Convergence Guarantees of Overparametrized Wide Deep Inverse Prior

Nathan Buskulic, Yvain Quéau, and Jalal Fadili

Normandie Univ., UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France

Abstract. Neural networks have become a prominent approach to solve inverse problems in recent years. Amongst the different existing methods, the Deep Image/Inverse Priors (DIPs) technique is an unsupervised approach that optimizes a highly overparametrized neural network to transform a random input into an object whose image under the forward model matches the observation. However, the level of overparametrization necessary for such methods remains an open problem. In this work, we aim to investigate this question for a two-layers neural network with a smooth activation function. We provide overparametrization bounds under which such network trained via continuous-time gradient descent will converge exponentially fast with high probability, from which recovery prediction bounds will be derived. This work is thus a first step towards a theoretical understanding of overparametrized DIP networks, and more broadly it participates to the theoretical understanding of neural networks in inverse problem settings.

Keywords: Inverse problems · Deep Image/Inverse Prior · Overparameterization · Gradient Flow.

1 Introduction

1.1 Problem Statement

A linear inverse problem consists in reliably recovering an object $\overline{\mathbf{x}} \in \mathbb{R}^n$ from noisy indirect observations

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \epsilon, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^m$ is the observation, $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a linear forward operator, and ϵ stands for some additive noise. We will denote $\overline{\mathbf{y}} = \mathbf{A}\overline{\mathbf{x}}$. Without loss of generality, we will assume throughout that $\mathbf{y} \in \text{Im}(\mathbf{A})$.

In recent years, the use of sophisticated machine learning algorithms, including deep learning, to solve inverse problems has gained a lot of momentum and provides promising results, see e.g., reviews [2, 12]. Most of these methods are supervised and require extensive datasets for training, which might not be available. An interesting unsupervised alternative [18] is known as Deep Image Prior, which is also named Deep Inverse Prior (DIP) as it is not confined to images. In the DIP framework, a generator network $\mathbf{g} : (\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbf{x} \in \mathbb{R}^n$, with

activation function ϕ, is optimized to transform some random input $\mathbf{u} \in \mathbb{R}^d$ into a vector in $\mathbf{x} \in \mathbb{R}^n$. The parameters $\boldsymbol{\theta}$ of the network are optimized via (possibly stochastic) gradient descent to minimize the squared Euclidean loss

$$\mathcal{L}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})) = \frac{1}{2m} \|\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) - \mathbf{y}\|^2 . \tag{2}$$

Theoretical understanding of recovery and convergence guarantees for deep learning-based methods is of paramount importance to make their routine usage in critical applications reliable [11]. Our goal in this paper is to participate to this endeavour by explaining when gradient descent consistently and provably finds global minima of (2), and how this translates into recovery guarantees of (1). For this, we focus on a continuous-time gradient flow applied to (2):

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))), \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases} \tag{3}$$

This is an idealistic setting which makes the presentation simpler and it is expected to reflect the behavior of practical and commonly encountered first-order descent algorithms, as they are known to approximate gradient flows.

1.2   Contributions

We will deliver a first theoretical analysis of DIP models in the overparametrized regime. We will first analyze (3) by providing sufficient conditions for $\mathbf{y}(t) := \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$ to converge exponentially fast to a globally optimal solution in the observation space. This result is then converted to a prediction error on $\overline{\mathbf{y}}$ through an early stopping strategy. Our conditions and bounds involve the conditioning of the forward operator, the minimum and maximum singular values of the Jacobian of the network, as well as its Lipschitz constant. We will then turn to evaluating these quantities for the case of a two-layer neural network

$$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) = \frac{1}{\sqrt{k}} \mathbf{V} \phi(\mathbf{W}\mathbf{u}), \tag{4}$$

with $\mathbf{V} \in \mathbb{R}^{n \times k}$ and $\mathbf{W} \times \mathbb{R}^{k \times d}$, and ϕ an element-wise nonlinear activation function. The scaling by $\sqrt{k}$ will become clearer later. In this context, the network will be optimized with respect to the first layer (i.e., $\mathbf{W}$) while keeping the second (i.e., $\mathbf{V}$) fixed. Consequently, $\boldsymbol{\theta} = \mathbf{W}$. We show that for a proper random initialization $\mathbf{W}(0)$ and sufficient overparametrization, all our conditions are in force and the smallest eigenvalue of the Jacobian is indeed bounded away from zero independently of time. We provide a characterization of the overparametrization needed in terms of (k, d, n) and the conditioning of $\mathbf{A}$. Lastly, we show empirically that the behavior of real-world DIP networks is consistent with our theoretical bounds.

1.3   Relation to Prior Work

Data-Driven Methods to Solve Inverse Problems Data-driven approaches
to solve inverse problems come in various forms [2, 12]. The first type trains an
end-to-end network to directly map the observations to the signals for a specific
problem. While they can provide impressive results, these methods can prove
very unstable as they do not use the physics of the problem which can be severely
ill-posed. To cope with these problems, several hybrid models that mix model-
and data-driven algorithms were developed in various ways. One can learn the
regularizer of a variational problem [15] or use Plug-and-Play methods [19] for
example. Another family of approaches, which takes inspiration from classical
iterative optimization algorithms, is based on unrolling (see [10] for a review of
these methods). Still, all these methods requires extensive amount of training
data, which may not always be available. Their theoretical recovery guarantees
are also not well understood [11].

Deep Inverse Prior The DIP model [18] (and its extensions that mitigate
some of its empirical issues [8, 9, 16, 20]) is an unsupervised alternative to
the supervised approches briefly reviewed above. The empirical idea is that the
architecture of the network acts as an implicit regularizer and will learn a more
meaningful transformation before overfitting to artefacts or noise. With an early
stopping strategy, one can get the network to generate a vector close to the
sought signal. However, this remains purely empirical and there is no guarantee
that a network trained in such manner converges in the observation space (and
even less in the signal space). Our work aims at reducing this theoretical gap,
by analyzing the behaviour of the network in the observation (prediction) space.

Theory of Overparametrized Networks In parallel to empirical studies,
there has been a lot of effort to develop some theoretical understanding of the
optimization of overparametrized networks [3, 6]. Amongst the theoretical mod-
els that emerged to analyze neural networks, the Neural Tangent Kernel (NTK)
captures the behavior of neural networks in the infinite width limit during op-
timization via gradient descent. In the NTK framework, the neural network be-
haves as its linearization around the initialization, thus yielding a model equiva-
lent to learning with a specific positive-definite kernel (so called NTK). In [7], it
was shown that in a highly overparametrized regime and random initialization,
parameters $\boldsymbol{\theta}(t)$ stay near the initialization, and are well approximated by their
linearized counterparts at all times (also called the "lazy" regime in [4]). With
a similar aim, several works characterized the overparametrization necessary to
obtain similar behaviour for shallow networks for different tasks, see e.g., [1, 5,
13, 14]. All these works provide lower bounds on the number of neurons from
which they can prove convergence rates to a zero-loss solution. Despite some
apparent similarities, our setting has important differences. On the one hand,
we have indirect measurements through (fixed) $\mathbf{A}$, the output is not scalar, and
there is no supervision. On the other hand, unlike all above works which deal
with a supervised training setting, in the DIP model the dimension d of the
input is a free parameter, while it is imposed in a supervised setting.

2   DIP Guarantees

2.1   Notations

For a matrix $\mathbf{M} \in \mathbb{R}^{a \times b}$ we denote, when dimension requirements are met, by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ (resp. $\sigma_{\min}(\mathbf{M})$ and $\sigma_{\max}(\mathbf{M})$) its smallest and largest eigenvalues (resp. non-zero singular values), and by $\kappa(\mathbf{M}) = \frac{\sigma_{\max}(\mathbf{M})}{\sigma_{\min}(\mathbf{M})}$ its condition number. We also denote by $\|\cdot\|_F$ the Frobenius norm and $\|\cdot\|$ the Euclidean norm of a vector (or operator norm of a matrix). We use $\mathbf{M}^i$ (resp. $\mathbf{M}_i$) as the i-th row (resp. column) of $\mathbf{M}$. We represent a ball of radius r and center x by $\mathbb{B}(x, r)$. We also define $\mathbf{y}(t) = \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$. The Jacobian of the network is denoted $\mathcal{J}(\boldsymbol{\theta}(t))$. The Lipschitz constant of a mapping is denoted $\mathrm{Lip}(\cdot)$. We set $C_\phi = \sqrt{\mathbb{E}_{g \sim \mathcal{N}(0,1)}[\phi(g)^2]}$ and $C_{\phi'} = \sqrt{\mathbb{E}_{g \sim \mathcal{N}(0,1)}[\phi'(g)^2]}$ with $\mathbb{E}[X]$ the expected value of X.

2.2   Main Result

Standing Assumptions In the rest of this work, we assume that:

> **A-1** $\mathbf{u}$ *is drawn uniformly on* $\mathbb{S}^{d-1}$;
>
> **A-2** $\mathbf{W}(0)$ *has iid entries from* $\mathcal{N}(0, 1)$;
>
> **A-3** $\mathbf{V}$ *has iid columns with identity covariance and D-bounded entries*;
>
> **A-4** ϕ *is a twice differentiable function with B-bounded derivatives*.

Assumptions A-1, A-2 and A-3 are standard. Assumptions A-4 is met by many activations such as the softmax, sigmoid or hyperbolic tangent. Including the ReLU would require more technicalities that will be avoided here.

Well-posedness In order for our analysis to hold, the Cauchy problem (3) needs to be well-defined. This is easy to prove upon observing that under (A-4), the gradient of the loss is both Lipschitz and continuous. Thus, the Cauchy-Lipschitz theorem applies, ensuring that (3) has a unique global continuously differentiable solution trajectory.

Our main result establishes the prediction error for the DIP model.

Theorem 1. *Consider a network* $\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$*, with* ϕ *obeying* (A-4)*, optimized via* (3)*.*

(i) Let $\sigma_{\mathbf{A}} = \inf_{\mathbf{z} \in \mathrm{Im}(\mathbf{A})} \|\mathbf{A}^\top \mathbf{z}\| / \|\mathbf{z}\| > 0$*. Suppose that*

$$\frac{\|\mathbf{y} - \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_0)\|}{\sigma_{\mathbf{A}}} < \frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))^2}{4\mathrm{Lip}(\mathcal{J})}. \tag{5}$$

Then for any $\epsilon > 0$

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \le 2\|\epsilon\| \qquad \textit{for all} \qquad t \ge \frac{4m \log(\|\mathbf{y} - \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_0)\| / \|\epsilon\|)}{\sigma_{\mathbf{A}}^2 \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))^2}. \tag{6}$$

(ii) Let the one-hidden layer network (4) *with architecture parameters obeying*

$$k \geq C_1 \kappa(\mathbf{A})^2 n \left(\sqrt{n} \left(\sqrt{\log(d)} + 1 \right) + \sqrt{m} \right)^2$$

Then

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2 \|\epsilon\| \qquad \text{for all} \qquad t \geq \frac{C_2 m \log \left(\|\mathbf{y} - \mathbf{A}\mathbf{g}(\mathbf{u}, \mathbf{W}(0))\| \right)}{\sigma_{\mathbf{A}}^2 C_{\phi'}^2}$$

with probability at least $1 - n^{-1} - d^{-1}$, *where* C_i *are positive constants that depend only on the activation function and the bound* D.

Before proceeding with the proof, a few remarks are in order.

Remark 1.   • We start with the scaling of the network architecture parameters required. First, the bound on k, the number of neurons of the hidden layer, scales quadratically in n^2 and linearly in m. We thus have the bound $k \gtrsim n^2 m$. The probability of success in our theorem is also dependent on the architecture parameters. More precisely, this probability increases with growing number of observations.
• The other scaling of the theorem is on the input size d and informs us that its influence is logarithmic. The bound is more demanding as $\mathbf{A}$ becomes more ill-conditioned. The latter dependency can be interpreted as follows: the more ill-conditioned the original problem is, the larger the network needs to be. Let us emphasize that, contrary to other learning settings in the literature, the size d of the random input $\mathbf{u}$ is free, and so far it has remained unclear how to choose it. Our result provides a first answer for shallow networks in the overparametrized setting: most of the overparametrization necessary for the optimization to converge is due to k.
• On our way to prove (1), we actually show that $\mathbf{y}(t)$ converges exponentially to $\mathbf{y}$, which is converted to a recovery of $\overline{\mathbf{y}}$ through an early stopping strategy. This ensures that the network does not overfit the noise and provides a solution in a ball around $\overline{\mathbf{y}}$ whose radius is linear in the noise level (so-called prediction linear convergence in the inverse problem literature). This provides a first result on convergence of wide DIP networks that ensures they behave well in the observation space.
• One has to keep in mind, however, that Theorem 1 does not say anything about the recovered vector generated by the network and its relation to $\overline{\mathbf{x}}$ (in absence of noise and at convergence, it might be any element of $\overline{\mathbf{x}} + \ker(\mathbf{A})$). Of course, when $\mathbf{A}$ is invertible, then we are done. In the general case, this is a much more challenging question which requires a more involved analysis and a restricted-type injectivity assumption. This will be the subject of a forthcoming paper.

3   Proof

The proof consists of two main steps. First, we prove that under (5), $\mathbf{y}(t)$ converges exponentially fast with a time-independent rate. We then use a triangle

inequality and an early stopping criterion to show our result. The proof of the second claim will consist in verifying that (5) holds with high probability for our random model of the two-layer network under our scaling.

(i) Since the solution trajectory $\boldsymbol{\theta}(t)$, hence $\mathbf{y}(t)$, is continuously differentiable, and thus

$$
\begin{aligned}
\frac{\mathrm{d}\frac{1}{2}\|\mathbf{y}(t)-\mathbf{y}\|^2}{\mathrm{d}t} &= (\mathbf{y}(t)-\mathbf{y})\dot{\mathbf{y}}(t) \\
&= (\mathbf{y}(t)-\mathbf{y})\mathbf{A}\mathcal{J}(\boldsymbol{\theta}(t))\dot{\boldsymbol{\theta}}(t) \\
&= -(\mathbf{y}(t)-\mathbf{y})\mathbf{A}\mathcal{J}(\boldsymbol{\theta}(t))\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t))) \\
&= -\frac{1}{m}(\mathbf{y}(t)-\mathbf{y})^{\top}\mathbf{A}\mathcal{J}(\boldsymbol{\theta}(t))\mathcal{J}(\boldsymbol{\theta}(t))^{\top}\mathbf{A}^{\top}(\mathbf{y}(t)-\mathbf{y}) \\
&= -\frac{1}{m}\left\|\mathcal{J}(\boldsymbol{\theta}(t))^{\top}\mathbf{A}^{\top}(\mathbf{y}(t)-\mathbf{y})\right\|^2 \le -\frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}(t)))^2\sigma_{\mathbf{A}}^2}{m}\|\mathbf{y}(t)-\mathbf{y}\|^2 .
\end{aligned}
\tag{7}
$$

where we used that $\mathbf{y}(t)-\mathbf{y} \in \mathrm{Im}(\mathbf{A})$. In view of Lemma 1(iii), we have $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}(t))) \ge \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))/2$ for all $t \ge 0$ if the initialization error verifies (5), and in turn

$$
\frac{\mathrm{d}\|\mathbf{y}(t)-\mathbf{y}\|^2}{\mathrm{d}t} \le -\frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2}{2m}\|\mathbf{y}(t)-\mathbf{y}\|^2 .
$$

Integrating, we obtain

$$
\|\mathbf{y}(t)-\mathbf{y}\| \le \|\mathbf{y}(0)-\mathbf{y}\|\, e^{-\frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2}{4m}t}.
\tag{8}
$$

Using

$$
\|\mathbf{y}(t)-\mathbf{y}\| \le \|\mathbf{y}(t)-\overline{\mathbf{y}}\| + \|\epsilon\| \le \|\mathbf{y}(0)-\mathbf{y}\|\, e^{-\frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2}{4m}t} + \|\epsilon\| ,
$$

we get the early stopping bound by bounding the exponential term by $\|\epsilon\|$.

(ii) To show the statement, it is sufficient to check that (5) holds under our scaling. From Lemma 2, we have

$$
\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \ge C_{\phi'}/2
$$

with probability at least $1 - n^{-1}$ provided $k \ge C_0 n \log(n)$ for $C_0 > 0$. Combining this with Lemma 3 and Lemma 4 and the union bound, it is sufficient for (5) to be fulfilled with probability at least $1 - n^{-1} - d^{-1}$, that

$$
C_1\kappa(\mathbf{A})\left(\sqrt{n}\left(\sqrt{\log(d)}+1\right)+\sqrt{m}\right) < \frac{C_{\phi'}^2\sqrt{k}}{16BD\sqrt{n}}.
$$

We now prove the main intermediate lemmas invoked in the proof.

Lemma 1. *(i) If $\boldsymbol{\theta} \in \mathbb{B}(\boldsymbol{\theta}_0, R)$ with $R = \frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))}{2\mathrm{Lip}(\mathcal{J})}$, then*

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta})) \geq \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))/2.$$

(ii) If for all $s \in [0, t]$, $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}(s))) \geq \frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)}{2}$, then

$$\boldsymbol{\theta}(t) \in \mathbb{B}(\boldsymbol{\theta}_0, R') \qquad with \qquad R' = \frac{2}{\sigma_{\mathbf{A}}\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))} \|\mathbf{y}(0) - \mathbf{y}\|.$$

(iii) If $R' < R$, then for all $t \geq 0$, $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}(t))) \geq \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))/2$.

Proof. (i) Since $\boldsymbol{\theta} \in \mathbb{B}(\boldsymbol{\theta}_0, R)$, we have

$$\|\mathcal{J}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta}_0)\| \leq \mathrm{Lip}(\mathcal{J}) \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \mathrm{Lip}(\mathcal{J})R.$$

By using that $\sigma_{\min}(A)$ is 1-Lipschitz, we obtain

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta})) \geq \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) - \|\mathcal{J}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta}_0)\| \geq \frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)}{2}.$$

(ii) From (7), we have for all $s \in [0, t]$

$$\begin{aligned}
\frac{\mathrm{d}\|\mathbf{y}(s) - \mathbf{y}\|}{\mathrm{d}s} &= -\frac{1}{m}\frac{\left\|\mathcal{J}(\boldsymbol{\theta}(s))^\top \mathbf{A}^\top (\mathbf{y}(s) - \mathbf{y})\right\|^2}{\|\mathbf{y}(s) - \mathbf{y}\|} \\
&\leq -\frac{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{2m}\left\|\mathcal{J}(\boldsymbol{\theta}(s))^\top \mathbf{A}^\top (\mathbf{y}(s) - \mathbf{y})\right\|.
\end{aligned}$$

The Cauchy-Schwarz inequality and (3) imply that

$$\frac{\mathrm{d}\|\boldsymbol{\theta}(s) - \boldsymbol{\theta}_0\|}{\mathrm{d}s} = \frac{\dot{\boldsymbol{\theta}}(s)^\top (\boldsymbol{\theta}(s) - \boldsymbol{\theta}_0)}{\|\boldsymbol{\theta}(s) - \boldsymbol{\theta}_0\|} \leq \left\|\dot{\boldsymbol{\theta}}(s)\right\| = \frac{1}{m}\left\|\mathcal{J}(\boldsymbol{\theta}(s))^\top \mathbf{A}^\top (\mathbf{y}(s) - \mathbf{y})\right\|.$$

We therefore get

$$\frac{\mathrm{d}\|\boldsymbol{\theta}(s) - \boldsymbol{\theta}_0\|}{\mathrm{d}s} + \frac{2}{\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}\frac{\mathrm{d}\|\mathbf{y}(s) - \mathbf{y}\|}{\mathrm{d}s} \leq 0.$$

Integration over $s \in [0, t]$, we get the claim.

(iii) Actually, we prove the stronger statement that $\boldsymbol{\theta}(t) \in \mathbb{B}(\boldsymbol{\theta}_0, R')$ for all $t \geq 0$, whence our claim will follow thanks to (i). Let us assume for contradiction that $R' < R$ and $\exists\, t < +\infty$ such that $\boldsymbol{\theta}(t) \notin \mathbb{B}(\boldsymbol{\theta}_0, R')$. By (ii), this means that $\exists\, s \leq t$ such that $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}(s))) < \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))/2$. In turn, (i) implies that $\boldsymbol{\theta}(s) \notin \mathbb{B}(\boldsymbol{\theta}_0, R)$. Let us define

$$t_0 = \inf\{\tau \geq 0 : \boldsymbol{\theta}(\tau) \notin \mathbb{B}(\boldsymbol{\theta}_0, R)\},$$

which is well-defined as it is at most s. Thus, for any small $\epsilon > 0$ and for all $t' \leq t_0 - \epsilon$, $\boldsymbol{\theta}(t') \in \mathbb{B}(\boldsymbol{\theta}_0, R)$ which, in view of (i) entails that $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta})(t')) \geq \sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))/2$. In turn, we get from (ii) that $\boldsymbol{\theta}(t_0 - \epsilon) \in \mathbb{B}(\boldsymbol{\theta}_0, R')$. Since ϵ is arbitrary and $\boldsymbol{\theta}$ is continuous, we pass to the limit as $\epsilon \to 0$ to deduce that $\boldsymbol{\theta}(t_0) \in \mathbb{B}(\boldsymbol{\theta}_0, R') \subsetneq \mathbb{B}(\boldsymbol{\theta}_0, R)$ hence contradicting the definition of t_0. $\qquad\square$

Lemma 2 (Bound on $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0))$). *For the one-hidden layer network* (4), *under assumptions* (A-1)-(A-4). *We have*

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \geq C_{\phi'}/2$$

with probability at least $1 - n^{-1}$ provided $k \geq Cn\log(n)$ for $C > 0$ large enough that depends only on ϕ and the bound on the entries of $\mathbf{V}$.

Proof. Define the matrix $\mathbf{H} = \mathcal{J}(\boldsymbol{\theta}_0)\mathcal{J}(\boldsymbol{\theta}_0)^\top$. For the two-layer network, and since $\mathbf{u}$ is on the unit sphere, $\mathbf{H}$ reads

$$\mathbf{H} = \frac{1}{k}\sum_{i=1}^{k}\phi'(\mathbf{W}^i(0)\mathbf{u})^2\mathbf{V}_i\mathbf{V}_i^\top.$$

It follows that

$$\mathbb{E}\left[\mathbf{H}\right] = \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\phi'(g)^2\right]\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\mathbf{V}_i\mathbf{V}_i^\top\right] = C_{\phi'}^2\mathbf{I}_n,$$

where we used A-1-A-2 and orthogonal invariance of the Gaussian distribution, hence $\mathbf{W}^i(0)\mathbf{u}$ are iid $\mathcal{N}(0,1)$, as well as A-3 and independence between $\mathbf{V}$ and $\mathbf{W}(0)$. Moreover,

$$\lambda_{\max}(\phi'(\mathbf{W}^i(0)\mathbf{u})^2\mathbf{V}_i\mathbf{V}_i^\top) \leq B^2D^2n.$$

We can then apply the matrix Chernoff inequality [17, Theorem 5.1.1] to get

$$\mathbb{P}\left(\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \leq \delta C_{\phi'}\right) \leq ne^{-\frac{(1-\delta)^2kC_{\phi'}^2}{2B^2D^2n}}.$$

Taking $\delta = 1/2$ and k as prescribed, we conclude. $\qquad\square$

Lemma 3 (Lipschitz constant of the Jacobian). *For the one-hidden layer network* (4), *under assumptions* (A-1), (A-2) *and* (A-4), *we have*

$$\mathrm{Lip}(\mathcal{J}) \leq BD\sqrt{\frac{n}{k}}.$$

Proof. We have for all $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathbb{R}^{k\times d}$,

$$\left\|\mathcal{J}(\mathbf{W}) - \mathcal{J}(\widetilde{\mathbf{W}})\right\|^2 \leq \frac{1}{k}\sum_{i=1}^{k}|\phi'(\mathbf{W}^i\mathbf{u}) - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})|^2\left\|\mathbf{V}_i\mathbf{u}^\top\right\|_F^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}|\phi'(\mathbf{W}^i\mathbf{u}) - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})|^2\left\|\mathbf{V}_i\right\|^2$$

$$\leq B^2D^2\frac{n}{k}\sum_{i=1}^{k}|\mathbf{W}^i\mathbf{u} - \widetilde{\mathbf{W}}^i\mathbf{u}|^2$$

$$\leq B^2D^2\frac{n}{k}\sum_{i=1}^{k}\left\|\mathbf{W}^i - \widetilde{\mathbf{W}}^i\right\|^2 = B^2D^2\frac{n}{k}\left\|\mathbf{W} - \widetilde{\mathbf{W}}\right\|_F^2.$$

$$\square$$

Lemma 4 (Bound on the initial error). *Under the main assumptions, the initial error of the network is bounded by*

$$\|\mathbf{y}(0) - \mathbf{y}\| \leq \|\mathbf{A}\| \left(C\sqrt{n\log(d)} + \sqrt{n}\,\|\mathbf{x}_0\|_\infty + \sqrt{m}\,\|\epsilon\|_\infty \right),$$

with probability at least $1 - d^{-1}$.

Proof. We first observe that

$$\|\mathbf{y}(0) - \mathbf{y}\| \leq \|\mathbf{A}\|\,\|\mathbf{g}(\mathbf{u}, \mathbf{W}(0))\| + \|\mathbf{A}\| \left(\sqrt{n}\,\|\mathbf{x}_0\|_\infty + \sqrt{m}\,\|\epsilon\|_\infty \right),$$

where $\mathbf{g}(\mathbf{u}, \mathbf{W}(0)) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \phi(\mathbf{W}^i \mathbf{u}) \mathbf{V}_i$. We now prove that this term concentrates around its expectation. First, we have by independence

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{u}, \mathbf{W}(0))\|\right]^2 \leq \frac{1}{k}\mathbb{E}\left[\left\|\sum_{i=1}^{k} \phi(\mathbf{W}^i \mathbf{u})\mathbf{V}_i\right\|^2\right] = \mathbb{E}\left[\phi(\mathbf{W}^1\mathbf{u})^2\,\|\mathbf{V}_1\|^2\right] = C_\phi^2 n.$$

In addition,

$$\left|\|\mathbf{g}(\mathbf{u}, \mathbf{W})\| - \|\mathbf{g}(\mathbf{u}, \widetilde{\mathbf{W}})\|\right| \leq \frac{1}{\sqrt{k}}\left\|\sum_{i=1}^{k}\left(\phi(\mathbf{W}^i\mathbf{u}) - \phi(\widetilde{\mathbf{W}}^i\mathbf{u})\right)\mathbf{V}_i\right\|$$

$$\leq BD\sqrt{n}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left\|\mathbf{W}^i - \widetilde{\mathbf{W}}^i\right\|\right) \leq BD\sqrt{n}\left\|\mathbf{W} - \widetilde{\mathbf{W}}\right\|_F.$$

We then get

$$\mathbb{P}\left(\|\mathbf{g}(\mathbf{u}, \mathbf{W}(0))\| \geq C_\phi\sqrt{n\log(d)} + \tau\right)$$

$$\leq \mathbb{P}\left(\|\mathbf{g}(\mathbf{u}, \mathbf{W}(0))\| \geq \mathbb{E}\left[\|\mathbf{g}(\mathbf{u}, \mathbf{W}(0))\|\right] + \tau\right)$$

$$\leq e^{-\frac{\tau^2}{2\mathrm{Lip}(\|\mathbf{g}(\mathbf{u}, \mathbf{W}(0))\|)^2}} \leq e^{-\frac{\tau^2}{2nB^2D^2}}.$$

Taking $\tau = \sqrt{2}BD\sqrt{n\log(d)}$, we get the desired claim. $\square$

4   Numerical Experiments

We realized numerical experiments to verify our theoretical finding by evaluating the convergence of networks with different architecture parameters in the noise-free context. Every network was initialized in accordance with the assumptions of our work and we used the sigmoid activation function. Both $\mathbf{A}$ and $\bar{\mathbf{x}}$ entries were drawn i.i.d from $\mathcal{N}(0, 1)$. We used gradient descent to optimize the networks with a fixed step size of 1. A network was trained until it reached a loss of 10^{-7} or after 25000 optimization steps. For each set of architecture parameters, we did 50 runs and calculated the frequency at which the network arrived at the error threshold of 10^{-7}. We present in Figure 1 two experiments: in the first one
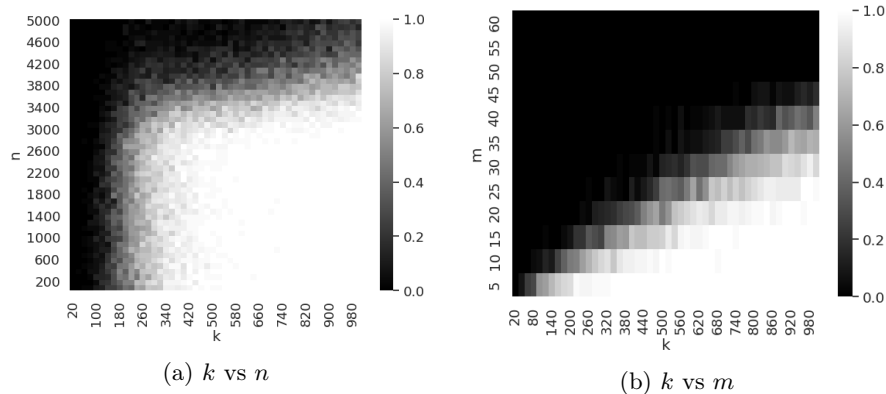
(a) k vs n

(b) k vs m

Fig. 1: Probability of arriving at a zero loss solution for networks with different architecture parameters. This emphasizes the linear dependence between k and m and the at least quadratic dependence between k and n.

we fix $m = 10$ and $d = 500$ and let k and n vary while in the second we fix $n = 60$, $d = 500$ and we let k and m vary.

We observe in Figure 1a the relationship between k and n for a fixed m. In this setup where $n \gg m$ and **A** is Gaussian, we expect, by our main theorem, a quadratic relationship which seems to be the case in the plot. It is however surprising that with values of k restricted to the range $[20, 1000]$, the network converges to a zero-loss solution with high probability for situations where $n > k$ which goes against our intuition for these underparametrized cases.

Additionally, the observation of Figure 1b provides a very different picture when the ratio m/n increases from 0. We first see clearly the expected linear relationship between k and m. However, contrary to Figure 1a where we observed convergence for high values of n (up to 3000) with a fixed m, this new setting where m grows gives very different results. For the same range of values for k, the networks only converge for small values of n (60 in this case). This indicates that the ratio m/n plays an important role in the level of overparametrization necessary for the network to converge. It is clear from these results that the order of our bounds is consistent with experiments, yet our bounds are not really tight, as we observe convergence for lower values of k than expected.

5   Conclusion and Future Work

This paper studied the convergence of shallow DIP networks and provided bounds on the level of overparametrization, both in the input dimension and the hidden layer dimension, under which the method converges exponentially fast to a zero-loss solution. The proof relies on bounding the minimum singular values of the Jacobian of the network through an overparametrization that ensures a good initialization of the network. These bounds are not tight, as demonstrated

by the numerical experiments, but they provide an important step towards the theoretical understanding of DIP methods, and neural networks for inverse problems resolution in general. In the future, this work will be extended in several directions. First, we will study recovery guarantees of the signal $\overline{x}$. Second, we will investigate the DIP model with unrestricted linear layers and possibly in the multilayer setting.

References

1. Allen-Zhu, Z., Li, Y., and Song, Z.: A Convergence Theory for Deep Learning via Over-Parameterization. In: ICML, pp. 242–252 (2019)
2. Arridge, S., Maass, P., Ozan, Ö., and Schönlieb, C.-B.: Solving inverse problems using data-driven models. Acta Numerica **28**, 1–174 (2019)
3. Bartlett, P.L., Montanari, A., and Rakhlin, A.: Deep learning: a statistical viewpoint. Acta Numerica **30**, 87–201 (2021). Publisher: Cambridge University Press
4. Chizat, L., Oyallon, E., and Bach, F.: On Lazy Training in Differentiable Programming. In: NeurIPS (2019)
5. Du, S.S., Zhai, X., Poczos, B., and Singh, A.: Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In: ICLR (2019)
6. Fang, C., Dong, H., and Zhang, T.: Mathematical Models of Overparameterized Neural Networks. Proceedings of the IEEE **109**(5), 683–703 (2021)
7. Jacot, A., Gabriel, F., and Hongler, C.: Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In: NeurIPS (2018)
8. Liu, J., Sun, Y., Xu, X., and Kamilov, U.S.: Image Restoration Using Total Variation Regularized Deep Image Prior. In: IEEE ICASSP, pp. 7715–7719 (2019)
9. Mataev, G., Milanfar, P., and Elad, M.: DeepRED: Deep Image Prior Powered by RED. In: ICCV (2019)
10. Monga, V., Li, Y., and Eldar, Y.C.: Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing. IEEE SPM **38**(2), 18–44 (2021)
11. Mukherjee, S., Hauptmann, A., Öktem, O., Pereyra, M., and Schönlieb, C.-B.: *Learned reconstruction methods with convergence guarantees*, (2022). arXiv:2206.05431 [cs]. Sept. 2022
12. Ongie, G., Jalal, A., Metzler, C.A., Baraniuk, R.G., Dimakis, A.G., and Willett, R.: Deep Learning Techniques for Inverse Problems in Imaging. IEEE J-SAIT (2020)
13. Oymak, S., and Soltanolkotabi, M.: Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path? In: ICML, pp. 4951–4960 (2019)
14. Oymak, S., and Soltanolkotabi, M.: Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks. J-SAIT (2020)
15. Prost, J., Houdard, A., Almansa, A., and Papadakis, N.: Learning Local Regularization for Variational Image Restoration. In: SSVM, pp. 358–370 (2021)
16. Shi, Z., Mettes, P., Maji, S., and Snoek, C.G.M.: On Measuring and Controlling the Spectral Bias of the Deep Image Prior. IJCV **130**(4), 885–908 (2022)
17. Tropp, J.A.: An Introduction to Matrix Concentration Inequalities. arXiv:1501.01571 [cs, math, stat] (2015). arXiv: 1501.01571

18. Ulyanov, D., Vedaldi, A., and Lempitsky, V.: Deep Image Prior. International Journal of Computer Vision **128**(7), 1867–1888 (2020). arXiv: 1711.10925
19. Venkatakrishnan, S.V., Bouman, C.A., and Wohlberg, B.: Plug-and-Play priors for model based reconstruction. In: GlobalSIP, pp. 945–948 (2013)
20. Zukerman, J., Tirer, T., and Giryes, R.: BP-DIP: A Backprojection based Deep Image Prior. In: EUSIPCO 2020, pp. 675–679 (2021). ISSN: 2076-1465