Provable Convolutional Sparse Coding via Nonconvex Optimization

Laixi Shi and Yuejie Chi

Department of Electrical and Computer Engineering, Carnegie Mellon University

Motivation

Convolutional Sparse Coding is a classical inverse problem that ubiquitously appears in various areas:

- Image Processing (Deblurring)
- Communications (Blind Channel Estimation)
- Array processing (Blind Gain and Phase Calibration)

Problem Formulation

- Let $\boldsymbol{y}_i \in \mathbb{R}^n$ be the convolution between a filter $\mathrm{g} \in \mathbb{R}^n$, and a sparse input $\boldsymbol{x}_i \in \mathbb{R}^n$:

$$\boldsymbol{y}_i = \mathrm{g} \circledast \boldsymbol{x}_i = \mathcal{C}(\mathrm{g})\boldsymbol{x}_i, \quad i = 1, \ldots, p, \tag{1}$$

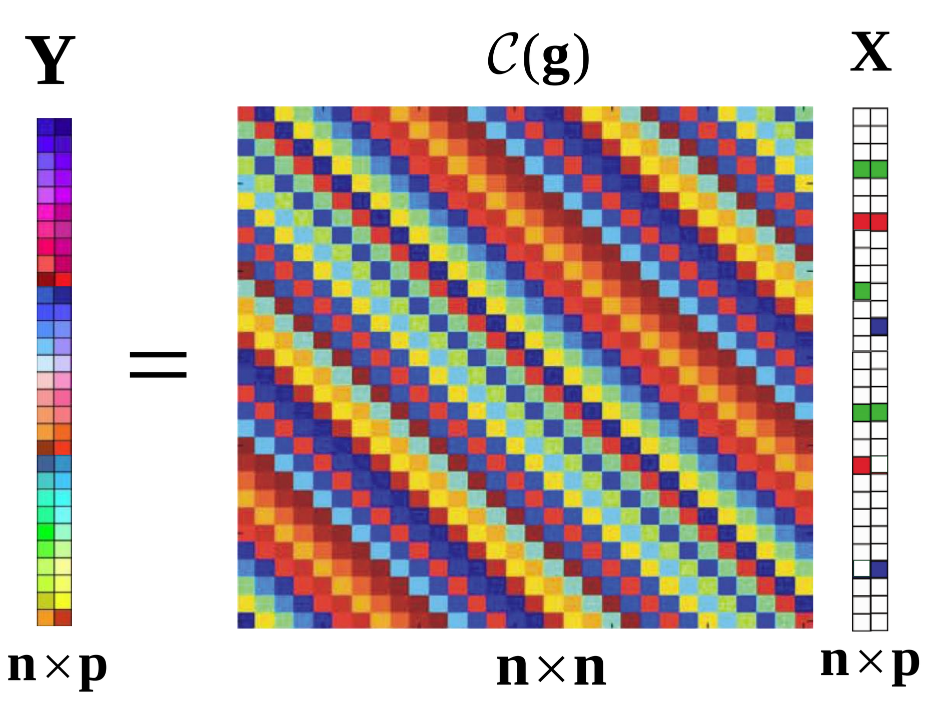
where the total number of observations is p, $\circledast$ denotes the circular convolution.

- $\mathcal{C}(\mathrm{g}) \in \mathbb{R}^{n\times n}$ is the circulant matrix spanned by $\mathrm{g} = [g_1, \ldots, g_n]^T$, given as

$$\mathcal{C}(\mathrm{g}) = \begin{bmatrix} g_1 & g_n & \cdots & g_2 \\ g_2 & g_1 & \cdots & g_3 \\ \vdots & \vdots & \ddots & \vdots \\ g_n & g_{n-1} & \cdots & g_1 \end{bmatrix}.$$

- Denote $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p] \in \mathbb{R}^{n\times p}$ and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$, the signal model can be rewritten as

$$\boldsymbol{Y} = \mathcal{C}(\mathrm{g})\boldsymbol{X}.$$



Goal: Provide a computationally-efficient algorithm with theoretical guarantees to recover both the signals $\{\boldsymbol{x}_i\}_{i=1}^p$ and the kernel g from their convolution $\{\boldsymbol{y}_i\}_{i=1}^p$, up to a circulant shift and a scaling factor.

- $\boldsymbol{y}_i = (\beta\mathcal{S}_j(\mathrm{g})) \circledast (\beta^{-1}\mathcal{S}_{-j}(\boldsymbol{x}_i))$, where $\mathcal{S}_j(\cdot)$ is a circulant shift by j positions, $j = 1, \ldots, n$, and $\beta \neq 0$ is an arbitrary scalar.

A Nonconvex Approach via Manifold Gradient Descent

- **Observation:** Assuming $\mathcal{C}(\mathrm{g})$ is invertible [Li, Lee, and Bresler], there exists a unique inverse filter $\mathrm{g}_{\mathrm{inv}} \in \mathbb{R}^n$ such that $\mathcal{C}(\mathrm{g})^{-1} := \mathcal{C}(\mathrm{g}_{\mathrm{inv}})$, which allows us to convert the bilinear form (1) into a linear form,

$$\mathcal{C}(\mathrm{g}_{\mathrm{inv}})\boldsymbol{y}_i = \mathcal{C}(\mathrm{g}_{\mathrm{inv}})\mathcal{C}(\mathrm{g})\boldsymbol{x}_i = \boldsymbol{x}_i, \quad i = 1, \ldots, p.$$

- Due to the sparsity of $\{\boldsymbol{x}_i\}_{i=1}^p$, we are motivated to recover $\mathrm{g}_{\mathrm{inv}}$ by seeking a vector $\boldsymbol{h}$ that minimizes the cardinality of $\mathcal{C}(\boldsymbol{h})\boldsymbol{y}_i = \mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}$:

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \frac{1}{p}\sum_{i=1}^{p} \|\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}\|_0 .$$

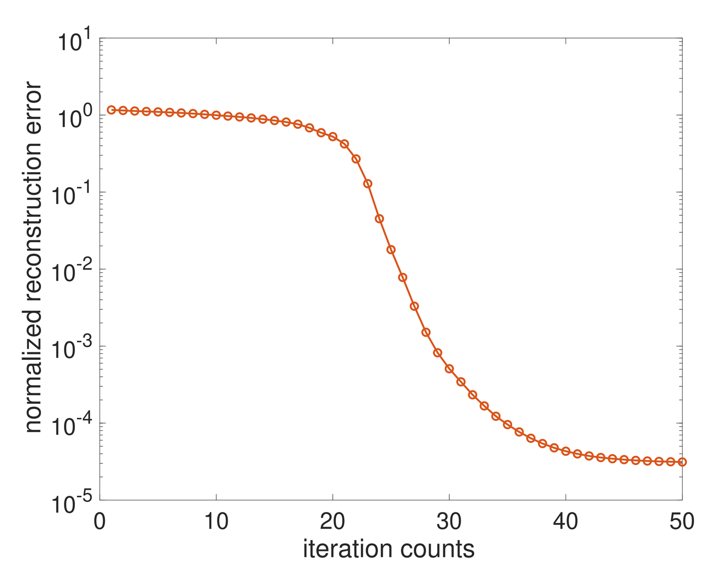
- **Problematic** with issues: (1) has a trivial solution $\boldsymbol{h} = \boldsymbol{0}$; (2) computationally intractable.
- We propose an alternative **nonconvex formulation** with pre-conditioning:

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} f(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{R}\boldsymbol{h}) \quad \text{s.t.} \quad \|\boldsymbol{h}\|_2 = 1,$$

where $\boldsymbol{R} = \left[\frac{1}{\theta np}\sum_{i=1}^p \mathcal{C}(\boldsymbol{y}_i)^\top \mathcal{C}(\boldsymbol{y}_i)\right]^{-1/2}$ is a preconditioning matrix depending only on $\{\boldsymbol{y}_i\}_{i=1}^p$, and $\psi_\mu(z) = \mu \log\cosh(z/\mu)$ is a convex surrogate with μ controlling the smoothness.

- **Manifold Gradient Descent (MGD):**

$$\boldsymbol{h}_{t+1} := (\boldsymbol{h}_t - \eta\partial f(\boldsymbol{h}_t)) \,/\, \|\boldsymbol{h}_t - \eta\partial f(\boldsymbol{h}_t)\|_2$$



- Motivated by such surprising success of MGD: ***can we establish theoretical guarantees of MGD to recover the filter for MSBD?***

Theoretical Analysis

Assumptions: Our theoretical guarantee relies on the following assumptions:

A1) The input sparse signal $\boldsymbol{X} = \boldsymbol{\Omega} \odot \boldsymbol{R}$ is assumed of a *Bernoulli-Gaussian model*,
- $\boldsymbol{\Omega}$ is an i.i.d. Bernoulli matrix with parameter θ,
- $\boldsymbol{R}$ is an independent random matrix with i.i.d. random Gaussian variables drawn from $\mathcal{N}(0,1)$.

A2) $\mathcal{C}(\mathrm{g})$ is invertible with the condition number κ, i.e. $\kappa = \sigma_1(\mathcal{C}(\mathrm{g}))/\sigma_n(\mathcal{C}(\mathrm{g}))$.

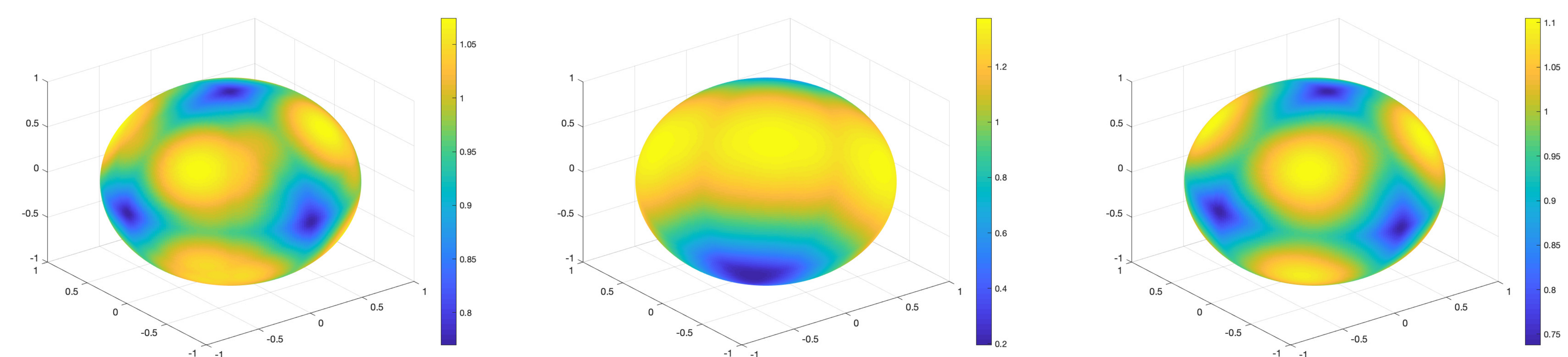
Theorem 1 (Benign Geometry of $f(\boldsymbol{h})$): Under the assumptions, let $\mathbf{w} = [h_1, \cdots, h_{n-1}]^\top$ and $\phi(\mathbf{w}) = f(\boldsymbol{h})$. For any $\theta \in (0, \frac{1}{3})$, when $\mu \lesssim \min\{\theta, n^{-3/4}\log^{-1} n\}$ and the number of measurements $p \gtrsim \kappa^8 n^{4.5}$ up to logarithmic factors, it holds with high probability that for $\boldsymbol{h}(\mathbf{w}) \in \mathcal{S}_{1/(4\log n)}^{(n+)}$, where $\mathcal{S}_\xi^{(i\pm)} = \left\{\boldsymbol{h} : h_i \gtrless 0, \frac{h_i^2}{\|\boldsymbol{h}_{\setminus i}\|_\infty^2} \geqslant 1 + \xi\right\}$:

$$\text{(large gradient)} \quad \mathbf{w}^\top \nabla\phi_o(\mathbf{w}) / \|\mathbf{w}\|_2 \gtrsim \theta/\log n, \quad \text{if } \|\mathbf{w}\|_2 > \mu/(4\sqrt{2}),$$

$$\text{(strong convexity)} \quad \nabla^2\phi_o(\mathbf{w}) \succeq \frac{n\theta}{5\sqrt{2\pi}\mu}\boldsymbol{I}, \quad \text{if } \|\mathbf{w}\|_2 \leq \mu/(4\sqrt{2}).$$

In addition, the function $\phi(\mathbf{w}) = f(\boldsymbol{h})$ has exactly one unique local minimizer $\mathbf{w}^\star$ near $\mathbf{0}$.

- Directly applying Theorem 1 for $\mathcal{S}_{1/(4\log n)}^{(n+)}$ to $2n$ subsets $\{\mathcal{S}_{1/(4\log n)}^{(i\pm)}\}_{i=1}^n$, we observe that there is no saddle points or spurious local minimizers except those corresponding to shift and sign-flipped ground truth with high probability.



(a) orthogonal filter $\mathcal{C}(\mathrm{g}_1)$ no pre-conditioning — (b) general filter $\mathcal{C}(\mathrm{g}_2)$ no pre-conditioning — (c) general filter $\mathcal{C}(\mathrm{g}_2)$ with pre-conditioning

Theorem 2 (Convergence Guarantee for MGD): Instate the assumptions of Theorem 1, with $O(\log n)$ independent random initializations selected uniformly over the sphere, it is guaranteed to obtain a vector $\boldsymbol{h}^{(T)}$ that accurately recover $\mathrm{g}_{\mathrm{inv}}$ up to scale and shift ambiguity:

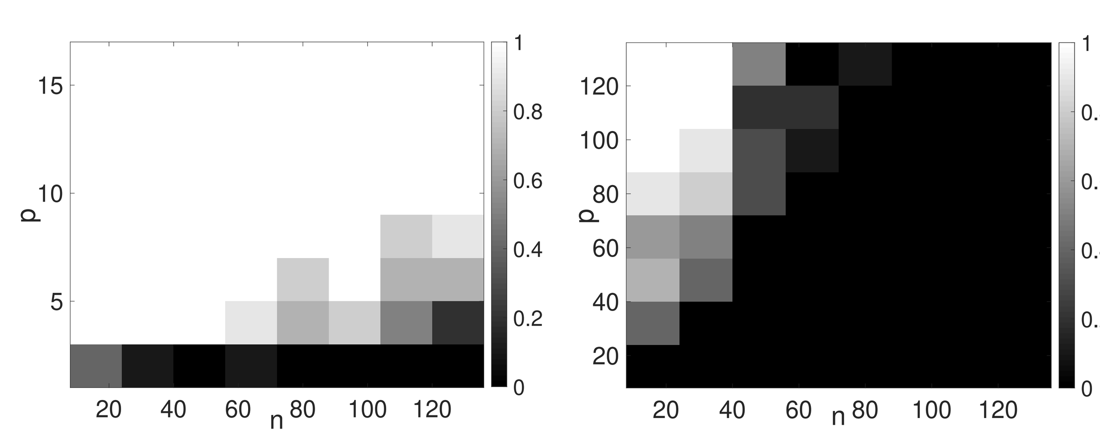
$$\min_{j\in[n]} \|\boldsymbol{h}^{(T)} \pm \mathcal{S}_j(\mathrm{g}_{\mathrm{inv}})\|_2 \lesssim \frac{\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p \log^2 n}{p}} + \epsilon$$

for any $\epsilon > 0$, in $T \lesssim \frac{n^3\log p}{\mu^2\theta^2} + \frac{n\log p}{\theta^2}\log\left(\frac{\mu}{\epsilon}\right)$ iterations of MGD.

Numerical Experiments

Blind deconvolution with Synthetic data:

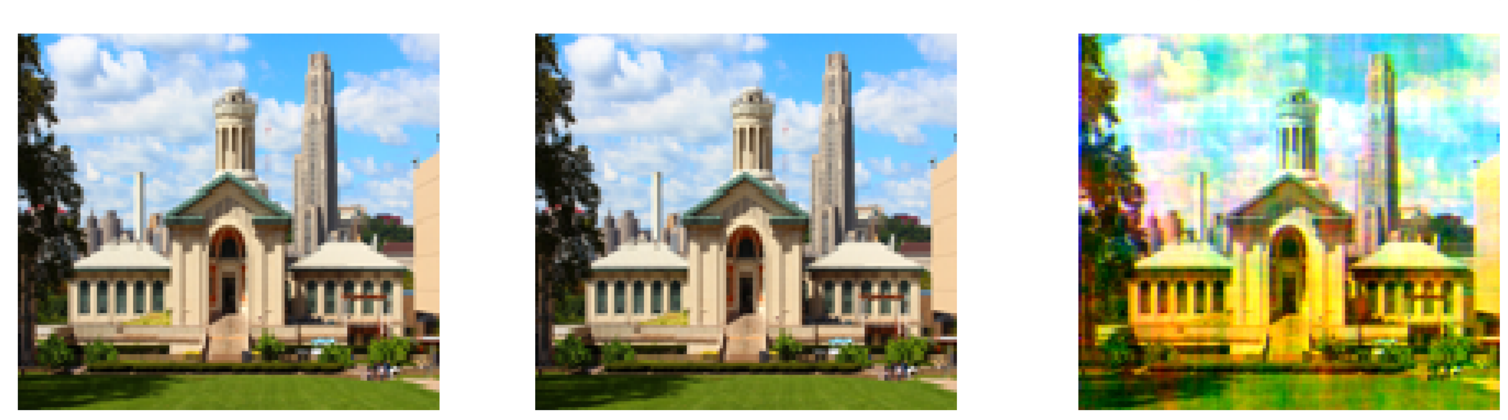
- The entries of $\boldsymbol{X}$ are drawn i.i.d. from a Bernoulli-Gaussian distribution with sparsity θ.
- We declare the recovery is successful if $\|\mathcal{C}(\mathrm{g})\boldsymbol{R}\boldsymbol{h}^{(T)}\|_\infty / \|\mathcal{C}(\mathrm{g})\boldsymbol{R}\boldsymbol{h}^{(T)}\|_2 > 0.99$.



(a) ours (n, p) — (b) [Li and Bresler] (n, p)

2D Image Blind Deconvolution:

- The size of the observations are $n = 128 \times 128$, $\theta = 0.1$, $p = 1000$ (significantly $< n$).



(a) true image — (b) recovery via ours — (c) recovery via [Li and Bresler]

Acknowledgment

This work is supported in part by ONR under the grants N00014-18-1-2142 and N00014-19-1-2404, and by NSF under the grants CAREER ECCS-1818571, CCF-1806154 and CCF-1901199.

Reference

Laixi Shi and Y. Chi, "Manifold Gradient Descent Solves Multi-Channel Sparse Blind Deconvolution Provably and Efficiently," *arXiv preprint, arXiv 1911.11167*, 2019.

Yanjun Li and Yoram Bresler, "Global geometry of multichannel sparse blind deconvolution on the sphere," in *Advances in Neural Information Processing Systems, 2018, pp. 1132–1143.*