Implicit Regularization of the Deep Inverse Prior Trained with Inertia

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

Solving inverse problems with neural networks benefits from very few theoretical guarantees when it comes to the recovery guarantees. We provide in this work convergence and recovery guarantees for self-supervised neural networks applied to inverse problems, such as Deep Image/Inverse Prior, and trained with inertia featuring both viscous and geometric Hessian-driven dampings. We study both the continuous-time case, i.e., the trajectory of a dynamical system, and the discrete case leading to an inertial algorithm with an adaptive step-size. We show in the continuous-time case that the network can be trained with an optimal accelerated exponential convergence rate compared to the rate obtained with gradient flow. We also show that training a network with our inertial algorithm enjoys similar recovery guarantees though with a less sharp linear convergence rate.

Keywords: Deep Inverse Prior Implicit regularization Self-supervised Inverse problems Momentum Hessian damping Convergence Stable recovery

1 Introduction

1.1 Motivation

A ubiquitous problem in science and engineering is to retrieve an unknown signal $\mathbf{x} \in \mathbb{R}^n$ from a noisy indirect observation $\mathbf{y} \in \mathbb{R}^m$. This inverse problem in the linear, finite-dimensional setting is formalized with a forward operator $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ and some additive noise $\boldsymbol{\varepsilon}$ as solving the following equation:

$$\mathbf{y} = \mathbf{A}\overline{\mathbf{x}} + \boldsymbol{\varepsilon}. \tag{1}$$

Throughout this paper, and without loss of generality, we will assume that $\mathbf{y} \in \mathrm{Im}\,(\mathbf{A})$.

While the variational model-based approach with hand-crafted regularizers has been the dominated approach for years to solve (1), data-driven approaches have emerged as powerful methods to solve inverse problems by capturing the prior information directly from data, either partly or completely, explicitly or implicitly. This trend has witnessed a dramatic increase with the rise of machine learning and notably (deep) neural networks [5, 31]. This type of approach has been applied to a variety of problems, and more specifically to solve imaging problems. These networks are simply parametrized functions where the parameters are learned through some gradient-based optimization algorithm to minimize a loss function that depends on the task at hand. Many works have been devoted to the practical aspects of neural networks for inverse problems (see our review later), from the best architecture for a given task to the evaluation of such models. However, while they now yield impressive results for various problems, the theoretical understanding of their recovery properties remains largely lacking.

Our focus in this chapter is the Deep Inverse/Image Prior (DIP), that was introduced in [41] for simple image processing tasks (denoising, super-resolution and in-painting). The central idea in the DIP is to train a neural network which acts as a generator with a randomly generated input that can be thought of as a latent random variable in dimension much smaller than n. The hope is that the architecture of this neural network will induce some "implicit regularization" and will add more and more detailed content during training before overfitting to noise. This already highlights the necessity of an early-stopping strategy that we will make rigorous later in this chapter. The DIP approach has some advantages as it is self-supervised, and does account for the forward model, hence ensuring consistency with observations. Furthermore, it is easy to implement with very good empirical results if an appropriate network architecture is chosen for the task at hand. Recently, we provided convergence and recovery guarantees of the DIP with general loss functions when the network's parameters are trained through gradient flow [10] or gradient descent [11]. In practice however, the parameters are trained through inertia-based methods (such as the widely used ADAM [25]) as they provide empirically faster convergence rates. Inertia-based methods have been actively studied and are known to provably lead to accelerated rates in the convex and strongly convex cases. Motivated by this, we propose to study the trajectories of the DIP neural network parameters when they are trained using inertial optimization dynamics, both in the continuous-time and discrete settings.

1.2 Problem statement

We will consider a feed-forward network $\mathbf{g} : (\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbf{x} \in \mathbb{R}^n$, equipped with some nonlinear activation function ϕ, that transforms an input $\mathbf{u} \in \mathbb{R}^d$ into a vector $\mathbf{x} \in \mathbb{R}^n$. We will restrict ourselves to fully connected multilayer networks that are defined as follows:

Definition 1.1. 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

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

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

$$\mathbf{g} : (\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^d \times \bigtimes_{l=1}^{L} \left(\left(\mathbb{R}^{N_l \times N_{l-1}}\right) \times \mathbb{R}^{N_l}\right) \mapsto \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) \in \mathbb{R}^{N_L}, \text{with } N_L = n,$$

which can be defined recursively as

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

The parameters $\boldsymbol{\theta}$ of the network are a solution of

$$\min_{\theta \in \mathbb{R}^p} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})) \tag{2}$$

where the loss function $\mathcal{L}_{\mathbf{y}} : \mathbb{R}^m \to \mathbb{R}_+, \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) \mapsto \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$ measures the discrepancy between the observation $\mathbf{y}$ and the observed solution of the network $\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta})$. In this work, we will use the Mean Square Error (MSE) as the loss function (see A-1).

We will first study the behavior of the network parameters trajectory in time when trained using the second-order ODE

$$\begin{cases} \ddot{\boldsymbol{\theta}}(t) + \alpha\dot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) = 0 \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0, \dot{\boldsymbol{\theta}}(0) = \mathbf{0}, \end{cases} \tag{DIN}$$

where $\alpha, \beta \geq 0$ and $\mathbf{y}(t) = \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$. This system is coined *Dynamical Inertial Newton-like* (DIN) after [1]. The parameter α corresponds to viscous damping while β is that of geometric Hessian-driven damping. When $\beta = 0$, one recovers the celebrated Polyak Heavy-Ball (HBF) method with friction [35]. Taking $\beta > 0$ has been

shown to attenuate the transversal oscillations that HBF can suffer from. The system (DIN) is known to achieve optimal accelerated convergence rates in both the convex and strongly convex cases when compared to gradient flow [6].

1.3 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 abuse this notation for the forward operator $\mathbf{A}$ and denote its minimum singular value as $\sigma_{\mathbf{A}}$. 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 $\mathbf{M}$. We denote the Kronecker product of matrices as $\otimes$. 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 $a \gtrsim b$ (resp. $a \lesssim b$) if there exists a constant $C > 0$ such that $a \geq Cb$ (resp. $a \leq Cb$).

We also define $\mathbf{x}(t) = \mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t))$ and $\overline{\mathbf{y}} = \mathbf{A}(\overline{\mathbf{x}})$, and we recall that $\mathbf{y}(t) = \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t)) = \mathbf{A}\mathbf{x}(t)$. The Jacobian of the network is denoted $\mathcal{J}_{\mathbf{g}}$. 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 some $\Theta \subset \mathbb{R}^p$, we define $\Sigma_\Theta = \{\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) : \ \boldsymbol{\theta} \in \Theta\}$ as the set of signals that the network $\mathbf{g}$ can generate for all θ in the parameter set Θ. Σ_Θ can thus be viewed as a parametric manifold. If Θ is closed (resp. compact), so is Σ_Θ. We denote $\mathrm{dist}(\cdot, \Sigma_\Theta)$ the distance to Σ_Θ which is well defined if Θ is closed and non-empty. For a vector $\mathbf{x}$, $\mathbf{x}_{\Sigma_\Theta}$ is its projection on Σ_Θ, i.e. $\mathbf{x}_{\Sigma_\Theta} \in \mathrm{Argmin}_{\mathbf{z}\in\Sigma_\Theta} \|\mathbf{x} - \mathbf{z}\|$. Observe that $\mathbf{x}_{\Sigma_\Theta}$ always exists but might not be unique. We also define $T_{\Sigma_\Theta}(\mathbf{x})$ the tangent cone of Σ_Θ at $\mathbf{x} \in \Sigma_\Theta$. The minimal (conic) singular value of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ w.r.t. the cone $T_{\Sigma_\Theta}(\mathbf{x})$ is then defined as

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

1.4 Contributions

We provide a theoretical analysis of the recovery properties of the DIP model for solving linear inverse problems when trained using the inertial system (DIN) in continuous-time, or the corresponding discretized algorithm. In the continuous-time setting, we show that the network can be trained to zero-loss with an (optimal) accelerated exponential convergence rate compared the gradient flow case, as seen in practice, at the cost of a slightly stronger condition on the initialization. We also give an early-stopping bound to avoid overfitting and an accelerated recovery result in the signal space. We show how a sufficiently overparametrized two layer Deep Inverse Prior (DIP) network [41] can meet the conditions to benefit from these guarantees. We also provide an inertial algorithm obtained by appropriate discretization of the continuous-time system. When the algorithm is run with an adaptive step-size to compensate for the lack of global Lipschitz smoothness, we demonstrate that the network can be trained while maintaining comparable recovery guarantees. However, unlike the continuous-time setting, the convergence rate we obtain in the algorithmic case, though linear, is not the optimal accelerated one.

2 Prior Work

Data-Driven Methods to Solve Inverse Problems Our review here is by no means exhaustive and the interested reader may refer to the reviews [5, 31] (among others). A natural, yet naive, way to solve (1) is to learn from pairs of $(\overline{\mathbf{x}}, \mathbf{y})$ a neural network that approximates an analytic "inverse" to the forward operator $\mathbf{A}$. While this approach can provide qualitatively satisfactory results, it does not take into account explicitly the physics of the problem (the forward model (1)), and lacks in particular data consistency. This approach lacks a deep understanding of its recovery guarantees with the only exception of the recent work of [34] who provided a generalization bound, which is motivated by a machine learning perspective rather than an inverse problem one. To overcome some of these shortcomings, the dominant state-of-the-art approach is hybrid, and consists in mixing model- and data-driven methods to get the best of both worlds. There exists a vast array of such hybrid methods among which the most prominent are Plug-and-Play (PnP, see the review in [24]), learned regularization of a variational problem [36], and "unrolling" or "unfolding" methods (see the review [29]). While PnP uses a denoiser network to restrict the range of acceptable signals, one could restrict the set of possible signals to the range of a generative model (see the survey in e.g., [15]). When no or not enough data is available, a well known alternative is the DIP framework [41], and its variants [26, 27, 38, 43, 40].

Several work have studied the theoretical aspects of DIP with various angles. In [20, 19], the authors show various recovery results for early-stopped convolutional networks trained with gradient descent under some training assumptions, and even show that early stopping might not be necessary in some compressive sensing settings. In [22], under similar assumptions, the authors expand on the theory of untrained convolutional networks and on theoretically sounded early-stopping criterion and the associated recovery guarantees. The author of [3] instead study a variant of DIP know as *analytical DIP* which study LISTA like networks, and they show that in that setting, training a network is very similar to solving a Tikhonov regularized problem. Our previous work [10, 11] give recovery guarantees and convergence rates of DIP trained with gradient flow and gradient descent. Our aim in this paper is to study recovery guarantees of the DIP when momentum-based inertial algorithms are used for training. To the best of our knowledge, none of the above reviewed work has studied this setting.

Implicit regularization, Training Dynamics and Overparametrization Neural networks are very high-dimensional non-linear parametric functions that are optimized/trained to minimize a given loss function. This should lead to highly non-convex optimization problems that are known to be challenging due to possibly many local minima and saddle points. Even more so in the context of inverse problems. In fact, even if the neural network is complex enough (overparametrized) to ensure zero empirical error, the set of minimizers may be large. Therefore, it may very well be the case that some minimizers are better than others (e.g. generalize, are stable, etc.). Optimization algorithms such as gradient descent introduce a bias in this choice: an iterative method is biased towards certain solutions of the problem it solves and thus may converge to a solution with certain properties. Since this bias is a by-product rather than an explicitly enforced property, it is known in the literature as *implicit regularization*. This clearly highlights the importance of the optimization algorithm as implicit regularizer, and has played an important role in understanding either statistical learning guarantees of such implicit regularization [8, 16], or the role of implicit regularization for inverse problems [23]. Understanding the role and implications of implicit regularization of an iterative algorithm for learning neural networks to solve inverse problems is at the heart of this chapter.

The modern approach to convergence of neural network training is based on gradient dominated inequalities from which one 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 [13] or gradient descent [14, 4, 32, 33]. 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 $\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}(t))\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}(t))^{\top}$, where $\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}(t))$ 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) [21], i.e., the Jacobian with respect to all the parameters except those of the last layer, and the Random Features Kernel (RF) [37], i.e., the Jacobian corresponding to the parameters of the last linear layer . The goal is then to control the eigenvalues of the kernel to ensure that they remain bounded away from zero, 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. This is also closely related to the celebrated Hartman-Grobman theorem in dynamical systems.

However, these works do not account for the inverse problem setting. Moreover, they only study the gradient flow or gradient descent while inertia or momentum-based algorithms are dominant now. Thus there is a clear need for an analysis targeting recovery guarantees of the DIP method for inverse problems by properly accommodating for the forward operator.

Inertia-based Optimization A large body of literature has been devoted to studying inertial optimization methods that we do not review for obvious space limitation. In the seminal work of Polyak [35], he proposed the HBF system (i.e., setting $\beta = 0$ in (DIN)) which achieves exponential convergence for strongly convex smooth functions with an optimal convergence rate when α is chosen as the square-root of the strong convexity modulus. This system is however no faster than the gradient flow for the non-strongly convex case. It is also known that HBF may suffer traverse oscillations which motivated the introduction of Hessian damping [1]. Note that the Hessian damping term appears as the derivative of the gradient with respect to time, which opens the door to first-order optimization algorithms after proper discretization. System (DIN) and its discretizations have been thoroughly studied in the convex and strongly convex case where α is an asymptotically vanishing viscous damping coefficient, see [6]. In the nonconvex case, (DIN) was studied in [28] and [12] with very promising performance when applied to neural network training. Our work brings together optimization results for inertial dynamics with overparametrization to obtain recovery results of the DIP method when solving linear inverse problems.

3 Continuous-time Setting

We will first analyze the trajectory of the parameters of a network trained through (DIN) as a continuous dynamical system. We start by showing that it is a well-posed system and then present our results showing accelerated convergence guarantees (and an associated recovery bound) compared to the gradient flow case for the right choice of (α, β). We also provide an overparametrization bound under which a two-layer network benefits from these guarantees. We will work under the following assumptions:

A-1. *$\mathcal{L}_{\mathbf{y}}$ is the MSE loss, i.e., $\mathcal{L}_{\mathbf{y}}(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}-\mathbf{y}\|^2$.*
A-2. *$\phi \in \mathcal{C}^1(\mathbb{R})$ and $\exists B > 0$ such that $\sup_{x\in\mathbb{R}} |\phi'(x)| \leq B$ and ϕ' is B-Lipschitz continuous.*

Note also that the MSE loss allows to easily link the loss to its gradient. The MSE case is widely used and we refrain from extending our results to a more general class of KL smooth losses as in [10, 11] to avoid unnecessary technicalities. The above two assumptions ensure that $\boldsymbol{\theta} \mapsto \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}))$ is locally Lipschitz continuous. This will be important when studying local well-posedness of (DIN). Handling rigorously non-smooth activation functions such as the ReLU requires more technicalities, including the use of involved generalized derivatives, that we choose to leave to a future work.

3.1 Well-Posedness

When $\beta > 0$, the second-order dynamical system given in (DIN) can be equivalently formulated as a first-order system both in time and space. We adapt the results given in [7] to show this equivalence.

Theorem 3.1. *Suppose that $\alpha \geq 0$ and $\beta > 0$. Then the following statement are equivalent:*

1. *$\boldsymbol{\theta} : [0, +\infty[\to \mathbb{R}^p$ is a solution trajectory of* (DIN) *with the initial conditions $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$ and $\dot{\boldsymbol{\theta}}(0) = \dot{\boldsymbol{\theta}}_0$.*
2. *$(\boldsymbol{\theta}, \mathbf{q}) : [0, +\infty[\to \mathbb{R}^p \times \mathbb{R}^p$ is a solution trajectory of the first-order system*

$$\begin{cases} \dot{\boldsymbol{\theta}}(t) + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) - \left(\frac{1}{\beta} - \alpha\right) \boldsymbol{\theta}(t) + \frac{1}{\beta}\mathbf{q}(t) & = 0 \\ \dot{\mathbf{q}}(t) - \left(\frac{1}{\beta} - \alpha\right) \boldsymbol{\theta}(t) + \frac{1}{\beta}\mathbf{q}(t) & = 0 \end{cases} \tag{3}$$

with initial conditions $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$ and $\mathbf{q}(0) = \mathbf{q}_0 = -\beta\left(\dot{\boldsymbol{\theta}}_0 + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\right) + (1 - \alpha\beta)\boldsymbol{\theta}_0$.

Proof. 2 $\implies$ 1. We start by differentiating the first equation of (3) which gives

$$\ddot{\boldsymbol{\theta}}(t) + \beta \frac{\mathrm{d}}{\mathrm{d}t} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) - \left(\frac{1}{\beta} - \alpha\right) \dot{\boldsymbol{\theta}}(t) + \frac{1}{\beta}\dot{\mathbf{q}}(t) = 0.$$

We replace $\dot{\mathbf{q}}(t)$ by using the second line of (3) and obtain that

$$\ddot{\boldsymbol{\theta}}(t) + \beta \frac{\mathrm{d}}{\mathrm{d}t} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) - \left(\frac{1}{\beta} - \alpha\right) \dot{\boldsymbol{\theta}}(t) + \frac{1}{\beta}\left(\left(\frac{1}{\beta} - \alpha\right) \boldsymbol{\theta}(t) - \frac{1}{\beta}\mathbf{q}(t)\right) = 0.$$

Now we replace $\mathbf{q}(t)$ by its expression from the first line of (3) and get

$$\ddot{\boldsymbol{\theta}}(t) + \beta \frac{\mathrm{d}}{\mathrm{d}t} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) - \left(\frac{1}{\beta} - \alpha\right) \dot{\boldsymbol{\theta}}(t) + \frac{1}{\beta}\left(\dot{\boldsymbol{\theta}}(t) + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right) = 0.$$

Once simplified, we obtain (DIN). The initial conditions are directly transferable as both $\boldsymbol{\theta}(0)$ and $\dot{\boldsymbol{\theta}}_0$ are defined the same way in both (DIN) and (3)

1 $\implies$ 2. Denoting $\mathbf{q}(t) = \beta\left(-\dot{\boldsymbol{\theta}}(t) - \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) + \left(\frac{1}{\beta} - \alpha\right) \boldsymbol{\theta}(t)\right)$ and differentiating, we get that

$$\dot{\mathbf{q}}(t) = \beta\left(-\ddot{\boldsymbol{\theta}}(t) - \beta \frac{\mathrm{d}}{\mathrm{d}t} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) + \left(\frac{1}{\beta} - \alpha\right) \dot{\boldsymbol{\theta}}(t)\right).$$

In view of $\ddot{\boldsymbol{\theta}}(t)$ in (DIN), we obtain that

$$\dot{\mathbf{q}}(t) = \dot{\boldsymbol{\theta}}(t) + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)).$$

By rearranging the terms and the definition of $\mathbf{q}(t)$, we obtain both expressions of (3). Furthermore, replacing in $\mathbf{q}(0)$ the initial conditions given in (DIN) gives the initial conditions of (3) concluding the proof. □

Theorem 3.1 is valid for any initial condition $\dot{\boldsymbol{\theta}}(0)$, which includes the special case of (DIN) where $\dot{\boldsymbol{\theta}}(0) = \mathbf{0}$. Therefore, from now on, and without loss of generality, we will take $\dot{\boldsymbol{\theta}}(0) = \mathbf{0}$. The reason for this choice will be transparent later. Thanks to this first-order reformulation, we will be able to invoke the Cauchy-Lipschitz theorem to show the existence and uniqueness of a solution of our original system. Towards this goal, we write (3) in the compact form

$$\begin{cases} \dot{\mathbf{z}}(t) + \nabla G\left(\mathbf{z}(t)\right) + D\left(\mathbf{z}(t)\right) = 0 \\ \mathbf{z}(0) = \left(\boldsymbol{\theta}_0, -\beta\left(\beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\right) + (1 - \alpha\beta)\,\boldsymbol{\theta}_0\right), \end{cases} \tag{4}$$

where $\mathbf{z}(t) = (\boldsymbol{\theta}(t), \mathbf{q}(t)) \in \mathbb{R}^p \times \mathbb{R}^p$, $G : \mathbb{R}^p \times \mathbb{R}^p \mapsto (\beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)), \mathbf{0}) \in \mathbb{R}^p \times \mathbb{R}^p$ and $D : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p \times \mathbb{R}^p$ is given by

$$D(\mathbf{z}(t)) = \left(-\left(\frac{1}{\beta} - \alpha\right)\boldsymbol{\theta}(t) + \frac{1}{\beta}\mathbf{q}(t), -\left(\frac{1}{\beta} - \alpha\right)\boldsymbol{\theta}(t) + \frac{1}{\beta}\mathbf{q}(t) \right).$$

Equipped with this condensed form we can show that (4) is well-posed and thus so is (DIN). We start by defining our notion of solution.

Definition 3.2. For $T > 0$, we will say that $\boldsymbol{\theta} : t \in [0, T] \to \mathbb{R}^p$ is a strong solution of (DIN) on $[0, T]$ if the following holds:

- $\boldsymbol{\theta}(\cdot) \in \mathcal{C}^0([0, T])$;
- $\boldsymbol{\theta}(\cdot) \in \mathcal{C}^1$ on every compact set of the interior of $[0, T[$;
- $\dot{\boldsymbol{\theta}}(\cdot)$ is absolutely continuous on every compact set of the interior of $[0, T[$;
- (DIN) holds for almost all $t \in]0, T[$.

A trajectory $\boldsymbol{\theta} : t \in [0, +\infty[\to \mathbb{R}^p$ is a strong global solution of (DIN) if it is a strong solution on $[0, T]$ for any $T > 0$.

Proposition 3.3. *Assume that A-1-A-2 hold and $\alpha \geq 0$ and $\beta \geq 0$. Then there exists $T(\boldsymbol{\theta}_0) \in [0, +\infty[$ and a unique strong solution trajectory $\boldsymbol{\theta}(\cdot)$ of* (DIN) *on $[0, T(\boldsymbol{\theta}_0)]$.*

Proof. Let us start with the case $\beta > 0$. We know by our assumptions and standard differential calculus on $\mathbf{g}(\mathbf{u}, \cdot)$ that $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))$ is locally Lipschitz continuous. Furthermore, the affine operator D is itself globally Lipschitz. Then by the Cauchy-Lipschitz Theorem [18, Theorem 0.4.1], we obtain that (4) has a unique maximal solution $\mathbf{z}(\cdot) \in \mathcal{C}^0([0, T(\boldsymbol{\theta}_0)])$, where the dependence is only on the initial condition $\boldsymbol{\theta}_0$ as we took $\dot{\boldsymbol{\theta}}_0 = 0$. Moreover, $\mathbf{z}(\cdot) \in \mathcal{C}^1$ on every compact set of the interior of $[0, T(\boldsymbol{\theta}_0)[$. This gives us the first item thanks to Theorem 3.1. Since $\boldsymbol{\theta} \in \mathcal{C}^1$ on every compact set of the interior of $[0, T(\boldsymbol{\theta}_0)[$ and $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \cdot))$ is locally Lipschitz continuous, we get that $t \mapsto \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}(t)))$ is Lipschitz continuous, hence absolutely continuous, on every compact set of the interior of $[0, T(\boldsymbol{\theta}_0)[$. The second and third claims then follow from the first line of (3).

For the case $\beta = 0$, we use the standard equivalent first-order system of (DIN) in phase-space (position-velocity) by introducing the velocity variable $\mathbf{v}(t) = \dot{\boldsymbol{\theta}}(t)$. The same reasoning as above yields the claim. □

The time $T(\boldsymbol{\theta}_0)$ is known as the maximal existence time of the solution. By the blow-up alternative, either $T(\boldsymbol{\theta}_0) = +\infty$ and in that case we say that the solution is *global*, or $T(\boldsymbol{\theta}_0) < +\infty$ and the solution blows-up in finite time i.e., $\|\boldsymbol{\theta}(t)\| \to +\infty$ when $t \to T(\boldsymbol{\theta}_0)$. In fact, thanks to Lemma 3.8 and Lemma 3.9 to be stated and proved later, we can show that the local strong solution is actually global provided that α and β are well-chosen and the dynamic is well initialized.

Proposition 3.4. *Assume that A-1-A-2 hold, $\alpha > 0$ and $0 < \beta < 2/\alpha$. Suppose also that* (5) *is verified. Then* (DIN) *has a unique global strong solution.*

Proof. By Proposition 3.3, we know that there exists a unique strong maximal solution to (DIN). Following the above discussion, it is sufficient to show that $\boldsymbol{\theta}(\cdot)$ is bounded. This follows from Lemma 3.8 and Lemma 3.9(iii). □

3.2 Convergence and Recovery Guarantees

We now state in the next theorem how, if a given network obeys some condition on its initialization and is trained with (DIN), we obtain accelerated convergence guarantees of the loss and the network parameters to a zero loss solution, with respect to the guarantees obtained with gradient flow. We also give the associated accelerated early-stopping bound and signal convergence bound.

Theorem 3.5. *Assume that A-1-A-2 hold. Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of* (DIN) *with $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}$ and $\beta = \frac{1}{2\alpha}$ where the initialization $\boldsymbol{\theta}_0$ is such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)) > 0 \ \text{ and } \ R' < R, \tag{5}$$

where R' and R obey

$$R' = \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \ \text{ and } \ R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})} \tag{6}$$

with

$$\xi = 1 + \frac{\kappa(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2\kappa(\mathbf{A})^2}{4} \ \text{ and } \ \eta = \frac{4\max\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}, \frac{1+\sqrt{2}}{2}\right)}{\min\left(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2, \frac{3}{4}\right)}.$$

Then, the following holds:

(i) the loss converges to 0 at the rate

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{2}t\right). \tag{7}$$

Moreover, $\boldsymbol{\theta}(t)$ converges to a global minimizer $\boldsymbol{\theta}_\infty$ at the rate

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| \leq \eta\sqrt{\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{4}t\right). \tag{8}$$

(ii) We have

$$\|\mathbf{y}(t) - \overline{\mathbf{y}}\| \leq 2\|\boldsymbol{\varepsilon}\| \quad \textit{when} \quad t \geq \frac{4}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}\ln\left(\frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}}{\|\boldsymbol{\varepsilon}\|}\right). \tag{9}$$

(iii) If, moreover,

A-3. $\ker(\mathbf{A}) \cap T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}) = \{0\}$ *with* $\Sigma' \overset{\text{def}}{=} \Sigma_{\mathbb{B}_{R'+\|\boldsymbol{\theta}_0\|}}$,

then

$$\begin{aligned}\|\mathbf{x}(t) - \overline{\mathbf{x}}\| \leq \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{4}t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \frac{\|\boldsymbol{\varepsilon}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} \\ + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma').\end{aligned} \tag{10}$$

Proof. See Section 3.5.1 □

3.3 Discussion and consequences

Role of α and accelerated rate The first result of our theorem shows that if the network training is well-initialized, i.e. according to (5), and with an appropriate choice of α and β, the network weights will converge to a zero-loss solution and the loss decreases at an exponential rate that depends on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}$. [10, Theorem 3.2], the authors proved that if $\boldsymbol{\theta}(\cdot)$ is a solution trajectory of the gradient flow

$$\begin{cases}\dot{\boldsymbol{\theta}} + \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) = 0 \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0,\end{cases}$$

and if

$$\frac{2\sqrt{2}}{\sigma_{\mathbf{F}}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))}\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} < R \ \text{ and } \ \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)) > 0,$$

then

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2}{4}t\right). \tag{11}$$

Moreover, $\boldsymbol{\theta}(t)$ converges to a global minimizer $\boldsymbol{\theta}_\infty$ of $\mathcal{L}_{\mathbf{y}}(\mathbf{A}(\mathbf{g}(\mathbf{u},\cdot)))$, at the rate

$$\|\boldsymbol{\theta}(t)-\boldsymbol{\theta}_\infty\| \leq \frac{2}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2}{4}t\right).$$

Clearly, training with gradient flow has also exponential convergence rates. However, compared to (11), our rate in Theorem 3.5 is provably accelerated in the ill-conditioned case. This is expected as it is known for optimization dynamics featuring inertia in the smooth and strongly convex case when α is appropriately tuned as the square-root of the strong convexity modulus [35, 6]. This rate is known to be optimal [30] for this class of objectives. Our setting is of course more intricate and general as our problem is nonconvex. One also observes that the effect of acceleration depends on the conditioning of the forward operator and specifically, the worse the conditioning, the better the acceleration. However, whereas in the gradient flow case the multiplying constant in the rate depends solely on $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))$, the extra-term ξ in the constant in the rates of Theorem 3.5 reveals a quadratic dependence on the condition numbers $\kappa(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))$ and $\kappa(\mathbf{A})$. This is a mild price to pay compared to the exponential gain in the rate.

The viscous and geometric damping parameters α and β were optimized to achieve the (optimal) accelerated exponential rate. However, one has to keep in mind that this rate only holds in the initialization regime where (5) is true (which will turn out to hold in the overparameterized regime as we will show in the forthcoming section). Since both R' and the bound (8) on the convergence of the network parameters grow linearly with η, it is tempting to make η as small as possible by adjusting α and β as η clearly depends on them, see (17). However, minimizing η in such a way should be done without harming the exponential convergence rate, particularly in the ill-conditioned setting i.e. $\sigma_{\mathbf{A}}$ small. In fact, minimizing (17) in α and β suggests to take $\beta = 1/\alpha$ and α as a constant. In turn, from (21), the convergence rate would be $O(1)$, which is vacuous. Even choosing β arbitrarily close to, but strictly less than, $1/\alpha$, would result in a rate $O\left(\exp(-c\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2 t)\right)$, for some $c>0$, when $\sigma_{\mathbf{A}}$ is small. This is the same rate as the gradient flow which cancels out the advantage brought by inertia. On the other hand, our choice of α and β leads to the (optimal) accelerated rate, but does not minimize η, which will scale as $O(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^{-2}\sigma_{\mathbf{A}}^{-2})$ for the ill-conditioned case. η would scale as $O(\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^{-1}\sigma_{\mathbf{A}}^{-1})$ when minimizing it in α and β.

Early stopping While (7) ensures convergence to a zero-loss solution, it does so by overfitting the noise inherent to the inverse problem. A classical way to avoid this is to use an early stopping strategy, hence ensuring that the solution in the observation space will lie in a ball around the sought after observation $\overline{\mathbf{y}}$. This is precisely what (9) states. It is worth mentioning that early stopping has been used by practitioners of the DIP model trained with gradient descent and our results give this intuition firm theoretical grounds. In view of our discussion on the rate accelerated above, it is clear that our early stopping bound is much better than that of [10, Theorem 3.2] for gradient flow.

Signal recovery Similarly to the case of gradient flow, see [10, Theorem 3.2], our recovery bound on $\overline{\mathbf{x}}$ in (10) is the sum of three terms. The last two ones correspond respectively to the "noise error" inherent to the forward model, and the "modeling error" which captures the expressivity of the trained network, i.e. its ability to generate solutions close to $\overline{\mathbf{x}}$. These two terms are exactly the same as those in [10, Theorem 3.2]. The first term (10) is an "optimization error". This is where the role of inertia is important and this error in our case is much smaller as discussed above.

The bounds in (10) depend on $\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))$, the minimal conic singular value, which is bounded away from zero thanks to the restricted injectivity condition (A-3). This is a classical and minimal assumption in the inverse problem literature if one hopes for recovering $\overline{\mathbf{x}}$ even in the noiseless case. Assuming the rows of $\mathbf{A}$ are linearly independent, one easily checks that (A-3) imposes that $m \geq \dim(T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))$. As it was also observed in [10], there is a trade-off between the restricted injectivity condition (A-3) and the expressivity of the network. If the model is highly expressive then $\operatorname{dist}(\overline{\mathbf{x}}, \Sigma')$ will be smaller. But this is likely to come at the cost of making $\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))$ decrease, as restricted injectivity may be required to hold on a larger subset (cone). Although this observation is to be tempered as the dimension $T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'})$ does not necessarily increase as Σ' gets larger. This discussion relates with the work on the instability phenomenon observed in learned reconstruction methods [2, 17]. In fact, one fundamental problem that creates these instabilities in the reconstruction is that $\ker(\mathbf{A})$ can be non-trivial. The restricted injectivity condition guarantees stable reconstruction but the error bound degrades with decreasing $\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))$.

3.4 Wide Two-Layer DIP Network

It is now natural to ask when a network obeys (5) and thus enjoys the convergence and reconstruction guarantees of Theorem 3.5. Our way of ensuring this is to be in a sufficiently overparametrized regime; see Section 2 for a review of the role of overparametrized when training neural networks. Informally, the question pertains to determining, for a network architecture and a random initialization, the number of neurons or parameters of the network to ensure the validity of (5) with high probability. Indeed, good statistical properties arise from overparametrized networks, enabling control over the eigenspace of the network Jacobian at initialization. Similar to other related works, we will primarily focus on studying shallow networks. Extensions to deeper network are beyond the scope of this chapter.

Recall that in our self-supervised DIP setting, the input $\mathbf{u}$ is sampled randomly and fixed during training. The network is then trained to map $\mathbf{u}$ to a signal $\mathbf{x}$ such that $\mathbf{A}\mathbf{x}$ is close to $\mathbf{y}$. We use a one-hidden layer network by taking $L = 2$ in Definition 1.1, which we write as

$$\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}) = \frac{1}{\sqrt{k}} \mathbf{V}\phi(\mathbf{W}\mathbf{u}) \tag{12}$$

with $\mathbf{V} \in \mathbb{R}^{n\times k}$ and $\mathbf{W} \in \mathbb{R}^{k\times d}$, and ϕ an element-wise nonlinear activation function. To establish our overparametrization bound, we will impose the following assumptions where we define $C_\phi = \sqrt{\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\phi(X)^2\right]}$ and $C_{\phi'} = \sqrt{\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\phi'(X)^2\right]}$[1]:

Assumptions on the network input and initialization

A-4. $\mathbf{u}$ *is a uniform vector on* $\mathbb{S}^{d-1}$*;*
A-5. $\mathbf{W}(0)$ *has iid entries from* $\mathcal{N}(0,1)$ *and* $C_\phi < +\infty$*;*
A-6. $\mathbf{V}(0)$ *is independent from* $\mathbf{W}(0)$ *and* $\mathbf{u}$*, and its entries are zero-mean independent* D*-bounded random variables of unit variance.*

These assumptions are quite standard for neural networks and very easy to verify. For A-6, as an example, one can use iid entries chosen from the uniform distribution on a compact interval. We can now state our overparametrization bound under which (5) holds, and thus, so do the guarantees of Theorem 3.5. We denote the signal-to-noise ratio as $\mathrm{SNR} = \|\mathbf{A}\overline{\mathbf{x}}\| \,/\, \|\boldsymbol{\varepsilon}\|$.

Theorem 3.6. *Suppose that assumptions A-1 and A-2 hold. Consider the one-hidden layer network* (12) *where both layers are trained with the initialization satisfying A-4 to A-6 and the architecture parameters obeying*

$$k \geq C(1+\kappa(\mathbf{A})^4)\frac{\max\left(\sigma_{\mathbf{A}}^4, c_1\right)}{\min\left(\sigma_{\mathbf{A}}^8, c_2\right)} n\left(\|\mathbf{A}\|^4 n^2 + \|\mathbf{A}\overline{\mathbf{x}}\|_\infty^4 \left(1+\mathrm{SNR}^{-1}\right)^4 m^2\right).$$

Then (5) *holds with probability at least* $1 - 5e^{-(n-1)} - 2n^{-1}$*. Here* $c_1, c_2, C > 0$ *are absolute constants that depend only on* $C_\phi, C_{\phi'}, B$ *and* D*.*

Proof. See Section 3.5.2. □

The overparametrization bound scales as $k \geq n^3 + nm^2$, which is similar to gradient flow [10, Theorem 4.1]. However, as we discussed in Section 3.3, training with (DIN) achieves an optimal exponential rate but at the price of the initialization condition which becomes more stringent as the conditioning of $\mathbf{A}$ degrades. This is clearly reflected in our overparametrization bound. Indeed, in the extremely ill-conditioned case, we have an extra multiplying factor that scales as $\kappa(\mathbf{A})^4$ compared to [10, Theorem 4.1]. Whether this can be improved to get the best of both worlds is an open question that we leave to a future work.

We observe again that the set Σ' on which A-3 is required to hold is random. Nevertheless, using similar arguments as in [10, Remark 4.2], one can show that $\Sigma' \subset \Sigma_{\mathbb{B}_\rho(0)}$, where

$$\rho \lesssim \frac{\max\left(\sigma_{\mathbf{A}}, (c_1)^{\frac{1}{4}}\right)}{\min\left(\sigma_{\mathbf{A}}^2, (c_2)^{\frac{1}{4}}\right)}(1+\kappa(\mathbf{A}))\left(\|\mathbf{A}\|\sqrt{n} + \|\mathbf{A}\overline{\mathbf{x}}\|_\infty\left(1+\mathrm{SNR}^{-1}\right)\sqrt{m}\right) + \sqrt{k}\left(\sqrt{n}+\sqrt{d}\right)$$

[1] Observe that $C_{\phi'} \leq B$ under A-2.

with probability at least $1 - 5e^{-(n-1)} - 2e^{-kd} - 2n^{-1}$. In the overparametrized regime, ρ scales as $O\left(\sqrt{k}\left(\sqrt{n}+\sqrt{d}\right)\right)$. This confirms the intuitively expected behaviour that expressivity of Σ' is better as the overparametrization increases.

3.5 Proofs

3.5.1 Proof of Theorem 3.5

We will start by showing some intermediate lemmas necessary to prove our main theorem. For these proofs, we will use the following Lyapunov function given in the original work of [1]:

$$V(t) = \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) + \frac{1}{2}\left\|\dot{\boldsymbol{\theta}}(t) + \beta\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2. \tag{13}$$

We prove in the following lemma that $V(t)$ converges which is then used in the proof of Proposition 3.4 to obtain that $\boldsymbol{\theta}$ is a global solution of (DIN).

Lemma 3.7. *Assume that A-1-A-2 hold, $\alpha > 0$ and $0 \leq \beta \leq \frac{2}{\alpha}$. Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of* (DIN). *Then,*

(i) $V(t)$ is nonincreasing and converges.

(ii) $\dot{\boldsymbol{\theta}}(\cdot) \in L^2([0,+\infty[)$. If $\beta \in]0, 2/\alpha[$, then $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(\cdot)) \in L^2([0,+\infty[)$.

(iii) If $\boldsymbol{\theta}(\cdot)$ is bounded and $\beta \in]0, 2/\alpha[$, then $\lim_{t\to+\infty}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\| = 0$.

Proof. We start by differentiating the Lyapunov function $V(t)$ and obtain that

$$\begin{aligned}
\dot{V}(t) &\leq \langle\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)), \dot{\boldsymbol{\theta}}(t)\rangle + \langle\dot{\boldsymbol{\theta}}(t) + \beta\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)), \ddot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\rangle \\
&\leq \langle\dot{\boldsymbol{\theta}}(t), \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) + \ddot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\rangle \\
&\quad + \beta\langle\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)), \ddot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\rangle.
\end{aligned}$$

We now replace $\ddot{\boldsymbol{\theta}}(t) + \beta\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))$ by using (DIN) which gives

$$\dot{V}(t) \leq -\alpha\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 - \beta\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 + \beta\alpha\langle\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)), -\dot{\boldsymbol{\theta}}(t)\rangle.$$

Applying Young's inequality we get

$$\begin{aligned}
\dot{V}(t) &\leq -\alpha\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 - \beta\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 + \frac{\beta^2\alpha}{2}\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 + \frac{\alpha}{2}\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 \\
&\leq -\frac{\alpha}{2}\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 - \beta\left(1 - \frac{\beta\alpha}{2}\right)\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2.
\end{aligned} \tag{14}$$

We get claim (i) as V is nonnegative and given the choice of β. Integrating (14), we also obtain claim (ii).

By our assumptions, we know that $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \cdot))$ is locally Lipschitz continuous. By the boundedness assumption, we have that $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(\cdot)) \in L^\infty([0,+\infty[)$. Moreover, since $\dot{\boldsymbol{\theta}}(\cdot) \in L^2([0,+\infty[)$ and is continuous, then $\dot{\boldsymbol{\theta}}(\cdot) \in L^\infty([0,+\infty[)$. These facts imply that there exists $L > 0$ such that for every $s, t \geq 0$

$$\begin{aligned}
&\left|\,\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(t))\|^2 - \|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(s))\|^2\,\right| \\
&\qquad \leq 2\sup_{\tau\geq 0}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(\tau))\|\,\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(t)) - \nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(s))\| \\
&\qquad \leq 2L\sup_{\tau\geq 0}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(\tau))\|\,\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}(s)\| \\
&\qquad \leq 2L\sup_{\tau\geq 0}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(\tau))\|\sup_{u\geq 0}\left\|\dot{\boldsymbol{\theta}}(u)\right\|\,|t - s|,
\end{aligned}$$

and thus $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{y}(\cdot))\|^2$ is uniformly continuous. Since it is also integrable, Barbălat's lemma [39, Lemma 1] yields claim (iii).

□

Lemma 3.8. *Assume that A-1 and A-2 hold, $\alpha > 0$ and $0 < \beta < \frac{2}{\alpha}$. Let $\boldsymbol{\theta}(\cdot)$ be a solution trajectory of* (DIN). *If for all $t \geq 0$, $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}(t))) \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))}{2} > 0$, then $\dot{\boldsymbol{\theta}}(\cdot) \in L^1([0, +\infty[)$. In turn, $\lim_{t\to+\infty} \boldsymbol{\theta}(t)$ exists.*

Proof. By assumption, we have for all $t > 0$ that

$$\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\|^2 = \left\|\mathcal{J}_{\mathbf{g}}(t)^\top \mathbf{A}^\top (\mathbf{y}(t) - \mathbf{y})\right\|^2 \geq 2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2 \sigma_{\mathbf{A}}^2 \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \tag{15}$$

where, in the inequality, we used that $\mathbf{y}(t) - \mathbf{y} \in \mathrm{Im}\,(\mathbf{A}) = \ker(\mathbf{A}^\top)^\perp$. This argument will be used repeatedly though we will not specify it. Now if we proceed to our Lyapunov function $V(t)$, we observe that

$$\begin{aligned}
V(t) &= \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) + \frac{1}{2}\left\|\dot{\boldsymbol{\theta}}(t) + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 \\
&\leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) + \left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 + \beta^2 \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\|^2 \\
&\overset{(15)}{\leq} \left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 + \left(\beta^2 + \left(2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2 \sigma_{\mathbf{A}}^2\right)^{-1}\right) \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\|^2 \\
&\leq \max\left(1, \beta^2 + \left(2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2 \sigma_{\mathbf{A}}^2\right)^{-1}\right) \left(\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 + \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\|^2\right).
\end{aligned} \tag{16}$$

We now look at $\frac{\mathrm{d}V(t)^{1/2}}{\mathrm{d}t}$. Without loss of generality, we assume that $V(t) \neq 0$ as otherwise $V(s) = 0$ for all $s \geq t$ (remember that V is nonincreasing), and thus $\dot{\boldsymbol{\theta}}(s) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(s)) = 0$ for all $s \geq t$, and there is nothing to prove. We have the following chain of inequalities

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{V(t)} &= \frac{\dot{V}(t)}{2\sqrt{V(t)}} \\
&\overset{(14)}{\leq} \frac{-\alpha/2\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 - \beta\left(1 - \frac{\beta\alpha}{2}\right)\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\|^2}{2\sqrt{V(t)}} \\
&\overset{(16)}{\leq} -\frac{\min\left(\alpha/2, \beta\left(1 - \frac{\beta\alpha}{2}\right)\right)\left(\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 + \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\|^2\right)^{1/2}}{2\max\left(1, \beta + \left(\sqrt{2}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}\right)^{-1}\right)} \\
&\leq -\eta^{-1}\left(\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 + \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\|^2\right)^{1/2},
\end{aligned}$$

where we let

$$\eta = \frac{2\max\left(1, \beta + \left(\sqrt{2}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}\right)^{-1}\right)}{\min\left(\alpha/2, \beta\left(1 - \frac{\beta\alpha}{2}\right)\right)}. \tag{17}$$

Integrating, we get

$$\begin{aligned}
\int_0^t \left\|\dot{\boldsymbol{\theta}}(s)\right\| \mathrm{d}s &\leq \int_0^t \left(\left\|\dot{\boldsymbol{\theta}}(s)\right\|^2 + \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(s))\|^2\right)^{1/2} \mathrm{d}s \\
&\leq -\eta \int_0^t \frac{\mathrm{d}V(s)^{1/2}}{\mathrm{d}s} \mathrm{d}s \\
&\leq \eta\sqrt{V(0)} \\
&\leq \eta\left(\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) + \frac{\beta^2}{2}\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\|^2\right)^{1/2}.
\end{aligned}$$

where in the last inequality we used that $\dot{\boldsymbol{\theta}}(0) = \mathbf{0}$. In view of A-1, we have

$$\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\| = \left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)^\top \mathbf{A}^\top (\mathbf{y}(t) - \mathbf{y})\right\| \leq \|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\|\,\|\mathbf{A}\|\sqrt{2\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}. \tag{18}$$

Combining this with (3.5.1), we obtain

$$\int_0^t \left\|\dot{\boldsymbol{\theta}}(s)\right\| \mathrm{d}s \leq \eta \sqrt{\left(1+\beta^2 \left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\right\|^2 \|\mathbf{A}\|^2\right) \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}. \tag{19}$$

Passing to the limit, we get that $\dot{\boldsymbol{\theta}}(\cdot) \in L^1([0,+\infty[)$ and thus, $\lim_{t\to+\infty} \boldsymbol{\theta}(t)$ exists by applying Cauchy's criterion to

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

□

Lemma 3.9. *Assume that A-1 and A-2 hold,* $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}$ *and* $\beta = \frac{1}{2\alpha}$. *Recall* R *and* R' *from* (6). *Let* $\boldsymbol{\theta}(\cdot)$ *be a solution trajectory of* (DIN).

(i) If $\boldsymbol{\theta} \in \mathbb{B}_R(\boldsymbol{\theta}_0)$ *then*

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

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

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

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

Proof. (i) Similar to [10, Lemma 3.11(i)].

(ii) Using (19), we have for $t > 0$

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\| \leq \int_0^t \left\|\dot{\boldsymbol{\theta}}(s)\right\| \mathrm{d}s \leq \eta \sqrt{\left(1+\beta^2 \left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\right\|^2 \|\mathbf{A}\|^2\right) \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}.$$

Replacing α and β in (17) by their values according to our choice, the rhs of the last inequality is precisely R', whence we get the claim.

(iii) Similar to [10, Lemma 3.11(iii)].

□

Proof of Theorem 3.5. (i) We follow a standard Lyapunov analysis. By Jensen's inequality,

$$-\left\|\dot{\boldsymbol{\theta}}(t)\right\|^2 \leq -\frac{1}{2}\left\|\dot{\boldsymbol{\theta}}(t) + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 + \beta^2 \left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2.$$

Combining this with (14) gives

$$\begin{aligned}
\dot{V}(t) &\leq -\frac{\alpha}{4}\left\|\dot{\boldsymbol{\theta}}(t) + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 - \beta\left(1-\beta\alpha\right)\left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 \\
&\overset{(15)}{\leq} -\frac{\alpha}{4}\left\|\dot{\boldsymbol{\theta}}(t) + \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))\right\|^2 \\
&\qquad - 2\beta\left(1-\beta\alpha\right)\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2 \sigma_{\mathbf{A}}^2 \mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \\
&\leq -\min\left(\frac{\alpha}{2}, 2\beta\left(1-\beta\alpha\right)\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2 \sigma_{\mathbf{A}}^2\right) V(t).
\end{aligned} \tag{20}$$

Integrating, we obtain

$$\begin{aligned}
\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) &\leq V(t) \\
&\leq V(0) \exp\left(-\min\left(\frac{\alpha}{2}, 2\beta\left(1-\beta\alpha\right)\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2 \sigma_{\mathbf{A}}^2\right) t\right).
\end{aligned} \tag{21}$$

The optimal rate is obtained by setting $\beta = \frac{1}{2\alpha}$ and $\alpha = \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}$ corresponding to the choice in the theorem, hence leading to

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t)) \leq V(t) \leq V(0) \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{2} t\right). \tag{22}$$

By assumption we set $\dot{\boldsymbol{\theta}}(0) = \mathbf{0}$ which means that

$$V(0) = \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0)) + \frac{\beta^2}{2} \left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))\right\|^2.$$

From (18), we get that

$$V(0) \leq \xi \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))$$

where $\xi = 1 + \beta^2 \left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\right\|^2 \left\|\mathbf{A}\right\|^2$. Replacing β with its value in the expression of ξ and plugging into (22) concludes the proof of (7).

By Lemma 3.8, we know that $\boldsymbol{\theta}(\cdot)$ converges to some θ_∞. We use (22) and a similar reasoning as for (3.5.1) to obtain

$$\begin{aligned}
\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_\infty\| &\leq \int_t^{+\infty} \left\|\dot{\boldsymbol{\theta}}(s)\right\| \mathrm{d}s \\
&\leq \eta \sqrt{V(t)} \\
&\leq \eta \sqrt{V(0)} \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{4} t\right) \\
&\leq \eta \sqrt{\xi \mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{4} t\right)
\end{aligned}$$

which shows (8).

(ii) Continuity of $\mathbf{A}$ and $\mathbf{g}(\mathbf{u}, \cdot)$ indicate that $\mathbf{y}(\cdot)$ also converges to $\mathbf{y}_\infty = \mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\infty)$. The early stopping bound can be obtained by using (7). Observe that

$$\begin{aligned}
\|\mathbf{y}(t) - \overline{\mathbf{y}}\| &\leq \|\mathbf{y}(t) - \mathbf{y}\| + \|\mathbf{y} - \overline{\mathbf{y}}\| \\
&\leq \sqrt{2\mathcal{L}_{\mathbf{y}}(\mathbf{y}(t))} + \|\boldsymbol{\varepsilon}\| \\
&\leq \sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{4} t\right) + \|\boldsymbol{\varepsilon}\|.
\end{aligned}$$

Thus, choosing $t \geq \frac{4}{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} \log\left(\frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))}}{\|\boldsymbol{\varepsilon}\|}\right)$ gives (9).

(iii) We recall that by Lemma 3.9, $\boldsymbol{\theta}(t) \in \mathbb{B}_{R'}(\boldsymbol{\theta}_0)$ for all $t \geq 0$, which in turn entails that $\mathbf{x}(t) \in \Sigma'$ for all $t \geq 0$. Then, we have using A-3 the following chain of inequalities:

$$\begin{aligned}
\|\mathbf{x}(t) - \overline{\mathbf{x}}\| &\leq \|\mathbf{x}(t) - \overline{\mathbf{x}}_{\Sigma'}\| + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma') \\
&\leq \lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))^{-1} \left(\|\mathbf{y}(t) - \mathbf{A}\overline{\mathbf{x}}_{\Sigma'}\|\right) + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma') \\
&\leq \lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))^{-1} (\|\mathbf{y}(t) - \mathbf{y}\| \\
&\quad + \|\mathbf{y} - \mathbf{A}\overline{\mathbf{x}}\| + \|\mathbf{A}(\overline{\mathbf{x}} - \overline{\mathbf{x}}_{\Sigma'})\|) + \operatorname{dist}(\overline{\mathbf{x}}, \Sigma') \\
&\leq \frac{\sqrt{2\xi\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \exp\left(-\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{4} t\right)}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} + \frac{\|\boldsymbol{\varepsilon}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} \\
&\quad + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right) \operatorname{dist}(\overline{\mathbf{x}}, \Sigma'),
\end{aligned}$$

which proves (10).

□

3.5.2 Proof of Theorem 3.6

Our proof is in the same vein as that of [10, Theorem 4.1]. However, we will improve not only the scaling but we will also accommodate better the linear operator, the new form of R' and the presence of η within it, since the latter depends on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))$.

We start by providing a bound on the Lipschitz constant of $\mathcal{J}_{\mathbf{g}}$ which is slightly tighter than the one in [10].

Lemma 3.10. *Suppose that assumptions A-2, A-4 and A-6 are satisfied. For the one-hidden layer network (12), we have for any $\boldsymbol{\theta}_0$ and $\rho > 0$:*

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

Proof. Let $\boldsymbol{\theta} \in \mathbb{R}^{k(d+n)}$ be the vectorized form of any parameters $(\mathbf{W}, \mathbf{V})$ of the network . The Jacobian $\mathcal{J}_{\mathbf{g}}$ at $\boldsymbol{\theta}$ reads

$$\frac{1}{\sqrt{k}}\left[\phi(\mathbf{W}^1\mathbf{u})\mathbf{I}_n \ \ldots \ \phi(\mathbf{W}^k\mathbf{u})\mathbf{I}_n \ \phi'(\mathbf{W}^1\mathbf{u})\mathbf{V}_1\mathbf{u}^\top \ \ldots \ \phi'(\mathbf{W}^k\mathbf{u})\mathbf{V}_k\mathbf{u}^\top\right]. \tag{23}$$

It then follows that $\forall \boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \mathbb{B}(\boldsymbol{\theta}_0, \rho)$,

$$\begin{aligned}
&\left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})\right\|^2 = \left\|\left(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})\right)\left(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}) - \mathcal{J}_{\mathbf{g}}(\widetilde{\boldsymbol{\theta}})\right)^\top\right\| \\
&= \frac{1}{k}\Big\|\sum_{i=1}^k \left(\phi(\mathbf{W}^i\mathbf{u}) - \phi(\widetilde{\mathbf{W}}^i\mathbf{u})\right)^2 \mathbf{I}_n + \\
&\qquad\qquad \left(\phi'(\mathbf{W}^i\mathbf{u})\mathbf{V}_i - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})\widetilde{\mathbf{V}}_i\right)\left(\phi'(\mathbf{W}^i\mathbf{u})\mathbf{V}_i - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})\widetilde{\mathbf{V}}_i\right)^\top\Big\| \\
&\leq \frac{1}{k}\sum_{i=1}^k \left(\left(\phi(\mathbf{W}^i\mathbf{u}) - \phi(\widetilde{\mathbf{W}}^i\mathbf{u})\right)^2 + \left\|\phi'(\mathbf{W}^i\mathbf{u})\mathbf{V}_i - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})\widetilde{\mathbf{V}}_i\right\|^2\right) \\
&\leq \frac{1}{k}\sum_{i=1}^k \left(\phi(\mathbf{W}^i\mathbf{u}) - \phi(\widetilde{\mathbf{W}}^i\mathbf{u})\right)^2 + 2\phi'(\mathbf{W}^i\mathbf{u})^2\left\|\mathbf{V}_i - \widetilde{\mathbf{V}}_i\right\|^2 \\
&\qquad\qquad + 2\left(\phi'(\mathbf{W}^i\mathbf{u}) - \phi'(\widetilde{\mathbf{W}}^i\mathbf{u})\right)^2\left\|\widetilde{\mathbf{V}}_i\right\|^2 \\
&\leq \frac{1}{k}\sum_{i=1}^k \left(B^2\left\|\mathbf{W}^i - \widetilde{\mathbf{W}}^i\right\|^2 + 2B^2\left\|\mathbf{V}_i - \widetilde{\mathbf{V}}_i\right\|^2 + 2B^2\left\|\mathbf{W}^i - \widetilde{\mathbf{W}}^i\right\|^2\left\|\widetilde{\mathbf{V}}_i\right\|^2\right) \\
&\leq \frac{2B^2}{k}\left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\right\|^2 + \frac{2B^2}{k}\left(\max_{i\in[k]}\left\|\widetilde{\mathbf{V}}_i\right\|^2\right)\left\|\mathbf{W} - \widetilde{\mathbf{W}}\right\|_F^2.
\end{aligned} \tag{24}$$

Now, for any $i \in [k]$, the following holds

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

Plugging this into (24) and taking the square-root, we conclude. □

We will also need to bound η and ξ, and control the spectrum of the Jacobian $\mathcal{J}_{\mathbf{g}}$ at the initial point $\boldsymbol{\theta}_0$.

Lemma 3.11. *Consider the one-hidden layer network (12) such that A-2 holds and the initialization $\boldsymbol{\theta}_0$ obeys A-4-A-6. We have*

$$\left\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\right\| \leq C_\phi + B + C\sqrt{\frac{n}{k}}$$

with probability at least $1 - 3e^{-n}$, and

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

with probability at least $1 - 2n^{-1}$ *provided that* $k/\log(k) \geq C' n \log(n)$. *Here,* C *and* $C' > 0$ *are large enough absolute constants that depend on* B, C_ϕ, $C_{\phi'}$ *and* D.

Proof. The second bound comes from [10, Lemma 4.9]. Let us now focus on the first one.

Arguing as in (24), we have

$$\|\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)\|^2 \leq \frac{1}{k}\|\phi(\mathbf{W}(0)\mathbf{u})\|^2 + \frac{1}{k}\left\|\sum_{i=1}^{k}\phi'(\mathbf{W}(0)^i\mathbf{u})^2\mathbf{V}(0)_i\mathbf{V}(0)_i^\top\right\|. \tag{25}$$

We first concentrate $\|\phi(\mathbf{W}(0)\mathbf{u})\|$ around its expectation. Using A-4, A-5 and orthogonal invariance of the Gaussian distribution, we have $\mathbf{W}(0)\mathbf{u}$ is $\mathcal{N}(0, \mathbf{I}_k)$. Therefore,

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

We also know from A-2 that $\|\phi(\cdot)\|$ is B-Lipschitz. Thus by the Gaussian concentration inequality,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{\sqrt{k}}\|\phi(\mathbf{W}(0)\mathbf{u})\| \geq C_\phi + \tau\right)\\ &\leq \mathbb{P}\left(\frac{1}{\sqrt{k}}\|\phi(\mathbf{W}(0)\mathbf{u})\| \geq \mathbb{E}\left[\frac{1}{\sqrt{k}}\|\phi(\mathbf{W}(0)\mathbf{u})\|\right] + \tau\right) \leq \exp\left(-\frac{\tau^2 k}{2B^2}\right).\end{aligned}$$

By choosing $\tau = B\sqrt{\frac{2n}{k}}$, we obtain that

$$\frac{1}{\sqrt{k}}\|\phi(\mathbf{W}(0)\mathbf{u})\| \leq C_\phi + B\sqrt{\frac{2n}{k}} \tag{26}$$

with probability at least $1 - e^{-n}$.

We now turn to bounding the second term of (25). We first note that by A-2, we have

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

By A-6 and [42, Example 5.8], the entries of $\mathbf{V}(0)$ are centered sub-gaussian random variables. Since they are also independent, we get from [42, Lemma 5.24] that the columns $\mathbf{V}(0)_i$ are independent centered sub-gaussian random vectors, with sub-gaussian norm $K \overset{\text{def}}{=} CD$, where C is an absolute constant. They are also isotropic thanks to A-6. We are then in position to invoke [42, Theorem 5.41] to assert that

$$\mathbb{P}\left(\frac{1}{\sqrt{k}}\left\|\mathbf{V}(0)^\top\right\| \geq 1 + (c_K^{-1/2} + C_K)\sqrt{\frac{n}{k}}\right) \leq 2e^{-n},$$

where $c_K, C_K > 0$ are absolute constants that depend only on the sub-gaussian norm K (hence on D). Plugging the last bounds into (25) and then into (18), and using a union bound, we get the claim. □

We finally need to provide a bound on the initial loss $\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))$, similarly to [10, Lemma 4.11]. Although the conclusion there was true, the proof used an independence argument to bound $\|\mathbf{x}(0)\|$ which was incorrect. Here, we will fix this using Hoeffding's inequality for sub-gaussian variables.

Lemma 3.12. *Suppose that A-2 holds and the initialization* $\boldsymbol{\theta}_0$ *obeys A-4-A-6. Then*

$$\|\mathbf{y}(0) - \mathbf{y}\| \leq C\|\mathbf{A}\|\left(C_\phi + B\sqrt{\frac{2n}{k}}\right)\sqrt{n} + \|\mathbf{A}\overline{\mathbf{x}}\|_\infty\left(1 + \mathrm{SNR}^{-1}\right)\sqrt{m},$$

with probability at least $1 - 2e^{-(n-1)}$, *where* $C > 0$ *is an absolute constant that depends on* D.

Proof. We have

$$\|\mathbf{y}(0)-\mathbf{y}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}(0)\| + \sqrt{m}\,\|\mathbf{A}\overline{\mathbf{x}}\|_\infty \left(1+\mathrm{SNR}^{-1}\right),$$

with $\mathbf{x}(0) = \mathbf{g}(\mathbf{u},\boldsymbol{\theta}(0)) = \frac{1}{\sqrt{k}}\sum_{i=1}^k \phi(\mathbf{W}^i(0)\mathbf{u})\mathbf{V}_i(0)$. Let us denote $\mathbf{a} \overset{\text{def}}{=} \frac{1}{\sqrt{k}}\phi(\mathbf{W}(0)\mathbf{u})$. We are going to use a covering argument to bound $\|\mathbf{x}(0)\|$. Let $\mathscr{N}_\epsilon$ be an ϵ-net of $\mathbb{S}^{n-1}$ for some $\epsilon \in]0,1[$. Let $\mathbf{s} \in \mathbb{S}^{n-1}$ such that $\|\mathbf{x}(0)\| = \langle \mathbf{x}(0), \mathbf{s}\rangle$. Let $\mathbf{z} \in \mathscr{N}_\epsilon$ which approximates $\mathbf{s}$ as $\|\mathbf{s}-\mathbf{z}\| \leq \epsilon$. We have

$$\left|\,|\langle \mathbf{x}(0),\mathbf{s}\rangle| - |\langle \mathbf{x}(0),\mathbf{z}\rangle|\,\right| \leq \epsilon\,\|\mathbf{x}(0)\|\,.$$

Thus

$$|\langle \mathbf{x}(0),\mathbf{z}\rangle| \geq |\langle \mathbf{x}(0),\mathbf{s}\rangle| - \epsilon\,\|\mathbf{x}(0)\| = (1-\epsilon)\,\|\mathbf{x}(0)\|\,.$$

This implies that

$$\|\mathbf{x}(0)\| \leq (1-\epsilon)^{-1}\sup_{\mathbf{z}\in\mathscr{N}_\epsilon}|\langle \mathbf{x}(0),\mathbf{z}\rangle| = (1-\epsilon)^{-1}\sup_{\mathbf{z}\in\mathscr{N}_\epsilon}\left|\sum_{i=1}^k \mathbf{a}_i\langle \mathbf{V}_i(0),\mathbf{z}\rangle\right|.$$

We then have

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{x}(0)\| \geq \delta\right) \leq & \mathbb{P}\left(\sup_{\mathbf{z}\in\mathscr{N}_\epsilon}\left|\sum_{i=1}^k \mathbf{a}_i\langle \mathbf{V}_i(0),\mathbf{z}\rangle\right| \geq (1-\epsilon)\delta \;\middle|\; \|\mathbf{a}\| < \nu\right) \\ & + \mathbb{P}\left(\|\mathbf{a}\| \geq \nu\right).\end{aligned}$$

Let us fix $\mathbf{z} \in \mathbb{S}^{n-1}$. By assumption A-6 and [42, Lemma 5.9], $\langle \mathbf{V}_i(0),\mathbf{z}\rangle$ are independent zero-mean sub-gaussian random variables with sub-gaussian norm $K = C'D$, where C' is an absolute constant. It then follows from Hoeffding's inequality ([42, Proposition 5.10]) that

$$\mathbb{P}\left(\left|\sum_{i=1}^k \mathbf{a}_i\langle \mathbf{V}_i(0),\mathbf{z}\rangle\right| \geq (1-\epsilon)\delta \;\middle|\; \|\mathbf{a}\| < \nu\right) \leq e^{-\frac{c(1-\epsilon)^2\delta^2}{K^2\nu^2}},$$

where $c > 0$ is an absolute constant. A union bound then yields

$$\mathbb{P}\left(\sup_{\mathbf{z}\in\mathscr{N}_\epsilon}\left|\sum_{i=1}^k \mathbf{a}_i\langle \mathbf{V}_i(0),\mathbf{z}\rangle\right| \geq (1-\epsilon)\delta \;\middle|\; \|\mathbf{a}\| < \nu\right) \leq |\mathscr{N}_\epsilon| e^{-\frac{c(1-\epsilon)^2\delta^2}{K^2\nu^2}}.$$

Taking $\epsilon = 1/2$, we have $|\mathscr{N}_\epsilon| \leq 5^n$; see [42, Lemma 5.2]. Moreover, we know from (26) that

$$\mathbb{P}\left(\|\mathbf{a}\| \geq C_\phi + B\sqrt{\frac{2n}{k}}\right) \leq e^{-n}.$$

Taking $\nu = C_\phi + B\sqrt{\frac{2n}{k}}$ and $\delta = 2K\nu\sqrt{\frac{3n}{c}}$, we get the claim. □

Proof of Theorem 3.6. The goal is to show that (5) holds with high probability under the given scaling. We start by upper-bounding R'. We can invoke Lemma 3.11 to infer that, whenever $k \gtrsim n\log(n)\log(k)$, with probability at least $1-3e^{-n}-2n^{-1}$,

$$\eta \lesssim \frac{\max\left(\sigma_{\mathbf{A}}, (c_1)^{\frac{1}{4}}\right)}{\min\left(\sigma_{\mathbf{A}}^2, (c_2)^{\frac{1}{4}}\right)} \quad \text{and} \quad \xi \lesssim 1+\kappa(\mathbf{A})^2.$$

Using Lemma 3.10 and Lemma 3.11, and arguing similarly to the first part of the proof of [10, Theorem 4.1], we have

$$R \gtrsim \left(\frac{k}{n}\right)^{1/4} \tag{27}$$

with the same probability as above. Now, Lemma 3.12 allows to assert that

$$\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}(0))} \lesssim \|\mathbf{A}\| \sqrt{n} + \|\mathbf{A}\bar{\mathbf{x}}\|_{\infty} \left(1 + \mathrm{SNR}^{-1}\right) \sqrt{m}$$

with probability at least $1 - 2e^{-(n-1)}$. Piecing all these bounds together, and using a union bound, one sees that

$$R' \lesssim \frac{\max\left(\sigma_{\mathbf{A}}, (c_1)^{\frac{1}{4}}\right)}{\min\left(\sigma_{\mathbf{A}}^2, (c_2)^{\frac{1}{4}}\right)} (1 + \kappa(\mathbf{A})) \left(\|\mathbf{A}\| \sqrt{n} + \|\mathbf{A}\bar{\mathbf{x}}\|_{\infty} \left(1 + \mathrm{SNR}^{-1}\right) \sqrt{m}\right) \tag{28}$$

with probability at least $1 - 5e^{-(n-1)} - 2n^{-1}$. Combining (27) and (28) and using that $(a+b)^4 \leq 8\left(a^4 + b^4\right)$ for $a, b \in \mathbb{R}$, we get the claim. □

4 Discrete Setting

Let us now turn to the discretization of (DIN) using explicit finite differences approximation. This gives a first-order (i.e., gradient-based) scheme summarized in Algorithm 1.

Algorithm 1:

Input: $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0$; $s_0 > 0$; $\delta \in]0, 2[$; $\rho \in]0, 1[$; $\alpha > 0$; $\beta > 0$.

for $\tau = 0, 1, \ldots$ **do**

Compute

$$\begin{aligned} \mathbf{q}_\tau &= \boldsymbol{\theta}_\tau + \alpha s_\tau \left(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_{\tau-1}\right) - \beta s_\tau^2 \left(\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_{\tau-1})\right), \\ \boldsymbol{\theta}_{\tau+1} &= \mathbf{q}_\tau - s_\tau \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \end{aligned} \tag{29}$$

with $s_\tau = \rho^{i_\tau} s_0$, where i_τ is the smallest nonnegative integer such that

$$\begin{aligned} &\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_{\tau+1}))) - \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\tau))) \\ &\qquad - \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\tau))), \boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau \rangle \leq \frac{\delta}{2 s_\tau} \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|^2 \end{aligned}$$

and

$$\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_{\tau+1}))) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u}, \boldsymbol{\theta}_\tau)))\| \leq \frac{\delta}{s_\tau} \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|.$$

end

As in the continuous case, α is the "momentum" parameter which controls the friction while β controls the geometric "Hessian"-driven[2] damping. The choice of the parameter sequences αs_τ and βs_τ^2 may seem cryptic at this stage, and is not stemming precisely from the time discretization of the continuous dynamic. We will however clarify later the reasons behind this choice which is flexible enough to get the desired convergence behaviour under solely local Lipschitz continuity of the objective gradient. Indeed, global Lipschitz continuity allows to take a standard upper-bound on the choice of the step-size s_τ. However, such an assumption is unrealistic when training neural networks. To cope with this, a line search procedure with backtracking is crucial which poses additional technical difficulties that we must deal with carefully.

Remark 4.1. It is worth mentioning at this stage that one can replace the backtracking update in our algorithm by $s_\tau = \rho^{i_\tau} s_{\tau-1}$. This update may have some benefits in practice. Our results and proofs extend readily to this case by a mild adaptation of Lemma 4.3. Therefore, we will not elaborate more on it.

[2]The quotation marks is because the Hessian does not appear explicitly but is rather approximated with the difference of gradients.

4.1 Convergence result

In the next theorem, we give sufficient conditions on (α, β, δ) that ensure linear convergence of the network training to a zero-loss solution. We also provide the convergence rates as well as the global convergence of the whole sequence $(\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}}$.

Theorem 4.2. *Assume that A-1 and A-2 hold. Let $(\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}}$ be the sequence generated by Algorithm 1 with the parameters (α, β, δ) satisfying $s_0 \geq 1$ and $0 < 2\delta_2 < s_0^{-1}(1-\delta/2)$, where $\delta_2 = \frac{\alpha+\beta\delta}{2}$. Moreover, let the initialization $\boldsymbol{\theta}_0$ be such that*

$$\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)) > 0 \text{ and } R' < R \tag{30}$$

with

$$R' = \frac{\sqrt{2}}{\delta_1(1-2s_0\delta_2)}\left(\frac{2}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \frac{1}{\sqrt{\delta_2}s_0}\right)\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}_0)} \text{ and} \tag{31}$$

$$R = \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))}{2\mathrm{Lip}_{\mathbb{B}(\boldsymbol{\theta}_0,R)}(\mathcal{J}_{\mathbf{g}})} \tag{32}$$

where $\delta_1 = s_0^{-1}\left(1-\frac{\delta}{2}\right) - 2\delta_2$ and $0 < \underline{s} \overset{\text{def}}{=} \inf_{\tau\in\mathbb{N}} s_\tau \leq \overline{s} \overset{\text{def}}{=} \sup_{\tau\in\mathbb{N}} s_\tau \leq s_0$. Then, the loss converges linearly to 0 with

$$\begin{aligned} &\mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \leq \frac{\delta R'^2}{2\underline{s}}\left(\frac{\rho}{1+\rho}\right)^\tau \\ &\text{where } \rho \leq 8\delta_1^{-1}(1-2s_0\delta_2)^{-1}\left(\frac{1}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \frac{1}{2\sqrt{\delta_2}s_0}\right)^2. \end{aligned} \tag{33}$$

In addition, $(\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}}$ converges linearly to a global minimizer $\boldsymbol{\theta}_\infty$ of (2) with

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\infty\| \leq R'\left(\frac{\rho}{1+\rho}\right)^{\tau/2}. \tag{34}$$

If, moreover, (A-3) holds, then

$$\begin{aligned} \|\mathbf{x}_\tau - \overline{\mathbf{x}}\| \leq \frac{\sqrt{\delta/\underline{s}}R'}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\left(\frac{\rho}{1+\rho}\right)^{\tau/2} + \frac{\|\boldsymbol{\varepsilon}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))} \\ + \left(1 + \frac{\|\mathbf{A}\|}{\lambda_{\min}(\mathbf{A}; T_{\Sigma'}(\overline{\mathbf{x}}_{\Sigma'}))}\right)\mathrm{dist}(\overline{\mathbf{x}}, \Sigma'). \end{aligned} \tag{35}$$

This theorem ensures that the neural network can be trained to zero loss using Algorithm 1 with a proper choice of α, β and δ. The condition $s_0(\alpha+\beta\delta) < 1-\frac{\delta}{2}$ balances the effect of viscous (momentum) and Hessian damping with respect to the user-chosen parameter δ of the backtracking procedure to ensure convergence of the network training.

Understanding more precisely the effect of the choice of α, β and δ, for fixed s_0, on the convergence guarantees in this discrete setting boils down to understanding their role in δ_1 and δ_2, which in turn influence both R' and our convergence rates. First, we see that the closer δ is to 2, the more limited is the choice of α and β in order to comply with the condition $s_0(\alpha+\beta\delta) < 1-\frac{\delta}{2}$. Furthermore, this means that both δ_1 and δ_2 would go towards 0, hence making R' scaling as $O(\delta^{-1}\delta_2^{-1/2}(1-\delta/2)^{-1})$ and ρ as $O(\delta^{-1}\delta_2^{-1}(1-\delta/2)^{-1})$. This regime is undesirable as it may induce a very slow training convergence rate. On the other hand, choosing δ smaller allows for larger choices of α and β to balance between δ_1 and δ_2. Indeed, for fixed s_0, one can decide to keep δ_2 larger at the expense of shrinking δ_1 and vice versa. Observe also that choosing δ small may have a cost by potentially increasing the backtracking procedure termination iteration. This would then make $\underline{s}$ smaller hence increasing ρ and R'. Thus, there is a clear tradeoff in the choice of δ, α and β.

Observe that under our initialization condition, our result states that the parameters of the network remain in a ball near that initialization on which $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}((\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}}))$ is bounded away from zero, hence verifying the Łojasiewicz inequality with exponent $1/2$, hence the linear convergence rate. For the ill-conditioned case, ρ scales as $O\left(\frac{1}{\delta_1\underline{s}^2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))^2\sigma_{\mathbf{A}}^2}\right)$, hence giving the convergence rate $\frac{1}{1+c\delta_1\underline{s}^2\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}{}^2}$ for some constant $c > 0$. Our estimate of the convergence rate seems overly pessimistic as it strictly larger than the convergence

rate $\frac{1}{1+c\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}$ known to be the optimal rate of first-order methods for strongly convex L-smooth objectives [30]. Note however that we are dealing with a nonconvex objective whose gradient is only locally Lipschitz continuous.

Whether our estimate of the rate can be improved or not is an open question. A possible way to have a tighter estimate is to to lift the problem to the product space $(\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\infty, \boldsymbol{\theta}_{\tau-1} - \boldsymbol{\theta}_\infty)$ and using a linearization of $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathbf{y}}(\mathbf{A}\mathbf{g}(\mathbf{u},\boldsymbol{\theta}))$ around $\boldsymbol{\theta}_\infty$, and then studying the spectral properties of the resulting matrix in the linearization, see [35]. We would like to explore this further in a future work.

Theorem 3.6 can be adapted to the new form of R' and R in Theorem 4.2 with minor modifications. The resulting scaling of the network architecture will be similar. We refrain from giving the details which are left to the reader.

4.2 Proofs

Lemma 4.3 (Finite termination and well-definedness)**.** *The backtracking procedure in Algorithm 1 terminates in a finite number of iterations and* $\overline{s} \stackrel{\text{def}}{=} \sup_{\tau\in\mathbb{N}} s_\tau \leq s_0$. *If the sequence* $(\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}}$ *is bounded, then* $\underline{s} \stackrel{\text{def}}{=} \inf_{\tau\in\mathbb{N}} s_\tau > 0$.

Proof. To lighten notation, let $f \stackrel{\text{def}}{=} \mathcal{L}_{\mathbf{y}} \circ \mathbf{A} \circ \mathbf{g}(\mathbf{u},\cdot)$, and denote the Bregman divergence of f as

$$D_f(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta}) \stackrel{\text{def}}{=} f(\widetilde{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}) - \langle \nabla f(\boldsymbol{\theta}), \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle.$$

We write generically each iteration of Algorithm 1 as

$$\boldsymbol{\theta}^+(\mu_i) \stackrel{\text{def}}{=} \boldsymbol{\theta} + \alpha_i(\boldsymbol{\theta} - \boldsymbol{\theta}_-) - \beta_i(\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}_-)) - \mu_i \nabla f(\boldsymbol{\theta}), \quad \text{where}$$
$$\mu_i = \rho^i s_0, \alpha_i = \alpha\mu_i, \beta_i = \beta\mu_i^2, \forall i \in \mathbb{N}.$$

Clearly, $\boldsymbol{\theta}^+(\mu_i) \to \boldsymbol{\theta}$ as $i \to \infty$. Thus $\forall \epsilon > 0$, $\exists l_\epsilon > 0$ such that $\boldsymbol{\theta}^+(\mu_i) \subset \mathbb{B}(\boldsymbol{\theta},\epsilon)$, $\forall i \geq l_\epsilon$. It then follows from the local Lipschitz continuity of ∇f (thanks to A-1 and A-2) and the descent lemma [9, Lemma 2.64(i)] that $\exists L_\epsilon > 0$ such that $\forall i \geq l_\epsilon$,

$$\begin{aligned} &\left\|\nabla f(\boldsymbol{\theta}^+(\mu_i)) - \nabla f(\boldsymbol{\theta})\right\| \leq L_\epsilon \left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\| \\ \text{and} \qquad &D_f(\boldsymbol{\theta}^+(\mu_i),\boldsymbol{\theta}) \leq \frac{L_\epsilon}{2}\left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\|^2. \end{aligned} \tag{36}$$

Assume by contradiction that the backtracking procedure does not terminate. That is, for all $i \geq 0$,

$$\begin{aligned} &\mu_i \left\|\nabla f(\boldsymbol{\theta}^+(\mu_i)) - \nabla f(\boldsymbol{\theta})\right\| > \delta \left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\| \\ \text{or} \qquad &\mu_i D_f(\boldsymbol{\theta}^+(\mu_i),\boldsymbol{\theta}) > \frac{\delta}{2}\left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\|^2. \end{aligned} \tag{37}$$

This together with (36) entails that for all $i \geq l_\epsilon$,

$$\delta \left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\| < \mu_i \left\|\nabla f(\boldsymbol{\theta}^+(\mu_i)) - \nabla f(\boldsymbol{\theta})\right\| \leq \mu_i L_\epsilon \left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\|$$
$$\text{or}$$
$$\frac{\delta}{2}\left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\|^2 < \mu_i D_f(\boldsymbol{\theta}^+(\mu_i),\boldsymbol{\theta}) \leq \frac{\mu_i L_\epsilon}{2}\left\|\boldsymbol{\theta}^+(\mu_i) - \boldsymbol{\theta}\right\|^2.$$

Simplifying gives in both cases that $\delta < \mu_i L_\epsilon$. Passing to the limit as $i \to \infty$ yields $\delta = 0$, a contradiction.

The fact that $\overline{s} \leq s_0$ is immediate. We will now show that $\underline{s} > 0$. We have by assumption that $(\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}} \subset \Omega$, for some convex bounded set Ω. The descent lemma used above implies that there exists $L_\Omega \geq 0$ such that for all $\tau \geq 0$,

$$D_f(\boldsymbol{\theta}_{\tau+1},\boldsymbol{\theta}_\tau) \leq \frac{L_\Omega}{2}\left\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\right\|^2.$$

We now show by induction that for all $\tau \geq 0$,

$$s_\tau \geq \min(s_0, \rho\delta L_\Omega^{-1}) > 0. \tag{38}$$

This is obviously true for $\tau = 0$. Assume that (38) holds at some $\tau \geq 1$. Recall that $s_{\tau+1} = \rho^{i_{\tau+1}} s_0$. If $i_{\tau+1} \leq i_\tau$ then $s_{\tau+1} \geq s_\tau$ and we are done. If $i_{\tau+1} \geq i_\tau + 1$, we suppose for contradiction that $s_{\tau+1} < \min(s_0, \rho\delta L_\Omega^{-1})$. Thus, the descent property above entails that

$$D_f(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau) < \frac{\delta}{2\rho^{i_{\tau+1}-1} s_0} \left\| \boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau \right\|^2,$$

meaning that the backtracking terminates at $i_{\tau+1} - 1$, leading to a contradiction as it was supposed to terminate at $i_{\tau+1}$. This concludes the proof. □

Proof of Theorem 4.2. We will first derive a Lyapunov function, then show how the parameters of the network remain bounded under (30) and the devised choice of (α, β, s_τ)) which gives a lower bound on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_\tau))$ for all $\tau \in \mathbb{N}^*$, which finally allow us to derive convergence rates.

Step 1: Lyapunov analysis. We first perform a Lyapunov analysis by designing an appropriate energy function. Let us now observe that the update (29) can be equivalently written

$$\boldsymbol{\theta}_{\tau+1} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2} \left\| \boldsymbol{\theta} - \mathbf{q}_\tau + s \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \right\|^2.$$

Using the 1-strong convexity of $\boldsymbol{\theta} \mapsto \frac{1}{2} \left\| \boldsymbol{\theta} - \mathbf{q}_\tau + s \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \right\|^2$, we get

$$\begin{aligned} \frac{1}{2} \left\| \boldsymbol{\theta}_{\tau+1} - \mathbf{q}_\tau + s \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \right\|^2 \leq \frac{1}{2} \left\| \boldsymbol{\theta}_\tau - \mathbf{q}_\tau + s \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \right\|^2 \\ - \frac{1}{2} \left\| \boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau \right\|^2. \end{aligned} \tag{39}$$

Let us denote for short $\alpha_\tau = \alpha s_\tau$, $\beta_\tau = \beta s_\tau^2$, $\mathbf{v}_\tau = \boldsymbol{\theta}_\tau - \boldsymbol{\theta}_{\tau-1}$ with $\mathbf{v}_0 = \mathbf{0}$ and $\mathbf{z}_\tau = \alpha_\tau \mathbf{v}_\tau - \beta_\tau \left(\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_{\tau-1}) \right)$. We have $\mathbf{q}_\tau = \boldsymbol{\theta}_\tau + \mathbf{z}_\tau$. Expanding the terms on both sides of (39), we obtain that

$$\langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau), \mathbf{v}_{\tau+1} \rangle \leq - \frac{\left\| \mathbf{v}_{\tau+1} \right\|^2}{s_\tau} + \frac{\langle \mathbf{z}_\tau, \mathbf{v}_{\tau+1} \rangle}{s_\tau}. \tag{40}$$

Combining (40) with the backtracking termination condition of Algorithm 1, which is well-defined thanks to Lemma 4.3, we have

$$\begin{aligned} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_{\tau+1}) &\leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) + \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau), \mathbf{v}_{\tau+1} \rangle + \frac{\delta}{2 s_\tau} \left\| \mathbf{v}_{\tau+1} \right\|^2 \\ &\leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) + \frac{\langle \mathbf{z}_\tau, \mathbf{v}_{\tau+1} \rangle}{s_\tau} - \frac{1}{s_\tau} \left(1 - \frac{\delta}{2} \right) \left\| \mathbf{v}_{\tau+1} \right\|^2 \\ &\leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) + \frac{\alpha_\tau}{s_\tau} \langle \mathbf{v}_{\tau+1}, \mathbf{v}_\tau \rangle - \frac{\beta_\tau}{s_\tau} \langle \mathbf{v}_{\tau+1}, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_{\tau-1}) \rangle \\ &\qquad - \frac{1}{s_\tau} \left(1 - \frac{\delta}{2} \right) \left\| \mathbf{v}_{\tau+1} \right\|^2 \\ &\leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) + \frac{\alpha_\tau}{s_\tau} \left\| \mathbf{v}_{\tau+1} \right\| \left\| \mathbf{v}_\tau \right\| + \frac{\delta \beta_\tau}{s_\tau^2} \left\| \mathbf{v}_{\tau+1} \right\| \left\| \mathbf{v}_\tau \right\| - \frac{1}{s_\tau} \left(1 - \frac{\delta}{2} \right) \left\| \mathbf{v}_{\tau+1} \right\|^2. \end{aligned}$$

Applying Young's inequality twice with $\epsilon, \epsilon' > 0$, and using that $s_\tau \leq s_0$, we get

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}_{\tau+1}) \leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) + \frac{\epsilon + \epsilon'}{2} \left\| \mathbf{v}_\tau \right\|^2 - \left(s_0^{-1} \left(1 - \frac{\delta}{2} \right) - \frac{\alpha^2}{2\epsilon} - \frac{\beta^2 \delta^2}{2\epsilon'} \right) \left\| \mathbf{v}_{\tau+1} \right\|^2.$$

Adding $\frac{\epsilon + \epsilon'}{2} \left\| \mathbf{v}_{\tau+1} \right\|^2$ on both sides gives

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}_{\tau+1}) + \frac{\epsilon + \epsilon'}{2} \left\| \mathbf{v}_{\tau+1} \right\|^2 \leq \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) + \frac{\epsilon + \epsilon'}{2} \left\| \mathbf{v}_\tau \right\|^2$$

$$- \left(s_0^{-1} \left(1 - \frac{\delta}{2} \right) - \frac{\alpha^2}{2\epsilon} - \frac{\beta^2 \delta^2}{2\epsilon'} - \frac{\epsilon + \epsilon'}{2} \right) \|\mathbf{v}_{\tau+1}\|^2 . \quad (41)$$

To ensure that the last term is nonpositive, we need that

$$s_0^{-1} \left(1 - \frac{\delta}{2} \right) - \frac{\alpha^2}{2\epsilon} - \frac{\beta^2 \delta^2}{2\epsilon'} - \frac{\epsilon + \epsilon'}{2} > 0.$$

Optimizing over ϵ and ϵ', we obtain $\epsilon = \alpha$ and $\epsilon' = \beta\delta$. Thus, the last condition is equivalent to $s_0(\alpha + \beta\delta) < 1 - \delta/2$, hence our condition imposed on the parameters. We are now in position to define our Lyapunov sequence as $V_\tau = \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) + \delta_2 \|\mathbf{v}_\tau\|^2$, where $\delta_2 = \frac{\alpha+\beta\delta}{2}$. (41) then yields

$$V_{\tau+1} \leq V_\tau - \delta_1 \|\mathbf{v}_{\tau+1}\|^2 \quad (42)$$

with $\delta_1 \stackrel{\text{def}}{=} s_0^{-1}(1 - \delta/2) - 2\delta_2 > 0$. Clearly V_τ is nonnegative decreasing sequence, and thus it converges. Moreover, as $\delta_1 > 0$, we get that $\sum_{\tau \in \mathbb{N}} \|\mathbf{v}_\tau\|^2 < +\infty$, entailing that $\lim_{\tau \to \infty} \|\mathbf{v}_\tau\| = 0$. Thus the loss $\mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau)$ converges to the same limit as V_τ.

Step 2: Network weights are bounded under initialization condition. Now that we have a Lyapunov function, we would like to have a similar result as in Lemma 3.9 for the continuous case, that is, where we show that $\boldsymbol{\theta}$ will be bounded in some ball of radius R given some initialization condition. These results adapted to the discrete setting are presented in the following lemma.

Lemma 4.4. *Assume A-1 and A-2 hold. Recall R and R' from (30) with $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0)) > 0$. Let $(\boldsymbol{\theta}_\tau)_{\tau \in \mathbb{N}}$ be the sequence given by Algorithm 1 and assume that $s_0(\alpha + \beta\delta) < 1 - \delta/2$.*

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

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

(ii) If $\forall\, l \in \{0, \ldots, \tau\}$, $(\boldsymbol{\theta}_l)_{l \leq \tau} \subset \mathbb{B}(\boldsymbol{\theta}_0, R)$ and $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_l)) \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))}{2}$, then

$$\boldsymbol{\theta}_{\tau+1} \in \mathbb{B}(\boldsymbol{\theta}_0, R').$$

(iii) If $R' < R$, then for all $\tau \in \mathbb{N}$, $(\boldsymbol{\theta}_\tau)_{\tau \in \mathbb{N}} \subset \mathbb{B}(\boldsymbol{\theta}_0, R)$ and $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}(\tau))) \geq \sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))/2$.

Proof. (i) The proof of this claim is the same as that of [10, Lemma 3.10(i)].

(ii) We know that $s_l \leq s_0$ for all $l \in \mathbb{N}$. Moreover, since $(\boldsymbol{\theta}_l)_{l \leq \tau}$, we can invoke Lemma 4.3 to deduce that there exists $\underline{s} > 0$ such that $s_l \geq \underline{s} > 0$ for all $l \leq \tau$. The update equation (29) then gives

$$\begin{aligned} \underline{s} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau)\| &\leq \|s_\tau \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau)\| \\ &= \|\boldsymbol{\theta}_{\tau+1} - \mathbf{q}_\tau\| \\ &\leq \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\| + \|\boldsymbol{\theta}_\tau - \mathbf{q}_\tau\| \\ &\leq \|\mathbf{v}_{\tau+1}\| + (\alpha + \beta\delta) s_0 \|\mathbf{v}_\tau\| \\ &= \|\mathbf{v}_{\tau+1}\| + 2 s_0 \delta_2 \|\mathbf{v}_\tau\| . \end{aligned}$$

Thus, we get from Step 1 that $\lim_{\tau \to \infty} \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau)\| = 0$. Let us observe that by the condition of the lemma on $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_l))$ for any $l \in \{0, \ldots, \tau\}$, we have that

$$\frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{\sqrt{2}} \sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}_l)} \leq \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_l)\| \leq \frac{1}{\underline{s}} \left(\|\mathbf{v}_{l+1}\| + 2 s_0 \delta_2 \|\mathbf{v}_l\| \right), \quad (43)$$

Without loss of generality, we assume that $V_l \neq 0$, as otherwise, the algorithm has converged and there is nothing to prove. By concavity of $\sqrt{\cdot}$, we have

$$\sqrt{V_{l+1}} - \sqrt{V_l} \leq \frac{1}{2\sqrt{V_l}} (V_{l+1} - V_l)$$

$$
\begin{aligned}
(42) &\leq \frac{-\delta_1 \left\|\mathbf{v}_{l+1}\right\|^2}{2\sqrt{V_l}} \\
(43) &\leq \frac{-\delta_1 \left\|\mathbf{v}_{l+1}\right\|^2}{2\left(\frac{\left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_l)\right\|}{2^{-1/2}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\left\|\mathbf{v}_l\right\|\right)}.
\end{aligned}
$$

Let us define $\left(\Delta\sqrt{V}\right)_l = \sqrt{V_l} - \sqrt{V_{l+1}}$. Then

$$
\begin{aligned}
\left\|\mathbf{v}_{l+1}\right\|^2 &\leq \frac{2}{\delta_1}\left(\frac{\left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbf{y}}(\mathbf{y}_l)\right\|}{2^{-1/2}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\left\|\mathbf{v}_l\right\|\right)\left(\Delta\sqrt{V}\right)_l \\
(43) &\leq \frac{2}{\delta_1}\left(\frac{\left\|\mathbf{v}_{l+1}\right\| + 2s_0\delta_2\left\|\mathbf{v}_l\right\|}{2^{-1/2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\left\|\mathbf{v}_l\right\|\right)\left(\Delta\sqrt{V}\right)_l \\
&\leq \frac{2\sqrt{2}}{\delta_1}\left(\frac{\left\|\mathbf{v}_{l+1}\right\|}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \left(\frac{2s_0\delta_2}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right)\left\|\mathbf{v}_l\right\|\right) \\
&\qquad \times \left(\Delta\sqrt{V}\right)_l.
\end{aligned}
$$

By Young's inequality, we have that for any $\epsilon > 0$

$$
\begin{aligned}
\left\|\mathbf{v}_{l+1}\right\| \leq \frac{\epsilon}{\sqrt{2}}\left(\frac{\left\|\mathbf{v}_{l+1}\right\|}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \left(\frac{2s_0\delta_2}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right)\left\|\mathbf{v}_l\right\|\right) \\
+ \frac{\left(\Delta\sqrt{V}\right)_l}{\delta_1\epsilon}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\left(1 - \frac{\epsilon}{\sqrt{2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}\right)\left\|\mathbf{v}_{l+1}\right\| \\
&\leq \frac{\epsilon}{\sqrt{2}}\left(\frac{2s_0\delta_2}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right)\left\|\mathbf{v}_l\right\| + \frac{\left(\Delta\sqrt{V}\right)_l}{\delta_1\epsilon}.
\end{aligned}
$$

For any $\epsilon \in]0, \sqrt{2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}[$, we divide by the factor on the left-hand $\left(1 - \frac{\epsilon}{\sqrt{2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}\right)$ on both sides. The goal is now to choose ϵ such that

$$
\frac{\epsilon\left(2s_0\delta_2 + \sqrt{\delta_2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}\right)}{\sqrt{2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}} - \epsilon} < 1,
$$

or equivalently

$$
\epsilon < \frac{\sqrt{2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{1 + 2s_0\delta_2 + \sqrt{\delta_2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}. \tag{44}
$$

It follows that any ϵ verifying this bound entails that there is $0 < \nu < 1$ and $C_1 > 0$ such that

$$
\left\|\mathbf{v}_{l+1}\right\| \leq (1-\nu)\left\|\mathbf{v}_l\right\| + C_1\left(\Delta\sqrt{V}\right)_l. \tag{45}
$$

We then choose

$$
\epsilon = \frac{\sqrt{2\delta_2}s_0\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}}{2\sqrt{\delta_2}s_0 + \underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}/2}.
$$

Using our condition that $2s_0\delta_2 \in]0, 1-\delta/2[$, one can easily check that such a choice obeys (44). It also gives us $\nu = 1 - 2s_0\delta_2 \in]\delta/2, 1[$ and

$$C_1 = \frac{\left(2\sqrt{\delta_2}s_0 + \underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}/2\right)^2}{\sqrt{2\delta_2}\delta_1 s_0 \underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}\left(\sqrt{\delta_2}s_0 + \underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}/2\right)}. \tag{46}$$

We then obtain

$$\begin{aligned} C_1 &\leq \frac{4\left(\sqrt{\delta_2}s_0 + \underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}/2\right)}{\sqrt{2\delta_2}\delta_1 s_0 \underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} \\ &\leq \frac{\sqrt{2}}{\delta_1}\left(\frac{2}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \frac{1}{\sqrt{\delta_2}s_0}\right). \end{aligned} \tag{47}$$

Since (45) holds for any $l \in \{0, \ldots, \tau\}$ and $\boldsymbol{\theta}_{-1} = \boldsymbol{\theta}_0$, we get that

$$\begin{aligned} \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\| \leq \sum_{l=0}^{\tau+1}\|\mathbf{v}_l\| &\leq \nu^{-1}\sum_{l=0}^{\tau+1}\left(\|\mathbf{v}_l\| - \|\mathbf{v}_{l+1}\| + C_1\left(\Delta\sqrt{V}\right)_l\right) \\ &\leq \nu^{-1}C_1\left(\sqrt{V_0}\right) + \nu^{-1}\|\mathbf{v}_0\| \\ &= \nu^{-1}C_1\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}_0)}. \end{aligned} \tag{48}$$

(iii) We prove this by induction. For $\tau = 0$, the claim trivially holds. Suppose now that $R' < R$ and that $(\boldsymbol{\theta}_l)_{l\leq\tau} \subset \mathbb{B}(\boldsymbol{\theta}_0, R)$ and $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_l)) \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))}{2}$ for all $l \in \{0, \ldots, \tau\}$. Let us now show that this also holds at $\boldsymbol{\theta}_{\tau+1}$. By the induction assumption and (ii), we have $\boldsymbol{\theta}_{\tau+1} \in \mathbb{B}(\boldsymbol{\theta}_0, R') \subset \mathbb{B}(\boldsymbol{\theta}_0, R)$. In view of (i), we get the claim.

□

In view of the condition of the theorem on (α, β, δ) and the assumption that $R' < R$, Lemma 4.4(iii) applies to ensure that $(\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}} \subset \mathbb{B}(\boldsymbol{\theta}_0, R)$ and $\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}(\tau))) \geq \frac{\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))}{2}$ for all $\tau \in \mathbb{N}$. Equipped with this result, we can embark from (45) and pass to the limit as $\tau \to +\infty$ to get that $\sum_{\tau\in\mathbb{N}}\|\mathbf{v}_\tau\| < +\infty$, i.e., the sequence $(\boldsymbol{\theta}_\tau)_{\tau\in\mathbb{N}}$ has finite length, and thus it converges to some point $\boldsymbol{\theta}_\infty$.

Step 3: Linear convergence rate. Let us define $\Delta_\tau = \sum_{l=\tau}^{+\infty}\|\mathbf{v}_{l+1}\|$. The triangle inequality yields $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\infty\| \leq \Delta_\tau$. Therefore it is sufficient to analyze the rate of Δ_τ to get that of the iterates. Summing (45) from $l = \tau$ to ∞, we have

$$\Delta_{\tau-1} = \sum_{l=\tau}^{+\infty}\|\mathbf{v}_l\| \leq \nu^{-1}\|\mathbf{v}_\tau\| + \nu^{-1}C_1\sqrt{V_\tau}.$$

Thus, we have the recursion

$$\begin{aligned} \Delta_\tau = \Delta_{\tau-1} - \|\mathbf{v}_\tau\| &\leq \frac{1-\nu}{\nu}\|\mathbf{v}_\tau\| + \nu^{-1}C_1\sqrt{V_\tau} \\ &\leq \frac{1-\nu}{\nu}(\Delta_{\tau-1} - \Delta_\tau) + \frac{C_1}{\nu}\sqrt{V_\tau}. \end{aligned} \tag{49}$$

From the definition of our Lyapunov function and (43) we get

$$\begin{aligned} \sqrt{V_\tau} &\leq \frac{\|\mathbf{v}_{\tau+1}\| + 2\delta_2 s_0\|\mathbf{v}_\tau\|}{2^{-1/2}\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\|\mathbf{v}_\tau\| \\ &\leq \left(\frac{\sqrt{2}}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right)(\|\mathbf{v}_{\tau+1}\| + \|\mathbf{v}_\tau\|) \\ &= \left(\frac{\sqrt{2}}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right)(\Delta_{\tau-1} - \Delta_{\tau+1}), \end{aligned}$$

where we used that $2\delta_2 s_0 < 1 - \delta/2 < 1$ by assumption. Plugging this into (49) we get

$$\Delta_\tau \leq \frac{1-\nu}{\nu}\left(\Delta_{\tau-1} - \Delta_\tau\right) + \frac{\left(\frac{\sqrt{2}}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right) C_1}{\nu}\left(\Delta_{\tau-1} - \Delta_{\tau+1}\right).$$

Let us denote $C_2 = \max\left(\left(\frac{\sqrt{2}}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right) C_1, 1-\nu\right)$. One can see, using (47), that $2s_0\delta_2 < 1$ and $1 - \nu = 2s_0\delta_2$, that

$$\begin{aligned} C_2 &\leq \max\left(4\delta_1^{-1}\left(\frac{1}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \sqrt{\delta_2}\right)\left(\frac{1}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \frac{1}{2\sqrt{\delta_2}s_0}\right), 2s_0\delta_2\right) \\ &\leq \max\left(4\delta_1^{-1}\left(\frac{1}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \frac{1}{2\sqrt{\delta_2}s_0}\right)^2, 2s_0\delta_2\right). \end{aligned}$$

We claim that the maximum in the rhs is given by the first term. Indeed,

$$4\delta_1^{-1}\left(\frac{1}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \frac{1}{2\sqrt{\delta_2}s_0}\right)^2 = \frac{1}{\delta_1\delta_2 s_0^2}\left(\frac{2\sqrt{\delta_2}s_0}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + 1\right)^2 \geq 2,$$

since by assumption $2\delta_2 s_0 < 1$ which also entails $s_0\delta_1 < 1$. Therefore

$$C_2 \leq 4\delta_1^{-1}\left(\frac{1}{\underline{s}\sigma_{\min}(\mathcal{J}_{\mathbf{g}}(\boldsymbol{\theta}_0))\sigma_{\mathbf{A}}} + \frac{1}{2\sqrt{\delta_2}s_0}\right)^2.$$

Using now that $\Delta_{\tau+1} \leq \Delta_\tau$, we obtain

$$\Delta_{\tau+1} \leq \Delta_\tau \leq \frac{C_2}{\nu}\left(\Delta_{\tau-1} - \Delta_{\tau+1} + \Delta_{\tau-1} - \Delta_\tau\right) \leq \frac{2C_2}{\nu}\left(\Delta_{\tau-1} - \Delta_{\tau+1}\right).$$

Equivalently,

$$\Delta_{\tau+1} \leq \frac{\rho}{1+\rho}\Delta_{\tau-1}$$

where $\rho = \frac{2C_2}{\nu}$. This implies

$$\Delta_\tau \leq \left(\frac{\rho}{1+\rho}\right)^{\tau/2}\Delta_0.$$

Observing that $\Delta_0 \leq \nu^{-1}C_1\sqrt{\mathcal{L}_{\mathbf{y}}(\mathbf{y}_0)} = R'$ (see (48)), it follows that

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\infty\| \leq \Delta_\tau \leq R'\left(\frac{\rho}{1+\rho}\right)^{\tau/2}.$$

By the well-posedness of the backtracking procedure, we obtain

$$\mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \leq \frac{\delta}{2\underline{s}}\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\infty\|^2 \leq \frac{\delta R'^2}{2\underline{s}}\left(\frac{\rho}{1+\rho}\right)^\tau.$$

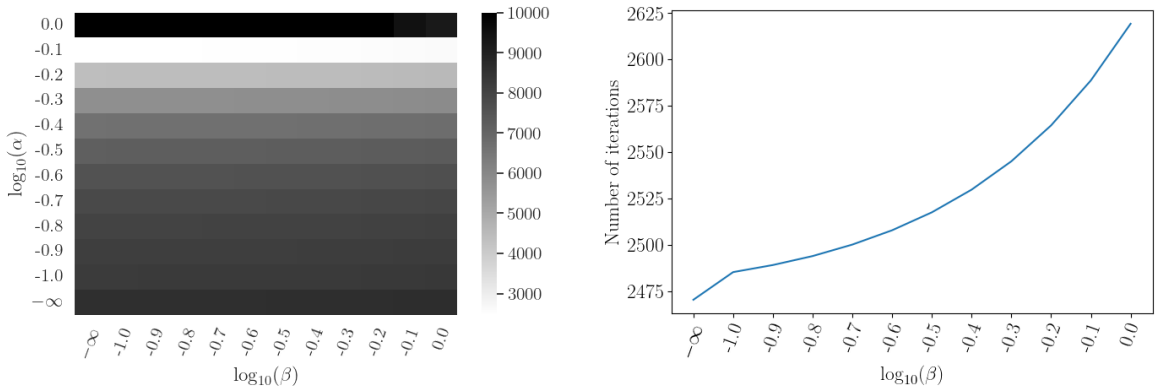
which concludes the proof.

□

5 Numerical Experiments

In order to validate our results, we performed different numerical experiments. Throughout these experiments, we used a two-layer neural network equipped with the sigmoid activation function and where the entries of $\mathbf{W}(0)$ are sampled from a standard normal distribution and the entries of $\mathbf{V}(0)$ from a uniform distribution between $-\sqrt{3}$ and $\sqrt{3}$. The networks then obviously obey our assumptions. We train both layers of the networks.

In our first experiment shown in Figure 1, our goal is to see the impact of the parameters α and β on the convergence speed of the network applied to a simple inverse problem of relatively low dimension with $n = 10$ and $m = 5$ and without noise. The entries of $\overline{\mathbf{x}}$ are iid samples from $\mathcal{N}(0, 1)$ and those of $\mathbf{A}$ from $\mathcal{N}(0, 1/\sqrt{n})$. Standard random matrix theory results ensure that the non-zero singular values of $\mathbf{A}$ are concentrated around 1. To solve this problem, we use networks where $k = 10^4$ and $d = 1$. We trained the network for each instance using our inertial algorithm with $s_0 = 0.1$ fixed, and varied α and β to assess their influence. We generate 50 different problems instances characterized by an operator $\mathbf{A}$, a signal $\overline{\mathbf{x}}$ and a network initialization. For each pair of parameters (α, β), we computed the average number of iterations over the 50 problem instances that were necessary to achieve machine precision accuracy (i.e., $\mathcal{L}_{\mathbf{y}}(\mathbf{y}_\tau) \leq 10^{-14}$).



(a) Number of iterations necessary for a network to converge on average, for different pairs (α, β)

(b) Effect of β on the number of iterations necessary to converge for $\alpha = 10^{-0.2}$

Fig. 1: Convergence rates of an inertial system for different α and β. Better α allows for much faster convergence while for this problem, β does not appear to be necessary.

For this problem, it is obvious that α is the driving factor of convergence speed. We see a clear acceleration of the training of the network as α progresses until the network start diverging when $\alpha = 1$. The acceleration we observe is very important as we go on average from 9000 iterations to converge when $\alpha = 0$ (the gradient descent case typically) to only 3000 when we chose $\alpha = 10^{-0.1}$. We see in Figure 1b the effect of β on the convergence when $\alpha = 10^{-2}$. We observe a slight degradation of the convergence rate when β progresses, revealing that for this problem, Hessian damping is not necessary and only the effect of viscous damping drives the acceleration. This is due to the fact that for this problem we do not observe oscillations around the minima, which is what Hessian damping helps to prevent.

In the next experiment, we kept the same setting but we fixed $\beta = 0.05$ and varied both k and α. We train once again 50 networks for each pair (k, α) and we plot in Figure 2 the probability of each network to achieve machine precision loss in less than 15000 iterations. From these results we see different regimes. When the network is too small, whatever α is chosen, the network will not converge to a zero-loss solution, which is in agreement with our theoretical predictions. On the other side, when α is too large, the algorithm will not converge either whatever the size of the network. However, when the network is big enough, the choice of α has a clear impact on the convergence speed. It is to be noted that in some cases, much more iterations would lead to convergence, but it seems that even by taking this into account, using the right α does help a network to find a zero-loss solution even when its size is relatively small. This might be related to a better trap-avoidance property of inertial dynamics that would be worth investigating precisely in the future.

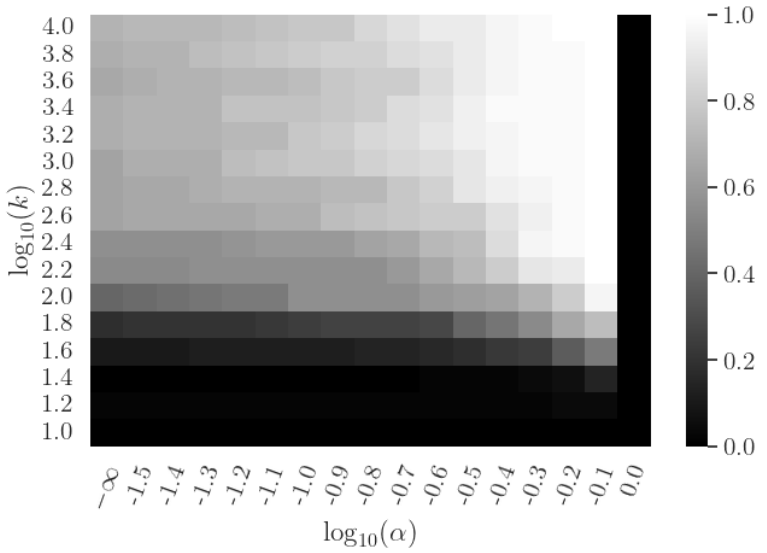


Fig. 2: Empirical probability of a network to be trained in less than 15000 iterations for different k and α. Choosing α correctly helps when k is above a certain threshold.

For our next experiment, we explore the effects of α and β on an imaging inverse problem. We consider the image as a vector in $[0, 255]^{4096}$ and use a network with $k = 7000$ hidden neurons, which is enough to achieve convergence empirically. We study a deconvolution problem where $\mathbf{A}$ is a Gaussian kernel of standard deviation 1, and added an $\mathcal{N}(0, 2.5^2)$ noise. We trained different networks using various α and β and show in Figure 3 the evolution of the loss and of the distance to the true solution for each pair of parameters. We also show in Figure 4 the final image obtained after a selected number of iterations for both gradient descent ($\alpha = 0$ and $\beta = 0$) and when $\alpha = 1$ and $\beta = 0.1$ which is one of the fastest converging combination.

Let us first focus on the evolution of the loss given in Figure 4. The first thing to note is that for the right combination of α and β ($\alpha = 1$ and $\beta \in \{0.1, 1\}$), there is considerable acceleration phenomenon compared to gradient descent. More precisely, the acceleration happens at some point, either in the beginning for $(\alpha = 1, \beta = 0.1)$ or later on for $(\alpha = 1, \beta = 1)$, and then the optimization seems to continue at a similar rate as the gradient descent. Contrary to the previous toy example, this time the acceleration can only be achieved by a good combination of α and β showing the interplay between viscous and Hessian dampings for more complex inverse problems. Indeed, for $(\alpha = 1, \beta = 0)$, we see that the loss oscillates without converging and on the other side $(\alpha = 0, \beta = 1)$ does converge but at a slower rate than gradient descent – note that such phenomenon also appears for $(\alpha = 0.1, \beta = 1)$. Finally, choosing small α and β will not provide the desired acceleration or can even hinder the convergence rate compared to gradient descent.

If we now observe the evolution of the error between the reconstructed signal and the true solution $\overline{\mathbf{x}}$, we see this error decreases and then starts increasing before reaching a plateau that fits the noise level. This effect is amplified here, as the convolution operator, while being injective, is very badly conditioned ($\sigma_{\min}(\mathbf{A}) \sim 10^{-5}$). This validates the need for an early stopping strategy. However in our case, we have that $\|\boldsymbol{\varepsilon}\| \sim 150$, which combined with the conditioning of the operator means that our early stopping bound is far from being reached in our experiments but we observe in Figure 4 that the reconstructed image has already severely overfitted the noise after 50000 iterations. Similarly, our bound on $\|\mathbf{x}_\tau - \overline{\mathbf{x}}\|$ is far from being reached because the noise term will be very large, showing the difficulties faced when using very badly conditioned operators. Finally, let us observe that in Figure 4, both gradient descent and the inertial system will overfit the noise, albeit at different times due to slower convergence of gradient descent.

We did a second imaging experiment where we changed the convolution operator to a better conditioned one and with higher level of noise to see how the optimization trajectories behave under these conditions. The operator $\mathbf{A}$ that we are using is built from the combination of two orthonormal matrices and a diagonal matrix with entries in the range $[1, 2]$ (the goal is to have the same matrices produced by an SVD). We used a Gaussian noise vector with entries iid sampled from $\mathcal{N}(0, 25)$. We plot the evolution of the loss for different parameters in Figure 5 and we see different behaviors than for the convolution operator. By choosing β too large (1 here), the algorithm diverges

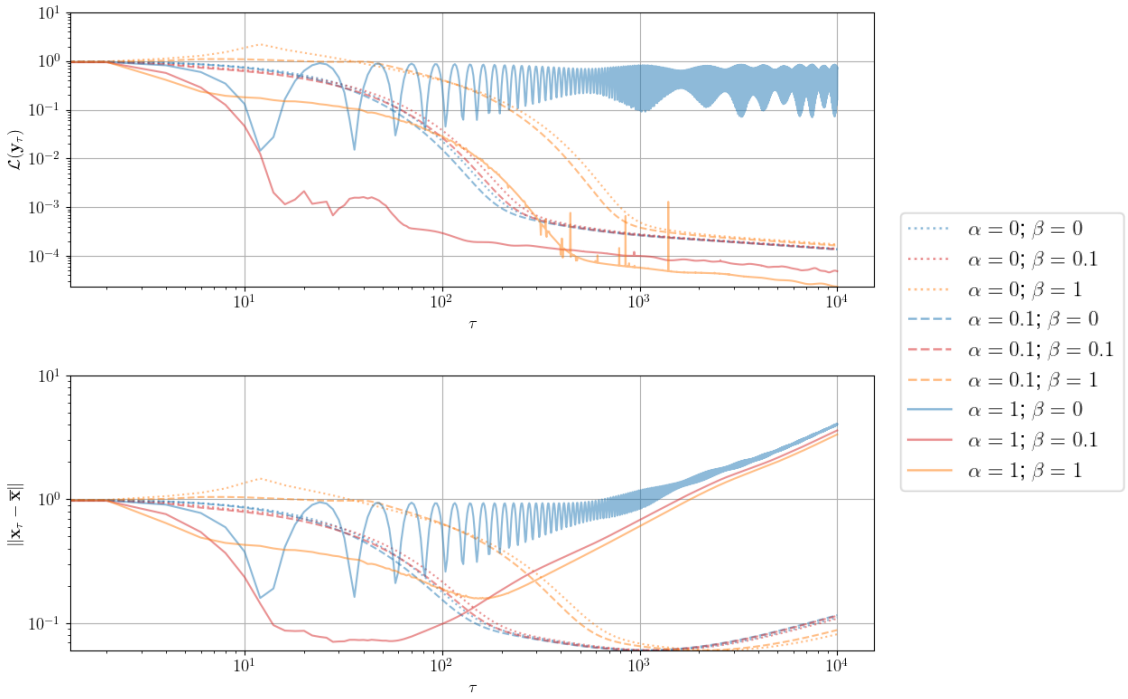


Fig. 3: Effects on the loss and the signal convergence of different combination of α and β for a deconvolution problem. Inertial systems can provide faster convergence but they are sensible to the parameters α and β and a wrong choice can prevent convergence.

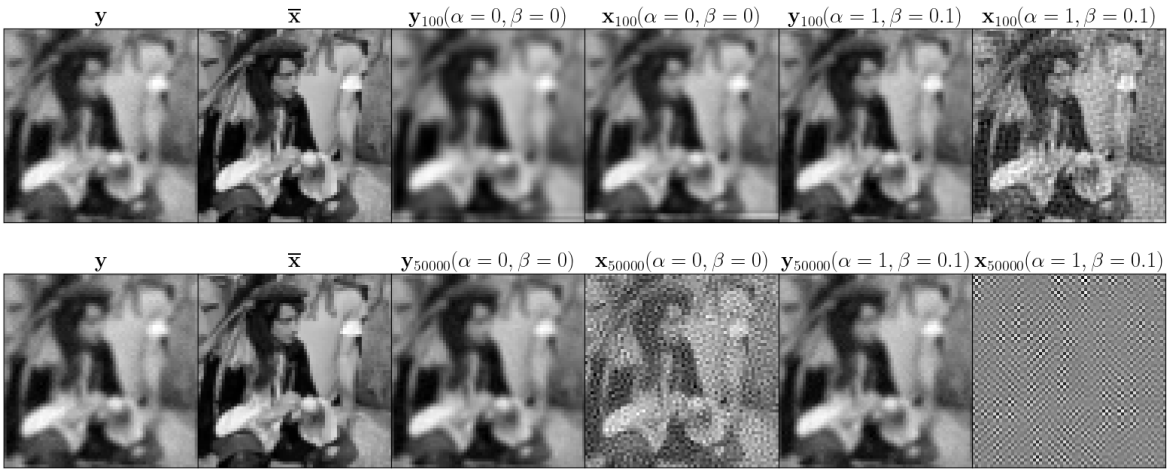


Fig. 4: Qualitative results of a deconvolution problem with low noise for different α and β. Both gradient descent and inertial system converge in observation space but they overfit the noise in signal space even for very reduced level of noise.

which did not happen in the previous experiment meaning that the conditioning of the operator plays an important role in the right choice of α and β to ensure convergence as we discussed after our theoretical results above. We also see that the choice $(\alpha = 1, \beta = 0.1)$ provides faster convergence at the beginning but then starts oscillating and reaches the convergence threshold later than gradient descent. It appears that choosing $(\alpha = 0.5, \beta = 0)$ provides the fastest convergence rate, indicating that maybe for this problem, Hessian damping is not as necessary and the solution landscape is smoother.

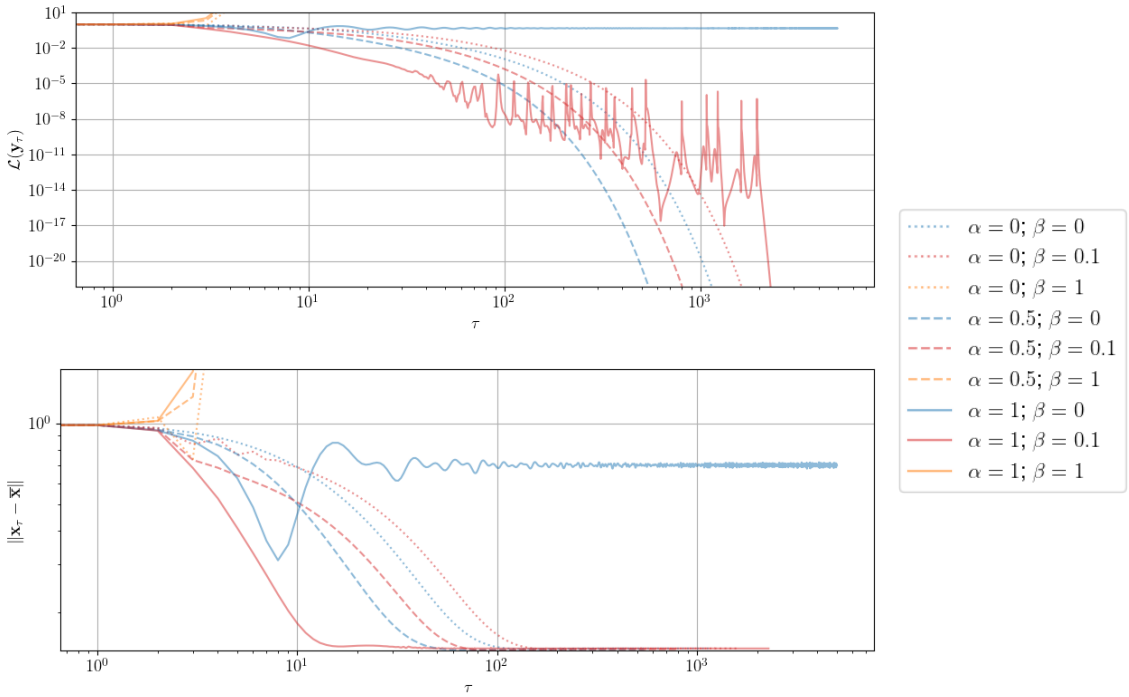


Fig. 5: Effects on the loss and the signal convergence of different combination of α and β for a well-conditioned operator. We observe faster convergence for a variety of parameters and the networks converge to the same signal as the problem is well-posed.

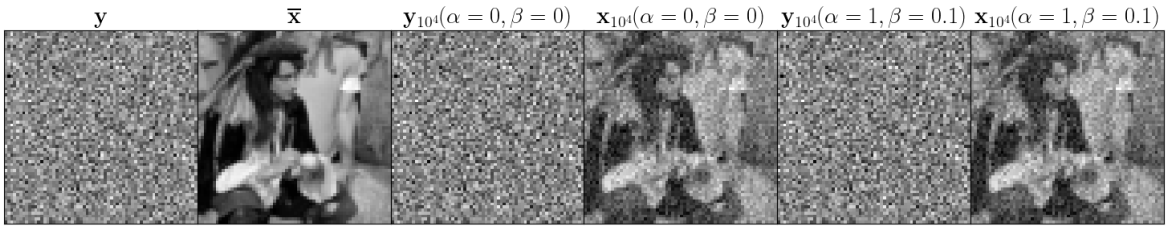


Fig. 6: Qualitative results with a well-conditioned operator and heavy noise for different α and β. The signal is well recovered despite the heavy noise level for both gradient descent and inertial systems.

When we look at the evolution of the curves in Figure 5, there are some surprises. Notably the case $(\alpha = 1, \beta = 0.1)$ which shows these swings in the loss, is the fastest in signal space. Furthermore, we see that for a lot of cases, the algorithm converges in a stable way in the signal space and stays around the solution before overfitting the noise. This was expected as the operator is well-conditioned and thus overfitting the noise, even if it is to quite high level like here, does not require big changes in signal space. We see this effect in the qualitative results of Figure 6 where the solution found by gradient descent and the inertial algorithm are very similar and do not show the same artifacts as was the case for the convolution operator.

References

[1] Alvarez, F., Attouch, H., Bolte, J., Redont, P.: A second-order gradient-like dissipative dynamical system with hessian-driven damping.: Application to optimization and mechanics. Journal de mathématiques pures et appliquées **81**(8), 747–779 (2002)

[2] 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)
[3] Arndt, C.: Regularization theory of the analytic deep prior approach. Inverse Problems **38**(11), 115005 (2022)
[4] 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)
[5] Arridge, S., Maass, P., Ozan, O., Schönlieb, C.B.: Solving inverse problems using data-driven models. Acta Numerica **28**, 1–174 (May 2019)
[6] Attouch, H., Chbani, Z., Fadili, J., Riahi, H.: First-order optimization algorithms via inertial systems with hessian driven damping. Mathematical Programming **193**, 1–43 (2022)
[7] Attouch, H., Fadili, J., Kungurtsev, V.: On the effect of perturbations in first-order optimization methods with inertia and Hessian driven damping. Evolution Equations and Control Theory **12**(1), 71 (2023)
[8] Bartlett, P.L., Montanari, A., Rakhlin, A.: Deep learning: a statistical viewpoint. Acta numerica **30**, 87–201 (2021)
[9] Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics, Springer International Publishing, Cham (2017)
[10] Buskulic, N., Fadili, J., Quéau, Y.: Convergence and recovery guarantees of unsupervised neural networks for inverse problems. Journal of Mathematical Imaging and Vision **66**, 1–22 (2024)
[11] Buskulic, N., Fadili, J., Quéau, Y.: Recovery guarantees of unsupervised neural networks for inverse problems trained with gradient descent. 32nd European Signal Processing Conference (2024)
[12] Castera, C., Bolte, J., Févotte, C., Pauwels, E.: An inertial Newton algorithm for deep learning. J. Mach. Learn. Res. **22**, 1–31 (2021)
[13] Chizat, L., Oyallon, E., Bach, F.: On lazy training in differentiable programming. Advances in neural information processing systems **32** (2019)
[14] Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In: ICLR. pp. 1–19 (2019)
[15] Duff, M., Campbell, N., Ehrhardt, M.J.: Regularising inverse problems with generative machine learning models. Journal of Mathematical Imaging and Vision **66**(1), 37–56 (2024)
[16] Fang, C., Dong, H., Zhang, T.: Mathematical models of overparameterized neural networks. Proceedings of the IEEE **109**(5), 683–703 (2021)
[17] 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)
[18] Haraux, A.: Systèmes dynamiques dissipatifs et applications, Recherches en Mathématiques Appliquées, vol. 17. Masson, Paris (1991)
[19] Heckel, R., Soltanolkotabi, M.: Compressive sensing with un-trained neural networks: Gradient descent finds a smooth approximation. In: International Conference on Machine Learning. pp. 4149–4158 (2020)
[20] Heckel, R., Soltanolkotabi, M.: Denoising and regularization via exploiting the structural bias of convolutional generators. In: International Conference on Learning Representations (2020)
[21] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. Advances in neural information processing systems **31** (2018)
[22] Jahn, T., Jin, B.: Early stopping of untrained convolutional neural networks. SIAM Journal on Imaging Sciences **17**(4), 2331–2361 (2024)
[23] Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative regularization methods for nonlinear ill-posed problems, vol. 6. Walter de Gruyter (2008)
[24] Kamilov, U.S., Bouman, C.A., Buzzard, G.T., Wohlberg, B.: Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. IEEE Signal Processing Magazine **40**(1), 85–97 (2023)
[25] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
[26] Liu, J., Sun, Y., Xu, X., Kamilov, U.S.: Image restoration using total variation regularized deep image prior. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7715–7719 (2019)
[27] 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)
[28] Maulen-Soto, R., Fadili, J., Ochs, P.: Inertial methods with viscous and Hessian driven damping for non-convex optimization. arXiv:2407.12518 (2024)
[29] 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)
[30] Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2013)
[31] 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 pp. 39–56 (May 2020)
[32] Oymak, S., Soltanolkotabi, M.: Overparameterized nonlinear learning: Gradient descent takes the shortest path? In: International Conference on Machine Learning. pp. 4951–4960 (2019)
[33] 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)
[34] Pineda, A.F.L., Petersen, P.C.: Deep neural networks can stably solve high-dimensional, noisy, non-linear inverse problems. Analysis and Applications **21**(01), 49–91 (2023)
[35] Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics **4**(5), 1–17 (Jan 1964)
[36] 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)
[37] Rahimi, A., Recht, B.: Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. Advances in neural information processing systems **21** (2008)
[38] 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)
[39] Tao, G.: A simple alternative to the barbalat lemma. IEEE Transactions on Automatic Control **42**(5), 698– (1997)
[40] Tirer, T., Giryes, R., Chun, S.Y., Eldar, Y.C.: Deep internal learning: Deep learning from a single input. IEEE Signal Processing Magazine **41**(4), 40–57 (2024)
[41] 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)
[42] Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Compressed Sensing: Theory and Applications, pp. 210–268. Cambridge University Press, Cambridge (2012)
[43] Zukerman, J., Tirer, T., Giryes, R.: Bp-dip: A backprojection based deep image prior. In: 28th European Signal Processing Conference. pp. 675–679. IEEE (2021)