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

Chong You† Zhihui Zhu‡ Qing Qu* Yi Ma†

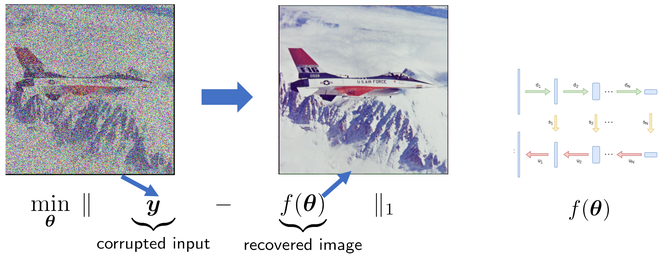
†University of California, Berkeley ‡Denver University *New York University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Deep neural networks are highly over-parameterized

- Background: Image recovery via *deep image prior* (**DIP**) [1]



Idea: network architecture (i.e., $f(\boldsymbol{\theta})$) encodes priors for clean images

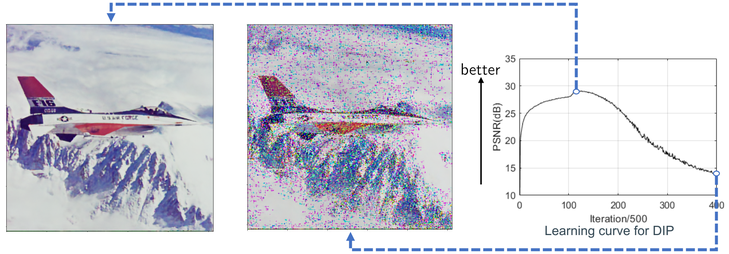
- Over-parameterization in DIP:

~ 2 million (#parameters in $f(\boldsymbol{\theta})$) $\gg$ ~ 0.1 million (# pixels in an image)

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

Challenge

Over-parameterization *causes* overfitting



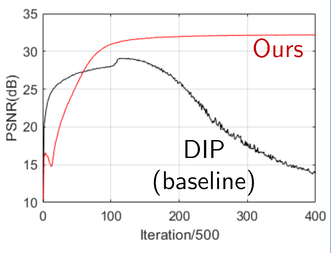
Early termination solution **(impractical!)** Global solution: $f(\boldsymbol{\theta}) \approx \boldsymbol{y}$ **(overfitting!)**

Contribution

Over-parameterization **WITHOUT** overfitting!

- Model: Double over-parameterization of both the image and the corruption
- Algorithm: Discrepant learning rates for different model parameters
- Theory: Correctness for low-rank recovery

No need to terminate early
No need to tune network width
No need to finetune parameters



Model: Double Over-Parameterization (DOP)

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 sparse noise

$$\ell(\boldsymbol{\theta}, \boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{\theta},\boldsymbol{g},\boldsymbol{h}} \| \underbrace{\mathbf{y}}_{} - (\underbrace{f(\boldsymbol{\theta})}_{} + \underbrace{\boldsymbol{g}\odot\boldsymbol{g} - \boldsymbol{h}\odot\boldsymbol{h}}_{})\|_2^2 \qquad (1)$$

- $f(\boldsymbol{\theta})$: U-shaped neural network (as in DIP) that *over-parameterizes* the image $\boldsymbol{X}$
- $\boldsymbol{g}\odot\boldsymbol{g} - \boldsymbol{h}\odot\boldsymbol{h}$ ($\odot$ - Hadamard product) *over-parameterizes* sparse corruption $\boldsymbol{s}$

(Seemingly) Problematic:

Even more parameters than DIP,
even more global solutions than DIP.
Hence, *even more* prone to overfitting than DIP?

Algorithm: Discrepant Learning Rates

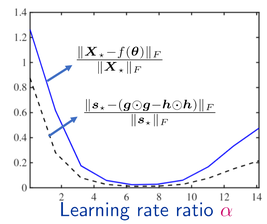
(Genuinely) a Blessing: Proper choice of algorithm leads to the desired solutions!

- Observation: Learning rate ratio α for $\boldsymbol{\theta}$ and $\{\boldsymbol{g}, \boldsymbol{h}\}$ in a gradient descent

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \tau\cdot\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_k, \boldsymbol{g}_k, \boldsymbol{h}_k)$$
$$\begin{bmatrix}\boldsymbol{g}_{k+1}\\ \boldsymbol{h}_{k+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{g}_k\\ \boldsymbol{h}_k\end{bmatrix} - \alpha\cdot\tau\cdot\begin{bmatrix}\nabla_{\boldsymbol{g}}\ell(\boldsymbol{\theta}_k, \boldsymbol{g}_k, \boldsymbol{h}_k)\\ \nabla_{\boldsymbol{h}}\ell(\boldsymbol{\theta}_k, \boldsymbol{g}_k, \boldsymbol{h}_k)\end{bmatrix} \qquad (2)$$

controls the quality of the solution

- No parameter tuning: Empirically, best α does NOT depend on a) test image $\boldsymbol{X}_\star$, b) sparsity of $\boldsymbol{s}_\star$, and c) width of network $f(\boldsymbol{\theta})$



Theory: Insights from Low-rank Modeling

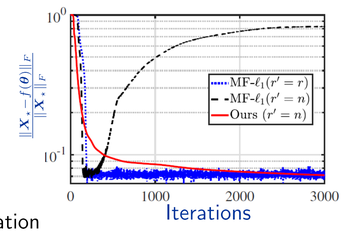
Consider the case $\boldsymbol{X}_\star \in \mathbb{R}^{n\times n}$ with rank $= r \ll n$. Let $f(\boldsymbol{\theta}) = \boldsymbol{U}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r'}$

- Classical method based on ℓ_1 loss:

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times r'}} \|\mathbf{y} - \mathbf{U}\mathbf{U}^\top\|_1$$

  – works for *exact*-param. models (i.e., $r' = r$)
  – fails for *over*-param. models (e.g., $r' = n$)

Failure of *classical* robust loss methods for *modern* over-parameterized models!



- Our method based on double over-parameterization

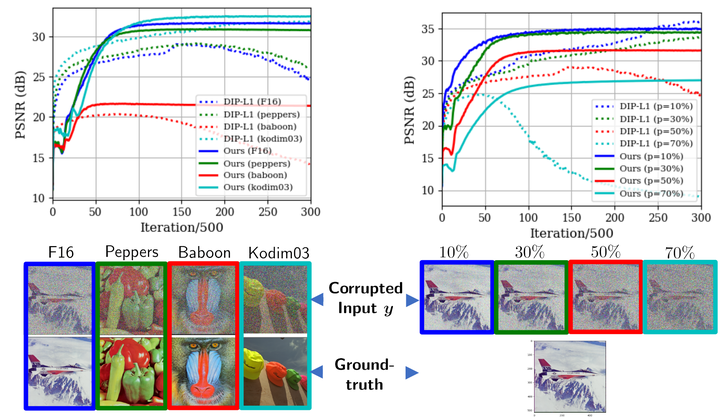
Theorem: Algorithm (2) for solving problem (1) with $r' = n$ converges to a solution to $\min_{\boldsymbol{X},\boldsymbol{s}} \|\boldsymbol{X}\|_* + 1/\alpha\cdot\|\boldsymbol{s}\|_1$ s.t. $\boldsymbol{y} = \boldsymbol{X} + \boldsymbol{s}$

Combined with [3], setting $\alpha = \sqrt{n}$ produces desired solutions (No tuning parameters!)

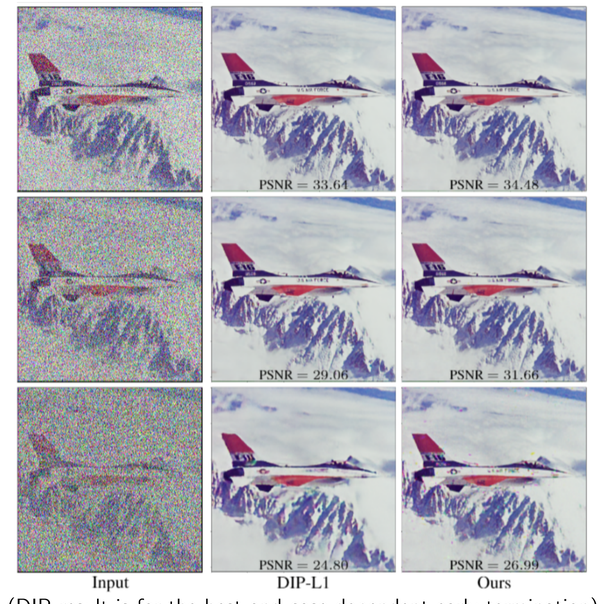
Experiments

Learning curves for varying images (left) and varying corruption levels (right)

- For DIP, best termination varies for different images and corruption levels
- For Ours, no need to terminate and no need to tune any parameters



Visualization of results with varying corruption levels



(DIP result is for the best and case-dependent early termination)

References

[1] Ulyanov, Vedaldi, Lempitsky (2018). "Deep Image Prior." In: CVPR
[2] Li, Zhu, So, Vidal (2019). "Robust Low-rank Matrix Recovery." In: SIAM Journal on Optimization
[3] Candes, Li, Ma, Wright (2011). "Robust Principal Component Analysis?" In: Journal of ACM

This work is supported by Tsinghua-Berkeley Shenzhen Institute Research Fund, ONR grant N00014-20-1-2002, Join Simons Foundation-NSF grant #2031899, NSF grant #200846, NSF DMS #2009752, Moore-Sloan Fellowship, Berkeley AI Research, Berkeley FHL Vive Center for Enhanced Reality, and Berkeley Center for Augmented Cognition