Multilook Coherent Imaging: Theoretical Guarantees and Algorithms

Xi Chen†, Soham Jana†, Christopher A. Metzler, Arian Maleki, Shirin Jalali

Abstract

Multilook coherent imaging is a widely used technique in applications such as digital holography, ultrasound imaging, and synthetic aperture radar. A central challenge in these systems is the presence of multiplicative noise, commonly known as speckle, which degrades image quality. Despite the widespread use of coherent imaging systems, their theoretical foundations remain relatively underexplored. In this paper, we study both the theoretical and algorithmic aspects of likelihood-based approaches for multilook coherent imaging, providing a rigorous framework for analysis and method development. Our theoretical contributions include establishing the first theoretical upper bound on the Mean Squared Error (MSE) of the maximum likelihood estimator under the deep image prior hypothesis. Our results capture the dependence of MSE on the number of parameters in the deep image prior, the number of looks, the signal dimension, and the number of measurements per look. On the algorithmic side, we employ projected gradient descent (PGD) as an efficient method for computing the maximum likelihood solution.

Furthermore, we introduce two key ideas to enhance the practical performance of PGD. First, we incorporate the Newton-Schulz algorithm to compute matrix inverses within the PGD iterations, significantly reducing computational complexity. Second, we develop a bagging strategy to mitigate projection errors introduced during PGD updates. We demonstrate that combining these techniques with PGD yields state-of-the-art performance. Our code is available at: https://github.com/Computational-Imaging-RU/Bagged-DIP-Speckle.

Keywords: Inverse Problems, Speckle Noise, Deep Image Prior

I. Introduction

One of the most fundamental and challenging issues faced by many coherent imaging systems is the presence of speckle noise. An imaging system with "fully-developed" speckle noise can be modeled as

$$\boldsymbol{y} = AX_o\boldsymbol{w} + \boldsymbol{z}. \tag{1}$$

Here, $X_o = \mathrm{diag}(\boldsymbol{x}_o)$, where $\boldsymbol{x}_o \in \mathbb{C}^n$ denotes the complex-valued signal of interest. $\boldsymbol{w} \in \mathbb{C}^n$ represents speckle (or multiplicative) noise, where $w_1, \ldots, w_n$ are independent and identically distributed (iid) $\mathcal{CN}(\mathbf{0}, \sigma_w^2 I_n)$, and finally $\boldsymbol{z} \in \mathcal{C}^m$ denotes the additive noise, often caused by the sensors, is modeled as iid $\mathcal{CN}(\mathbf{0}, \sigma_z^2)$. In this paper, we explore the scenario where $m \leq n$, allowing imaging systems to capture higher resolution images than constrained by the number of sensors. Considering $m < n$ for simpler imaging systems (with no speckle noise) has led to the development of the fields of compressed sensing and compressive phase retrieval [1]–[6].

As is clear from (1), the multiplicative nature of the speckle noise poses a challenge in extracting accurate information from measurements, especially when the measurement matrix A is ill-conditioned. To alleviate this issue, many practical systems employ a technique known as multilook or multishot [7], [8]. Instead of taking a single measurement of the image, multilook systems capture multiple measurements, aiming for each group of measurements to have independent speckle and additive noise. In an L look system, the measurements captured at look ℓ, $\ell = 1, \ldots, L$, can be represented as

$$\boldsymbol{y}_\ell = AX_o\boldsymbol{w}_\ell + \boldsymbol{z}_\ell,$$

†Equal contributions. X. C. and S. Jalali are with the Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ, USA. S. Jana is with the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, USA, (Correspondence to: soham.jana@nd.edu). C. A. M. is with the Department of Computer Science, University of Maryland, College Park, MD, USA. A. M. is with the Department of Statistics, Columbia University, NY, USA. An earlier version of this paper was presented in part at the Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

where, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_L \in \mathbb{C}^n$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \in \mathbb{C}^m$ denote the independent speckle noise and additive noise vectors, respectively. In this model, we have assumed that the measurement kernel A remains constant across the looks. This assumption holds true in multilooking for several imaging systems, such as when the sensors' locations change slightly for different looks.

Since fully-developed noises are complex-valued Gaussian and have uniform phases, the phase of $\boldsymbol{x}_o$ cannot be recovered. Hence, the goal of a multilook system is to obtain a precise estimate of $|\boldsymbol{x}_o|$ based on the L observations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L\}$, given the measurement matrix A. (Here, $|\cdot|$ denotes the element-wise absolute value operation.) Therefore, since the phase of $\boldsymbol{x}_o$ is not recoverable, in the rest of the paper, we assume that $\boldsymbol{x}_o$ is real-valued.

A standard approach for estimating $\boldsymbol{x}_o$ is to minimize the negative log-likelihood function subject to the signal structure constraint. More precisely, in a constrained-likelihood-based approach, one aims to solve the following optimization problem:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \in \mathcal{C}} f_L(\boldsymbol{x}), \tag{2}$$

where $\mathcal{C}$ represents the set encompassing all conceivable images and $f_L(\boldsymbol{x})$ is defined as:

$$f_L(\boldsymbol{x}) = \log\det(B(\boldsymbol{x})) + \frac{1}{L}\sum_{\ell=1}^{L} \widetilde{\boldsymbol{y}}_\ell^\top (B(\boldsymbol{x}))^{-1} \widetilde{\boldsymbol{y}}_\ell, \tag{3}$$

where

$$B(\boldsymbol{x}) = \begin{bmatrix} \sigma_z^2 I_n + \sigma_w^2 \Re(U(\boldsymbol{x})) & -\sigma_w^2 \Im(U(\boldsymbol{x})) \\ \sigma_w^2 \Im(U(\boldsymbol{x})) & \sigma_z^2 I_n + \sigma_w^2 \Re(U(\boldsymbol{x})) \end{bmatrix},$$

and $\widetilde{\boldsymbol{y}}_\ell^\top = \begin{bmatrix} \Re(\boldsymbol{y}_\ell^\top) & \Im(\boldsymbol{y}_\ell^\top) \end{bmatrix}$, with $X = \mathrm{diag}(\boldsymbol{x})$ and

$$U(\boldsymbol{x}) = AX^2\bar{A}^\top.$$

Here, $\Re(\cdot)$ and $\Im(\cdot)$ denote element-wise real and imaginary parts, respectively. (Appendix A presents the derivation of the log likelihood function and its gradient.)

It is important to note that the set $\mathcal{C}$ in (2) is not known explicitly in practice. Hence, in this paper we work with the following hypothesis that was put forward in [9], [10].

- **Deep image prior (DIP) hypothesis** [9], [10]: Natural images can be embedded within the range of untrained neural networks that have substantially fewer parameters than the total number of pixels, and use iid noises as inputs.

Inspired by this hypothesis, we define $\mathcal{C}$ as the range of a deep image prior. More specifically, we assume that for every $\boldsymbol{x} \in \mathcal{C}$, there exists $\boldsymbol{\theta} \in \mathbb{R}^k$ such that $\boldsymbol{x} = g_{\boldsymbol{\theta}}(\boldsymbol{u})$, where $\boldsymbol{u}$ is generated iid $\mathcal{N}(0,1)$, and $\boldsymbol{\theta} \in \mathbb{R}^k$ denotes the parameters of the DIP neural network. There are two main challenges that we address:

- Theoretical challenge: Assuming that we can solve the optimization problem (2) under the DIP hypothesis, the following question arises: Can we theoretically characterize the corresponding reconstruction quality? Moreover, what is the relationship between the reconstruction error and key parameters such as k (the number of parameters of the DIP neural network), m, n and L? Specifically, in the scenario where the scene is static, and we can acquire as many looks as necessary, what is the achievable level of accuracy?
- Practical challenge: Given the challenging nature of the likelihood and the DIP hypothesis, can we design a computationally-efficient algorithm for solving (2) under the DIP hypothesis?

Here is a summary of our contributions:

On the theoretical front, we establish the first theoretical result on the performance of multilook coherent imaging systems. Compared to the earlier version of this work presented at the International Conference on Machine Learning, the theoretical results in the current paper are significantly sharper. Our findings unveil intriguing characteristics of such imaging systems. A special case of our result, corresponding to

$L = 1$, is directly comparable to the findings in [11]. As we will show, in this setting, our bounds on the mean squared error are significantly tighter than those presented in [11].

On the practical side, we start with vanilla projected gradient descent (PGD) [12], which faces two challenges diminishing its effectiveness on this problem:

Challenge 1: As will be described in Section IV-B, in the PGD, the signal to be projected on the range of $g_{\boldsymbol{\theta}}(\boldsymbol{u})$ is burried in "noise". Hence, DIPs with large number of parameters will overfit to the noise and will not allow the PGD algorithm to obtain a reliable estimate [10], [13]. On the other hand, the low accuracy of simpler DIPs becomes a bottleneck as the algorithm progresses through iterations, limiting the overall performance. To alleviate this issue, we propose **Bagged-DIP**. This is a simple idea with roots in classical literature of ensemble methods [14]. Bagged-DIP idea enables us to use complex DIPs at every iteration and yet obtain accurate results.

Challenge 2: As will be clarified in Section IV-A, PGD requires the inversion of large matrices at every iteration, which is a computationally challenging problem. We alleviate this issue by using the Newton-Schulz algorithm [15], and empirically demonstrating that **only one** step of this algorithm is sufficient for the PGD algorithm. This significantly reduces the computational complexity of each iteration of PGD.

II. Related Work

Eliminating speckle noise has been extensively explored in the literature [16]–[18]. Current technology relies on gathering enough measurements to ensure the invertibility of matrix A and subsequently inverting A to represent the measurements in the following form: $\boldsymbol{y}_\ell = X\boldsymbol{w}_\ell + \boldsymbol{z}_\ell$. However, as matrix A deviates from the identity, the elements of the vector $\boldsymbol{z}$ become dependent. In practice, these dependencies are often overlooked, simplifying the likelihood. This simplification allows researchers to leverage various denoising methods, spanning from classical low-pass filtering to application of convolutional neural networks [19] and transformers [20]. A series of papers have considered the impact of the measurement kernel in the algorithms. By using single-shot digital holography, the authors in [21], [22] develop heuristic method to obtain maximum a posteriori estimate of the real-valued speckle-free object reflectance. They later extend this method to handle multi-shot measurements and incorporate more accurate image priors [8], [23], [24]. While these methods can work with non-identity A's, they still require A to be well-conditioned.

Our paper is different from the existing literature, mainly because we study scenarios where the matrix A is under-sampled ($m < n$). In a few recent papers, researchers have explored similar problems [11], [25]. The paper [25] aligns closely in scope and approach with our work. The authors addressed a similar problem and advocated for the use of DIP-based PGD. Addressing the concerns highlighted in the last section (further elucidated in Section V-B), our Bagged-DIP-based PGD employing the Newton-Schulz algorithm significantly outperforms [25] in both reconstruction quality and computational complexity. We will provide more information in our simulation studies. Furthermore, we should emphasize that [25] did not offer any theoretical results regarding the performance of DIP-based MLE.

The authors in [11] theoretically demonstrated the feasibility of accurate recovery of $\boldsymbol{x}_o$ even for $m < n$ measurements. While our theoretical results build upon the contributions of [11], our paper extends significantly in three key aspects: (1) We address the multilook problem and investigate the influence of the number of looks on our bounds. To ensure sharp bounds, especially when L is large, we derive sharper bounds than those presented in [11]. These requires a different proof technique as detailed in our proof. (2) In contrast to the use of compression codes' codewords for the set $\mathcal{C}$ in [11], we leverage the range of a deep image prior, inspired by recent advances in machine learning. Despite presenting new challenges in proving our results, this approach enables us to simplify and establish the relationship between Mean Squared Error (MSE) and problem specification parameters such as n, m, k, L. (3) On the empirical side, the experiments in [11] were restricted to a few toy examples due to their limiting assumptions. In contrast, by leveraging Deep Image Priors (DIPs) together with the Newton–Schulz method and bagging, we are able to evaluate our algorithms on natural images and achieve state-of-the-art results.

Given DIP's flexibility, it has been employed for various imaging and (blind) inverse problems, e.g., compressed sensing, phase retrieval etc. [26]–[31]. To boost the performance of DIP in these applications, researchers have explored several ideas, including, introducing explicit regularization [32], incorporating prior on network weights by introducing a learned regularization method into the DIP structure [33], combining with pre-trained denoisers in a Plug-and-Play fashion [34], [35], and exploring the effect of changing DIP structures and input noise settings to speed up DIP training [36].

Lastly, it's important to note our work can be situated within the realm of compressed sensing (CS) [1], [2], [37]–[41], where the objective is to derive high-resolution images from lower-resolution measurements. However, notably, the specific challenge of recovery in the presence of speckle noise has not been explored in the literatures before, except in [11] that we discussed before.

III. Main Theoretical Result

A. Assumptions and their justifications

As we described in the last section, in our theoretical work, we consider the cases in which $m < n$. m can even be much smaller than n. Furthermore, for notational simplicity, in our theoretical work only, we assume that the measurements and noises are real-valued.[1] Hence, we work with the following likelihood function:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}), \tag{4}$$

where

$$f(\boldsymbol{x}) = \log\det\left(\sigma_z^2 I_m + \sigma_w^2 A X^2 A^\top\right) + \frac{1}{L}\sum_{\ell=1}^{L} \boldsymbol{y}_\ell^\top \left(\sigma_z^2 I_m + \sigma_w^2 A X^2 A^\top\right)^{-1} \boldsymbol{y}_\ell. \tag{5}$$

Note that we omit subscript L from the likelihood as a way to distinguish between the negative loglikelihood of real-valued measurements from the complex-valued ones. The following theorem is the main theoretical result of the paper. Consider the case of no additve noise, i.e. $\sigma_z = 0$, and that for all i, we have $0 < x_{\min} \leq x_{o,i} \leq x_{\max}$.

As we discussed before, our theoretical results are based on the "Deep image prior" hypothesis, mentioned in the introduction. However, the following aspects of the hypothesis are not mathematically rigorous and should be carefully discussed:

1) What are the mathematical properties of $g_{\boldsymbol{\theta}}(\boldsymbol{u})$ as a function of $\boldsymbol{\theta} \in \mathbb{R}^k$?
2) If $\boldsymbol{\theta}$ belongs to a set Θ, what assumptions should we have about Θ based on the assumptions we have for $\boldsymbol{x}$, i.e., the assumption $0 < x_{\min} \leq x_{o,i} \leq x_{\max}$?

Clearly, the answers to the two questions raised above are closely related, and various reasonable choices can be made in addressing them. In this paper, we adopt a natural and relatively mild assumption that has been commonly used for $g_{\boldsymbol{\theta}}(\boldsymbol{u})$ as a function of $\boldsymbol{\theta}$. Specifically, we assume that $g_{\boldsymbol{\theta}}(\boldsymbol{u})$ is a Lipschitz function with respect to $\boldsymbol{\theta} \in \Theta$, with Lipschitz constant 1. Under this assumption, we must carefully consider the constraints we impose on the set Θ.

Assume that $g_{\mathbf{0}}(\boldsymbol{u}) = \mathbf{0}$, which holds for all networks typically used in the Deep Image Prior (DIP) framework. Suppose Θ is a compact set, and define its radius as

$$r_\Theta = \sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\theta}\|_2.$$

Since $g_{\boldsymbol{\theta}}(\boldsymbol{u})$ is a Lipschitz function, we can conclude that

$$\|g_{\boldsymbol{\theta}}(\boldsymbol{u})\|_2 \leq \|\boldsymbol{\theta}\|_2 \leq r_\Theta.$$

[1]For the complex-valued problem, since the phases of the elements of $\boldsymbol{x}_o$ are not recoverable, we can assume that $\boldsymbol{x}_o$ is real-valued. Even though in this case, the problem is similar to the problem we study in this paper, given that we have to deal with real and imaginary parts of the measurement matrices and noises, they are notationally more involved.

On the other hand we know that since for $\boldsymbol{x} = g_{\boldsymbol{\theta}}(\boldsymbol{u})$, we have $0 < x_{\min} \leq x_{o,i} \leq x_{\max}$, then $x_{\min}\sqrt{n} \leq \|\boldsymbol{x}\|_2 = \|g_{\boldsymbol{\theta}}(\boldsymbol{u})\|_2 \leq x_{\max}\sqrt{n}$. This gives a lower bound for r_Θ. Hence, we will assume that $r_\Theta = x_{\max}\sqrt{n}$. In summary, defining

$$B_k\left(\mathbf{0}, r\right) = \{\theta \in \mathbb{R}^k \mid \|\theta\|_2 \leq r\}.$$

we assume that

$$\Theta = B_k\left(\mathbf{0}, x_{\max}\sqrt{\frac{n}{k}}\right).$$

Note that based on the DIP hypothesis we will have

$$\mathcal{C} \subset \{\boldsymbol{x} \mid \boldsymbol{x} = g_{\boldsymbol{\theta}}(\boldsymbol{u}), \boldsymbol{\theta} \in \Theta\}.$$

As a final remark, we should clarify that under our assumptions $\mathcal{C}$ cannot be equal to $\{\boldsymbol{x} \mid \boldsymbol{x} = g_{\boldsymbol{\theta}}(\boldsymbol{u}), \boldsymbol{\theta} \in \Theta\}$. This is mainly because $\mathbf{0} \notin \mathcal{C}$, while

$$\mathbf{0} \in \{\boldsymbol{x} \mid \boldsymbol{x} = g_{\boldsymbol{\theta}}(\boldsymbol{u}), \boldsymbol{\theta} \in \Theta\}.$$

B. Main theorem and its implications

We are now ready to state our main theorem. Our main theorem captures the interplay between the accuracy of our maximum likelihood-based recovery, the number of measurements m, the number of looks L, the ambient dimension of the signals to be recovered n, and the number of the parameters in the DIP model.

Theorem III.1. *Let the elements of the measurement matrix A_{ij} be iid $\mathcal{N}(0,1)$. Suppose that $m < n$ and that the function $g_{\boldsymbol{\theta}}(\boldsymbol{u})$, as a function of $\boldsymbol{\theta} \in B_k\left(\mathbf{0}, x_{\max}\sqrt{\frac{n}{k}}\right)$, is Lipschitz with Lipschitz constant 1. Then*

$$\frac{1}{n}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 = C_1\left(\frac{n}{m^2}\cdot\frac{k\log n}{L} + \frac{\sqrt{k\log n}}{m}\right), \tag{6}$$

with probability $1 - C_2\left(e^{-\frac{m}{2}} + e^{-\frac{Ln}{8}} + e^{-C_3 k\log n} + e^{k\log n - \frac{n}{2}}\right)$ for some constants $C_1, C_2, C_3 > 0$.

Before discussing the proof sketch and the technical novelties of our proof strategy, we explain some of the conclusions that can be drawn from this theorem, provide some intuition, and compare with some of the existing results. Our first remark compares our result with the only existing theoretical work on this problem:

Remark III.2. A much weaker version of Theorem III.1 appeared in the shorter version of this paper published at the International Conference on Machine Learning. Specifically, the leading term in the upper bound was $\frac{n}{m}\cdot\frac{\sqrt{k\log n}}{\sqrt{Lm}}$, which is substantially looser than the bound obtained here, particularly since $k\log n \ll Lm$. The sharper result presented in this paper is derived using a different proof strategy.

The only existing results on the recovery performance of coherent imaging systems are those presented in [11], [42]. The authors of [42] focus exclusively on the despeckling problem and do not address the central challenge considered in our work—namely, the existence of the measurement matrix A and the fact that $m < n$. As a result, their results are not directly comparable to ours. However, the results of [11] are more closely related to our work. But there are a few major differences between the theoretical result presented in [11] and Theorem III.1:

1) The theoretical results in [11] rely on the existence of a compression algorithm tailored to the set of images, whereas our work is based on the Deep Image Prior (DIP) hypothesis. The DIP hypothesis is more flexible and, as demonstrated in the simulation section, allows for efficient solutions to the associated optimization problem.
2) While there are some major differences, we interpret the α-dimension of the sequence of compression algorithms as the effective number of parameters k so that we can provide a more accurate comparison

between our result and the one presented in [11], [43]. The authors in the above work consider the setup where m, n are of the same orders, $L = 1$, and their upper bound [43, Corollary 1] takes the form $\sqrt{\frac{k \log n}{m}}$. Comparing this with our bound in the special case $L = 1$, we observe that when $\frac{k \log n}{m} \ll 1$, our bounds are significantly sharper than those presented in [11]. In fact, it is straightforward to see that since $n > m$ and n is proportional to m, our upper bound in Theorem III.1 simplifies to $\frac{k \log n}{m}$. One source of looseness in the proof of [11] arises from an early step that involves working with the expected log-likelihood. To overcome this issue, we develop a proof strategy that entirely avoids relying on the expected log-likelihood.

3) This paper also establishes the dependence of the error on the number of looks.

Now, we provide more information on the upper bound we have obtained in Theorem III.1. As is clear in (6), there are two terms in the MSE. One that does not change with L and the other term that decreases with L. To understand these two terms, we provide further explanation in the following remarks.

Remark III.3. As the number of parameters of DIP, k, increases (while keeping m, n, and L fixed), both error terms in the upper bound of MSE grow. This aligns with intuition, as increasing the number of parameters in $g_{\boldsymbol{\theta}}(\boldsymbol{u})$ allows the DIP model to generate more intricate images. Consequently, distinguishing between these diverse alternatives based on the noisy measurements becomes more challenging.

Remark III.4. The main interesting feature of the first term in the MSE, i.e., $\frac{nk \log n}{Lm^2}$, is the fact that it grows rapidly as a function of n. In imaging systems with only additive noise, the growth is often logarithmic in n [44], contrasting with polynomial growth observed here. This can be attributed to the fact that as we increase n, the number of speckle noise elements present in our measurements also increases. Hence, it is reasonable to expect the error term to grow faster in n compared with additive noise models. However, the exact rate at which the error increases is yet unclear. Nonetheless, we believe the dependency on L for the term $\frac{nk \log n}{Lm^2}$ is sharp as it aligns with the notion of parametric error rate $\frac{1}{L}$ for an estimation problem with L samples.

Remark III.5. As $L \to \infty$, the first term in the upper bound of MSE converges to zero, and the dominant term becomes $\sqrt{k \log n}/m$. Note that since we are considering a fixed matrix A across the looks, even when L goes to infinity, we should not expect to be able to recover $\boldsymbol{x}_o$ independent of the value of m. One heuristic way to see this is to calculate

$$\frac{1}{L} \sum_{\ell=1}^{L} \boldsymbol{y}_\ell \boldsymbol{y}_\ell^\top = A X_o \frac{1}{L} \sum_{\ell=1}^{L} \boldsymbol{w}_\ell \boldsymbol{w}_\ell^\top X_o A^\top. \tag{7}$$

If we heuristically apply the weak law of large numbers and approximate $\frac{1}{L} \sum_\ell \boldsymbol{w}_\ell \boldsymbol{w}_\ell^\top$ with I, we get

$$\frac{1}{L} \sum_{\ell=1}^{L} \boldsymbol{y}_\ell \boldsymbol{y}_\ell^\top \approx A X_o^2 A^\top.$$

Under these approximations, the matirx $\frac{1}{L} \sum_\ell \boldsymbol{y}_\ell \boldsymbol{y}_\ell^\top$ provides $m(m+1)/2$ (due to symmetry) linear measurements of X_o^2. Hence, inspired by classic results in compressed sensing [45], intuitively, we expect the accurate recovery of $\boldsymbol{x}_o^2$ to be possible when $m^2 \gg k \log n$. The first error term in MSE is negligible when $m^2 \gg k \log n$, which is consistent with our conclusion based on the limit of $\frac{1}{L} \sum_\ell \boldsymbol{y}_\ell \boldsymbol{y}_\ell^\top$.

We next provide the key steps of the proof to highlight the technical novelties of our proof and also to enable the readers to navigate through the detailed proof more easily. We follow it up with the details to end this section.

C. Key steps in the proof of Theorem III.1

We first prove Denote the objective function f as

$$f(\boldsymbol{x}) = f(\Sigma(\boldsymbol{x})) = -\log\det\Sigma + \frac{1}{L\sigma_w^2} \sum_{\ell=1}^{L} \mathrm{Tr}(\Sigma \boldsymbol{y}_\ell \boldsymbol{y}_\ell^\top), \tag{8}$$

with $\Sigma = \Sigma(\boldsymbol{x}) = (AX^2A^\top)^{-1}$ and $X = \mathrm{diag}(\boldsymbol{x})$. Note that, f can be written as

$$f(\Sigma) = -\log\det\Sigma + \frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\mathrm{Tr}(\Sigma A X_o \boldsymbol{w}_\ell \boldsymbol{w}_\ell^\top X_o A^\top).$$

Let $\widehat{\boldsymbol{x}}$ denote the minimizer of the objective f, i.e.,

$$f(\widehat{\Sigma}) \leq f(\Sigma_o), \tag{9}$$

where $\widehat{\Sigma} = \Sigma(\widehat{\boldsymbol{x}})$ and $\Sigma_o = \Sigma(\boldsymbol{x}_o)$. For a given Σ, let $\bar{f}(\Sigma)$ denote the expected value of $f(\Sigma)$ with respect to $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_\ell$. It is straightforward to show

$$\bar{f}(\Sigma) = -\log\det\Sigma + \mathrm{Tr}(\Sigma A X_o^2 A^\top). \tag{10}$$

Expanding the terms in (9) we have

$$\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\widehat{\Sigma}\boldsymbol{y}_\ell - \log\det(\widehat{\Sigma}) \leq \frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\Sigma_o\boldsymbol{y}_\ell - \log\det(\Sigma_o). \tag{11}$$

Define

$$\Delta\Sigma = \widehat{\Sigma} - \Sigma_o. \tag{12}$$

Using the above definition and (10), we reorganize the terms in (11) to obtain

$$\bar{f}(\widehat{\Sigma}) - \bar{f}(\Sigma_o) \leq -\left[\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\Delta\Sigma\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\Sigma\Sigma_o^{-1})\right]. \tag{13}$$

The rest of the proof can be summarized in the following steps:

1) We establish a lower bound for $\bar{f}(\widehat{\Sigma}) - \bar{f}(\Sigma_o)$ in terms of $\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)$.
2) We obtain an upper bound for $-\left[\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\Delta\Sigma\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\Sigma\Sigma_o^{-1})\right]$ in terms of $\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)$.
3) We use the two bounds derived in Steps 1 and 2 to obtain an upper bound for $\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)$.
4) We use concentration arguments to obtain a high probability lower bound (in terms of the randomness in A) for $\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)$ in terms of $\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2$. By combining this with Step 3, we establish the proof. We provide more details for each of the above four steps below.

D. Details of the proof of Theorem III.1

1) **Lower bounding** $\bar{f}(\widehat{\Sigma}) - \bar{f}(\Sigma_o)$:
In Lemma VI.6 that will be presented in Section VI we show that

$$\bar{f}(\widehat{\Sigma}) - \bar{f}(\Sigma_o) \geq \frac{\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)}{2(1+\lambda_{\max})^2}, \tag{14}$$

where $\lambda_{\max}$ denote the maximum eigenvalue of $\Sigma_o^{-\frac{1}{2}}\Delta\Sigma\Sigma_o^{-\frac{1}{2}}$. To make this bound useful we have to find a data-independent upper bound for $\lambda_{\max}$. Note that

$$\lambda_{\max} = \max_{\boldsymbol{u}\in\mathbb{R}^n}\frac{\boldsymbol{u}^\top\Sigma_o^{-\frac{1}{2}}\Delta\Sigma\Sigma_o^{-\frac{1}{2}}\boldsymbol{u}}{\|\boldsymbol{u}\|_2^2} = \max_{\boldsymbol{u}\in\mathbb{R}^n}\frac{\left|\boldsymbol{u}^\top\left(\Sigma_o^{-\frac{1}{2}}\widehat{\Sigma}\Sigma_o^{-\frac{1}{2}} - I\right)\boldsymbol{u}\right|}{\|\boldsymbol{u}\|_2^2} \leq \left|\lambda_{\max}(\Sigma_o^{-1})\lambda_{\max}(\widehat{\Sigma})\right| + 1. \tag{15}$$

But $\widehat{\Sigma} = (A\widehat{X}^2A^\top)^{-1}$ and $\widehat{X} = \mathrm{diag}(\widehat{\boldsymbol{x}})$. Therefore,

$$\lambda_{\max}(\widehat{\Sigma}) = (\lambda_{\min}(A\widehat{X}^2A^\top))^{-1} \leq (\lambda_{\min}(AA^\top)x_{\min}^2)^{-1},$$

and

$$\lambda_{\max}\left(\Sigma_o^{-1}\right) = \lambda_{\max}(AX_o^2A^\top) \le \lambda_{\max}(AA^\top)x_{\max}^2.$$

To make the upper bound indepednet of the choice of matrix A, consider the event

$$\mathcal{E}_4 = \{\sqrt{n} - 2\sqrt{m} \le \sigma_{\min}(A) \le \sigma_{\max}(A) \le \sqrt{n} + 2\sqrt{m}\}. \tag{16}$$

In Lemma VI.2, presented in Section VI, we show tha $\mathbb{P}[\mathcal{E}_4] \ge 1 - 2e^{-\frac{m}{2}}$. Hence, conditioned on $\mathcal{E}_4$ we continue (15) to get

$$\lambda_{\max} \le \frac{\lambda_{\max}(AA^\top)x_{\max}^2}{\lambda_{\min}(AA^\top)x_{\min}^2} + 1 \le \frac{\left(\sqrt{n}+2\sqrt{m}\right)^2 x_{\max}^2}{\left(\sqrt{n}-2\sqrt{m}\right)^2 x_{\min}^2} + 1 \le \widetilde{c} \tag{17}$$

for some constant $\widetilde{c} > 0$ whenever $\frac{m}{n} \le \frac{1}{5}$. Combining (13), (14), and (17), we have

$$\frac{\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)}{2(1+\widetilde{c})^2} \le -\left[\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\Delta\Sigma\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\Sigma\Sigma_o^{-1})\right]. \tag{18}$$

2) **Upper bounding $-\left[\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\Delta\Sigma\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\Sigma\Sigma_o^{-1})\right]$:** Note that if we assume that $\Delta\Sigma$ is independent of $\boldsymbol{y}_\ell$, then we have $\mathbb{E}(\boldsymbol{y}_\ell^\top\Delta\Sigma\boldsymbol{y}_\ell) = \mathrm{Tr}(\Delta\Sigma\Sigma_o^{-1})$, and we could use standard concentration results to bound the difference. However, the main issue is that $\Delta\Sigma$ depends on $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L$. To resolve this issue, we use a δ-net argument, as will be clarified below. Consider a δ-net of the set $B_k(\mathbf{0}, x_{\max}\sqrt{\frac{n}{k}})$. We call the mapping of the δ-net under g, $\mathcal{C}_\delta$. The choice of δ will be discussed later. Define $\widetilde{\boldsymbol{x}}$ as the closest vector in $\mathcal{C}_\delta$ to $\widehat{\boldsymbol{x}}$, i.e.,

$$\widetilde{\boldsymbol{x}} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{C}_\delta} \|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|.$$

For $\widetilde{\boldsymbol{x}} \in \mathcal{C}_\delta$ define

$$\widetilde{X} = \mathrm{diag}(\widetilde{\boldsymbol{x}}), \quad \widetilde{\Sigma} = (A\widetilde{X}^2A^\top)^{-1}, \quad \Delta\widetilde{\Sigma} = \widetilde{\Sigma} - \Sigma_o.$$

Then, Lemma VI.8 proves that there is a constant $c > 0$ such that for every $\widetilde{\boldsymbol{x}} \in \mathcal{C}_\delta$

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\Delta\widetilde{\Sigma}\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\widetilde{\Sigma}\Sigma_o^{-1})\right| > t\right] \\
&\le 2\exp\left(-c\cdot\min\left\{\frac{L^2t^2}{\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})}, \frac{Lt\cdot x_{\min}^4\lambda_{\min}^2(AA^\top)}{x_{\max}^2(\sigma_{\max}(A))^2\lambda_{\max}(AA^\top)\|\boldsymbol{x}_o^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}\right\}\right).
\end{aligned} \tag{19}$$

Conditioned on the event $\mathcal{E}_4$ in (16) we note whenever $m \le n$, we have

$$\begin{aligned}
\sigma_{\max} &\le \sqrt{n} + \sqrt{m} \\
\lambda_{\min}(AA^\top) = \sigma_{\min}(AA^\top) &\ge (\sigma_{\min}(A))^2 \ge (\sqrt{n} - \sqrt{m})^2 \\
\lambda_{\max}(AA^\top) = \sigma_{\max}(AA^\top) &\le (\sigma_{\max}(A))^2 \le (\sqrt{n} + \sqrt{m})^2 \\
\|\boldsymbol{x}_o^2 - \widetilde{\boldsymbol{x}}^2\|_\infty &\le x_{\max}^2.
\end{aligned}$$

The above implies for $n \ge 2m$, there is a constant $c > 0$ for which

$$\frac{x_{\min}^4\lambda_{\min}^2(AA^\top)}{x_{\max}^2(\sigma_{\max}(A))^2\lambda_{\max}(AA^\top)\|\boldsymbol{x}_o^2 - \widetilde{\boldsymbol{x}}^2\|_\infty} \ge \frac{x_{\min}^4(\sqrt{n}-\sqrt{m})^4}{x_{\max}^4(\sqrt{n}+\sqrt{m})^4} \ge c.$$

In view of above, we choose

$$t = \mathcal{R}\cdot\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})} + \mathcal{R}^2, \quad \mathcal{R} = c_2\sqrt{\frac{k\log n}{L}} \tag{20}$$

for a suitably large constant $c_2 > 0$ to be picked later. Then we have from (19) and (20), with a probability at least $1 - e^{-c_3 k \log n}$, for some constant $c_3 > 0$ depending on c_2

$$\mathbb{P}\left[\left|\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top \Delta\widetilde{\Sigma}\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\widetilde{\Sigma}\Sigma_o^{-1})\right| > \mathcal{R}\cdot\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})} + \mathcal{R}^2\right] \le 2e^{-c_3 k\log n}. \tag{21}$$

Next, note that from [46, Chapter 27] we have for any set $\mathcal{S} \subseteq \mathbb{R}^k$ with $\sup_{\boldsymbol{x}\in\mathcal{S}} \|\boldsymbol{x}\|_\infty < \tau$, the size of any δ-covering set is bounded as

$$|\mathcal{C}_\delta| \le \left(\frac{2\tau\sqrt{k}}{\delta}\right)^k. \tag{22}$$

In view of the above, by choosing $\delta = \frac{1}{n^8}, \tau = x_{\max}\sqrt{\frac{n}{k}}$ we can show that it is possible to construct a δ-net $\mathcal{C}_\delta$ of the set $B_k(\mathbf{0}, x_{\max}\sqrt{\frac{n}{k}})$ of size at most $|\mathcal{C}_\delta| \le e^{c_0 k\log n}$ for a constant $c_0 > 0$. In view of the above covering number bound we can pick $c_2 > 0$ in (21) large enough such that $c_3 > 2c_0$ and we continue (21) to get

$$\mathbb{P}\left[\left|\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top \Delta\widetilde{\Sigma}\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\widetilde{\Sigma}\Sigma_o^{-1})\right| > \mathcal{R}\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})} + \mathcal{R}^2 \text{ for all } \widetilde{\boldsymbol{x}} \in \mathcal{C}_\delta\right] \le 2e^{-c_0 k\log n}. \tag{23}$$

In view of the covering set argument, there exists $\widetilde{\boldsymbol{x}} \in \mathcal{C}_\delta$ such that $\|\widehat{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}\|_2 \le \delta$. Then we use Lemma VI.10 (with $\Delta\Sigma = \widehat{\Sigma} - \Sigma_o$) to get that the event

$$\mathcal{E}_2 = \left\{\left|\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top \Delta\Sigma\boldsymbol{y}_\ell - \mathrm{Tr}(\Delta\Sigma\Sigma_o^{-1})\right| \le \mathcal{R}\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)} + \mathcal{R}^2 + \frac{\widetilde{C}}{n^3}\right\} \tag{24}$$

satisfies

$$\mathbb{P}\left[\mathcal{E}_2\right] \ge 1 - O\left(e^{-c_4 k\log n} + e^{-\frac{m}{2}} + e^{-\frac{Ln}{8}}\right). \tag{25}$$

3) **Upper bounding** $\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)$: Combining (18) and (25), we can see that conditioned on the event $\mathcal{E}_2$ we have

$$\frac{1}{2(1+\widetilde{c})^2}\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma) \le \mathcal{R}\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)} + \mathcal{R}^2 + \frac{\widetilde{C}}{n^3}. \tag{26}$$

Defining

$$z = \sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)}, \quad a = \frac{1}{2(1+\widetilde{c}^2)}, \quad b = \mathcal{R}, \quad c = \mathcal{R}^2 + \frac{\widetilde{C}}{n^3}, \tag{27}$$

the last inequality reduces to

$$az^2 - bz - c \le 0 \tag{28}$$

which implies,

$$\frac{b - \sqrt{b^2+4ac}}{2a} \le z \le \frac{b+\sqrt{b^2+4ac}}{2a},$$

or in other words, as z in (27) is positive, we get

$$z^2 \le 2\left(\frac{b^2}{4a^2} + \frac{b^2+4ac}{4a^2}\right) = \frac{b^2}{a^2} + \frac{2c}{a}.$$

Recalling the definitions from (27) the above inequality transforms into

$$\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma) \leq 4(1+\widetilde{c}^2)^2\cdot\mathcal{R}^2 + 8(1+\widetilde{c}^2)(\mathcal{R}^2+\frac{\widetilde{C}}{n^3}) \leq 4(1+\widetilde{c}^2)^2\left\{3\mathcal{R}^2+\frac{\widetilde{C}}{n^3}\right\}. \quad (29)$$

4) **Lower bounding** $\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)$ **in terms of** $\|\boldsymbol{x}_o-\widehat{\boldsymbol{x}}\|_2$**:** Using Lemma VI.7 we get

$$\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma) \geq \frac{x_{\min}^4\lambda_{\min}^2(AA^\top)}{x_{\max}^8\lambda_{\max}^4(AA^\top)}\|A(\widehat{X}^2-X_o^2)A^\top\|_{\mathsf{HS}}^2. \quad (30)$$

To obtain a lower bound for $\|A(\widehat{X}^2-X_o^2)A^\top\|_{\mathsf{HS}}^2$ we first write:

$$\widehat{X}^2 = \widehat{X}^2-\widetilde{X}^2+\widetilde{X}^2,$$

where $\widetilde{X}=\mathrm{diag}(\widetilde{\boldsymbol{x}})$ is chosen from the δ-net with $\delta=\frac{1}{n^8}$,such that $\|\widehat{\boldsymbol{x}}-\widetilde{\boldsymbol{x}}\|_2\leq\delta$. Using $(\|B+C\|_{\mathsf{HS}}^2 \leq 2\left(\|B\|_{\mathsf{HS}}^2+\|C\|_{\mathsf{HS}}^2\right))$ we get

$$\|A(\widehat{X}^2-X_o^2)A^\top\|_{\mathsf{HS}}^2 \geq \frac{1}{2}\|A(\widetilde{X}^2-X_o^2)A^\top\|_{\mathsf{HS}}^2 - \|A(\widetilde{X}_o^2-\widehat{X}^2)A^\top\|_{\mathsf{HS}}^2 \quad (31)$$

We bound $\|A(\widetilde{X}^2-\widehat{X}^2)A^\top\|_{\mathsf{HS}}^2$ from above via inequalities on the Hilbert-Schmidt norm and the operator norm, [47, Chapter IX], conditioned on the high probability event $\mathcal{E}_4$ as in (16)

$$\|A(\widetilde{X}^2-\widehat{X}^2)A^\top\|_{\mathsf{HS}}^2 \leq (\sigma_{\max}(A))^4\|\widetilde{X}^2-\widehat{X}^2\|_{\mathsf{HS}}^2 \leq (2x_{\max})^2n^2\|\widetilde{\boldsymbol{x}}-\widehat{\boldsymbol{x}}\|_2^2 \leq (2x_{\max})^2n^2\delta^2. \quad (32)$$

Then it remains to find a lower bound for $\|A(\widetilde{X}^2-X_o^2)A^\top\|_{\mathsf{HS}}$ in terms of $\|\boldsymbol{x}_o-\widehat{\boldsymbol{x}}\|_2$.
Towards this goal, for $\gamma>0$ define $\mathcal{E}_1(\gamma)$ as the event that

$$\|A(\widetilde{X}^2-X_o^2)A^\top\|_{\mathsf{HS}}^2 \geq m(m-1)\|\boldsymbol{x}_o^2-\widetilde{\boldsymbol{x}}^2\|_2^2 - m^2n\gamma,$$

We show that for an appropriate value of γ this event holds with high probability. In Lemma VI.11, we will prove that

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &\overset{(a)}{\leq} 2Ce^{k\log\frac{2k}{\delta}}\exp\left(-c\cdot\min\left(\frac{\check{\alpha}_{m,n}\gamma^2}{x_{\max}^8},\frac{\check{\beta}_{m,n}\gamma}{x_{\max}^4}\right)\right)+2e^{k\log\frac{2k}{\delta}-n/2}\\
&\leq 2Ce^{k\log\frac{2k}{\delta}}\left(e^{-c\frac{\check{\alpha}_{m,n}\gamma^2}{x_{\max}^8}}+e^{-c\frac{\check{\beta}_{m,n}\gamma}{x_{\max}^4}}\right)+2e^{k\log\frac{2k}{\delta}-\frac{n}{2}}\\
&= 2Ce^{k\log\frac{2k}{\delta}}e^{-\frac{cm^2\gamma^2}{C^2x_{\max}^8(2+\sqrt{m/n})^4}}+2Ce^{k\log\frac{2k}{\delta}}e^{-\frac{cm^2\gamma}{Cx_{\max}^4(2+\sqrt{m/n})^2}}+2e^{k\log\frac{2k}{\delta}-n/2},
\end{aligned} \quad (33)$$

where

$$\begin{aligned}
\check{\alpha}_{m,n} &\triangleq \frac{m^4n^2}{C^2m^2(2\sqrt{n}+\sqrt{m})^4} = \frac{m^2n^2}{C^2(2\sqrt{n}+\sqrt{m})^4},\\
\check{\beta}_{m,n} &\triangleq \frac{m^2n}{C(2\sqrt{n}+\sqrt{m})^2}.
\end{aligned} \quad (34)$$

By combining the result (22), the union bound on the choice of $\widetilde{\boldsymbol{x}}$ and Lemma VI.11 we reach Inequality (a). By setting

$$\gamma = 2C\frac{x_{\max}^4(2+\sqrt{m/n})^2}{m\sqrt{c}}\sqrt{k\log\frac{2k}{\delta}}, \quad (35)$$

we have on the event $\mathcal{E}_1$

$$\begin{aligned}\|A(\widetilde{X}^2 - X_o^2)A^\top\|_{\mathsf{HS}}^2 &\geq m(m-1)\sum_i^n (\widetilde{x}_i^2 - x_{o,i}^2)^2 - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} \\ &= m(m-1)\sum_i^n (\widetilde{x}_i - x_{o,i})^2(\widetilde{x}_i + x_{o,i})^2 - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} \\ &\geq 4m(m-1)x_{\min}^2\sum_i^n (\widetilde{x}_i - x_{o,i})^2 - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} \\ &= 4m(m-1)x_{\min}^2\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} \end{aligned} \tag{36}$$

with probability

$$\mathbb{P}(\mathcal{E}_1^c) \leq O(e^{-k\log\frac{k}{\delta}} + e^{k\log\frac{k}{\delta} - \frac{n}{2}}). \tag{37}$$

In the above equations $\widetilde{C}$ is a constant that does not depend on m, n or δ. Furthermore, in the last display we have assumed that m is large enough (and hence γ is small enough) to make the inequality $\frac{m^2\gamma^2}{C^2x_{\max}^8(2+\sqrt{m/n})^4} < \frac{m^2\gamma}{Cx_{\max}^4(2+\sqrt{m/n})^2}$ true. Simplifying the above, with

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 \leq \|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 + \|\widehat{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}\|_2 \leq \|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_o\|_2 + \delta,$$

we use $(b+c)^2 \leq 2(b^2+c^2)$ with $b = \|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_o\|_2, c = \delta$ to get

$$\begin{aligned}\|A(\widetilde{X}^2 - X_o^2)A^\top\|_{\mathsf{HS}}^2 &\geq 4m(m-1)x_{\min}^2\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} \\ &\geq 4m(m-1)x_{\min}^2\left(\frac{1}{2}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 - \delta^2\right) - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} \end{aligned} \tag{38}$$

In view of (31), we combine (30),(32),(38) to get

$$\begin{aligned}&\frac{x_{\max}^8\lambda_{\max}^4(AA^\top)}{x_{\min}^4\lambda_{\min}^2(AA^\top)}\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma) \geq \frac{1}{2}\|A(\widetilde{X}^2 - X_o^2)A^\top\|_{\mathsf{HS}}^2 - (2x_{\max})^2n^2\delta^2 \\ &\geq \frac{4m(m-1)x_{\min}^2}{2}\left(\frac{1}{2}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 - \delta^2\right) - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} - (2x_{\max})^2n^2\delta^2.\end{aligned}$$

5) **Combining the results:** In view of (29) and the last display, on the event $\mathcal{E}_4$ as in (16) we get

$$\begin{aligned}&\frac{4m(m-1)x_{\min}^2}{2}\left(\frac{1}{2}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 - \delta^2\right) - \widetilde{C}mn\sqrt{k\log\frac{2k}{\delta}} - (2x_{\max})^2n^2\delta^2 \\ &\leq \widetilde{C}_2\frac{x_{\max}^8\lambda_{\max}^4(AA^\top)}{x_{\min}^4\lambda_{\min}^2(AA^\top)}\cdot\left\{\mathcal{R}^2 + \frac{1}{n^3}\right\} \leq \widetilde{C}_3n^2\left(\frac{k\log n}{L} + \frac{1}{n^3}\right),\end{aligned}$$

for constants $C_1, \widetilde{C} > 0$ depending on $x_{\min}, x_{\max}$. Simplifying with $\delta = \frac{1}{n^8}$, and (37),(25),(16) we get

$$\frac{1}{n}\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_o\|_2^2 \leq \widetilde{C}_4\left(\frac{n}{m^2}\cdot\frac{k\log n}{L} + \frac{\sqrt{k\log n}}{m}\right).$$

for a constant $\widetilde{C}_4 > 0$ depending on $x_{\min}, x_{\max}$, on the event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$ with $\mathbb{P}[\mathcal{E}] \geq 1 - C_2\left(e^{-\frac{m}{2}} + e^{-\frac{Ln}{8}} + e^{-C_3k\log n} + e^{k\log n - \frac{n}{2}}\right)$ for some constants $C_2, C_3 > 0$. This completes our proof of Theorem III.1.

IV. Main Algorithmic Contributions

A. Summary of Projected Gradient Descent and DIP

As discussed in Section I, we aim to solve the optimization problem (5) under the DIP hypothesis. A popular heuristic for achieving this goal is using projected gradient descent (PGD). At each iteration t, the estimate $\boldsymbol{x}^t$ is updated as follows:

$$\boldsymbol{x}^{t+1} = \text{Proj}(\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t)), \tag{39}$$

where $\text{Proj}(\cdot)$ projects its input onto the range of the function $g_{\boldsymbol{\theta}}(\boldsymbol{u})$, and μ_t denotes the step size. The details of the calculation of $\nabla f_L(\boldsymbol{x}^t)$ are outlined in Appendix A.

An outstanding question in the implementation pertains to the nature of the projection operation $\text{Proj}(\cdot)$. If $g_\theta(\boldsymbol{u})$, in which θ denotes the parameters of the neural network and $\boldsymbol{u}$ denotes the input Gaussian noise, represents the reconstruction of the DIP, during training, DIP learns to reconstruct images by performing the following two steps:

$$\begin{aligned} \widehat{\theta}^t &= \underset{\theta}{\text{argmin}} \, \|g_\theta(\boldsymbol{u}) - (\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t))\|, \\ \boldsymbol{x}^{t+1} &= g_{\widehat{\theta}^t}(\boldsymbol{u}), \end{aligned} \tag{40}$$

where to obtain a local minima in the first optimization problem, we use Adam [48] . One of the main challenges in using DIPs in PGD is that the performance of DIP $g_{\boldsymbol{\theta}}(\boldsymbol{u})$ is affected by the structure choices, training iterations as well as the statistical properties of $\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t)$ [13]. We will discuss this issue in the next section.

B. Challenges of DIP-based PGD

In this section, we examine two primary challenges encountered by DIP-based PGD and present novel perspectives for addressing them.

1) Challenge 1: Right choice of DIP: Designing PGD, as described in Section IV-A, is particularly challenging when it comes to selecting the appropriate network structure for DIP. Figure 1 clarifies the main reason. In this figure, four DIP networks are used for fitting to the clean image (left panel) and an image corrupted by the Gaussian noise (right panel). As is clear, the sophisticated networks fit the clean image very well. However, they are more susceptible to overfitting when the image is corrupted with noise. On the other hand, the networks with simpler structures do not fit the clean image well but are less susceptible to noise than the sophisticated DIPs. This issue has been observed in previous work [10], [13].

The problem outlined above poses a challenge for the DIP-based PGD. Note that if $\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t)$ closely approximates $\boldsymbol{x}_o$, fitting a highly intricate DIP to $\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t)$ will yield an estimate that remains close to $\boldsymbol{x}_o$. Conversely, if overly simplistic networks are employed in this scenario, their final estimate may fail to closely approach $\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t)$, resulting in a low-quality estimate. In the converse scenario, where $\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t)$ is significantly distant from $\boldsymbol{x}_o$, a complex network may overfit to the noise. On the contrary, a simpler network, capable of learning only fundamental features of the image, may generate an estimate that incorporates essential image features, bringing it closer to the true image.

The above argument suggests the following approach: initiate DIP-PGD with simpler networks and progressively shift towards more complex structures as the estimate quality improves[2]. However, finding the right complexity level of the DIP for each iteration of PGD, in which the statistics of the error in the estimate $\boldsymbol{x}^t - \mu_t \nabla f_L(\boldsymbol{x}^t)$ is not known and may be image dependent, is a challenging problem. In the next section, we propose a novel approach to addressing this issue.

[2]A somewhat weaker approach would be to use intricate networks at every iteration, but then use some regularization approach such as early stopping to control the complexity of the estimates.

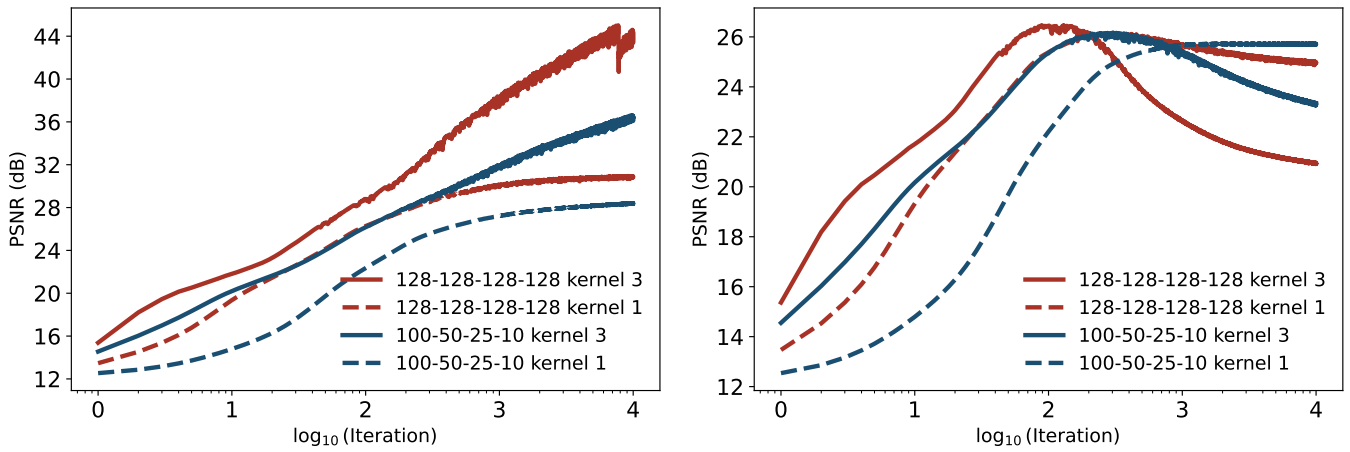

Fig. 1. PSNR (averaged over 8 images) versus iteration count is depicted for four DIP models fitted to both clean (left panel) and noisy images with only additive noise, noise level $\sigma = 25$ (right panel). The 4-layer networks used in DIP are specified in the legend.

2) Solution to Challenge 1: Bagged-DIP: Our new approach is based on a classical idea in statistics and machine learning: Bagging. Rather than finding the right complexity level for the DIP at each iteration, which is a computationally demanding and statistically challenging problem, we use bagging. The idea of bagging is that in the case of challenging estimation problems, we create several low-bias and, hopefully, weakly dependent estimates (we are overloading the phrase weakly dependent to refer to situations in which the cross-correlations of different estimates are not close to 1 or -1) of a quantity and then calculate the average of those quantities to obtain a lower-variance estimate. In order to obtain weakly dependent estimates, a common practice in the literature is to apply the same learning scheme to multiple datasets, each of which is a random perturbation of the original training set, e.g., the construction of random forests.

While there are many ways to create Bagged-DIP estimates, in this paper, we explore a few very simple estimates, leaving other choices for future research. First, we select a network that is sophisticated enough to fit nicely into real-world images. The details of the network we use for this paper can be found in Appendix C. Using the neural network provides our initial estimate of the image from the noisy observation. To generate a new estimate, we begin by selecting an integer number k, partitioning an image of size $(H \times W)$ into non-overlapping patches of sizes $(h_k \times w_k)$. Independent DIPs, with the same structure as the main one, are then employed to reconstruct each of these $(h_k \times w_k)$ patches. Essentially, the estimation of the entire image involves learning $\frac{HW}{h_k w_k}$ DIP models. By placing these $\frac{HW}{h_k w_k}$ patches back into their original positions, we obtain the estimate of the entire image, denoted as $\check{\boldsymbol{x}}_k$. A crucial aspect of this estimate is that the estimation of a pixel relies solely on the $(h_k \times w_k)$ patch to which the pixel belongs and no other pixel. By iterating this process for K different values of $(h_k \times w_k)$, we derive K estimates denoted as $\check{\boldsymbol{x}}_1, \ldots, \check{\boldsymbol{x}}_K$. The final sought-after estimate is obtained by averaging the individual estimates.

The estimation of a pixel in $\check{\boldsymbol{x}}_k$ is only dependent on the $(h_k \times w_k)$ patch to which the pixel belongs. As our estimates for different values of k utilize distinct regions of the image to derive their pixel estimates, we anticipate these estimates to be weakly dependent (again, in the sense that the cross-correlations are not close to 1 or -1).

3) Challenge 2: Matrix inversion: As shown in Appendix A, the gradient of $f_L(\boldsymbol{x})$ defined in (2) can be written as

$$\frac{\partial f_L}{\partial x_j} = 2\boldsymbol{x}_j \sigma_w^2 \left(\widetilde{\boldsymbol{a}}_{\cdot,j}^{+T} B^{-1} \widetilde{\boldsymbol{a}}_{\cdot,j}^{+} + \widetilde{\boldsymbol{a}}_{\cdot,j}^{-T} B^{-1} \widetilde{\boldsymbol{a}}_{\cdot,j}^{-} \right) - \frac{2\boldsymbol{x}_j \sigma_w^2}{L} \sum_{\ell=1}^{L} \left[\left(\widetilde{\boldsymbol{a}}_{\cdot,j}^{+T} B^{-1} \widetilde{\boldsymbol{y}}_\ell \right)^2 + \left(\widetilde{\boldsymbol{a}}_{\cdot,j}^{-T} B^{-1} \widetilde{\boldsymbol{y}}_\ell \right)^2 \right], \tag{41}$$

TABLE I
PSNR (DB) / SSIM ↑ OF 8 TEST IMAGES AT SAMPLING RATES $m/n = 0.125, 0.25, 0.5$, WITH NUMBER OF LOOKS $L = 1, 2, 4, 8, 16, 32, 64, 128$.

m/n	#Looks	Barbara	Peppers	House	Foreman	Boats	Parrots	Cameraman	Monarch	Average
0.125	1	13.05/0.126	13.34/0.132	12.90/0.127	11.20/0.083	12.92/0.098	10.40/0.090	9.08/0.069	12.36/0.161	11.91/0.111
	2	14.62/0.179	15.03/0.175	15.33/0.190	12.83/0.131	15.54/0.161	12.35/0.146	10.49/0.112	14.15/0.241	13.79/0.167
	4	16.11/0.273	16.80/0.322	16.49/0.274	14.29/0.268	17.32/0.247	13.64/0.228	11.75/0.156	15.60/0.327	15.25/0.262
	8	16.84/0.376	17.37/0.465	17.51/0.429	14.63/0.446	18.50/0.356	14.19/0.311	12.26/0.203	16.18/0.419	15.94/0.376
	16	17.78/0.293	18.11/0.305	18.20/0.286	17.98/0.257	17.86/0.259	15.36/0.279	13.01/0.226	17.26/0.463	16.95/0.296
	32	20.88/0.462	20.41/0.438	21.55/0.445	19.64/0.448	20.74/0.398	18.11/0.411	15.68/0.318	19.78/0.570	19.60/0.436
	64	21.71/0.587	21.92/0.585	23.52/0.571	20.80/0.590	22.21/0.503	19.27/0.522	17.73/0.407	21.72/0.675	21.11/0.555
	128	22.67/0.647	22.70/0.667	24.09/0.651	21.07/0.684	22.33/0.550	19.42/0.587	19.06/0.506	22.77/0.740	21.76/0.629
0.25	1	14.04/0.157	14.41/0.159	14.01/0.146	12.61/0.100	14.24/0.123	11.50/0.125	9.97/0.117	13.28/0.280	13.01/0.151
	2	16.23/0.235	16.85/0.245	17.17/0.277	14.74/0.195	16.89/0.208	13.87/0.226	11.96/0.189	16.02/0.392	15.47/0.246
	4	18.25/0.388	18.78/0.394	18.67/0.368	16.25/0.321	18.73/0.313	15.31/0.323	13.82/0.256	17.84/0.504	17.21/0.358
	8	18.94/0.480	19.41/0.516	20.02/0.530	17.50/0.544	19.90/0.430	16.57/0.436	15.09/0.328	18.84/0.591	18.28/0.482
	16	21.76/0.491	21.41/0.445	21.70/0.417	21.01/0.385	20.79/0.398	19.25/0.429	19.13/0.417	20.55/0.612	20.70/0.449
	32	23.96/0.611	24.05/0.600	25.17/0.573	23.88/0.593	23.13/0.528	22.18/0.571	22.33/0.526	23.29/0.713	23.50/0.590
	64	25.81/0.734	25.74/0.719	27.66/0.683	25.27/0.714	24.84/0.632	23.99/0.671	25.04/0.656	25.23/0.804	25.45/0.702
	128	26.70/0.774	26.34/0.782	28.76/0.741	26.54/0.791	25.59/0.683	24.80/0.741	27.01/0.780	26.44/0.857	26.52/0.769
0.5	1	16.31/0.242	16.77/0.246	16.42/0.225	15.70/0.164	16.36/0.213	13.87/0.255	14.00/0.317	15.73/0.413	15.65/0.259
	2	19.17/0.388	19.70/0.376	19.57/0.360	17.61/0.297	19.26/0.315	17.03/0.381	18.23/0.503	19.07/0.556	18.70/0.397
	4	21.03/0.533	21.57/0.549	22.03/0.518	20.08/0.491	21.74/0.456	19.32/0.515	21.28/0.639	21.78/0.683	21.10/0.548
	8	22.19/0.650	22.73/0.666	23.84/0.639	21.51/0.683	22.72/0.565	20.22/0.620	24.27/0.773	23.08/0.760	22.57/0.669
	16	25.85/0.691	25.67/0.636	26.53/0.609	25.77/0.621	24.83/0.596	24.52/0.664	25.87/0.675	25.48/0.789	25.56/0.660
	32	27.77/0.780	27.70/0.753	29.26/0.725	27.76/0.758	26.81/0.699	26.23/0.752	28.47/0.778	27.84/0.868	27.73/0.764
	64	28.91/0.823	28.70/0.825	30.90/0.790	28.99/0.842	27.98/0.763	27.64/0.822	30.47/0.859	29.30/0.914	29.11/0.830
	128	29.43/0.846	29.37/0.861	31.71/0.824	29.33/0.885	28.57/0.792	28.28/0.859	31.76/0.911	30.07/0.937	29.82/0.864

where $\widetilde{\mathbf{a}}^{+}_{\cdot,j} = \begin{bmatrix} \Re(\mathbf{a}_{\cdot,j}) \\ \Im(\mathbf{a}_{\cdot,j}) \end{bmatrix}$, $\widetilde{\mathbf{a}}^{-}_{\cdot,j} = \begin{bmatrix} -\Im(\mathbf{a}_{\cdot,j}) \\ \Re(\mathbf{a}_{\cdot,j}) \end{bmatrix}$, $\widetilde{\boldsymbol{y}}_{\ell} = \begin{bmatrix} \Re(\boldsymbol{y}_{\ell}) \\ \Im(\boldsymbol{y}_{\ell}) \end{bmatrix}$, $\mathbf{a}_{\cdot,j}$ denotes the j-th column of matrix A. It's important to highlight that in each iteration of the PGD, the matrix B changes because it depends on the current estimate $\boldsymbol{x}^t$. This leads to the computation of the inverse of a large matrix $B \in \mathbb{R}^{2m \times 2m}$ at each iteration, posing a considerable computational challenge and a significant obstacle in applying DIP-based PGD for this problem. In the next section, we present a solution to address this issue.

4) Solution to Challenge 2: To address the challenge mentioned in the previous section, we propose using the Newton-Schulz algorithm. Newton-Schulz is an iterative algorithm for obtaining a matrix inverse. The iterations of Newton-Schulz for finding $(B_t)^{-1}$ is given by

$$M^k = M^{k-1} + M^{k-1}(I - B_t M^{k-1}), \tag{42}$$

where M^k is the approximation of $(B_t)^{-1}$ at iteration k. $M^0 = (B_{t-1})^{-1}$. It is shown that if $\sigma_{\max}(I - M^0 B_t) < 1$, the Newton-Schulz converges to B_t^{-1} quadratically fast [49], [50].

An observation to alleviate the issue mentioned in the previous section is that, given the nature of gradient descent, we don't anticipate significant changes in the matrix X_t^2 from one iteration to the next. Consequently, we expect B_t and B_{t-1}, as well as their inverses, to be close to each other. Hence, instead of calculating the full inverse at iteration $t+1$, we can employ the Newton-Schulz algorithm with M^0 set to $(B_t)^{-1}$ from the previous iteration. Our simulations will show that **one** step of the Newton-Schulz algorithm suffices.

V. Simulation Results

A. Study of the Impacts of Different Modules

1) Newton-Schulz iterations: In this section, we aim to answer the following questions: (1) Is the Newton-Schulz algorithm effective in our Bagged-DIP-based PGD? (2) What is the minimum number of iterations for the Newton-Schulz algorithm to have good performance in Bagged-DIP-based PGD? (3) How does the computation time differ when using the Newton-Schulz algorithm compared to exact inverse computation?

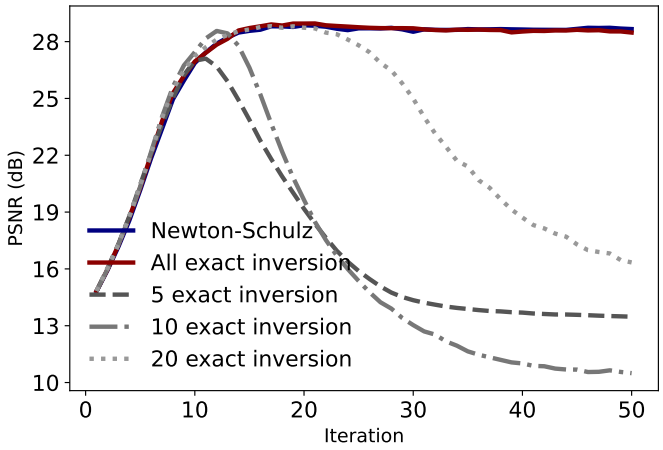

Fig. 2. Newton-Schulz approximation compared with computing exact inverse for all interations, the rest of the curves correspond to stopping the update of the inverse after the first 5, 10, and 20 iterations respectively. The number of looks is $L = 32$, sampling rate is $m/n = 0.5$. The test image is "Cameraman".

Figure 2 shows one of the simulations we ran to address the first two questions. In this figure, we have chosen $L = 32$ and $m/n = 0.5$, and the learning rate of PGD is 0.01. The result of Bagged-DIP-based PGD with a **single step** of Newton-Schulz is virtually identical to PGD with the exact inverse. To investigate the impact of the Newton-Schulz algorithm further, we next checked if applying even one step of Newton-Schulz is necessary. Hence, in three different simulations we stopped the matrix inverse update at iterations 5, 10, and 20. As is clear from Figure 2, a few iterations after stopping the update, PGD starts diverging. Hence, we conclude that a single step of the Newton-Schulz is necessary and sufficient for PGD.

To address the last question raised above, we evaluated how much time the calculation of the gradient takes if we use one step of the Newton-Schulz compared to the full matrix inversion. Our results are reported in Table II. Our simulations are for sampling rate 50%, and number of looks $L = 50$ and three different images sizes.[3] As is clear the Newton-Schulz is much faster.

TABLE II
RUNTIME (IN SECONDS) FOR MATRIX INVERSION AND ITS NEWTON–SCHULZ APPROXIMATION IN EACH PGD STEP.

Method	**32×32**	**64×64**	**128×128**
GD w/ Newton–Schulz	∼7e-5	∼8e-5	∼1e-4
GD w/o Newton–Schulz	∼0.3	∼1.2	∼52.8

In our final algorithm, i.e. Bagged DIP-based PGD, if the difference between $\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_\infty > \delta_{\boldsymbol{x}}$, then we use the exact inverse update. $\delta_{\boldsymbol{x}}$ is set to 0.12 (please refer to Appendix C for details) in all our simulations. Based on this updating criterion, we observe that the exact matrix inverse is only required for the first 2-3 iterations, and it is adaptive enough to guarantee the convergence of PGD.

2) Bagged-DIP: Intuitively speaking, in bagging, the more weakly dependent estimates one generate the better the average estimate will be. In the context of DIPs, there appear to be many different ways to create weakly dependent samples. The goal of this section is not to explore the full-potential of Bagged-DIPs. Instead, we aim to demonstrate that even a few weakly dependent samples can offer noticeable improvements. Hence, unlike the classical applications of bagging in which thousands of bagged samples are generated, to keep the computations managable, we have only considered three bagged estimates.

[3]Our algorithm still faces memory limitations on a single GPU when processing 256×256 images. Addressing this issue through approaches like parallelization remains subject for future research.

Figure 3 shows one of our simulations. In this simulation we have chosen $K = 3$, i.e. we have only three weakly-dependent estimates. These estimates are constructed according to the recipe presented in Section IV-B2 with the following patch sizes: $h_1 = w_1 = 32$, $h_2 = w_2 = 64$, and $h_3 = w_3 = 128$. As is clear from the left panel of Figure 3, even with these very few samples, Bagged-DIPs has offered between 0.5dB and 1dB over the three estimates it has combined.

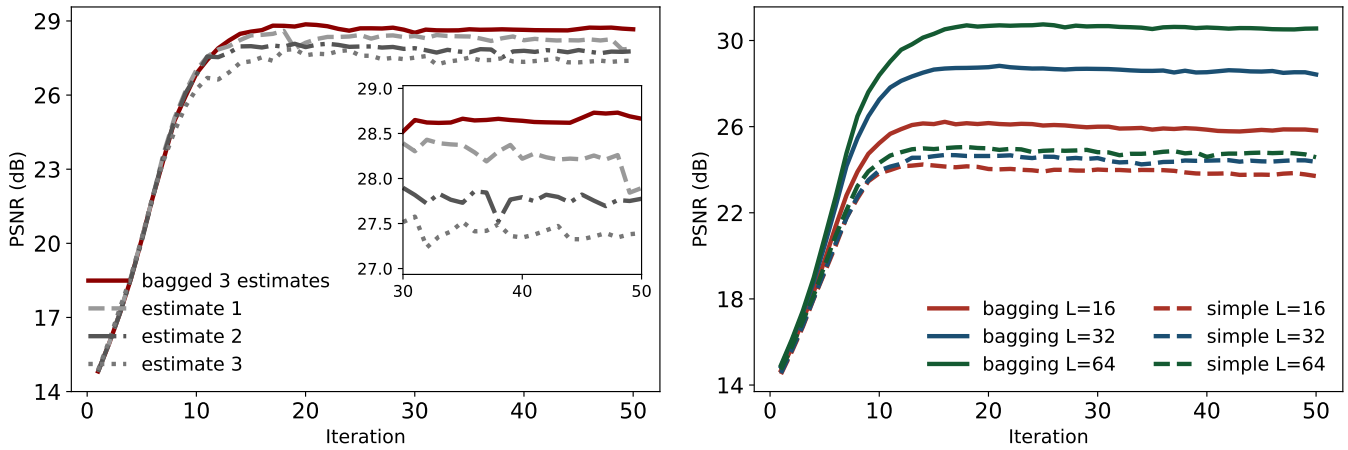

Fig. 3. (Left) We compare a Bagged-DIP with three sophisticated DIP estimates, where $L = 32$, $m/n = 0.5$. (Right) We compare PGD with simple and Bagged-DIP across different looks $L = 16, 32, 64$. The test image is "Cameraman".

3) Simple architectures versus Bagged-DIPs: So far our simulations have been focused on sophisticated networks. Are simpler networks that trade variance for the bias able to offer better performance? The right panel of Figure 3 compares the performance of Bagged-DIP-based PGD with that of PGD with a simple DIP. Not only this figure shows the major improvement that is offered by using more complicated networks (in addition to bagging), but also it clarifies one of the serious limitations of the simple networks. Note that as L increases, the performance of PGD with simple DIP is not improving. In such cases, the low-accuracy of DIP blocks the algorithm from taking advantage of extra information offered by the new looks.

B. Performance of Bagged-DIP-based PGD

In this section, we offer a comprehensive simulaion study to evaluate the performance of the Bagged-DIP-based PGD on several images. We explore the following settings in our simulations:

- Number of looks (L): $L = 1, 2, 4, 8, 16, 32, 64, 128$.
- Undersampling rate ($\frac{m}{n}$): $\frac{m}{n} = 0.125, 0.25, 0.5$.

For each combination of L and m/n, we pick one of the 8, 128×128 images mentioned in Table I.[4] We then generate the matrix $A \in \mathbb{C}^{m \times n}$ by selecting the first m rows of a matrix that is drawn from the Haar measure on the space of orthogonal matrices. We then generate $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_L \sim \mathcal{CN}(0, 1)$, and for $\ell = 1, 2, \ldots, L$, calculate $\boldsymbol{y}_\ell = AX_o\boldsymbol{w}_\ell$.

For our implementation of Bagged-DIP-based PGD, we have made the following choices:

- Initialization: We initialize our algorithm with $\boldsymbol{x}_0 = \frac{1}{L}\sum_{\ell=1}^{L} |\bar{A}^\top \mathbf{y}_\ell|$. However, the final performance of DIP-based PGD is robust to the choice of initialization.
- Learning rate: We have selected a learning rate of 0.001 for the gradient desent of the likelihood when $L \leq 8$, and 0.01 otherwise, and learning rate of 0.001 in the training of DIPs.
- Number of iterations of SGD for training DIP: The details are presented in Table III in the appendix.
- Number of iterations of PGD: We run the outer loop (gradient descent of likelihood) for $100, 200, 300$ iterations when $m/n = 0.5, 0.25, 0.125$ respectively.

[4]Images from the Set11 [51] are chosen and cropped to 128×128 for computational manageability in Table I.

The peak signal-to-noise-ratio (PSNR) and structural index similarity (SSIM) of our reconstructions are all reported in Table I.

There are no other existing algorithms that are applicable in the undersampled regime ($m < n$) considered in this paper. The only algorithm addressing speckle noise in ill-conditioned and undersampled scenarios prior to our work is the vanilla PGD proposed in [25]. It can be seen that, for $L = 100$, $L = 50$, and $L = 25$ on average (being averaged over $m/n = 0.125, 0.25, 0.5$, and across all images) our algorithm outperforms the one presented in [25] by 1.09 dB, 1.47 dB, and 1.27 dB, respectively.

VI. Technical results

We present a few lemmas that are used in the proof of Theorem III.1.

Lemma VI.1. *Let B and C denote two $n \times n$ symmetric and invertible matrices. Then, if λ_i represents the i^{th} eigenvalue of $B^{-1} - C^{-1}$, we have $|\lambda_i| \in [-\frac{\sigma_{\max}(B-C)}{\sigma_{\min}(B)\sigma_{\min}(C)}, \frac{\sigma_{\max}(B-C)}{\sigma_{\min}(B)\sigma_{\min}(C)}]$.*

Proof. Suppose λ_i is the i^{th} eigenvalue of $B^{-1} - C^{-1}$. Then, there exists a norm 1 vector $\boldsymbol{v} \in \mathbb{R}^n$ such that

$$(B^{-1} - C^{-1})\boldsymbol{v} = \lambda_i \boldsymbol{v}.$$

Multiplying both sides by B, we have

$$(I - BC^{-1})\boldsymbol{v} = \lambda_i B\boldsymbol{v}.$$

Define $\boldsymbol{u} = C^{-1}\boldsymbol{v}$. Then, we have $(C - B)\boldsymbol{u} = \lambda_i BC\boldsymbol{u}$, or equivalently

$$\lambda_i \boldsymbol{u} = (BC)^{-1}(C - B)\boldsymbol{u}.$$

Hence,

$$|\lambda_i| \leq \frac{\sigma_{\max}(C - B)}{\sigma_{\min}(B)\sigma_{\min}(C)}.$$

□

Lemma VI.2. *[52] Let the elements of an $m \times n$ ($m < n$) matrix A be drawn independently from $\mathcal{N}(0, 1)$. Then, for any $t > 0$,*

$$\mathbb{P}(\sqrt{n} - \sqrt{m} - t \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \sqrt{n} + \sqrt{m} + t) \geq 1 - 2e^{-\frac{t^2}{2}}. \tag{43}$$

Lemma VI.3 (Concentration of χ^2 [53])**.** *Let $Z_1, Z_2, \ldots, Z_n$ denote a sequence of independent $\mathcal{N}(0, 1)$ random variables. Then, for any $t \in (0, 1)$, we have*

$$\mathbb{P}(\sum_{i=1}^{n} Z_i^2 \leq n(1 - t)) \leq e^{\frac{n}{2}(t + \log(1-t))}.$$

Also, for any $t > 0$,

$$\mathbb{P}(\sum_{i=1}^{n} Z_i^2 \geq n(1 + t)) \leq e^{-\frac{n}{2}(t - \log(1+t))}.$$

Theorem VI.4 (Hanson-Wright inequality)**.** *Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a random vector with independent components with $\mathbb{E}[X_i] = 0$ and $\|X_i\|_{\Psi_2} \leq K$. Let A be an $n \times n$ matrix. Then, for $t > 0$,*

$$\mathbb{P}\Big(|\boldsymbol{X}^\top A\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^\top A\boldsymbol{X}]| > t\Big) \leq 2\exp\left(-c\min\left(\frac{t^2}{K^4\|A\|_{\mathsf{HS}}^2}, \frac{t}{K^2\|A\|_2}\right)\right), \tag{44}$$

where c is a constant, and $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}(\exp(X^2/t^2)) \leq 2\}$.

Theorem VI.5 (Decoupling of U-processes, Theorem 3.4.1. of [54])**.** *Let $X_1, X_2, \ldots, X_n$ denote random variables with values in measurable space $(S, \mathcal{S})$. Let $(\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n)$ denote an independent copy of*

$X_1, X_2, \ldots, X_n$. *For* $i \neq j$ *let* $h_{i,j} : S^2 \to \mathbb{R}$. *Then, there exists a constant* C *such that for every* $t > 0$ *we have*

$$\mathbb{P}\Big(|\sum_{i\neq j} h_{i,j}(X_i, X_j)| > t\Big) \leq C\mathbb{P}\Big(C|\sum_{i\neq j} h_{i,j}(X_i, \widetilde{X}_j)| > t\Big).$$

Lemma VI.6. *[11] For* $\widetilde{\boldsymbol{x}} \in \mathbb{R}^n$ *and* $\boldsymbol{x}_o \in \mathbb{R}^n$, *let* $\widetilde{X} = \mathrm{diag}(\widetilde{\boldsymbol{x}})$, $X_o = \mathrm{diag}(\boldsymbol{x}_o)$. *Assume that* $A\widetilde{X}^2A^\top$ *and* $AX_o^2A^\top$ *are both invertible and define* $\widetilde{\Sigma} = (A\widetilde{X}^2A^\top)^{-1}, \Sigma_o = (AX_o^2A^\top)^{-1}, \Delta\widetilde{\Sigma} = \widetilde{\Sigma} - \Sigma_o$. *Let* $\lambda_{\max}$ *be the maximum eigen value of* $\Sigma_o^{-\frac{1}{2}}\Delta\widetilde{\Sigma}\Sigma_o^{-\frac{1}{2}}$. *Then,*

$$\bar{f}(\widetilde{\Sigma}) - \bar{f}(\Sigma_o) \geq \frac{1}{2(1+\lambda_{\max})^2}\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma}), \tag{45}$$

Lemma VI.7. *[11] Consider two* $m \times m$ *matrices* $\widetilde{\Sigma} = (A\widetilde{X}^2A^\top)^{-1}$ *and* $\Sigma = (AX^2A^\top)^{-1}$ *and define* $\Delta\Sigma = \widetilde{\Sigma} - \Sigma$. *Then,*

$$\mathrm{Tr}(\Sigma^{-1}\Delta\Sigma\Sigma^{-1}\Delta\Sigma) \geq \frac{x_{\min}^4\lambda_{\min}^2(AA^\top)}{x_{\max}^8\lambda_{\max}^4(AA^\top)}\|A(\widetilde{X}^2 - X^2)A^\top\|_{\mathsf{HS}}^2 \tag{46}$$

$$\mathrm{Tr}(\Sigma^{-1}\Delta\Sigma\Sigma^{-1}\Delta\Sigma) \leq \frac{x_{\max}^4\lambda_{\max}^2(AA^\top)}{x_{\min}^8\lambda_{\min}^4(AA^\top)}\|A(\widetilde{X}^2 - X^2)A^\top\|_{\mathsf{HS}}^2. \tag{47}$$

Lemma VI.8. *Assume that* $\widetilde{\Sigma} = A\widetilde{X}^2A^\top$ *and* $\Sigma_o = AX_o^2A^\top$ *are both invertible. Define* $\Delta\widetilde{\Sigma} = \widetilde{\Sigma} - \Sigma_o$. *Let*

$$\delta f(\Sigma) = \frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\Sigma\boldsymbol{y}_\ell - \mathrm{Tr}(\Sigma\Sigma_o^{-1}).$$

Then, for $t > 0$, *there exists a constant* c *independent of* $m, n, x_{\min}$, *and* $x_{\max}$, *such that*

$$\mathbb{P}(|\delta f(\Delta\widetilde{\Sigma})| \geq t|A) \leq 2\exp\left(-c\cdot\min\left\{\frac{L^2t^2}{\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})}, \frac{Lt\cdot x_{\min}^4\lambda_{\min}^2(AA^\top)}{x_{\max}^2(\sigma_{\max}(A))^2\lambda_{\max}(AA^\top)\|\boldsymbol{x}_o^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}\right\}\right).$$

Proof. Define matrix $B = X_oA^\top\Delta\widetilde{\Sigma}AX_o \in \mathbb{R}^{n\times n}$ and $\widetilde{B} \in \mathbb{R}^{Ln\times Ln}$ as

$$\widetilde{B} = \begin{bmatrix} B & 0 & \cdots & 0 \\ 0 & B & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & B \end{bmatrix}.$$

Furthermore, fr $\boldsymbol{w}_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \sigma_w^2\boldsymbol{I}_n)$ define

$$\boldsymbol{W}^\top = [\boldsymbol{w}_1^\top, \boldsymbol{w}_2^\top, \ldots, \boldsymbol{w}_L^\top].$$

Note that for any fixed $\Delta\widetilde{\Sigma}$ we have,

$$\mathbb{E}\left[\boldsymbol{W}^\top B\boldsymbol{W}\right] = L\mathbb{E}[\boldsymbol{w}_1^\top B\boldsymbol{w}_1] = L\mathbb{E}[\mathrm{Tr}(B\boldsymbol{w}_1\boldsymbol{w}_1^\top)] \overset{(a)}{=} L\sigma_w^2\mathrm{Tr}(\Delta\widetilde{\Sigma}AX_o^2A^\top) = L\sigma_w^2\mathrm{Tr}(\Delta\widetilde{\Sigma}\Sigma_o^{-1}).$$

where (a) followed using $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for two matrices A, B. Then, by the Hanson-Wright inequality (Theorem VI.4), we have

$$\mathbb{P}(|\frac{1}{L\sigma_w^2}\boldsymbol{W}^\top\widetilde{B}\boldsymbol{W} - \mathrm{Tr}(\Delta\widetilde{\Sigma}\Sigma_o^{-1})| > t) \leq 2\exp\left(-c\cdot\min\left\{\frac{L^2t^2}{\|\widetilde{B}\|_{\mathsf{HS}}^2}, \frac{Lt}{\|\widetilde{B}\|_2}\right\}\right). \tag{48}$$

Next, note that

$$\|\widetilde{B}\|_{\mathsf{HS}}^2 = L\mathrm{Tr}(B^2) = L\mathrm{Tr}(X_oA^\top\Delta\widetilde{\Sigma}AX_o^2A^\top\Delta\widetilde{\Sigma}AX_o) = L\mathrm{Tr}(AX_o^2A^\top\Delta\widetilde{\Sigma}AX_o^2A^\top\Delta\widetilde{\Sigma}) \tag{49}$$

and using $\|\widetilde{B}\|_2 = \|B\|_2$

$$\begin{aligned}\|\widetilde{B}\|_2 &= \|X_o A^\top \Delta\widetilde{\Sigma} A X_o\|_2 \\ &\le x_{\max}^2 \sigma_{\max}(A)\sigma_{\max}(A^\top)\sigma_{\max}(\Delta\widetilde{\Sigma}) \\ &\overset{(a)}{\le} \frac{x_{\max}^2(\sigma_{\max}(A))^2\sigma_{\max}(AX_o^2A^\top - A\widetilde{X}^2A^\top)}{\sigma_{\min}(A\widehat{X}^2A^\top)\sigma_{\min}(A\widetilde{X}^2A^\top)} \le \frac{x_{\max}^2(\sigma_{\max}(A))^2\lambda_{\max}(AA^\top)\|\boldsymbol{x}_o^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^4\lambda_{\min}^2(AA^\top)},\end{aligned} \tag{50}$$

where (a) followed using Lemma VI.1. Substituting the above in (48) we conclude the result. □

Lemma VI.9. *Assume that both $\widetilde{\Sigma} = (A\widetilde{X}^2A^\top)^{-1}$ and $\widehat{\Sigma} = (A\widehat{X}^2A^\top)^{-1}$ exists. Then,*

$$|\lambda_i(\Sigma_o^{-\frac{1}{2}}(\widehat{\Sigma} - \widetilde{\Sigma})\Sigma_o^{-\frac{1}{2}})| \in [0, \frac{x_{\max}^2\lambda_{\max}^2(AA^\top)\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^4\lambda_{\min}^2(AA^\top)}].$$

Furthermore, we have

$$\left|\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top(\widehat{\Sigma} - \widetilde{\Sigma})\boldsymbol{y}_\ell\right| \le \frac{x_{\max}^2\lambda_{\max}^2(AA^\top)\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^4\lambda_{\min}^2(AA^\top)}\Big(1 + \frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{w}_\ell^\top\boldsymbol{w}_\ell\Big). \tag{51}$$

Proof. To prove the first inequality, we note that

$$\begin{aligned}|\lambda_i(\Sigma_o^{-\frac{1}{2}}(\widehat{\Sigma} - \widetilde{\Sigma})\Sigma_o^{-\frac{1}{2}})| &\le \frac{\left|\sigma_{\max}(\widehat{\Sigma} - \widetilde{\Sigma})\right|}{\sigma_{\min}(\Sigma_o)} = \frac{\left|\sigma_{\max}((A\widehat{X}^2A^\top)^{-1} - (A\widetilde{X}^2A^\top)^{-1})\right|}{\sigma_{\min}(\Sigma_o)} \\ &\overset{(a)}{\le} \frac{\sigma_{\max}((A\widehat{X}^2A^\top) - (A\widetilde{X}^2A^\top))}{\sigma_{\min}(\Sigma_o)\sigma_{\min}(A\widehat{X}^2A^\top)\sigma_{\min}(A\widetilde{X}^2A^\top)} \\ &= \frac{\sigma_{\max}(AX_o^2A^\top)\left|\sigma_{\max}(A(\widehat{X}^2 - \widetilde{X}^2)A^\top)\right|}{\sigma_{\min}(A\widehat{X}^2A^\top)\sigma_{\min}(A\widetilde{X}^2A^\top)} \le \frac{x_{\max}^2\lambda_{\max}^2(AA^\top)\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^4\lambda_{\min}^2(AA^\top)}.\end{aligned}$$

To obtain inequality (a) we have used Lemma VI.1. To prove the second inequality, note that

$$\begin{aligned}&\left|\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top(\widehat{\Sigma} - \widetilde{\Sigma})\boldsymbol{y}_\ell\right| \\ &\le \left|\sigma_{\max}((A\widehat{X}^2A^\top)^{-1} - (A\widetilde{X}^2A^\top)^{-1})\right|\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\boldsymbol{y}_\ell \\ &\le \frac{\lambda_{\max}(AA^\top)\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^4\lambda_{\min}^2(AA^\top)}\frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top\boldsymbol{y}_\ell \le \frac{x_{\max}^2\lambda_{\max}^2(AA^\top)\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{L\sigma_w^2x_{\min}^4\lambda_{\min}^2(AA^\top)}\sum_{\ell=1}^{L}\boldsymbol{w}_\ell^\top\boldsymbol{w}_\ell.\end{aligned}$$

□

Lemma VI.10. *Define $h(\Sigma) = \frac{1}{L\sigma_w^2}\sum_{\ell=1}^{L}\boldsymbol{y}_\ell^\top(\Sigma - \Sigma_o)\boldsymbol{y}_\ell - \mathrm{Tr}((\Sigma - \Sigma_o)\Sigma_o^{-1})$ and denote $\widehat{\Sigma} = (A\widehat{X}^2A^\top)^{-1}, \Sigma_o = (AX_o^2A^\top)^{-1}, \widetilde{\Sigma} = (A\widetilde{X}^2A^\top)^{-1}$ as usual. There exists a constant $\widetilde{C}$ such that given any $\widehat{\boldsymbol{x}} = g_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{u})$ and $\widetilde{\boldsymbol{x}} = g_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{u})$ with $\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_2 \le \delta$ the following holds.*

(a.) $\mathbb{P}\left[|h(\widetilde{\Sigma}_o) - h(\widehat{\Sigma}_o)| \le \widetilde{C}n\delta\right] \ge 1 - O(e^{-\frac{m}{2}} + e^{-\frac{Ln}{8}})$

(b.) $\mathbb{P}\left[\left|\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})} - \sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\widetilde{\Sigma})}\right| \le \frac{x_{\max}^3\lambda_{\max}^3(AA^\top)\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^5\lambda_{\min}^3(AA^\top)}\right] \ge 1 - 2e^{-\frac{m}{2}}$

Proof. (a.) We start with proving the first result. In view of VI.9, the main objective of this proof strategy is to obtain upper bound for the following three terms:

- $\frac{1}{L\sigma_w^2}\sum_\ell \boldsymbol{w}_\ell^\top\boldsymbol{w}_\ell$:

Consider the following event:

$$\mathcal{E}_3 = \left\{ \frac{1}{L\sigma_w^2} \sum_{\ell=1}^{L} \boldsymbol{w}_\ell^\top \boldsymbol{w}_\ell \leq 2n \right\}$$

It is straightforward to use Lemma VI.3 to see that

$$\mathbb{P}(\mathcal{E}_3) \geq 1 - e^{-\frac{Ln}{8}}.$$

Hence, for the rest of the proof we will consider the high probability event $\mathcal{E}_3$.

- $\frac{\lambda_{\max}^2(AA^\top)}{\lambda_{\min}^2(AA^\top)}$:
Consider the event $\mathcal{E}_4$ described in (16)

$$\mathcal{E}_4 = \{\sqrt{n} - 2\sqrt{m} \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \sqrt{n} + 2\sqrt{m}\}$$

Then, in view of Lemma VI.2 we get $\mathbb{P}[\mathcal{E}_4] \geq 1 - 2e^{-\frac{m}{2}}$. Hence, conditioned on $\mathcal{E}_4$ it is straightforward to see that

$$\frac{\lambda_{\max}^2(AA^\top)}{\lambda_{\min}^2(AA^\top)} \leq \frac{(\sqrt{n} + 2\sqrt{m})^2}{(\sqrt{n} - 2\sqrt{m})^2}, \tag{52}$$

with probability

$$\mathbb{P}(\mathcal{E}_4) \geq 1 - 2\exp(-\frac{m}{2}). \tag{53}$$

- $\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty$: It is straightforward to use Lipschitzness of $g_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{u})$ to prove that

$$\|\widetilde{\boldsymbol{x}} - \widehat{\boldsymbol{x}}\|_2 \leq \|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\|_2 \leq \delta, \tag{54}$$

and

$$\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty \leq 2x_{\max}\|\widehat{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}\|_\infty \leq 2x_{\max}\|\widehat{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}\|_2 \leq 2x_{\max}\delta,$$

Summarizing the discussions of this section, we can conclude from VI.9

$$\left| \frac{1}{L\sigma_w^2} \sum_{\ell=1}^{L} \boldsymbol{y}_\ell^\top ((A\widehat{X}^2A^\top)^{-1} - (A\widetilde{X}^2A^\top)^{-1})\boldsymbol{y}_\ell \right| \leq \widetilde{C}_1 n\delta \tag{55}$$

for a constant $\widetilde{C}_1 > 0$. To complete the proof we exhibit the following bound on $\left|\lambda_{\max}((\widehat{\Sigma} - \widetilde{\Sigma})\Sigma_o^{-1})\right|$

$$\left|\lambda_{\max}((\widehat{\Sigma} - \widetilde{\Sigma})\Sigma_o^{-1})\right| \overset{(a)}{=} \left|\lambda_{\max}(\Sigma_o^{-\frac{1}{2}}(\widehat{\Sigma} - \widetilde{\Sigma})\Sigma_o^{-\frac{1}{2}})\right| \leq \frac{x_{\max}^2\lambda_{\max}^2(AA^\top)\|\widehat{\boldsymbol{x}}^2 - \widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^4\lambda_{\min}^2(AA^\top)}. \tag{56}$$

where (a) followed as the maximum absolute eigenvalues of AB and BA are equal. Hence, on the event $\mathcal{E}_4$ as in (16) we use (52) to get

$$\left|\mathrm{Tr}((\widehat{\Sigma} - \widetilde{\Sigma})\Sigma_o^{-1})\right| = \widetilde{C}_2 n\delta. \tag{57}$$

(b.) To prove the second result in view of (57) we note

$$\begin{aligned}
&\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma}) - \mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma) \\
&= \mathrm{Tr}((\Sigma_o^{-1}\Delta\widetilde{\Sigma})^2) - \mathrm{Tr}((\Sigma_o^{-1}\Delta\Sigma)^2) \\
&= \mathrm{Tr}((\Sigma_o^{-1}(\widehat{\Sigma} - \widetilde{\Sigma}))(\Sigma_o^{-1}(\Delta\widetilde{\Sigma} + \Delta\Sigma))) \\
&\leq n|\lambda_{\max}(\Sigma_o^{-1}(\widehat{\Sigma} - \widetilde{\Sigma}))| \cdot |\lambda_{\max}(\Sigma_o^{-1}(\Delta\widetilde{\Sigma} + \Delta\Sigma))| \leq \widetilde{C}_2 n\delta \cdot |\lambda_{\max}(\Sigma_o^{-1}(\Delta\widetilde{\Sigma} + \Delta\Sigma))|.
\end{aligned} \tag{58}$$

Finally, we note that

$$\begin{aligned}
|\lambda_{\max}(\Sigma_o^{-1}(\Delta\widetilde{\Sigma}+\Delta\Sigma))| &\le \frac{\sigma_{\max}(\Delta\widetilde{\Sigma}+\Delta\Sigma)}{\sigma_{\min}(\Sigma_o)} \\
&= \frac{\left|\sigma_{\max}((A\widehat{X}^2A^\top)^{-1}+(A\widetilde{X}^2A^\top)^{-1}-2(AX_o^2A^\top)^{-1})\right|}{\sigma_{\min}(\Sigma_o)} \\
&\le \frac{\left|\sigma_{\max}((A\widehat{X}^2A^\top)^{-1})\right|+\left|\sigma_{\max}((A\widetilde{X}^2A^\top)^{-1})\right|+2\left|\sigma_{\max}((AX_o^2A^\top)^{-1})\right|}{\sigma_{\min}(\Sigma_o)} \\
&= \sigma_{\max}(AX_o^2A^\top)\left(\frac{1}{\sigma_{\min}(A\widehat{X}^2A^\top)}+\frac{1}{\sigma_{\min}(A\widetilde{X}^2A^\top)}+\frac{2}{\sigma_{\min}(AX_o^2A^\top)}\right) \\
&\le \frac{4x_{\max}^2\lambda_{\max}(AA^\top)\|\widehat{\boldsymbol{x}}^2-\widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^2\lambda_{\min}(AA^\top)}. \qquad (59)
\end{aligned}$$

Hence, we continue (58) to get

$$\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})-\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)\le\frac{4\widetilde{C}_2n\delta x_{\max}^2\lambda_{\max}(AA^\top)\|\widehat{\boldsymbol{x}}^2-\widetilde{\boldsymbol{x}}^2\|_\infty}{x_{\min}^2\lambda_{\min}(AA^\top)}\le\widetilde{C}_3n\delta.$$

where the last inequality holds on the event $\mathcal{E}_4$ as in (16) for a suitable constant $\widetilde{C}_3$. Next we note the fact that for $a,b,c\ge 0$, we have $a-b<c$ means $\sqrt{a}-\sqrt{b}\le\sqrt{c}$ (as otherwise $\sqrt{a}-\sqrt{b}>\sqrt{c}$ will mean $a>b+c+2\sqrt{bc}$, leading to a contradiction). Hence we have on event $\mathcal{E}_4$

$$\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\widetilde{\Sigma}\Sigma_o^{-1}\Delta\widetilde{\Sigma})}-\sqrt{\mathrm{Tr}(\Sigma_o^{-1}\Delta\Sigma\Sigma_o^{-1}\Delta\Sigma)}\le\sqrt{\widetilde{C}_3n\delta}.$$

This completes the proof. □

Lemma VI.11. *Let the elements of $m\times n$ matrix A be drawn i.i.d. $\mathcal{N}(0,1)$. For any given $\boldsymbol{d}\in\mathbb{R}^n$, define $D=\mathrm{diag}(\boldsymbol{d})$. Then,*

$$\mathbb{P}(\|ADA^\top\|_{\mathsf{HS}}^2\le m(m-1)\sum_{i=1}^n d_i^2-t)\le 2C\exp\left(-c\min\left(\frac{t^2}{C^2\|\boldsymbol{d}\|_\infty^4 q_{m,n}},\frac{t}{C\|\boldsymbol{d}\|_\infty^2\widetilde{q}_{m,n}}\right)\right)+2e^{-\frac{n}{2}}, \qquad (60)$$

where C and c are the constants that appeared in Lemmas VI.5 and VI.4, and

$$\begin{aligned}
q_{m,n} &\triangleq m^2(2\sqrt{n}+\sqrt{m})^4, \\
\widetilde{q}_{m,n} &\triangleq (2\sqrt{n}+\sqrt{m})^2. \qquad (61)
\end{aligned}$$

Proof. Let $\boldsymbol{a}_i^\top$ denote the i^{th} row of matrix A. We have

$$\|ADA^\top\|_{\mathsf{HS}}^2=\sum_{i=1}^m\sum_{j=1}^m|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2\ge\sum_{i=1}^m\sum_{j\ne i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2. \qquad (62)$$

Note that

$$\mathbb{E}(\sum_{i=1}^m\sum_{j\ne i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2)=m(m-1)\sum_{i=1}^n d_i^2.$$

Using Theorem VI.5 we conclude that there exists a constant C such that

$$\begin{aligned}
&\mathbb{P}(|\sum_{i=1}^{m}\sum_{j\neq i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2 - m(m-1)\sum_{i=1}^{n}d_i^2| > t)\\
&\leq C\mathbb{P}(C|\sum_{i=1}^{m}\sum_{j\neq i}|\boldsymbol{a}_i^\top D\widetilde{\boldsymbol{a}}_j|^2 - m(m-1)\sum_{i=1}^{n}d_i^2| > t)\\
&= C\mathbb{P}(C|\sum_{i=1}^{m}\boldsymbol{a}_i^\top D\sum_{j\neq i}\widetilde{\boldsymbol{a}}_j\widetilde{\boldsymbol{a}}_j^\top D\boldsymbol{a}_i - m(m-1)\sum_{i=1}^{n}d_i^2| > t),
\end{aligned} \tag{63}$$

where $\widetilde{\boldsymbol{a}}_1, \widetilde{\boldsymbol{a}}_2, \ldots, \widetilde{\boldsymbol{a}}_m$ denote independent copies of $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$. Define $\widetilde{A}$ as the matrix whose rows are $\widetilde{\boldsymbol{a}}_1^\top, \widetilde{\boldsymbol{a}}_2^\top, \ldots, \widetilde{\boldsymbol{a}}_m^\top$. Also, let $\widetilde{A}_{\setminus i}$ denote the matrix that is constructed by removing the i^{th} row of $\widetilde{A}$. Define

$$F \triangleq \begin{bmatrix} D\widetilde{A}_{\setminus 1}^\top\widetilde{A}_{\setminus 1}D & 0 & \ldots & 0\\ 0 & D\widetilde{A}_{\setminus 2}^\top\widetilde{A}_{\setminus 2}D & \ldots & 0\\ 0 & 0 & \ldots & D\widetilde{A}_{\setminus m}^\top\widetilde{A}_{\setminus m}D \end{bmatrix}.$$

and

$$\boldsymbol{v}^\top = [\boldsymbol{a}_1^\top, \boldsymbol{a}_2^\top, \ldots, \boldsymbol{a}_m^\top].$$

Using Theorem VI.4 we have

$$\begin{aligned}
&\mathbb{P}(C|\sum_{i=1}^{m}\boldsymbol{a}_i^\top D\sum_{j\neq i}\widetilde{\boldsymbol{a}}_j\widetilde{\boldsymbol{a}}_j^\top D\boldsymbol{a}_i - m(m-1)\sum_{i=1}^{n}d_i^2| > t \mid \widetilde{A})\\
&= \mathbb{P}(C|\boldsymbol{v}^\top F\boldsymbol{v} - \mathbb{E}\boldsymbol{v}^\top F\boldsymbol{v}| > t \mid \widetilde{A})\\
&\leq 2\exp\left(-c\min(\frac{t^2}{C^2\|F\|_{\mathsf{HS}}^2}, \frac{t}{C\|F\|_2})\right)
\end{aligned} \tag{64}$$

Hence, in order to obtain a more explicit upper bound, we have to find upper bounds for $\|F\|_2$ and $\|F\|_{\mathsf{HS}}^2$. First note that

$$\lambda_{\max}(F) = \max_i(\lambda_{\max}(D\widetilde{A}_{\setminus i}^\top\widetilde{A}_{\setminus i}D)) \leq \lambda_{\max}(D\widetilde{A}^\top\widetilde{A}D) \leq \|\boldsymbol{d}\|_\infty^2\lambda_{\max}(\widetilde{A}^\top\widetilde{A}). \tag{65}$$

Similarly,

$$\|F\|_{\mathsf{HS}}^2 = \sum_{i=1}^{m}\|D\widetilde{A}_{\setminus i}^\top\widetilde{A}_{\setminus i}D\|_{\mathsf{HS}}^2 \overset{(a)}{\leq} \sum_{i=1}^{m}m\lambda_{\max}^2(D\widetilde{A}_{\setminus i}^\top\widetilde{A}_{\setminus i}D) \overset{(b)}{\leq} m^2\|\boldsymbol{d}\|_\infty^4\lambda_{\max}^2(\widetilde{A}^\top\widetilde{A}), \tag{66}$$

where Inequality (a) uses the fact that the rank of matrix $D\widetilde{A}_{\setminus i}^\top\widetilde{A}_{\setminus i}D$ is $m-1$, and Inequality (b) uses (65). Finally, using Lemma VI.2 we have

$$\mathbb{P}(\sigma_{\max}(\widetilde{A}) > 2\sqrt{n} + \sqrt{m}) \leq 2e^{-\frac{n}{2}}, \tag{67}$$

and hence

$$\mathbb{P}(\lambda_{\max}(\widetilde{A}^\top A) > (2\sqrt{n} + \sqrt{m})^2) \leq 2e^{-\frac{n}{2}}. \tag{68}$$

By combining (63) and (64) we obtain

$$\mathbb{P}(|\sum_{i=1}^{m}\sum_{j\neq i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2 - m(m-1)\sum_{i=1}^{n}d_i^2| > t \mid \widetilde{A}) \leq 2C\mathbb{E}\left(\exp\left(-c\min\left(\frac{t^2}{C^2\|F\|_{\mathsf{HS}}^2}, \frac{t}{C\|F\|_2}\right)\right)\right), \tag{69}$$

where the expected value is with respect to the randomness in F or equivalently $\widetilde{A}$.

Let the event $\mathcal{E}$ denote the event of $\sigma_{\max}(\widetilde{A}) \leq 2\sqrt{n} + \sqrt{m}$, and $\mathbb{I}_{\mathcal{E}}$ denote the indicator function of the event $\mathcal{E}$. Then, using (69) we have

$$\begin{aligned}
&\mathbb{P}(|\sum_{i=1}^{m}\sum_{j\neq i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2 - m(m-1)\sum_{i=1}^{n} d_i^2| > t) \\
&= \mathbb{P}(\{|\sum_{i=1}^{m}\sum_{j\neq i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2 - m(m-1)\sum_{i=1}^{n} d_i^2| > t\} \cap \mathcal{E}) \\
&+ \mathbb{P}(\{|\sum_{i=1}^{m}\sum_{j\neq i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2 - m(m-1)\sum_{i=1}^{n} d_i^2| > t\} \cap \mathcal{E}^c) \\
&\leq \mathbb{E}\left(\mathbb{P}(|\sum_{i=1}^{m}\sum_{j\neq i}|\boldsymbol{a}_i^\top D\boldsymbol{a}_j|^2 - m(m-1)\sum_{i=1}^{n} d_i^2| > t \mid \widetilde{A})\mathbb{I}_{\mathcal{E}}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\leq 2C\mathbb{E}\left(\exp\left(-c\min\left(\frac{t^2}{C^2\|F\|_{\mathsf{HS}}^2}, \frac{t}{C\|F\|_2}\right)\right)\mathbb{I}_{\mathcal{E}}\right) + \mathbb{P}(\mathcal{E}^c) \\
&\leq 2C\exp\left(-c\min\left(\frac{t^2}{C^2\|\boldsymbol{d}\|_\infty^4 q_{m,n}}, \frac{t}{C\|\boldsymbol{d}\|_\infty^2 \widetilde{q}_{m,n}}\right)\right) + 2e^{-\frac{n}{2}}. \qquad (70)
\end{aligned}$$

□

VII. Conclusion

We have explored the theoretical and algorithmic aspects of the problem of signal recovery from multiple sets of measurements, termed as looks, amidst the presence of speckle noise. We established an upper bound on the MSE of such imaging systems, effectively capturing the MSE's dependence on the number of measurements, image complexity, and number of looks. Drawing inspiration from our theoretical framework, we introduce the bagged deep image prior (Bagged-DIP) projected gradient descent (PGD) algorithm. Through extensive experimentation, we demonstrate that our algorithm attains state-of-the-art performance.

Acknowledgements

X.C., S.Jalali and A.M. were supported in part by ONR award no. N00014-23-1-2371. S.Jalali was supported in part by NSF CCF-2237538. C.A.M. was supported in part by SAAB, Inc., AFOSR Young Investigator Program Award no. FA9550-22-1-0208, and ONR award no. N00014-23-1-2752.

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
[2] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.
[3] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, "Compressive phase retrieval," in *Wavelets XII*, vol. 6701, pp. 712–722, SPIE, 2007.
[4] S. Jalali, A. Maleki, and R. G. Baraniuk, "Minimum complexity pursuit for universal compressed sensing," vol. 60, pp. 2253–2268, Apr. 2014.
[5] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1043–1055, 2014.
[6] M. Bakhshizadeh, A. Maleki, and S. Jalali, "Using black-box compression algorithms for phase retrieval," *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7978–8001, 2020.
[7] F. Argenti, A. Lapini, T. Bianchi, and L. Alparone, "A tutorial on speckle reduction in synthetic aperture radar images," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 3, pp. 6–35, 2013.
[8] T. Bate, D. O'Keefe, M. F. Spencer, and C. J. Pellizzari, "Experimental validation of model-based digital holographic imaging using multi-shot data," in *Unconventional Imaging and Adaptive Optics 2022*, vol. 12239, pp. 83–94, SPIE, 2022.
[9] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

[10] R. Heckel and P. Hand, "Deep decoder: Concise image representations from untrained non-convolutional networks," in *International Conference on Learning Representations*, 2018.

[11] W. Zhou, S. Jalali, and A. Maleki, "Compressed sensing in the presence of speckle noise," *IEEE Transactions on Information Theory*, vol. 68, no. 10, pp. 6964–6980, 2022.

[12] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. SIAM, 1995.

[13] R. Heckel and M. Soltanolkotabi, "Denoising and regularization via exploiting the structural bias of convolutional generators," in *International Conference on Learning Representations*, 2019.

[14] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[15] G. Schulz, "Iterative berechung der reziproken matrix," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 13, no. 1, pp. 57–59, 1933.

[16] J. S. Lim and H. Nawab, "Techniques for speckle noise removal," in *Applications of speckle phenomena*, vol. 243, pp. 35–45, SPIE, 1980.

[17] L. Gagnon and A. Jouan, "Speckle filtering of sar images: a comparative study between complex-wavelet-based and standard filters," in *Wavelet Applications in Signal and Image Processing V*, vol. 3169, pp. 80–91, SPIE, 1997.

[18] Y. Tounsi, M. Kumar, A. Nassim, F. Mendoza-Santoyo, and O. Matoba, "Speckle denoising by variant nonlocal means methods," *Applied optics*, vol. 58, no. 26, pp. 7110–7120, 2019.

[19] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided cnn for image denoising," *Neural Networks*, vol. 124, pp. 117–129, 2020.

[20] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "Sunet: swin transformer unet for image denoising," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2333–2337, IEEE, 2022.

[21] C. J. Pellizzari, M. F. Spencer, and C. A. Bouman, "Phase-error estimation and image reconstruction from digital-holography data using a bayesian framework," *JOSA A*, vol. 34, no. 9, pp. 1659–1669, 2017.

[22] C. J. Pellizzari, M. F. Spencer, and C. A. Bouman, "Optically coherent image reconstruction in the presence of phase errors using advanced-prior models," in *Long-range imaging III*, vol. 10650, pp. 68–82, SPIE, 2018.

[23] C. J. Pellizzari, M. F. Spencer, and C. A. Bouman, "Coherent plug-and-play: digital holographic imaging through atmospheric turbulence using model-based iterative reconstruction and convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1607–1621, 2020.

[24] C. J. Pellizzari, T. J. Bate, K. P. Donnelly, and M. F. Spencer, "Solving coherent-imaging inverse problems using deep neural networks: an experimental demonstration," in *Unconventional Imaging and Adaptive Optics 2022*, vol. 12239, pp. 57–65, SPIE, 2022.

[25] X. Chen, Z. Hou, C. Metzler, A. Maleki, and S. Jalali, "Multilook compressive sensing in the presence of speckle noise," in *NeurIPS 2023 Workshop on Deep Learning and Inverse Problems*, 2023.

[26] G. Jagatap and C. Hegde, "Algorithmic guarantees for inverse imaging with untrained network priors," *Advances in neural information processing systems*, vol. 32, 2019.

[27] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 39–56, 2020.

[28] M. Z. Darestani and R. Heckel, "Accelerated mri with un-trained neural networks," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 724–733, 2021.

[29] S. Ravula and A. G. Dimakis, "One-dimensional deep image prior for time series inverse problems," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pp. 1005–1009, IEEE, 2022.

[30] Z. Zhuang, D. Yang, F. Hofmann, D. Barmherzig, and J. Sun, "Practical phase retrieval using double deep image priors," *arXiv preprint arXiv:2211.00799*, 2022.

[31] Z. Zhuang, T. Li, H. Wang, and J. Sun, "Blind image deblurring with unknown kernel size and substantial noise," *International Journal of Computer Vision*, pp. 1–30, 2023.

[32] G. Mataev, P. Milanfar, and M. Elad, "Deepred: Deep image prior powered by red," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[33] D. Van Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, and A. G. Dimakis, "Compressed sensing with deep image prior and learned regularization," *arXiv preprint arXiv:1806.06438*, 2018.

[34] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6360–6376, 2021.

[35] Z. Sun, F. Latorre, T. Sanchez, and V. Cevher, "A plug-and-play deep image prior," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8103–8107, IEEE, 2021.

[36] T. Li, H. Wang, Z. Zhuang, and J. Sun, "Deep random projector: Accelerated deep image prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18176–18185, 2023.

[37] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing.," 2012.

[38] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International conference on machine learning*, pp. 537–546, PMLR, 2017.

[39] P. Peng, S. Jalali, and X. Yuan, "Solving inverse problems via auto-encoders," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 312–323, 2020.

[40] B. Joshi, X. Li, Y. Plan, and O. Yilmaz, "Plugin-cs: A simple algorithm for compressive sensing with generative prior," in *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021.

[41] T. V. Nguyen, G. Jagatap, and C. Hegde, "Provable compressed sensing with generative priors via langevin dynamics," *IEEE Transactions on Information Theory*, vol. 68, no. 11, pp. 7410–7422, 2022.

[42] R. Malekian and A. Maleki, "Is speckle noise more challenging to mitigate than additive noise?," *arXiv preprint arXiv:2409.16585*, 2024.

[43] W. Zhou, S. Jalali, and A. Maleki, "Correction to" compressed sensing in the presence of speckle noise"," *IEEE Transactions on Information Theory*, 2024.

[44] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," 2009.

[45] E. J. Candes and M. A. Davenport, "How well can we estimate a sparse vector?," *Applied and Computational Harmonic Analysis*, vol. 34, no. 2, pp. 317–323, 2013.

[46] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[47] J. B. Conway, *A course in functional analysis*, vol. 96. Springer, 2019.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[49] R. M. Gower and P. Richtárik, "Randomized quasi-newton updates are linearly convergent matrix inversion algorithms," *SIAM Journal on Matrix Analysis and Applications*, vol. 38, no. 4, pp. 1380–1409, 2017.

[50] A. Stotsky, "Convergence rate improvement of richardson and newton-schulz iterations," *arXiv preprint arXiv:2008.11480*, 2020.

[51] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 449–458, 2016.

[52] M. Rudelson and R. Vershynin, "Non-asymptotic theory of random matrices: extreme singular values," in *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pp. 1576–1602, World Scientific, 2010.

[53] S. Jalali, A. Maleki, and R. G. Baraniuk, "Minimum complexity pursuit for universal compressed sensing," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2253–2268, 2014.

[54] V. De la Pena and E. Giné, *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

APPENDIX

A. Caculation of the likelihood function

The aim of this section is to derive the loglikelihood for our model,

$$\boldsymbol{y}_\ell = AX\boldsymbol{w}_\ell + \boldsymbol{z}_\ell, \qquad \text{for} \quad \ell = 1, \dots, L,$$

where $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_L$, and $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_L$ are independent and identically distributed $\mathcal{CN}(0, \sigma_w^2 I_n)$ and $\mathcal{CN}(0, \sigma_z^2 I_p)$ respectively. Since the noises are indepednet across the looks, we can write the loglikelihood for one of the looks, and then add the loglikelihoods to obtain the likelihood for all the looks. For notational simplicity, we write the measurements of one of the looks as:

$$\boldsymbol{y} = AX\boldsymbol{w} + \boldsymbol{z}$$

Note that $\boldsymbol{y}$ is a linear combination of two Gaussian random vectors and is hence Gaussian. Hence, by writing the real and imaginary parts of $\boldsymbol{y}$ seperately we will have

$$\Re(\boldsymbol{y}) + \Im(\boldsymbol{y}) = (\Re(AX) + i\Im(AX))(\boldsymbol{w}^{(1)} + i\boldsymbol{w}^{(2)}) + (\boldsymbol{z}^{(1)} + i\boldsymbol{z}^{(2)}),$$

and defining

$$\widetilde{\boldsymbol{y}} \triangleq \begin{bmatrix} \Re(\boldsymbol{y}) \\ \Im(\boldsymbol{y}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, B \right),$$

where

$$B = \begin{bmatrix} \sigma_z^2 I_n + \sigma_w^2 \Re(AX^2\bar{A}^\top) & -\sigma_w^2 \Im(AX^2\bar{A}^\top) \\ \sigma_w^2 \Im(AX^2\bar{A}^\top) & \sigma_z^2 I_n + \sigma_w^2 \Re(AX^2\bar{A}^\top) \end{bmatrix}.$$

Hence, the log-likelihood of our data $\boldsymbol{y}$ as a function of $\boldsymbol{x}$ is

$$\ell(\boldsymbol{x}) = -\frac{1}{2}\log\det(B) - \frac{1}{2}\begin{bmatrix} \Re(\boldsymbol{y}^\top) & \Im(\boldsymbol{y}^\top) \end{bmatrix}(B)^{-1}\begin{bmatrix} \Re(\boldsymbol{y}) \\ \Im(\boldsymbol{y}) \end{bmatrix} + C. \tag{71}$$

Note that equation (71) is for a single look. Hence the loglikelihood of $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_L$ as a function of $\boldsymbol{x}$ is:

$$\ell(\boldsymbol{x}) = -\frac{L}{2}\log\det(B) - \frac{1}{2}\sum_{\ell=1}^{L} \widetilde{\boldsymbol{y}}_\ell^\top B^{-1}\widetilde{\boldsymbol{y}}_\ell + C, \tag{72}$$

Since we would like to maximize $\ell(\boldsymbol{x})$ as a function of $\boldsymbol{x}$, for notational simplicty we define the cost function $f_L(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$:

$$f_L(\boldsymbol{x}) = \log\det(B) + \frac{1}{L}\sum_{\ell=1}^{L} \widetilde{\boldsymbol{y}}_\ell^\top B^{-1}\widetilde{\boldsymbol{y}}_\ell, \tag{73}$$

that we will minimize to obtain the maximum likelihood estimate.

B. Calculation of the gradient of the likelihood function

As discussed in the main text, to execute the projected gradient descent, it is necessary to compute the gradient of the negative log-likelihood function ∂f_L. The derivatives of f_L with respect to each element $\boldsymbol{x}_j$ of $\boldsymbol{x}$ is given by:

$$\begin{aligned}\frac{\partial f_L}{\partial \boldsymbol{x}_j} =& 2\boldsymbol{x}_j\sigma_w^2\left(\begin{bmatrix}\Re(\boldsymbol{a}_{\cdot,j}^\top) & \Im(\boldsymbol{a}_{\cdot,j}^\top)\end{bmatrix} B^{-1}\begin{bmatrix}\Re(\boldsymbol{a}_{\cdot,j})\\ \Im(\boldsymbol{a}_{\cdot,j})\end{bmatrix} + \begin{bmatrix}-\Im(\boldsymbol{a}_{\cdot,j}^\top) & \Re(\boldsymbol{a}_{\cdot,j}^\top)\end{bmatrix} B^{-1}\begin{bmatrix}-\Im(\boldsymbol{a}_{\cdot,j})\\ \Re(\boldsymbol{a}_{\cdot,j})\end{bmatrix}\right)\\ &- \frac{2\boldsymbol{x}_j\sigma_w^2}{L}\sum_{\ell=1}^{L}\left[\left(\begin{bmatrix}\Re(\boldsymbol{a}_{\cdot,j}^\top) & \Im(\boldsymbol{a}_{\cdot,j}^\top)\end{bmatrix} B^{-1}\begin{bmatrix}\Re(\boldsymbol{y}_\ell)\\ \Im(\boldsymbol{y}_\ell)\end{bmatrix}\right)^2 + \left(\begin{bmatrix}-\Im(\boldsymbol{a}_{\cdot,j}^\top) & \Re(\boldsymbol{a}_{\cdot,j}^\top)\end{bmatrix} B^{-1}\begin{bmatrix}\Re(\boldsymbol{y}_\ell)\\ \Im(\boldsymbol{y}_\ell)\end{bmatrix}\right)^2\right]\\ =& 2\boldsymbol{x}_j\sigma_w^2\left(\widetilde{\boldsymbol{a}}_{\cdot,j}^{+T}B^{-1}\widetilde{\boldsymbol{a}}_{\cdot,j}^{+} + \widetilde{\boldsymbol{a}}_{\cdot,j}^{-T}B^{-1}\widetilde{\boldsymbol{a}}_{\cdot,j}^{-}\right) - \frac{2\boldsymbol{x}_j\sigma_w^2}{L}\sum_{\ell=1}^{L}\left[\left(\widetilde{\boldsymbol{a}}_{\cdot,j}^{+T}B^{-1}\widetilde{\boldsymbol{y}}_\ell\right)^2 + \left(\widetilde{\boldsymbol{a}}_{\cdot,j}^{-T}B^{-1}\widetilde{\boldsymbol{y}}_\ell\right)^2\right], \end{aligned} \tag{74}$$

where $\boldsymbol{a}_{\cdot,j}$ denotes the j-th column of matrix A, $\widetilde{\boldsymbol{a}}_{\cdot,j}^{+} = \begin{bmatrix}\Re(\boldsymbol{a}_{\cdot,j})\\ \Im(\boldsymbol{a}_{\cdot,j})\end{bmatrix}$ and $\widetilde{\boldsymbol{a}}_{\cdot,j}^{-} = \begin{bmatrix}-\Im(\boldsymbol{a}_{\cdot,j})\\ \Re(\boldsymbol{a}_{\cdot,j})\end{bmatrix}$.

C. More simplification of the gradient

The special form of the matrix B enables us to do the calculations more efficiently. To see this point, define:

$$U + iV \triangleq \left(\sigma_z^2 I_n + \sigma_w^2 A X^2 \bar{A}^\top\right)^{-1},$$

where $U, V \in R^{m\times m}$. These two matrices should satisfy:

$$\begin{aligned}\left(\sigma_z^2 I_n + \sigma_w^2\Re(AX^2\bar{A}^\top)\right)U - \sigma_w^2\Im(AX^2\bar{A}^\top)V = I_n\\ \sigma_w^2\Im(AX^2\bar{A}^\top)U + \left(\sigma_z^2 I_n + \sigma_w^2\Re(AX^2\bar{A}^\top)\right)V = 0.\end{aligned}$$

These two equations imply that:

$$B^{-1} = \begin{bmatrix}U & -V\\ V & U\end{bmatrix}. \tag{75}$$

This simple observation, enables us to reduce the number of multiplications required for the Newton-Schulz algorithm. More specifically, instead of requiring to multiply two $2m \times 2m$ matrices, we can do 4 multiplications of $m \times m$ matrices. This helps us have a factor of 2 reduction in the cost of matrix-matrix multiplication in our Newton-Schulz algorithm.

In cases the exact inverse calculation is required, again this property enables us to reduce the inversion of matrix $B \in \mathbb{R}^{2m\times 2m}$ to the inversion of two $m \times m$ matrices (albeit a few $m \times m$ matrix multiplications are required as well).

Plugging (75) into (74), we obtain a simplified form for the gradient of $f_L(\boldsymbol{x})$:

$$\begin{aligned}\frac{\partial f_L}{\partial \boldsymbol{x}_j} =& 4\boldsymbol{x}_j\sigma_w^2\Re\left(\bar{\boldsymbol{a}}_{\cdot,j}^\top(U+iV)\boldsymbol{a}_{\cdot,j}\right) - \frac{2\boldsymbol{x}_j\sigma_w^2}{L}\sum_{\ell=1}^{L}\left[\Re^2\left(\bar{\boldsymbol{a}}_{\cdot,j}^\top(U+iV)\boldsymbol{y}_\ell\right) + \Im^2\left(\bar{\boldsymbol{a}}_{\cdot,j}^\top(U+iV)\boldsymbol{y}_\ell\right)\right]\\ =& 4\boldsymbol{x}_j\sigma_w^2\Re\left(\bar{\boldsymbol{a}}_{\cdot,j}^\top(U+iV)\boldsymbol{a}_{\cdot,j}\right) - \frac{2\boldsymbol{x}_j\sigma_w^2}{L}\sum_{\ell=1}^{L}\left\|\bar{\boldsymbol{a}}_{\cdot,j}^\top(U+iV)\boldsymbol{y}_\ell\right\|_2^2. \end{aligned} \tag{76}$$

Algorithm 1 shows a detailed version of the final algorithm we execute for recovering images from their multilook, speckle-corrupted, undersampled measurements. In one of the steps of the algorithm, we ensure that all the pixel values of our estimate are within the range $[0, 1]$. This is because we have assumed that the image pixels take values within $[0, 1]$.

Algorithm 1 Projected gradient descent algorithm

Input: $\{\mathbf{y}_l\}_{l=1}^L, A, \boldsymbol{x}_0 = \frac{1}{L}\sum_{l=1}^L |A^\top \mathbf{y}_l|, g_\theta(\cdot, \delta_{\boldsymbol{x}})$.
Output: Reconstructed $\widehat{\boldsymbol{x}}$.
for $t = 1, \ \ldots, \ T$ **do**
  [Gradient Descent Step]
  if $t = 1$ **or** $\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_\infty > \delta_{\boldsymbol{x}}$ **then**
    Calculate exact $B_t^{-1} = (AX_t^2 A^\top)^{-1}$.
  else
    [Newton-Schulz matrix inverse approximation]
    $M^0 = B_{t-1}^{-1}$,
    for $k = 1, \ \ldots, \ K$ **do**
      $M^k = M^{k-1} + M^{k-1}(I_m - B_t M^{k-1})$,
    end for
    $\widetilde{B}_t^{-1} = M^K$.
  end if
  Gradient calculation at coordinate j as $\nabla f_L(\boldsymbol{x}_{t-1,j})$ using B_t^{-1} or $\widetilde{B}_t^{-1}$, and update $\boldsymbol{x}_{t,j}^G$: $\boldsymbol{x}_{t,j}^G \leftarrow \boldsymbol{x}_{t-1,j} - \mu_t \nabla f_L(\boldsymbol{x}_{t-1,j})$.
  Save matrix inverse B_t^{-1} or $\widetilde{B}_t^{-1}$.
  Truncate $\boldsymbol{x}_t^G$ into range $(0, 1)$, $\boldsymbol{x}_t^G = \text{clip}(\boldsymbol{x}_t^G, 0, 1)$.
  [Bagged-DIPs Projection Step]
  Generate random image given randomly generated noise $\boldsymbol{u} \sim \mathcal{N}(0, 1)$ as $g_\theta(\boldsymbol{u})$.
  Update θ_t by optimizing over $\|g_\theta(\boldsymbol{u}) - \boldsymbol{x}_t^G\|_2^2$ till converges: $\widehat{\theta}_t \leftarrow \text{argmin}_\theta \|g_\theta(\boldsymbol{u}) - \boldsymbol{x}_t^G\|_2^2$.
  Generate $\boldsymbol{x}_t^P$ using trained $g_{\widehat{\theta}_t}(\cdot)$ as $\boldsymbol{x}_t^P \leftarrow g_{\widehat{\theta}_t}(\boldsymbol{u})$.
  Obtain $\boldsymbol{x}_t = \boldsymbol{x}_t^P$.
end for
Reconstruct image as $\widehat{\boldsymbol{x}} = \boldsymbol{x}_T$.

The only remaining parts of the algorithm we need to clarify are (1) our hyperparameter choices and (2) the implementation details of the Bagged-DIP module. As described in the main text, in each (outer) iteration of PGD, we learn three DIPs and then take the average of their outputs. Let us now consider one of these DIPs that is applied to one of the $h_k \times w_k$ patches.

Inspired by the deep decoder paper [10], we construct our neural network using four blocks: we call the first three blocks DIP-blocks and the last one output block. The structures of the blocks are shown in Figure 4. As is clear from the figure, each DIP block is composed of the following components:

- Up sample: This unit increases the height and width of the datacube that receives by a factor of 2. To interpolate the missing elements, it uses the simple bilinear interpolation. Hence, if the size of the image is 128×128, then the height and width of the input to DIP-block3 will be 64×64, the input of DIP-Block2 will be 32×32, and so on.
- ReLU: this module is quite standard and does not require further explanation.
- Convolution: For all our simulations, we have either used 1×1 or 3×3 convolutions. Additionally, we provide details on the number of channels for the data cubes entering each block in our simulations. The channel numbers are [128, 128, 128, 128] for the four blocks.

The output block is simpler than the other three blocks. It only has a 2D convolution that uses the same size as the convolutions of the other DIP blocks. The nonlinearity used here is a sigmoid function, as we assume that the pixel values are between $[0, 1]$.

Finally, we should mention that each element of the input noise $\boldsymbol{u}$ of DIP (as described before DIP function is $g_{\boldsymbol{\theta}}(\boldsymbol{u})$) is generated independently from Normal distribution $\mathcal{N}(0, 1)$.

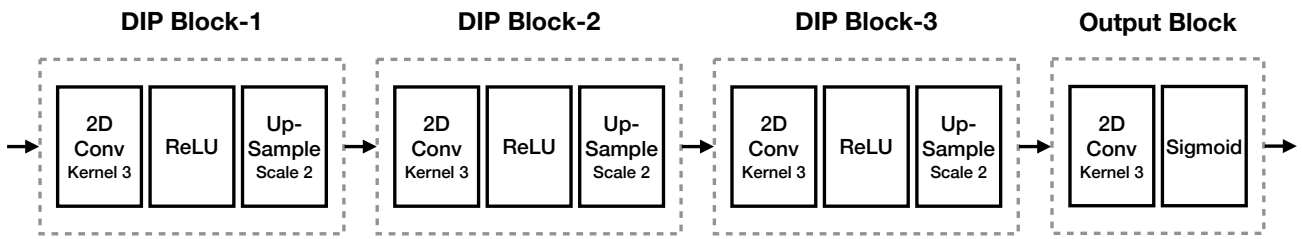

Fig. 4. The structure of DIP and Output Blocks.

TABLE III
NUMBER OF ITERATIONS USED IN TRAINING BAGGED-DIPS FOR DIFFERENT ESTIMATES.

Patch size	Barbara	Peppers	House	Foreman	Boats	Parrots	Cameraman	Monarch
128	400	400	400	400	400	800	4000	800
64	300	300	300	300	300	600	2000	600
32	200	200	200	200	200	400	1000	400

The other hyperparameters that are used in the DIP-based PGD algorithm are set in the following way: The learning rate of the loglikelihood gradient descent step (in PGD) is set to $\mu = 0.01$. For training the Bagged-DIPs, we use Adam [48] with the learning rate set to 0.001 and weight decay set to 0. The number of iterations used for training Bagged-DIPs for different estimates on images are mentioned in Table III. We run the outer loop (gradient descent of likelihood) for $100, 200, 300$ iterations when $m/n = 0.5, 0.25, 0.125$ respectively. For "Cameraman" only, when $m/n = 0.125$, since the convergence rate is slow, we run 800 outer iterations.

The Newton-Schulz algorithm, utilized for approximating the inverse of the matrix B_t, has a quadratic convergence when the maximum singular value $\sigma_{\max}(I - M^0 B_t) < 1$. Hence, ideally, if this condition does not hold, we do not want to use the Newton-Schulz algorithm and may prefer the exact inversion. Unfortunately, checking the condition $\sigma_{\max}(I - M^0 B_t) < 1$ is also computationally demanding. However, the special form of B_t enables us to have an easier heuristic evaluation of this condition.

For our problems, we establish an empirical sufficient condition for convergence: $\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_\infty < \delta_{\boldsymbol{x}}$, where $\delta_{\boldsymbol{x}}$ is a predetermined constant. To determine the most robust value for $\delta_{\boldsymbol{x}}$, we conducted simple experiments. We set $n = 128 \times 128$ and $m/n = 0.5$. The sensing matrix A is generated as described in the main part of the paper (see Section V-B). Each element of $\boldsymbol{x}_o$ is independently drawn from a uniform distribution $U[0.001, 1]$. Furthermore, each element of $\Delta \boldsymbol{x}_o$ is independently sampled from a two-point distribution. In this distribution, the probability of the variable X being $\delta_{\boldsymbol{x}}$ is equal to the probability of X being $-\delta_{\boldsymbol{x}}$, both with a probability of 0.5, ensuring $\|\Delta \boldsymbol{x}_o\|_\infty = \delta_{\boldsymbol{x}}$. We define B as $A(X + \Delta X_o)^2 \bar{A}^\top$, and M^0 as $(AX^2\bar{A}^\top)^{-1}$. We then assess the convergence of the Newton-Schulz algorithm for calculating B^{-1}. For various values of $\delta_{\boldsymbol{x}}$, we ran the simulation 100 times each, recording the convergence success rate. As indicated in Table IV, the algorithm demonstrates instability when $\delta_{\boldsymbol{x}} \geq 0.13$. Consequently, we set $\delta_{\boldsymbol{x}}$ to 0.12 in all our simulations to ensure the reliable convergence of the Newton-Schulz algorithm.

TABLE IV
CONVERGENCE SUCCESS RATE UNDER VARYING THRESHOLD $\delta_{\boldsymbol{x}}$.

Threshold $\delta_{\boldsymbol{x}}$	Convergence Success Rate
0.10	100%
0.11	100%
0.12	100%
0.13	38%
0.14	0%
0.15	0%