Robust Recovery via Implicit Bias of Discrepant Learning Rates for Double Over-parameterization

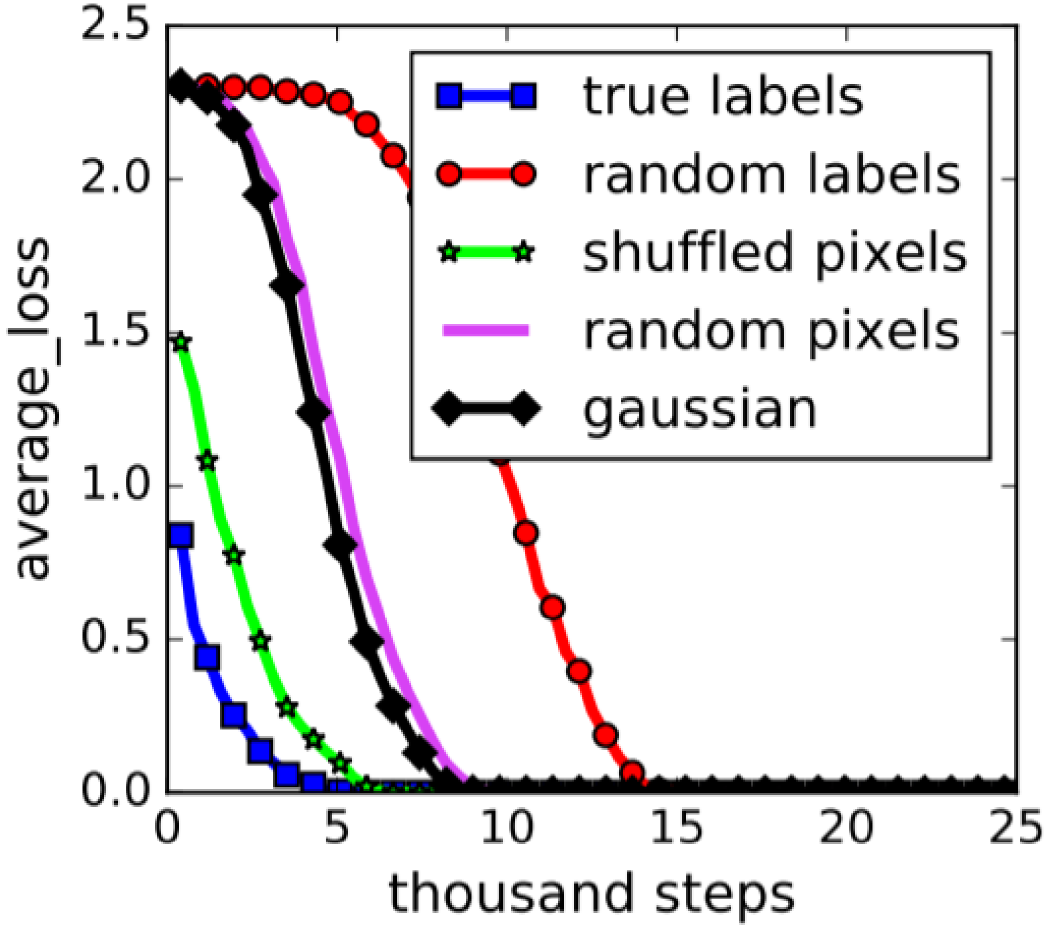
Chong You[†], Zhihui Zhu[‡], Qing Qu[#], Yi Ma[†]

[†]Department of EECS, University of California, Berkeley

[‡]Department of Electrical and Computer Engineering, University of Denver

[#]Center for Data Science, New York University

Over-parameterization and Overfitting



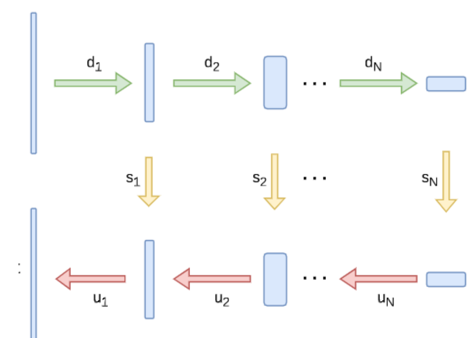
(Zhang et. al., ICLR'17)

- A classification network can overfit to label corruption

Deep Image Prior (DIP) for Image Recovery

$$\min_{\boldsymbol{\theta}} \| \underbrace{\boldsymbol{y}}_{\text{corrupted input}} - \underbrace{f(\boldsymbol{\theta})}_{\text{recovered image}} \|_1$$

- **Idea**: CNN architecture encodes priors for clean images



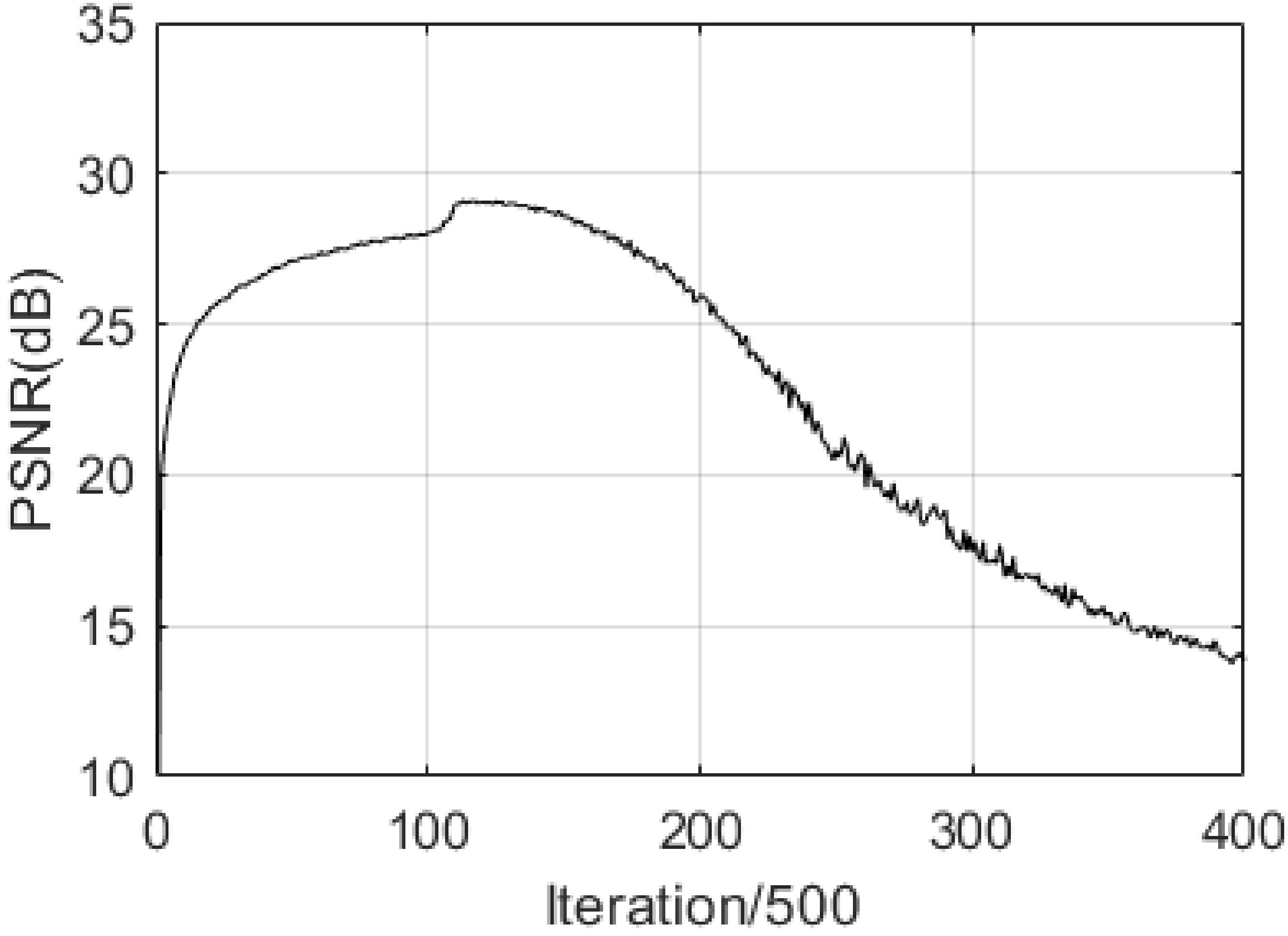
Ulyanov, Vedaldi, Lempitsky, Deep Image Prior, *CVPR 2018*

Over-Parameterization

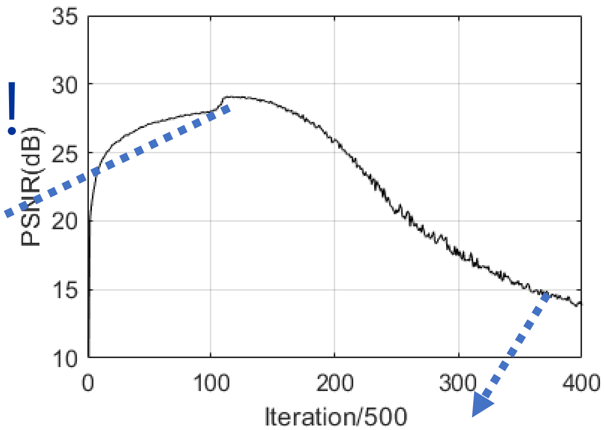
~ 2 million (#parameters in model) $\gg$ ~ 0.1 million (DOF in an image)

In principle, $f(\boldsymbol{\theta})$ can generate any image!
(i.e., not only clean, but also corrupted images)

Over-Parameterization -> Overfitting!



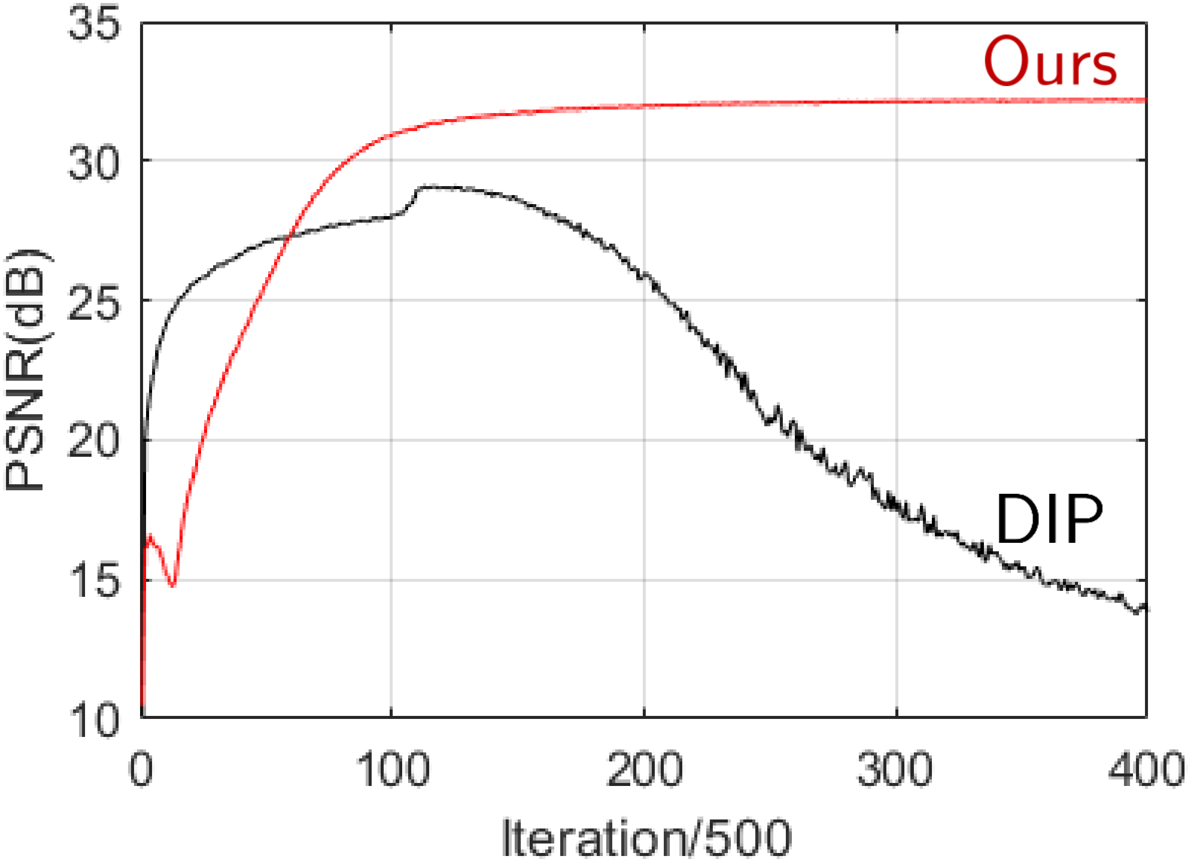
Over-Parameterization -> Overfitting!



Early termination solution
(impractical!)

Global solution: $f(\boldsymbol{\theta}) \approx \boldsymbol{y}$
(overfitting!)

This Work: Over-Parameterization *Without* Overfitting!



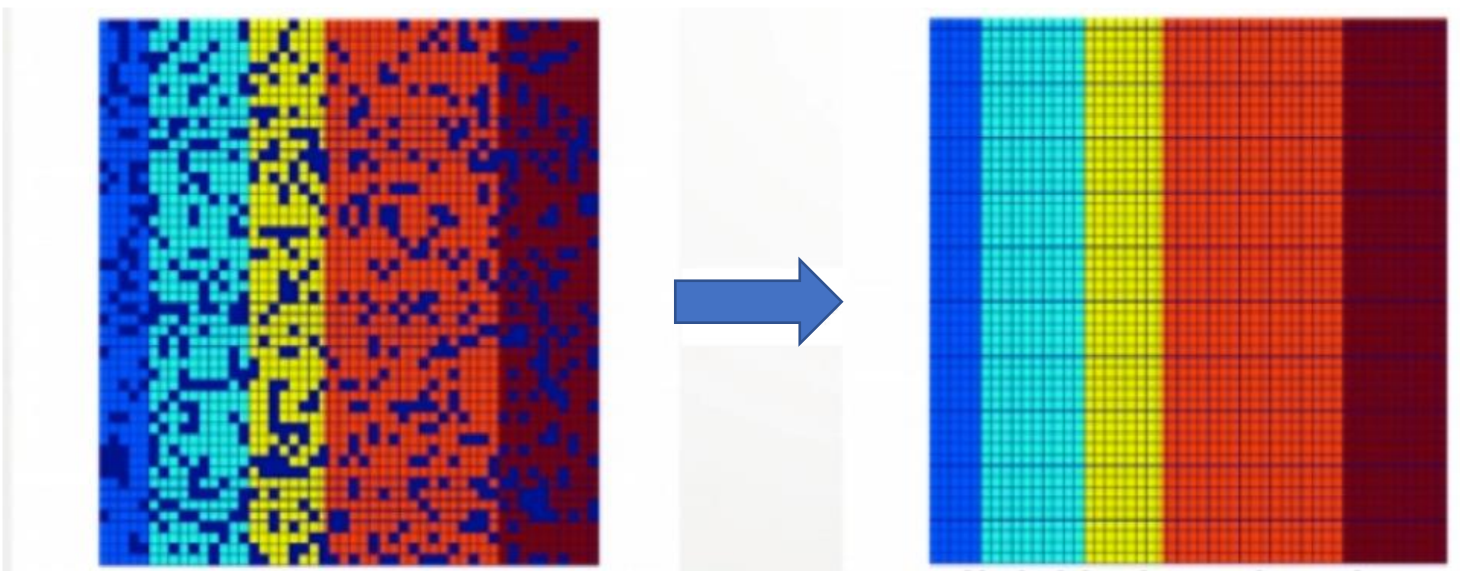
Robust recovery of natural images

Simplification ↓ ↑ Implication

Robust recovery of low-rank matrices

Robust Recovery of Low-Rank Matrices

- **Goal**: Recover a rank-r matrix $\mathbf{X}_\star \in \mathbb{R}^{n \times n}$ from (possibly corrupted) linear measurements $\mathbf{y} = \mathcal{A}(\mathbf{X}_\star) + \mathbf{s}_\star$



Matrix Factorization – *Exact* Parameterization

	(Classical) Exact-Parameterization
$\mathbf{s}_\star = 0$ (noiseless)	$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_2^2$ – Gradient descent finds $\mathbf{X}_\star$
	⬇ (Classical) robust loss
sparse $\mathbf{s}_\star$	$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_1$ – Sub-grad method finds $\mathbf{X}_\star$ Require knowing $r = \text{rank}(\boldsymbol{X}_\star)$!

- Li, Zhu, So, Vidal, Nonconvex Robust Low-rank Matrix Recovery, SIAM Journal on Optimization 2019
- Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, Implicit Regularization in Matrix Factorization, NeurIPS 2017

Matrix Factorization – *Over* Parameterization

	(Classical) Exact-Parameterization	(Modern) Over-Parameterization
$\mathbf{s}_\star = 0$ (noiseless)	$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_2^2$ – Gradient descent finds $\mathbf{X}_\star$	$\min_{\mathbf{U}\in\mathbb{R}^{n\times n}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_2^2$ – Gradient descent finds $\mathbf{X}_\star$
	(Classical) robust loss	
sparse $\mathbf{s}_\star$	$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_1$ – Sub-grad method finds $\mathbf{X}_\star$ Require knowing $r = \text{rank}(\boldsymbol{X}_\star)$!	?

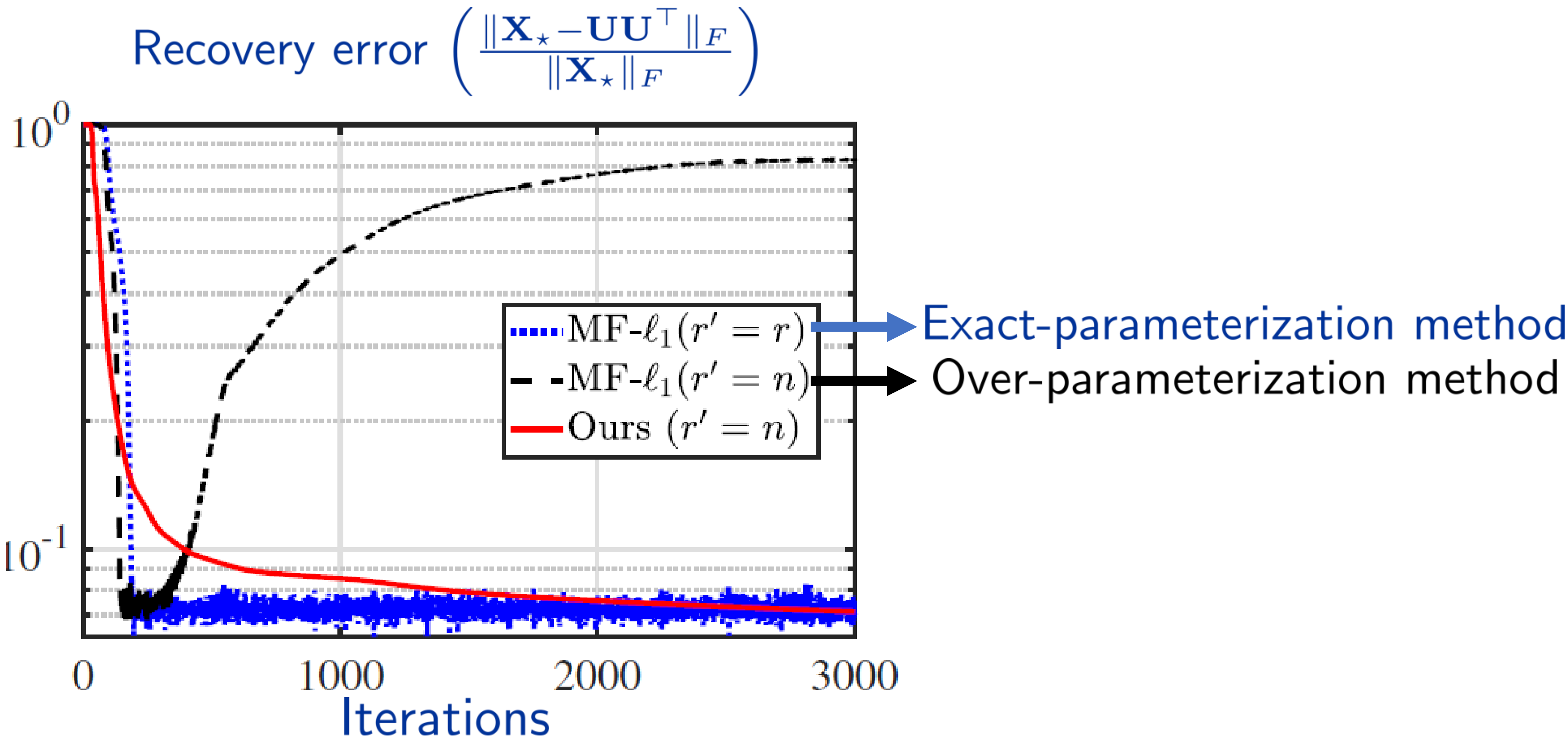
- Li, Zhu, So, Vidal, Nonconvex Robust Low-rank Matrix Recovery, SIAM Journal on Optimization 2019
- Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, Implicit Regularization in Matrix Factorization, NeurIPS 2017

Matrix Factorization – *Over* Parameterization

	(Classical) Exact-Parameterization		(Modern) Over-Parameterization
$\mathbf{s}_\star = 0$ (noiseless)	$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_2^2$ – Gradient descent finds $\mathbf{X}_\star$	→	$\min_{\mathbf{U}\in\mathbb{R}^{n\times n}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_2^2$ – Gradient descent finds $\mathbf{X}_\star$
	↓ (Classical) robust loss		↓ Use classical approach?
sparse $\mathbf{s}_\star$	$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_1$ – Sub-grad method finds $\mathbf{X}_\star$ Require knowing $r = \text{rank}(\boldsymbol{X}_\star)$!	→	$\min_{\mathbf{U}\in\mathbb{R}^{n\times n}} \|\mathbf{y} - \mathcal{A}(\mathbf{U}\mathbf{U}^\top)\|_1$ – **Does this work?**

- Li, Zhu, So, Vidal, Nonconvex Robust Low-rank Matrix Recovery, SIAM Journal on Optimization 2019
- Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, Implicit Regularization in Matrix Factorization, NeurIPS 2017

Failure of Classical Approach for Over-Param. Models



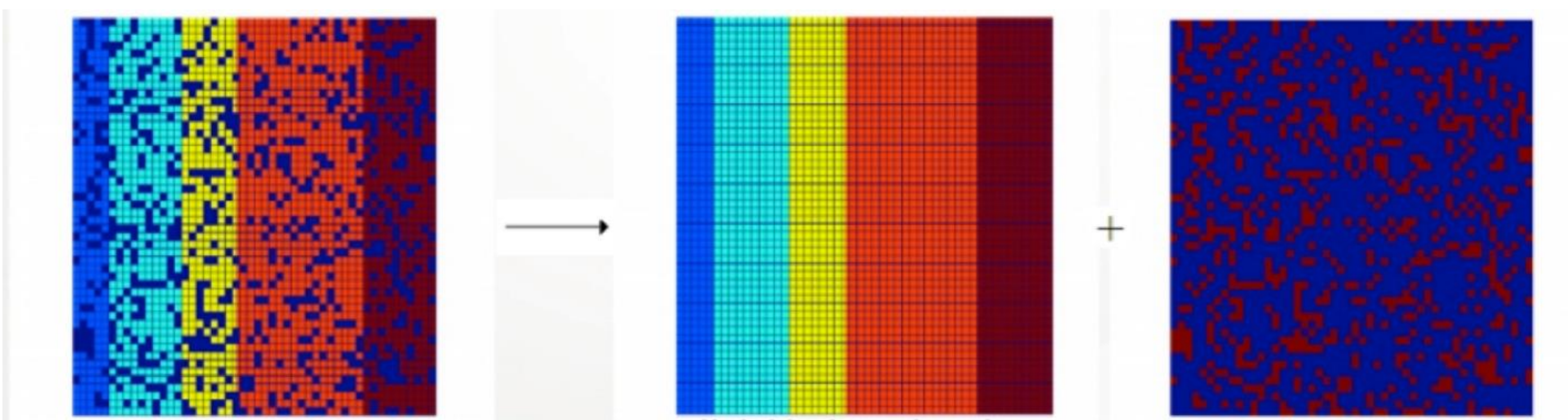
Key Challenge: How to robustify over-parameterized models?

A Double Over-Parameterization Method

- **Our Strategy:** Over-parameterize the noise $\boldsymbol{s}$

Hadamard product

$$\min_{\boldsymbol{U}\in\mathbb{R}^{n\times n},\boldsymbol{g},\boldsymbol{h}} \ell(\boldsymbol{U},\boldsymbol{g},\boldsymbol{h}) := \| \underbrace{\boldsymbol{y}} - \Big(\mathcal{A}\big(\underbrace{\boldsymbol{U}\boldsymbol{U}^\top}_{\text{over-parameterizes } \boldsymbol{X}}\big) + \underbrace{\boldsymbol{g}\odot\boldsymbol{g} - \boldsymbol{h}\odot\boldsymbol{h}}_{\text{over-parameterizes } \boldsymbol{s}} \Big)\|_2^2$$



- (Seemingly) Problematic: Even more parameters, infinitely many global solutions

Implicit Algorithmic Bias

- A gradient descent with discrepant learning rates algorithm:

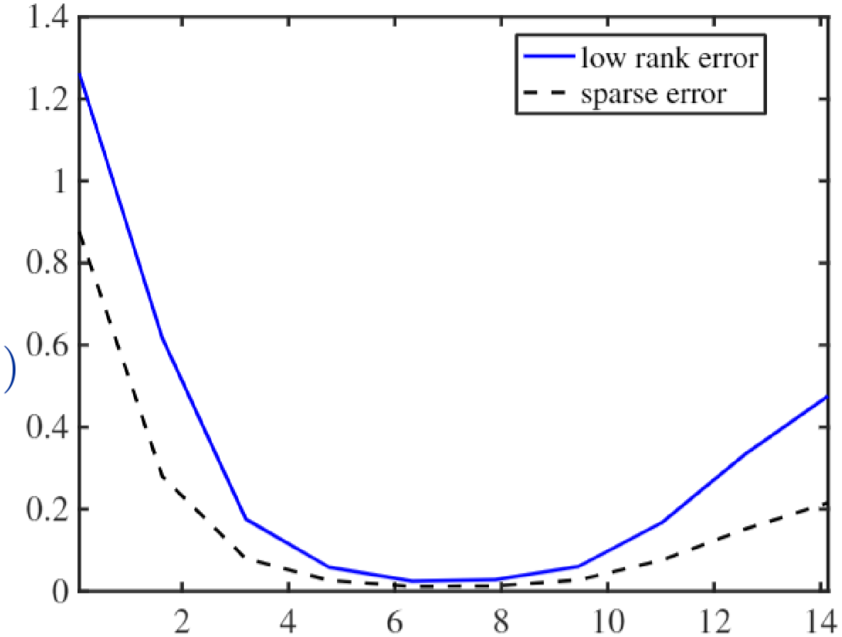
$$\boldsymbol{U}_{k+1} = \boldsymbol{U}_k - \tau \cdot \nabla \ell(\boldsymbol{U}_k, \boldsymbol{g}_k, \boldsymbol{h}_k)$$

$$\begin{bmatrix} \boldsymbol{g}_{k+1} \\ \boldsymbol{h}_{k+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}_k \\ \boldsymbol{h}_k \end{bmatrix} - \alpha \cdot \tau \cdot \begin{bmatrix} \nabla_{\boldsymbol{g}} \ell(\boldsymbol{U}_k, \boldsymbol{g}_k, \boldsymbol{h}_k) \\ \nabla_{\boldsymbol{h}} \ell(\boldsymbol{U}_k, \boldsymbol{g}_k, \boldsymbol{h}_k) \end{bmatrix}$$

- Implicit bias of learning rate ratio α

Low rank error $\left(\frac{\|\boldsymbol{X}_\star - \boldsymbol{U}\boldsymbol{U}^\top\|_F}{\|\boldsymbol{X}_\star\|_F}\right)$

Sparse error $\left(\frac{\|\boldsymbol{s}_\star - (\boldsymbol{g} \odot \boldsymbol{g} - \boldsymbol{h} \odot \boldsymbol{h})\|_F}{\|\boldsymbol{s}_\star\|_F}\right)$



Learning rate ratio α

Implicit Algorithmic Bias: Main Theory

- **Q1:** How does α (learning rate ratio) control the solution?

Theorem: [You*, Zhu*, Qu, Ma'2020]

- $\boldsymbol{U}_0, \boldsymbol{g}_0, \boldsymbol{h}_0$ are infinitesimally small, τ is infinitesimally small
- $\mathcal{A}$ is symmetric and commutable (i.e., $A_i A_j = A_j A_i$ for each $i \neq j$)
- $\boldsymbol{X}_\infty = \lim_{k\to\infty} \boldsymbol{U}_k \boldsymbol{U}_k^\top$ and $\boldsymbol{s}_\infty = \lim_{k\to\infty} (\boldsymbol{g}_k \odot \boldsymbol{g}_k - \boldsymbol{h}_k \odot \boldsymbol{h}_k)$ exist and produce a global solution (i.e., $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\infty) + \boldsymbol{s}_\infty$),

then $(\boldsymbol{X}_\infty, \boldsymbol{s}_\infty)$ is a solution to the following problem

$$\min_{\boldsymbol{X},\, \boldsymbol{s}} \|\boldsymbol{X}\|_* + \frac{1}{\alpha}\|\boldsymbol{s}\|_1 \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{s}$$

- **Q2:** What value of α should I use?

  Theorem [Candes et al. '09] Using $\alpha = \sqrt{n}$, the solution to the convex optimization above is $(\boldsymbol{X}_\star, \boldsymbol{s}_\star)$ under certain conditions

  – No parameter tuning is required

Extension to Image Recovery

- **Goal:** Recover an image $\mathbf{X}_\star \in \mathbb{R}^{H \times W \times 3}$ from $\mathbf{y} = \mathbf{X}_\star + \mathbf{s}_\star$, where $\mathbf{s}_\star$ is salt-and-pepper (sparse) noise

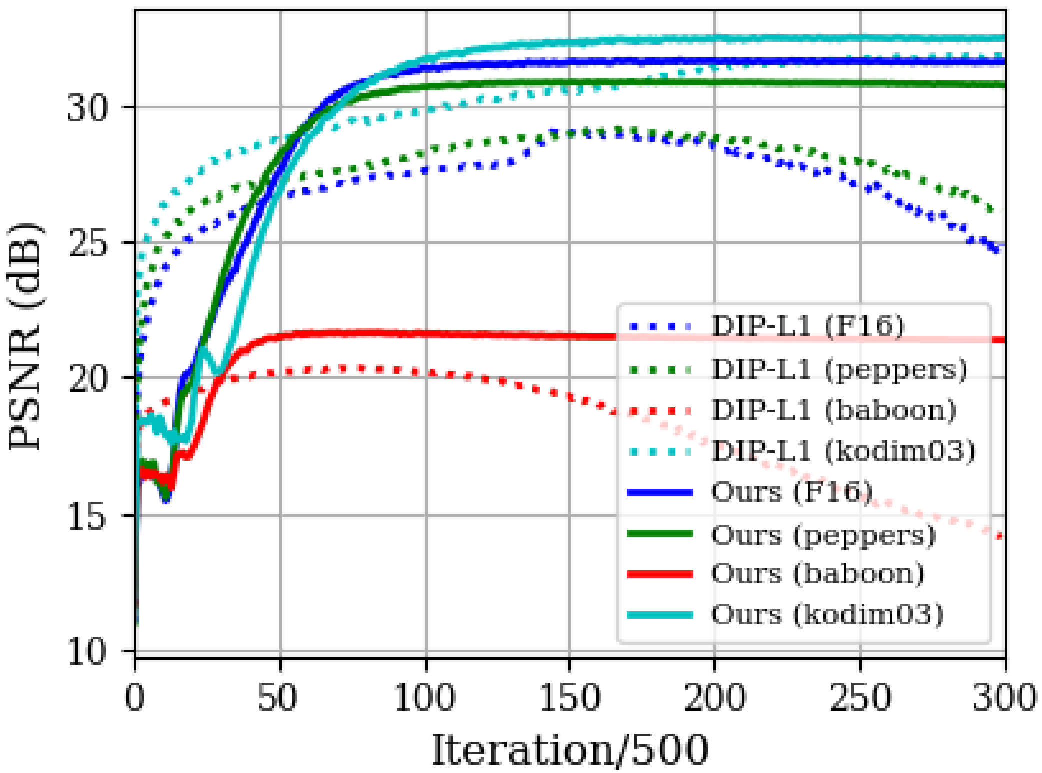
- **Method:**

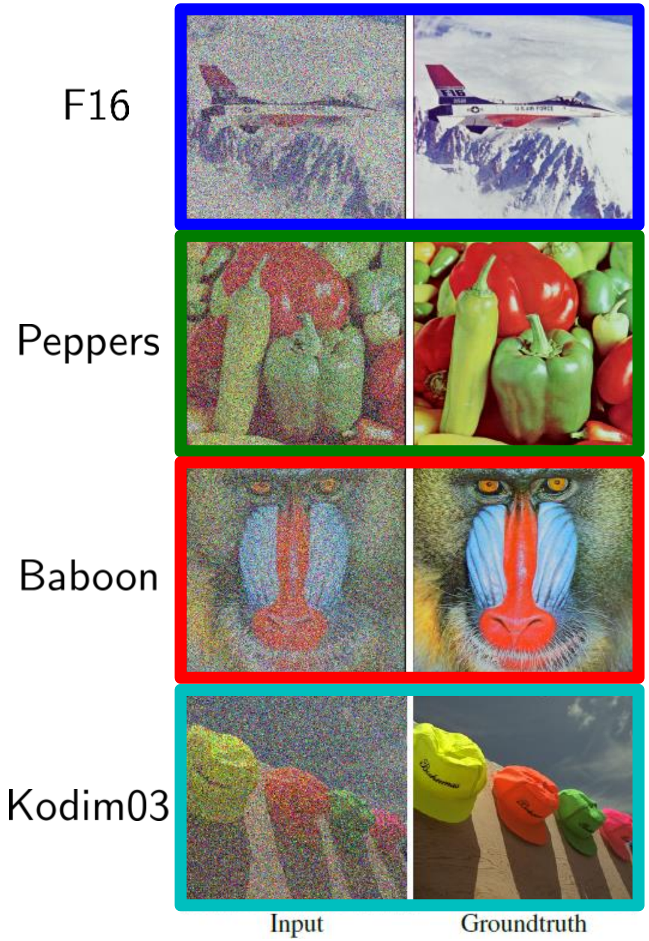
$$\min_{\boldsymbol{\theta}, \boldsymbol{g}, \boldsymbol{h}} \| \underbrace{\mathbf{y}} - (\underbrace{f(\boldsymbol{\theta})} + \underbrace{\boldsymbol{g} \odot \boldsymbol{g} - \boldsymbol{h} \odot \boldsymbol{h}}) \|_2^2$$

  - Solve this problem using gradient descent with learning rate ratio α

- No parameter tuning is required:
  - No theory for optimal α, but $\alpha = 500$ works well for all cases (i.e., with different images/corruption levels/network widths)

Results for Varying Images





No early stop, no parameter tuning!

Take-Home: Over-parameterization without Overfitting!

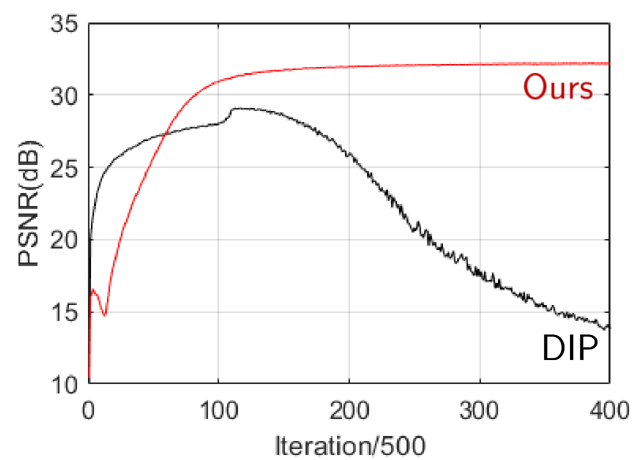
- Classical robust loss function approaches will overfit for over-parameterized models

- This work proposes a

  double over-parameterization model

  which does not overfit by exploiting

  implicit bias of discrepant learning rates



Thank you for your attention!

Acknowledgement:

NSF

MSDSE
Moore - Sloan
Data Science Environments

Office of Naval Research
ONR
Science & Technology

TBSI
Tsinghua-Berkeley Shenzhen Institute

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

FHL VIVE CENTER
FOR ENHANCED REALITY