# Provably and Efficient Manifold Gradient Descent for Multi-Channel Sparse Blind Deconvolution

Laixi Shi and Yuejie Chi

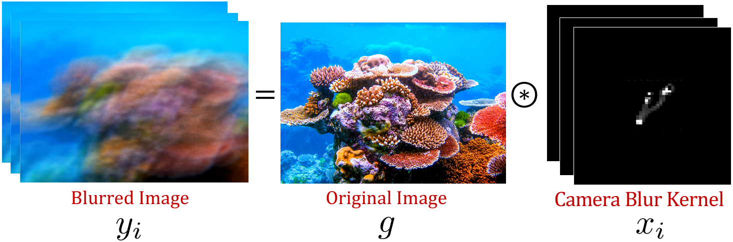
Dept.of Electrical and Computer Engineering,
Carnegie Mellon University

ICASSP 2020

Carnegie Mellon University

# Motivation

# Image superresolution/deblurring




*How to find the high-resolution original image and the blurring kernels simultaneously?*

# Formulation

# Multi-channel sparse blind deconvolution (MSBD)

**Problem Formulation**: the $i$-th observed signal $\boldsymbol{y}_i \in \mathbb{R}^n$ can be expressed as:

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

- $\boldsymbol{g}$ is a filter, and $\boldsymbol{x}_i \in \mathbb{R}^n$ is a sparse input signal.
- $p$ is the total number of observations, and $\circledast$ denote the circulant convolution.

# Multi-channel sparse blind deconvolution (MSBD)

**Problem Formulation**: the $i$-th observed signal $\boldsymbol{y}_i \in \mathbb{R}^n$ can be expressed as:

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

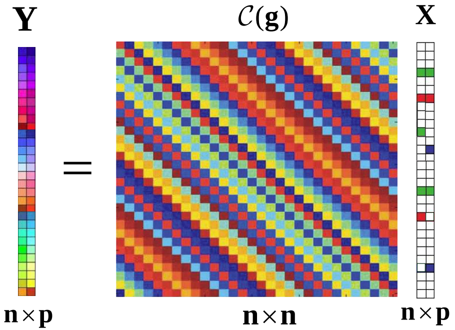
- $\boldsymbol{g}$ is a filter, and $\boldsymbol{x}_i \in \mathbb{R}^n$ is a sparse input signal.
- $p$ is the total number of observations, and $\circledast$ denote the circulant convolution.
- $\boldsymbol{g} = [g_1, g_2, \cdots, g_n]^\top$ and circulant matrix $\mathcal{C}(\boldsymbol{g}) \in \mathbb{R}^{n \times n}$:

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

# Multi-channel sparse blind deconvolution (MSBD)

- $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p] \in \mathbb{R}^{n\times p}$, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$ :

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




- **Goal:** recover both the unknown signals $\{\boldsymbol{x}_i\}_{i=1}^p$ and the kernel $\boldsymbol{g}$ from multiple observations $\{\boldsymbol{y}_i\}_{i=1}^p$

# Ambiguities

- The bilinear form of the observations:

$$\boldsymbol{y}_i = (\beta \cdot \mathcal{S}_j(\boldsymbol{g})) \circledast \frac{\mathcal{S}_{-j}(\boldsymbol{x}_i)}{\beta},$$

  where $\mathcal{S}_j(\boldsymbol{z})$ is the $j$-th circulant shift of the vector $\boldsymbol{z}$, $\beta \neq 0$ is an arbitrary scalar.

# Ambiguities

- The bilinear form of the observations:

$$\boldsymbol{y}_i = (\beta \cdot \mathcal{S}_j(\boldsymbol{g})) \circledast \frac{\mathcal{S}_{-j}(\boldsymbol{x}_i)}{\beta},$$

  where $\mathcal{S}_j(\boldsymbol{z})$ is the $j$-th circulant shift of the vector $\boldsymbol{z}$, $\beta \neq 0$ is an arbitrary scalar.
- **Challenge:** Scaling and shift ambiguities $\rightarrow$ $\boldsymbol{g}$ and $\{\boldsymbol{x}_i\}_{i=1}^{p}$ are not uniquely identifiable.

# Ambiguities

- The bilinear form of the observations:

$$\boldsymbol{y}_i = (\beta \cdot \mathcal{S}_j(\boldsymbol{g})) \circledast \frac{\mathcal{S}_{-j}(\boldsymbol{x}_i)}{\beta},$$

  where $\mathcal{S}_j(\boldsymbol{z})$ is the $j$-th circulant shift of the vector $\boldsymbol{z}$, $\beta \neq 0$ is an arbitrary scalar.
- **Challenge:** Scaling and shift ambiguities $\rightarrow \boldsymbol{g}$ and $\{\boldsymbol{x}_i\}_{i=1}^p$ are not uniquely identifiable.
- **Goal:** recover filter $\boldsymbol{g}$ and sparse inputs $\{\boldsymbol{x}_i\}_{i=1}^p$, up to scaling and shift ambiguity.

# Bilinear to linear

- $\mathcal{C}(\boldsymbol{g})$ is invertible $\rightarrow$ a unique inverse filter $\boldsymbol{g}_{\text{inv}}$:

$$\mathcal{C}(\boldsymbol{g}_{\text{inv}})\mathcal{C}(\boldsymbol{g}) = \mathcal{C}(\boldsymbol{g})\mathcal{C}(\boldsymbol{g}_{\text{inv}}) = \boldsymbol{I}.$$

# Bilinear to linear

- $\mathcal{C}(\boldsymbol{g})$ is invertible $\rightarrow$ a unique inverse filter $\boldsymbol{g}_{\text{inv}}$:

$$\mathcal{C}(\boldsymbol{g}_{\text{inv}})\mathcal{C}(\boldsymbol{g}) = \mathcal{C}(\boldsymbol{g})\mathcal{C}(\boldsymbol{g}_{\text{inv}}) = \boldsymbol{I}.$$

- **Bilinear to linear**: multiply $\mathcal{C}(\boldsymbol{g}_{\text{inv}})$ on both side,

$$\begin{aligned} \boldsymbol{y}_i &= \mathcal{C}(\boldsymbol{g})\boldsymbol{x}_i \rightarrow \\ \mathcal{C}(\boldsymbol{g}_{\text{inv}})\boldsymbol{y}_i &= \mathcal{C}(\boldsymbol{g}_{\text{inv}})\mathcal{C}(\boldsymbol{g})\boldsymbol{x}_i = \underbrace{\boldsymbol{x}_i}_{\text{sparse}} \quad i = 1, \ldots, p. \end{aligned}$$

# A natural formulation

- **Exploiting the sparsity of $\{x_i\}_{i=1}^{p}$**: seek $\boldsymbol{h}$ that minimize the cardinality of $\mathcal{C}(\boldsymbol{h})\boldsymbol{y}_i = \mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}$:

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

  - $\|\cdot\|_0$ is the pseudo-$\ell_0$ norm: counts the cardinality of the nonzero entries of the input vector.

# A natural formulation

- **Exploiting the sparsity of $\{x_i\}_{i=1}^{p}$**: seek $\boldsymbol{h}$ that minimize the cardinality of $\mathcal{C}(\boldsymbol{h})\boldsymbol{y}_i = \mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}$:

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

  - $\|\cdot\|_0$ is the pseudo-$\ell_0$ norm: counts the cardinality of the nonzero entries of the input vector.

- Problematic for two reasons:
  1. has a trivial solution $\boldsymbol{h} = \mathbf{0}$.
  2. the cardinality minimization is computationally intractable.

# A natural formulation

- **Exploiting the sparsity of $\{x_i\}_{i=1}^p$**: seek $\boldsymbol{h}$ that minimize the cardinality of $\mathcal{C}(\boldsymbol{h})\boldsymbol{y}_i = \mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}$:

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

  - $\|\cdot\|_0$ is the pseudo-$\ell_0$ norm: counts the cardinality of the nonzero entries of the input vector.

- Problematic for two reasons:
  1. has a trivial solution $\boldsymbol{h} = \boldsymbol{0}$.
  2. the cardinality minimization is computationally intractable.

*How to recover $\boldsymbol{g}_{\mathrm{inv}}$ provably and efficiently ?*

# Our nonconvex formulation

- We propose a nonconvex optimization formulation (following [Sun, et al, 2017][1] ,[Li and Bresler, 2019][2]) :

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \; f_o(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^{p} \underbrace{\psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h})}_{\text{convex surrogate}} \quad \text{s.t} \quad \underbrace{\|\boldsymbol{h}\|_2 = 1}_{\text{nonconvex}}$$

  - Add a spherical constraint.

[1] Ju Sun, Qing Qu, and John Wright. "Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture". In: *IEEE Transactions on Information Theory* 63.2 (2017), pp. 853–884.

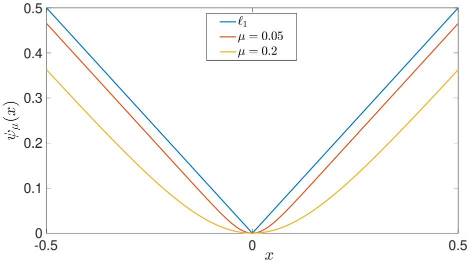
[2] Yanjun Li and Yoram Bresler. "Multichannel sparse blind deconvolution on the sphere". In: *IEEE Transactions on Information Theory* 65.11 (2019), pp. 7415–7436.

# Our nonconvex formulation

- We propose a nonconvex optimization formulation (following [Sun, et al, 2017][1] ,[Li and Bresler, 2019][2]) :

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \; f_o(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^{p} \underbrace{\psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h})}_{\text{convex surrogate}} \quad \text{s.t} \quad \underbrace{\|\boldsymbol{h}\|_2 = 1}_{\text{nonconvex}}$$

  - Add a spherical constraint.
  - Relax to a convex smooth surrogate: $\psi_\mu(z) = \mu \log\cosh(z/\mu)$, where $\mu$ controls the smoothness of the surrogate.

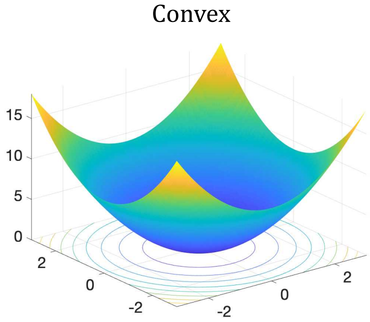



[1] Ju Sun, Qing Qu, and John Wright. "Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture". In: *IEEE Transactions on Information Theory* 63.2 (2017), pp. 853–884.

[2] Yanjun Li and Yoram Bresler. "Multichannel sparse blind deconvolution on the sphere". In: *IEEE Transactions on Information Theory* 65.11 (2019), pp. 7415–7436.

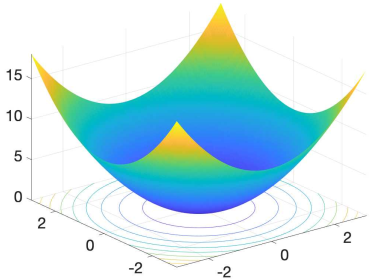
# Optimization Geometry

# *Convex vs nonconvex: optimization geometry*

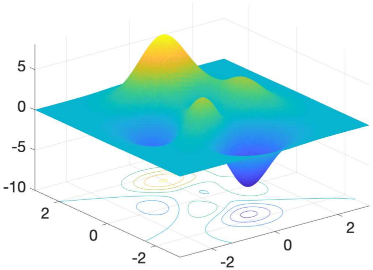


# *Convex vs nonconvex: optimization geometry*

Convex



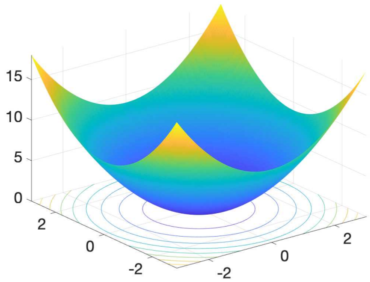

Nonconvex




Unique global minimizer

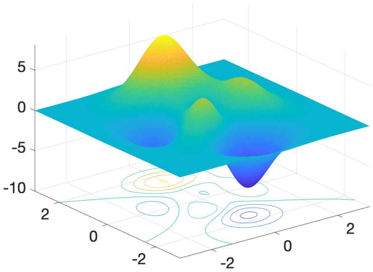
# *Convex vs nonconvex: optimization geometry*

Convex




Unique global minimizer

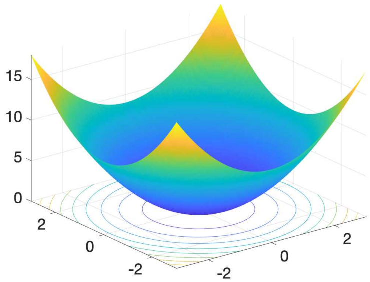
Nonconvex




saddle points and spurious local minimizers

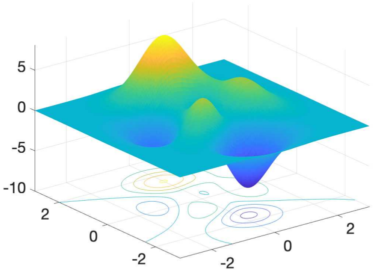
# Convex vs nonconvex: optimization geometry

Convex




Unique global minimizer

Nonconvex




saddle points and spurious local minimizers

*Is our objective landscape geometry of MSBD bad ?*

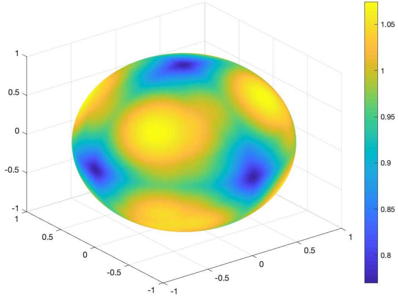
# Benign geometry in the orthogonal case

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

# Benign geometry in the orthogonal case

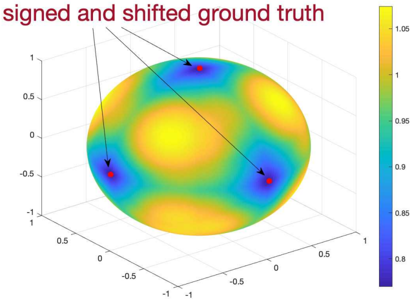
$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \; f_o(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}) \quad \text{s.t} \quad \|\boldsymbol{h}\|_2 = 1$$

- The landscape of the loss value $f_o(\boldsymbol{h})$ with respect to $\boldsymbol{h}$:
  - $\mathcal{C}(\boldsymbol{g}) = \boldsymbol{I}$.
  - $n = 3, p = 30$.



# Benign geometry in the orthogonal case

- The landscape of the loss value $f_o(\boldsymbol{h})$ with respect to $\boldsymbol{h}$:
  - $\mathcal{C}(\boldsymbol{g}) = \boldsymbol{I}$.
  - $2n = 6$ ground truth $\{\pm \boldsymbol{e}_i\}_{i=1}^{3}$




- **Benign geometry**: $2n$ local minimizers are approximately all shift and signed variants of the ground truth ($\{\pm \boldsymbol{e}_i\}_{i=1}^{3}$), and symmetrically distributed over the sphere.

# Manifold gradient descent (MGD)

- Manifold gradient descent:

$$\boldsymbol{h}_{t+1} := \frac{\boldsymbol{h}_t - \eta_t \partial f_o(\boldsymbol{h}_t)}{\|\boldsymbol{h}_t - \eta_t \partial f_o(\boldsymbol{h}_t)\|_2},$$

where $\eta_t$ is the stepsize, $\partial f_o(\boldsymbol{h}) = (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^\top)\nabla f_o(\boldsymbol{h})$, and $\nabla f_o(\boldsymbol{h})$ is the Euclidean gradient of $f_o(\boldsymbol{h})$.

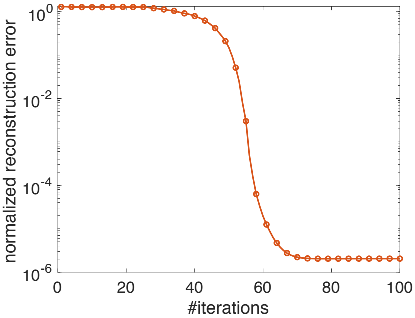
# Manifold gradient descent (MGD)

- Manifold gradient descent:

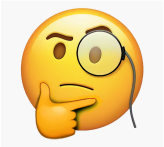
$$\boldsymbol{h}_{t+1} := \frac{\boldsymbol{h}_t - \eta_t \partial f_o(\boldsymbol{h}_t)}{\|\boldsymbol{h}_t - \eta_t \partial f_o(\boldsymbol{h}_t)\|_2},$$

where $\eta_t$ is the stepsize, $\partial f_o(\boldsymbol{h}) = (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^\top)\nabla f_o(\boldsymbol{h})$, and $\nabla f_o(\boldsymbol{h})$ is the Euclidean gradient of $f_o(\boldsymbol{h})$.

- With random initialization, $n = 128, p = 16$.




Surprising success of nonconvex optimization



# Theoretical guarantee

*Can we establish theoretical guarantee for the simple and efficient MGD based on nonconvex optimization formulation?*

# Theoretical guarantee

*Can we establish theoretical guarantee for the simple and efficient MGD based on nonconvex optimization formulation?*

**Yes. The statistical model will help !**

# Main Theoretical Results

# Assumptions

- **Inputs are sparse**: the inputs $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p]$ is under Bernoulli-Gaussian[3] model $\mathrm{BG}(\theta)$.
  - Each entry $x$ in $\boldsymbol{X}$ is an i.i.d variable satisfing $x = \Omega \cdot z$, where $\Omega$ is a Bernoulli variable with parameter $\theta$ and $z \sim \mathcal{N}(0, 1)$.

[3] Qing Qu et al. "Analysis of the Optimization Landscapes for Overcomplete Representation Learning". In: *arXiv preprint arXiv:1912.02427* (2019).

[4] Yanjun Li, Kiryung Lee, and Yoram Bresler. "A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models". In: *arXiv preprint arXiv:1501.06120* (2015).

# Assumptions

- **Inputs are sparse**: the inputs $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p]$ is under Bernoulli-Gaussian[3] model $\mathrm{BG}(\theta)$.
  - Each entry $x$ in $\boldsymbol{X}$ is an i.i.d variable satisfing $x = \Omega \cdot z$, where $\Omega$ is a Bernoulli variable with parameter $\theta$ and $z \sim \mathcal{N}(0, 1)$.
- **$\mathcal{C}(\boldsymbol{g})$ is invertible**[4]: ensure the identifiability of the filter $\boldsymbol{g}$.
  - The condition number of $\mathcal{C}(\boldsymbol{g})$ is $\kappa$, i.e.

$$\kappa = \sigma_1(\mathcal{C}(\boldsymbol{g}))/\sigma_n(\mathcal{C}(\boldsymbol{g}))$$

.

[3] Qing Qu et al. "Analysis of the Optimization Landscapes for Overcomplete Representation Learning". In: *arXiv preprint arXiv:1912.02427* (2019).

[4] Yanjun Li, Kiryung Lee, and Yoram Bresler. "A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models". In: *arXiv preprint arXiv:1501.06120* (2015).

# Main results

- Distance metric to measure the success recovery:

$$\operatorname{dist}(\boldsymbol{h}, \boldsymbol{g}_{\text{inv}}) = \min_{j \in [n]} \|\boldsymbol{g}_{\text{inv}} \pm \mathcal{S}_j(\boldsymbol{h})\|_2.$$

# Main results

- Distance metric to measure the success recovery:

$$\operatorname{dist}(\boldsymbol{h}, \boldsymbol{g}_{\text{inv}}) = \min_{j \in [n]} \|\boldsymbol{g}_{\text{inv}} \pm \mathcal{S}_j(\boldsymbol{h})\|_2.$$

**Theorem (Shi and Chi, 2019)**

*Instate the assumptions above, for $\theta \in (0, \frac{1}{3})$, when $\mu$ is small enough, with $O(\log n)$ random initializations, the output $\hat{\boldsymbol{h}}$ of MGD with a proper step size will satisfy:*

$$\operatorname{dist}(\hat{\boldsymbol{h}}, \boldsymbol{g}_{\text{inv}}) \lesssim \frac{\kappa^4}{\theta^2}\sqrt{\frac{n}{p}}$$

*in polynomial iterations, provided* $p \gtrsim \frac{\kappa^8 n^{4.5} \log^4 p \log^2 n}{\theta^4}$

# Prior work

Table: Comparison with existing methods for solving MSBD

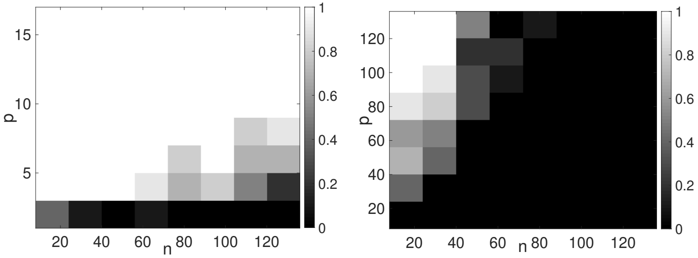
| Methods | [Wang and Chi, 2016] | [Li and Bresler, 2019] | **Ours** |
|---|---|---|---|
| Assumptions | **filter $\boldsymbol{g}$ spiky $\&$ $\mathcal{C}(\boldsymbol{g})$ invertible**, $\boldsymbol{X} \sim \mathrm{BG}(\theta)$ | $\mathcal{C}(\boldsymbol{g})$ invertible, $\boldsymbol{X} \sim \mathrm{BR}(\theta)$ | $\mathcal{C}(\boldsymbol{g})$ invertible, $\boldsymbol{X} \sim \mathrm{BG}(\theta)$ |
| Formulation | Convex $\min_{\boldsymbol{e}_1^\top \boldsymbol{h}=1} \|\mathcal{C}(\boldsymbol{h})\boldsymbol{Y}\|_1$ | Nonconvex $\max_{\|\boldsymbol{h}\|_2=1} \|\mathcal{C}(\boldsymbol{h})\boldsymbol{R}\boldsymbol{h}\|_4^4$ | Nonconvex $\min_{\|\boldsymbol{h}\|_2=1} \psi_\mu(\mathcal{C}(\boldsymbol{h})\boldsymbol{R}\boldsymbol{h})$ |
| Algorithm | linear programming | *noisy* MGD | *vanilla* MGD |
| Recovery Condition | $\boldsymbol{\theta \in O(1/\sqrt{n})}$, $p \geq O(n)$ | $\theta \in O(1)$, $\boldsymbol{p \geq O(n^9)}$ | $\boldsymbol{\theta \in O(1)}$, $\boldsymbol{p \geq O(n^{4.5})}$ |

- For order of $p$, assuming $\theta, \kappa$ are constants, the order of sample complexity $p$ is shown up to logarithmic factors.

## Practical Experiment Results

# Numerical experiments: synthetic data

- Success rate of recovering the filter $\boldsymbol{g}$:
  - 10 Monte Carlo for success rate $\in [0, 1]$.
  - Fix sparsity $\theta = 0.3$.

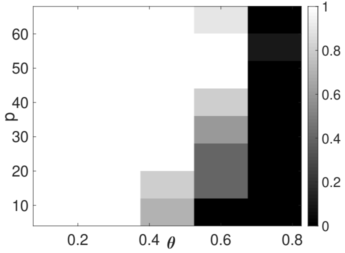



(a) Ours (b) [Li and Bresler, 2019]

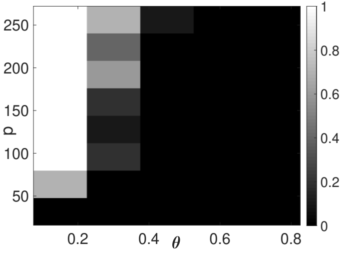
Figure: Requirement of sample complexity $p$ with respect to $n$.

# Numerical experiments: synthetic data

- Success rate of recovering the filter $\boldsymbol{g}$:
    - 10 Monte Carlo for success rate $\in [0, 1]$.
    - Fix $n = 64$.




(a) Ours

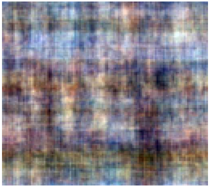



(b) [Li and Bresler, 2019]

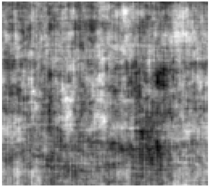
Figure: Requirement of sample complexity $p$ with respect to $\theta$.

# Numerical experiments: blind image deconvolution

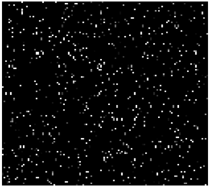
- Experimental setting:
  - The filter size is $n = 128 \times 128$.
  - The number of observations is $p = 1000$.
  - Sparsity level $\theta = 0.1$: $\boldsymbol{X} \in \mathrm{BG}(\theta)$



(a) Observation (RGB)



(b) Observation (R)



(c) Sparse input

# Numerical experiments: blind image deconvolution

Comparisons of the recovered filter $\boldsymbol{g}$:

(d) True image

(e) Recovery via ours

(f) Recovery via [Li, et al., 2019]

# Conclusion

- We propose a novel nonconvex approach for MSBD problem based on MGD with random initializations.

# Conclusion

- We propose a novel nonconvex approach for MSBD problem based on MGD with random initializations.
- Under mild statistical model for sparse inputs, we provide theoretical characterizations for benign geometric landscape of the loss function $\rightarrow$ ensures the global convergence of MGD.

# Conclusion

- We propose a novel nonconvex approach for MSBD problem based on MGD with random initializations.
- Under mild statistical model for sparse inputs, we provide theoretical characterizations for benign geometric landscape of the loss function $\rightarrow$ ensures the global convergence of MGD.
- Comparisons with prior work:
    1. significant improvement of sample complexity $p$: from $p \gtrsim O(n^9) \rightarrow p \gtrsim O(n^{4.5})$.
    2. better practical performance in a much larger range of the sparsity level.

# References


- Yanjun Li and Yoram Bresler. "Multichannel sparse blind deconvolution on the sphere". In: *IEEE Transactions on Information Theory* 65.11 (2019), pp. 7415–7436.
- Yanjun Li, Kiryung Lee, and Yoram Bresler. "A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models". In: *arXiv preprint arXiv:1501.06120* (2015).
- Qing Qu et al. "Analysis of the Optimization Landscapes for Overcomplete Representation Learning". In: *arXiv preprint arXiv:1912.02427* (2019).
- Ju Sun, Qing Qu, and John Wright. "Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture". In: *IEEE Transactions on Information Theory* 63.2 (2017), pp. 853–884.

Thank you!

Email: {laixishi, yuejiechi}@cmu.edu

Paper link: `https://arxiv.org/pdf/1911.11167.pdf`