Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks

Gauthier Gidel
Mila & DIRO
Université de Montréal

Francis Bach
INRIA & École Normale Supérieure
PSL Research University, Paris

Simon Lacoste-Julien[*]
Mila & DIRO
Université de Montréal

Abstract

When optimizing over-parameterized models, such as deep neural networks, a large set of parameters can achieve zero training error. In such cases, the choice of the optimization algorithm and its respective hyper-parameters introduces biases that will lead to convergence to specific minimizers of the objective. Consequently, this choice can be considered as an implicit regularization for the training of over-parametrized models. In this work, we push this idea further by studying the discrete gradient dynamics of the training of a two-layer linear network with the least-squares loss. Using a time rescaling, we show that, with a vanishing initialization and a small enough step size, this dynamics sequentially learns the solutions of a reduced-rank regression with a gradually increasing rank.

1 Introduction

When optimizing over-parameterized models, such as deep neural networks, a large set of parameters leads to a zero training error. However they lead to different values for the test error and thus have distinct generalization properties. More specifically, Neyshabur [2017, Part II] argues that the choice of the optimization algorithm (and its respective hyperparameters) provides an implicit regularization with respect to its geometry: it biases the training, finding a particular minimizer of the objective.

In this work, we use the same setting as Saxe et al. [2018]: a regression problem with least-squares loss on a multi-dimensional output. Our prediction is made either by a linear model or by a two-layer linear neural network [Saxe et al., 2018]. We extend their work which covered the *continuous* gradient dynamics, to weaker assumptions as well as analyze the behavior of the *discrete* gradient updates

We show that with a vanishing initialization and a small enough step-size, the gradient dynamics of a two-layer linear neural network sequentially learns components that can be ranked according to a hierarchical structure whereas the gradient dynamics induced by the same regression problem but with a linear prediction model instead learns these components simultaneously, missing this notion of hierarchy between components. The path followed by the two-layer formulation actually corresponds to successively solving the initial regression problem with a growing low rank constraint which is also know as reduced-rank regression [Izenman, 1975]. Note that this notion of path followed by *the dynamics of a whole network* is different from the notion of path introduced by Neyshabur et al. [2015a] which corresponds to a path followed *inside* a fixed network, i.e., one corresponds to training dynamics whereas the other corresponds to the propagation of information inside a network.

To sum-up, in our framework, the path followed by the gradient dynamics of a two-layer linear network provides an implicit regularization that may lead to potentially better generalization properties. Our contributions are the following:

- Under some assumptions (see Assumption 1), we prove that both the *discrete* and continuous gradient dynamics sequentially learn the solutions of a gradually less regularized version of reduced-rank regression (Corollary 2 and 3). Among the close related work, such result on implicit regularization regarding discrete dynamics is novel. For the continuous case, we weaken the standard commutativity assumption using perturbation analysis.

- We experimentally verify the reasonableness of our assumption and observe improvements in terms of generalization (matrix reconstruction in our case) using the gradient dynamics of the two-layer linear network when compared against the linear model.

1.1 Related Work

The implicit regularization provided by the choice of the optimization algorithm has recently become an active area of research in machine learning, putting lot of interest on the behavior of gradient descent on deep over-parametrized models [Neyshabur et al., 2015b, 2017, Zhang et al., 2017].

Several works show that gradient descent on *unregularized* problems actually finds a minimum norm solution with respect to a particular norm that drastically depends on the problem of interest. Soudry et al. [2018] look at a logistic regression problem and show that the predictor does converge to the max-margin solution. A similar idea has been developed in the context of matrix factorization [Gunasekar et al., 2017]. Under the assumption that the observation matrices commute, they prove that gradient descent on this non-convex problem finds the minimum nuclear norm solution of the reconstruction problem, they also conjecture that this result would still hold without the commutativity assumption. This conjecture has been later partially solved by Li et al. [2018] under mild assumptions (namely the restricted isometry property). This work has some similarities with ours, since both focus on a least-squares regression problem over matrices with a form of matrix factorization that induces a non convex landscape. Their problem is more general than ours (see Uschmajew and Vandereycken [2018] for an even more general setting) but they are showing a result of a different kind from ours: they focus on the properties of the limit solution the continuous dynamics whereas we show some properties on the whole dynamics (continuous *and* discrete), proving that it actually visits points *during* the optimization that may provide good generalization. Interestingly, both results actually share common assumptions such as a commutativity assumption (which is less restrictive in our case since it is always true in some realistic settings such as linear autoencoders), vanishing initialization and a small enough step size.

Nar and Sastry [2018] also analyzed the gradient descent algorithm on a least-squares linear network model as a discrete time dynamical system, and derived certain necessary (but not sufficient) properties of the local optima that the algorithm can converge to with a non-vanishing step size. In this work, instead of looking at the properties of the limit solutions, we focus on the path followed by the gradient dynamics and precisely caracterize the weights learned along this path.

Combes et al. [2018] studied the continuous dynamics of some non-linear networks under relatively strong assumptions such as the linear separability of the data. Conversely, in this work, we do not make such separability assumption on the data but focus on linear networks.

Finally, Gunasekar et al. [2018] compared the implicit regularization provided by gradient descent in deep linear *convolutional* and *fully connected* networks. They show that the solution found by gradient descent is the minimum norm for both networks but according to a different norm. In this work, the fact that gradient descent finds the minimum norm solution is almost straightforward using standard results on least-squares. But the path followed by the gradient dynamics reveals interesting properties for generalization. As developed earlier, instead of focusing on the properties of the solution found by gradient descent, our goal is to study the path followed by the *discrete* gradient dynamics in the case of a two-layer linear network.

Prior work [Saxe et al., 2013, 2014, Advani and Saxe, 2017, Saxe et al., 2018, Lampinen and Ganguli, 2019] studied the gradient dynamics of two-layer linear networks and proved a result similar to our Thm. 2. We consider Saxe et al. [2018] as the closest related work, we re-use their notion of *simple deep linear neural network*, that we call two-layer neural networks, and use some elements of their proofs to extend their results. However, note that their work comes from a different perspective: through a mathematical analysis of a simple non-linear dynamics, they intend to highlight *continuous* dynamics of learning where one observes the sequential emergence of hierarchically structured notions to explain the regularities in representation of human semantic knowledge. In this work, we

are also considering a two-layer neural network but with an optimization perspective. We are able to extend Saxe et al. [2018, Eq. 6 and 7] weakening the commutativity assumption considered in Saxe et al. [2018] using perturbation analysis. In §4.1, we test to what extent our weaker assumption holds. Our main contribution is to show a similar result on the *discrete* gradient dynamics, that is important in our perspective since we aim to study the dynamics of gradient descent. This result cannot be trivially extended from the result on the *continuous* dynamics. We provide details on the difficulties of the proof in §3.2.

2 A Simple Deep Linear Model

In this work, we are interested in analyzing a least-squares model with multi-dimensional outputs. Given a *finite* number n of inputs $\boldsymbol{x}_i \in \mathbb{R}^d$, $1 \leq i \leq n$ we want to predict a *multi-dimensional outputs* $\boldsymbol{y}_i \in \mathbb{R}^p$, $1 \leq i \leq n$ with a *deep linear network* [Saxe et al., 2018, Gunasekar et al., 2018],

$$\text{Deep linear model:} \quad \hat{\boldsymbol{y}}^d(\boldsymbol{x}) := \boldsymbol{W}_L^\top \cdots \boldsymbol{W}_1^\top \boldsymbol{x}, \tag{1}$$

where $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L$ are learned through a MSE formulation with the least-squares loss f,

$$(\boldsymbol{W}_1^*, \ldots, \boldsymbol{W}_L^*) \in \underset{\substack{\boldsymbol{W}_l \in \mathbb{R}^{r_{l-1} \times r_l} \\ 1 \leq l \leq L}}{\arg\min} \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_1 \cdots \boldsymbol{W}_L\|_2^2 =: f(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_L), \tag{2}$$

where $r_0 = d$, $r_l \in \mathbb{N}$, $1 \leq l \leq L - 1$ and $r_L = p$, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$ are such that,

$$\boldsymbol{X}^\top := (\boldsymbol{x}_1 \quad \cdots \quad \boldsymbol{x}_n) \quad \text{and} \quad \boldsymbol{Y}^\top := (\boldsymbol{y}_1 \quad \cdots \quad \boldsymbol{y}_n), \tag{3}$$

are the *design matrices* of $(\boldsymbol{x}_i)_{1 \leq i \leq n}$ and $(\boldsymbol{y}_i)_{1 \leq i \leq n}$. The *deep linear model* (1) is a L-layer deep linear neural network where we see $\boldsymbol{h}_l := \boldsymbol{W}_l \cdots \boldsymbol{W}_1 \boldsymbol{x}$ for $1 \leq l \leq L - 1$ as the lth *hidden layer*. At first, since this deep linear network cannot represent more than a linear transformation, we could think that there is no reason to use a deeper representation $L = 1$. However, in terms of learning flow, we will see in §3 that for $L = 2$ this model has a completely different dynamics from $L = 1$.

Increasing L may induce a low rank constraint when $r := \min\{r_l : 1 \leq l \leq L - 1\} < \min(d, p)$. In that case, (2) is equivalent to a reduced-rank regression,

$$\boldsymbol{W}^{k,*} \in \underset{\substack{\boldsymbol{W} \in \mathbb{R}^{p \times d} \\ \text{rank}(\boldsymbol{W}) \leq r}}{\arg\min} \frac{1}{2n} \sum_{i=1}^n \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_2^2. \tag{4}$$

These problems have explicit solutions depending on $\boldsymbol{X}$ and $\boldsymbol{Y}$ [Reinsel and Velu, 1998, Thm. 2.2].

Note that, in this work we are interested in the implicit regularization provided in the context of *over-parametrized models*, i.e., when $r > \min(p, d)$. In that case,

$$\{\boldsymbol{W}_1 \cdots \boldsymbol{W}_L : \boldsymbol{W}_l \in \mathbb{R}^{r \times l-1, r_l}, 1 \leq l \leq L\} = \mathbb{R}^{p \times d}.$$

3 Gradient Dynamics as a Regularizer

In this section we would like to study the *discrete* dynamics of the gradient flow of (2), i.e.,

$$\boldsymbol{W}_l^{(t+1)} = \boldsymbol{W}_l^{(t)} - \eta \nabla_{\boldsymbol{W}_l} f\big(\boldsymbol{W}_{[L]}^{(t)}\big) \qquad \boldsymbol{W}_l^{(0)} \in \mathbb{R}^{r_{l-1} \times r_l}, 1 \leq l \leq L, \tag{5}$$

where we use the notation $\boldsymbol{W}_{[L]}^{(t)} := (\boldsymbol{W}_1^{(t)}, \ldots, \boldsymbol{W}_L^{(t)})$. The quantity η is usually called the *step-size*. In order to get intuitions on the discrete dynamics we also consider its respective *continuous* version,

$$\dot{\boldsymbol{W}}_l(t) = -\nabla_{\boldsymbol{W}_l} f\big(\boldsymbol{W}_{[L]}(t)\big) \qquad \boldsymbol{W}_l(0) \in \mathbb{R}^{r_{l-1} \times r_l}, 1 \leq l \leq L, \tag{6}$$

where for $1 \leq l \leq L$, $\dot{\boldsymbol{W}}_l(t)$ is the temporal derivative of $\boldsymbol{W}_l(t)$. Note that there is no step-size in the continuous time dynamics since it actually corresponds to the limit of (5) when $\eta \to 0$. The continuous dynamics may be more convenient to study because such differential equations may have closed form solutions. In §3.1, we will see that under reasonable assumptions it is the case for (6).

3.1 Continuous dynamics

Linear model: $L = 1$. We start with the study of the continuous linear model, its gradient is,

$$\nabla f(\boldsymbol{W}) = \boldsymbol{\Sigma}_x \boldsymbol{W} - \boldsymbol{\Sigma}_{xy}, \tag{7}$$

where $\boldsymbol{\Sigma}_{xy} := \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{Y}$ and $\boldsymbol{\Sigma}_x := \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}$. Thus, $\boldsymbol{W}(t)$ is the solution of the differential equation,

$$\dot{\boldsymbol{W}}(t) = \boldsymbol{\Sigma}_{xy} - \boldsymbol{\Sigma}_x \boldsymbol{W}(t), \quad \boldsymbol{W}(0) = \boldsymbol{W}_0. \tag{8}$$

Proposition 1. *For any $\boldsymbol{W}_0 \in \mathbb{R}^{d \times p}$, the solution to the linear differential equation* (8) *is*

$$\boldsymbol{W}(t) = e^{-t\boldsymbol{\Sigma}_x}(\boldsymbol{W}_0 - \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy}) + \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy}, \tag{9}$$

where $\boldsymbol{\Sigma}_x^\dagger$ is the pseudoinverse of $\boldsymbol{\Sigma}_x$.

This standard result on ODE is provided in §B.1. Note that when $\boldsymbol{W}_0 \to \boldsymbol{0}$ we have

$$\boldsymbol{W}(t) \underset{\boldsymbol{W}_0 \to 0}{\to} (\boldsymbol{I}_d - e^{-t\boldsymbol{\Sigma}_x})\boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy}. \tag{10}$$

Deep linear network: $L \geq 2$. The study of the deep linear model is more challenging since for $L \geq 2$, the landscape of the objective function f is non-convex. The gradient flow of (2) is

$$\nabla f_{\boldsymbol{W}_l}(\boldsymbol{W}_{[L]}) = \boldsymbol{W}_{1:l-1}^\top(\boldsymbol{\Sigma}_x \boldsymbol{W} - \boldsymbol{\Sigma}_{xy})\boldsymbol{W}_{l+1:L}^\top \quad \text{where} \quad \boldsymbol{W}_{i:j} := \boldsymbol{W}_i \cdots \boldsymbol{W}_j, \ 1 \leq l \leq L, \tag{11}$$

where we used the convention that $\boldsymbol{W}_{1,0} = \boldsymbol{I}_d$ and $\boldsymbol{W}_{L+1,L} = \boldsymbol{I}_p$. Thus (6) becomes

$$\dot{\boldsymbol{W}}_l(t) = \boldsymbol{W}_{1:l-1}(t)^\top(\boldsymbol{\Sigma}_{xy} - \boldsymbol{\Sigma}_x \boldsymbol{W}(t))\boldsymbol{W}_{l+1:L}(t)^\top, \quad \boldsymbol{W}_l(0) \in \mathbb{R}^{d \times p}, \quad 1 \leq l \leq L. \tag{12}$$

We obtain a *coupled* differential equation (12) that is harder to solve than the previous linear differential equation (8) due, at the same time, to its non-linear components and to the coupling between $\boldsymbol{W}_l$, $1 \leq l \leq L$. However, in the case $L = 2$, Saxe et al. [2018] managed to find an explicit solution to this coupled differential equation under the assumption that "perceptual correlation is minimal" ($\boldsymbol{\Sigma}_x = \boldsymbol{I}_d$).[2] In this work we extend Saxe et al. [2018, Eq. 7] (for $L = 2$) under weaker assumptions. More precisely, we do not require the covariance matrix $\boldsymbol{\Sigma}_x$ to be the identity matrix. Let $(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D})$ be the SVD of $\boldsymbol{\Sigma}_{xy}$, our assumption is the following:

Assumption 1. *There exist two orthogonal matrices $\boldsymbol{U}$, $\boldsymbol{V}$ such that we have the joint decomposition,*

$$\boldsymbol{\Sigma}_x = \boldsymbol{U}(\boldsymbol{D}_x + \boldsymbol{B})\boldsymbol{U}^\top \quad and \quad \boldsymbol{\Sigma}_{xy} = \boldsymbol{U}\boldsymbol{D}_{xy}\boldsymbol{V}^\top, \tag{13}$$

where $\boldsymbol{B}$ is such that $\|\boldsymbol{B}\|_2 \leq \epsilon$ and $\boldsymbol{D}_x$, $\boldsymbol{D}_{xy}$ are matrices only with diagonal coefficients. We note $\sigma_1 \geq \cdots \geq \sigma_{r_{xy}} > 0$ the singular values of $\boldsymbol{\Sigma}_{xy}$ and $\lambda_1, \ldots, \lambda_{r_x}$ the diagonal entries of $\boldsymbol{D}_x$.

Since two matrices commute if and only if they are co-diagonalizable [Horn et al., 1985, Thm. 1.3.21], the quantity ϵ represent to what extend $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{xy}^\top$ do not commute. Before solving (12) under Assump. 1, we describe some motivating examples where the quantity ϵ is small or zero:

- **Linear autoencoder**: If $\boldsymbol{Y}$ is set to $\boldsymbol{X}$ and $L = 2$, we recover a linear autoencoder: $\hat{\boldsymbol{x}}(\boldsymbol{x}) = \boldsymbol{W}_2^\top \boldsymbol{W}_1^\top \boldsymbol{x}$, where $\boldsymbol{h} := \boldsymbol{W}_1^\top \boldsymbol{x}$ is the *encoded* representation of $\boldsymbol{x}$,

$$\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{xy}^\top = \left(\tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{X}\right)^2 = \boldsymbol{\Sigma}_x^2. \qquad \text{Thus, } \boldsymbol{B} = 0. \tag{14}$$

  Note that this linear autoencoder can also be interpreted as a form of principal component analysis. Actually, if we initialize with $\boldsymbol{W}_1 = \boldsymbol{W}_2^\top$, the gradient dynamics exactly recovers the PCA of $\boldsymbol{X}$, which is closely related to the matrix factorization problem of Gunasekar et al. [2017]. See §A where this derivation is detailed.

- **Deep linear *multi*-class prediction:** In that case, p is the number of classes and $\boldsymbol{y}_i$ is a one-hot encoding of the class with, in practice, $p \ll d$. The intuition on why we may expect $\|\boldsymbol{B}\|_2$ to be small is because $\mathrm{rank}(\boldsymbol{Y}) \ll \mathrm{rank}(\boldsymbol{X})$ and thus the matrices of interest only have to almost commute on a small space in comparison to the whole space, thus $\boldsymbol{B}$ would be close to 0. We verify this intuition by computing $\|\boldsymbol{B}\|_2$ for several classification datasets in Table 1.

- **Minimal influence of perceptual correlation:** $\boldsymbol{\Sigma}_x \approx \boldsymbol{I}_d$. It is the setting discussed by Saxe et al. [2018]. We compare this assumption for some classification datasets with our Assump. 1 in §4.1.

---

[2]By a rescaling of the data, their proof is valid for any matrix $\boldsymbol{\Sigma}_x$ *proportional* to the identity matrix.

An explicit solution for $L = 2$. Under Assump. 1 and specifying the initialization, one can solve the matrix differential equation for $\epsilon = 0$ and then use perturbation analysis to assess how close the solution of (8) is to the closed form solution derived for $\epsilon = 0$. This result is summarized in the following theorem proved in §B.2.

Theorem 1. *When $L = 2$, under Assump. 1, if we initialize with $\boldsymbol{W}_1(0) = \boldsymbol{U} \operatorname{diag}(e^{-\delta_1}, \dots, e^{-\delta_p})\boldsymbol{Q}$ and $\boldsymbol{W}_2(0) = \boldsymbol{Q}^{-1} \operatorname{diag}(e^{-\delta_1}, \dots, e^{-\delta_d})\boldsymbol{V}^\top$ where $\boldsymbol{Q}$ is an arbitrary invertible matrix, then the solution of (12) can be decomposed as the sum of the solution for $\epsilon = 0$ and a perturbation term,*

$$\begin{cases} \boldsymbol{W}_1(t) = \boldsymbol{W}_1^0(t) + \boldsymbol{W}_1^\epsilon(t) & where \quad \boldsymbol{W}_1^0(t) := \boldsymbol{U} \operatorname{diag}\left(\sqrt{w_1(t)}, \dots, \sqrt{w_p(t)}\right)\boldsymbol{Q} \\ \boldsymbol{W}_2(t) = \boldsymbol{W}_1^0(t) + \boldsymbol{W}_2^\epsilon(t) & where \quad \boldsymbol{W}_2^0(t) := \boldsymbol{Q}^{-1} \operatorname{diag}\left(\sqrt{w_1(t)}, \dots, \sqrt{w_d(t)}\right)\boldsymbol{V}^\top \end{cases} \quad (15)$$

where we have $c > 0$ such that $\|\boldsymbol{W}_i^\epsilon(t)\| \le \epsilon \cdot e^{ct^2}$ and,

$$w_i(t) = \frac{\sigma_i e^{2\sigma_i t - 2\delta_i}}{\lambda_i(e^{2\sigma_i t - 2\delta_i} - e^{-2\delta_i}) + \sigma_i}, \ 1 \le i \le r_{xy}, \ w_i(t) = \frac{e^{-2\delta_i}}{1 + 2e^{-\delta_i}\lambda_i t}, \ r_{xy} < i \le r_x \quad (16)$$

where (σ_i) and (λ_i) are defined is Assump. 1. Note that $\operatorname{rank}(\boldsymbol{\Sigma}_{xy}) := r_{xy} \le \operatorname{rank}(\boldsymbol{\Sigma}_x) := r_x$.

The main difficulty in this result is the perturbation analysis for which we use a consequence of Grönwall's inequality [Gronwall, 1919] (Lemma 4). The proof can be sketched in three parts: first showing the result for $\epsilon = 0$, then showing that in the case $\epsilon > 0$, the matrices $\boldsymbol{W}_1(t)/t$ and $\boldsymbol{W}_2(t)/t$ are bounded and finally use Lemma 4 to get the perturbation bound.

This result is more general than the one provided by Saxe et al. [2018] because it requires a weaker assumption than $\boldsymbol{\Sigma}_x = \boldsymbol{I}_d$ and $\epsilon = 0$. In doing so, we obtain a result that takes into account the influence of correlations of the input samples. Note that Thm. 1 is only valid if the initialization $\boldsymbol{W}_1(0)\boldsymbol{W}_2(0)$ has the same singular vectors as $\boldsymbol{\Sigma}_{xy}$. However, making such assumptions on the initialization is standard in the literature and, in practice, we can set the initialization of the optimization algorithm in order to also ensure that property. For instance, in the case of the linear autoencoder, one can set $\boldsymbol{W}_1(0) = \boldsymbol{W}_2(0) = e^{-\delta}\boldsymbol{I}_d$.

In the following subsection we will use Thm. 1 to show that the components $[\boldsymbol{U}]_i$, $1 \le i \le r_{xy}$ in the order defined by the decreasing singular values of $\boldsymbol{\Sigma}_{xy}$ are learned sequentially by the gradient dynamics.

Sequential learning of components. The sequential learning of the left singular vectors of $\boldsymbol{\Sigma}_{xy}$ (sorted by the magnitude of its singular values) by the *continuous* gradient dynamics of deep linear networks has been highlighted by Saxe et al. [2018]. They note in their Eq. (10) that the ith phase transition happens approximately after a time T_i defined as (using our notation),

$$T_i := \frac{\delta_i}{\sigma_i} \ln(\sigma_i) \quad where \quad \boldsymbol{\Sigma}_{xy} = \sum_{i=1}^{r_{xy}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top . \quad (17)$$

They argue that as $\delta_i \to \infty$, the time T_i is roughly $O(1/\sigma_i)$. The intuition is that a vanishing initialization increases the gap between the phase transition times T_i and thus tends to separate the learning of each components. However, a vanishing initialization just formally leads to $T_i \to \infty$.

In this work, we introduce a notion of *time rescaling* in order to formalize this notion of phase transition and we show that, after this time rescaling, the point visited between two phase transitions is the solution of a low rank regularized version (4) of the initial problem (2) with the low rank constraint that loosens sequentially.

The intuition behind time rescaling is that it counterbalances the vanishing initialization in (17): Since T_i grows as fast as δ_i we need to multiply the time by δ_i, in order to grow at the same pace as T_i.

Using this rescaling we can present our theorem, proved in §B.3, which says that a vanishing initialization tends to force the sequential learning of the component of $\boldsymbol{X}$ associated with the largest singular value of $\boldsymbol{\Sigma}_{xy}$. Note that we need to rescale the time *uniformly* for each component. That is why in the following we set $\delta_i = \delta$, $1 \le i \le \max(p, d)$.

Theorem 2. *Let us denote $w_i(t)$, the values defined in (16). If $w_i(0) = e^{-\delta}$, $1 \le i \le r$, and $\epsilon = e^{-\delta^2 \ln(\delta)}$ then we have that $w_i(\delta t)$ converge to a step function as $\delta \to \infty$:*

$$w_i(\delta t) \underset{\delta \to \infty}{\to} \frac{\sigma_i}{\lambda_i + \sigma_i} \mathbb{1}\{t = T_i\} + \frac{\sigma_i}{\lambda_i}\mathbb{1}\{t > T_i\} . \quad (18)$$

where $T_i := 1/\sigma_i$, $\mathbb{1}\{t \in A\} = 1$ if $t \in A$ and 0 otherwise.

Notice how the ith components of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are inactive, i.e., $w_i(t)$ is zero, for small t and is suddenly learned when t reaches the phase transition time $T_i := 1/\sigma_i$. As shown in Prop. 1 and illustrated in Fig. 1, this sequential learning behavior does not occur for the non-factorized formulation. Gunasekar et al. [2017] observed similar differences between their factorized and not factorized formulations of matrix regression. Note that, the time rescaling we introduced is $t \to \delta t$, in order to compensate the vanishing initialization, rescaling the time and taking the limit this way for (8) would lead to a constant function.

Gunasekar et al. [2017] also had to consider a vanishing initialization in order to show that on a simple matrix factorization problem the continuous dynamics of gradient descent does converge to the minimum nuclear norm solution. This assumption is necessary in such proofs in order to avoid to initialize with wrong components. However one cannot consider an initialization with the null matrix since it is a stationary point of the dynamics, that is why this notion of double limit (vanishing initialization and $t \to \infty$) is used.

From Thm. 2, two corollaries follow directly. The first one regards the nuclear norm of the product $\boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)$. This corollary says that $\|\boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)\|_*$ is a step function and that each increment of this integer value corresponds to the learning of a new component of $\boldsymbol{X}$. These components are leaned by order of relevance, i.e., by order of magnitude of their respective eigenvalues and the learning of a new component can be easily noticed by an incremental gap in the nuclear norm of the matrix product $\boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)$,

Corollary 1. *Let $\boldsymbol{W}_1(t)$ and $\boldsymbol{W}_2(t)$ be the matrices solution of (12) defined in (15). The limit of the squared euclidean norm of $\boldsymbol{W}_1(t)\boldsymbol{W}_2(t)$ when $\delta \to \infty$ is a step function defined as,*

$$\|\boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)\|_2^2 \underset{\delta \to \infty}{\to} \sum_{i=1}^{r_{xy}} \frac{\sigma_i^2}{\lambda_i^2} \mathbb{1}\{T_i < t\} + \frac{\sigma_i^2}{(\lambda_i+\sigma_i)^2} \mathbb{1}\{T_i = t\} \tag{19}$$

where $T_i := 1/\sigma_i$ and $\sigma_1 > \cdots > \sigma_{r_{xy}} > 0$ are the positive singular values of $\boldsymbol{\Sigma}_{xy}$.

It is natural to look at the norm of the product $\boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)$ since in Thm. 2, $(w_i(t))$ are its singular values. However, since the rank of $\boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)$ is discontinuously increasing after each phase transition, any norm would jump with respect to the rank increments. We illustrate these jumps in Fig. 1 where we plot the closed form of the squared ℓ_2 norms of $t \mapsto \boldsymbol{W}(\delta t)$ and $t \mapsto \boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)$ for vanishing initializations with $\boldsymbol{\Sigma}_{yx} = \text{diag}(10^{-1}, 10^{-2}, 10^{-3})$ and $\boldsymbol{\Sigma}_x = \boldsymbol{I}_d$.

Figure 1: Closed form solution of squared ℓ_2 norm of $\boldsymbol{W}(\delta t)$ and $\boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t)$ respectively for a linear model and a two-layer linear autoencoder, depending on $\boldsymbol{W}(0) = \boldsymbol{W}_1(0)\boldsymbol{W}_2(0) = e^{-\delta}\boldsymbol{I}_d$. Note that for an autoencoder $\lambda_i = \sigma_i$ and thus the trace norm has integer values. According to Thm. 2, the integer trace norm increment represents the learning of a new component.

From Thm. 2, we can notice that, between time T_k and T_{k+1}, the rank of the limit matrix $\boldsymbol{W}_1\boldsymbol{W}_2$ is actually equal to k, meaning that at each phase transition, the rank of $\boldsymbol{W}_1\boldsymbol{W}_2$ is increased by 1. Moreover, this matrix product contains the k components of $\boldsymbol{X}$ corresponding to the k largest singular values of $\boldsymbol{\Sigma}_{xy}$. Thus, we can show that this matrix product is the solution of the k-low rank constrained version (4) of the initial problem (2),

Corollary 2. *Let $\boldsymbol{W}_1(t)$ and $\boldsymbol{W}_2(t)$ be the matrices solution of (12) defined in (15). We have that,*

$$\frac{1}{\sigma_k} < t < \frac{1}{\sigma_{k+1}} \quad \Rightarrow \quad \boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t) \underset{\delta \to \infty}{\to} \boldsymbol{W}^{k,*}, \qquad 1 \le k \le r_{xy}. \tag{20}$$

where $\boldsymbol{W}^{k,*}$ is the minimum ℓ_2 norm solution of the reduced-rank-k regression problem (4) .
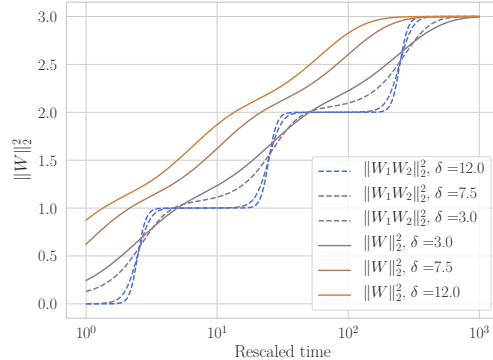
6

3.2 Discrete dynamics

We are interested in the behavior of optimization methods. Thus, the gradient dynamics of interest is the *discrete* one (5). A major contribution of our work is thus contained in this section. The continuous case studied in § 3.1 provided good intuitions and insights on the behavior of the potential discrete dynamics that we can use for our analysis.

Why the discrete analysis is challenging. Previous related work [Gunasekar et al., 2017, Saxe et al., 2018] only provide results on the continuous dynamics. Their proofs use the fact that their respective continuous dynamics of interest have a closed form solution (e.g., Thm.1). To our knowledge, no closed form solution is known for the discrete dynamics (5). Thus its analysis requires a new proof technique. Moreover, using Euler's integration methods, one can show that both dynamics are close but only for a vanishing step size depending on a *finite* horizon. Such dependence on the horizon is problematic since the time rescaling used in Thm. 2 would make any finite horizon go to infinity. In this section, we consider realistic conditions on the step-size (namely, it has to be smaller than the Lipschitz constant and some notion of eigen-gap) without any dependence on the horizon. Such assumption is relevant since we want to study the dynamics of practical optimization algorithms (i.e., with a step size as large as possible and without knowing in advance the horizon).

Single layer linear model. In this paragraph, we consider the discrete update for the linear model. Since $L = 1$, for notational compactness, we call $\boldsymbol{W}_t$ the matrix updated according to (5). Using the gradient derivation (7), the discrete update scheme for the *linear model* is,

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta(\boldsymbol{\Sigma}_x \boldsymbol{W}_t - \boldsymbol{\Sigma}_{xy}) = (\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)\boldsymbol{W}_t + \eta\boldsymbol{\Sigma}_{xy}.$$

Noticing that for $1/\lambda_{\max}(\boldsymbol{\Sigma}_x) > \eta > 0$, $\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x$ is invertible, this recursion (see §B.4) leads to,

$$\boldsymbol{W}_t = (\boldsymbol{W}_0 - \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy})(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)^t + \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy}. \tag{21}$$

We obtain a similar result as the solution of the differential equation given in Prop. 1. With a vanishing initialization we reach a function that does not sequentially learn some components.

Two-layer linear model. The discrete update scheme for the *two-layer linear network* (2) is,

$$\boldsymbol{W}_1^{(t+1)} = \boldsymbol{W}_1^{(t)} - \eta(\boldsymbol{\Sigma}_x \boldsymbol{W}^{(t)} - \boldsymbol{\Sigma}_{xy})(\boldsymbol{W}_2^{(t)})^\top, \quad \boldsymbol{W}_2^{(t+1)} = \boldsymbol{W}_2^{(t)} - \eta(\boldsymbol{W}_1^{(t)})^\top(\boldsymbol{\Sigma}_x \boldsymbol{W}^{(t)} - \boldsymbol{\Sigma}_{xy}).$$

When $\epsilon = 0$, by a change of basis and a proper initialization, we can reduce the study of this matrix equation to r independant dynamics (see §B.5 for more details), for $1 \leq i \leq r$,

$$w_i^{(t+1)} = w_i^{(t)} + \eta w_i^{(t)}(\sigma_i - \lambda_i w_i^{(t)} w_i^{(t)}). \tag{22}$$

Thus we can derive a bound on the iterate $w_i^{(t)}$ leading to the following theorem,

Theorem 3. *Under the same assumptions as Thm. 1 and $\epsilon = 0$, we have*

$$\boldsymbol{W}_1^{(t)} = \boldsymbol{U} \operatorname{diag}\left(\sqrt{w_1^{(t)}}, \ldots, \sqrt{w_p^{(t)}}\right)\boldsymbol{Q} \quad and \quad \boldsymbol{W}_2^{(t)} = \boldsymbol{Q}^{-1} \operatorname{diag}\left(\sqrt{w_1^{(t)}}, \ldots, \sqrt{w_d^{(t)}}\right)\boldsymbol{V}^\top.$$

Moreover, for any $1 \leq i \leq r_{xy}$, if $1 > w_i^{(0)} > 0$ and $2\eta\sigma_i < 1$, then $\forall t \geq 0$, $1 \leq i \leq r_x$ we have,

$$\frac{w_i^{(0)}}{(\sigma_i - \lambda_i w_i^{(0)})e^{(-2\eta\sigma_i + 4\eta^2\sigma_i^2)t} + w_i^{(0)}\lambda_i} \leq w_i^{(t)} \leq \frac{w_i^{(0)}}{(\sigma_i - \lambda_i w_i^{(0)})e^{(-2\eta\sigma_i - \eta^2\sigma_i^2)t} + w_i^{(0)}\lambda_i}, \tag{23}$$

and $w_i^{(t)} \leq \frac{w_i^{(0)}}{1 + w_i^{(0)}\lambda_i \eta t}$ for $r_{xy} \leq i \leq r_x$. The differences with the continuous case (16) are in red.

Proof sketch. The solution of the continuous dynamics lets us think directly studying the sequence $w_i^{(t)}$ might be quite challenging since the solution of the continuous dynamics $w_i(t)^{-1}$ has a non-linear and non-convex behavior.

The main insight from this proof is that one can treat the discrete case using the right transformation, to show that some sequence doee converge linearly.
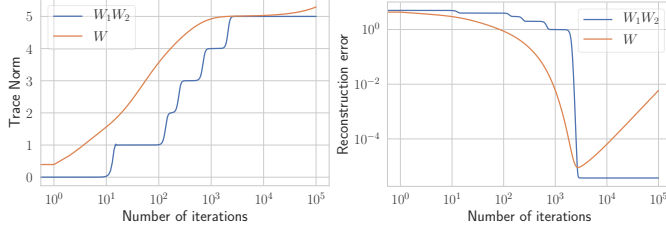
7

Figure 2: Trace norm and reconstruction errors of $\boldsymbol{W}^{(t)}$ for $L = 1$ and 2 as a function of t.

Table 1: Value of the quantities Δ_{xy} and Δ_x defined in (27).

Dataset	Δ_{xy}	Δ_x
MNIST	2.8×10^{-2}	.70
CIFAR-10	3.0×10^{-2}	.68
ImageNet	1.7×10^{-1}	.70

Thm. 2 indicates the quantity $w_i(t)^{-1} - \frac{\lambda_i}{\sigma_i}$ is the good candidate to show linear convergence to 0,

$$w_i(t)^{-1} - \frac{\sigma_i}{\lambda_i} = (w_i(0)^{-1} - \frac{\sigma_i}{\lambda_i})e^{-2\eta\sigma_i t}. \tag{24}$$

What we can expect is thus to show that the sequence $(w_i^{(t)})^{-1} - \frac{\sigma_i}{\lambda_i}$ has similar properties. The first step of the proof is to show that $(w_i^{(t)})$ is an increasing sequence smaller than one. The second step is then to use (22) to get,

$$\frac{1}{w_i^{(t+1)}} - \frac{\lambda_i}{\sigma_i} = \frac{1}{w_i^{(t)}}\left(\frac{1}{1+2(\sigma_i - \lambda_i w_i^{(t)}) + \eta^2(\sigma_i - \lambda_i w_i^{(t)})^2}\right) - \frac{\lambda_i}{\sigma_i}$$

Then using that $1 - x \leq \frac{1}{1+x} \leq 1 - x + x^2$ for any $1 \leq x \leq 0$ we can derive the upper and lower bounds on the linear convergence rate. See §B.5 for full proof. □

In order to get a similar interpretation of Thm. 3 in terms of implicit regularization, we use the intuitions from Thm. 2. The analogy between continuous and discrete time is that the discrete time dynamics is doing t time-steps of size η, meaning that we have $\boldsymbol{W}(\eta t) \approx \boldsymbol{W}_t$, the time rescaling in continuous time consists in multiplying the time by δ thus we get the analog phase transition time,

$$\eta T_i := \frac{1}{\sigma_i} \quad \Rightarrow \quad T_i := \frac{1}{\eta\sigma_i}. \tag{25}$$

Recall that we assumed that $m_i^{(0)} = n_i^{(0)} = e^{-\delta}$. Thus, we show that the ith component is learned around time T_i, and consequently that the components are learned sequentially,

Corollary 3. *If* $\eta < \frac{1}{2\sigma_1}$, $\eta < 2\frac{\sigma_i - \sigma_{i+1}}{\sigma_i^2}$ *and* $\eta < \frac{\sigma_i - \sigma_{i+1}}{2\sigma_{i+1}^2}$, *for* $1 \leq i \leq r_{xy} - 1$, *then for* $1 \leq i < r_x$,

$$w_i^{(\delta T_j)} \underset{\delta\to\infty}{\to} \begin{cases} 0 & \text{if } i > r_{xy} \quad \text{or} \quad j < i \\ \frac{\sigma_i}{\lambda_i} & \text{if } i \leq r_{xy} \quad \text{and} \quad j > i. \end{cases} \tag{26}$$

where $T_0 := 0$, $T_j := \frac{1}{\sigma_j \eta}$, $1 \leq j \leq r_{xy}$ *and* $T_j := +\infty$ *if* $j > r_{xy}$.

This result is proved in §B.5. The quantities $\frac{\sigma_i - \sigma_{i+1}}{\sigma_i^2}$ and $\frac{\sigma_i - \sigma_{i+1}}{\sigma_{i+1}^2}$ can be interpreted as relative *eigen-gaps*. Note that they are well defined since we assumed that the eigenspaces were unidimensional. The intuition behind this condition is that the step-size cannot be larger than the eigen-gaps because otherwise the discrete optimization algorithm would not be able to distinguish some components.

4 Experiments

4.1 Assump. 1 for Classification Datasets

In this section we verify to what extent Assump. 1 is true on standard classification datasets. For this, we compute the normalized quantities Δ_{xy} and Δ_x representing how much Assump. 1 and the assumption that $\boldsymbol{\Sigma}_x \approx \boldsymbol{I}_d$ are respectively broken. We compute $\boldsymbol{B}$ by computing $\boldsymbol{U}$, the left singular vector of $\boldsymbol{\Sigma}_{xy}$ and looking at the non-diagonal coefficients of $\boldsymbol{U}^\top\boldsymbol{\Sigma}_x\boldsymbol{U}$,

$$\Delta_{xy} := \frac{\|\boldsymbol{B}\|_2}{\|\boldsymbol{\Sigma}_x\|_2}, \quad \Delta_x := \frac{1}{2}\|\hat{\boldsymbol{\Sigma}}_x - \hat{\boldsymbol{I}}_d\|_2, \tag{27}$$

where $\|\cdot\|$ is the Frobenius norm, the $\hat{\Sigma}$ expressions correspond to $\hat{X} := X/\|X\|$ and $\hat{I}_d := I_d/\|I_d\|$. These normalized quantities are between 0 and 1. The closer to 1, the less the assumption hold and conversely, the closer to 0, the more the assumption approximately holds. We present the results on three standard classification datasets, MNIST [LeCun et al., 2010], CIFAR10 [Krizhevsky et al., 2014] and ImageNet [Deng et al., 2009], a down-sampled version of ImageNet with images of size 64×64. In Table 1, we can see that the quantities Δ_x and Δ_{xy} do not vary much among the datasets and that the Δ associated with our our Assump. 1 is two orders of magnitude smaller than the Δ associated with Saxe et al. [2018]'s assumption indicating the relevance of our assumption.

4.2 Linear Autoencoder

For an auto-encoder, we have, $Y = X$. We want to compare the reconstruction properties of $W^{(t)}$ computed though (21) and of the matrix product $W_1^{(t)} W_2^{(t)}$ where $W_1^{(t)}$ and $W_2^{(t)}$ are computed though (22). In this experiment, we have $p = d = 20, n = 1000, r = 5$ and we generated synthetic data. First we generate a *fixed* matrix $B \in \mathbb{R}^{d \times r}$ such that, $B_{kl} \sim \mathcal{U}([0, 1]), 1 \leq k, l \leq n$. Then, for $1 \leq i \leq n$, we sample $x_i \sim B z_i + \epsilon_i$ where $z_i \sim \mathcal{N}(\mathbf{0}, D := \mathrm{diag}(4, 2, 1, 1/2, 1/4))$ and $\epsilon_i \sim 10^{-3}\mathcal{N}(\mathbf{0}, I_d)$. In Fig. 2, we plot the trace norm of $W^{(t)}$ and $W_1^{(t)} W_2^{(t)}$ as well as their respective reconstruction errors as a function of t the number of iterations,

$$\|W^{(t)} - BDB^\top\|_2. \tag{28}$$

We can see that the experimental results are very close to the theoretical behavior predicted with the continuous dynamics in Figure 1. Contrary to the dynamics induced by the linear model formulation ($L = 1$), the dynamics induced by the two-layer linear network ($L = 2$) is very close to a step function: each step corresponds to the learning to a new component: They are learned sequentially.

5 Discussion

There is a growing body of empirical and theoretical evidence that the implicit regularization induced by gradient methods is key in the training of deep neural networks. Yet, as noted by Zhang et al. [2017], even for linear models, our understanding of the origin of generalization is limited. In this work, we focus on a simple non-convex objective that is parametrized by a two-layer linear network. In the case of linear regression we show that the *discrete* gradient dynamics also visits points that are implicitly regularized solutions of the initial optimization problem. In that sense, in the context of machine learning, applying gradient descent on the overparametrized model of interest, provides a form of implicit regularization: it sequentially learns the hierarchical components of our problem which could help for generalization. Our setting does not pretend to solve generalization in deep neural networks; many majors components of the standard neural network training are omitted such as the non-linearities, large values of L and the stochasticity in the learning procedure (SGD). Nevertheless, it provides useful insights about the source of generalization in deep learning.

References

M. S. Advani and A. M. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

N. Berglund. Perturbation theory of dynamical systems. *arXiv preprint math/0111178*, 2001.

E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. Tata McGraw-Hill Education, 1955.

R. T. d. Combes, M. Pezeshki, S. Shabanian, A. Courville, and Y. Bengio. On the learning dynamics of deep neural networks. *arXiv preprint arXiv:1809.06848*, 2018.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 1919.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *NIPS*, 2017.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018.

R. A. Horn, R. A. Horn, and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 1975.

A. Krizhevsky, V. Nair, and G. Hinton. The CIFAR-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.

A. K. Lampinen and S. Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *ICLR*, 2019.

Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2010.

Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

K. Nar and S. Sastry. Step size matters in deep learning. In *NeurIPS*, 2018.

B. Neyshabur. *Implicit Regularization in Deep Learning*. PhD thesis, TTIC, 2017.

B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *NIPS*, 2015a.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR*, 2015b.

B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

G. C. Reinsel and R. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer Science & Business Media, 1998.

A. M. Saxe, J. L. McClellans, and S. Ganguli. Learning hierarchical categories in deep neural networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2013.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, 2014.

A. M. Saxe, J. L. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *arXiv preprint arXiv:1810.10531*, 2018.

D. Soudry, E. Hoffer, and N. Srebro. The implicit bias of gradient descent on separable data. In *ICLR*, 2018.

A. Uschmajew and B. Vandereycken. On critical points of quadratic low-rank matrix optimization problems. Tech. report (submitted), July 2018.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. 2017.

A Deep Linear Autoencoder Recovers PCA.

Let us recall that the two-layer linear autoencoder can be formulated as,

$$(W_2^*, W_1^*) \in \underset{\substack{W_2 \in \mathbb{R}^{r \times p} \\ W_1 \in \mathbb{R}^{d \times r}}}{\arg\min} \frac{1}{2n} \|X - XW_1 W_2\|_2^2. \tag{29}$$

Thus, the gradients of the objective are,

$$\nabla f_{W_2}(W_2, W_1) = W_1^\top (\Sigma_x W_1 W_2 - \Sigma_x) \quad \text{and} \quad \nabla f_{W_1}(W_2, W_1) = (\Sigma_x W_1 W_2 - \Sigma_x) W_2^\top.$$

Thus, if $W_2^{(0)} = (W_1^{(0)})^\top$ and if $[W_1^{(0)} W_2^{(0)}, \Sigma_x] = 0$, we have that,

$$\nabla f_M(W_2^{(0)}, W_1^{(0)}) = (W_1^{(0)})^\top (\Sigma_x W_1^{(0)} W_2^{(0)} - \Sigma_x) \tag{30}$$

$$= ((\Sigma_x W_1^{(0)} W_2^{(0)} - \Sigma_x)(W_2^{(0)})^\top)^\top \tag{31}$$

$$= \nabla f_N(W_2^{(0)}, W_1^{(0)})^\top. \tag{32}$$

Thus, for the discrete case, by a recurrence we have that, $W_1^{(t)} = (W_2^{(t)})^\top$, $t \geq 0$ and for the continuous case, invoking the Cauchy-Lipschitz theorem, we have that $W_1(t) = W_2(t)^\top$, $t \geq 0$. Consequently, the limit solution is such that

$$W_1^* \in \underset{W_1 \in \mathbb{R}^{d \times r}}{\arg\min} \frac{1}{2n} \|X - XW_1 W_1^\top\|_2^2, \tag{33}$$

which is a formulation of the PCA.

B Proof of Theorems and Propositions

B.1 Proof of Prop. 1

Proposition' 1. *For any $W_0 \in \mathbb{R}^{d \times p}$, the solution to the linear differential equation (8) is,*

$$W(t) = e^{-t\Sigma_x}(W_0 - \Sigma_x^\dagger \Sigma_{xy}) + \Sigma_x^\dagger \Sigma_{xy}, \tag{34}$$

where $\Sigma_x^\dagger$ is the pseudoinverse of Σ_x.

We can differentiate (34) and check if it verifies (8). In order to do that, we just need to notice that $\Sigma_x \Sigma_x^\dagger \Sigma_{xy} = \Sigma_{xy}$. To see that we compute the SVD of $X^\top = U^\top D V$ where D is a rectangular matrix with only diagonal coefficients such that,

$$DD^\top = \mathrm{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0). \tag{35}$$

Thus, we have $\Sigma_x = U^\top \mathrm{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0) U$ and $\Sigma_x^\dagger = U^\top \mathrm{diag}(1/\lambda_1, \dots, 1/\lambda_r, 0, \dots, 0) U$. Leading to,

$$\Sigma_x \Sigma_x^\dagger \Sigma_{xy} = U^\top D V Y = \Sigma_{xy}.$$

Consequently, the matrix valued function $W(t)$ defined in (34) verifies (8). Now we just need to use Cauchy-Lipschitz theorem [Coddington and Levinson, 1955] (a.k.a. Picard–Lindelöf theorem) to say that this solution is the unique solution of the ODE (8).

B.2 proof of Thm. 1

Commutative case, $\epsilon = 0$: We use ideas from [Saxe et al., 2018] and combine it with Assum. 1 for $\epsilon = 0$. Note that $\epsilon = 0$ if and only if Σ_x and Σ_{xy} commute. thus, we have that,

$$\Sigma_{xy} = U D_{xy} V^\top \quad \text{and} \quad \Sigma_x = U D_x U^\top. \tag{36}$$

Let us consider a generalization of the linear transformation proposed by Saxe et al. [2018, Eq. S6,S7],

$$\bar{W}_1 = U^\top W_1 Q_1, \quad \bar{W}_l = Q_{l-1}^{-1} W_l Q_l, \ 2 \leq l \leq L-1, \quad \text{and} \quad W_L = Q_{L-1}^{-1} W_L V, \tag{37}$$

where $\mathbf{Q}_l$, $1 \leq l \leq L-1$ are arbitrary invertible matrices. Then, noting $\mathbf{Q}_0 := \mathbf{U}$ and $\mathbf{Q}_L := \mathbf{V}$, we get the following dynamics,

$$\frac{d\bar{\mathbf{W}}_l(t)}{dt} = \mathbf{Q}_{l-1}^{-1} \mathbf{W}_{1:l-1}(t)^\top (\mathbf{\Sigma}_{xy} - \mathbf{\Sigma}_x \mathbf{W}(t)) \mathbf{W}_{l+1:L}(t)^\top \mathbf{Q}_l, \quad 1 \leq l \leq L. \tag{38}$$

Thus using (36), the fact that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_d$ and that for any invertible matrix $\mathbf{Q}$, we have $(\mathbf{Q}^{-1})^\top = (\mathbf{Q}^\top)^{-1}$, we get that,

$$\frac{d\bar{\mathbf{W}}_l(t)}{dt} = \bar{\mathbf{W}}_{1:l-1}(t)^\top (\mathbf{D}_{xy} - \mathbf{D}_x \bar{\mathbf{W}}(t)) \bar{\mathbf{W}}_{l+1:L}(t)^\top, \quad \mathbf{W}_l(0) = \mathbf{W}_l^{(0)} \ 1 \leq l \leq L. \tag{39}$$

Using the same argument as [Saxe et al., 2018], if $\bar{\mathbf{W}}_l(t)$, $1 \leq l \leq L$ only have diagonal coefficients then their derivative also only have diagonal coefficients. Thus, if we initialize $\mathbf{W}_l^{(0)}$, $1 \leq l \leq L$, only with diagonal coefficients we have a decoupled solution for each diagonal coefficient. This argument can be formalized using Cauchy-Lipschitz theorem: (39) has a unique solution which is the one we will exhibit in the following.

Recall that we noted $r_0 = d$ and $r_L = p$ and that $\mathbf{W}_l \in \mathbb{R}^{r \times l-1, r_l}$. Let us note, $r = \min\{r_l : 0 \leq l \leq L-1\}$ and $w_{l,i}(t)$, $1 \leq i \leq r$ the respective diagonal coefficients of $\mathbf{W}_l(t)$ for $1 \leq l \leq L$. Note that for $i \geq r$ one could define diagonal coefficients for some of the matrices $\mathbf{W}_l$ but their gradient will be equal to 0, thus non-trivial dynamics only occur for $i \leq r$. They follow the equation,

$$\dot{w}_{l,i}(t) = w_{-l,i}(t)(\sigma_i - \lambda_i w_i(t)), \ w_{l,i}(0) \in \mathbb{R}, \quad 1 \leq l \leq L, \quad 1 \leq i \leq r, \tag{40}$$

where the notation $w_{-l,i}(t)$ stands for the product of the $w_{k,i}(t)$, $1 \leq k \leq L$ omitting $w_{l,i}(t)$, i.e.,

$$w_{-l,i}(t) := \prod_{\substack{k=1 \\ k \neq l}}^{L} w_{k,i}(t), \tag{41}$$

and $w_i(t)$ stands for the product of the $w_{k,i}(t)$, $1 \leq k \leq L$. The difference with [Saxe et al., 2018] is that, since they only consider the case $\mathbf{\Sigma}_x = \mathbf{I}_d$ they have $\lambda_i = 1$, they also only consider the case $L = 2$. The use of Assumption 1 allowed us to work in a more general case.

We will assume that if $w_{l,i}(t) = w_{k,i}(t)$, $1 \leq l, k \leq L$, to find an analytic solution and then show that if $w_{l,i}(0) = w_{k,i}(0)$, $1 \leq k, l \leq L$ then this analytic solution verifies (40) and thus, by Cauchy-Lipschitz theorem, is the unique solution of the non-linear differential equation.

Thus, considering $w_i(t) := w_{1,i}(t) \cdots w_{L,i}(t)$, and assuming that $w_{l,i}(t) = w_{k,i}(t)$, $1 \leq l, k \leq L$, we get that, for $1 \leq i \leq r$,

$$\dot{w}_i(t) = \sum_{l=1}^{L} w_{1,i}(t) \cdots w_{l-1,i}(t) \dot{w}_{l,i}(t) w_{l+1,i}(t) \cdots w_{L,i}(t) \tag{42}$$

$$= L w_i(t)^{2-2/L}(\sigma_i - \lambda_i w_i(t)), \ w_i(0) \in \mathbb{R}. \tag{43}$$

Lemma 1. *If $w_i(0) \in (0, \frac{\sigma_i}{\lambda_i})$, then the differential equations has a unique solution that is increasing and $w_i(t) \in (0, \frac{\sigma_i}{\lambda_i})$, $\forall t \in \mathbb{R}$.*

Proof. If at a time $t \in \mathbb{R}$, we have $w_i(t) = 0$ and thus $\dot{w}_i(t)$. Noticing that then the constant function $w_i(t) = 0 \ t \in \mathbb{R}$ is a solution of (1), by Cauchy-Lipschitz it is the only one. We can use the same argument to say that if there exists a time $t \in \mathbb{R}$, such we have $w_i(t) = 0$ then $w_i(t) = 0 \ \forall t \in \mathbb{R}$. Thus by continuity of $w_i(t)$ we have that if $w_i(0) \in (0, \frac{\sigma_i}{\lambda_i})$ then, $w_i(t) \in (0, \frac{\sigma_i}{\lambda_i})$, $\forall t \in \mathbb{R}$. $\qquad \square$

Case $L = 2$: in that case we have two situations, $\sigma_i > 0$ and $\sigma_i = 0$, $\lambda_i > 0$ (the case $\sigma_i = \lambda_i = 0$ give a constant functions).

For $\sigma_i > 0$ we have that,

$$t = \int_0^t \frac{dw_i(t)}{2w_i(t)(\sigma_i - \lambda_i w_i(t))} \tag{44}$$

$$= \frac{1}{2\sigma_i} \int_0^t \frac{dw_i(t}{w_i(t)} + \frac{\lambda_i dw_i(t}{\sigma_i - \lambda_i w_i(t)} \tag{45}$$

$$= \frac{1}{2\sigma_i} \ln \frac{w_i(t)(\sigma_i - \lambda_i w_i(0))}{w_i(0)(\sigma_i - \lambda_i w_i(t))}. \tag{46}$$

Leading to,

$$w_i(t) = \frac{w_i(0)\sigma_i e^{2\sigma_i t}}{w_i(0)\lambda_i(e^{2\sigma_i t} - 1) + \sigma_i}. \tag{47}$$

In order to get a solution for $w_{2,i}(t)$ and $w_{1,i}(t)$, we will use Cauchy-Lipschitz theorem [Coddington and Levinson, 1955]. The idea is that if we find a solution of (40), it is the only one. Let us set, $m_i(0) = n_i(0) = e^{-\delta_i}$, then we can set,

$$w_{2,i}(t) = w_{1,i}(t) = \sqrt{\frac{\sigma_i e^{2\sigma_i t - 2\delta_i}}{\lambda_i(e^{2\sigma_i t - 2\delta_i} - e^{-2\delta_i}) + \sigma_i}}, \tag{48}$$

and verify that we have,

$$\dot{w}_{2,i}(t) = w_{1,i}(t)(\sigma_i - \lambda_i w_{1,i}(t)w_{2,i}(t)), \ m_i(0) = e^{-\delta_i} \tag{49}$$

$$\dot{w}_{1,i}(t) = w_{2,i}(t)(\sigma_i - \lambda_i w_{1,i}(t)w_{2,i}(t)), \ n_i(0) = e^{-\delta_i} \quad 1 \le i \le r. \tag{50}$$

Thus, this is the unique solution of (40).

For $\sigma_i = 0$, $\lambda_i > 0$ we have that,

$$t = \int_0^t \frac{\dot{w}_i(t)}{-2\lambda_i w_i(t)^2} dt = \frac{1}{2\lambda_i}\left(\frac{1}{w_i(t)} - \frac{1}{w_i(0)}\right). \tag{51}$$

Thus,

$$w_i(t) = \frac{w_i(0)}{1 + 2w_i(0)\lambda_i t}. \tag{52}$$

Thus, if we initialize with $m_i(0) = n_i(0) = e^{-\delta_i}$ we get,

$$w_{1,i}(t) = w_{2,i}(t) = \frac{e^{-\delta_i}}{\sqrt{1 + 2e^{-\delta_i}\lambda_i t}}. \tag{53}$$

Non commutative case $\epsilon > 0$. Now, we will consider Assumption 1 with $\epsilon > 0$ and $L = 2$.

First let us proove two lemmas usefull for later,

Lemma 2. *The matrix valued function $\boldsymbol{W}(t)$ converge to $\boldsymbol{X}^\dagger \boldsymbol{Y}$ and thus is bounded for $t > 0$.*

Proof. Since (6) is a *gradient dynamics*, it only moves in the span of the gradient of f (the explicit expressions of ∇f is derived in (7) for $L = 1$ and (11) for a general L). We use this property to characterize the solution found by these dynamics. We can study each column of the predictors $\boldsymbol{W} := \boldsymbol{W}_1 \cdots \boldsymbol{W}_L$. If we look at the columns of $\nabla_{\boldsymbol{W}_L} f$, they are included in $\boldsymbol{X}^\top$, thus it means that if we initialize the columns of $\boldsymbol{W}_L^{(0)}$ in that span, then the columns of $\boldsymbol{W}$ will belong to that span during the whole learning process,

$$[\boldsymbol{W}]_i \in \text{span}(\nabla_{\boldsymbol{W}_L} f) \subset \text{span}(\boldsymbol{X}^\top), \ 1 \le i \le n, \tag{54}$$

where $\boldsymbol{W}$ is $\boldsymbol{W}(t)$. Thus, if the dynamics (6) converge, then they converge to a matrix with the ith column vector being in the intersection,

$$\text{span}(\boldsymbol{X}^\top) \cap \{\boldsymbol{u} : \boldsymbol{X}\boldsymbol{u} = [\boldsymbol{Y}]_i\} = \{\boldsymbol{X}^\dagger[\boldsymbol{Y}]_i\}. \tag{55}$$

Finally, we have $\boldsymbol{X}\boldsymbol{W}(t) \to \boldsymbol{Y}$ by definition of the gradient dynamics,

$$\frac{d\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_1(t)\boldsymbol{W}_2(t)\|^2}{dt} = -\|\nabla_{\boldsymbol{W}_1} f(\boldsymbol{W}_1(t), \boldsymbol{W}_2(t))\|^2 - \|\nabla_{\boldsymbol{W}_2} f(\boldsymbol{W}_1(t), \boldsymbol{W}_2(t))\|^2 < 0 \tag{56}$$

$\square$

Lemma 3. *We have that* $\|\boldsymbol{W}_1(t)\|^2 = O(t)$ *and* $\|\boldsymbol{W}_2(t)\|^2 = O(t)$.

Proof. We have that,

$$\frac{d\|\boldsymbol{W}_1(t)\|^2}{dt} = \langle \boldsymbol{W}_1(t), (\boldsymbol{\Sigma}_{xy} - \boldsymbol{\Sigma}_x \boldsymbol{W}(t)) \boldsymbol{W}_2(t)^\top \rangle = Tr((\boldsymbol{\Sigma}_{xy} - \boldsymbol{\Sigma}_x \boldsymbol{W}(t)) \boldsymbol{W}(t)^\top). \quad (57)$$

Since $\boldsymbol{W}(t)$ is bounded then $\|\boldsymbol{W}_1(t)\|^2 = O(t)$. The same way we have $\|\boldsymbol{W}_2(t)\|^2 = O(t)$ $\quad\square$

After the same change of basis as in the commutative case The matrices $\bar{\boldsymbol{W}}_l(t)$ follow the differencial equations

$$\frac{d\bar{\boldsymbol{W}}_1(t)}{dt} = \left(\boldsymbol{D}_{xy} - (\boldsymbol{D}_x + \boldsymbol{B})\bar{\boldsymbol{W}}(t)\right)\bar{\boldsymbol{W}}_2(t)^\top, \quad \frac{d\bar{\boldsymbol{W}}_2(t)}{dt} = \bar{\boldsymbol{W}}_1(t)^\top\left(\boldsymbol{D}_{xy} - (\boldsymbol{D}_x + \boldsymbol{B})\bar{\boldsymbol{W}}(t)\right) \quad (58)$$

In order to perform pertrubation analysis we will use a consequence of Grönwall's inequality [Gronwall, 1919].

Lemma 4. *Let* β *be a non negative function and* α *a non decreasing function. Let* u *be a function defined on an interval* $I = [a, \infty)$ *such that*

$$u(t) \leq \alpha(t) + \int_a^t \beta(s)u(s)ds, \quad \forall t \in I. \quad (59)$$

then we have that

$$u(t) \leq \alpha(t)\exp\left(\int_a^t \beta(s)ds\right), \quad \forall t \in I. \quad (60)$$

Proof. The proof can be found for instance in [Berglund, 2001, Lemma 3.1.6] $\quad\square$

Thus, let us consider $\bar{\boldsymbol{W}}_1(t)$ and $\bar{\boldsymbol{W}}_2(t)$ the solutions of (58) as well as $\bar{\boldsymbol{W}}_1^0(t)$ and $\bar{\boldsymbol{W}}_2^0(t)$ the solutions of the very same differential equation but with $\boldsymbol{B} = 0$. For notational simplicity we will omit the bar on the matrices $\boldsymbol{W}$ in the following. We have that,

$$\boldsymbol{W}_1(t) - \boldsymbol{W}_1^0(t) = \int_0^t [-\boldsymbol{B}\boldsymbol{W}(s)\boldsymbol{W}_2(s) + (\boldsymbol{D}_{xy} - \boldsymbol{D}_x\boldsymbol{W}(s))\boldsymbol{W}_2(s)^\top - (\boldsymbol{D}_{xy} - \boldsymbol{D}_x\boldsymbol{W}^0(s))\boldsymbol{W}_2^0(s)^\top]ds$$

Leading to

$$\|\boldsymbol{W}_1(t) - \boldsymbol{W}_1^0(t)\| \leq \int_0^t \|D_x\|\|\boldsymbol{W}_1(s)\boldsymbol{W}_2(s)\boldsymbol{W}_2(s)^\top - \boldsymbol{W}_1^0(s)\boldsymbol{W}_2^0(s)\boldsymbol{W}_2^0(s)^\top\|ds$$

$$+ \int_0^t \|D_xy\|\|\boldsymbol{W}_2(s) - \boldsymbol{W}_2^0(t)\|ds + \|\boldsymbol{B}\boldsymbol{W}(t)\boldsymbol{W}_2(t)\|$$

In order to upper bound the first integral we will consider the function $F(\boldsymbol{A}, \boldsymbol{B}) := \boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^\top$. This function is Lipschitz on any compact because this function is infinitely differentiable. Thus we have that (omiting the t in the notation),

$$\|\boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_2^\top - \boldsymbol{W}_1^0\boldsymbol{W}_2^0(\boldsymbol{W}_2^0)^\top\| = t^{3/2}\|\frac{\boldsymbol{W}_1}{\sqrt{t}}\frac{\boldsymbol{W}_2}{\sqrt{t}}\frac{\boldsymbol{W}_2}{\sqrt{t}}^\top - \frac{\boldsymbol{W}_1^0}{\sqrt{t}}\frac{\boldsymbol{W}_2^0}{\sqrt{t}}(\frac{\boldsymbol{W}_2^0}{\sqrt{t}})^\top\| \quad (61)$$

$$= t^{3/2}\|F(\frac{\boldsymbol{W}_1}{\sqrt{t}}, \frac{\boldsymbol{W}_2}{\sqrt{t}}) - F(\frac{\boldsymbol{W}_1^0}{\sqrt{t}}, \frac{\boldsymbol{W}_2^0}{\sqrt{t}})\|$$

$$\leq t^{3/2}L(\|\frac{\boldsymbol{W}_2}{\sqrt{t}} - \frac{\boldsymbol{W}_2^0}{\sqrt{t}}\| + \|\frac{\boldsymbol{W}_1}{\sqrt{t}} - \frac{\boldsymbol{W}_1^0}{\sqrt{t}}\|)$$

$$\leq tL(\|\boldsymbol{W}_2 - \boldsymbol{W}_2^0\| + \|\boldsymbol{W}_1 - \boldsymbol{W}_1^0\|)$$

Using the fact that $\frac{\boldsymbol{W}_1(t)}{\sqrt{t}}, \frac{\boldsymbol{W}_1^0(t)}{\sqrt{t}}, \frac{\boldsymbol{W}_2^0(t)}{\sqrt{t}}$ and $\frac{\boldsymbol{W}_2(t)}{\sqrt{t}}$, $t \geq 0$ live in a compact set (Lemma 3) and that F is Lipschitz on any compact. Thus, using that $\|S\| = O(\epsilon)$, we have

$$\|\boldsymbol{W}_1(t) - \boldsymbol{W}_1^0(t)\| \leq O(\epsilon) + O(1)\int_0^t s(\|\boldsymbol{W}_2(s) - \boldsymbol{W}_2^0(s)\| + \|\boldsymbol{W}_1(s) - \boldsymbol{W}_1(s)^0\|)ds$$

14

The same way we can prove that,

$$\|\boldsymbol{W}_2(t) - \boldsymbol{W}_2^0(t)\| \leq O(\epsilon) + O(1) \int_0^t s(\|\boldsymbol{W}_2(t) - \boldsymbol{W}_2^0(t)\| + \|\boldsymbol{W}_1(t) - \boldsymbol{W}_1^0(t)\|)dt$$

And consequently we can sum these two inequalities and apply Grönwall's inequality with the quantities $u(t) = \|\boldsymbol{W}_1(t) - \boldsymbol{W}_1^0(t)\| + \|\boldsymbol{W}_2(t) - \boldsymbol{W}_2^0(t)\|$, $\alpha(t) = O(\epsilon)$ and $\beta(s) = O(s)$ to get,

$$\|\boldsymbol{W}_1(t) - \boldsymbol{W}_1^0(t)\| + \|\boldsymbol{W}_2(t) - \boldsymbol{W}_2^0(t)\| \leq \epsilon \cdot e^{O(t^2)} \tag{62}$$

B.3 Proof of Thm. 2

Theorem' 2. *Let us denotes $w_i(t)$, the values defined in* (16). *If $m_i(0) = e^{-\delta}$, $1 \leq i \leq r$, then we have,*

$$w_i(\delta t) \underset{\delta \to \infty}{\to} \begin{cases} 0 & \text{if} \quad t < 1/\sigma_i \\ \sqrt{\frac{\sigma_i}{\lambda_i + \sigma_i}} & \text{if} \quad t = 1/\sigma_i \\ \sqrt{\frac{\sigma_i}{\lambda_i}} & \text{otherwise}. \end{cases}$$

Proof. Using (16) we get that,

$$w_i(\delta t) = \sqrt{\frac{\sigma_i e^{2\delta(\sigma_i t - 1)}}{\lambda_i (e^{2\delta(\sigma_i t - 1)} - e^{-2\delta}) + \sigma_i}}. \tag{63}$$

Then we can conclude saying that for any i and $t \geq 0$,

$$e^{2\delta(\sigma_i t - 1)} \underset{\delta \to \infty}{\to} \begin{cases} 0 & \text{if} \quad t < 1/\sigma_i \\ 1 & \text{if} \quad t = 1/\sigma_i \\ +\infty & \text{otherwise}, \end{cases} \tag{64}$$

and that when $\delta \to \infty$,

$$\|\boldsymbol{W}_i^0(\delta t) - \boldsymbol{W}_i^\epsilon(\delta t)\| \leq \epsilon \cdot e^{ct^2} = e^{\delta^2(ct^2 - \ln(\delta))} \to 0 \tag{65}$$

$\square$

B.4 Proof of Eq. 21

Let us recall (21),

$$\boldsymbol{W}_t = (\boldsymbol{W}_0 - \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy})(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)^t + \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy}. \tag{21}$$

Thus we have that,

$$\boldsymbol{W}_t = \boldsymbol{W}_0(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)^t + \eta\boldsymbol{\Sigma}_{xy} \sum_{s=0}^{t-1} (\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)^s \tag{66}$$

$$= \boldsymbol{W}_0(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)^t + \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy} - \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy}(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)^t$$

$$= (\boldsymbol{W}_0 - \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy})(\boldsymbol{I}_d - \eta\boldsymbol{\Sigma}_x)^t + \boldsymbol{\Sigma}_x^\dagger \boldsymbol{\Sigma}_{xy}. \tag{67}$$

B.5 Proof of Thm. 3

Case $\epsilon = 0$. If we define $\boldsymbol{W}^{(t)} := \boldsymbol{W}_1^{(t)} \boldsymbol{W}_2^{(t)}$, the discrete update scheme for the *two-layer linear neural network* (2) is,

$$\begin{cases} \boldsymbol{W}_1^{(t+1)} = \boldsymbol{W}_1^{(t)} - \eta(\boldsymbol{\Sigma}_x \boldsymbol{W}^{(t)} - \boldsymbol{\Sigma}_{xy})(\boldsymbol{W}_2^{(t)})^\top \\ \boldsymbol{W}_2^{(t+1)} = \boldsymbol{W}_2^{(t)} - \eta(\boldsymbol{W}_1^{(t)})^\top(\boldsymbol{\Sigma}_x \boldsymbol{W}^{(t)} - \boldsymbol{\Sigma}_{xy}). \end{cases} \tag{68}$$

Using the same transformation (37) as in §B.2 we get that,

$$\begin{cases} \bar{\boldsymbol{W}}_1^{(t+1)} = \bar{\boldsymbol{W}}_1^{(t)} - \eta(\boldsymbol{D}\bar{\boldsymbol{W}}^{(t)} - \boldsymbol{S})(\bar{\boldsymbol{W}}_2^{(t)})^\top \\ \bar{\boldsymbol{W}}_2^{(t+1)} = \bar{\boldsymbol{W}}_2^{(t)} - \eta(\bar{\boldsymbol{W}}_1^{(t)})^\top(\boldsymbol{D}\bar{\boldsymbol{W}}^{(t)} - \boldsymbol{S}). \end{cases} \tag{69}$$

where D and S only have diagonal coefficients. Thus, by an immediate recurrence we can show that if $\bar{W}_1^{(0)}$ and $\bar{W}_2^{(0)}$ only have diagonal coefficients then $\bar{W}_1^{(t)}$ and $\bar{W}_2^{(t)}$ for $t \in \mathbb{N}$ only have diagonal coefficients.

let us note, $r = \min(p,d)$ and $m_1^{(t)}, \ldots, m_r^{(t)}$ and $n_1^{(t)}, \ldots, n_r^{(t)}$ the respective diagonal coefficients of $\bar{W}_1^{(t)}$ and $\bar{W}_2^{(t)}$, they follow the equation,

$$m_i^{(t+1)} = m_i^{(t)} + \eta n_i^{(t)}(\sigma_i - \lambda_i n_i^{(t)} m_i^{(t)}) \quad \text{and} \quad n_i^{(t+1)} = n_i^{(t)} + \eta m_i^{(t)}(\sigma_i - \lambda_i n_i^{(t)} m_i^{(t)}), \quad 1 \leq i \leq r. \tag{70}$$

In order to prove Thm. 3, we will prove several properties on the sequences $(m_i^{(t)})_{t \geq 0}$ and $(n_i^{(t)})_{t \geq 0}$, $1 \leq i \leq d$. First let us introduce the sequence $(a_i^{(t)})_{t \geq 0}$ defined as $a_i^{(t)} := m_i^{(t)} n_i^{(t)}, t \geq 0$.

Lemma 5. *If $m_i^{(0)} = n_i^{(0)}$ then,*

$$m_i^{(t)} = n_i^{(t)} = \sqrt{a_i^{(t)}} \qquad \forall t \in \mathbb{N}. \tag{71}$$

Proof. By a straightforward recurrence we have that if at time t, $m_i^{(t)} = n_i^{(t)}$ then by (70), we have $m_i^{(t+1)} = n_i^{(t+1)}$. $\qquad \square$

Thus, we will now focus on the sequence $(a_i^{(t)})_{t \geq 0}$, by (70), we have that

$$a_i^{(t+1)} = a_i^{(t)} + 2\eta a_i^{(t)}(\sigma_i - \lambda_i a_i^{(t)}) + \eta^2 a_i^{(t)}(\sigma_i - \lambda_i a_i^{(t)})^2 = a_i^{(t)} + \eta a_i^{(t)}(\sigma_i - \lambda_i a_i^{(t)})(2 + \eta(\sigma_i - \lambda_i a_i^{(t)})). \tag{72}$$

Similarly as for the continuous case, there is two different behavior $\sigma_i > 0$ ad $\sigma_i = 0$, $\lambda_i > 0$. In the following we assume that $\eta > 0$.

For $\sigma_i > 0$ we can derive the following results,

Lemma 6. *For any $1 \leq i \leq r_{xy}$, if $0 < a_i^{(0)} < \frac{\sigma_i}{\lambda_i}$ and $2\eta\sigma_i < 1$, then, the sequence $(a_i^{(t)})$ is increasing and*

$$0 < a_i^{(t)} < \frac{\sigma_i}{\lambda_i}, \quad \forall t \geq 0. \tag{73}$$

Proof. By assumption (73) is true for $t = 0$.

Let us assume that (73) is true for a time-step t and let us prove that it is still true at time-step $t+1$.

Using the recursive definition (72) of $a_i^{(t)}$ we get for $t \geq 0$,

$$a_i^{(t+1)} = a_i^{(t)} + \eta a_i^{(t)}(\sigma_i - \lambda_i a_i^{(t)})(2 + \eta(\sigma_i - \lambda_i a_i^{(t)})) \tag{74}$$

$$> a_i^{(t)} > 0. \tag{75}$$

For the upper bound we need to notice that $a_i^{(t+1)} = f_i(a_i^{(t)})$ where $f_i : x \mapsto x + \eta x(\sigma_i - \lambda_i x)(2 + \eta(\sigma_i - \lambda_i x))$ where $\eta > 0$. Since we assumed that $2\eta\sigma_i < 1$, we have that,

$$f_i(x) < x + \eta x(\sigma_i - \lambda_i x)(2 + \frac{\sigma_i - \lambda_i x}{2\sigma_i}), \quad \forall x \in (0,1) \tag{76}$$

$$< x + \frac{x(1 - \frac{\lambda_i}{\sigma_i}x)(\frac{5}{2} - \frac{\lambda_i}{2\sigma_i}x)}{2} =: g_i(x), \quad \forall x \in (0,1). \tag{77}$$

$$\tag{78}$$

Then we just need to show that $g(x) < \frac{\lambda_i}{\sigma_i}$, $\forall x \in (0, \frac{\lambda_i}{\sigma_i})$.

$$4g'(x) = 9 - 12\frac{\lambda_i}{\sigma_i}x + 3\frac{\lambda_i^2}{\sigma_i^2}x^2 > 0, \quad \forall x \in (0, \frac{\lambda_i}{\sigma_i}). \tag{79}$$

Thus g is non-decreasing on $(0,1)$ and consequently, $g(x) < g(\frac{\lambda_i}{\sigma_i}) = \frac{\lambda_i}{\sigma_i}$, $\forall x \in (0,1)$.

Finally, we get that,

$$a_i^{(t+1)} = f_i(a_i^{(t)}) < g(a_i^{(t)}) < g(\tfrac{\lambda_i}{\sigma_i}) = \tfrac{\lambda_i}{\sigma_i} . \tag{80}$$

$\square$

With this lemma we can proof Thm. 3. Let us first recall this theorem.

Theorem' 3. *For any* $1 \leq i \leq r_{xy}$, *if* $\tfrac{\sigma_i}{\lambda_i} > a_i^{(0)} > 0$ *and* $2\eta\sigma_i < 1$, *then* $\forall t \geq 0$, $1 \leq i \leq r$ *we have,*

$$a_i^{(t)} \leq \frac{a_i^{(0)}}{(\sigma_i - \lambda_i a_i^{(0)})e^{(-2\eta\sigma_i - \eta^2\sigma_i^2)t} + a_i^{(0)}\lambda_i} \tag{81}$$

$$\text{and} \quad a_i^{(t)} \geq \frac{a_i^{(0)}}{(\sigma_i - \lambda_i a_i^{(0)})e^{(-2\eta\sigma_i + 4\eta^2\sigma_i^2)t} + a_i^{(0)}\lambda_i} , \tag{82}$$

and for $r_{xy} \leq i \leq r_x$,

$$a_i^{(t)} \leq \frac{a_i^{(0)}}{1 + a_i^{(0)}\lambda_i\eta t} .$$

Proof. In this proof for notational compactness we will remove the index i.

We first prove (81), we work with $1/a^{(t+1)} - \tfrac{\lambda}{\sigma}$, Using (72) we get,

$$1/a^{(t+1)} - \frac{\lambda}{\sigma} = \frac{1}{a^{(t)}} \left(\frac{1}{1 + 2\eta\sigma(1 - \tfrac{\lambda}{\sigma}a^{(t)}) + \eta^2\sigma^2(1 - \tfrac{\lambda}{\sigma}a^{(t)})^2} \right) - \frac{\lambda}{\sigma} \tag{83}$$

$$\geq \frac{1}{a^{(t)}} \left(\frac{1}{1 + (2\eta\sigma + \eta^2\sigma^2)(1 - \tfrac{\lambda}{\sigma}a^{(t)})} \right) - \frac{\lambda}{\sigma} \tag{84}$$

$$\geq \frac{1}{a^{(t)}} - \frac{\lambda}{\sigma} - \frac{2\eta\sigma + \eta^2\sigma^2}{a^{(t)}}(1 - \tfrac{\lambda}{\sigma}a^{(t)}) , \tag{85}$$

where we used that $\frac{1}{1+x} \geq 1 - x$, $\forall x \geq 0$. Thus we have,

$$1/a^{(t)} - \frac{\lambda}{\sigma} \geq (\frac{1}{a^{(t-1)}} - \frac{\lambda}{\sigma})(1 - 2\eta\sigma - \eta^2\sigma^2) \tag{86}$$

$$\geq (\frac{1}{a^{(0)}} - \frac{\lambda}{\sigma})(1 - 2\eta\sigma - \eta^2\sigma^2)^t \tag{87}$$

$$\geq (\frac{1}{a^{(0)}} - \frac{\lambda}{\sigma})e^{(-2\eta\sigma - \eta^2\sigma^2)t} . \tag{88}$$

Thus Leads to,

$$a^{(t)} \leq \frac{\sigma a^{(0)}}{(\sigma - \lambda a^{(0)})e^{t(-2\eta\sigma - \eta^2\sigma^2)} + a^{(0)}\lambda} . \tag{89}$$

To prove (81) we will once again work with $1/a^{(t+1)} - \tfrac{\lambda}{\sigma}$. Using (72) we get

$$1/a^{(t+1)} - \frac{\lambda}{\sigma} = \frac{1}{a^{(t)}} \left(\frac{1}{1 + 2\sigma(1 - \tfrac{\lambda}{\sigma}a^{(t)}) + \sigma^2(1 - \tfrac{\lambda}{\sigma}a^{(t)})^2} \right) - \frac{\lambda}{\sigma} \tag{90}$$

$$\leq \frac{1}{a^{(t)}} \left(\frac{1}{1 + 2\sigma(1 - \tfrac{\lambda}{\sigma}a^{(t)})} \right) - \frac{\lambda}{\sigma} \tag{91}$$

$$\leq \frac{1}{a^{(t)}} - \frac{\lambda}{\sigma} - \frac{2\eta\sigma}{a^{(t)}}(1 - \tfrac{\lambda}{\sigma}a^{(t)}) + \frac{4\eta^2\sigma^2}{a^{(t)}}(1 - \tfrac{\lambda}{\sigma}a^{(t)})^2 , \tag{92}$$

17

where we used that $\frac{1}{1+x} \leq 1 - x + x^2$, $\forall x \geq 0$. Thus we have,

$$1/a^{(t)} - \frac{\lambda}{\sigma} \leq (\frac{1}{a^{(t)}} - \frac{\lambda}{\sigma})(1 - 2\sigma\eta + 4\eta^2\sigma^2) \tag{93}$$

$$\leq (\frac{1}{a^{(0)}} - \frac{\lambda}{\sigma})(1 - 2\sigma\eta + 4\eta^2\sigma^2)^t \tag{94}$$

$$\leq (\frac{1}{a^{(0)}} - \frac{\lambda}{\sigma})e^{(-2\sigma\eta + 4\eta^2\sigma^2)t}. \tag{95}$$

This leads to,

$$a^{(t)} \geq \frac{\sigma a^{(0)}}{(\sigma - \lambda a^{(0)})e^{t(-2\sigma + 4\eta^2\sigma^2)} + \lambda a^{(0)}}. \tag{96}$$

Now for $\sigma = 0$ and $\lambda > 0$ we have that,

$$a_i^{(t+1)} = a_i^{(t)}(1 - 2\lambda\eta a_i^{(t)} + \lambda^2\eta^2(a_i^{(t)})^2). \tag{97}$$

Thus, considering $(a_i^{(t)})^{-1}$ we get

$$1/a_i^{(t+1)} = 1/a_i^{(t)}(1 - 2\lambda\eta a_i^{(t)} + \lambda^2\eta^2(a_i^{(t)})^2)^{-1} \tag{98}$$

$$\geq 1/a_i^{(t)}(1 + 2\lambda\eta a_i^{(t)} - \lambda^2\eta^2(a_i^{(t)})^2) \tag{99}$$

$$= 1/a_i^{(t)} + 2\lambda\eta - \lambda^2\eta^2 a_i^{(t)}. \tag{100}$$

Thus, if we assume that $1/a_i^{(0)} \geq \lambda\eta$ we have that $(1/a_i^{(t)})$ is a increasing sequence and that,

$$1/a_i^{(t)} \geq 1/a_i^{(t-1)} + \lambda\eta \geq 1/a_i^{(0)} + \lambda\eta t, \tag{101}$$

leading to,

$$a_i^{(t)} \leq \frac{a_i^{(0)}}{1 + a_i^{(0)}\lambda\eta t}. \tag{102}$$

$\square$

Case $\epsilon > 0$. If we are able to show that all the sequences $W_1^{(t)}$ and $W_2^{(t)}$ are bounded

From this theorem we can deduce the following corollary,

Proof of Corollary 3 Let us recall Corollary 3.

Corollary' 3. If $\eta < \frac{1}{2\sigma_1}$, $\eta < 2\frac{\sigma_i - \sigma_{i+1}}{\sigma_i^2}$ and $\eta < \frac{\sigma_i - \sigma_{i+1}}{2\sigma_{i+1}^2}$, $\forall i\ 1 \leq i \leq r_{xy} - 1$ then for $1 \leq i < r_x$,

$$a_i^{(\delta T_j)} \xrightarrow[\delta \to \infty]{} \begin{cases} 0 & \text{if}\quad i > r_{xy}\quad \text{or}\quad j < i \\ \frac{\sigma_i}{\lambda_i} & \text{if}\quad i \leq r_{xy}\quad \text{and}\quad j > i, \end{cases}$$

where $T_j := \frac{1}{\sigma_j\eta}$, $1 \leq j \leq r_{xy}$ and $T_j = +\infty$ if $j > r_{xy}$ and $T_0 = 0$.

Proof. First let us notice that since $\sigma_1 > \ldots > \sigma_{r_{xy}} > 0$, the assumption $\eta < 1/(2\sigma_1)$ implies $\eta < 1/(2\sigma_i)$, $1 \leq i \leq d$.

Let $i \leq r_{xy}$. Let us first prove that if $j < i$, then $a_i^{(T_j)} \xrightarrow[\delta \to \infty]{} 0$.

Using (81) and recalling that in Thm. 3, we assume that $a_i^{(0)} = e^{-2\delta}$, we have for $1 \leq j < i$,

$$0 < a_i^{(\delta T_j)} \leq a_i^{(\delta T_{i-1})} < \frac{\sigma_i}{(\sigma_i e^{2\delta} - \lambda_i)e^{\delta(-2\sigma_i/\sigma_{i-1} - \eta\sigma_i^2/\sigma_{i-1})} + \sigma_i} \xrightarrow[\delta \to \infty]{} 0. \tag{103}$$

We have $(2 + \eta\sigma_i)\sigma_i/\sigma_{i-1} < 2$, because we assumed that $\eta < 2\frac{\sigma_i - \sigma_{i+1}}{\sigma_i^2}$, $\forall i\ 1 \leq i \leq d$. Note that for $i = r_{xy}$ we have $a_i^{(\delta T_{i+1})} = 0$, $\forall \delta > 0$.

18

Let us now prove that if $j > i$, then $a_i^{(\delta T_j)} \underset{\delta \to \infty}{\to} \frac{\sigma_i}{\lambda_i}$.

Using (82) and recalling that in Thm. 3, we assume that $a_i^{(0)} = e^{-2\delta}$, we have for $j > i \geq 1$,

$$\frac{\lambda_i}{\sigma_i} > a_i^{(\delta T_j)} \geq a_i^{(\delta T_{i+1})} > \frac{\sigma_i}{(\sigma_i e^{2\delta} - \lambda_i) e^{\delta(-2\sigma_i/\sigma_{i+1} + 4\eta\sigma_i^2/\sigma_{i+1})} + \sigma_i} \underset{\delta \to \infty}{\to} \frac{\sigma_i}{\lambda_i} . \tag{104}$$

where we have that $(e^{2\delta} - 1)e^{\delta(-2\sigma_i/\sigma_j + 4\eta\sigma_i^2/\sigma_j)} \to 0$ because

$$1 - \sigma_i/\sigma_{i+1} + 2\eta\sigma_i^2/\sigma_{i+1} < 0 \Leftrightarrow \eta < (\sigma_i - \sigma_{i+1})/(2\sigma_i^2) .$$

Now for $i \geq r_{xy} + 1$ we just need to use, (83) to get,

$$a_i^{(t)} \leq \frac{a_i^{(0)}}{1 + \lambda\eta a_i^{(0)} t} \underset{\delta \to \infty}{\to} 0 . \tag{105}$$