Convergence and Recovery Guarantees of Unsupervised Neural Networks for Inverse Problems

Nathan Buskulic, Jalal M. Fadili, Yvain Quéau

Convergence and Recovery Guarantees of Unsupervised Neural Networks for Inverse Problems

Nathan Buskulic, Jalal Fadili and Yvain Quéau

Greyc, Normandie Univ., UNICAEN, ENSICAEN, CNRS, 6 Boulevard Maréchal Juin, Caen, 14000, France.

*Corresponding author(s). E-mail(s): nathan.buskulic@unicaen.fr;
Contributing authors: Jalal.Fadili@ensicaen.fr; yvain.queau@ensicaen.fr;

Abstract

Neural networks have become a prominent approach to solve inverse problems in recent years. While a plethora of such methods was developed to solve inverse problems empirically, we are still lacking clear theoretical guarantees for these methods. On the other hand, many works proved convergence to optimal solutions of neural networks in a more general setting using overparametrization as a way to control the Neural Tangent Kernel. In this work we investigate how to bridge these two worlds and we provide deterministic convergence and recovery guarantees for the class of unsupervised feedforward multilayer neural networks trained to solve inverse problems. We also derive overparametrization bounds under which a two-layers Deep Inverse Prior network with smooth activation function will benefit from our guarantees.

Keywords: Inverse problems, Deep Image/Inverse Prior, Overparametrization, Gradient flow, Unsupervised learning

1 Introduction

1.1 Problem Statement

An inverse problem consists in reliably recovering a signal $\overline{\mathbf{x}} \in \mathbb{R}^n$ from noisy indirect observations

$$\mathbf{y} = \mathbf{F}(\overline{\mathbf{x}}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^m$ is the observation, $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is a forward operator, and $\boldsymbol{\varepsilon}$ stands for some additive noise. We will denote by $\overline{\mathbf{y}} = \mathbf{F}(\overline{\mathbf{x}})$ the ideal observations i.e., those obtained in the absence of noise.

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., the reviews [1, 2]. The general framework of these methods is to optimize a generator network $\mathbf{g} : (\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^d \times \Theta \mapsto \mathbf{x} \in \mathbb{R}^n$, $\Theta \subset \mathbb{R}^p$, with some activation function ϕ, to transform a given input $\mathbf{u} \in \mathbb{R}^d$ into a vector $\mathbf{x} \in \mathbb{R}^n$. The parameters $\boldsymbol{\theta}$ of the network are optimized via (possibly stochastic) gradient descent to minimize a loss function $\mathcal{L}_{\mathbf{y}} : \mathbb{R}^m \to \mathbb{R}_+, \mathbf{y}(t) \mapsto \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))$ which measures the discrepancy between the observation $\mathbf{y}$ and the solution $\mathbf{y}(t) = \mathbf{F}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t)))$ generated by the network at time $t \geq 0$.

Theoretical understanding of recovery and convergence guarantees for deep learning-based methods is of paramount importance to make their routine use in critical applications reliable [3]. While there is a considerable amount of work on the understanding of optimization dynamics of neural network training, especially through the lens of overparametrization, recovery guarantees when using neural networks for inverse problem remains elusive. Some attempts have been made in that direction but they are usually restricted to very specific settings. One kind of results that was obtained [4–6] is convergence towards the optimal points of a regularized problem, typically with a learned regularizer. However this does not give guarantees about the real sought-after vector. Another approach is used in Plug-and-Play [7] to show that under strong assumptions on the pre-trained denoiser, one can prove convergence to the true vector. This work is however limited by the constraints on the denoiser which are not met in many settings.

Our aim in this paper is to help close this gap by explaining when gradient descent consistently and provably finds global minima of $\mathcal{L}$, and how this translates into recovery guarantees for both $\overline{\mathbf{y}}$ and $\overline{\mathbf{x}}$ i.e., in both the observation and the signal spaces. For this, we focus on a continuous-time gradient flow applied to $\mathcal{L}$:

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t)))) \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases} \tag{2}$$

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

In this work, our focus in on an unsupervised method known as Deep Image Prior [8], that we also coin Deep Inverse Prior (DIP) as it is not confined to images. A chief advantage of this method is that it does not need any training data, while the latter is mandatory in most supervised deep learning-based methods used in the literature. In the DIP method, $\mathbf{u}$ is fixed throughout the optimization/training process, usually a realization of a random variable. By taking out the need of training data, this method focuses on the generation capabilities of the network trained through gradient descent. In turn, this will allow us to get insight into the effect of network architecture on the reconstruction quality.

1.2 Contributions

We deliver a theoretical analysis of gradient flow optimization of neural networks, i.e. (2), in the context of inverse problems and provide various recovery guarantees for general loss functions verifying the Kurdyka-Łojasewicz (KL) property. We first prove that the trained network with a properly initialized gradient flow will converge to an optimal solution in the observation space with a rate characterized by the desingularizing function appearing in the KL property of the loss function. This result is then converted to a prediction error on $\overline{\mathbf{y}}$ through an early stopping strategy. More importantly, we present a recovery result in the signal space with an upper bound on the reconstruction error of $\overline{\mathbf{x}}$. The latter result involves for instance a restricted injectivity condition on the forward operator.

We then turn to showing how these results can be applied to the case of a two-layer neural network in the DIP setting where

$$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) = \frac{1}{\sqrt{k}} \mathbf{V} \phi(\mathbf{W}\mathbf{u}), \quad \boldsymbol{\theta} \stackrel{\text{def}}{=} (\mathbf{V}, \mathbf{W}), \tag{3}$$

with $\mathbf{V} \in \mathbb{R}^{n \times k}$, $\mathbf{W} \times \mathbb{R}^{k \times d}$, and ϕ an element-wise nonlinear activation function. The scaling by $\sqrt{k}$ will become clearer later. We show that for a proper random initialization $\mathbf{W}(0)$, $\mathbf{V}(0)$ and sufficient overparametrization, all our conditions are in force to control the eigenspace of the Jacobian of the network as required to obtain the aforementioned convergence properties. We provide a characterization of the overparametrization needed in terms of (k, d, n) and the conditioning of $\mathbf{F}$.

1.3 Relation to Prior Work

Data-Driven Methods to Solve Inverse Problems

Data-driven approaches to solve inverse problems come in various forms [1, 2]. 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 [9] or use Plug-and-Play methods [10] for example. Another family of approaches, which takes inspiration from classical iterative optimization algorithms, is based on unrolling (see [11] for a review of these methods). Still, all these methods require an extensive amount of training data, which may not always be available.

Deep Inverse Prior

The DIP model [8] (and its extensions that mitigate some of its empirical issues [12–15]) 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). The theoretical recovery

guarantees of these methods are not well understood [3] and our work aims at reducing this theoretical gap by analyzing the behaviour of such networks in both the observation and the signal space under some overparametrization condition.

Theory of Overparametrized Networks

To construct our analysis, we build upon previous theoretical work of overparametrized networks and their optimization trajectories [16, 17]. The first works that proved convergence to an optimal solution were based on a strong convexity assumption of the loss which is typically not the case when it is composed with a neural network. A more recent approach is based on a gradient dominated inequality from which we can deduce by simple integration an exponential convergence of the gradient flow to a zero-loss solution. This allows to obtain convergence guarantees for networks trained to minimize a mean square error by gradient flow [18] or its discrete counterpart (i.e., gradient descent with fixed step) [19–22]. The work that we present here is inspired by these works but it goes far beyond them. Amongst other differences, we are interested in the challenging situation of inverse problems (presence of a foward operator), and we deal with more general loss functions that obey the Kurdyka-Łojasewicz inequality (e.g., any semi-algebraic function or even definable on an o-minimal structure) [23–25].

Recently, it has been found that some kernels play a very important role in the analysis of convergence of the gradient flow when used to train neural networks. In particular the semi-positive definite kernel given by $\mathcal{J}_{\mathbf{g}}(t)\mathcal{J}_{\mathbf{g}}(t)^\top$, where $\mathcal{J}_{\mathbf{g}}(t)$ is the Jacobian of the network at time t. When all the layers of a network are trained, this kernel is a combination of the *Neural Tangent Kernel* (NTK) [26] and the Random Features Kernel (RF) [27]. If one decides to fix the last layer of the network, then this amounts to just looking at the NTK which is what most of the previously cited works do. The goal is then to control the eigenvalues of the kernel to ensure that it stays positive definite during training, which entails convergence to a zero-loss solution at an exponential rate. The control of the eigenvalues of the kernel is done through a random initialization and the overparametrization of the network. Indeed, for a sufficiently wide network, the parameters $\boldsymbol{\theta}(t)$ will stay near their initialization and they will be well approximated by their linearization (so-called "lazy" regime [18]). The overparametrization bounds that were obtained are mostly for two-layers networks as the control of deep networks is much more complex.

However, even if there are theoretical works on the gradient flow-based optimization of neural networks as reviewed above, similar analysis that would accommodate for the forward operator as in inverse problems remain challenging and open. Our aim is to participate in this endeavour by providing theoretical understanding of recovery guarantees with neural network-based methods.

This paper is an extension of our previous one in [28]. There are however several distinctive and new results in the present work. For instance, the work [28] only dealt with linear inverse problems while our results here apply to non-linear ones. Moreover, we here provide a much more general analysis under which we obtain convergence guarantees for a wider class of models than just the DIP one and for a general class of loss functions, not just the MSE. More importantly we show convergence not only in the observation space but also in the signal space now. When particularized to the DIP case, we also provide overparametrization bounds for the case when the linear layer of the network is not fixed which is also an additional novelty.

Paper organization

The rest of this work is organized as follows. In Section 2 we give the necessary notations and definitions useful for this work. In Section 3 we present our main result with the associated assumptions and proof. In Section 4 we present the overparametrization bound on the DIP model. Finally, in Section 5, we show some numerical experiments that validate our findings, before drawing our conclusions in Section 6.

2 Preliminaries

2.1 General Notations

For a matrix $\mathbf{M} \in \mathbb{R}^{a\times b}$ we denote by $\sigma_{\min}(\mathbf{M})$ and $\sigma_{\max}(\mathbf{M})$ its smallest and largest 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 $\langle , \rangle$ the Euclidean scalar product, $\|\cdot\|$ the associated norm (the dimension is implicit from the context), and $\|\cdot\|_F$ the Frobenius norm of a matrix. With a slight abuse of notation $\|\cdot\|$ will also denote the spectral norm of a matrix. We use $\mathbf{M}^i$ (resp. $\mathbf{M}_i$) as the i-th row (resp. column) of $\mathbf{M}$. For two vectors $\mathbf{x}, \mathbf{z}$, $[\mathbf{x}, \mathbf{z}] = \{(1-\rho)\mathbf{x} + \rho\mathbf{z} : \ \rho \in [0,1]\}$ is the closed segment joining them. We use the notation $a \gtrsim b$ if there exists a constant $C > 0$ such that $a \geq Cb$.

We also define $\mathbf{y}(t) = \mathbf{F}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t)))$ and $\mathbf{x}(t) = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$ and we recall $\overline{\mathbf{y}} = \mathbf{F}(\overline{\mathbf{x}})$. The Jacobian of the network is denoted $\mathcal{J}_{\mathbf{g}}$. $\mathcal{J}_{\mathbf{g}}(t)$ is a shorthand notation of $\mathcal{J}_{\mathbf{g}}$ evaluated at $\boldsymbol{\theta}(t)$. $\mathcal{J}_{\mathbf{F}}(t)$ is the Jacobian of the forward operator $\mathbf{F}$ evaluated at $\mathbf{x}(t)$. The local Lipschitz constant of a mapping on a ball of radius $R > 0$ around a point $\mathbf{z}$ is denoted $\mathrm{Lip}_{\mathbb{B}(\mathbf{z},R)}(\cdot)$. We omit R in the notation when the Lipschitz constant is global. For a function $f : \mathbb{R}^n \to \mathbb{R}$, we use the notation for the sublevel set $[f < c] = \{\mathbf{z} \in \mathbb{R}^n : \ f(\mathbf{z}) < c\}$ and $[c_1 < f < c_2] = \{\mathbf{z} \in \mathbb{R}^n : \ c_1 < f(\mathbf{z}) < c_2\}$.

Given $\mathbf{z} \in \mathcal{C}^0(]0, +\infty[; \mathbb{R}^a)$, the set of cluster points of $\mathbf{z}$ is defined as

$$\mathfrak{W}(\mathbf{z}(\cdot)) = \left\{ \widetilde{\mathbf{z}} \in \mathbb{R}^a : \ \exists (t_k)_{k\in\mathbb{N}} \to +\infty \text{ s.t. } \lim_{k\to\infty} \mathbf{z}(t_k) = \widetilde{\mathbf{z}} \right\}.$$

We define $\Sigma = \{\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) : \ \boldsymbol{\theta} \in \Theta\}$ the set of signals that the network $\mathbf{g}$ can generate for all θ in the parameter set Θ. Σ can thus be viewed as a parametric manifold. If Θ is closed (resp. compact), so is Σ. We denote $\mathrm{dist}(\cdot, \Sigma)$ the distance to Σ which is well defined if Θ is closed and non-empty. For a vector $\mathbf{x}$, $\mathbf{x}_\Sigma$ is its projection on Σ. We also define $T_\Sigma(\mathbf{x}) = \overline{\mathrm{conv}}\left(\mathbb{R}_+(\Sigma - \mathbf{x})\right)$ the tangent cone of Σ at $\mathbf{x} \in \Sigma$.

The minimal (conic) singular value of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ w.r.t. the cone $T_\Sigma(\mathbf{x})$ is then defined as

$$\lambda_{\min}(\mathbf{A}; T_\Sigma(\mathbf{x})) = \inf\{\|\mathbf{A}\mathbf{z}\| \,/\, \|\mathbf{z}\| : \mathbf{z} \in T_\Sigma(\mathbf{x})\}.$$

2.2 Multilayer Neural Networks

Neural networks produce structured parametric families of functions that have been studied and used for almost 70 years, going back to the late 1950's [29].

Definition 2.1. Let $d, L \in \mathbb{N}$ and $\phi : \mathbb{R} \to \mathbb{R}$ an activation map which acts componentwise on the entries of a vector. A fully connected multilayer neural network with input dimension d, L layers and activation ϕ, is a collection of weight matrices $\left(\mathbf{W}^{(l)}\right)_{l\in[L]}$ and bias vectors

$\left(\mathbf{b}^{(l)}\right)_{l\in[L]}$, where $\mathbf{W}^{(l)} \in \mathbb{R}^{N_l\times N_{l-1}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{N_l}$, with $N_0 = d$, and $N_l \in \mathbb{N}$ is the number of neurons for layer $l \in [L]$. Let us gather these parameters as

$$\boldsymbol{\theta} = \left((\mathbf{W}^{(1)}, \mathbf{b}^{(1)}), \dots, (\mathbf{W}^{(L)}, \mathbf{b}^{(L)})\right) \in \Theta \stackrel{\text{def}}{=} \bigtimes_{l=1}^{L} \left(\mathbb{R}^{N_l\times N_{l-1}} \times \mathbb{R}^{N_l}\right).$$

Then, a neural network parametrized by $\boldsymbol{\theta}$ produces a function

$$\mathbf{g} : (\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^d \times \Theta \mapsto \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^{N_L}, \quad \text{with} \quad N_L = n,$$

which can be defined recursively as

$$\begin{cases} \mathbf{g}^{(0)}(\mathbf{u}, \boldsymbol{\theta}) & = \mathbf{u}, \\ \mathbf{g}^{(l)}(\mathbf{u}, \boldsymbol{\theta}) & = \phi\left(\mathbf{W}^{(l)}\mathbf{g}^{(l-1)}(\mathbf{u}, \boldsymbol{\theta}) + \mathbf{b}^{(l)}\right), \quad \text{for } l = 1, \dots, L-1, \\ \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) & = \mathbf{W}^{(L)}\mathbf{g}^{(L-1)}(\mathbf{u}, \boldsymbol{\theta}) + \mathbf{b}^{(L)}. \end{cases}$$

In the rest of this work, $\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$ is always defined as just described. In Section 4 for example, the network we study is defined with $L = 2$ and $\mathbf{b}^{(1)} = \mathbf{b}^{(2)} = 0$.

2.3 KL Functions

We will work under a general condition of the loss function $\mathcal{L}$ which includes non-convex ones. More precisely, we will suppose that $\mathcal{L}$ verifies a Kurdyka-Łojasewicz-type (KL for short) inequality [25, Theorem 1].

Definition 2.2 (KL inequality)**.** A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the KL inequality if there exists $r_0 > 0$ and a strictly increasing function $\psi \in \mathcal{C}^0([0, r_0[) \cap \mathcal{C}^1(]0, r_0[)$ with $\psi(0) = 0$ such that

$$\psi'(f(\mathbf{z}) - \min f)\, \|\nabla f(\mathbf{z})\| \geq 1, \quad \text{for all} \quad \mathbf{z} \in [\min f < f < \min f + r_0]. \tag{4}$$

We use the shorthand notation $f \in \text{KŁ}_\psi(r_0)$ for a function satisfying this inequality.

The KL property basically expresses the fact that the function f is sharp under a reparameterization of its values. Functions satisfying the KL inequality are also sometimes called gradient dominated functions [30]. The function ψ is known as the desingularizing function for f. The Łojasiewicz inequality [23, 24] corresponds to the case where the desingularizing function takes the form $\psi(s) = cs^\alpha$ with $\alpha \in [0, 1]$. The KL inequality plays a fundamental role in several fields of applied mathematics among which convergence behaviour of (sub-)gradient-like systems and minimization algorithms [31–36], neural networks [37], partial differential equations [38–40], to cite a few. The KL inequality is closely related to error bounds that also play a key role to derive complexity bounds of gradient descent-like algorithms [41].

Let us give some examples of functions satisfying (4); see also [35].

Example 2.3 (Convex functions with sufficient growth)**.** Let f be a differentiable convex function on $\mathbb{R}^n$ such that $\operatorname{Argmin}(f) \neq \emptyset$. Assume that f verifies the growth condition

$$f(\mathbf{z}) \geq \min f + \varphi(\operatorname{dist}(\mathbf{z}, \operatorname{Argmin}(f))), \quad \text{for all} \quad \mathbf{z} \in [\min f < f < \min f + r], \tag{5}$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous, increasing, $\varphi(0) = 0$ and $\int_0^r \frac{\varphi^{-1}(s)}{s} ds < +\infty$. Then by [36, Theorem 30], $f \in \text{KŁ}_\psi(r)$ with $\psi(r) = \int_0^r \frac{\varphi^{-1}(s)}{s} ds$.

Example 2.4 (Uniformly convex functions)**.** Suppose that f is a differentiable uniformly convex function, i.e., $\forall \mathbf{z}, \mathbf{x} \in \mathbb{R}^n$,

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle + \varphi\left(\|\mathbf{x} - \mathbf{z}\|\right) \tag{6}$$

for an increasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ that vanishes only at 0. Thus f has a unique minimizer, say $\mathbf{z}^\star$, see [42, Proposition 17.26]. This example can then be deduced from the previous one since a uniformly convex function obviously obeys (5). However, we here provide an alternative and sharper characterization. We may assume without loss of generality that $\min f = 0$. Applying inequality (6) at $\mathbf{x} = \mathbf{z}^\star$ and any $\mathbf{z} \in [0 < f]$, we get

$$\begin{aligned} f(\mathbf{z}) &\leq \langle \nabla f(\mathbf{z}), \mathbf{z} - \mathbf{x} \rangle - \varphi\left(\|\mathbf{x} - \mathbf{z}\|\right) \\ &\leq \|\nabla f(\mathbf{z})\| \, \|\mathbf{x} - \mathbf{z}\| - \varphi\left(\|\mathbf{x} - \mathbf{z}\|\right) \\ &\leq \varphi_+(\|\nabla f(\mathbf{z})\|), \end{aligned}$$

where $\varphi_+ : a \in \mathbb{R}_+ \mapsto \varphi^+(a) = \sup_{x \geq 0} ax - \varphi(x)$ is known as the monotone conjugate of φ. φ_+ is a proper closed convex and non-decreasing function on $\mathbb{R}_+$ that vanishes at 0. When φ is strictly convex and supercoercive, so is φ_+ which implies that φ_+ is also strictly increasing on $\mathbb{R}_+$. Thus f verifies Definition 2.2 at any $\mathbf{z} \in [0 < f]$ with ψ a primitive of $\frac{1}{\varphi_+^{-1}}$, and ψ is indeed strictly increasing, vanishes at 0 and is even concave. A prominent example is the case where $\varphi : s \in \mathbb{R}_+ \mapsto \frac{1}{p} s^p$, for $p \in]1, +\infty[$, in which case $\psi : s \in \mathbb{R}_+ \mapsto q^{-1/q} s^{1/p}$, where $1/p + 1/q = 1$.

Example 2.5. In finite-dimensional spaces, deep results from algebraic geometry have shown that the KL inequality is satisfied by a large class of functions, namely, real semi-algebraic functions and more generally, function definable on an o-minimal structure or even functions belonging to analytic-geometric categories [23–25, 43, 44]. Many popular losses used in machine learning and signal processing turn out to be KL functions (MSE, Kullback-Leibler divergence and cross-entropy to cite a few).

3 Recovery Guarantees

3.1 Main Assumptions

Throughout this paper, we will work under the following standing assumptions.

Assumptions on the loss

A-1. $\mathcal{L}_{\mathbf{y}}(\cdot) \in \mathcal{C}^1(\mathbb{R}^m)$ *whose gradient is Lipschitz continuous on the bounded sets of* $\mathbb{R}^m$.
A-2. $\mathcal{L}_{\mathbf{y}}(\cdot) \in \text{KŁ}_{\psi}(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) + \eta)$ *for some* $\eta > 0$.
A-3. $\min \mathcal{L}_{\mathbf{y}}(\cdot) = 0$.
A-4. $\nabla_{\mathbf{v}} \mathcal{L}_{\mathbf{y}}(\mathbf{v}) \in \operatorname{Im}(\mathcal{J}_{\mathbf{F}}(\mathbf{x}))$ *for any* $\mathbf{v} = \mathbf{F}(\mathbf{x})$ *with* $\mathbf{x} \in \Sigma$.

Assumption on the activation

A-5. $\phi \in \mathcal{C}^1(\mathbb{R})$ *and* $\exists B > 0$ *such that* $\sup_{x \in \mathbb{R}} |\phi'(x)| \leq B$ *and* ϕ' *is* B*-Lipschitz continuous.*

Assumption on the forward operator

A-6. $\mathbf{F} \in \mathcal{C}^1(\mathbb{R}^n; \mathbb{R}^m)$ *whose Jacobian is Lipschitz continuous on the bounded sets of* $\mathbb{R}^n$.

Let us now discuss the meaning and effects of these assumptions. First, A-1 is made for simplicity to ensure existence and uniqueness of a strong maximal solution (in fact even global thanks to our estimates) of (2) thanks to the Cauchy-Lipschitz theorem (see hereafter). We think this could be relaxed to cover non-smooth losses if we assume path differentiability, hence existence of an absolutely continuous trajectory. This is left to a future work. A notable point in A-2 is that convexity is not always needed for the loss (see the statements of the theorem). Regarding A-3, it is natural yet it would be straightforward to relax it. Finally, Assumption A-4 allows us to leverage the fact that

$$\sigma_{\mathbf{F}} := \inf_{\mathbf{x} \in \Sigma, \mathbf{z} \in \operatorname{Im}(\mathcal{J}_{\mathbf{F}}(\mathbf{x}))} \left\| \mathcal{J}_{\mathbf{F}}(\mathbf{x})^{\top} \mathbf{z} \right\| / \|\mathbf{z}\| > 0. \tag{7}$$

One case when this assumption is verified is when $\mathbf{F}$ is an immersion, which implies that $\mathcal{J}_{\mathbf{F}}(\mathbf{x})$ is surjective for all $\mathbf{x}$. Other interesting cases are when $\mathcal{L}_{\mathbf{y}}(\mathbf{v}) = \eta\left(\|\mathbf{v} - \mathbf{y}\|^2\right)$ and $\mathbf{F}$ is linear, where $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ is differentiable and vanishes only at 0, in which case $\nabla_{\mathbf{v}} \mathcal{L}_{\mathbf{y}}(\mathbf{v}) = 2\eta'\left(\|\mathbf{v} - \mathbf{y}\|^2\right)(\mathbf{v} - \mathbf{y})$ for $\mathbf{v} = \mathbf{A}\mathbf{x}$. It is then sufficient to require that $\mathbf{y} \in \operatorname{Im}(\mathbf{A})$ for the assumption to hold.

Assumption A-5 is key in well-posedness as it ensures, by Definition 2.1 which $\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$ follows, that $\mathbf{g}(\mathbf{u}, \cdot)$ is $\mathcal{C}^1(\mathbb{R}^p; \mathbb{R}^p)$ whose Jacobian is Lipschitz continuous on bounded sets, which is necessary for the Cauchy-Lipschitz theorem. This constraint on ϕ is met by many activations such as the softmax, sigmoid or hyperbolic tangent. Including the ReLU requires more technicalities that will be avoided here.

Finally, Assumption A-6 on local Lipschitz continuity on $\mathbf{F}$ is not only important for well-posedness of (2), but it turns out to be instrumental when deriving recovery rates (as a function of the noise) in the literature of regularized nonlinear inverse problems; see [45] and references therein.

3.2 Well-posedness

In order for our analysis to hold, the Cauchy problem (2) needs to be well-posed. We start by showing that (2) has a unique maximal solution.

Proposition 3.1. *Assume that A-1, A-5 and A-6 hold. There there exists* $T(\boldsymbol{\theta}_0) \in]0, +\infty]$ *and a unique maximal solution* $\boldsymbol{\theta}(\cdot) \in \mathcal{C}^0([0, T(\boldsymbol{\theta}_0)[)$ *of (2), and* $\boldsymbol{\theta}(\cdot)$ *is* $\mathcal{C}^1$ *on every compact set of the interior of* $[0, T(\boldsymbol{\theta}_0)[$.

Proof. Thanks to A-5, one can verify with standard differential calculus applied to $\mathbf{g}(\mathbf{u}, \cdot)$, as given in Definition 2.1, that $\mathcal{J}_{\mathbf{g}}$ is Lipschitz continuous on the bounded sets of $\mathbb{R}^p$. This together with A-1 and A-6 entails that $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u}, \cdot))$ is also Lipschitz continuous on the bounded sets of $\mathbb{R}^p$. The claim is then a consequence of the Cauchy-Lipschitz theorem [46, Theorem 0.4.1]. □

$T(\boldsymbol{\theta}_0)$ is known as the maximal existence time of the solution and verifies the alternative: either $T(\boldsymbol{\theta}_0) = +\infty$ and the solution is called *global*; or $T(\boldsymbol{\theta}_0) < +\infty$ and the solution blows-up in finite time, i.e., $\|\boldsymbol{\theta}(t)\| \to +\infty$ as $t \to T(\boldsymbol{\theta}_0)$. We will show later that the maximal solution of (2) is indeed global; see Section 3.4.4.

3.3 Main Results

We are now in position to state our recovery results.

Theorem 3.2. *Recall* $\sigma_{\mathbf{F}}$ *from (7). Consider a network* $\mathbf{g}(\mathbf{u}, \cdot)$*, a forward operator* $\mathbf{F}$ *and a loss* $\mathcal{L}$*, such that A-1 to A-6 hold. Let* $\boldsymbol{\theta}(\cdot)$ *be a solution trajectory of (2) where the initialization* $\boldsymbol{\theta}_0$ *is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) > 0 \quad \text{and} \quad R' < R \tag{8}$$

where R' *and* R *obey*

$$R' = \frac{2}{\sigma_{\mathbf{F}} \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))} \psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))) \quad \text{and} \quad R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0, R)}(\mathcal{J}_{\mathbf{g}})}. \tag{9}$$

Then the following holds:

(i) the loss converges to 0 *at the rate*

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \Psi^{-1}(\gamma(t)) \tag{10}$$

with Ψ *a primitive of* $-\psi'^2$ *and* $\gamma(t) = \frac{\sigma_{\mathbf{F}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4} t + \Psi\left(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\right)$*. Moreover,* $\boldsymbol{\theta}(t)$ *converges to a global minimizer* $\boldsymbol{\theta}_\infty$ *of* $\mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u}, \cdot)))$*, at the rate*

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| \leq \frac{2}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) \sigma_{\mathbf{F}}} \psi\left(\Psi^{-1}\left(\gamma(t)\right)\right). \tag{11}$$

(ii) If $\mathrm{Argmin}(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$*, then* $\lim_{t \to +\infty} \mathbf{y}(t) = \mathbf{y}$*. In addition, if* $\mathcal{L}$ *is convex then*

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\|\boldsymbol{\varepsilon}\| \quad \text{when} \quad t \geq \frac{4\Psi(\psi^{-1}(\|\boldsymbol{\varepsilon}\|))}{\sigma_{\mathbf{F}}^2 \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2} - \Psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))). \tag{12}$$

(iii) Assume that $\mathrm{Argmin}(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$*,* $\mathcal{L}$ *is convex, and that*

A-7. $\mu_{\mathbf{F},\Sigma} > 0$ *where* $\mu_{\mathbf{F},\Sigma} = \inf_{\mathbf{x} \in \Sigma} \frac{\|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma)\|}{\|\mathbf{x} - \overline{\mathbf{x}}_\Sigma\|}$.

Let $L_{\mathbf{F}} = \max_{\mathbf{x}\in\mathbb{B}(0,2\|\overline{\mathbf{x}}\|)} \|\mathcal{J}_{\mathbf{F}}(\mathbf{x})\| < +\infty$. *Then*

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{2\psi\left(\Psi^{-1}\left(\gamma(t)\right)\right)}{\mu_{\mathbf{F},\Sigma}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{F}}} + \left(1 + \frac{L_{\mathbf{F}}}{\mu_{\mathbf{F},\Sigma}}\right)\operatorname{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\|\boldsymbol{\varepsilon}\|}{\mu_{\mathbf{F},\Sigma}}. \tag{13}$$

3.4 Discussion and Consequences

We first discuss the meaning of the initialization condition $R' < R$. This dictates that $\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)))$ must be smaller than some constant that depends on the operator $\mathbf{F}$ and the Jacobian of the network at initialization. Intuitively, this requires the initialization of the network to be in an appropriate convergence basin i.e., we start close enough from an optimal solution.

3.4.1 Convergence Rate

The first result ensures that under the conditions of the theorem, the network converges towards a zero-loss solution. The convergence speed is given by the application of Ψ^{-1}, which is (strictly) decreasing by definition, on an affine function w.r.t time. The function Ψ only depends on the chosen loss function and its associated Kurdyka-Łojasewiecz inequality. This inequality is verified for a wide class of functions, including all the semi-algebraic ones [25], but it is not always obvious to know the exact formulation of ψ (see section 2.3).

In the case where the KL inequality is respected with $\psi = cs^{\alpha}$ (the Łojasiewicz case), we obtain by direct computation the following decay rate of the loss and convergence rate for the parameters:

Corollary 3.3. *If $\mathcal{L}$ satisfies the Łojasiewicz inequality, that is A-2 holds with $\psi(s) = cs^{\alpha}$ and $\alpha \in [0, 1]$, then, $\exists t_0 \in R^+$ such that $\forall t > t_0, \gamma(t) > 0$ and the loss and the parameters converge with rate:*

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \begin{cases} \left(\frac{1-2\alpha}{\alpha^2 c^2}\gamma(t)\right)^{-\frac{1}{1-2\alpha}} & \text{if } 0 < \alpha < \frac{1}{2}, \\ \left(\frac{2\alpha-1}{\alpha^2 c^2}\gamma(t)\right)^{-\frac{1}{2\alpha-1}} & \text{if } \frac{1}{2} < \alpha < 1 \\ \exp\left(-\frac{4}{c^2}\gamma(t)\right) & \text{if } \alpha = \frac{1}{2} \end{cases}$$

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_{\infty}\| \leq \begin{cases} \left(\frac{1-2\alpha}{\alpha^2 c^2}\gamma(t)\right)^{-\frac{\alpha}{1-2\alpha}} & \text{if } 0 < \alpha < \frac{1}{2}, \\ \left(\frac{2\alpha-1}{\alpha^2 c^2}\gamma(t)\right)^{-\frac{\alpha}{2\alpha-1}} & \text{if } \frac{1}{2} < \alpha < 1 \\ \exp\left(-\frac{4}{c^2}\gamma(t)\right) & \text{if } \alpha = \frac{1}{2} \end{cases}$$

These results allow to see precise convergence rates of the loss for a wide variety of functions. First let us observe the particular case when $\alpha = 1/2$ which gives exponential convergence to the solution. In practice a function that matches such Łojasiewicz inequality is the Mean Squared Error (MSE). For other values of α, we obtain convergence rates in $O(t^{-\frac{1}{1-2\alpha}})$ or $O(t^{-\frac{1}{2\alpha-1}})$ depending on the interval of α that was chosen. Furthermore, in theory, the parameters of the model will converge slightly slower than the loss with their convergence speed modulated by α.

3.4.2 Early stopping strategy

While the first result allows us to obtain convergence rates to a zero-loss solution, it does so by overfitting the noise inherent to the problem. A classical way to avoid this to happen is to use an early stopping strategy to ensure that our solution will lie in a ball around the desired solution. The bound on the time given in (12) will verify that all the solutions found past that time will be no more than $2\|\boldsymbol{\varepsilon}\|$ away from the noiseless solution. This bound is given by balancing the convergence rate offered by the KL properties of the loss, the loss of the model at initialization and the level of noise in the problem.

3.4.3 Signal Recovery Guarantees

Our third result provides a bound on the distance between the solution found at time t and the true solution $\overline{\mathbf{x}}$. This bound is a sum of three terms representing three kinds of errors. The first term is an "optimization error", which represents how far $\mathbf{x}(t)$ is from the solution found at the end of the optimization process. Of course, this decreases to 0 as t goes to infinity. The second error is a "modelization error" which represents how well the chosen model can generate solutions close to $\overline{\mathbf{x}}$. Finally, the third term is a "noise error" that depends on $\|\boldsymbol{\varepsilon}\|$ which is inherent to the problem at hand.

Obviously, the operator $\mathbf{F}$ also plays a key role in this bound where its effects are given by three quantities of interest. First, $\sigma_{\mathbf{F}}$ ensures that on the subspace Σ, the jacobian of $\mathbf{F}$ will not be ill-conditioned. Second, $\mu_{\mathbf{F},\Sigma}$ yields a restricted injectivity condition on Σ, which is classical and necessary if we hope to be able to solve the problem. This constraint implies that the dimension of $\mathrm{Im}\,(\mathbf{F})$ when $\mathbf{F}$ is applied on $\mathbf{x} \in \Sigma$ is the same as the dimension of Σ. Third, $L_{\mathbf{F}}$, which is the Lipschitz constant of the jacobian of $\mathbf{F}$ on Σ. In particular, in the case where $\mathbf{F}$ is a linear operator $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mu_{\mathbf{F},\Sigma}$ becomes the minimal conic singular value $\lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}))$ and $L_{\mathbf{F}}$ is replaced by $\|\mathbf{A}\|$. (A-7) then amounts to assuming that

$$\ker(\mathbf{A}) \cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}) = \{0\}\,. \tag{14}$$

From the above description, it appears that there is a trade-off between the expressivity of the model and the injectivity of the forward operator. If one chooses a very expressive model, then $\mathrm{dist}(\overline{\mathbf{x}}, \Sigma)$ will tend to 0. But this is likely to come at the cost of making $\mu_{\mathbf{F},\Sigma}$ decrease, as restricted injectivity is likely to be required on a larger subset (cone). In the worst case, if $\mathbf{g}$ spans the entire space, this enforces the global injectivity of $\mathbf{F}$ which is very restrictive.

This discussion relates with the work on the instability phenomenon observed in learned reconstruction methods as discussed in [47, 48]. For instance, when $\mathbf{F}$ is a linear operator $\mathbf{A}$, the fundamental problem that creates these instabilities and/or hallucinations in the reconstruction is due to the fact that the kernel of $\mathbf{A}$ is non-trivial. Thus a method that can correctly learn to reconstruct signals whose difference lies in or close to the kernel of $\mathbf{A}$ will necessarily be unstable or hallucinate. In our setting, this is manifested through the restricted injectivity condition, that imposes that the smallest conic singular value is bounded away from 0, i.e. $\mu_{\mathbf{F},\Sigma} = \lambda_{\min}(\mathbf{A}; T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma})) > 0$. This is a natural (and minimal) condition in the context of inverse problems to have stable reconstruction guarantees. Note that our condition is non-uniform as it is only required to hold at $\overline{\mathbf{x}}_{\Sigma}$ and not at all points of Σ.

In A-11, we generalize the restricted injectivity condition (14) beyond the linear case provided that $\mathcal{J}_{\mathbf{F}}$ is Lipschitz continuous. This covers many practical cases, for instance that of

phase retrieval. Observe that whereas assumption A-7 requires a uniform control of injectivity of $\mathbf{F}$ on the whole signal class Σ, A-11 is less demanding and only requires injectivity of the Jacobian of $\mathbf{F}$ at $\overline{\mathbf{x}}_\Sigma$ on the tangent space of Σ at $\overline{\mathbf{x}}_\Sigma$. However the price is that the recovery bound in Theorem A.1 is only valid for high signal-to-noise regime and $\operatorname{dist}(\overline{\mathbf{x}}, \Sigma)$ is small enough. Moreover, the convergence rate in noise becomes $O(\sqrt{\|\varepsilon\|})$ which is worse than $O(\|\varepsilon\|)$ of Theorem 3.2.

Example 3.4 (Compressed sensing with sub-Gaussian measurements)**.** Controlling the minimum conic singular value is not easy in general. Amongst the cases where results are available, we will look at the compressed sensing framework with linear random measurements. In this setting, the forward operator $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a random sensing matrix. Exploiting the randomness of $\mathbf{A}$, a natural question is then how many measurements are sufficient to ensure that $\lambda_{\min}(\mathbf{A}; T_\Sigma(\overline{\mathbf{x}}_\Sigma)) > 0$ with high probability. In the case of Gaussian and sub-Gaussian measurements, we can exploit the non-uniform results of [49, 50] to derive sample complexity bounds, i.e. lower bounds on m, for this to hold. By using [50, Theorem 6.3], we have the following proposition:

Proposition 3.5. *Assume that each row $\mathbf{A}^i$ is an independent sub-Gaussian vector, that is*

$$
\begin{aligned}
&(i) \quad \mathbb{E}[\mathbf{A}^i] = 0,\\
&(ii) \quad \alpha \le \mathbb{E}[\left|\langle \mathbf{A}^i, \mathbf{w}\rangle\right|] \text{ for each } \mathbf{w} \in \mathbb{S}^{n-1}, \text{ with } \alpha > 0,\\
&(iii) \quad \mathbb{P}\left(\left|\langle \mathbf{A}^i, \mathbf{w}\rangle\right| \ge \tau\right) \le 2e^{-\tau^2/(2\sigma^2)} \text{ for each } \mathbf{w} \in \mathbb{S}^{n-1}, \text{ with } \sigma > 0.
\end{aligned}
$$

Let C and C' be positive constants and $w(K)$ the Gaussian width of the cone K defined as:

$$
w(K) = \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{I})}\left[\sup_{\mathbf{w}\in K\cap\mathbb{S}^{d-1}} \langle \mathbf{z}, \mathbf{w}\rangle\right].
$$

If

$$
m \ge C'\left(\frac{\sigma}{\alpha}\right)^6 w(T_\Sigma(\overline{\mathbf{x}}_\Sigma))^2 + 2C^{-2}\frac{\sigma^2}{\alpha^4}\tau^2,
$$

then $\lambda_{\min}(\mathbf{A}, T_\Sigma(\overline{\mathbf{x}}_\Sigma)) > 0$ with probability at least $1 - \exp(-C\tau^2)$.

The Gaussian width is an important tool in high-dimensional convex geometry and can be interpreted as a measure of the "dimension" of a cone. Except in some specific settings (such as when K is a descent cone of a convex function and other special cases), it is notoriously difficult to compute this quantity; see the discussion in [49]. Another "generic" tool for computing Gaussian widths is based on Dudley's inequality which bounds the width of a set in terms of the covering number of the set at all scales. Estimating the covering number is not easy either in general. This shows the difficulty of computing $w(T_\Sigma(\overline{\mathbf{x}}_\Sigma))$ which we leave to a future work.

Analyzing recovery guarantees in the compressed sensing framework using unsupervised neural networks such as DIP was proposed in [51, 52]. In [51], the authors restricted their analysis to the case of networks without non-linear activations nor training/optimization. The authors of [52] studied the case of the DIP method but their optimization algorithms is

prohibitively intensive necessitating at each iteration retraining the DIP network. Another distinctive difference with our work is that these existing results are uniform relying on RIP-type arguments and their specialization for Gaussian measurements.

3.4.4 Existence and Uniqueness of a Global Strong Solution

We have already stated in Section 3.2 that (2) admits a unique maximal solution. Assumption (8) allows us to further specify this solution as strong and global. Indeed, (11) ensures that the set Θ is uniformly bounded. Let us start by recalling the notion of a strong solution.

Definition 3.6. Denote $\boldsymbol{\theta} : t \in [0, +\infty[\mapsto \boldsymbol{\theta}(t) \in \mathbb{R}^p$. The function $\boldsymbol{\theta}(\cdot)$ is a strong global solution of (2) if it satisfies the following properties:

- $\boldsymbol{\theta}$ is in $\mathcal{C}^1([0, +\infty[; \mathbb{R}^p)$;
- for almost all $t \in [0, +\infty[$, (2) holds with $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$.

Proposition 3.7. *Assume that A-1-A-6 and (8) are satisfied. Then, for any initial condition $\boldsymbol{\theta}_0$, the evolution system (2) has a unique strong global solution.*

Proof. Proposition 3.1 ensures the existence and uniqueness of a maximal solution. Following the discussion after the proof of Proposition 3.1, if $\boldsymbol{\theta}(t)$ is bounded, then we are done. This is precisely what is ensured by Theorem 3.2 under our conditions. □

3.5 Proofs

We start with the following lemmas that will be instrumental in the proof of Theorem 3.2.

Lemma 3.8. *Assume that A-1, A-3, A-5 and A-6 hold. Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (2). Then,*

(i) $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(\cdot)))$ is nonincreasing, and thus converges.
(ii) If $\boldsymbol{\theta}(\cdot)$ is bounded, $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(\cdot)))$ is constant on $\mathfrak{W}(\boldsymbol{\theta}(\cdot))$.

Proof. Let $V(t) = \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))$.

(i) Differentiating $V(\cdot)$, we have for $t > 0$:

$$\begin{aligned}
\dot{V}(t) &= \langle \dot{\mathbf{y}}(t), \nabla_{\mathbf{y}(t)} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \rangle \\
&= \langle \mathcal{J}_{\mathbf{F}}(t) \mathcal{J}_{\mathbf{g}}(t) \dot{\boldsymbol{\theta}}(t), \nabla_{\mathbf{y}(t)} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \rangle \\
&= -\langle \mathcal{J}_{\mathbf{F}}(t) \mathcal{J}_{\mathbf{g}}(t) \mathcal{J}_{\mathbf{g}}(t)^\top \mathcal{J}_{\mathbf{F}}(t)^\top \nabla_{\mathbf{y}(t)} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)), \nabla_{\mathbf{y}(t)} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \rangle \\
&= -\left\| \mathcal{J}_{\mathbf{g}}(t)^\top \mathcal{J}_{\mathbf{F}}(t)^\top \nabla_{\mathbf{y}(t)} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \right\|^2 = -\left\| \dot{\boldsymbol{\theta}}(t) \right\|^2, \qquad (15)
\end{aligned}$$

and thus $V(\cdot)$ is decreasing. Since it is bounded from below (by 0 by assumption), it converges to say $\mathcal{L}_\infty$ (0 in our case).

(ii) Since $\boldsymbol{\theta}(\cdot)$ is bounded, $\mathfrak{W}(\boldsymbol{\theta}(\cdot))$ is non-empty. Let $\boldsymbol{\theta}_\infty \in \mathfrak{W}(\boldsymbol{\theta}(\cdot))$. Then $\exists t_k \to +\infty$ such that $\boldsymbol{\theta}(t_k) \to \boldsymbol{\theta}_\infty$ as $k \to +\infty$. Combining claim (i) with continuity of $\mathcal{L}$, $\mathbf{F}$ and $\mathbf{g}(\cdot, \mathbf{u})$, we have

$$\mathcal{L}_\infty = \lim_{k \to +\infty} \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t_k)))) = \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\infty))).$$

Since this is true for any cluster point, the claim is proved.

□

Lemma 3.9. *Assume that A-1 to A-6 hold. Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (2). If for all $t \geq 0$, $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(t)) \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2} > 0$, then $\|\dot{\boldsymbol{\theta}}(\cdot)\| \in L^1([0,+\infty[)$. In turn, $\lim_{t\to+\infty} \boldsymbol{\theta}(t)$ exists.*

Proof. From Lemma 3.8(i), we have for $t \geq 0$:

$$\mathbf{y}(t) \in [0 \leq \mathcal{L}_{\mathbf{y}}(\cdot) \leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))].$$

We may assume without loss of generality that $\mathbf{y}(t) \in [0 < \mathcal{L}_{\mathbf{y}}(\cdot) \leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))]$ since otherwise $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(\cdot))$ is eventually zero which implies, by Lemma 3.8, that $\dot{\boldsymbol{\theta}}$ is eventually zero, in which case there is nothing to prove.

We are now in position to use the KL property on $\mathbf{y}(\cdot)$. We have for $t > 0$:

$$\begin{aligned}
\frac{\mathrm{d}\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)))}{\mathrm{d}t} &= \psi'(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)))\frac{\mathrm{d}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))}{\mathrm{d}t} \\
&= -\psi'(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)))\left\|\mathcal{J}_{\mathbf{g}}(t)^\top \mathcal{J}_{\mathbf{F}}(t)^\top \nabla_{\mathbf{y}(t)}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 \\
&\leq -\frac{\left\|\mathcal{J}_{\mathbf{g}}(t)^\top \mathcal{J}_{\mathbf{F}}(t)^\top \nabla_{\mathbf{y}(t)}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2}{\left\|\nabla_{\mathbf{y}(t)}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|} \\
&\leq -\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(t))\sigma_{\mathbf{F}}\left\|\mathcal{J}_{\mathbf{g}}(t)^\top \mathcal{J}_{\mathbf{F}}(t)^\top \nabla_{\mathbf{y}(t)}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\| \\
&\leq -\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{F}}}{2}\left\|\dot{\boldsymbol{\theta}}(t)\right\|. \qquad (16)
\end{aligned}$$

where we used A-4 and that $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(t)) \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2} > 0$. Integrating, we get

$$\int_0^t \left\|\dot{\boldsymbol{\theta}}(s)\right\| \mathrm{d}s \leq \frac{2}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{F}}}\left(\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))) - \psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)))\right). \qquad (17)$$

Since $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))$ converges thanks to Lemma 3.8(i) and ψ is continuous and increasing, the right hand side in (17) has a limit. Thus passing to the limit as $t \to +\infty$, we get that $\dot{\boldsymbol{\theta}} \in L^1([0,+\infty[)$. This in turn implies that $\lim_{t\to+\infty} \boldsymbol{\theta}(t)$ exists, say $\boldsymbol{\theta}_\infty$, by applying Cauchy's criterion to

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}_0 + \int_0^t \dot{\boldsymbol{\theta}}(s)\mathrm{d}s.$$

□

Lemma 3.10. *Assume that A-1 to A-6 hold. Recall R and R' from (9). Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of (2).*

(i) If $\boldsymbol{\theta} \in \mathbb{B}(\boldsymbol{\theta}_0, R)$ then

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta})) \geq \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))/2.$$

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

$$\boldsymbol{\theta}(t) \in \mathbb{B}(\boldsymbol{\theta}_0, R').$$

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

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

$$\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\| \leq \mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})\,\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})R \leq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2}.$$

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

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta})) \geq \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) - \|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\| \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2}.$$

(ii) We have for $t > 0$

$$\frac{1}{2}\frac{\mathrm{d}\,\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\|^2}{\mathrm{d}t} = \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\|\,\frac{\mathrm{d}\,\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\|}{\mathrm{d}t} = \langle \dot{\boldsymbol{\theta}}(t), \boldsymbol{\theta}(t) - \boldsymbol{\theta}_0 \rangle,$$

and Cauchy-Schwarz inequality then implies

$$\frac{\mathrm{d}\,\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\|}{\mathrm{d}t} \leq \left\|\dot{\boldsymbol{\theta}}(t)\right\|.$$

Combining this with (17) yields

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\| \leq \int_0^t \left\|\dot{\boldsymbol{\theta}}(s)\right\| \mathrm{d}s \leq \frac{2}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{F}}}\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))),$$

where we argue that $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))$ is positive and bounded and ψ is positive and increasing.

(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 $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(s)) < \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))/2$. 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 $\varepsilon > 0$ and for all $t' \leq t_0 - \varepsilon$, $\boldsymbol{\theta}(t') \in \mathbb{B}(\boldsymbol{\theta}_0, R)$ which, in view of (i) entails that $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta})(t')) \geq \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))/2$. In turn, we get from (ii) that $\boldsymbol{\theta}(t_0 - \varepsilon) \in \mathbb{B}(\boldsymbol{\theta}_0, R')$. Since ε is arbitrary and $\boldsymbol{\theta}$ is continuous, we pass to the limit as $\varepsilon \to 0$ 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.

□

Proof of Theorem 3.2. (i) We here use a standard Lyapunov analysis with several energy functions. Let us reuse $V(t)$. Embarking from (15), we have for $t > 0$

$$\begin{aligned}\dot{V}(t) &= -\left\|\mathcal{J}_{\mathbf{g}}(t)^\top \mathcal{J}_{\mathbf{F}}(t)^\top \nabla_{\mathbf{y}(t)}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 \\ &\leq -\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(t))^2\sigma_{\mathbf{F}}^2\left\|\nabla_{\mathbf{y}(t)}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2,\end{aligned}$$

where we used A-4. In view of Lemma 3.10(iii), we have $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(t)) \geq \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))/2 > 0$ for all $t \geq 0$ if the initialization error verifies (8). Using once again A-2, we get

$$\begin{aligned}\dot{V}(t) &\leq -\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2\sigma_{\mathbf{F}}^2}{4}\left\|\nabla_{\mathbf{y}(t)}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 \\ &\leq -\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2\sigma_{\mathbf{F}}^2}{4\psi'(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)))^2}.\end{aligned}$$

Let Ψ be a primitive of $-\psi'^2$. Then, the last inequality gives

$$\begin{aligned}\dot{\Psi}(V(t)) &= \Psi'(V(t))\dot{V}(t) \\ &\geq \frac{\sigma_{\mathbf{F}}^2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4}.\end{aligned}$$

By integration on $s \in [0,t]$ alongside the fact that Ψ and Ψ^{-1} are (strictly) decreasing functions, we get

$$\begin{aligned}\Psi(V(t)) - \Psi(V(0)) &\geq \frac{\sigma_{\mathbf{F}}^2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4}t \\ V(t) &\leq \Psi^{-1}\left(\frac{\sigma_{\mathbf{F}}^2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))^2}{4}t + \Psi(V(0))\right),\end{aligned}$$

which gives (10).

By Lemma 3.9, $\boldsymbol{\theta}(t)$ converges to some $\boldsymbol{\theta}_\infty$. Continuity of $\mathcal{L}_{\mathbf{y}}(\cdot)$, $\mathbf{F}$ and $\mathbf{g}(\mathbf{u},\cdot)$ implies that

$$0 = \lim_{t\to+\infty}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) = \lim_{t\to+\infty}\mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(t)))) = \mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u},\boldsymbol{\theta}_\infty))),$$

and thus $\boldsymbol{\theta}_\infty \in \operatorname{Argmin}(\mathcal{L}_{\mathbf{y}}(\mathbf{F}(\mathbf{g}(\mathbf{u},\cdot))))$. To get the rate, we argue as in the proof of Lemma 3.10 (ii), replacing $\boldsymbol{\theta}_0$ by $\boldsymbol{\theta}_\infty$, to obtain

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| \leq \int_t^{+\infty}\left\|\dot{\boldsymbol{\theta}}(s)\right\|\mathrm{d}s.$$

We then get by integrating (16) that

$$\begin{aligned}\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| &\leq -\frac{2}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{F}}}\int_t^{+\infty}\frac{\mathrm{d}\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(s)))}{\mathrm{d}s}\mathrm{d}s \\ &\leq \frac{2}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{F}}}\psi(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))).\end{aligned}$$

Thanks to (10), and using that ψ is increasing, we arrive at (11).

(ii) By Lemma 3.9 and continuity of $\mathbf{F}$ and $\mathbf{g}(\mathbf{u}, \cdot)$, we can infer that $\mathbf{y}(\cdot)$ also converges to $\mathbf{y}_\infty = \mathbf{F}(\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\infty))$, where $\boldsymbol{\theta}_\infty = \lim_{t\to+\infty} \boldsymbol{\theta}(t)$. Thus using also continuity of $\mathcal{L}_\mathbf{y}(\cdot)$, we have

$$0 = \lim_{t\to+\infty} \mathcal{L}_\mathbf{y}(\mathbf{y}(t)) = \mathcal{L}_\mathbf{y}(\mathbf{y}_\infty),$$

and thus $\mathbf{y}_\infty \in \operatorname{Argmin}(\mathcal{L}_\mathbf{y}(\cdot))$. Since the latter is the singleton $\{\mathbf{y}\}$ by assumption, we conclude.

In order to obtain the early stopping bound, we use [41, Theorem 5] that links the KL property of $\mathcal{L}_\mathbf{y}(\cdot)$ with an error bound. In our case, this reads

$$\operatorname{dist}(\mathbf{y}(t), \operatorname{Argmin}(\mathcal{L}_\mathbf{y}(\cdot))) = \|\mathbf{y}(t) - \mathbf{y}\| \leq \psi(\mathcal{L}_\mathbf{y}(\mathbf{y}(t))). \tag{18}$$

It then follows that

$$\begin{aligned}
\|\mathbf{y}(t) - \overline{\mathbf{y}}\| &\leq \|\mathbf{y}(t) - \mathbf{y}\| + \|\mathbf{y} - \overline{\mathbf{y}}\| \\
&\leq \psi(\mathcal{L}_\mathbf{y}(\mathbf{y}(t))) + \|\boldsymbol{\varepsilon}\| \\
&\leq \psi\left(\Psi^{-1}\left(\frac{\sigma_\mathbf{F}^2 \sigma_{\min}(\mathcal{J}_\mathbf{g}(t))^2}{4} t + \Psi(V(0))\right)\right) + \|\boldsymbol{\varepsilon}\|.
\end{aligned}$$

Using that ψ is increasing and Ψ is decreasing, the first is bounded by $\|\boldsymbol{\varepsilon}\|$ for all $t \geq \frac{4\Psi(\psi^{-1}(\|\boldsymbol{\varepsilon}\|))}{\sigma_\mathbf{F}^2 \sigma_{\min}(\mathcal{J}_\mathbf{g}(0))^2} - \Psi(V(0))$.

(iii) We have the chain of inequalities

$$\begin{aligned}
\|\mathbf{x}(t) - \overline{\mathbf{x}}\| &\leq \|\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma\| + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma) \\
\text{A-7} &\leq \mu_{\mathbf{F},\Sigma}^{-1} \|\mathbf{y}(t) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma)\| + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma) \\
&\leq \mu_{\mathbf{F},\Sigma}^{-1} (\|\mathbf{y}(t) - \mathbf{y}\| + \|\mathbf{y} - \mathbf{F}(\overline{\mathbf{x}})\| + \|\mathbf{F}(\overline{\mathbf{x}}) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma)\|) + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma) \\
\text{(1), (10), (18)} &\leq \frac{2\psi\left(\Psi^{-1}\left(\gamma(t)\right)\right)}{\mu_{\mathbf{F},\Sigma} \sigma_{\min}(\mathcal{J}_\mathbf{g}(0)) \sigma_\mathbf{F}} + \mu_{\mathbf{F},\Sigma}^{-1} \|\boldsymbol{\varepsilon}\| + \|\mathbf{F}(\overline{\mathbf{x}}) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma)\| + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma).
\end{aligned}$$

By assumption A-6 and the mean value theorem, we have

$$\|\mathbf{F}(\overline{\mathbf{x}}) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma)\| \leq \max_{\mathbf{z}\in[\overline{\mathbf{x}}, \overline{\mathbf{x}}_\Sigma]} \|\mathcal{J}_\mathbf{F}(\mathbf{z})\| \operatorname{dist}(\overline{\mathbf{x}}, \Sigma).$$

Since $0 \in \Sigma$, by Jensen's inequality, we have for all $\mathbf{z} \in [\overline{\mathbf{x}}, \overline{\mathbf{x}}_\Sigma]$, with $\rho \in [0, 1]$:

$$\|\mathbf{z}\| \leq \|\overline{\mathbf{x}}\| + \rho \operatorname{dist}(\overline{\mathbf{x}}, \Sigma) \leq 2\|\overline{\mathbf{x}}\|,$$

meaning that $[\overline{\mathbf{x}}, \overline{\mathbf{x}}_\Sigma] \subset \mathbb{B}(0, 2\|\overline{\mathbf{x}}\|)$. Thus

$$\|\mathbf{F}(\overline{\mathbf{x}}) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma)\| \leq \max_{\mathbf{z}\in\mathbb{B}(0,2\|\overline{\mathbf{x}}\|)} \|\mathcal{J}_\mathbf{F}(\mathbf{z})\| \operatorname{dist}(\overline{\mathbf{x}}, \Sigma). \tag{19}$$

□

4 Case of The Two-Layer DIP Network

This section is devoted to studying under which conditions on the neural network architecture the key condition in (8) is fulfilled. Towards this goal, we consider the case of a two-layer DIP network. Therein, $\mathbf{u}$ is randomly set and kept fixed during the training, and the network is trained to transform this input into a signal that matches the observation $\mathbf{y}$. In particular, we will provide bounds on the level of overparametrization ensuring that (8) holds, which in turn will provide the subsequent recovery guarantees in Theorem 3.2.

4.1 The Two-Layer Neural Network

We take $L = 2$ in Definition 2.1 and thus consider the network defined in (3):

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

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. Observe that it is immediate to account for the bias vector in the hidden layer by considering the bias as a column of the weight matrices $\mathbf{W}$, augmenting $\mathbf{u}$ by 1 and then normalizing to unit norm to comply with forthcoming A-8. The role of the scaling by $\sqrt{k}$ will become apparent shortly, but it will be instrumental to concentrate the kernel stemming from the Jacobian of the network.

In the sequel, we set $C_\phi = \sqrt{\mathbb{E}_{X \sim \mathcal{N}(0,1)} [\phi(X)^2]}$ and $C_{\phi'} = \sqrt{\mathbb{E}_{X \sim \mathcal{N}(0,1)} [\phi'(X)^2]}$. We will assume without loss of generality that $\mathbf{F}(0) = 0$. This is a very mild assumption that is natural in the context of inverse problems, but can be easily removed if needed. We will also need the following assumptions:

Assumptions on the network input and intialization

A-8. $\mathbf{u}$ *is a uniform vector on* $\mathbb{S}^{d-1}$*;*
A-9. $\mathbf{W}(0)$ *has iid entries from* $\mathcal{N}(0, 1)$ *and* $C_\phi, C_{\phi'} < +\infty$*;*
A-10. $\mathbf{V}(0)$ *is independent from* $\mathbf{W}(0)$ *and* $\mathbf{u}$ *and has iid columns with identity covariance and* D*-bounded centered entries.*

4.2 Recovery Guarantees in the Overparametrized Regime

Our main result gives a bound on the level of overparameterization which is sufficient for (8) to hold.

Theorem 4.1. *Suppose that assumptions A-1, A-3, A-5 and A-6 hold. Let* C*,* C' *two positive constants that depend only on the activation function and* D*. Let:*

$$L_{\mathbf{F},0} = \max_{\mathbf{x} \in \mathbb{B}\left(0, C\sqrt{n \log(d)}\right)} \|\mathcal{J}_{\mathbf{F}}(\mathbf{x})\|$$

and

$$L_{\mathcal{L},0} = \max_{\mathbf{v} \in \mathbb{B}\left(0, C L_{\mathbf{F},0} \sqrt{n \log(d)} + \sqrt{m}\left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty + \|\boldsymbol{\varepsilon}\|_\infty\right)\right)} \frac{\|\nabla_{\mathbf{v}} \mathcal{L}_{\mathbf{y}}(\mathbf{v})\|}{\|\mathbf{v} - \mathbf{y}\|}.$$

Consider the one-hidden layer network (3) where both layers are trained with the initialization satisfying A-8 to A-10 and the architecture parameters obeying

$$k \geq C' \sigma_{\mathbf{F}}^{-4} n \psi \left(\frac{L_{\mathcal{L},0}}{2} \left(C L_{\mathbf{F},0} \sqrt{n \log(d)} + \sqrt{m} \left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty + \|\boldsymbol{\varepsilon}\|_\infty \right) \right)^2 \right)^4 .$$

Then (8) holds with probability at least $1 - n^{-1} - d^{-1}$.

Before proving Theorem 4.1, a few remarks are in order.

Remark 4.2 (Dependence on $L_{\mathcal{L},0}$ and $L_{\mathbf{F},0}$). The overparametrization bound on k depends on $L_{\mathcal{L},0}$ and $L_{\mathbf{F},0}$ which in turn may depend on (n, m, d). Their estimate is therefore important. For instance, if $\mathbf{F}$ is globally Lipschitz, as is the case when it is linear, then $L_{\mathbf{F},0}$ is independent of (n, m, d). As far as $L_{\mathcal{L},0}$ is concerned, it is of course independent of (n, m, d) if the loss gradient is globally Lipschitz continuous. Another situation of interest is when $\nabla_{\mathbf{v}} \mathcal{L}_{\mathbf{y}}(\mathbf{v})$ verifies

$$\|\nabla_{\mathbf{v}} \mathcal{L}_{\mathbf{y}}(\mathbf{v}) - \nabla_{\mathbf{z}} \mathcal{L}_{\mathbf{y}}(\mathbf{z})\| \leq \varphi\left(\|\mathbf{v} - \mathbf{z}\|\right), \quad \forall \mathbf{v}, \mathbf{z} \in \mathbb{R}^m,$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing and vanishes at 0. This is clearly weaker than global Lipschitz continuity and covers it as a special case. It also encompasses many important situations such as e.g. losses with Hölderian gradients. It then easily follows, see e.g. [42, Theorem 18.13], that for all $\mathbf{v} \in \mathbb{R}^m$:

$$\mathcal{L}_{\mathbf{y}}(\mathbf{v}) \leq \Phi\left(\|\mathbf{v} - \mathbf{y}\|\right) \quad \text{where} \quad \Phi(s) = \int_0^1 \frac{\varphi(st)}{t} \mathrm{d}t.$$

In this situation, and if $\mathbf{F}$ is also globally $L_{\mathbf{F}}$-Lipschitz, following our line of proof, the overparametrization bound of Theorem 4.1 reads

$$k \geq C' \sigma_{\mathbf{F}}^{-4} n \psi \left(\Phi \left(C L_{\mathbf{F}} \sqrt{n \log(d)} + \sqrt{m} \left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty + \|\boldsymbol{\varepsilon}\|_\infty \right) \right) \right)^4 .$$

Remark 4.3 (Dependence on the desingularizing function). If we now take interest in the scaling of the overparametrization bound on k with respect to (n, m, d) in the general case we obtain that $k \gtrsim \sigma_{\mathbf{F}}^{-4} n \psi (L_{\mathcal{L},0} (L_{\mathbf{F},0}^2 n + m))^4$. Aside from the possible dependence of $L_{\mathcal{L},0}$ and $L_{\mathbf{F},0}$ on the parameters (n, m, d) discussed before, we observe that this bound is highly dependent on the desingularizing function ψ given by the loss function. In the Łojasiewicz case where $\psi = cs^\alpha$ with $\alpha \in [0, 1]$, one can choose to use a sufficiently small α to reduce the scaling on the parameters but then one would slow the convergence rate as described in Corollary 3.3 which implies a tradeoff between the convergence rate and the number of parameters to ensure this convergence.

In the special case where $\alpha = \frac{1}{2}$ which corresponds to the MSE loss, and where $L_{\mathbf{F},0}$ is of constant order and independent of (n, m, d), then the overperametrization of k necessary for ensuring convergence to a zero-loss is $k \gtrsim n^3 m^2$. Another interesting case is when $\mathbf{F}$ is linear. In that setting, the overparametrization bound becomes $k \gtrsim \sigma_{\mathbf{F}}^{-4} n \psi (L_{\mathcal{L},0} (\|\mathbf{F}\|^2 n + m))^4$. By choosing the MSE loss, and thus controlling ψ to be a square root operator, then we obtain that we need $k \gtrsim \kappa(\mathbf{F})^4 n^3 m^2$. The bound is thus more demanding as $\mathbf{F}$ becomes more and 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.

Remark 4.4 (Scaling when $\mathbf{V}$ is fixed). When the linear layer $\mathbf{V}$ is fixed and only $\mathbf{W}$ is trained, the overparametrization bound to guarantee convergence can be improved (see Appendix B and the results in [28]). In this case, one needs $k \gtrsim \sigma_{\mathbf{F}}^{-2} n\psi(L_{\mathcal{L},0}(L_{\mathbf{F},0}^2 n + m))^2$. In particular, for the MSE loss and an operator such that $L_{\mathbf{F},0}$ is of constant order (as is the case when $\mathbf{F}$ is linear), we only need $k \gtrsim n^2 m$. The main reason underlying this improvement is that there is no need in this case to control the deviation of $\mathbf{V}$ from its initial point to compute the local Lipschitz constant of the jacobian of the network. This allows to have a far better Lipschitz constant estimate which turns out to be even global in this case.
Remark 4.5 (Effect of input dimension d). Finally, the dependence on d is far smaller (by a log factor) than the one on n and m. In the way we presented the theorem, it does also affect the probability obtained but it is possible to write the same probability without d and with a stronger impact of n. This indicates that d plays a very minor role on the overparametrization level whereas k is the key to reaching the overparametrized regime we are looking for. In fact, this is demonstrated by our numerical experiments where we obtained the same results by using very small $d \in [1, 10]$ or larger values up to 500, for all our experiments with potentially large n.

4.3 Proofs

We start with the following lemmas that will be instrumental in the proof of Theorem 4.1.
Lemma 4.6 (Bound on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))$ with both layers trained)**.** *Consider the one-hidden layer network* (3) *with both layers trained under assumptions A-5 and A-8-A-10. We have*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) \geq \sqrt{C_\phi^2 + C_{\phi'}^2}/2$$

with probability at least $1 - n^{-1}$ *provided that* $k/\log(k) \geq Cn\log(n)$ *for* $C > 0$ *large enough that depends only on* B, C_ϕ, $C_{\phi'}$ *and* D.

Proof. Define the matrix $\mathbf{H} = \mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)^\top$. Since $\mathbf{u}$ is on the unit sphere, $\mathbf{H}$ reads

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

It then follows that

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

where we used A-8, A-9 and orthogonal invariance of the Gaussian distribution, hence $\mathbf{W}^i(0)\mathbf{u}$ are iid in $\mathcal{N}(0,1)$, as well as A-10 and independence between $\mathbf{V}(0)$ and $\mathbf{W}(0)$. Moreover, since $X \sim \mathcal{N}(0,1)$, we can upper-bound $\phi(X)$ using the Gaussian concentration

inequality as follows:

$$\mathbb{P}\left(\phi(X) \geq \mathbb{E}\left[\phi(X)\right]\sqrt{\log(nk)} + \tau\right) \leq \mathbb{P}\left(\phi(X) \geq \mathbb{E}\left[\phi(X)\right] + \tau\right) \leq \exp\left(-\frac{\tau^2}{2B^2}\right). \tag{20}$$

By choosing $\tau = C\sqrt{2}B\sqrt{\log(nk)}$ with C a positive constant depending on ϕ, we get with a union bound that $\forall i \in [1, k]$:

$$\mathbb{P}\left(\phi(\mathbf{W}^i\mathbf{u})^2 > C\log(nk)\right) < (2n)^{-1}.$$

Thus, with the same probability we observe that

$$\begin{aligned}\lambda_{\max}\left(\frac{1}{k}(\phi'(\mathbf{W}^i(0)\mathbf{u})^2\mathbf{V}_i(0)\mathbf{V}_i(0)^\top + \phi(\mathbf{W}^i(0)\mathbf{u})^2\mathbf{I}_n)\right) &\leq B^2D^2n + C\log(nk)\\ &\leq Cn\log(k).\end{aligned}$$

We can then apply the matrix Chernoff inequality [53, Theorem 5.1.1] to get

$$\mathbb{P}\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) \leq \delta\sqrt{C_{\phi'}^2 + C_\phi^2}\right) \leq ne^{-\frac{(1-\delta)^2k(C_{\phi'}^2+C_\phi^2)}{Cn\log(k)}}.$$

Taking $\delta = 1/2$ and k as prescribed with a sufficiently large constant C, we conclude. □

Lemma 4.7 (Local Lipschitz constant of $\mathcal{J}_{\mathbf{g}}$ with both layers trained)**.** *Suppose that assumptions A-5, A-8 and A-10 are satisfied. For the one-hidden layer network (3) with both layers trained, we have for $n \geq 2$ and any $\rho > 0$:*

$$\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,\rho)}(\mathcal{J}_{\mathbf{g}}) \leq B(1 + 2(D + \rho))\sqrt{\frac{n}{k}}.$$

Proof. Let $\boldsymbol{\theta} \in \mathbb{R}^{k(d+n)}$ (resp. $\widetilde{\boldsymbol{\theta}}$) be the vectorized form of the parameters of the network $(\mathbf{W}, \mathbf{V})$ (resp. $(\widetilde{\mathbf{W}}, \widetilde{\mathbf{V}})$). For $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \mathbb{B}(R, \boldsymbol{\theta}_0)$, we have

$$\begin{aligned}\left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})\right\|^2 &\leq \frac{1}{k}\left(\sum_{i=1}^k \left\|\phi'(\mathbf{W}^i\mathbf{u})\mathbf{V}_i\mathbf{u}^\top - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})\widetilde{\mathbf{V}}_i\mathbf{u}^\top\right\|_F^2 + \left\|\mathrm{ddiag}_n\left(\phi(\mathbf{W}\mathbf{u}) - \phi(\widetilde{\mathbf{W}}\mathbf{u})\right)\right\|_F^2\right)\\ &\leq \frac{1}{k}\left(2\sum_{i=1}^k\left(\left\|\phi'(\mathbf{W}^i\mathbf{u})\left(\mathbf{V}_i - \widetilde{\mathbf{V}}_i\right)\mathbf{u}^\top\right\|_F^2 + \left\|\left(\phi'(\mathbf{W}^i\mathbf{u}) - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})\right)\widetilde{\mathbf{V}}_i\mathbf{u}^\top\right\|_F^2\right)\right.\\ &\qquad\left. + \left\|\mathrm{ddiag}_n\left(\phi(\mathbf{W}\mathbf{u}) - \phi(\widetilde{\mathbf{W}}\mathbf{u})\right)\right\|_F^2\right)\\ &\leq \frac{1}{k}\left(2B^2\sum_{i=1}^k\left(\left\|\mathbf{V}_i - \widetilde{\mathbf{V}}_i\right\|^2 + \left\|\mathbf{W}^i - \widetilde{\mathbf{W}}^i\right\|^2\left\|\widetilde{\mathbf{V}}_i\right\|^2\right) + n\left\|\phi(\mathbf{W}\mathbf{u}) - \phi(\widetilde{\mathbf{W}}\mathbf{u})\right\|^2\right)\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{k}\left(2B^2\left\|\mathbf{V}-\widetilde{\mathbf{V}}\right\|_F^2 + 2B^2\sum_{i=1}^{k}\left\|\mathbf{W}^i-\widetilde{\mathbf{W}}^i\right\|^2\left\|\widetilde{\mathbf{V}}_i\right\|^2 + B^2 n\left\|(\mathbf{W}-\widetilde{\mathbf{W}})\mathbf{u}\right\|^2\right)\\
&\leq \frac{1}{k}\left(2B^2\left\|\mathbf{V}-\widetilde{\mathbf{V}}\right\|_F^2 + 2B^2\max_i\left\|\widetilde{\mathbf{V}}_i\right\|^2\left\|\mathbf{W}-\widetilde{\mathbf{W}}\right\|_F^2 + B^2 n\left\|\mathbf{W}-\widetilde{\mathbf{W}}\right\|_F^2\right)\\
&\leq \frac{n}{k}B^2\left(\left\|\mathbf{V}-\widetilde{\mathbf{V}}\right\|_F^2 + \left\|\mathbf{W}-\widetilde{\mathbf{W}}\right\|_F^2\right) + \frac{2}{k}B^2\max_i\left\|\widetilde{\mathbf{V}}_i\right\|^2\left\|\mathbf{W}-\widetilde{\mathbf{W}}\right\|_F^2\\
&= \frac{n}{k}B^2\left\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right\|^2 + \frac{2}{k}B^2\max_i\left\|\widetilde{\mathbf{V}}_i\right\|^2\left\|\mathbf{W}-\widetilde{\mathbf{W}}\right\|_F^2.
\end{aligned}
$$

Moreover, for any $i \in [k]$:

$$
\left\|\widetilde{\mathbf{V}}_i\right\|^2 \leq 2\left\|\mathbf{V}_i(0)\right\|^2 + 2\left\|\widetilde{\mathbf{V}}_i - \mathbf{V}_i(0)\right\|^2 \leq 2\left\|\mathbf{V}_i(0)\right\|^2 + 2\left\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\right\|^2 \leq 2nD^2 + 2\rho^2,
$$

where we used A-10. Thus

$$
\left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})\right\|^2 \leq \frac{n}{k}B^2\left(1+4D^2+2\rho^2\right)\left\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right\|^2.
$$

□

Lemma 4.8 (Bound on the initial error)**.** *Under assumptions A-5, A-6 and A-8 to A-10, the initial error of the network satisfies*

$$
\|\mathbf{y}(0)-\mathbf{y}\| \leq CL_{\mathbf{F},0}\sqrt{n\log(d)} + \sqrt{m}\left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty + \|\boldsymbol{\varepsilon}\|_\infty\right),
$$

with probability at least $1-d^{-1}$*, where* C *is a constant that depends only on* B*,* C_ϕ*, and* D*.*

Proof. By A-6 and the mean value theorem, we have

$$
\|\mathbf{y}(0)-\mathbf{y}\| \leq \max_{\mathbf{x}\in\mathbb{B}(0,\|\mathbf{x}(0)\|)}\|\mathcal{J}_{\mathbf{F}}(\mathbf{x})\|\,\|\mathbf{x}(0)\| + \sqrt{m}\left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty + \|\boldsymbol{\varepsilon}\|_\infty\right),
$$

where $\mathbf{x}(0) = \mathbf{g}(\mathbf{u},\boldsymbol{\theta}(0)) = \frac{1}{\sqrt{k}}\sum_{i=1}^{k}\phi(\mathbf{W}^i(0)\mathbf{u})\mathbf{V}_i(0)$. Moreover, by A-10:

$$
\|\mathbf{g}(\mathbf{u},\boldsymbol{\theta}(0))\| \leq \max_i\|\mathbf{V}_i(0)\|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left|\phi(\mathbf{W}^i(0)\mathbf{u})\right| \leq D\sqrt{n}\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left|\phi(\mathbf{W}^i(0)\mathbf{u})\right|.
$$

We now prove that the last term concentrates around its expectation. First, owing to A-8 and A-9, we can argue using orthogonal invariance of the Gaussian distribution and independence to infer that

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

In addition, the triangle inequality and Lipschitz continuity of ϕ (see A-5) yields

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

We then get using the Gaussian concentration inequality that

$$\mathbb{P}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left|\phi(\mathbf{W}^i(0)\mathbf{u})\right| \geq C_\phi\sqrt{\log(d)} + \tau\right)$$
$$\leq \mathbb{P}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left|\phi(\mathbf{W}^i(0)\mathbf{u})\right| \geq \mathbb{E}\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left|\phi(\mathbf{W}^i(0)\mathbf{u})\right|\right] + \tau\right) \leq e^{-\frac{\tau^2}{2B^2D^2}}.$$

Taking $\tau = \sqrt{2}BD\sqrt{\log(d)}$, we get

$$\|\mathbf{x}(0)\| \leq C\sqrt{n\log(d)}$$

with probability at least $1 - d^{-1}$. Since the event above implies $\mathbb{B}(0, \|\mathbf{x}(0)\|) \subset \mathbb{B}\left(0, C\sqrt{n\log(d)}\right)$, we conclude. □

Proof of Theorem 4.1. Proving Theorem 4.1 amounts to showing that (8) holds with high probability under our scaling. This will be achieved by combining Lemma 4.6, Lemma 4.7 and Lemma 4.8 as well as the union bound.

From Lemma 4.6, we have

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0)) \geq \sqrt{C_\phi^2 + C_{\phi'}^2}/2$$

with probability at least $1 - n^{-1}$ provided $k \geq C_0 n\log(n)$ for $C_0 > 0$. On the other hand, from Lemma 4.7, and recalling R from (9), we have that R must obey

$$R \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))}{2B((1+2D)+2R))}\sqrt{\frac{k}{n}} \geq \frac{\sqrt{C_\phi^2 + C_{\phi'}^2}}{8B((1/2+D)+R))}\sqrt{\frac{k}{n}}.$$

Solving for R, we arrive at

$$R \geq \frac{\sqrt{(1/2+D)^2 + \frac{\sqrt{(C_\phi^2+C_{\phi'}^2)\frac{k}{n}}}{2B}} - (1/2+D)}{2}.$$

Simple algebraic computations and standard bounds on $\sqrt{1+a}$ for $a \in [0,1]$ show that

$$R \geq C_1 \left(\frac{k}{n}\right)^{1/4}$$

whenever $k \gtrsim n$, C_1 being a positive constant that depends only on B, C_ϕ, $C_{\phi'}$ and D.

Thanks to A-1 and A-3, we have by the descent lemma, see e.g. [42, Lemma 2.64], that

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) \leq \max_{\mathbf{v} \in [\mathbf{y}, \mathbf{y}(0)]} \frac{\|\nabla \mathcal{L}_{\mathbf{y}}(\mathbf{v})\|}{\|\mathbf{v} - \mathbf{y}\|} \frac{\|\mathbf{y}(0) - \mathbf{y}\|^2}{2}.$$

Combining Lemma 4.8 and the fact that

$$[\mathbf{y}, \mathbf{y}(0)] \subset \mathbb{B}(0, \|\mathbf{y}\| + \|\mathbf{y}(0)\|)$$

then allows to deduce that with probability at least $1 - n^{-1}$, we have

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) \leq \frac{L_{\mathcal{L},0}}{2} \left(C L_{\mathbf{F},0} \sqrt{n \log(d)} + \sqrt{m} \left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty + \|\boldsymbol{\varepsilon}\|_\infty \right) \right)^2.$$

Therefore, using the union bound and the fact that ψ is increasing, it is sufficient for (8) to be fulfilled with probability at least $1 - n^{-1} - d^{-1}$, that

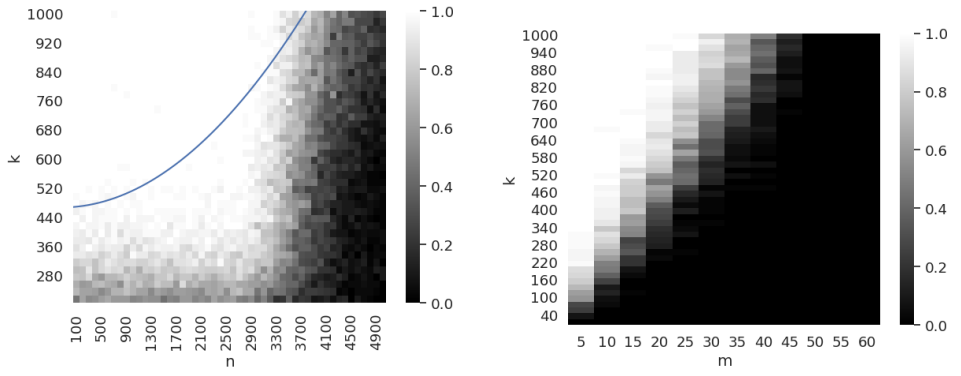
$$\sigma_{\mathbf{F}}^{-1} \psi \left(\frac{L_{\mathcal{L},0}}{2} \left(C L_{\mathbf{F},0} \sqrt{n \log(d)} + \sqrt{m} \left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty + \|\boldsymbol{\varepsilon}\|_\infty \right) \right)^2 \right) < C_1 \left(\frac{k}{n}\right)^{1/4},$$

whence we deduce the claimed scaling. □

5 Numerical Experiments

To validate our theoretical findings, we carried out a series of experiments on two-layer neural networks in the DIP setting. Therein, 25000 gradient descent iterations with a fixed step-size were performed. If the loss reached a value smaller than 10^{-7}, we stopped the training and considered it has converged. For these networks, we only trained the first layer, $\mathbf{W}$, and fixed the second layer, $\mathbf{V}$, as it allows to have better theoretical scalings as discussed in Remark 4.4. Every network was initialized with respect to the assumption of this work where we used sigmoid activation function. The entries of $\overline{\mathbf{x}}$ are drawn from $\mathcal{N}(0,1)$ while the entries of the linear forward operator $\mathbf{F}$ are drawn from $\mathcal{N}(0, 1/\sqrt{n})$ to ensure that $L_{\mathbf{F},0}$ is of constant order.

Our first experiment in Figure 1 studies the convergence to a zero-loss solution of networks with different architecture parameters in a noise-free context. The absence of noise allows the networks to converge faster which is helpful to check convergence in 25000 iterations. We used $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) = \frac{1}{2}\|\mathbf{y}(t) - \mathbf{y}\|^2$ as it should gives good exponential decay. 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 two experiments, in the first one 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.



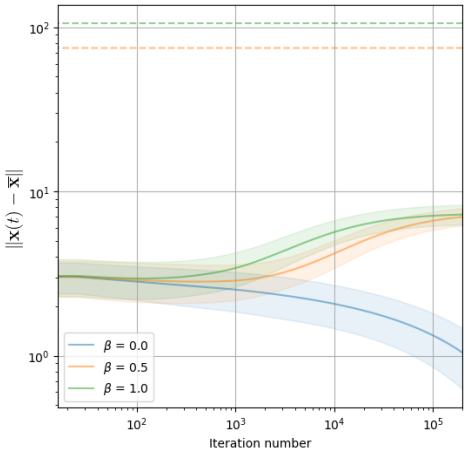
(a) k vs n (b) k vs m

Fig. 1: Probability of converging to a zero-loss solution for networks with different architecture parameters confirming our theoretical predictions: linear dependency between k and m and at least quadratic dependency between k and n. The blue line is a quadratic function representing the phase transition fitted on the data.

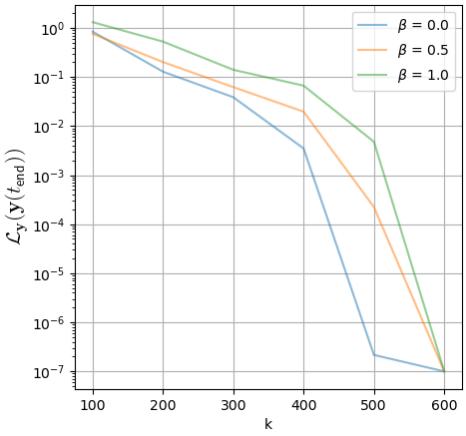
Based on Remark 4.4 concerning Theorem B.1 which is a specialisation of Theorem 4.1, for our experimental setting (MSE loss with $L_{\mathbf{F},0}$ of constant order), one should expect to observe convergence to zero-loss solutions when $k \gtrsim n^2 m$. We observe in Figure 1a the relationship between k and n for a fixed m. In this setup where $n \gg m$ and $\mathbf{A}$ is Gaussian, we expect 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 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 goes away from 0. We first see clearly the expected linear relationship between k and m. However, we used in this experiment $n = 60$ and we can see that for the same range of values of k, the method has much more difficulty to converge with already small m. 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 our bounds are not tight as we observe convergence for lower values of k than expected.

In our second experiment presented in Figure 2a, we look at the signal evolution under different noise levels when the restricted injectivity constraint A-7 is met to verify our theoretical bound on the signal loss. Due to the fact that our networks can span the entirety of the space $\mathbb{R}^n$, this injectivity constraint becomes a global one, which forces us to use a square matrix as our forward operator, we thus chose to use $n = m = 10$. Following the discussion about assumption A-4, we choose to use $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) = \eta(\|\mathbf{y}(t) - \mathbf{y}\|^2)$ with $\eta(s) = s^{p+1}/(2(p+1))$ where $p \in [0, 1]$ with $p = 0.2$ for this specific experiment. We generated once a forward operator with singular values in $\{\frac{1}{z^2+1} \mid z \in [0, 9]\}$ and kept the same one for all the runs. To better see the convergence of the signal, we ran these experiments for 200000 iterations. Furthermore ϵ is a noise vector with entries drawn from a uniform distribution $U(-\beta, \beta)$ with β representing the level of noise.



(a) Signal distance to $\overline{\mathbf{x}}$ for different noise levels. The mean and standard deviation of 50 runs are plotted. The dashed line represents the expectation of the theoretical upper bound of this distance when $t \to +\infty$.



(b) Loss found at time t_{end}, which correspond to the end of the optimization process for networks with varying number of neurons k with three noise levels. Each point is averaged from 50 runs.

Fig. 2: Effect of the noise on both the signal and the loss convergence in different contexts.

In this figure, we plot the mean and the standard deviation of 50 runs for each noise level. For comparison we also show with the dashed line the expectation of the theoretical upper bound, corresponding to $\mathbb{E}\left[\|\varepsilon\| / \mu_{\mathbf{F},\Sigma}\right] \geq \frac{\sqrt{m\beta}}{\sqrt{6}\mu_{\mathbf{F},\Sigma}}$. We observe that the gap between this theoretical bound and the mean of the signal loss is growing as the noise level grows. This indicates that the more noise, the less tighter our bound becomes. We also see different convergence profiles of the signal depending on the noise level which is to be expected as the network will fit this noise to optimize its loss. Of course, when there is no noise, the signal tends to the ground truth thanks to the injectivity of the forward operator.

We continue the study of the effect of the noise on the convergence of the networks in Figure 2b. We show the convergence profile of the loss depending on the noise level and k. For that we fixed $n = 1000$, $m = 10$, $d = 10$, $p = 0.1$ and ran the optimization of networks with different k and β values and we took the loss value obtained at the end of the optimization. The results are averaged from 50 runs and help to see that even if a network with insufficient overparametrization does not converge to a zero-loss solution, the more neurons it has, the better in average the solution in term of loss value. Moreover, this effect seems to stay true even with noise. It is interesting to see the behavior of the loss in such cases that are not treated by our theoretical framework.

For our fourth experiment, we are interested by the effect on the convergence speed of the parameter p of the loss previously described. We fixed $n = 1000$, $m = 10$ and $k = 800$ and varied p between 0 and 1. For each choice of p, we trained 50 networks and show the mean value of the loss at each iteration in Figure 3. We chose to use 10^6 iteration steps and let the optimization reach a limit of 10^{-14}. As expected by corollary 3.3, smaller p values lead

to faster convergence rate in general. Indeed, smaller p values are closer to the case where $\alpha = 1/2$ in the corollary and higher p values means that α will grow away from $1/2$ which worsens the theoretical rate of convergence.

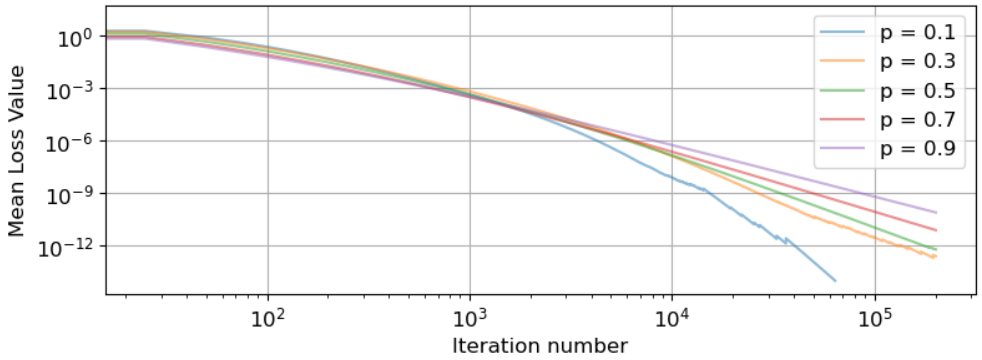


Fig. 3: Convergence profile of different losses parametrized by p. The mean loss values at each iteration of 50 networks are plotted.

6 Conclusion and Future Work

This paper studied the optimization trajectories of neural networks in the inverse problem setting and provided both convergence guarantees for the network and recovery guarantees of the solution. Our results hold for a broad class of loss functions thanks to the Kurdyka-Łojasewiecz inequality. We also demonstrate that for a two-layers DIP network with smooth activation and sufficient overparametrization, we obtain with high probability our theoretical guarantees. Our 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. Then the recovery guarantees are obtained by decomposing the distance to the signal in different error terms explained by the noise, the optimization and the architecture. Although our bounds are not tight as demonstrated by the numerical experiments, they provide a step towards the theoretical understanding of neural networks for inverse problem resolution. In the future we would like to study more thorougly the multilayer case and adapt our result to take into account the ReLU function. Another future direction is to adapt our analysis to the supervised setting and to provide a similar analysis with accelerated optimization methods.

References

[1] Arridge, S., Maass, P., Ozan, O., Schönlieb, C.-B.: Solving inverse problems using data-driven models. Acta Numerica **28**, 1–174 (2019)

[2] Ongie, G., Jalal, A., Metzler, C.A., Baraniuk, R.G., Dimakis, A.G., Willett, R.: Deep Learning Techniques for Inverse Problems in Imaging. IEEE Journal on Selected Areas in Information Theory, 39–56 (2020)

[3] Mukherjee, S., Hauptmann, A., Öktem, O., Pereyra, M., Schönlieb, C.-B.: Learned reconstruction methods with convergence guarantees: a survey of concepts and applications. IEEE Signal Processing Magazine **40**(1), 164–182 (2023)

[4] Li, H., Schwab, J., Antholzer, S., Haltmeier, M.: NETT: solving inverse problems with deep neural networks. Inverse Problems **36**(6), 065005 (2020)

[5] Mukherjee, S., Dittmer, S., Shumaylov, Z., Lunz, S., Öktem, O., Schönlieb, C.-B.: Learned convex regularizers for inverse problems. arXiv preprint arXiv:2008.02839 (2020)

[6] Schwab, J., Antholzer, S., Haltmeier, M.: Deep null space learning for inverse problems: convergence analysis and rates. Inverse Problems **35**(2), 025008 (2019)

[7] Liu, J., Asif, S., Wohlberg, B., Kamilov, U.: Recovery analysis for plug-and-play priors using the restricted eigenvalue condition. Advances in Neural Information Processing Systems **34**, 5921–5933 (2021)

[8] Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9446–9454 (2018)

[9] Prost, J., Houdard, A., Almansa, A., Papadakis, N.: Learning local regularization for variational image restoration. In: International Conference on Scale Space and Variational Methods in Computer Vision, pp. 358–370 (2021)

[10] Venkatakrishnan, S.V., Bouman, C.A., Wohlberg, B.: Plug-and-play priors for model based reconstruction. In: 2013 IEEE Global Conference on Signal and Information Processing, pp. 945–948 (2013)

[11] Monga, V., Li, Y., Eldar, Y.C.: Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. IEEE Signal Processing Magazine **38**(2), 18–44 (2021)

[12] Liu, J., Sun, Y., Xu, X., Kamilov, U.S.: Image restoration using total variation regularized deep image prior. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7715–7719 (2019)

[13] Mataev, G., Milanfar, P., Elad, M.: Deepred: Deep image prior powered by red. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)

[14] Shi, Z., Mettes, P., Maji, S., Snoek, C.G.: On measuring and controlling the spectral bias of the deep image prior. International Journal of Computer Vision **130**(4), 885–908 (2022)

[15] Zukerman, J., Tirer, T., Giryes, R.: Bp-dip: A backprojection based deep image prior. In: 2020 28th European Signal Processing Conference (EUSIPCO), pp. 675–679 (2021). IEEE

[16] Bartlett, P.L., Montanari, A., Rakhlin, A.: Deep learning: a statistical viewpoint. Acta numerica **30**, 87–201 (2021)

[17] Fang, C., Dong, H., Zhang, T.: Mathematical models of overparameterized neural networks. Proceedings of the IEEE **109**(5), 683–703 (2021)

[18] Chizat, L., Oyallon, E., Bach, F.: On lazy training in differentiable programming. Advances in neural information processing systems **32** (2019)

[19] Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In: International Conference on Learning Representations (2019)

[20] Arora, S., Du, S., Hu, W., Li, Z., Wang, R.: Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In: International Conference on Machine Learning, pp. 322–332 (2019)

[21] Oymak, S., Soltanolkotabi, M.: Overparameterized nonlinear learning: Gradient descent takes the shortest path? In: International Conference on Machine Learning, pp. 4951–4960 (2019)

[22] Oymak, S., Soltanolkotabi, M.: Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. IEEE Journal on Selected Areas in Information Theory **1**(1), 84–105 (2020)

[23] Lojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. Coll. du CNRS, Les équations aux dérivées partielles **117**(87-89), 2 (1963)

[24] Łojasiewicz, S.: Sur les trajectoires du gradient d'une fonction analytique. Semin. Geom., Univ. Studi Bologna **1982/1983**, 115–117 (1984)

[25] Kurdyka, K.: On gradients of functions definable in o-minimal structures. In: Annales de L'institut Fourier, vol. 48, pp. 769–783 (1998). Issue: 3

[26] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. Advances in neural information processing systems **31** (2018)

[27] Rahimi, A., Recht, B.: Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. Advances in neural information processing systems **21** (2008)

[28] Buskulic, N., Quéau, Y., Fadili, J.: Convergence guarantees of overparametrized wide deep inverse prior. In: International Conference on Scale Space and Variational Methods in Computer Vision, pp. 406–417 (2023)

[29] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6), 386 (1958)

[30] Nesterov, Y., Polyak, B.T.: Cubic regularization of newton method and its global performance. Mathematical Programming **108**(1), 177–205 (2006)

[31] Absil, P.A., Mahony, R., Andrews, B.: Convergence of the iterates of descent methods for analytic cost functions. SIAM Journal on Optimization **16**(2), 531–547 (2005)

[32] Huang, S.-Z.: Gradient Inequalities. With Applications to Asymptotic Behavior and Stability of Gradient-like Systems. Mathematical Surveys and Monographs, vol. 126. American Mathematical Society, Providence, RI (2006)

[33] Bolte, J., Daniilidis, A., Lewis, A.: The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization **17**(4), 1205–1223 (2007)

[34] Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Mathematical Programming **116**(1), 5–16 (2009)

[35] Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. Mathematics of operations research **35**(2), 438–457 (2010)

[36] Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. Transactions of the American Mathematical Society **362**(6), 3319–3363 (2010)

[37] Forti, M., Nistri, P., Quincampoix, M.: Convergence of neural networks for programming problems via a nonsmooth Łojasiewicz inequality. IEEE Transactions on Neural Networks **17**(6), 1471–1486 (2006)

[38] Simon, L.: Asymptotics for a class of non-linear evolution equations, with applications to geometric problems. Annals of Mathematics **118**(3), 525–571 (1983)

[39] Haraux, A.: A hyperbolic variant of Simon's convergence theorem. In: Lumer, G. (ed.) Evolution equations and their applications in physical and life sciences. Lecture Notes in Pure and Appl. Math., vol. 215 (2001)

[40] Chill, R., Fiorenza, A.: Convergence and decay rate to equilibrium of bounded solutions of quasilinear parabolic equations. Journal of Differential Equations **228**(2), 611–632 (2006)

[41] Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Mathematical Programming **165**, 471–507 (2017)

[42] Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer, Cham (2017)

[43] Coste, M.: An Introduction to O-minimal Geometry. Istituti editoriali e poligrafici

internazionali Pisa, Pisa (2000)

[44] Dries, L.: Tame Topology and O-minimal Structures vol. 248. Cambridge university press, Cambridge (1998)

[45] Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging, 1st edn. Applied Mathematical Sciences. Springer, Cham (2009)

[46] Haraux, A.: Systèmes Dynamiques Dissipatifs et Applications. Recherches en Mathématiques Appliquées, vol. 17. Masson, Paris (1991)

[47] Antun, V., Renna, F., Poon, C., Adcock, B., Hansen, A.C.: On instabilities of deep learning in image reconstruction and the potential costs of ai. Proceedings of the National Academy of Sciences **117**(48), 30088–30095 (2020)

[48] Gottschling, N.M., Antun, V., Adcock, B., Hansen, A.C.: The troublesome kernel—on hallucinations no free lunches and the accuracy-stability trade-off in inverse problems. arXiv preprint arXiv:2001.01258 (2020)

[49] Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.: The convex geometry of linear inverse problems. Foundations of Computational Mathematics **12**(6), 805–849 (2012)

[50] Tropp, J.A.: Convex recovery of a structured signal from independent random linear measurements. Sampling Theory, a Renaissance: Compressive Sensing and Other Developments, 67–101 (2015)

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

[52] Jagatap, G., Hegde, C.: Algorithmic guarantees for inverse imaging with untrained network priors. Advances in neural information processing systems **32** (2019)

[53] Tropp, J.A., *et al.*: An introduction to matrix concentration inequalities. Foundations and Trends® in Machine Learning **8**(1-2), 1–230 (2015)

A Reconstruction Bound in High Signal/Noise Ratio

The reconstruction bound given in (13) relies on assumption A-7 which requires injectivity of $\mathbf{F}$ on Σ. An alternative way of deriving a similar bound only requires to impose restricted injectivity of the jacobian of $\mathbf{F}$ at one point. The trade-off is that it is only valid for low noise level.

Theorem A.1. *Under the setting of Theorem 3.2, let $\mathcal{L}$ be convex and* $\operatorname{Argmin}(\mathcal{L}_{\mathbf{y}}(\cdot)) = \{\mathbf{y}\}$. *Assume that*

A-11. $\ker(\mathcal{J}_{\mathbf{F}}(\overline{\mathbf{x}}_{\Sigma})) \cap T_{\Sigma}(\overline{\mathbf{x}}_{\Sigma}) = 0$.

Denote

$$L_{\mathbf{F}} \stackrel{\text{def}}{=} \max_{\mathbf{x}\in\mathbb{B}(0,2\|\overline{\mathbf{x}}\|)} \|\mathcal{J}_{\mathbf{F}}(\mathbf{x})\| < +\infty. \quad \textit{and} \quad \delta(t) \stackrel{\text{def}}{=} \frac{2\psi\left(\Psi^{-1}\left(\gamma(t)\right)\right)}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(0))\sigma_{\mathbf{F}}}.$$

Then for $\operatorname{dist}(\overline{\mathbf{x}}, \Sigma)$ *and* $\|\varepsilon\|$ *small enough and all* $t > 0$ *sufficiently large, we have*

$$\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \sqrt{\frac{\mu_{\mathbf{F},\Sigma}}{L_{\mathcal{J}_{\mathbf{F}}}}\left(\frac{L_{\mathcal{J}_{\mathbf{F}}}}{\mu_{\mathbf{F},\Sigma}}\operatorname{dist}(\overline{\mathbf{x}}, \Sigma)^2 + \left(\frac{L_{\mathbf{F}}}{\mu_{\mathbf{F},\Sigma}} + 1\right)\operatorname{dist}(\overline{\mathbf{x}}, \Sigma) + \frac{\delta(t) + \|\varepsilon\|}{\mu_{\mathbf{F},\Sigma}}\right)}, \tag{21}$$

where $L_{\mathcal{J}_{\mathbf{F}}} > 0$ *is a constant.*

Proof. Observe that Assumption A-11 implies that $\mu_{\mathbf{F},\Sigma} = \lambda_{\min}(\mathcal{J}_{\mathbf{F}}(\overline{\mathbf{x}}_\Sigma); T_\Sigma(\overline{\mathbf{x}}_\Sigma)) > 0$. Thus we have

$$\begin{aligned}\|\mathbf{x}(t) - \overline{\mathbf{x}}\| &\leq \|\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma\| + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma)\\ &\leq \mu_{\mathbf{F},\Sigma}^{-1}\|\mathcal{J}_{\mathbf{F}}(\overline{\mathbf{x}}_\Sigma)(\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma)\| + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma).\end{aligned}$$

Recall from Theorem 3.2(i) that $\theta(\cdot)$ is bounded, and therefore so is $\mathbf{x}(\cdot)$ by continuity of $\mathbf{g}(\mathbf{u}, \cdot)$; i.e. $\mathbf{x}(t) \in \mathbb{B}(0, \rho)$ for some $\rho > 0$. It then follows from the local Lipschitz continuity assumption on $\mathcal{J}_{\mathbf{F}}$ in A-6 that there exists $L_{\mathcal{J}_{\mathbf{F}}} > 0$ such that for all $\mathbf{z}, \mathbf{z}' \in \mathbb{B}(0, \rho + 2\|\overline{\mathbf{x}}\|)$

$$\|\mathcal{J}_{\mathbf{F}}(\mathbf{z}) - \mathcal{J}_{\mathbf{F}}(\mathbf{z}')\| \leq L_{\mathcal{J}_{\mathbf{F}}}\|\mathbf{z} - \mathbf{z}'\|.$$

In turn, we have

$$\begin{aligned}&\|\mathbf{F}(\mathbf{x}(t)) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma) - \mathcal{J}_{\mathbf{F}}(\overline{\mathbf{x}}_\Sigma)(\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma)\|\\ &= \left\|\int_0^1 \left(\mathcal{J}_{\mathbf{F}}\left(\overline{\mathbf{x}}_\Sigma + t(\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma)\right) - \mathcal{J}_{\mathbf{F}}(\overline{\mathbf{x}}_\Sigma)\right)(\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma)dt\right\|\\ &\leq \frac{L_{\mathcal{J}_{\mathbf{F}}}}{2}\|\mathbf{x}(t) - \overline{\mathbf{x}}\|^2.\end{aligned}$$

Thus

$$\begin{aligned}\|\mathbf{x}(t) - \overline{\mathbf{x}}\| &\leq \mu_{\mathbf{F},\Sigma}^{-1}\left(\|\mathbf{F}(\mathbf{x}(t)) - \mathbf{F}(\overline{\mathbf{x}}_\Sigma)\| + \frac{L_{\mathcal{J}_{\mathbf{F}}}}{2}\|\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma\|^2\right) + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma)\\ &\leq \mu_{\mathbf{F},\Sigma}^{-1}(\|\mathbf{y}(t) - \mathbf{y}\| + \|\mathbf{F}(\overline{\mathbf{x}}_\Sigma) - \mathbf{y}\| + \frac{L_{\mathcal{J}_{\mathbf{F}}}}{2}\|\mathbf{x}(t) - \overline{\mathbf{x}}_\Sigma\|^2) + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma).\end{aligned}$$

From (18), we get

$$\begin{aligned}\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \mu_{\mathbf{F},\Sigma}^{-1}(\delta(t) + \|\varepsilon\| + \|\mathbf{F}(\overline{\mathbf{x}}_\Sigma) - \mathbf{F}(\overline{\mathbf{x}})\| + L_{\mathcal{J}_{\mathbf{F}}}(\|\mathbf{x}(t) - \overline{\mathbf{x}}\|^2 + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma)^2))\\ + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma).\end{aligned}$$

Using (19), we obtain the following second-order polynomial inequality

$$-\frac{L_{\mathcal{J}_{\mathbf{F}}}}{\mu_{\mathbf{F},\Sigma}}\|\mathbf{x}(t)-\overline{\mathbf{x}}\|^2+\|\mathbf{x}(t)-\overline{\mathbf{x}}\| \\ -\mu_{\mathbf{F},\Sigma}^{-1}(\delta(t)+\|\varepsilon\|)-\frac{L_{\mathcal{J}_{\mathbf{F}}}}{\mu_{\mathbf{F},\Sigma}}\operatorname{dist}(\overline{\mathbf{x}},\Sigma)^2-\left(1+\frac{L_{\mathbf{F}}}{\mu_{\mathbf{F},\Sigma}}\right)\operatorname{dist}(\overline{\mathbf{x}},\Sigma)\leq 0.$$

Since $\delta(t)\to 0$, there exists $\tilde{t}>0$ such that $\delta(t)$ is small enough for all $t\geq\tilde{t}$. Thus for all such t and for sufficiently small $\operatorname{dist}(\overline{\mathbf{x}},\Sigma)$ and $\|\varepsilon\|$, we know that the above polynomial has two real positive roots. Solving for $\|\mathbf{x}(t)-\overline{\mathbf{x}}\|$, we get for $\operatorname{dist}(\overline{\mathbf{x}},\Sigma)$ and $\|\varepsilon\|$ small enough and $t\geq\tilde{t}$, that

$$\|\mathbf{x}(t)-\overline{\mathbf{x}}\|\leq\frac{\mu_{\mathbf{F},\Sigma}}{2L_{\mathcal{J}_{\mathbf{F}}}}-\frac{\mu_{\mathbf{F},\Sigma}}{2L_{\mathcal{J}_{\mathbf{F}}}}\left(\sqrt{1-4\frac{L_{\mathcal{J}_{\mathbf{F}}}}{\mu_{\mathbf{F},\Sigma}}(\mu_{\mathbf{F},\Sigma}^{-1}(\delta(t)+\|\varepsilon\|)+\frac{L_{\mathcal{J}_{\mathbf{F}}}}{\mu_{\mathbf{F},\Sigma}}\operatorname{dist}(\overline{\mathbf{x}},\Sigma)^2+\left(1+\frac{L_{\mathbf{F}}}{\mu_{\mathbf{F},\Sigma}}\right)\operatorname{dist}(\overline{\mathbf{x}},\Sigma)}\right) \\ \leq\sqrt{\frac{\mu_{\mathbf{F},\Sigma}}{L_{\mathcal{J}_{\mathbf{F}}}}(\mu_{\mathbf{F},\Sigma}^{-1}(\delta(t)+\|\varepsilon\|)+\frac{L_{\mathcal{J}_{\mathbf{F}}}}{\mu_{\mathbf{F},\Sigma}}\operatorname{dist}(\overline{\mathbf{x}},\Sigma)^2+\left(1+\frac{L_{\mathbf{F}}}{\mu_{\mathbf{F},\Sigma}}\right)\operatorname{dist}(\overline{\mathbf{x}},\Sigma)}.$$

□

B Overparametrization Bound When the Linear Layer is Fixed

In the setting described in Section 4, if one fixes the linear layer, as is usually done in the literature, a better overparametrization bound can be derived.

Theorem B.1. *Under the setting of Theorem 4.1 where*

$$L_{\mathbf{F},0}=\max_{\mathbf{x}\in\mathbb{B}\left(0,C\sqrt{n\log(d)}\right)}\|\mathcal{J}_{\mathbf{F}}(\mathbf{x})\|$$

and

$$L_{\mathcal{L},0}=\max_{\mathbf{v}\in\mathbb{B}\left(0,CL_{\mathbf{F},0}\sqrt{n\log(d)}+\sqrt{m}\left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty+\|\boldsymbol{\varepsilon}\|_\infty\right)\right)}\frac{\|\nabla_{\mathbf{v}}\mathcal{L}_{\mathbf{y}}(\mathbf{v})\|}{\|\mathbf{v}-\mathbf{y}\|},$$

consider the one-hidden layer network (3) where only the first layer is trained with the initialization satisfying A-8-A-10 and the architecture parameters obeying

$$k\geq C'\sigma_{\mathbf{F}}^{-2}n\psi\left(\frac{L_{\mathcal{L},0}}{2}\left(CL_{\mathbf{F},0}\sqrt{n\log(d)}+\sqrt{m}\left(\|\mathbf{F}(\overline{\mathbf{x}})\|_\infty+\|\boldsymbol{\varepsilon}\|_\infty\right)\right)^2\right)^2.$$

Then (8) holds with probability at least $1-n^{-1}-d^{-1}$, where C and C' are positive constants that depend only on the activation function and D.

Proof. The proof follows a very similar pattern as in the case where both layers are trained. The two main changes happen in the bounds on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))$ and $\operatorname{Lip}(\mathcal{J}_{\mathbf{g}})$. First, the constant

on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))$ changes slightly but is still in $O(1)$ as described in lemma B.2. The main change from the previous setting is that $\mathrm{Lip}(\mathcal{J}_{\mathbf{g}})$ is now a global constant given in lemma B.3. We now follow the same structure as in the proof of theorem 4.1 and see that by Lemma B.2 and Lemma B.3 we have that

$$R \geq \frac{C_{\phi'}}{2BD}\sqrt{\frac{k}{n}} \quad \text{thus,} \quad R \geq C_1 \left(\frac{k}{n}\right)^{1/2}.$$

Moreover, let us recall that by combining lemma 4.8 and the fact that

$$[\mathbf{y}, \mathbf{y}(0)] \subset \mathbb{B}(0, \|\mathbf{y}\| + \|\mathbf{y}(0)\|)$$

we can deduce that with probability at least $1 - n^{-1}$,

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) \leq \frac{L_{\mathcal{L},0}}{2}\left(CL_{\mathbf{F},0}\sqrt{n\log(d)} + \sqrt{m}\left(\|\mathbf{F}(\overline{\mathbf{x}})\|_{\infty} + \|\boldsymbol{\varepsilon}\|_{\infty}\right)\right)^2.$$

Therefore, by using a union bound and that ψ is increasing, for (8) to hold with probability $1 - n^{-1} - d^{-1}$, we need

$$\sigma_{\mathbf{F}}^{-1}\psi\left(\frac{L_{\mathcal{L},0}}{2}\left(CL_{\mathbf{F},0}\sqrt{n\log(d)} + \sqrt{m}\left(\|\mathbf{F}(\overline{\mathbf{x}})\|_{\infty} + \|\boldsymbol{\varepsilon}\|_{\infty}\right)\right)^2\right) \leq C_1\left(\frac{k}{n}\right)^{1/2}$$

which gives the claim. □

Lemma B.2 (Bound on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))$)**.** *For the one-hidden layer network (3), under assumptions A-5 and A-8 to A-10, we have*

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

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 $\mathbf{H} = \mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)^{\top}$. 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

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

where we used A-8-A-9 and orthogonal invariance of the Gaussian distribution, hence $\mathbf{W}^i(0)\mathbf{u}$ are iid $\mathcal{N}(0,1)$, as well as A-10 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 [53, Theorem 5.1.1] to get

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

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

Lemma B.3 (Global Lipschitz constant of $\mathcal{J}_{\mathbf{g}}$ with linear layer fixed)**.** *Suppose that assumptions* (A-5)*,* (A-8) *and* (A-10) *are satisfied. For the one-hidden layer network* (3) *with both layers trained, we have for* $n \geq 2$ *and any* $\rho > 0$

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

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

$$\begin{aligned}
\left\|\mathcal{J}(\mathbf{W}) - \mathcal{J}(\widetilde{\mathbf{W}})\right\|^2 &\leq \frac{1}{k}\sum_{i=1}^{k} |\phi'(\mathbf{W}^i\mathbf{u}) - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})|^2 \left\|\mathbf{V}_i\mathbf{u}^\top\right\|_F^2 \\
&= \frac{1}{k}\sum_{i=1}^{k} |\phi'(\mathbf{W}^i\mathbf{u}) - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})|^2 \left\|\mathbf{V}_i\right\|^2 \\
&\leq B^2D^2\frac{n}{k}\sum_{i=1}^{k} |\mathbf{W}^i\mathbf{u} - \widetilde{\mathbf{W}}^i\mathbf{u}|^2 \\
&\leq B^2D^2\frac{n}{k}\sum_{i=1}^{k} \left\|\mathbf{W}^i - \widetilde{\mathbf{W}}^i\right\|^2 = B^2D^2\frac{n}{k}\left\|\mathbf{W} - \widetilde{\mathbf{W}}\right\|_F^2.
\end{aligned}$$

□