcal{M}_\phi$, for any $\phi \in \{0,1\}$.

Suppose for the moment that a policy estimate $\widehat{\pi}$ achieves

$$\mathbb{P}_\phi\left\{ \left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\widehat{\pi},\sigma,\phi} \right\rangle \le \varepsilon \right\} \ge \frac{7}{8}, \tag{E.84}$$

then in view of (E.83), we necessarily have $\widehat{\pi}_1(\phi \,|\, 0) \geq \frac{1}{2}$ with probability at least $\frac{7}{8}$. With this in mind, we are motivated to construct the following estimate $\widehat{\phi}$ for $\phi \in \{0, 1\}$:

$$\widehat{\phi} = \arg \max_{a \in \{0,1\}} \widehat{\pi}_1(a \,|\, 0), \tag{E.85}$$

which obeys

$$\mathbb{P}_\phi\big\{\widehat{\phi} = \phi\big\} \geq \mathbb{P}_\phi\big\{\widehat{\pi}_1(\phi \,|\, 0) > 1/2\big\} \geq \frac{7}{8}. \tag{E.86}$$

In what follows, we would like to show (E.86) cannot happen without enough samples, which would in turn contradict (E.83).

Step 2: probability of error in testing two hypotheses. Armed with the above preparation, we shall focus on differentiating the two hypotheses $\phi \in \{0, 1\}$. Towards this, consider the minimax probability of error defined as follows:

$$p_{\mathrm{e}} \coloneqq \inf_{\psi} \max\big\{\mathbb{P}_0(\psi \neq 0),\, \mathbb{P}_1(\psi \neq 1)\big\}, \tag{E.87}$$

where the infimum is taken over all possible tests ψ constructed from the batch dataset.

Let $\mu^{\mathsf{b},\phi}$ (resp. $\mu_h^{\mathsf{b},\phi}(s_h)$) be the distribution of a sample trajectory $\{s_h, a_h\}_{h=1}^H$ (resp. a sample (a_h, s_{h+1}) conditional on s_h) for the MDP $\mathcal{M}_\phi$. Following standard results from Tsybakov (2009, Theorem 2.2) and the additivity of the KL divergence (cf. Tsybakov (2009, Page 85)), we obtain

$$\begin{aligned} p_{\mathrm{e}} &\geq \frac{1}{4} \exp\Big(- K \mathsf{KL}\big(\mu^{\mathsf{b},0} \,\|\, \mu^{\mathsf{b},1}\big)\Big) \\ &\geq \frac{1}{4} \exp\bigg\{ - \frac{1}{2} K \mu(0) \Big(\mathsf{KL}\big(P_1^0(\cdot \,|\, 0, 0) \,\|\, P_1^1(\cdot \,|\, 0, 0)\big) + \mathsf{KL}\big(P_1^0(\cdot \,|\, 0, 1) \,\|\, P_1^1(\cdot \,|\, 0, 1)\big)\Big)\bigg\}, \end{aligned} \tag{E.88}$$

where we also use the independence of the K trajectories in the batch dataset in the first line. Here, the second line arises from the chain rule of the KL divergence (Duchi, 2018, Lemma 5.2.8) and the Markov property of the sample trajectories (recall that $d_h^{\mathsf{b},P^0} = d_h^{\mathsf{b},P^1}$) according to

$$\begin{aligned} \mathsf{KL}\big(\mu^{\mathsf{b},0} \,\|\, \mu^{\mathsf{b},1}\big) &= \sum_{h=1}^H \mathop{\mathbb{E}}_{s_h \sim d_h^{\mathsf{b},P^0}} \Big[\mathsf{KL}\big(\mu_h^{\mathsf{b},0}(s_h) \,\|\, \mu_h^{\mathsf{b},1}(s_h)\big)\Big] \\ &= \sum_{a \in \{0,1\}} d_1^{\mathsf{b},P^0}(0, a) \mathsf{KL}\left(P_1^0(\cdot \,|\, 0, a) \,\|\, P_1^1(\cdot \,|\, 0, a)\right) \\ &= \frac{1}{2}\mu(0) \sum_{a \in \{0,1\}} \mathsf{KL}\big(P_1^0(\cdot \,|\, 0, a) \,\|\, P_1^1(\cdot \,|\, 0, a)\big), \end{aligned}$$

where the penultimate equality holds by the fact that $P_h^0(\cdot \,|\, s, a)$ and $P_h^1(\cdot \,|\, s, a)$ only differ when

$h = 1$ and $s = 0$, and the last equality follows from (E.69).

It remains to control the KL divergence terms in (E.88). Given $p \geq q \geq 1/2$ (cf. (E.64)), applying Lemma 60 (cf. (E.11)) yields

$$\begin{aligned}\mathsf{KL}\big(P_1^0(\cdot\,|\,0,0)\;\|\;P_1^1(\cdot\,|\,0,0)\big) = \mathsf{KL}\left(p\;\|\;q\right) \leq \frac{(p-q)^2}{(1-p)p} &\overset{\text{(i)}}{=} \frac{\Delta^2}{p(1-p)} \\ &\overset{\text{(ii)}}{=} \frac{128^2 e^{12}\sigma^2(1-q)^2\varepsilon^2}{H^2 p(1-p)} \\ &\overset{\text{(iii)}}{\leq} \frac{c_1\sigma^2 P_{\mathsf{min}}^\star \varepsilon^2}{H^2},\end{aligned} \tag{E.89}$$

where (i) follows from the definition (E.61), (ii) holds by plugging in the expression of Δ in (E.82), (iii) arises from $1 - q \leq 2(1-p) = 2P_{\mathsf{min}}^\star$ (see (E.62) and (E.81)), $p > \frac{1}{2}$, as long as c_1 is a large enough constant. It can be shown that $\mathsf{KL}\big(P_1^0(\cdot\,|\,0,1)\;\|\;P_1^1(\cdot\,|\,0,1)\big)$ can be upper bounded in the same way. Substituting (E.89) back into (E.88) demonstrates that: if the sample size is chosen as

$$KH \leq \frac{H^3 S C_{\mathsf{rob}}^\star \log 2}{4c_1 P_{\mathsf{min}}^\star \sigma^2 \varepsilon^2}, \tag{E.90}$$

then one necessarily has

$$\begin{aligned}p_{\mathrm{e}} \geq \frac{1}{4}\exp\left\{-\frac{1}{2}K\mu(0)\cdot 2\frac{c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{H^2}\right\} &\overset{\text{(i)}}{=} \frac{1}{4}\exp\left\{-K\frac{c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{SCH^2}\right\} \\ &\overset{\text{(ii)}}{\geq} \frac{1}{4}\exp\left\{-K\frac{4c_1\sigma^2 P_{\mathsf{min}}^\star\varepsilon^2}{SC_{\mathsf{rob}}^\star H^2}\right\} \geq \frac{1}{8},\end{aligned} \tag{E.91}$$

where (i) follows from (E.67) and (ii) holds by (E.80).

Step 3: putting things together. Finally, suppose that there exists an estimator $\widehat{\pi}$ such that

$$\mathbb{P}_0\big\{\big\langle\rho, V_1^{\star,\sigma,0} - V_1^{\widehat{\pi},\sigma,0}\big\rangle > \varepsilon\big\} < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big\{\big\langle\rho, V_1^{\star,\sigma,1} - V_1^{\widehat{\pi},\sigma,1}\big\rangle > \varepsilon\big\} < \frac{1}{8}.$$

Then Step 1 tells us that the estimator $\widehat{\phi}$ defined in (E.85) must satisfy

$$\mathbb{P}_0\big(\widehat{\phi} \neq 0\big) < \frac{1}{8} \qquad \text{and} \qquad \mathbb{P}_1\big(\widehat{\phi} \neq 1\big) < \frac{1}{8},$$

which cannot happen under the sample size condition (E.90) to avoid contradition with (E.91). The proof is thus finished.

E.2.3.3 Proof of (E.69)

With the initial state distribution and behavior policy defined in (E.66), we have for any MDP $\mathcal{M}_\phi$ with $\phi \in \{0, 1\}$,

$$d_1^{\mathsf{b},P^\phi}(s) = \rho^{\mathsf{b}}(s) = \mu(s),$$

which leads to

$$\forall a \in \mathcal{A}: \quad d_1^{\mathsf{b},P^\phi}(0, a) = \frac{1}{2}\mu(0). \tag{E.92}$$

In view of (E.60a), the state occupancy distribution at step $h = 2$ obeys

$$d_2^{\mathsf{b},P^\phi}(0) = \mathbb{P}\left\{s_2 = 0 \,|\, s_1 \sim d_1^{\mathsf{b},P^\phi}; \pi^{\mathsf{b}}\right\} = \mu(0)\left[\pi_1^{\mathsf{b}}(\phi \,|\, 0)p + \pi_1^{\mathsf{b}}(1-\phi \,|\, 0)q\right] = \frac{(p+q)\mu(0)}{2},$$

and

$$\begin{aligned} d_2^{\mathsf{b},P^\phi}(1) &= \mathbb{P}\left\{s_2 = 1 \,|\, s_1 \sim d_1^{\mathsf{b},P^\phi}; \pi^{\mathsf{b}}\right\} \\ &= \mu(0)\left[\pi_1^{\mathsf{b}}(\phi \,|\, 0)(1-p) + \pi_1^{\mathsf{b}}(1-\phi \,|\, 0)(1-q)\right] + \mu(1) = \mu(1) + \frac{(2-p-q)\mu(0)}{2}. \end{aligned}$$

With the above result in mind and recalling the assumption in (E.64), we arrive at

$$\frac{\mu(0)}{2} \le d_2^{\mathsf{b},P^\phi}(0) \le \mu(0), \qquad \mu(1) \le d_2^{\mathsf{b},P^\phi}(1) \overset{\text{(i)}}{\le} 2\mu(1), \tag{E.93}$$

where (i) holds by applying (E.64) and (E.68) (which implies $\mu(0) \le \mu(1)$ by the assumption in (E.68))

$$d_2^{\mathsf{b},P^\phi}(1) = \mu(1) + \frac{(2-p-q)\mu(0)}{2} \le \mu(1) + \mu(0) \le 2\mu(1).$$

Finally, from the definitions of $P_h^\phi(\cdot \,|\, s, a)$ in (E.60b) and the Markov property, we arrive at for any $(h, s) \in [H] \times \mathcal{S}$,

$$\frac{\mu(s)}{2} \le d_h^{\mathsf{b},P^\phi}(s) \le 2\mu(s), \tag{E.94}$$

which directly leads to

$$\frac{\mu(s)}{4} \le d_h^{\mathsf{b},P^\phi}(s, a) = \pi_1^{\mathsf{b}}(a \,|\, s) d_h^{\mathsf{b},P^\phi}(s) \le \mu(s). \tag{E.95}$$

E.2.3.4 Proof of Lemma 62

Note that $\underline{p} \geq \underline{q}$ can be easily verified since $p > q$, which indicates that the first assertion is true. So we will focus on the second assertion in (E.76). Towards this, invoking the definition in (E.10), let σ' be the KL divergence from $\mathsf{Ber}\left(\frac{1}{\beta}\right)$ to $\mathsf{Ber}(q)$, defined as follows

$$\begin{aligned}\sigma' &:= \mathsf{KL}\left(\mathsf{Ber}\left(\frac{1}{\beta}\right) \parallel \mathsf{Ber}(q)\right) = \frac{1}{\beta}\log\frac{\frac{1}{\beta}}{q} + \left(1-\frac{1}{\beta}\right)\log\frac{\left(1-\frac{1}{\beta}\right)}{1-q} \\ &= \left(\frac{1}{\beta}\right)\log\left(\frac{1}{\beta}\right) - \left(\frac{1}{\beta}\right)\log(q) + \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) + \left(1-\frac{1}{\beta}\right)\log\left(1-\frac{1}{\beta}\right), \end{aligned} \tag{E.96}$$

where the second line uses the definition of q in (E.61). We claim that σ' satisfies the following relation with σ, which will be proven at the end of this proof:

$$\left(1-\frac{3}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \leq \sigma \leq \left(1-\frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \leq \sigma' \leq \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right). \tag{E.97}$$

Recalling the definition of the transition kernel in (E.60a)

$$P_1^{\phi}(0\,|\,0,1-\phi) = q, \quad P_1^{\phi}(1\,|\,0,1-\phi) = 1-q, \quad P_1^{\phi}(s\,|\,0,1-\phi) = 0, \quad \forall s \in \mathcal{S}\setminus\{0,1\},$$

the uncertainty set of the transition kernel with radius σ is thus given as

$$\mathcal{U}^{\sigma}(P^{\phi}_{1,0,1-\phi}) = \left\{P_{1,0,1-\phi} \in \Delta(\mathcal{S}) : P(0\,|\,0,1-\phi) = q', P(1\,|\,0,1-\phi) = 1-q', \mathsf{KL}\left(\mathsf{Ber}\left(q'\right) \parallel \mathsf{Ber}(q)\right) \leq \sigma\right\}. \tag{E.98}$$

Recalling the definition of $\underline{q}$ in (E.75), we can bound

$$\begin{aligned}\underline{q} &= \inf_{P_{1,0,1-\phi}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,1-\phi})} P(0\,|\,0,1-\phi) = \inf_{q':\mathsf{KL}(\mathsf{Ber}(q')\|\mathsf{Ber}(q))\leq\sigma} q' \\ &\overset{\text{(i)}}{\geq} \inf_{q':\mathsf{KL}(\mathsf{Ber}(q')\|\mathsf{Ber}(q))\leq\sigma'} q' = \frac{1}{\beta},\end{aligned}$$

where (i) holds by $\sigma \leq \sigma'$ (cf. (E.97)) and the last equality follows from applying Lemma 60 (cf. (E.12)) and (E.96) to arrive at

$$\forall 0 \leq q' < \frac{1}{\beta}: \qquad \mathsf{KL}\left(\mathsf{Ber}\left(q'\right) \parallel \mathsf{Ber}(q)\right) > \mathsf{KL}\left(\mathsf{Ber}\left(\frac{1}{\beta}\right) \parallel \mathsf{Ber}(q)\right) = \sigma'.$$

Proof of (E.97)**.** To control σ', we plug in the assumptions in (E.64) and $\beta \geq 4$ and arrive at the trivial facts

$$\left(\frac{1}{\beta}\right)\log\left(\frac{1}{\beta}\right) - \left(\frac{1}{\beta}\right)\log(q) < 0, \quad \left(1-\frac{1}{\beta}\right)\log\left(1-\frac{1}{\beta}\right) < 0.$$

The above facts directly lead to

$$\sigma' \leq \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right). \tag{E.99}$$

Similarly, observing

$$-1 \leq \left(\frac{1}{\beta}\right)\log\left(\frac{1}{\beta}\right) + \left(1-\frac{1}{\beta}\right)\log\left(1-\frac{1}{\beta}\right) \leq 0, \quad -\left(\frac{1}{\beta}\right)\log(q) \geq 0,$$

we arrive at

$$\sigma' \geq -1 + \left(1-\frac{1}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \geq \left(1-\frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) \tag{E.100}$$

as long as $\log\left(\frac{1}{\alpha+\Delta}\right) \geq \beta$ (cf. (E.63)). With (E.99) and (E.100) in hand, it is straightforward to see that the choice of the uncertainty radius σ in (E.73) obeys the advertised bound (E.97).

E.2.3.5 Proof of Lemma 63

For notational conciseness, we shall drop the superscript ϕ and use the shorthand $V_h^{\pi,\sigma} = V_h^{\pi,\sigma,\phi}$ and $V_h^{\star,\sigma} = V_h^{\star,\sigma,\phi}$ whenever it is clear from the context. We begin by deriving the robust value function for any policy π. Starting with state 1, at any step $h \in [H]$, it obeys

$$V_h^{\pi,\sigma}(1) = \mathbb{E}_{a\sim\pi_h(\cdot\,|\,1)}\left[r_h(1,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{h,1,a})} \mathcal{P}V_{h+1}^{\pi,\sigma}\right] = 0 + V_{h+1}^{\pi,\sigma}(1),$$

where the first equality follows from the robust Bellman consistency equation (cf. (2.18)), and the second equality follows from the observation that the distribution $P^\phi_{h,1,a}$ is supported solely on state 1 in view of (E.60a), therefore $\mathcal{U}^\sigma(P^\phi_{h,1,a}) = P^\phi_{h,1,a}$. Leveraging the terminal condition $V_{H+1}^{\pi,\sigma}(1) = 0$, and recursively applying the previous relation, we have

$$V_h^{\star,\sigma}(1) = V_h^{\pi,\sigma}(1) = 0, \qquad \forall h \in [H]. \tag{E.101}$$

Similarly, turning to state 0, at any step $h > 1$, the robust value function satisfies

$$V_h^{\pi,\sigma}(0) = \mathbb{E}_{a\sim\pi_h(\cdot\,|\,0)}\left[r_h(0,a) + \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{h,0,a})} \mathcal{P}V_{h+1}^{\pi,\sigma}\right] = 1 + V_{h+1}^{\pi,\sigma}(0),$$

which again uses the fact that the distribution $P^{\phi}_{h,0,a}$ is supported solely on state 0 in view of (E.60b), therefore $\mathcal{U}^{\sigma}(P^{\phi}_{h,0,a}) = P^{\phi}_{h,0,a}$. Leveraging the terminal condition $V^{\pi,\sigma}_{H+1}(0) = 0$, and recursively applying the previous relation, we have

$$V^{\star,\sigma}_{h}(0) = V^{\pi,\sigma}_{h}(0) = H - h + 1, \qquad 2 \le h \le H. \tag{E.102}$$

Taking (E.101) and (E.102) together, it follows that

$$\forall\, 2 \le h \le H: \qquad V^{\pi,\sigma}_{h}(0) > V^{\pi,\sigma}_{h}(1). \tag{E.103}$$

Consequently, the robust value function of state 0 at step $h = 1$ satisfies

$$\begin{aligned}
V^{\pi,\sigma}_{1}(0) &= \mathbb{E}_{a\sim\pi_1(\cdot\,|\,0)}\left[r_1(0,a) + \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,a})} \mathcal{P}V^{\pi,\sigma}_{2}\right] \\
&\overset{\text{(i)}}{=} 1 + \pi_1(\phi\,|\,0)\Big(\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,\phi})} \mathcal{P}V^{\pi,\sigma}_{2}\Big) + \pi_1(1-\phi\,|\,0)\Big(\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,1-\phi})} \mathcal{P}V^{\pi,\sigma}_{2}\Big) \\
&\overset{\text{(ii)}}{=} 1 + \pi_1(\phi\,|\,0)\Big[\underline{p}V^{\pi,\sigma}_{2}(0) + \big(1-\underline{p}\big)\,V^{\pi,\sigma}_{2}(1)\Big] + \pi_1(1-\phi\,|\,0)\Big[\underline{q}V^{\pi,\sigma}_{2}(0) + \big(1-\underline{q}\big)\,V^{\pi,\sigma}_{2}(1)\Big] \\
&\overset{\text{(iii)}}{=} 1 + V^{\pi,\sigma}_{2}(1) + z^{\pi}_{\phi}\left[V^{\pi,\sigma}_{2}(0) - V^{\pi,\sigma}_{2}(1)\right] \\
&= 1 + z^{\pi}_{\phi}V^{\pi,\sigma}_{2}(0)
\end{aligned} \tag{E.104}$$

where (i) uses the definition of the reward function in (E.65), (ii) uses (E.103) so that the infimum is attained by picking the choice specified in (E.75) with a smallest probability mass imposed on the transition to state 0. Finally, we plug in the definition (E.77) of z^{π}_{ϕ} in (iii), and the last line follows from (E.101).

Therefore, taking $\pi = \pi^{\star,\phi}$ in the previous relation directly leads to

$$V^{\star,\sigma}_{1}(0) = 1 + z^{\pi^{\star,\phi}}_{\phi}V^{\star,\sigma}_{2}(0) = 1 + z^{\pi^{\star,\phi}}_{\phi}(H-1), \tag{E.105}$$

where the second equality follows from (E.102). Observing that the function $(H-1)z$ is increasing in z and that z^{π}_{ϕ} is increasing in $\pi_1(\phi\,|\,0)$ (due to the fact $\underline{p} \ge \underline{q}$ in (E.76)). As a result, the optimal policy obeys

$$\pi^{\star,\phi}_{1}(\phi\,|\,0) = 1 \tag{E.106}$$

at state 0, and plugging back to (E.105) gives

$$V^{\star,\sigma}_{1}(0) = 1 + z^{\pi^{\star,\phi}}_{\phi}(H-1) = 1 + \underline{p}(H-1),$$

where $z^{\pi^{\star,\phi}}_{\phi} = \underline{p}\pi^{\star,\phi}_{1}(\phi\,|\,0) + \underline{q}\pi^{\star,\phi}_{1}(1-\phi\,|\,0) = \underline{p}$. For the rest of the states, without loss of generality,

we choose the optimal policy obeying

$$\forall h \in [H]: \qquad \pi_h^{\star,\phi}(\phi \,|\, 0) = 1, \quad \pi_h^{\star,\phi}(\phi \,|\, 1) = 1. \tag{E.107}$$

Proof of claim (E.80). Given that $\pi_h^{\star,\phi}(\phi \,|\, 0) = 1$ for all $h \in [H]$ and $\rho(0) = 1$, for any $P \in \mathcal{U}^\sigma(P^\phi)$, we have

$$\begin{aligned} d_2^{\star,P}(0,\phi) &= d_2^{\star,P}(0)\pi_2^{\star,\phi}(\phi \,|\, 0) = d_2^{\star,P}(0) = \mathbb{P}_{s_2 \sim P(\cdot \,|\, s_1, \pi_1^{\star,\phi}(s_1))}\big\{s_2 = 0 \,|\, s_1 \sim \rho; \pi^{\star,\phi}\big\} \\ &= P_1(0 \,|\, 0, \phi) \overset{\text{(i)}}{\geq} \underline{P}_1^{\phi}(0 \,|\, 0, \phi) \overset{\text{(ii)}}{=} \underline{p} \geq \frac{1}{\beta}, \end{aligned} \tag{E.108}$$

which (i) holds by plugging in the definition (E.74), (ii) follows from the definition (E.75), and the final inequality arises from Lemma 62. Hence, for all $2 \leq h \leq H$, by the Markov property and $P_h^{\phi}(0 \,|\, 0, \phi) = 1$, we have

$$d_h^{\star,P}(0,\phi) = d_2^{\star,P}(0,\phi) \geq \frac{1}{\beta}. \tag{E.109}$$

Examining the definition of $C_{\mathsf{rob}}^{\star}$ in (7.5), we make the following observations.

- For $h = 1$, we have

$$\begin{aligned} \max_{(s,a,P) \in \mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_1^{\star,P}(s,a), \frac{1}{S}\big\}}{d_1^{\mathsf{b},P^\phi}(s,a)} &\overset{\text{(i)}}{=} \max_{P \in \mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_1^{\star,P}(0,\phi), \frac{1}{S}\big\}}{d_1^{\mathsf{b},P^\phi}(0,\phi)} \overset{\text{(ii)}}{=} \max_{P \in \mathcal{U}^\sigma(P^\phi)} \frac{1}{S d_1^{\mathsf{b},P^\phi}(0,\phi)} \\ &\overset{\text{(iii)}}{=} \frac{2}{S\mu(0)} = 2C, \end{aligned} \tag{E.110}$$

  where (i) holds by $d_1^{\star,P}(s) = \rho(s) = 0$ for all $s \in \mathcal{S} \setminus \{0\}$ (see (E.70)) and $\pi_h^{\star,\phi}(\phi \,|\, 0) = 1$ for all $h \in [H]$, (ii) follows from the fact $d_1^{\star,P}(0,\phi) = 1$, (iii) is verified in (E.69), and the last equality arises from the definition in (E.67).

- Similarly, for $h = 2$, we arrive at

$$\begin{aligned} \max_{(s,a,P) \in \mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_2^{\star,P}(s,a), \frac{1}{S}\big\}}{d_2^{\mathsf{b},P^\phi}(s,a)} &\overset{\text{(i)}}{=} \max_{s \in \{0,1\}, P \in \mathcal{U}^\sigma(P^\phi)} \frac{\min\big\{d_2^{\star,P}(s,\phi), \frac{1}{S}\big\}}{d_2^{\mathsf{b},P^\phi}(s,\phi)} \\ &\leq \max_{s \in \{0,1\}, P \in \mathcal{U}^\sigma(P^\phi)} \frac{1}{S d_2^{\mathsf{b},P^\phi}(s,\phi)} \overset{\text{(ii)}}{\leq} \frac{4}{S\mu(0)} = 4C, \end{aligned} \tag{E.111}$$

  where (i) holds by the optimal policy in (E.79) and the trivial fact that $d_2^{\star,P}(s) = 0$ for all $s \in \mathcal{S} \setminus \{0,1\}$ (see (E.70) and (E.60a)), (ii) arises from (E.69), and the last equality comes from (E.67).

- For all other steps $h = 3, \ldots, H$, observing from the deterministic transition kernels in (E.60b), it can be easily verified that

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^{\phi}}(s,a)} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_2^{\star,P}(s,a), \frac{1}{S}\right\}}{d_2^{\mathsf{b},P^{\phi}}(s,a)} \le 4C. \tag{E.112}$$

Combining the above cases, we complete the proof by

$$2C \le C^{\star}_{\mathsf{rob}} = \max_{(h,s,a,P)\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{U}^{\sigma}(P^{\phi})} \frac{\min\left\{d_h^{\star,P}(s,a), \frac{1}{S}\right\}}{d_h^{\mathsf{b},P^{\phi}}(s,a)} \le 4C.$$

E.2.3.6 Proof of the claim (E.83)

Recall that by virtue of (E.77) and (E.79), we arrive at

$$z^{\star}_{\phi} \coloneqq z^{\pi^{\star,\phi}}_{\phi} = \underline{p}\pi_1^{\star,\phi}(\phi\,|\,0) + \underline{q}\pi_1^{\star,\phi}(1-\phi\,|\,0) = \underline{p}.$$

Applying (E.78) yields

$$\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\pi,\sigma,\phi} \right\rangle = V_h^{\star,\sigma,\phi}(0) - V_h^{\pi,\sigma,\phi}(0) = \left(\underline{p} - z^{\pi}_{\phi}\right)(H-1) = \left(\underline{p} - \underline{q}\right)(H-1)\left(1 - \pi_1(\phi\,|\,0)\right), \tag{E.113}$$

where the last equality uses the definition (E.77). Therefore, it boils down to control $\underline{p} - \underline{q}$.

To continue, we define an auxiliary value function vector $\overline{V} \in \mathbb{R}^{S\times 1}$ obeying

$$\overline{V}(0) = H - 1 \quad \text{and} \quad \overline{V}(s) = 0, \quad \forall s \in \mathcal{S}\setminus\{0\}. \tag{E.114}$$

With this in hand, applying Lemma 56 gives

$$\begin{aligned}
&(H-1)\left(\underline{p} - \underline{q}\right) \\
&\overset{\text{(i)}}{=} \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,\phi})} \mathcal{P}\overline{V} - \inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,1-\phi})} \mathcal{P}\overline{V} \\
&= \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^{\phi}_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\} - \sup_{\lambda\ge 0}\left\{-\lambda\log\left(P^{\phi}_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right\} \\
&\overset{\text{(ii)}}{\ge} \left\{-\lambda^{\star}\log\left(P^{\phi}_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^{\star}}\right)\right) - \lambda^{\star}\sigma\right\} - \left\{-\lambda^{\star}\log\left(P^{\phi}_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^{\star}}\right)\right) - \lambda^{\star}\sigma\right\} \\
&= -\lambda^{\star}\left[\log\left(P^{\phi}_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^{\star}}\right)\right) - \log\left(P^{\phi}_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^{\star}}\right)\right)\right],
\end{aligned} \tag{E.115}$$

where (i) follows from (see the definition of $\underline{p}$ in (E.75))

$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,\phi})} \mathcal{P}\overline{V} = \underline{P}^{\phi}_{1}(0\,|\,0,\phi)\overline{V}(0) = (H-1)\underline{p},$$
$$\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{\phi}_{1,0,1-\phi})} \mathcal{P}\overline{V} = \underline{P}^{\phi}_{1}(0\,|\,0,1-\phi)\overline{V}(0) = (H-1)\underline{q}.$$

Here, (ii) holds by letting

$$\lambda^{\star} \coloneqq \arg\max_{\lambda\geq 0} f(\lambda) \coloneqq \arg\max_{\lambda\geq 0}\left\{-\lambda\log\left(P^{\phi}_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda}\right)\right)-\lambda\sigma\right\}. \tag{E.116}$$

The rest of the proof is then to control (E.115). We start with the observation that $\lambda^{\star}>0$; this is because in view of Lemma 57 (cf. (E.5)), it suffices to verify that

$$\log(1-q)+\sigma \overset{\text{(i)}}{\leq} \log(\alpha+\Delta)+\left(1-\frac{2}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right) = -\frac{2}{\beta}\log\left(\frac{1}{\alpha+\Delta}\right) < 0, \tag{E.117}$$

where (i) holds by (E.73). We now claim the following bound for $\lambda^{\star}$ holds, whose proof is postponed to the end:

$$\frac{H}{16\sigma} \leq \frac{H-1}{\log\left(\frac{\beta}{\alpha+\Delta}\right)} \leq \lambda^{\star} \leq \frac{H-1}{\left(1-\frac{3}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right)}, \tag{E.118}$$

which immediately implies the following by taking exponential maps given $\lambda^{\star}>0$:

$$\frac{\alpha+\Delta}{\beta} \leq e^{-(H-1)/\lambda^{\star}} \leq (\alpha+\Delta)^{1-3/\beta}. \tag{E.119}$$

Moving to the second term of (E.115), it follows that

$$\begin{aligned}
\log\left(P^{\phi}_{1,0,\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^{\star}}\right)\right)-\log\left(P^{\phi}_{1,0,1-\phi}\cdot\exp\left(\frac{-\overline{V}}{\lambda^{\star}}\right)\right) &\overset{\text{(i)}}{=} \log\frac{pe^{-(H-1)/\lambda^{\star}}+(1-p)}{qe^{-(H-1)/\lambda^{\star}}+(1-q)} \\
&= \log\left(1+\frac{(p-q)\left(e^{-(H-1)/\lambda^{\star}}-1\right)}{qe^{-(H-1)/\lambda^{\star}}+(1-q)}\right) \\
&\overset{\text{(ii)}}{<} -\frac{\Delta\left(1-e^{-(H-1)/\lambda^{\star}}\right)}{qe^{-(H-1)/\lambda^{\star}}+(1-q)} \\
&\overset{\text{(iii)}}{\leq} -\frac{1}{2}\frac{\Delta}{\left(\frac{1}{\alpha+\Delta}\right)^{\frac{3}{\beta}}(1-q)+(1-q)} \\
&\leq -\frac{\Delta}{4e^{6}(1-q)},
\end{aligned} \tag{E.120}$$

where (i) follows from the definitions in (E.60) and (E.114), (ii) holds by $\log(1+x) < x$ for $x \in (-1, \infty)$, (iii) can be verified by (E.119), $\beta \geq 4$, and (E.62):

$$1 - e^{-(H-1)/\lambda^\star} \geq 1 - (\alpha+\Delta)^{1-3/\beta} \geq 1 - (\alpha+\Delta)^{1/4} \geq 1 - \left(\frac{3}{2H}\right)^{1/4} \geq \frac{1}{2},$$

and the last line uses $\left(\frac{1}{\alpha+\Delta}\right)^{\frac{3}{\beta}} = \left(\frac{1}{\alpha+\Delta}\right)^{6/\log\left(\frac{1}{\alpha+\Delta}\right)} = e^6$ by the definition of β in (E.63). Plugging (E.118) and (E.120) back into (E.115) and (E.113), we arrive at

$$\begin{aligned}\left\langle \rho, V_1^{\star,\sigma,\phi} - V_1^{\pi,\sigma,\phi} \right\rangle &= (H-1)\left(\underline{p} - \underline{q}\right)(1 - \pi_1(\phi \,|\, 0)) \\ &\overset{\text{(i)}}{\geq} \frac{H\Delta}{64e^6\sigma(1-q)}(1 - \pi_1(\phi \,|\, 0)) = 2\varepsilon\,(1 - \pi_1(\phi \,|\, 0))\,,\end{aligned}$$

where (i) holds by (E.118) and the last equality follows directly from the choice of Δ in (E.82).

Proof of inequality (E.118). Applying (E.4) in Lemma 57 to $\lambda^\star$ in (E.116) leads to the upper bound in (E.118):

$$\lambda^\star \leq \frac{H-1}{\sigma} \leq \frac{H-1}{\left(1 - \frac{3}{\beta}\right)\log\left(\frac{1}{\alpha+\Delta}\right)}, \tag{E.121}$$

where the last inequality holds by (E.73). As a result, we shall focus on showing the lower bounds in (E.118) in the remainder of the proof.

Recalling the definition of q in (E.61), we can reparameterize $1-q$ using two positive variables c_q and λ_q (whose choices will be made clearer soon) as follows:

$$1 - q = \alpha = c_q e^{-(H-1)/\lambda_q}. \tag{E.122}$$

Deriving the first derivative of the function of interest $f(\lambda)$ in (E.116) as follows:

$$\begin{aligned}\nabla_\lambda f(\lambda) &= \nabla_\lambda\left(-\lambda \log\left(P^\phi_{1,0,1-\phi} \cdot \exp\left(\frac{-\overline{V}}{\lambda}\right)\right) - \lambda\sigma\right) \\ &\overset{\text{(i)}}{=} \nabla_\lambda\left(-\lambda\log\left(qe^{-(H-1)/\lambda} + 1 - q\right) - \lambda\sigma\right) \\ &= -\sigma - \log\left(qe^{-(H-1)/\lambda} + 1 - q\right) - \frac{1}{\lambda}\cdot\frac{q(H-1)e^{-(H-1)/\lambda}}{qe^{-(H-1)/\lambda} + 1 - q},\end{aligned} \tag{E.123}$$

where (i) holds by the chosen transition kernels in (E.60) and the last line arises from basic calculus.

To continue, when $\lambda = \lambda_q$, the derivative of the function $f(\lambda)$ can be expressed as

$$\begin{aligned}
\nabla_\lambda f(\lambda)\,|_{\,\lambda=\lambda_q} &= -\sigma - \log\left((1-q)\frac{q}{c_q} + 1 - q\right) + \frac{(1-q)\frac{q}{c_q}\log\frac{1-q}{c_q}}{(1-q)\frac{q}{c_q}+1-q} \\
&= -\sigma - \log(1-q) - \log\left(1+\frac{q}{c_q}\right) + \frac{\frac{q}{c_q}\log\frac{1-q}{c_q}}{\frac{q}{c_q}+1} \\
&= -\sigma - \log(1-q)\left(1-\frac{q/c_q}{q/c_q+1}\right) - \log\left(1+\frac{q}{c_q}\right) - \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q} \\
&\overset{\text{(i)}}{=} -\sigma + \log\left(\frac{1}{\alpha+\Delta}\right)\left(1-\frac{q/c_q}{q/c_q+1}\right) - \log\left(1+\frac{q}{c_q}\right) - \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q} \qquad \text{(E.124)} \\
&\overset{\text{(ii)}}{\geq} \log\left(\frac{1}{\alpha+\Delta}\right)\left(\frac{2}{\beta}-\frac{q/c_q}{q/c_q+1}\right) - \log\left(1+\frac{q}{c_q}\right) - \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q} \\
&\overset{\text{(iii)}}{\geq} \frac{1}{\beta}\log\left(\frac{1}{\alpha+\Delta}\right) - \log(1+\frac{1}{\beta}) - 1 \\
&\geq \frac{1}{\beta}\log\left(\frac{1}{\alpha+\Delta}\right) - 2 = 0, \qquad \text{(E.125)}
\end{aligned}$$

where (i) holds by (E.122), (ii) follows from the bound of σ in (E.73), (iii) arises from letting $c_q = \beta \geq 4$ and noting the fact $1/2 \leq q < 1$ (see (E.64)), leading to

$$\frac{1}{2\beta} \leq \frac{q}{c_q} < \frac{1}{\beta}, \qquad \frac{q/c_q}{q/c_q+1} \leq \frac{1}{\beta}, \qquad \frac{\frac{q}{c_q}\log(c_q)}{1+q/c_q} < 1. \qquad \text{(E.126)}$$

Finally, the last line holds by $1/\beta \leq \frac{1}{4}$ and $\log\left(\frac{1}{\alpha+\Delta}\right) = 2\beta$ (see (E.63)).

To proceed, note that the function $f(\lambda)$ is concave with respect to λ. Therefore, observing $\nabla_\lambda f(\lambda)\,|_{\,\lambda=\lambda_q} \geq 0$ with $c_q = \beta$, we have $\lambda_q \leq \lambda^\star$, which implies (see (E.122))

$$1 - q = \alpha + \Delta = \beta e^{-(H-1)/\lambda_q} \leq \beta e^{-(H-1)/\lambda^\star}. \qquad \text{(E.127)}$$

The above assertion directly gives

$$\lambda^\star \geq \frac{H-1}{\log\left(\frac{\beta}{\alpha+\Delta}\right)}.$$

The proof is completed by noticing

$$\frac{H-1}{\log\left(\frac{\beta}{\alpha+\Delta}\right)} = \frac{H-1}{\log\left(\frac{1}{\alpha+\Delta}\right)+\log\beta} \overset{\text{(i)}}{\geq} \frac{H-1}{2\log\left(\frac{1}{\alpha+\Delta}\right)} \geq \frac{H}{16\sigma},$$

where (i) follows from (E.63), and the last inequality follows from (E.73) and the fact $\beta \in [4, \infty)$.

E.3 Analysis: discounted infinite-horizon RMDPs

E.3.1 Proof of Lemma 38

We shall first show that the operator $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ (cf. (7.30)) is a γ-contraction, which will in turn imply the existence of the unique fixed point of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$. Before starting, suppose that the entries of $Q_1, Q_2 \in \mathbb{R}^{SA}$ are all bounded in $\left[0, \frac{1}{1-\gamma}\right]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Denote that

$$\forall s \in \mathcal{S}: \quad V_1(s) \coloneqq \max_a Q_1(s,a), \quad V_2(s) \coloneqq \max_a Q_2(s,a). \tag{E.128}$$

Proof of γ-contraction. We first show that $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ is a γ-contraction. Towards this, instead of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$, we begin with a simpler operator $\widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$, defined as follows:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q)(s,a) = r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big), \tag{E.129}$$

which consequently leads to

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q)(s,a) = \max\left\{\widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q)(s,a), 0\right\}. \tag{E.130}$$

It follows straightforwardly that

$$\left\|\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_1) - \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_2)\right\|_{\infty} \leq \left\|\widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_2)\right\|_{\infty}, \tag{E.131}$$

and hence it suffices to establish the γ-contraction of $\widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$. With this in mind, we observe that

$$\begin{aligned}
\left\|\widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_1) - \widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(Q_2)\right\|_{\infty} &= \gamma \left\|\inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V_1 - \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V_2\right\|_{\infty} \overset{\text{(i)}}{\leq} \gamma \left\|V_1 - V_2\right\|_{\infty} \\
&\overset{\text{(ii)}}{=} \gamma \max_s \left|\max_a Q_1(s,a) - \max_a Q_2(s,a)\right| \\
&\leq \gamma \max_{(s,a)} \left|Q_1(s,a) - Q_2(s,a)\right| = \gamma \left\|Q_1 - Q_2\right\|_{\infty},
\end{aligned} \tag{E.132}$$

where the first equality holds by the definition of $\widetilde{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ (cf. (E.129)), (i) follows from that the infimum operator is a 1-contraction w.r.t. $\|\cdot\|_{\infty}$ and $\|\mathcal{P}V_1 - \mathcal{P}V_2\|_{\infty} \leq \|V_1 - V_2\|_{\infty}$ for all $\mathcal{P} \in \Delta(\mathcal{S})$, (ii) arises from the definitions in (E.128), and the last inequality is due to the maximum operator is also a 1-contraction w.r.t. $\|\cdot\|_{\infty}$. Combining the above two inequalities establish the desired statement.

Existence of the unique fixed point. To continue, we shall first claim that there exists at least one fixed point of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$. This is a standard argument, which we omit for brevity; interested readers

are encouraged to refer to, e.g. Li et al. (2022a), for details.

To prove the uniqueness of the fixed points of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$, suppose that there exist two fixed points Q' and Q'' obeying obeying $Q' = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q')$ and $Q'' = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q'')$. Moreover, the definition of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ directly implies $0 \leq Q', Q'' \leq \frac{1}{1-\gamma}$, since for any $0 \leq Q \leq \frac{1}{1-\gamma}$, it follows that $0 \leq \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q) \leq \frac{1}{1-\gamma}$. By the γ-contraction property, it follows that

$$\left\|Q' - Q''\right\|_{\infty} = \left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q') - \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q'')\right\|_{\infty} \leq \gamma \left\|Q' - Q''\right\|_{\infty}. \tag{E.133}$$

However, (E.133) can't happen given $\gamma \in \left[\frac{1}{2}, 1\right)$, indicating the uniqueness of the fixed points of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$.

E.3.2 Proof of Lemma 41

To begin with, considering any Q, Q' obeying $Q \leq Q'$, and $0 \leq Q, Q' \leq \frac{1}{1-\gamma}$. We observe that the operator $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$ (cf. (7.30)) has the monotone non-decreasing property, namely,

$$\begin{aligned}
\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q)(s,a) &= \max\left\{r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P}V - b\big(s,a\big),\, 0\right\} \\
&= \max\left\{r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P} \max_{a'} Q(\cdot, a') - b\big(s,a\big),\, 0\right\} \\
&\leq \max\left\{r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})} \mathcal{P} \max_{a'} Q'(\cdot, a') - b\big(s,a\big),\, 0\right\} = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(Q')(s,a),
\end{aligned} \tag{E.134}$$

where the last line uses $Q \leq Q'$. Recalling the fixed point $\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}$ of $\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\cdot)$, armed with (E.134) and the initialization $\widehat{Q}_0 = 0$, we arrive at

$$\widehat{Q}_1 = \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_0) \leq \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}) = \widehat{Q}_{\mathsf{pe}}^{\star,\sigma},$$

where the inequality follows from $\widehat{Q}_0 = 0 \leq \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}$. Implementing the above result recursively gives

$$\forall\, m \geq 0: \quad \widehat{Q}_m \leq \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}.$$

Applying the γ-contraction property in Lemma 38 thus yields that for any $m \geq 0$,

$$\begin{aligned}
\|\widehat{Q}_m - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_{\infty} = \left\|\widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{m-1}) - \widehat{\mathcal{T}}_{\mathsf{pe}}^{\sigma}(\widehat{Q}_{\mathsf{pe}}^{\star,\sigma})\right\|_{\infty} &\leq \gamma \|\widehat{Q}_{m-1} - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_{\infty} \\
&\leq \cdots \leq \gamma^m \|\widehat{Q}_0 - \widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_{\infty} = \gamma^m \|\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_{\infty} \leq \frac{\gamma^m}{1-\gamma},
\end{aligned}$$

where the last inequality holds by the fact $\|\widehat{Q}_{\mathsf{pe}}^{\star,\sigma}\|_{\infty} \leq \frac{1}{1-\gamma}$ (see Lemma 38).

E.3.3 Proof of Theorem 16

To begin, we introduce some additional notation that will be useful throughout the analysis. We denote the state-action space covered by the batch dataset $\mathcal{D}$ as

$$\mathcal{C}^{\mathsf{b}} = \left\{(s,a) : d^{\mathsf{b},P^0}(s,a) > 0\right\}. \tag{E.135}$$

In addition, recalling the definition in (7.31), we define a similar one based on the true nominal model P^0 as

$$P_{\mathsf{min}}(s,a) \coloneqq \min_{s'} \left\{P^0(s' \,|\, s,a) : \; P^0(s' \,|\, s,a) > 0\right\}, \tag{E.136}$$

which directly indicates that

$$P^{\star}_{\mathsf{min}} = \min_{s} \; P_{\mathsf{min}}(s, \pi^{\star}(s)), \qquad P^{\mathsf{b}}_{\mathsf{min}} = \min_{(s,a)\in\mathcal{C}^{\mathsf{b}}} \; P_{\mathsf{min}}(s,a). \tag{E.137}$$

Next, we denote the set of possible state occupancy distributions associated with the optimal policy $\pi^{\star}$ in a model within the uncertainty set $P \in \mathcal{U}^{\sigma}\left(P^0\right)$ as

$$\mathcal{D}^{\star} \coloneqq \left\{\left[d^{\star,P}(s)\right]_{s\in\mathcal{S}} : P \in \mathcal{U}^{\sigma}\left(P^0\right)\right\} = \left\{\left[d^{\star,P}\left(s,\pi^{\star}(s)\right)\right]_{s\in\mathcal{S}} : P \in \mathcal{U}^{\sigma}\left(P^0\right)\right\}, \tag{E.138}$$

where the second equality is due to the fact that $\pi^{\star}$ is chosen to be deterministic.

We are now ready to embark on the proof of Theorem 16. We first introduce a fact that is used throughout the proof; the proof is postponed to Appendix E.3.3.2:

$$\forall (s,a) \in \mathcal{C}^{\mathsf{b}} : \qquad N(s,a) \geq \frac{N d^{\mathsf{b},P^0}(s,a)}{12} \geq \frac{c_1 \log(NS/\delta)}{12 P_{\mathsf{min}}(s,a)} \geq -\frac{\log \frac{2NS}{\delta}}{\log(1 - P_{\mathsf{min}}(s,a))} \tag{E.139}$$

as long as (7.42) holds.

For notation simplicity, denote the output Q-function and value function from Algorithm 14 as $\widehat{Q} = \widehat{Q}_M$ and $\widehat{V} = \widehat{V}_M$. Invoking Lemma 41 with $M \geq \frac{\log \frac{\sigma N}{1-\gamma}}{\log \frac{1}{\gamma}}$ directly leads to

$$\left\|\widehat{Q} - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\right\|_{\infty} \leq \frac{1}{\sigma N} \tag{E.140}$$

and therefore

$$\left\|\widehat{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_{\infty} \leq \max_{s} \left|\max_{a} \widehat{Q}(s,a) - \max_{a} \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a)\right| \leq \left\|\widehat{Q} - \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\right\|_{\infty} \leq \frac{1}{\sigma N}. \tag{E.141}$$

The proof of Theorem 16 is separated into several key steps as follows.

Step 1: controlling the uncertainty via leave-one-out analysis. Given access to only a finite number of samples for estimating the nominal transition kernel P^0, we need to efficiently control

$$\left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V} \right|$$

across the robust value iterations, where $\widehat{V}$ is statistically dependent on $\widehat{P}^0_{s,a}$ (since $\widehat{P}^0_{s,a}$ will be reused in the update rule (cf. (7.35)) for all the iterations). A naive treatment via the standard covering arguments will unfortunately lead to rather loose bounds (Panaganti and Kalathil, 2022; Yang et al., 2022; Zhou et al., 2021). To overcome this challenge, we resort to the leave-one-out analysis—pioneered by Agarwal et al. (2020b); Li et al. (2022a, 2023c) in the context of model-based RL—to decouple the statistical dependency. The results are summarized in the following lemma, with the proof provided in Appendix E.3.3.1.

Lemma 64. *Instate the assumptions in Theorem 16. Then for all vector $\widetilde{V}$ obeying $\left\|\widetilde{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \le \frac{1}{\sigma N}$ and $\|\widetilde{V}\|_\infty \le \frac{1}{1-\gamma}$, with probability at least $1-\delta$, one has*

$$\left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V} \right| \le \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta})}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)},\ \frac{1}{1-\gamma} \right\} \tag{E.142}$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. In addition, for all $(s,a)\in\mathcal{C}^{\mathsf{b}}$, with probability at least $1-\delta$, one has

$$\frac{P_{\mathsf{min}}(s,a)}{8\log(NS/\delta)} \le \widehat{P}_{\mathsf{min}}(s,a) \le e^2 P_{\mathsf{min}}(s,a). \tag{E.143}$$

Step 2: establishing the pessimism property. Armed with Lemma 64, we aim to show the key property that

$$\forall (s,a)\in\mathcal{S}\times\mathcal{A}: \qquad \widehat{Q}(s,a) \le Q^{\widehat{\pi},\sigma}(s,a), \qquad \widehat{V}(s) \le V^{\widehat{\pi},\sigma}(s). \tag{E.144}$$

Similar to the finite-horizon setting, it suffices to focus on verifying the former assertion in (E.144). Towards this, we first recall that the fixed point $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ of the pessimistic robust Bellman operator $\widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\cdot)$ (cf. (7.30)) obeys

$$\widehat{Q}^{\star,\sigma}_{\mathsf{pe}} = \widehat{\mathcal{T}}^\sigma_{\mathsf{pe}}(\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}) = \max\left\{ r(s,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b\big(s,a\big),\, 0 \right\}. \tag{E.145}$$

If $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a)=0$. Given the initialization $\widehat{Q}_0=0$, invoking Lemma 41 gives

$$\widehat{Q}(s,a)=\widehat{Q}_M(s,a)\leq\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a)=0.$$

As a result, $Q^{\widehat{\pi},\sigma}(s,a)\geq 0=\widehat{Q}(s,a)$ as desired. Therefore, it boils down to examine the case when $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a)>0$. One has

$$\begin{aligned}
\widehat{Q}(s,a)&\overset{\text{(i)}}{\leq}\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s,a)+\frac{1}{\sigma N}=r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}-b\big(s,a\big)+\frac{1}{\sigma N}\\
&\leq r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}-b(s,a)+\frac{1}{\sigma N}+\gamma\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}}-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}\right|\\
&\overset{\text{(ii)}}{\leq} r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}-b(s,a)+\frac{2}{\sigma N}\\
&\leq r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}\widehat{V}-b(s,a)+\frac{2}{\sigma N}+\gamma\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}\widehat{V}-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}\widehat{V}\right|\\
&\leq r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}\widehat{V},
\end{aligned}\tag{E.146}$$

where (i) follows from (E.140), (ii) arises from (E.141) and the basic fact that infimum operator is 1-contraction w.r.t $\|\cdot\|_\infty$, and the last inequality holds by the definition of $b(s,a)$ (cf. (7.32)) and Lemma 64. Putting the above inequality together with the robust Bellman equation (cf. (2.27a)) pertaining to $Q^{\widehat{\pi},\sigma}(s,a)$, we arrive at

$$\begin{aligned}
Q^{\widehat{\pi},\sigma}(s,a)-\widehat{Q}(s,a)&\geq r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V^{\widehat{\pi},\sigma}-\left(r(s,a)+\gamma\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}\widehat{V}\right)\\
&=\gamma\left(\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V^{\widehat{\pi},\sigma}-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}\widehat{V}\right)\\
&\overset{\text{(i)}}{=}\gamma\left(\widetilde{P}_{s,a}V^{\widehat{\pi},\sigma}-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}\widehat{V}\right)\geq\gamma\widetilde{P}_{s,a}\left(V^{\widehat{\pi},\sigma}-\widehat{V}\right),
\end{aligned}$$

where (i) holds by setting $\widetilde{P}_{s,a}=\operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V^{\widehat{\pi},\sigma}$. Consequently, one has

$$\begin{aligned}
\min_{s,a}\left[Q^{\widehat{\pi},\sigma}(s,a)-\widehat{Q}(s,a)\right]&\geq\min_{s,a}\left[\gamma\widetilde{P}_{s,a}\left(V^{\widehat{\pi},\sigma}-\widehat{V}\right)\right]\overset{\text{(i)}}{\geq}\gamma\min_{s}\left[V^{\widehat{\pi},\sigma}(s)-\widehat{V}(s)\right]\\
&=\gamma\min_{s}\left[Q^{\widehat{\pi},\sigma}\big(s,\widehat{\pi}(s)\big)-\widehat{Q}\big(s,\widehat{\pi}(s)\big)\right]\\
&\geq\gamma\min_{s,a}\left[Q^{\widehat{\pi},\sigma}\big(s,a\big)-\widehat{Q}\big(s,a\big)\right],
\end{aligned}\tag{E.147}$$

where (i) follows from $\widetilde{P}_{s,a} \in \Delta(\mathcal{S})$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Noting that $0 \leq \gamma < 1$, we conclude $Q^{\widehat{\pi},\sigma}(s,a) - \widehat{Q}(s,a) \geq 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. This establishes the claim (E.144).

Step 3: bounding $V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho)$. In view of the pessimistic property (cf. (E.144)), it follows that

$$V^{\star,\sigma}(s) - V^{\widehat{\pi},\sigma}(s) \leq V^{\star,\sigma}(s) - \widehat{V}(s). \tag{E.148}$$

Towards this, note that

$$\begin{aligned}
\widehat{V}(s) = \max_a \widehat{Q}(s,a) \geq \widehat{Q}\big(s,\pi^\star(s)\big) &\overset{\text{(i)}}{\geq} \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N} \\
&\overset{\text{(ii)}}{\geq} r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N} \\
&\geq r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - b\big(s,\pi^\star(s)\big) - \frac{1}{\sigma N} - \gamma \left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} \right| \\
&\overset{\text{(iii)}}{\geq} r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - b\big(s,\pi^\star(s)\big) - \frac{2}{\sigma N} \\
&\geq r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - b\big(s,\pi^\star(s)\big) - \frac{2}{\sigma N} - \gamma \left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(\widehat{P}^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} \right| \\
&\geq r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} - 2b\big(s,\pi^\star(s)\big),
\end{aligned} \tag{E.149}$$

where (i) follows from (E.140), (ii) holds by applying (E.145), (iii) arises from (E.141), and the basic fact that the infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$, and the final inequality holds by the definition of $b(s,a)$ (see (7.32)) and Lemma 64.

To continue, invoking the robust Bellman optimality equation in (2.27b) gives

$$V^{\star,\sigma}(s) = Q^{\star,\sigma}\big(s,\pi^\star(s)\big) = r\big(s,\pi^\star(s)\big) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}V^{\star,\sigma}.$$

Combining the above relation with (E.149), we arrive at

$$\begin{aligned}
V^{\star,\sigma}(s) - \widehat{V}(s) &\leq \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}V^{\star,\sigma} - \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V} + 2b\big(s,\pi^\star(s)\big) \\
&\leq \gamma \widehat{P}^{\inf}_{s,\pi^\star(s)}\left(V^{\star,\sigma} - \widehat{V}\right) + 2b\big(s,\pi^\star(s)\big),
\end{aligned} \tag{E.150}$$

where the final inequality holds evidently, by introducing

$$\widehat{P}^{\inf}_{s,\pi^\star(s)} \coloneqq \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^\sigma\left(P^0_{s,\pi^\star(s)}\right)} \mathcal{P}\widehat{V}. \tag{E.151}$$

Before continuing, for convenience, let us introduce a matrix $\widehat{P}^{\inf} \in \mathbb{R}^{S\times\mathcal{S}}$ and a vector $b^\star \in \mathbb{R}^{\mathcal{S}}$, where their s-th rows (resp. entries) are defined as

$$\left[\widehat{P}^{\inf}\right]_{s,\cdot} = \widehat{P}^{\inf}_{s,\pi^\star(s)}, \qquad \text{and} \qquad b^\star(s) = b\big(s,\pi^\star(s)\big). \tag{E.152}$$

With these notation in hand, averaging (E.150) over the initial state distribution ρ leads to

$$\begin{aligned} V^{\star,\sigma}(\rho) - \widehat{V}(\rho) &= \sum_{s\in\mathcal{S}} \rho(s)\left(V^{\star,\sigma}(s) - \widehat{V}(s)\right) \\ &\le \gamma \sum_{s\in\mathcal{S}} \rho(s)\widehat{P}^{\inf}_{s,\pi^\star(s)}\left(V^{\star,\sigma} - \widehat{V}\right) + 2\sum_{s\in\mathcal{S}}\rho(s)b\big(s,\pi^\star(s)\big) \\ &= \gamma\rho^\top \widehat{P}^{\inf}\left(V^{\star,\sigma} - \widehat{V}\right) + 2\rho^\top b^\star. \end{aligned} \tag{E.153}$$

Applying the above result recursively gives

$$\begin{aligned} V^{\star,\sigma}(\rho) - \widehat{V}(\rho) &\le \gamma\rho^\top \widehat{P}^{\inf}\left(V^{\star,\sigma} - \widehat{V}\right) + 2\rho^\top b^\star \\ &\le \gamma\left(\gamma\rho^\top\widehat{P}^{\inf}\right)\widehat{P}^{\inf}\left(V^{\star,\sigma} - \widehat{V}\right) + 2\left(\gamma\rho^\top\widehat{P}^{\inf}\right)b^\star + 2\rho^\top b^\star \\ &\le \cdots \le \left\{\lim_{i\to\infty}\gamma^i\rho^\top\left(\widehat{P}^{\inf}\right)^i\left(V^{\star,\sigma} - \widehat{V}\right)\right\} + 2\rho^\top\sum_{i=0}^{\infty}\gamma^i\left(\widehat{P}^{\inf}\right)^i b^\star \\ &\overset{\text{(i)}}{\le} 2\rho^\top\sum_{i=0}^{\infty}\gamma^i\left(\widehat{P}^{\inf}\right)^i b^\star = 2\rho^\top\left(I - \gamma\widehat{P}^{\inf}\right)^{-1} b^\star, \end{aligned} \tag{E.154}$$

where (i) holds by $\left|\rho^\top\left(\widehat{P}^{\inf}\right)^i\left(V^{\star,\sigma} - \widehat{V}\right)\right| \le \frac{1}{1-\gamma}$ for all $i \ge 0$, and that $\lim_{i\to\infty}\gamma^i\rho^\top\left(\widehat{P}^{\inf}\right)^i\left(V^{\star,\sigma} - \widehat{V}\right) = 0$ since $\lim_{i\to\infty}\gamma^i = 0$ for all $0 \le \gamma < 1$.

To further characterize the above performance gap, invoking the definition of $d^{\star,P}$ (cf. (2.22) and (2.28a)), we arrive at

$$\left(d^{\star,\widehat{P}^{\inf}}\right)^\top = (1-\gamma)\rho^\top\sum_{t=0}^{\infty}\gamma^t\left(\widehat{P}^{\inf}\right)^t = (1-\gamma)\rho^\top\left(I - \gamma\widehat{P}^{\inf}\right)^{-1}. \tag{E.155}$$

Plugging the above expression back into (E.154), and combining with(E.148), yields

$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \le V^{\star,\sigma}(\rho) - \widehat{V}(\rho) \le \frac{2}{1-\gamma}\left\langle d^{\star,\widehat{P}^{\inf}}, b^\star\right\rangle. \tag{E.156}$$

Step 4: controlling $\left\langle d^{\star,\widehat{P}^{\inf}}, b^\star \right\rangle$ using concentrability. Note that $\widehat{P}^{\inf} \in \mathcal{U}^\sigma(P^0)$ (see (E.151) and (E.152)), which in words means $\widehat{P}^{\inf}$ is some transition kernel inside $\mathcal{U}^\sigma(P^0)$ — the uncertainty set around the nominal kernel P^0. Similar to the finite-horizon case, observing that we can express $\left\langle d^{\star,\widehat{P}^{\inf}}, b^\star \right\rangle = \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\inf}}(s) b^\star(s)$, we divide the states into two cases and control them separately.

- **Case 1: $s \in \mathcal{S}$ where $\max_{P\in\mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\big(s,\pi^\star(s)\big) = 0$.** Since $\widehat{P}^{\inf} \in \mathcal{U}^\sigma(P^0)$, one has

$$0 \le d^{\star,\widehat{P}^{\inf}}(s) = d^{\star,\widehat{P}^{\inf}}\big(s,\pi^\star(s)\big) \le \max_{P\in\mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\big(s,\pi^\star(s)\big) = 0,$$

  which consequently indicates

$$d^{\star,\widehat{P}^{\inf}}(s) = 0. \tag{E.157}$$

- **Case 2: $s \in \mathcal{S}$ where $\max_{P\in\mathcal{U}^\sigma\left(P^0\right)} d^{\star,P}\big(s,\pi^\star(s)\big) > 0$.** For any such state s, we claim that

$$d^{\mathsf{b},P^0}\big(s,\pi^\star(s)\big) > 0 \quad \text{and} \quad \big(s,\pi^\star(s)\big) \in \mathcal{C}^{\mathsf{b}}. \tag{E.158}$$

  This is due to Assumption 6, which requires $C^\star_{\mathsf{rob}}$ to be finite given the numerator is positive:

$$\max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d^{\star,P}\big(s,\pi^\star(s)\big), \frac{1}{S}\right\}}{d^{\mathsf{b},P^0}\big(s,\pi^\star(s)\big)} = \max_{P\in\mathcal{U}^\sigma(P^0)} \frac{\min\left\{d^{\star,P}(s), \frac{1}{S}\right\}}{d^{\mathsf{b},P^0}(s,a)} \le C^\star_{\mathsf{rob}} < \infty. \tag{E.159}$$

  To continue, invoking the fact in (E.139) with $\big(s,\pi^\star(s)\big) \in \mathcal{C}^{\mathsf{b}}$ gives

$$\begin{aligned} N\big(s,\pi^\star(s)\big) &\ge \frac{N d^{\mathsf{b},P^0}\big(s,\pi^\star(s)\big)}{12} \\ &\overset{\text{(i)}}{\ge} \frac{N \max_{P\in\mathcal{U}^\sigma(P^0)} \min\left\{d^{\star,P}\big(s,\pi^\star(s)\big), \frac{1}{S}\right\}}{12 C^\star_{\mathsf{rob}}} \ge \frac{N \min\left\{d^{\star,\widehat{P}^{\inf}}(s), \frac{1}{S}\right\}}{12 C^\star_{\mathsf{rob}}}, \end{aligned} \tag{E.160}$$

  where (i) holds by Assumption 6, and the last inequality holds by $\widehat{P}^{\inf} \in \mathcal{U}^\sigma(P^0)$. With this in mind, we can control the pessimistic penalty $b^\star(s)$ (cf. (7.32)) by

$$\begin{aligned} b^\star(s) &\le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}\big(s,\pi^\star(s)\big) N\big(s,\pi^\star(s)\big)}} + \frac{4}{\sigma N(1-\gamma)} + \frac{2}{\sigma N} \\ &\overset{\text{(i)}}{\le} \frac{4c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^\star(s)\big) N\big(s,\pi^\star(s)\big)}} + \frac{4}{\sigma N(1-\gamma)} + \frac{2}{\sigma N} \end{aligned}$$

$$\leq \frac{16c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}}+\frac{6}{\sigma N(1-\gamma)}$$
$$\leq \frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}},$$

where (i) arises from (E.143), the penultimate inequality follows from (E.160), and the last inequality holds as long as c_{b} is large enough.

Summing up the above two cases, we arrive at

$$\begin{aligned}\left\langle d^{\star,\widehat{P}^{\mathrm{inf}}},b^{\star}\right\rangle &= \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)b^{\star}(s)\\ &\leq \sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)\frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}}\\ &\overset{\text{(i)}}{\leq} \frac{20c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)\frac{C^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)N\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}}\sqrt{\sum_{s\in\mathcal{S}} d^{\star,\widehat{P}^{\mathrm{inf}}}(s)}\\ &\leq \frac{40c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{SC^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}}N}},\end{aligned}\tag{E.161}$$

where (i) arises from Cauchy-Schwarz inequality, and the last inequality holds since $P_{\mathsf{min}}\big(s,\pi^{\star}(s)\big)\geq P^{\star}_{\mathsf{min}}$ for all $s\in\mathcal{S}$ (see (E.137)) and the following fact (which has been established in (E.36)):

$$\sum_{s\in\mathcal{S}}\frac{d^{\star,\widehat{P}^{\mathrm{inf}}}(s)}{\min\left\{d^{\star,\widehat{P}^{\mathrm{inf}}}(s),\frac{1}{S}\right\}}\leq 2S.$$

Finally, inserting (E.161) back into (E.156), with probability at least $1-2\delta$, one has

$$V^{\star,\sigma}(\rho)-V^{\widehat{\pi},\sigma}(\rho)\leq\frac{2}{1-\gamma}\left\langle d^{\star,\widehat{P}^{\mathrm{inf}}},b^{\star}\right\rangle\leq\frac{80c_{\mathrm{b}}}{\sigma(1-\gamma)^2}\sqrt{\frac{SC^{\star}_{\mathsf{rob}}\log^2\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{\star}_{\mathsf{min}}N}},$$

which concludes the proof.

E.3.3.1 Proof of Lemma 64

We first note that the second assertion in (E.143) is the counterpart of (E.18), which can be verified following the same argument in Appendix E.2.2.1. For brevity, we omit its proof, and shall focus on

verifying (E.142).

To begin with, we consider the situation when $N(s,a)=0$. In this case, (E.142) can be easily verified since

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V\right|\overset{\text{(i)}}{=}\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V\le\|V\|_\infty\overset{\text{(ii)}}{\le}\frac{1}{1-\gamma}, \tag{E.162}$$

where (i) follows from the fact $\widehat{P}^0_{s,a}=0$ when $N(s,a)=0$ (see (7.28)), and (ii) arises from the assumption $\|V\|_\infty\le\frac{1}{1-\gamma}$. Consequently, in the remainder of the proof, we focus on verifying (E.142) when $N(s,a)>0$. Let us first introduce the counterpart of the claim (E.17) in Lemma 61 as follows.

Lemma 65. *For all $(s,a)\in\mathcal{S}\times\mathcal{A}$ with $N(s,a)>0$, consider any vector $V\in\mathbb{R}^S$ independent of $\widehat{P}^0_{s,a}$ obeying $\|V\|_\infty\le\frac{1}{1-\gamma}$. With probability at least $1-\delta$, one has*

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^0_{s,a})}\mathcal{P}V-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^0_{s,a})}\mathcal{P}V\right|\le\frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log(\frac{NS}{\delta})}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}}. \tag{E.163}$$

Proof. The proof follows from the same arguments in Appendix E.2.2.2, with small modifications to adapt to the infinite-horizon setting; we omit the details for conciseness. □

Armed with the above point-wise concentration bound, we are now ready to derive the uniform concentration bound desired as in Lemma 64, counting on a leave-one-out argument divided into the following steps. The crux of the analysis is to construct a set of auxiliary RMDPs, each different from the empirical RMDP only at a single state but possessing crucial statistical independence that facilitates the concentration arguments, which can then be transferred back to the empirical RMDP via a simple triangle inequality.

Step 1: construction of auxiliary RMDPs with state-absorbing empirical nominal transitions. Denote the empirical infinite-horizon robust MDP with the nominal transition kernel $\widehat{P}^0$ as $\widehat{\mathcal{M}}_{\mathsf{rob}}$. Then, for each state s and each scalar $u\ge0$, we can construct an auxiliary robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ so that it is the same as $\widehat{\mathcal{M}}_{\mathsf{rob}}$ except the properties in state s. To be precise, let the nominal transition kernel and reward function of $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ be $P^{s,u}$ and $r^{s,u}$, which are given respectively as

$$\begin{cases}P^{s,u}(s'\,|\,s,a)=\mathbb{1}(s'=s) & \text{for all } (s',a)\in\mathcal{S}\times\mathcal{A},\\ P^{s,u}(\cdot\,|\,\widetilde{s},a)=\widehat{P}^0(\cdot\,|\,\widetilde{s},a) & \text{for all } (\widetilde{s},a)\in\mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s}\neq s,\end{cases} \tag{E.164}$$

and

$$\begin{cases}r^{s,u}(s,a)=u & \text{for all } a\in\mathcal{A},\\ r^{s,u}(\widetilde{s},a)=r(\widetilde{s},a) & \text{for all } (\widetilde{s},a)\in\mathcal{S}\times\mathcal{A} \text{ and } \widetilde{s}\neq s.\end{cases} \tag{E.165}$$

Clearly, state s of the auxiliary $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is absorbing, meaning that the state stays at s once entering it. This removes the randomness of $\widehat{P}^0_{s,a}$ for all $a \in \mathcal{A}$ in state s, a key property we will leverage later.

With the robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ in hand, we still need to complete the design by defining the corresponding penalty term for all $(\widetilde{s}, a) \in \mathcal{S} \times \mathcal{A}$, which is given as follows

$$b^{s,u}(\widetilde{s}, a) \coloneqq \begin{cases} \min\left\{ \frac{c_{\mathsf{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3 S}{(1-\gamma)\delta}\right)}{P^{s,u}_{\mathsf{min}}(s,a) N(\widetilde{s},a)}} + \frac{4}{N\sigma(1-\gamma)} \ , \ \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N} & \text{if } N(\widetilde{s}, a) > 0, \\ \frac{1}{1-\gamma} + \frac{2}{\sigma N} & \text{otherwise,} \end{cases} \tag{E.166}$$

where $P^{s,u}_{\mathsf{min}}(\widetilde{s}, a)$ is defined as the smallest positive state transition probability over the nominal kernel $P^{s,u}(\cdot \,|\, \widetilde{s}, a)$:

$$\forall (\widetilde{s}, a) \in \mathcal{S} \times \mathcal{A}: \quad P^{s,u}_{\mathsf{min}}(\widetilde{s}, a) \coloneqq \min_{s'} \Big\{ P^{s,u}(s' \,|\, \widetilde{s}, a) : \ P^{s,u}(s' \,|\, \widetilde{s}, a) > 0 \Big\}. \tag{E.167}$$

In view of (E.164) and (7.31), it holds that $P^{s,u}_{\mathsf{min}}(\widetilde{s}, a) = \widehat{P}_{\mathsf{min}}(\widetilde{s}, a)$, and therefore $b^{s,u}(\widetilde{s}, a) = b(\widetilde{s}, a)$, when $\widetilde{s} \neq s$ for any $u \geq 0$. Armed with the above definitions, the pessimistic robust Bellman operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(\cdot)$ of the RMDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is defined as

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}: \quad \widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(s, a) = \max\left\{ r(s, a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P^{s,u}_{s,a})} \mathcal{P}V - b^{s,u}(s, a),\, 0 \right\}. \tag{E.168}$$

Step 2: fixed-point equivalence between $\widehat{\mathcal{M}}_{\mathsf{rob}}$ and the auxiliary RMDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$. Recall that $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ is the unique fixed point of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$ with the corresponding value $\widehat{V}^{\star,\sigma}_{\mathsf{pe}}$. We claim that there exists some choice of u such that the fixed point of $\widehat{\mathcal{T}}^{\sigma}_{s,u}(Q)(\cdot)$ coincides with that of $\widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\cdot)$. In particular, given a state s, we show the following choice of u suffices:

$$u^\star \coloneqq (1-\gamma)\widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) + \min\left\{ \frac{c_{\mathsf{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3 S}{(1-\gamma)\delta}\right)}{P^{s,u}_{\mathsf{min}}(s,a) N(s,a)}} + \frac{4}{N\sigma(1-\gamma)} \ , \ \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N}. \tag{E.169}$$

Towards this, we shall break our arguments in two different cases.

- **For state $s' \neq s$.** In this case, for any $a \in \mathcal{A}$, it can be verified that

$$\begin{aligned} &\max\left\{ r^{s,u^\star}(s', a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P^{s,u^\star}_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b^{s,u^\star}(s', a),\, 0 \right\} \\ &= \max\left\{ r(s', a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^0_{s',a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b(s', a),\, 0 \right\} \end{aligned}$$

$$= \widehat{\mathcal{T}}^{\sigma}_{\mathsf{pe}}(\widehat{Q}^{\star,\sigma}_{\mathsf{pe}})(s',a) = \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s',a), \tag{E.170}$$

where the second line follows from the definitions in (E.165) and (E.164) as well as $b^{s,u^\star}(s',a) = b(s',a)$ when $s' \neq s$, the last line arises from the definition of the pessimistic Bellman operator (7.30), and that $\widehat{Q}^{\star,\sigma}_{\mathsf{pe}}$ is the fixed point.

- **For state s.** In this case, for any u and $a \in \mathcal{A}$, observing that $P^{s,u}(s' \,|\, s,a)$ has only one positive entry equal to 1 (cf. (E.164)), applying (E.167) yields

$$P^{s,u}_{\mathsf{min}}(s,a) = 1. \tag{E.171}$$

Plugging the above fact into (E.166) leads to

$$b^{s,u}(s,a) = \begin{cases} \min\left\{ \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{N(s,a)}} + \frac{4}{N\sigma(1-\gamma)} \,,\, \frac{1}{1-\gamma} \right\} + \frac{2}{\sigma N} & \text{if } N(s,a) > 0, \\ \frac{1}{1-\gamma} & \text{otherwise} \end{cases} \tag{E.172}$$

for all $a \in \mathcal{A}$. As a result, we have for any $a \in \mathcal{A}$:

$$\begin{aligned} &\max\left\{ r^{s,u^\star}(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^\sigma(P^{s,u^\star}_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - b^{s,u^\star}(s,a),\, 0 \right\} \\ &= \max\left\{ u^\star + \gamma \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) - b^{s,u^\star}(s,a),\, 0 \right\} \\ &= \max\left\{ (1-\gamma)\widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) + \gamma \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s),\, 0 \right\} = \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s), \end{aligned} \tag{E.173}$$

where the second line follows from the fact that $P^{s,u^\star}_{s,a}$ is a singleton distribution at state s, and hence $\mathcal{U}^\sigma(P^{s,u^\star}_{s,a}) = P^{s,u^\star}_{s,a}$ by the definition of the KL uncertainty set, and the second line follows from plugging in the definition of $u^\star$ in (E.169) and $b^{s,u^\star}(s,a)$ in (E.172).

Summing up the above two cases, we establish that there exists a fixed point $\widehat{Q}^{\star,\sigma}_{s,u^\star}$ of the operator $\widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\cdot)$ if we let

$$\begin{cases} \widehat{Q}^{\star,\sigma}_{s,u^\star}(s,a) = \widehat{V}^{\star,\sigma}_{\mathsf{pe}}(s) & \text{for all } a \in \mathcal{A}, \\ \widehat{Q}^{\star,\sigma}_{s,u^\star}(s',a) = \widehat{Q}^{\star,\sigma}_{\mathsf{pe}}(s',a) & \text{for all } s' \neq s \text{ and } a \in \mathcal{A}. \end{cases} \tag{E.174}$$

Consequently, we confirm the existence of a fixed point of the operator $\widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\cdot)$. In addition, its corresponding value function $\widehat{V}^{\star,\sigma}_{s,u^\star}$ also coincides with $\widehat{V}^{\star,\sigma}_{\mathsf{pe}}$.

Step 3: building an ε-net for all reward values u. It is easily verified that the reward $u^\star$ obeys

$$u^\star \le 1 + \min\left\{\frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{P^{s,u}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{\sigma N(1-\gamma)}\,,\ \frac{1}{1-\gamma}\right\} + \frac{2}{\sigma N} \le \frac{2}{\sigma} + \frac{2}{1-\gamma}. \tag{E.175}$$

As a result, we construct an ε-net (Vershynin, 2018) of the line segment within the range $\left[0, \frac{2}{\sigma} + \frac{2}{1-\gamma}\right]$ with $\varepsilon = \frac{1}{\sigma N}$ as follows:

$$\mathcal{U}_\varepsilon := \left\{\frac{i}{\sigma N} \,|\, 1 \le i \le \left\lfloor \sigma N\left(\frac{2}{\sigma} + \frac{2}{1-\gamma}\right)\right\rfloor\right\}. \tag{E.176}$$

Armed with this covering net $\mathcal{U}_\varepsilon$, we can construct an auxiliary robust MDP $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ and its corresponding pessimistic robust Bellman operator for each $u \in \mathcal{U}_\varepsilon$ (see Step 1). Following the same arguments in the proof of Lemma 38 (cf. Appendix E.3.1), for each $u \in \mathcal{U}_\varepsilon$, it can be verified that there exists a unique fixed point $\widehat{Q}^{\star,\sigma}_{s,u}$ of the operator $\widehat{\mathcal{T}}^{\sigma}_{s,u}(\cdot)$, which satisfies $0 \le \widehat{Q}^{\star,\sigma}_{s,u} \le \frac{1}{1-\gamma}\cdot 1$. In turn, the corresponding value function also satisfies $\|\widehat{V}^{\star,\sigma}_{s,u}\|_\infty \le \frac{1}{1-\gamma}$.

In view of the definitions in (E.164) and (E.165), for all $u \in \mathcal{U}_\varepsilon$, $\widehat{\mathcal{M}}^{s,u}_{\mathsf{rob}}$ is statistically independent from $\widehat{P}^0_{s,a}$, which indicates the independence between $\widehat{V}^{\star,\sigma}_{s,u}$ and $\widehat{P}^0_{s,a}$. This makes it possible to invoke Lemma 65, and taking the union bound over all samples N and $u \in \mathcal{U}_\varepsilon$ give that, with probability at least $1-\delta$,

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u}\right| \le \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)}\sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} \tag{E.177}$$

hold simultaneously for all $(s,a,u) \in \mathcal{S}\times\mathcal{A}\times\mathcal{U}_\varepsilon$ with $N(s,a) > 0$.

Step 4: a covering argument. Recalling that $u^\star \in \left[0, \frac{2}{\sigma} + \frac{2}{1-\gamma}\right]$ (see (E.175)), we can always find some $\widetilde{u} \in \mathcal{U}_\varepsilon$ such that $|\widetilde{u} - u^\star| \le \frac{1}{\sigma N}$. Consequently, plugging in the operator in (E.168) yields

$$\forall Q \in \mathbb{R}^{SA}: \quad \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(Q) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(Q)\right\|_\infty \overset{\text{(i)}}{\le} |\widetilde{u} - u^\star| \le \frac{1}{\sigma N}, \tag{E.178}$$

where (i) holds by $b^{s,\widetilde{u}}(s,a) = b^{s,u^\star}(s,a)$ for s (see (E.172)) and $b^{s,\widetilde{u}}(s',a) = b^{s,u^\star}(s',a) = b(s',a)$ for all $s' \ne s$.

With this in mind, we observe that the fixed points of $\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\cdot)$ and $\widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\cdot)$ obey

$$\left\|\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star}\right\|_\infty = \left\|\widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star})\right\|_\infty$$

$$\leq \left\| \widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,\widetilde{u}}) - \widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star}) \right\|_\infty + \left\| \widehat{\mathcal{T}}^{\sigma}_{s,\widetilde{u}}(\widehat{Q}^{\star,\sigma}_{s,u^\star}) - \widehat{\mathcal{T}}^{\sigma}_{s,u^\star}(\widehat{Q}^{\star,\sigma}_{s,u^\star}) \right\|_\infty$$
$$\leq \gamma \left\| \widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star} \right\|_\infty + \frac{1}{\sigma N}, \tag{E.179}$$

which directly indicates that

$$\left\| \widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star} \right\|_\infty \leq \frac{1}{(1-\gamma)\sigma N} \tag{E.180}$$

and

$$\left\| \widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{V}^{\star,\sigma}_{s,u^\star} \right\|_\infty \leq \left\| \widehat{Q}^{\star,\sigma}_{s,\widetilde{u}} - \widehat{Q}^{\star,\sigma}_{s,u^\star} \right\|_\infty \leq \frac{1}{(1-\gamma)\sigma N}. \tag{E.181}$$

Armed with the above facts, invoking the identity $\widehat{V}^{\star,\sigma}_{\mathsf{pe}} = \widehat{V}^{\star,\sigma}_{s,u^\star}$ established in Step 2 gives

$$\begin{aligned}
&\left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} \right| = \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star} \right| \\
&\overset{\text{(i)}}{\leq} \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} \right| \\
&\qquad + \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star} \right| + \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,u^\star} \right| \\
&\overset{\text{(ii)}}{\leq} \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{s,\widetilde{u}} \right| + \frac{2}{N\sigma(1-\gamma)} \\
&\leq \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{2}{N\sigma(1-\gamma)},
\end{aligned} \tag{E.182}$$

where (i) holds by applying the triangle inequality, (ii) arises from (E.181) and the basic fact that infimum operator is a 1-contraction w.r.t. $\|\cdot\|_\infty$, and the final inequality follows from (E.177).

Step 5: finishing up. Now we are positioned to finish up the proof. For all vector $\widetilde{V}$ obeying $\left\|\widetilde{V} - \widehat{V}^{\star,\sigma}_{\mathsf{pe}}\right\|_\infty \leq \frac{1}{\sigma N}$ and $\|\widetilde{V}\|_\infty \leq \frac{1}{1-\gamma}$, we apply the triangle inequality and invoke (E.182) to reach

$$\begin{aligned}
&\left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V} \right| \leq \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} \right| \\
&\qquad + \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(\widehat{P}^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} \right| + \left| \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widetilde{V} - \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^0_{s,a})} \mathcal{P}\widehat{V}^{\star,\sigma}_{\mathsf{pe}} \right| \\
&\leq \frac{c_{\mathrm{b}}}{\sigma(1-\gamma)} \sqrt{\frac{\log\left(\frac{2(1+\sigma)N^3S}{(1-\gamma)\delta}\right)}{\widehat{P}_{\mathsf{min}}(s,a)N(s,a)}} + \frac{4}{N\sigma(1-\gamma)}.
\end{aligned} \tag{E.183}$$

Finally, we complete the proof by verifying that

$$\left|\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(\widehat{P}^{0}_{s,a})}\mathcal{P}\widetilde{V}-\inf_{\mathcal{P}\in\mathcal{U}^{\sigma}(P^{0}_{s,a})}\mathcal{P}\widetilde{V}\right|\leq\left\|\widetilde{V}\right\|_{\infty}\leq\frac{1}{1-\gamma}. \tag{E.184}$$

E.3.3.2 Proof of (E.139)

For all $(s,a)\in\mathcal{C}^{\mathsf{b}}$, one has

$$Nd^{\mathsf{b},P^0}(s,a)\overset{\text{(i)}}{\geq}\frac{c_1 d^{\mathsf{b},P^0}(s,a)\log(NS/\delta)}{d^{\mathsf{b}}_{\mathsf{min}}P^{\mathsf{b}}_{\mathsf{min}}}\overset{\text{(ii)}}{\geq}\frac{c_1\log(NS/\delta)}{P^{\mathsf{b}}_{\mathsf{min}}}\overset{\text{(iii)}}{\geq}\frac{c_1\log(NS/\delta)}{P_{\mathsf{min}}(s,a)}, \tag{E.185}$$

where (i) follows from the condition (7.42), (ii) arises from the definition that $d^{\mathsf{b}}_{\mathsf{min}}\leq d^{\mathsf{b},P^0}(s,a)$ for all $(s,a)\in\mathcal{C}^{\mathsf{b}}$, and (iii) follows from the definition in (E.137). In particular, when c_1 is large enough, one has $\frac{2}{3}\log\frac{NS}{\delta}<\frac{Nd^{\mathsf{b},P^0}(s,a)}{12}$. To continue, we recall a key property of $N(s,a)$ (cf. (7.27)) in the following lemma.

Lemma 66 ((Li et al., 2022a, Lemma 7)). *Fix $\delta\in(0,1)$. With probability at least $1-\delta$, the quantities $\{N(s,a)\}$ in (7.27) obey*

$$\max\left\{N(s,a),\frac{2}{3}\log\frac{NS}{\delta}\right\}\geq\frac{Nd^{\mathsf{b},P^0}(s,a)}{12} \tag{E.186}$$

simultaneously for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

Consequently, Lemma 66 tells us that with probability at least $1-\delta$,

$$N(s,a)\geq\frac{Nd^{\mathsf{b},P^0}(s,a)}{12}\geq\frac{c_1\log(NS/\delta)}{12P_{\mathsf{min}}(s,a)} \tag{E.187}$$

as long as c_1 is large enough. Last but not least, taking the basic fact $x\leq-\log(1-x)$ for all $x\in[0,1]$, the last inequality of (E.139) can be verified by

$$\frac{c_1\log(NS/\delta)}{12P_{\mathsf{min}}(s,a)}\geq-\frac{\log\frac{2NS}{\delta}}{\log(1-P_{\mathsf{min}}(s,a))}. \tag{E.188}$$

E.3.4 Proof of Theorem 17

Similar to the finite-horizon case, we shall first construct some hard discounted infinite-horizon RMDP instances and then characterize the sample complexity requirements over these instances.

E.3.4.1 Construction of hard problem instances

Construction of a collection of hard MDPs. Suppose there are two MDPs

$$\left\{\mathcal{M}_\phi = \left(\mathcal{S}, \mathcal{A}, P^\phi, r, \gamma\right) \mid \phi = \{0, 1\}\right\}.$$

Here, γ is the discount parameter, $\mathcal{S} = \{0, 1, \ldots, S-1\}$ is the state space, and $\mathcal{A} = \{0, 1\}$ is the action space. The transition kernel P^ϕ of either constructed MDP $\mathcal{M}_\phi$ is defined as

$$P^\phi(s' \mid s, a) = \begin{cases} p\mathbb{1}(s'=2) + (1-p)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0, \phi) \\ q\mathbb{1}(s'=2) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad (s,a) = (0, 1-\phi) \\ \mathbb{1}(s'=s) & \text{if} \quad s = 1 \text{ or } s = 2 \\ q\mathbb{1}(s'=s) + (1-q)\mathbb{1}(s'=1) & \text{if} \quad s > 2 \end{cases}, \tag{E.189}$$

where p and q are set as

$$p = 1 - \alpha \quad \text{and} \quad q = 1 - \alpha - \Delta \tag{E.190}$$

for some γ, α and Δ obeying

$$0 < \alpha \leq 1 - \gamma \leq 1/(2e^8) \leq \frac{1}{2} \quad \text{and} \quad \Delta \leq \frac{\alpha}{2}. \tag{E.191}$$

Here, α and Δ are some values that will be introduced later. Consequently, applying (E.190) directly leads to

$$1 \geq p \geq q \geq \gamma \geq \frac{1}{2}. \tag{E.192}$$

Note that state 1 and 2 are absorbing states. In addition, if the initial distribution is supported on states $\{0, 1, 2\}$, the MDP will always stay in the state $\{1, 2\}$ after the first transition.

Finally, we define the reward function as

$$r(s, a) = \begin{cases} 1 & \text{if } s = 0 \text{ or } s = 2 \\ 0 & \text{otherwise} \end{cases}. \tag{E.193}$$

Construction of the history/batch dataset. Define a useful state distribution (only supported on the state subset $\{0, 1, 2\}$) as

$$\mu(s) = \frac{1}{CS}\mathbb{1}(s=0) + \frac{1}{CS}\mathbb{1}(s=2) + \left(1 - \frac{2}{CS}\right)\mathbb{1}(s=1), \tag{E.194}$$

where $C > 0$ is some constant that determines the robust concentrability coefficient $C^\star_{\mathsf{rob}}$ (which will be made clear soon) and obeys

$$\frac{1}{CS} \leq \frac{1}{4}. \tag{E.195}$$

A batch dataset—consists of N i.i.d samples $\{(s_i, a_i, s_i')\}_{1 \leq i \leq N}$—is generated over the nominal environment $\mathcal{M}_\phi$ according to (7.24), with the behavior distribution chosen to be:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad d^{\mathsf{b}}(s,a) = \frac{\mu(s)}{2}. \tag{E.196}$$

Additionally, we choose the following initial state distribution:

$$\rho(s) = \begin{cases} 1, & \text{if } s = 0 \\ 0, & \text{otherwise} \end{cases}. \tag{E.197}$$

Uncertainty set of the transition kernels. We next describe the radius σ of the uncertainty set in our construction of the robust MDPs, along with some useful properties, which are similar to the finite-horizon case. To begin with, with slight abuse of notation, we introduce an important constant β defined as

$$\beta \coloneqq \frac{1}{2} \log \frac{1}{\alpha + \Delta} \geq 4. \tag{E.198}$$

The perturbed transition kernels in $\mathcal{M}_\phi$ is limited to the following uncertainty set

$$\mathcal{U}^\sigma(P^\phi) \coloneqq \otimes \mathcal{U}^\sigma\left(P^\phi_{s,a}\right), \qquad \mathcal{U}^\sigma(P^\phi_{s,a}) \coloneqq \left\{P_{s,a} \in \Delta(\mathcal{S}) : \mathsf{KL}\left(P_{s,a} \parallel P^\phi_{s,a}\right) \leq \sigma\right\}, \tag{E.199}$$

where $P^\phi_{s,a} \coloneqq P^\phi(\cdot \,|\, s,a) \in [0,1]^{1 \times S}$. Moreover, the radius of the uncertainty set σ obeys

$$\left(1 - \frac{3}{\beta}\right) \log \frac{1}{\alpha + \Delta} \leq \sigma \leq \left(1 - \frac{2}{\beta}\right) \log \frac{1}{\alpha + \Delta}. \tag{E.200}$$

For any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, we denote the infimum entry of the perturbed transition kernel $P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})$ moving to the next state s' as

$$\underline{P}^\phi(s' \,|\, s,a) \coloneqq \inf_{P_{s,a} \in \mathcal{U}^\sigma(P^\phi_{s,a})} P(s' \,|\, s,a). \tag{E.201}$$

As shall be seen, the transition from state 0 to state 2 plays an important role in the analysis, for convenience, we denote

$$\underline{p} \coloneqq \underline{P}^\phi(2 \,|\, 0, \phi), \qquad \underline{q} \coloneqq \underline{P}^\phi(2 \,|\, 0, 1-\phi). \tag{E.202}$$

With these definitions in place, we summarize some useful properties of the uncertainty set in the following lemma, which parallels Lemma 62 in the finite-horizon case.

Lemma 67. *Suppose β satisfies (E.198) and the uncertainty level σ satisfies (E.200). The perturbed transition kernels obey*

$$\underline{p} \geq \underline{q} \geq \frac{1}{\beta}. \tag{E.203}$$

Proof. The proof follows from the same arguments as Appendix E.2.3.4 by replacing H with $\frac{1}{1-\gamma}$; we omit the details for brevity. □

Value functions and optimal policies. Now we are positioned to derive the corresponding robust value functions and identify the optimal policies. For any MDP $\mathcal{M}_\phi$ with the above uncertainty set, denote $\pi_\phi^\star$ as the optimal policy. In addition, we denote the robust value function of any policy π (resp. the optimal policy $\pi_\phi^\star$) as $V_\phi^{\pi,\sigma}$ (resp. $V_\phi^{\star,\sigma}$). Then, we introduce the following lemma which describes some important properties of the robust value functions and optimal policies.

Lemma 68. *For any $\phi = \{0, 1\}$ and any policy π, one has*

$$V_\phi^{\pi,\sigma}(0) = 1 + \frac{\gamma}{1-\gamma} z_\phi^\pi, \tag{E.204}$$

where z_ϕ^π is defined as

$$z_\phi^\pi \coloneqq \underline{p}\pi(\phi \,|\, 0) + \underline{q}\pi(1-\phi \,|\, 0). \tag{E.205}$$

In addition, the optimal value functions and the optimal policies obey

$$V_\phi^{\star,\sigma}(0) = 1 + \frac{\gamma}{1-\gamma}\underline{p}, \quad V_\phi^{\star,\sigma}(2) = \frac{1}{1-\gamma}, \quad V_\phi^{\star,\sigma}(s) = 0 \quad \text{for } s = 1 \text{ or } s > 2, \tag{E.206a}$$

$$\pi_\phi^\star(\phi \,|\, s) = 1, \qquad \text{for } s \in \mathcal{S}. \tag{E.206b}$$

Moreover, choosing $S \geq 2\beta$, the robust single-policy clipped concentrability coefficient $C_{\mathsf{rob}}^\star$ obeys

$$C_{\mathsf{rob}}^\star = 2C. \tag{E.207}$$

Proof. See Appendix E.3.4.3. □

E.3.4.2 Establishing the minimax lower bound

Now we are positioned to provide the sample complexity lower bound. In view of Lemma 68, the smallest positive state transition probability of the optimal policy $\pi_\phi^\star$ under any nominal transition

kernel P^ϕ with $\phi \in \{0,1\}$ satisfies:

$$P^\star_{\mathsf{min}} \coloneqq \min_{s,s'} \left\{ P^\phi\left(s' \,|\, s, \pi^\star_\phi(s)\right) :\ P^\phi\left(s' \,|\, s, \pi^\star_\phi(s)\right) > 0 \right\} = P^\phi\left(1|0,\phi\right) = 1-p. \tag{E.208}$$

Our goal is to control the quantity w.r.t. any policy estimator $\widehat{\pi}$ based on the batch dataset and the chosen initial distribution ρ in (E.197), which gives

$$V_\phi^{\star,\sigma}(\rho) - V_\phi^{\widehat{\pi},\sigma}(\rho) = V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0). \tag{E.209}$$

Towards this, we first introduce the following lemma, which parallels the claim in (E.82)-(E.83) in the finite-horizon case.

Lemma 69. *Given* $\varepsilon \leq \frac{1}{384e^6(1-\gamma)\log\left(\frac{1}{\alpha}\right)} \leq \frac{1}{384e^6(1-\gamma)\log\left(\frac{1}{\alpha+\Delta}\right)}$, *choosing* $\Delta = 128e^6\sigma(1-q)\varepsilon(1-\gamma) \leq 128e^6(\alpha+\Delta)\varepsilon\log\left(\frac{1}{\alpha+\Delta}\right)(1-\gamma) \leq \frac{\alpha}{2}$, *one has for any policy* $\widehat{\pi}$,

$$V_\phi^{\star,\sigma}(0) - V_\phi^{\widehat{\pi},\sigma}(0) \geq 2\varepsilon\big(1-\widehat{\pi}(\phi\,|\,0)\big).$$

Proof. This lemma follows from the same arguments as Appendix E.2.3.6 except replacing H with $\frac{1}{1-\gamma}$ under the additional condition $\gamma \geq \frac{1}{2}$; we omit the details for brevity. □

Armed with this lemma, following the same arguments in Appendix E.2.3.2, we can complete the proof by observing that: let c_1 be some sufficient large constant, as long as the sample size is beneath

$$N \leq \frac{SC^\star_{\mathsf{rob}}\log 2}{4c_1 P^\star_{\mathsf{min}}\sigma^2(1-\gamma)^2\varepsilon^2}, \tag{E.210}$$

then we necessarily have

$$\inf_{\widehat{\pi}} \max_{\phi\in\{0,1\}} \mathbb{P}_\phi\left\{V_\phi^{\star,\sigma}(\rho) - V_\phi^{\widehat{\pi},\sigma}(\rho) \geq \varepsilon\right\} \geq \frac{1}{8}, \tag{E.211}$$

where $\mathbb{P}_\phi$ denote the probability conditioned on that the MDP is $\mathcal{M}_\phi$. We omit the details for brevity and complete the proof.

E.3.4.3 Proof of Lemma 68

For any $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, we first characterize the robust value function for any policy π over different states. due to state absorbing, the uncertainty set becomes a singleton containing the nominal distribution at state $s=1$ and $s=2$. It is easily observed that for any policy π, the robust

value functions at state $s = 1$ and $s = 2$ obey

$$V_\phi^{\pi,\sigma}(1) = \sum_{t=0}^{\infty} \gamma^t \cdot 0 = 0, \tag{E.212a}$$

$$V_\phi^{\pi,\sigma}(2) = \sum_{t=0}^{\infty} \gamma^t \cdot 1 = \frac{1}{1-\gamma}, \tag{E.212b}$$

since $r(1,a) = 0$ and $r(2,a) = 1$. In addition, for state $s > 2$, the perturbed transition kernel is supported on itself and state 1, both of which receive a reward of 0 by design (E.193), leading to

$$V_\phi^{\pi,\sigma}(s) = \sum_{t=0}^{\infty} \gamma^t \cdot 0 = 0, \qquad \text{for } s > 2. \tag{E.212c}$$

Moving onto the remaining states, the robust value function of state 0 satisfies

$$\begin{aligned}
V_\phi^{\pi,\sigma}(0) &= \mathbb{E}_{a\sim\pi(\cdot\,|\,0)}\left[r(0,a) + \gamma \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,a})} \mathcal{P}V_\phi^{\pi,\sigma}\right] \\
&\overset{\text{(i)}}{=} 1 + \gamma\pi(\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,\phi})} \mathcal{P}V_\phi^{\pi,\sigma} + \gamma\pi(1-\phi\,|\,0) \inf_{\mathcal{P}\in\mathcal{U}^\sigma(P^\phi_{0,1-\phi})} \mathcal{P}V_\phi^{\pi,\sigma} \\
&\overset{\text{(ii)}}{=} 1 + \gamma\pi(\phi\,|\,0)\Big[\underline{p}V_\phi^{\pi,\sigma}(2) + \big(1-\underline{p}\big)\,V_\phi^{\pi,\sigma}(1)\Big] + \gamma\pi(1-\phi\,|\,0)\Big[\underline{q}V_\phi^{\pi,\sigma}(2) + \big(1-\underline{q}\big)\,V_\phi^{\pi,\sigma}(1)\Big] \\
&\overset{\text{(iii)}}{=} 1 + \gamma V_\phi^{\pi,\sigma}(1) + \gamma z_\phi^\pi\left[V_\phi^{\pi,\sigma}(2) - V_\phi^{\pi,\sigma}(1)\right] \\
&= 1 + \frac{\gamma}{1-\gamma} z_\phi^\pi,
\end{aligned} \tag{E.213}$$

where (i) holds by the reward function defined in (E.193). To see (ii), note that (E.212) indicates $V_\phi^{\pi,\sigma}(2) \geq V_\phi^{\pi,\sigma}(1)$, so that the infimum is obtained by picking the smallest possible mass on the transition to state 2, provided by the definition in (E.202). Last but not least, (iii) follows by plugging in the definition of z_ϕ^π in (E.205), and the last identity is due to (E.212). Consequently, taking $\pi = \pi_\phi^\star$, we directly arrive at

$$V_\phi^{\star,\sigma}(0) = 1 + \frac{\gamma}{1-\gamma} z_\phi^{\pi^\star}. \tag{E.214}$$

Observing that the function $z\frac{\gamma}{1-\gamma}$ is increasing in z and z_ϕ^π is also increasing in $\pi(\phi\,|\,0)$ (see the fact $\underline{p} \geq \underline{q}$ in (E.203)), the optimal policy in state 0 thus obeys

$$\pi_\phi^\star(\phi\,|\,0) = 1. \tag{E.215}$$

Finally, plugging the above fact back into (E.205) leads to

$$z_\phi^\star \coloneqq z_\phi^{\pi^\star} = \underline{p}\pi_\phi^\star(\phi \,|\, 0) + \underline{q}\pi_\phi^\star(1-\phi \,|\, 0) = \underline{p}, \tag{E.216}$$

which combined with (E.214) yields

$$V_\phi^{\star,\sigma}(0) = 1 + \frac{\gamma}{1-\gamma}\underline{p}. \tag{E.217}$$

Regarding the optimal policy for the remaining states $s > 0$, since the action does not influence the state transition, without loss of generality, we choose the optimal policy to obey

$$\forall s > 0: \quad \pi_\phi^\star(\phi \,|\, s) = 1. \tag{E.218}$$

Proof of (E.207). To begin with, for any MDP $\mathcal{M}_\phi$ with $\phi \in \{0,1\}$, recall the definition of $C_{\mathsf{rob}}^\star$ as

$$C_{\mathsf{rob}}^\star = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)}. \tag{E.219}$$

Given $\pi_\phi^\star(\phi \,|\, s) = 1$ for all $s \in \mathcal{S}$ and the initial distribution $\rho(0) = 1$, for any $P \in \mathcal{U}^\sigma(P^\phi)$, we arrive at

$$d^{\star,P}(0,\phi) = (1-\gamma)\rho(0)\pi_\phi^\star(\phi \,|\, 0) = (1-\gamma), \tag{E.220}$$

which holds due to that the agent transits from state 0 to other states at the first step and then will never go back to state 0. In addition, one has for any $P \in \mathcal{U}^\sigma(P^\phi)$,

$$\begin{aligned} d^{\star,P}(2,\phi) &= (1-\gamma)P(2 \,|\, 0,\phi)\sum_{t=1}^{\infty}\gamma^t\big(P(2 \,|\, 2,\phi)\big)^t \\ &= (1-\gamma)P(2 \,|\, 0,\phi)\sum_{t=1}^{\infty}\gamma^t \overset{\text{(i)}}{\geq} \gamma\underline{p} \geq \frac{1}{2\beta}, \end{aligned} \tag{E.221}$$

where (i) holds by (E.202) and the final inequality follows from (E.203) and $\gamma \geq 1/2$. Armed with the above facts, we observe that

$$\max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(s,a), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)} = \max_{s\in\{0,1,2\},P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(s,\phi), \frac{1}{S}\right\}}{d^{\mathsf{b}}(s,\phi)} \tag{E.222}$$

which follows from the properties of the optimal policy in (E.218) and consequently $d^{\star,P}(s) = d^{\star,P}(s,\phi) = 0$ for all $s > 2$ and all $P \in \mathcal{U}^\sigma(P^\phi)$.

To continue, we control the term in states $\{0,1,2\}$ separately:

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(2,\phi),\frac{1}{S}\right\}}{d^{\mathsf{b}}(2,\phi)} \overset{\text{(i)}}{=} \frac{1}{Sd^{\mathsf{b}}(2,\phi)} \overset{\text{(ii)}}{=} \frac{2}{S\mu(2)} = 2C, \tag{E.223a}$$

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(0,\phi),\frac{1}{S}\right\}}{d^{\mathsf{b}}(0,\phi)} \leq \frac{1}{Sd^{\mathsf{b}}(0,\phi)} \overset{\text{(iii)}}{=} \frac{2}{S\mu(0)} = 2C, \tag{E.223b}$$

$$\max_{P\in\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(1,\phi),\frac{1}{S}\right\}}{d^{\mathsf{b}}(1,\phi)} \leq \frac{1}{Sd^{\mathsf{b}}(1,\phi)} \overset{\text{(iv)}}{=} \frac{2}{S\left(1-\frac{2}{CS}\right)} \overset{\text{(v)}}{\leq} \frac{4}{S} \overset{\text{(vi)}}{\leq} C, \tag{E.223c}$$

where (i) holds by (E.221) and $S \geq 2\beta$, (ii), (iii) and (iv) follow from the definitions in (E.196) and (E.194), (v) and (vi) arise from the assumption in (E.195). Plugging the above results back into (E.222) directly completes the proof of

$$C^\star_{\mathsf{rob}} = \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^\sigma(P^\phi)} \frac{\min\left\{d^{\star,P}(s,a),\frac{1}{S}\right\}}{d^{\mathsf{b}}(s,a)} = 2C.$$

Bibliography

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019). Reinforcement learning: Theory and algorithms. 19, 26

Agarwal, A., Kakade, S., and Yang, L. F. (2020a). Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR. 113, 298, 316

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2020b). Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pages 64–66. PMLR. 23, 24, 98, 129, 383

Agarwal, R. P., Meehan, M., and O'regan, D. (2001). *Fixed point theory and applications*, volume 141. Cambridge university press. 250

Agrawal, S. and Jia, R. (2023). Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. *Mathematics of Operations Research*, 48(1):363–392. 4

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77. 3, 20

Auer, P. and Ortner, R. (2006). Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19. 20

Azar, M. G., Kappen, H. J., Ghavamzadeh, M., and Munos, R. (2011). Speedy Q-learning. In *Advances in neural information processing systems*, pages 2411–2419. 24

Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349. 13, 23, 86, 113

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 263–272. JMLR. org. 4, 5, 6, 8, 20, 24, 74, 78, 84

Badrinath, K. P. and Kalathil, D. (2021). Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR. 23

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. (2019). Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pages 8002–8011. 4, 6, 24

Bartlett, P. and Tewari, A. (2009). Regal: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*, pages 35–42. AUAI Press. 21

Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208. 24

Bertsekas, D. P. (2017). *Dynamic programming and optimal control (4th edition)*. Athena Scientific. 28

Bertsimas, D., Gupta, V., and Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292. 13, 23

Best, A., Narang, S., Pasqualin, L., Barber, D., and Manocha, D. (2018). Autonovi-sim: Autonomous vehicle simulation platform with weather, sensing, and traffic control. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1048–1056. 2

Blanchet, J., Lu, M., Zhang, T., and Zhong, H. (2023). Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*. 23

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600. 13, 23

Bourel, H., Maillard, O., and Talebi, M. S. (2020). Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, pages 1056–1066. PMLR. 8, 20

Buckman, J., Gelada, C., and Bellemare, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations*. 8, 22

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR. 52, 79

Chen, M., Li, Y., Wang, E., Yang, Z., Wang, Z., and Zhao, T. (2021a). Pessimism meets invariance: Provably efficient offline mean-field multi-agent RL. *Advances in Neural Information Processing Systems*, 34. 8

Chen, R., Huang, P., and Shi, L. (2021b). Latent goal allocation for multi-agent goal-conditioned self-supervised imitation learning. *NeurIPS Workshop on Bayesian Deep Learning*. 133

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234. 24

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2021c). A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*. 24

Clavier, P., Pennec, E. L., and Geist, M. (2023). Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*. 14, 15, 113

Cui, Q. and Du, S. S. (2022). When are offline two-player zero-sum markov games solvable? *Advances in Neural Information Processing Systems*, 35:25779–25791. 8

Cui, Q. and Yang, L. F. (2021). Minimax sample complexity for turn-based stochastic game. In *Uncertainty in Artificial Intelligence*, pages 1496–1504. PMLR. 24

Dadashi, R., Rezaeifar, S., Vieillard, N., Hussenot, L., Pietquin, O., and Geist, M. (2021). Offline reinforcement learning with pseudometric learning. In *International Conference on Machine Learning*, pages 2307–2318. PMLR. 22

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30. 2

Derman, E. and Mannor, S. (2020). Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*. 23

Diehl, C., Sievernich, T., Krüger, M., Hoffmann, F., and Bertran, T. (2021). Umbrella: Uncertainty-aware model-based offline reinforcement learning leveraging planning. *arXiv preprint arXiv:2111.11097*. 7

Ding, W., Shi, L., Chi, Y., and Zhao, D. (2023). Seeing is not believing: Robust reinforcement learning against spurious correlation. *In submission. A short version at ICML Workshop on Spurious Correlations, Invariance and Stability*. 133

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021). Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR. 3, 6, 21, 57

Dong, J., Li, J., Wang, B., and Zhang, J. (2022). Online policy optimization for robust MDP. *arXiv preprint arXiv:2209.13841*. 23

Dong, K., Wang, Y., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*. 20, 24

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. (2017). Stochastic variance reduction methods for policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1049–1058. JMLR. org. 24

Duan, Y. and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*. 21

Duan, Y., Wang, M., and Wainwright, M. J. (2021). Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*. 21

Duchi, J. C. (2018). Introductory lectures on stochastic optimization. *The mathematics of data*, 25:99–186. 265, 369

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406. 13, 23

Ebert, F., Finn, C., Dasari, S., Xie, A., Lee, A., and Levine, S. (2018). Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*. 7

Efroni, Y., Merlis, N., Ghavamzadeh, M., and Mannor, S. (2019). Tight regret bounds for model-based reinforcement learning with greedy policies. *Advances in Neural Information Processing Systems*, 32. 4

Even-Dar, E. and Mansour, Y. (2003). Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25. 24

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for dynamics and control*, pages 486–489. PMLR. 79

Farahmand, A.-m., Szepesvári, C., and Munos, R. (2010). Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23. 79

Fatemi, M., Killian, T. W., Subramanian, J., and Ghassemi, M. (2021). Medical dead-ends and learning to identify high-risk states and treatments. *Advances in Neural Information Processing Systems*, 34:4856–4870. 1

Filippi, S., Cappé, O., and Garivier, A. (2010). Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. IEEE. 20

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118. 134

Fruit, R., Pirotta, M., and Lazaric, A. (2020). Improved analysis of UCRL2 with empirical Bernstein inequality. *arXiv preprint arXiv:2007.05456*. 20, 74, 78

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR. 22

Fulbright, N. R. (2017). The privacy implications of autonomous vehicles. 2

Gao, R. (2022). Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*. 13, 23

Gilbert, E. N. (1952). A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522. 262

Gower, R. M., Schmidt, M., Bach, F., and Richtárik, P. (2020). Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983. 24

Goyal, V. and Grand-Clement, J. (2022). Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*. 23

Han, S., Su, S., He, S., Han, S., Yang, H., and Miao, F. (2022). What is the solution for state adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*. 22

He, J., Zhou, D., and Gu, Q. (2021). Nearly minimax optimal reinforcement learning for discounted MDPs. *Advances in Neural Information Processing Systems*, 34:22288–22300. 24, 84, 86

Ho, C. P., Petrik, M., and Wiesemann, W. (2018). Fast bellman updates for robust MDPs. In *International Conference on Machine Learning*, pages 1979–1988. PMLR. 23

Ho, C. P., Petrik, M., and Wiesemann, W. (2021). Partial policy iteration for l1-robust Markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46. 23

Hu, Z. and Hong, L. J. (2013). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pages 1695–1724. 120, 127, 351

Huang, P., Xu, M., Zhu, J., Shi, L., Fang, F., and Zhao, D. (2022). Curriculum reinforcement learning using optimal transport via gradual domain adaptation. *Advances in Neural Information Processing Systems*. 133

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280. 13, 18, 22, 23, 31, 33, 34, 111, 120, 125, 127, 293, 294, 295

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710. 24

Jafarnia-Jahromi, M., Wei, C.-Y., Jain, R., and Luo, H. (2020). A model-free learning algorithm for infinite-horizon average-reward MDPs with near-optimal regret. *arXiv preprint arXiv:2006.04354*. 20, 24

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4). 3, 4, 8, 19, 20, 21, 84

Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33:2747–2758. 21

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR. 21

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873. 3, 4, 5, 6, 8, 20, 21, 24, 35, 36, 39, 42, 50, 66, 74, 78, 139, 140, 157, 158, 175, 176

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR. 24, 133

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. 8, 22, 53, 73, 120, 121

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323. 9, 24, 36, 55

Kakade, S. (2003). *On the sample complexity of reinforcement learning*. PhD thesis, University of London. 2

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63. 21

Kaufman, D. L. and Schaefer, A. J. (2013). Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410. 23

Kearns, M. J. and Singh, S. P. (1999). Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002. 109

Khamaru, K., Pananjady, A., Ruan, F., Wainwright, M. J., and Jordan, M. I. (2021). Is temporal difference learning optimal? an instance-dependent analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1013–1040. 24

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823. 22

Klopp, O., Lounici, K., and Tsybakov, A. B. (2017). Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564. 11

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274. 1

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc. 8, 22

Kumar, N., Derman, E., Geist, M., Levy, K., and Mannor, S. (2023). Policy gradient for s-rectangular robust Markov decision processes. *arXiv preprint arXiv:2301.13589*. 23

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22. 8, 19, 36, 84

Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer. 7

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press. 19, 36

Lee, J., Jeon, W., Lee, B., Pineau, J., and Kim, K.-E. (2021). Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR. 14

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*. 7, 22

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023a). Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*. 24, 134

Li, G., Shi, L., Chen, Y., and Chi, Y. (2023b). Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Information and Inference: A Journal of the IMA*, 12(2):969–1043. 20, 24, 55, 66, 74, 78

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022a). Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*. 21, 120, 121, 129, 316, 323, 353, 364, 381, 383, 394

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2023c). Breaking the sample size barrier in model-based reinforcement learning with a generative model. *accepted to Operations Research*. 13, 16, 23, 24, 98, 113, 129, 298, 383

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2021). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473. 24, 36, 52, 55, 133, 284, 285, 286

Li, L., Munos, R., and Szepesvári, C. (2014). On minimax optimal offline policy evaluation. *arXiv preprint arXiv:1409.3653*. 21

Li, Y., Zhao, T., and Lan, G. (2022b). First-order policy optimization for robust Markov decision process. *arXiv preprint arXiv:2209.10579*. 23

Liu, S., Ngiam, K. Y., and Feng, M. (2019). Deep reinforcement learning for clinical decision support: a brief survey. *arXiv preprint arXiv:1907.09475*. 1

Liu, S. and Su, H. (2020). γ-regret for non-episodic reinforcement learning. *arXiv:2002.05138*. 20

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274. 8, 22

Low, T. M., Chi, Y., Hoe, J., Kumar, S., Prabhakara, A., Shi, L., Sridhar, U., Tukanov, N., Wang, C., and Wu, Y. (2022). Zoom out: Abstractions for efficient radar algorithms on cots architectures. In *2022 IEEE International Symposium on Phased Array Systems & Technology (PAST)*, pages 1–6. IEEE. 133

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., and Bergstra, J. (2018). Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR. 2, 11

McGinnis, J. M., Olsen, L., Goolsby, W. A., Grossmann, C., et al. (2011). *Clinical data as the basic staple of health learning: Creating and protecting a public good: Workshop summary*. National Academies Press. 2

Ménard, P., Domingues, O. D., Shang, X., and Valko, M. (2021). UCB momentum Q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618. PMLR. 5, 6, 42

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*. 1

Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., and Peters, J. (2022). Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1):276–315. 22

Munos, R. (2007). Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561. 79

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR. 24

Nguyen-Tang, T., Gupta, S., and Venkatesh, S. (2021). Sample complexity of offline reinforcement learning with deep ReLU networks. *arXiv preprint arXiv:2103.06671*. 22, 133

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798. 13, 18, 31, 33, 34, 125

OpenAI (2023). Gpt-4 technical report. 1

Osband, I. and Van Roy, B. (2016). On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*. 21

Pacchiano, A., Ball, P., Parker-Holder, J., Choromanski, K., and Roberts, S. (2021). Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1413–1423. PMLR. 4

Panaganti, K. and Kalathil, D. (2022). Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR. 14, 15, 16, 17, 18, 19, 21, 23, 106, 109, 116, 118, 120, 124, 127, 128, 129, 303, 334, 383

Pananjady, A. and Wainwright, M. J. (2020). Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1):566–585. 23, 24

Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20. 283

Prudencio, R. F., Maximo, M. R., and Colombini, E. L. (2023). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*. 7

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons. 28

Qian, J., Fruit, R., Pirotta, M., and Lazaric, A. (2019). Exploration bonus for regret minimization in discrete and continuous average reward MDPs. *Advances in Neural Information Processing Systems*, 32. 20

Qiaoben, Y., Zhou, X., Ying, C., and Zhu, J. (2021). Strategically-timed state-observation attacks on deep reinforcement learning agents. In *ICML 2021 Workshop on Adversarial Machine Learning*. 22

Qu, G. and Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. *Conference on Learning Theory*, pages 3185–3205. 24

Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*. 13, 23

Rajaraman, N., Yang, L., Jiao, J., and Ramchandran, K. (2020). Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33. 52

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Neural Information Processing Systems (NeurIPS)*. 8, 9, 12, 22, 52, 53, 54, 79, 83, 84, 85, 86, 118, 120, 121

Ren, T., Li, J., Dai, B., Du, S. S., and Sanghavi, S. (2021). Nearly horizon-free offline reinforcement learning. *Advances in neural information processing systems*, 34. 21, 79

Rezaeifar, S., Dadashi, R., Vieillard, N., Hussenot, L., Bachem, O., Pietquin, O., and Geist, M. (2022). Offline reinforcement learning as anti-exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8106–8114. 22

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407. 4, 36

Roy, A., Xu, H., and Pokutta, S. (2017). Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30. 23

Saengkyongam, S., Thams, N., Peters, J., and Pfister, N. (2023). Invariant policy learning: A causal perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2

Sang, Y., Shi, L., and Liu, Y. (2018). Micro hand gesture recognition system using ultrasonic active sensing. *IEEE Access*, 6:49339–49347. 133

Shi, L. and Chi, Y. (2021). Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *IEEE Transactions on Information Theory*, 67(7):4784–4811. 133

Shi, L. and Chi, Y. (2022). Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*. 106, 316

Shi, L., Dadashi, R., Chi, Y., Castro, P. S., and Geist, M. (2023a). Offline reinforcement learning with on-policy Q-function regularization. *European Conference on Machine Learning*. 133

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 19967–20025. PMLR. 21, 22, 24, 118, 121

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. (2023b). The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*. xii, 15

Shi, L., Liu, D., and Thornton, J. (2021a). Robust camera pose estimation for image stitching. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2838–2842. IEEE. 133

Shi, L., Liu, D., Umeda, M., and Hana, N. (2021b). Fusion-based digital image correlation framework for strain measurement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1400–1404. IEEE. 133

Shi, L., Mirshekari, M., Fagert, J., Chi, Y., Noh, H. Y., Zhang, P., and Pan, S. (2019). Device-free multiple people localization through floor vibration. In *Proceedings of the 1st ACM International Workshop on Device-Free Human Sensing*, pages 57–61. ACM. 133

Shi, L., Zhang, Y., Pan, S., and Chi, Y. (2020). Data quality-informed multiple occupant localization using floor vibration sensing. In *Proceedings of the Twenty-first International Workshop on Mobile Computing Systems and Applications*. ACM. 133

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. (2018a). Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196. 24, 78

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2018b). Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM. 24, 36, 55

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359. 1

Smirnova, E., Dohmatob, E., and Mary, J. (2019). Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*. 23

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. (2006). PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888. 4, 5

Sun, K., Liu, Y., Zhao, Y., Yao, H., Jui, S., and Kong, L. (2021). Exploring the training robustness of distributional reinforcement learning against noisy state observations. *arXiv preprint arXiv:2109.08776*. 22

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press. 106, 129, 130

Szepesvári, C. (1997). The asymptotic convergence-rate of Q-learning. In *NIPS*, volume 10, pages 1064–1070. Citeseer. 24

Talebi, M. S. and Maillard, O.-A. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805. PMLR. 20, 74

Tamar, A., Mannor, S., and Xu, H. (2014). Scaling up robust MDPs using function approximation. In *International conference on machine learning*, pages 181–189. PMLR. 23

Tan, K. L., Esfandiari, Y., Lee, X. Y., and Sarkar, S. (2020). Robustifying reinforcement learning agents via action space adversarial training. In *2020 American control conference (ACC)*, pages 3959–3964. IEEE. 22

Tang, S. and Wiens, J. (2021). Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR. 7

Tao, T. (2012). *Topics in Random Matrix Theory*. Graduate Studies in Mathematics. American Mathematical Society, Providence, Rhode Island. 182

Tessler, C., Efroni, Y., and Mannor, S. (2019). Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pages 6215–6224. PMLR. 22

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR. 21

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270. 134

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202. 24

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*, volume 11. Springer. 109, 117, 265, 273, 277, 292, 325, 369

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR. 21

Uehara, M. and Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*. 8, 22

Uehara, M., Zhang, X., and Sun, W. (2022). Representation learning for online and offline RL in low-rank MDPs. 22

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press. 188, 189, 254, 313, 318, 335, 336, 392

Wai, H.-T., Hong, M., Yang, Z., Wang, Z., and Tang, K. (2019). Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems*, 32:5784–5795. 24

Wainwright, M. J. (2019a). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*. 24

Wainwright, M. J. (2019b). Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*. 24, 36, 55

Wang, S., Si, N., Blanchet, J., and Zhou, Z. (2023a). A finite sample complexity bound for distributionally robust Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR. 23

Wang, Y., Dong, K., Chen, X., and Wang, L. (2019). Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations*. 84

Wang, Y., Xu, M., Shi, L., and Chi, Y. (2023b). A trajectory is worth three sentences: Multimodal transformer for offline reinforcement learning. *The Conference on Uncertainty in Artificial Intelligence*. 133

Wang, Y. and Zou, S. (2021). Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34. 23

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292. 4, 24, 36, 52

Watkins, C. J. C. H. (1989). Learning from delayed rewards. *PhD thesis, King's College, University of Cambridge*. 36, 52

Weng, B., Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020). Momentum Q-learning with finite-sample convergence guarantee. *arXiv preprint arXiv:2007.15418*. 24

Wiesemann, W., Kuhn, D., and Rustem, B. (2013). Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183. 13, 33

Wolff, E. M., Topcu, U., and Murray, R. M. (2012). Robust control of uncertain Markov decision processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3372–3379. IEEE. 23

Woo, J., Joshi, G., and Chi, Y. (2023). The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *arXiv preprint arXiv:2305.10697*. 24

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021a). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694. 22

Xie, T. and Jiang, N. (2021). Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR. 79

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. (2021b). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34. 8, 9, 10, 12, 22, 52, 54, 55, 57, 66, 73, 74, 78, 118, 120, 121, 188

Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020). Finite-time analysis for double Q-learning. *Advances in Neural Information Processing Systems*, 33. 24

Xiong, Z., Eappen, J., Zhu, H., and Jagannathan, S. (2022). Defending observation attacks in deep reinforcement learning via detection and denoising. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 235–250. Springer. 22

Xu, H. and Mannor, S. (2012). Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300. 23

Xu, T., Wang, Z., Zhou, Y., and Liang, Y. (2019). Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations*. 24

Xu, T., Yang, Z., Wang, Z., and Liang, Y. (2021). A unified off-policy evaluation approach for general value function. *arXiv preprint arXiv:2107.02711*. 21

Xu, Z., Panaganti, K., and Kalathil, D. (2023). Improved sample complexity bounds for distributionally robust reinforcement learning. *arXiv preprint arXiv:2303.02783*. 23

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022a). The efficacy of pessimism in asynchronous Q-learning. *arXiv preprint arXiv:2203.07368*. 12, 22, 24, 283, 286

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022b). Model-based reinforcement learning is minimax optimal for offline zero-sum Markov games. *arXiv preprint arXiv:2206.04044*. 22

Yang, K., Yang, L., and Du, S. (2021). Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR. 4, 20, 24, 158

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized Lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561. 21

Yang, W., Wang, H., Kozuno, T., Jordan, S. M., and Zhang, Z. (2023). Avoiding model estimation in robust Markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*. 23

Yang, W., Zhang, L., and Zhang, Z. (2022). Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248. 14, 15, 16, 17, 18, 19, 21, 23, 114, 116, 118, 120, 124, 127, 128, 129, 322, 344, 383

Yin, M., Bai, Y., and Wang, Y.-X. (2021a). Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34:7677–7688. 24, 52

Yin, M., Bai, Y., and Wang, Y.-X. (2021b). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR. 24, 52

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. (2022). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*. 22

Yin, M. and Wang, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34. xii, 8, 9, 12, 22

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., and Finn, C. (2021a). Conservative data sharing for multi-task offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:11501–11516. 22

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. (2021b). Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967. 22

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. (2020). MOPO: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142. 22

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR. 24

Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34. 8, 22

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. (2022). Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR. 22

Zhang, H., Chen, H., Boning, D., and Hsieh, C.-J. (2021a). Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*. 2, 22

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. (2020a). Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037. 11, 22

Zhang, K., Kakade, S., Basar, T., and Yang, L. (2020b). Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33. 24

Zhang, Z., Ji, X., and Du, S. (2021b). Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. pages 4528–4531. 24

Zhang, Z., Zhou, Y., and Ji, X. (2020c). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33. 5, 6, 7, 20, 24, 35, 36, 38, 39, 42, 55, 66, 74, 78, 132

Zhang, Z., Zhou, Y., and Ji, X. (2021c). Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, pages 12653–12662. PMLR. 24, 86

Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z., and Yang, Z. (2022). Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pages 27117–27142. PMLR. 8

Zhou, Z., Bai, Q., Zhou, Z., Qiu, L., Blanchet, J., and Glynn, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR. 13, 18, 19, 21, 23, 33, 106, 109, 125, 128, 129, 130, 351, 352, 383

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*. 1