Almost Global Convergence to Global Minima for Gradient Descent in Deep Linear Networks

Zhenyu Liao, Yacine Chitour, Romain Couillet
Laboratoire des Signaux et Systèmes
CentraleSupélec, Université Paris-Saclay
Gif-sur-Yvette, France

Abstract

In this article we prove that, for almost all training data-target pairs, a linear deep neural network exhibits convergence for almost all initializations to a global minimum when using the classical gradient descent algorithm. This global result is obtained through an original geometric framework relying on a key invariance property induced by the network structure and providing, as a fundamental side result, a clearer picture of the loss landscape. We further argue that the presented framework is sufficiently powerful to envision extensions to nonlinear deep networks.

1 Introduction

Despite the rapid growing list of successful applications of deep neural networks trained with back-propagation in various fields from computer vision [15] to speech recognition [19] and natural language processing [7], our theoretical understanding on these elaborate systems, however, is developing at a more modest pace.

One of the major difficulties in the design of deep neural networks today is that, to obtain networks with greater expressive power, we cascade more and more layers to make them "deeper" and hope to extract more "abstract" features from the (numerous) training data so as to improve the networks in terms of generalization performance. Nonetheless, from an optimization viewpoint, this "deeper" structure poses problems because it gives rise to non-convex loss functions and makes the optimization seemingly intractable. In general finding a global minimum of a *generic* non-convex function is an NP-complete problem [20] and it is unfortunately the case for neural networks as it was shown in [3] that even training a very simple network is indeed NP-complete.

Yet, many non-convex problems such as phase retrieval, independent component analysis and orthogonal tensor decomposition are known to obey the important properties [23] that 1) all local minima are also global; and 2) around any saddle point the objective function has a negative directional curvature (i.e., the possibility to continue to descend) and thus allow for the possibility to find some way to fall into a "basin" with a (comparably) low loss "with high probability". In this regard, the loss surfaces of deep neural networks are receiving an unprecedented research interest: in the pioneering work of Baldi & Hornik [2] the landscape of mean square losses was studied in the case of linear auto-encoders (i.e., the same dimension for input data and output targets) of depth one; more recently in the work of Saxe et al. [22] the dynamics of the corresponding gradient descent system was first studied, by assuming the input data $\mathbf{X}$ empirical correlation matrix $\mathbf{X}\mathbf{X}^\mathsf{T}$ to be identity, in a linear deep neural networks, so as to propose a novel initialization method. Then in [13] the author proved that under some appropriate rank condition on the (cascading) matrix product, all critical points of a deep linear neural networks are either global minima or saddle points with Hessian admitting eigenvalues with different signs, meaning that linear deep networks are somehow "close" to those examples mentioned at the beginning of this paragraph. Nonetheless, the results in [22, 13] are incomplete in the sense that they do not provide enough (global) information regarding when and how can gradient descent trajectories result in these global minima (recall that, to escape from saddle points within a reasonable time one may alternatively use second-order methods with information from the Hessian, artificially perturb the gradient with noise as in [12], etc.). Concretely speaking, previous analyses in [13, 16] only focus on the *local* behavior of each critical point and a "global picture" on the *whole* space occupied by the network weights is still in demand.

In this paper, we elaborate on the model from [22, 13] and evaluate the dynamics of the associated gradient system in a "continuous" manner. We prove that, for almost every choice of training data-target pair $(\mathbf{X}, \mathbf{Y})$ and almost every initialization for the weight matrices $\mathbf{W}_i$, the corresponding trajectory of the gradient system converges to a global minimum of the loss function. Based on a cornerstone "invariant" in the parameter space induced by the network cascading structure, we establish a generic framework for the geometric understanding of deep neural networks and provide the aforementioned sought-for global picture of the gradient descent dynamics in the specific case of linear networks. Due to space limitation, only proof sketches are provided in the $H = 1$ layer scenario, along with generalization intuitions to the $H > 1$ case. The complete derivations for the case $H \geq 1$ are available in an extended version.

2 System Model and Main Result

2.1 Problem setup

We start with a deep linear neural network with H hidden layers as illustrated in Figure 1. To begin with, the network structure as well as associated notations are presented as follows.

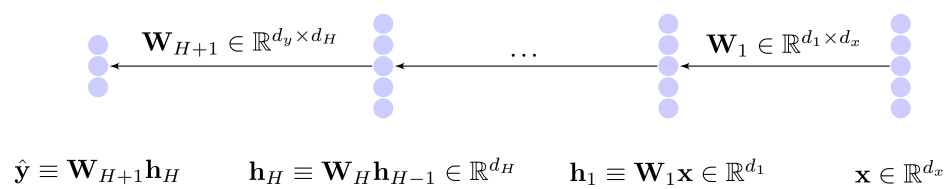

Figure 1: Illustration of the H-hidden-layer linear neural network

Let the pair $(\mathbf{X}, \mathbf{Y})$ denote the training data and associated targets, with $\mathbf{X} \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{d_x \times m}$ and $\mathbf{Y} \equiv [\mathbf{y}_1, \ldots, \mathbf{y}_m] \in \mathbb{R}^{d_y \times m}$, where m denotes the number of instances in the training set and d_x, d_y the dimensions of data and targets, respectively. We denote the weight matrix $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ that connects $\mathbf{h}_{i-1}$ to $\mathbf{h}_i$ for $i = 1, \ldots H+1$ and set $\mathbf{h}_0 = \mathbf{x}$, $\mathbf{h}_{H+1} = \hat{\mathbf{y}}$ as in Figure 1. The network output is thus given by $\hat{\mathbf{Y}} = \mathbf{W}_{H+1} \ldots \mathbf{W}_1 \mathbf{X}$. We denote $\mathbf{W}$ the $(H+1)$-tuple of $(\mathbf{W}_1, \ldots, \mathbf{W}_{H+1})$ for simplicity and work on the mean squared error $\mathcal{L}(\mathbf{W})$ given by the Frobenius norm below,

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2}\|\mathbf{Y} - \hat{\mathbf{Y}}\|_F^2 = \frac{1}{2}\|\mathbf{Y} - \mathbf{W}_{H+1} \ldots \mathbf{W}_1 \mathbf{X}\|_F^2 \tag{1}$$

under the following assumptions:

Assumption 1 (Dimension Condition)**.** $m \geq d_x \geq \max(d_1, \ldots, d_H) \geq \min(d_1, \ldots, d_H) > d_y$. *In particular in the case $H = 1$ this condition yields $m \geq d_x \geq d_1 > d_y$.*

Assumption 2 (Full Rank Data and Targets)**.** *The matrices $\mathbf{X}$ and $\mathbf{Y}$ are of full (row) rank, i.e., of rank d_x and d_y, respectively, accordingly with Assumption 1.*

Assumption 1 and 2 on the dimension and rank of the training data are realistic and practically easy to satisfy, as discussed in previous works [2, 13].[1]

Under Assumptions 1 and 2, with the singular value decomposition on $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$ with $\mathbf{V}_\mathbf{X} = \left[\begin{array}{c|c} \mathbf{V}_\mathbf{X}^1 & \mathbf{V}_\mathbf{X}^2 \end{array}\right]$, $\mathbf{V}_\mathbf{X}^1 \in \mathbb{R}^{m \times d_x}$ and then on $\mathbf{Y}\mathbf{V}_\mathbf{X}^1 \equiv \overline{\mathbf{Y}} = \mathbf{U}_\mathbf{Y}\mathbf{\Sigma}_\mathbf{Y}\mathbf{V}_\mathbf{Y}^\mathsf{T}$, together with an immediate change of variable, we get $\mathcal{L}(\mathbf{W}) = L(\mathbf{W}) + \frac{1}{2}\|\mathbf{Y}\mathbf{V}_\mathbf{X}^2\|_F^2$ with

$$L(\mathbf{W}) \equiv \frac{1}{2}\|\mathbf{\Sigma}_\mathbf{Y} - \overline{\mathbf{W}}_{H+1}\overline{\mathbf{W}}_H \ldots \overline{\mathbf{W}}_2\overline{\mathbf{W}}_1\|_F^2 \tag{2}$$

[1]Assumption 1 is demanded here for convenience and our results can be extended to handle more elaborate dimension settings. Similarly, when the training data is rank deficient, the learning problem can be reduced to a lower-dimensional one by removing these non-informative data in such a way that Assumption 2 holds.

where $\mathbf{\Sigma_X} \equiv [\ \mathbf{S_X} \mid \mathbf{0}\ ] \in \mathbb{R}^{d_x \times m}$, $\mathbf{\Sigma_Y} \in \mathbb{R}^{d_y \times d_x}$ and we denote $\overline{\mathbf{W}}_{H+1} \equiv \mathbf{U}_\mathbf{Y}^\mathsf{T} \mathbf{W}_{H+1} \in \mathbb{R}^{d_y \times d_H}$, $\overline{\mathbf{W}}_1 \equiv \mathbf{W}_1 \mathbf{U_X S_X V_Y} \in \mathbb{R}^{d_1 \times d_x}$ and $\overline{\mathbf{W}}_i = \mathbf{W}_i$ for $i = 2, \ldots, H$. Therefore the state space[2] of $\mathbf{\Xi} \equiv (\overline{\mathbf{W}}_{H+1}, \ldots, \overline{\mathbf{W}}_1)$ is equal to $\mathcal{X} = \mathbb{R}^{d_y \times d_H} \times \ldots \times \mathbb{R}^{d_1 \times d_x}$. In particular, for $H = 1$ we have $d_H = d_1$ and $\mathcal{X}$ has dimension $d_1(d_x + d_y)$.

With the above notations, we demand in addition the following assumption on the target $\overline{\mathbf{Y}}$.

Assumption 3 (Distinct Singular Values)**.** *The target $\overline{\mathbf{Y}}$ has d_y distinct singular values.*

Although seemingly restrictive, Assumption 3 actually holds for an open and dense subset of $\mathbb{R}^{d_y \times d_x}$.

The objective of this article is to study the gradient descent [5] dynamics (GDD) defined as

Definition 1 (GDD)**.** *The Gradient Descent Dynamic of L is the dynamical system defined on $\mathcal{X}$ by*

$$\frac{d\mathbf{\Xi}}{dt} = -\nabla_{\mathbf{\Xi}} L(\mathbf{\Xi}) \tag{3}$$

where $\nabla_{\mathbf{\Xi}} L(\mathbf{\Xi})$ denotes the gradient of the loss function L with respect to $\mathbf{\Xi}$. A point $\mathbf{\Xi} \in \mathcal{X}$ is a critical point of L if and only if $\nabla_{\mathbf{\Xi}} L(\mathbf{\Xi}) = \mathbf{0}$ and we denote $\operatorname{Crit}(L)$ the set of critical points.

In the following, we work directly on the equivalent equation (2) and start by evaluating the gradient of L. With the previous notations, for $\boldsymbol{\xi} \equiv (\overline{\mathbf{w}}_{H+1}, \ldots, \overline{\mathbf{w}}_1)$, we expand the variation of $L(\mathbf{\Xi} + \boldsymbol{\xi})$ as

$$L(\mathbf{\Xi} + \boldsymbol{\xi}) = L(\mathbf{\Xi}) + \Delta_{\mathbf{\Xi}}(\boldsymbol{\xi}) + O\left(\|\boldsymbol{\xi}\|^2\right)$$

with $\overline{\mathbf{M}} \equiv \mathbf{\Sigma_Y} - \overline{\mathbf{W}}_{H+1} \ldots \overline{\mathbf{W}}_1$, $L(\mathbf{\Xi}) = \frac{1}{2}\|\overline{\mathbf{M}}\|_F^2$ and the differential $\Delta_{\mathbf{\Xi}}(\boldsymbol{\xi})$ given by $\Delta_{\mathbf{\Xi}}(\boldsymbol{\xi}) \equiv -\sum_{j=1}^{H+1} \operatorname{tr}\left(\overline{\mathbf{M}}^\mathsf{T} \overline{\mathbf{W}}_{H+1} \ldots \overline{\mathbf{W}}_{j+1} \overline{\mathbf{w}}_j \overline{\mathbf{W}}_{j-1} \ldots \overline{\mathbf{W}}_1\right)$. We thus derive from Definition 1 the dynamics of L, for $j = 1, \ldots, H+1$, as

$$\frac{d\overline{\mathbf{W}}_j}{dt} \equiv -\nabla_{\overline{\mathbf{W}}_j} L(\mathbf{\Xi}) = \left(\overline{\mathbf{W}}_{H+1} \ldots \overline{\mathbf{W}}_{j+1}\right)^\mathsf{T} \overline{\mathbf{M}} \left(\overline{\mathbf{W}}_{j-1} \ldots \overline{\mathbf{W}}_1\right)^\mathsf{T}. \tag{4}$$

We first remark the following interesting (and crucial to what follows) property of the gradient system (4), inspired by [22] which essentially considered the case where all dimensions are equal to one.

Lemma 1 (Invariant in GDD)**.** *Consider any trajectory of the gradient system given by* (4)*. Then, for $j = 1, \ldots, H$, the value of $\overline{\mathbf{W}}_{j+1}^\mathsf{T} \overline{\mathbf{W}}_{j+1} - \overline{\mathbf{W}}_j \overline{\mathbf{W}}_j^\mathsf{T}$ remains constant, i.e.,*

$$\overline{\mathbf{W}}_{j+1}^\mathsf{T}(t) \overline{\mathbf{W}}_{j+1}(t) - \overline{\mathbf{W}}_j(t) \overline{\mathbf{W}}_j^\mathsf{T}(t) = \mathbf{C}_j^0 \equiv (\overline{\mathbf{W}}_{j+1}^\mathsf{T} \overline{\mathbf{W}}_{j+1} - \overline{\mathbf{W}}_j \overline{\mathbf{W}}_j^\mathsf{T})\big|_{t=0}, \quad \forall t \geq 0.$$

In particular, in the case of $H = 1$ we get $\overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2 - \overline{\mathbf{W}}_1 \overline{\mathbf{W}}_1^\mathsf{T} = \mathbf{C}^0 \equiv (\overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2 - \overline{\mathbf{W}}_1 \overline{\mathbf{W}}_1^\mathsf{T})\big|_{t=0}$.

Proof. Simply check that $\frac{d}{dt}\left(\overline{\mathbf{W}}_{j+1}^\mathsf{T} \overline{\mathbf{W}}_{j+1} - \overline{\mathbf{W}}_j \overline{\mathbf{W}}_j^\mathsf{T}\right) = 0$. □

Lemma 1 provides a key structural property of the GDD that is instrumental to ensure the boundedness of the gradient descent trajectories and thus in turn to prove the convergence to global minima.

2.2 Main Results

Our main result is Theorem 1 which provides information on the convergence of the GDD trajectories to global minima under reasonable conditions. This result is obtained as follows: we first prove that every trajectory of the GDD converges to a critical point of the loss function L. Thanks to this crucial fact, it remains to show that the union of the basins of attraction over all the critical points which are not global minima is "small", in the sense that this union is contained in a subset of the state space $\mathcal{X}$ that is of positive codimension. To this end, we perform a precise characterization and classification of the critical points, followed by a *local* analysis of (4) around *each* critical point.

Concretely, we focus on the state space $\mathcal{X}$ and evaluate "how much" is occupied by the saddle points (it turns out that every local minimum is global and there is no local maximum [13]): we stratify

[2]The network (weight) parameters $\mathbf{\Xi}$ evolve through time and are considered to be *state variables* of the dynamical system, while the pair $(\mathbf{X}, \mathbf{Y})$ is fixed and thus referred as the "parameters" of the given system.

the set of critical points $\mathrm{Crit}(L)$ in d_y+1 subsets, one of them corresponding to the set of global minima and the d_y others, $\mathrm{Crit}_r(L)$ with $r=0,\ldots,d_y-1$, corresponding to the set of saddle points. Then, we perform the aforementioned fine study on the local behavior of gradient descent trajectories "around" each saddle point, so as to measure its basin of attraction (i.e., the set of initializations for which the GDD trajectories converge to that given saddle point). Precisely speaking, we determine, for a given $0 \leq r \leq d_y-1$, an upper bound D_S^r on the dimension of a differentiable manifold $\mathcal{M}(\boldsymbol{\Xi})$ containing the basin of attraction of each saddle point $\boldsymbol{\Xi} \in \mathrm{Crit}_r(L)$ (note that D_S^r does not depend on $\boldsymbol{\Xi} \in \mathrm{Crit}_r(L)$). To ensure *global* convergence from the *local* analysis, we further show that the sum of the dimension of $\mathrm{Crit}_r(L)$ (which is defined since the latter is an algebraic variety) and D_S^r is smaller than the dimension of the state space minus two. This allows us to reach the conclusion of almost global convergence to global minima by means of transversality arguments [10].

We start with the global convergence to critical points of all gradient descent trajectories. While one expects the gradient descent algorithm to converge to critical points, this may not always be the case. Two possible (undesirable) situations are 1) a trajectory is unbounded or 2) it oscillates "around" several critical points without convergence, i.e., along an ω-limit set made of a continuum of critical points (see [24] for notions on ω-limit sets). The property of an iterative algorithm (like gradient descent) to converge to a critical point for any initialization is referred to as "global convergence" [25]. However, it is very important to stress the fact that it does not imply (contrary to what the name might suggest) convergence to a global (or good) minimum for all initializations.

To answer the convergence question, we resort to Lojasiewicz's theorem[3] for the convergence of a gradient descent flow of the type of (4) with real analytic right-hand side [17]. Since the loss function $L(\boldsymbol{\Xi})$ is a polynomial of degree $(H+1)^2$ in the components of $\boldsymbol{\Xi}$, Lojasiewicz's theorem ensures that if a given trajectory of the gradient descent flow is bounded (i.e., it remains in a compact set for every $t \geq 0$) it must converge to a critical point with a guaranteed rate of convergence. In particular, the previous phenomenon of "oscillation" cannot occur and we are left to ensure the absence of unbounded trajectories. Lemma 1 is the core argument to show that all trajectories of the GDD are indeed bounded, leading to the first result of this article as follows.

Proposition 1 (Global Convergence of GDD to Critical Points)**.** *Let* $(\mathbf{X}, \mathbf{Y})$ *be a data-target pair satisfying Assumptions 1 and 2. Then, every trajectory of the corresponding gradient flow* (4) *converges to a critical point as* $t \to \infty$*, at rate at least* $t^{-\alpha}$*, for some fixed* $\alpha > 0$ *only depending on the dimensions of the problem.*

A first consequence of Proposition 1 is that it provides a rigorous justification for the appropriate discretization of the GDD given in (3). Indeed the step size can be chosen in terms of an a priori bound for the whole trajectory, which is explicitly determined only with the initial condition (see [4]). This is in contrast with [16] in which the discretization step size of the GDD is determined with a bound on the Hessian norm of a critical point, however, the latter assumption supposes that the trajectory converges to a "well-known" critical point (with a prior information on the Hessian for example) while no such information is available at the initial stage of training.

To provide guarantees of global convergence to a "good" critical point, we then carry out the aforementioned analysis of the dimension of the sets $\mathrm{Crit}_r(L)$ to obtain our main result as follows.

Theorem 1 (GDD Converges to a Global Minimum for Almost All Initializations)**.** *Let Assumptions 1-3 hold. Then there exists an open and dense subset* $\overline{\mathcal{P}}$ *of the parameter (data) space* $\mathcal{P}$ *so that, for every pair* $(\mathbf{X}, \mathbf{Y})$ *in* $\overline{\mathcal{P}}$*, there exists an open and dense subset* $\overline{\mathcal{X}}$ *of the state space* $\mathcal{X}$ *such that every trajectory of the GDD in* (4) *corresponding to* $(\mathbf{X}, \mathbf{Y})$ *and starting in* $\overline{\mathcal{X}}$ *converges to a global minimum with* $L(\boldsymbol{\Xi}) = 0$*.*

Previous works [13, 16] only studied *local* properties of critical points by establishing that the basin of attraction of each saddle point, i.e., the set of initializations of the GDD trajectories converging to that saddle point, is of measure zero. However, to obtain a *global* picture, one must estimate how "big" is the *union* of all these basins of attraction. For that purpose, first note that the set of saddle points, being an algebraic variety of positive dimension (see Item iii) of Proposition 2 below), is therefore uncountable. This is why the previous works of local nature left open the possibility that a *global* convergence result may not hold since the uncountable union of measure zero sets may sum up to a set of positive measure. We solve this issue here by proving that the union of all the basins of attraction associated with the saddle points is in fact contained in a codimension two subset of

[3] We defer the readers to Section A in Supplementary Material for a detailed description of the theorem.

the state space $\mathcal{X}$. In the next section, a more advanced sketch of proof of Theorem 1 is provided. For the sake of readability and to avoid cumbersome technical details, only the case of $H = 1$ is elaborated. This proof provides the main arguments for the more technical analysis of the $H \geq 1$ scenario, available in an extended version of this article.

3 Detailed Analysis of the case $H = 1$

In this section, we provide a detailed proof of Theorem 1 in the case of a single-hidden-layer linear network (i.e., $H = 1$). To this end, we start with Proposition 1 which states the global convergence of gradient flows in (4) to critical points, with a polynomial convergence rate in the worst case. In the following, when state variables are concerned, we frequently drop the argument t for simplicity.

3.1 Global Convergence to Critical Points in GDD

Proof of Proposition 1 for $H = 1$. First note that for $H = 1$ the loss function L is a polynomial of degree four in the elements of $\mathbf{\Xi}$. According to Lojasiewicz's theorem, it is enough to show that every trajectory is bounded. To this end, we note from (4) that

$$\begin{cases} \frac{d\overline{\mathbf{W}}_1}{dt} = \overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{M}} \\ \frac{d\overline{\mathbf{W}}_2}{dt} = \overline{\mathbf{M}}\, \overline{\mathbf{W}}_1^\mathsf{T} \end{cases} \Rightarrow \frac{d \operatorname{tr}\left(\overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_1 + \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_2^\mathsf{T} \right)}{dt} = 4 \operatorname{tr}\left(\overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{M}} \right),$$

where we recall $\overline{\mathbf{M}} \equiv \mathbf{\Sigma_Y} - \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1$ and therefore (since $\operatorname{tr} \mathbf{A}\mathbf{A}^\mathsf{T} = \|\mathbf{A}\|_F^2$)

$$\begin{aligned}
&\frac{d\left(\|\overline{\mathbf{W}}_1\|_F^2 + \|\overline{\mathbf{W}}_2\|_F^2\right)}{dt} = 4 \operatorname{tr} \overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_2^\mathsf{T} \mathbf{\Sigma_Y} - 2 \operatorname{tr} \overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1 - 2 \operatorname{tr} \overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1 \\
&= 4 \operatorname{tr} \overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_2^\mathsf{T} \mathbf{\Sigma_Y} - \operatorname{tr} \overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2 \left(\overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2 - \mathbf{C}^0 \right) - \operatorname{tr} \overline{\mathbf{W}}_1 \overline{\mathbf{W}}_1^\mathsf{T} \left(\overline{\mathbf{W}}_1 \overline{\mathbf{W}}_1^\mathsf{T} + \mathbf{C}^0 \right) - 2\|\overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1\|_F^2 \\
&\leq 4 \operatorname{tr} \overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_2^\mathsf{T} \mathbf{\Sigma_Y} - \frac{1}{d_1} \left(\|\overline{\mathbf{W}}_1\|_F^4 + \|\overline{\mathbf{W}}_2\|_F^4 \right) - 2\|\overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1\|_F^2 + \lambda_{\max}(\mathbf{C}^0) \left(\|\overline{\mathbf{W}}_1\|_F^2 + \|\overline{\mathbf{W}}_2\|_F^2 \right) \\
&\leq c_1 \left(\|\overline{\mathbf{W}}_1\|_F^2 + \|\overline{\mathbf{W}}_2\|_F^2 \right) - c_2 \left(\|\overline{\mathbf{W}}_1\|_F^2 + \|\overline{\mathbf{W}}_2\|_F^2 \right)^2
\end{aligned}$$

for some $c_1, c_2 > 0$, where we used Cauchy–Schwarz inequality $(\operatorname{tr} \mathbf{A}\mathbf{A}^\mathsf{T})^2 \leq \operatorname{tr}(\mathbf{A}\mathbf{A}^\mathsf{T})^2 \cdot \operatorname{tr} \mathbf{I}$ along with $\left|\operatorname{tr} \mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{B}\right| \leq \lambda_{\max}(\mathbf{B}) \operatorname{tr} \mathbf{A}\mathbf{A}^\mathsf{T}$. Setting $F \equiv \|\overline{\mathbf{W}}_1\|_F^2 + \|\overline{\mathbf{W}}_2\|_F^2$ the above inequality reads $\frac{dF}{dt} \leq c_1 F - c_2 F^2$ with $F \geq 0$ and hence the sum $\|\overline{\mathbf{W}}_1\|_F^2 + \|\overline{\mathbf{W}}_2\|_F^2$ is uniformly bounded for all $t \geq 0$. With Lemma 1 we know that the difference $\|\overline{\mathbf{W}}_2\|_F^2 - \|\overline{\mathbf{W}}_1\|_F^2$ is also uniformly bounded, which further leads to the boundedness of all trajectories of both $\overline{\mathbf{W}}_1$ and $\overline{\mathbf{W}}_2$. Since the trajectories of $\overline{\mathbf{W}}_1, \overline{\mathbf{W}}_2$ (and thus $\overline{\mathbf{M}}$) are uniformly bounded for all $t \geq 0$, the norm of the gradient $\|\Delta_{\mathbf{\Xi}}\|_F$ as well as all trajectories in the GDD are bounded. The guaranteed rate of convergence can be obtained from estimates associated with polynomial gradient systems [8]. □

3.2 Characterization of Critical Points

Proposition 1 ensures, for all initializations, the convergence of the gradient descent to a critical point, i.e., a point $\mathbf{\Xi}$ in the state space $\mathcal{X}$ verifying $\Delta_{\mathbf{\Xi}}(\boldsymbol{\xi}) = 0$. Nonetheless, the information on the "quality" of the solution achieved by the algorithm is still missing. To obtain a clearer picture, we now focus on the set of *all* critical points by further decomposing the loss L with $\mathbf{\Sigma_Y} \equiv \left[\begin{array}{c|c} \mathbf{S_Y} & \mathbf{0} \end{array}\right]$ for diagonal $\mathbf{S_Y} \in \mathbb{R}^{d_y \times d_y}$ with $[\mathbf{S_Y}]_{ii} > 0$ as

$$L(\overline{\mathbf{W}}_1, \overline{\mathbf{W}}_2) = \frac{1}{2}\|\mathbf{\Sigma_Y} - \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1\|_F^2 = \frac{1}{2}\|\mathbf{S_Y} - \mathbf{C}\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{C}\mathbf{B}\|_F^2 \tag{5}$$

with $\mathbf{C} \equiv \overline{\mathbf{W}}_2 \in \mathbb{R}^{d_y \times d_1}$, $\mathbf{A} \in \mathbb{R}^{d_1 \times d_y}$ and $\mathbf{B} \in \mathbb{R}^{d_1 \times (d_x - d_y)}$ such that $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \end{array}\right] \equiv \overline{\mathbf{W}}_1$.

Under the notations above, we further expand $L(\mathbf{\Xi} + \boldsymbol{\xi})$ to obtain its higher order variation as

$$L(\mathbf{A} + \mathbf{a}, \mathbf{B} + \mathbf{b}, \mathbf{C} + \mathbf{c}) \equiv L(\mathbf{\Xi} + \boldsymbol{\xi}) = L(\mathbf{\Xi}) + \Delta_{\mathbf{\Xi}}(\boldsymbol{\xi}) + H_{\mathbf{\Xi}}(\boldsymbol{\xi}) + O\left(\|\boldsymbol{\xi}\|^3\right)$$

with $\mathbf{M} \equiv \mathbf{S_Y} - \mathbf{CA}$, $L(\boldsymbol{\Xi}) = \frac{1}{2}\|\mathbf{M}\|_F^2 + \frac{1}{2}\|\mathbf{CB}\|_F^2$ and

$$\Delta_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) \equiv -\operatorname{tr}\left(\mathbf{M}^\mathsf{T}(\mathbf{Ca} + \mathbf{cA})\right) + \operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{C}^\mathsf{T}(\mathbf{Cb} + \mathbf{cB})\right) = O(\|\boldsymbol{\xi}\|)$$

$$H_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) \equiv -\operatorname{tr}(\mathbf{M}^\mathsf{T}\mathbf{ca}) + \frac{1}{2}\|\mathbf{Ca} + \mathbf{cA}\|_F^2 + \operatorname{tr}(\mathbf{B}^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{cb}) + \frac{1}{2}\|\mathbf{Cb} + \mathbf{cB}\|_F^2 = O(\|\boldsymbol{\xi}\|^2)$$

that give the differential and the Hessian of L, respectively. Recall that $\operatorname{Crit}(L) \equiv \{\boldsymbol{\Xi} \mid \Delta_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) = 0\}$ and denote $\mathbf{M} \equiv \mathbf{S_Y} - \mathbf{CA}$, so that, by Definition 1,

$$\begin{cases} \frac{d\mathbf{A}}{dt} & \equiv -\nabla_{\mathbf{A}} L(\boldsymbol{\Xi}) = \mathbf{C}^\mathsf{T}\mathbf{M} = \mathbf{0} \\ \frac{d\mathbf{B}}{dt} & \equiv -\nabla_{\mathbf{B}} L(\boldsymbol{\Xi}) = -\mathbf{C}^\mathsf{T}\mathbf{CB} = \mathbf{0} \\ \frac{d\mathbf{C}}{dt} & \equiv -\nabla_{\mathbf{C}} L(\boldsymbol{\Xi}) = \mathbf{MA}^\mathsf{T} - \mathbf{CBB}^\mathsf{T} = \mathbf{0} \end{cases} \Leftrightarrow \begin{cases} \mathbf{C}^\mathsf{T}\mathbf{S_Y} = \mathbf{C}^\mathsf{T}\mathbf{CA} \\ \mathbf{CB} = \mathbf{0} \\ \mathbf{AS_Y} = \mathbf{AA}^\mathsf{T}\mathbf{C}^\mathsf{T}. \end{cases} \tag{6}$$

Observing the symmetric structure of $\mathbf{A}, \mathbf{C}$ in (6) we have the following lemma.

Lemma 2 (Same Kernel for $\mathbf{A}$ and $\mathbf{C}^\mathsf{T}$)**.** *Let Assumption 1 and 2 hold. Then for all* $\boldsymbol{\Xi} \in \operatorname{Crit}(L)$*,*

$$\operatorname{Ker}\mathbf{A} = \operatorname{Ker}\mathbf{C}^\mathsf{T}, \textit{ with } \operatorname{Ker}\mathbf{A} \equiv \{\mathbf{x}, \mathbf{Ax} = \mathbf{0}\}.$$

Moreover, denote r the common rank of $\mathbf{A}$ and $\mathbf{C}$ with $0 \le r \le d_y$. Then there exists some orthogonal matrix $\mathbf{U} \in \mathbb{R}^{d_y \times d_y}$ such that

$$\begin{cases} \mathbf{AU} = \left[\begin{array}{c|c} \overline{\mathbf{A}} & \mathbf{0}_{d_1 \times (d_y - r)} \end{array}\right] \\ \mathbf{C}^\mathsf{T}\mathbf{U} = \left[\begin{array}{c|c} \overline{\mathbf{C}}^\mathsf{T} & \mathbf{0}_{d_1 \times (d_y - r)} \end{array}\right] \\ \mathbf{U}^{-1}\mathbf{S_Y}\mathbf{U} = \mathbf{S_Y} \end{cases} \tag{7}$$

with $\overline{\mathbf{A}}, \overline{\mathbf{C}}^\mathsf{T} \in \mathbb{R}^{d_1 \times r}$. Moreover, if $\mathbf{S_Y}$ has distinct eigenvalues (i.e., $\overline{\mathbf{Y}}$ has d_y distinct singular values, as demanded in Assumption 3), then $\mathbf{U}$ is a permutation matrix.

Sketch of proof. It can be shown with basic algebraic manipulations that the eigenvectors of $\mathbf{S}_\mathbf{Y}^2$ (thus of $\mathbf{S_Y}$) form a basis of both $\operatorname{Ker}\mathbf{A}$ and $\operatorname{Ker}\mathbf{C}^\mathsf{T}$. Therefore $\operatorname{Ker}\mathbf{A} = \operatorname{Ker}\mathbf{C}^\mathsf{T}$ and in particular $\dim\operatorname{Ker}\mathbf{A} = \dim\operatorname{Ker}\mathbf{C}^\mathsf{T}$. We denote this dimension $d_y - r$ and $\mathbf{A}, \mathbf{C}$ are thus both of rank r. Choose $\mathbf{U}_2$ from $\operatorname{Ker}\mathbf{A}$ and $\mathbf{U}_1 \perp \operatorname{Ker}\mathbf{A}$; we deduce $\mathbf{U} = \left[\begin{array}{c|c} \mathbf{U}_1 & \mathbf{U}_2 \end{array}\right]$ so that (7) holds. $\square$

Remark from (7) in Lemma 2 that, for arbitrary $\mathbf{S_Y}$, there are infinitely many possibilities on the choice of $\mathbf{U}$ with the risk of occupying too much of the state space $\mathcal{X}$, since, with the change of variable in Lemma 2 the state variable now becomes the tuple $(\overline{\mathbf{A}}, \mathbf{B}, \overline{\mathbf{C}}, \mathbf{U})$. Using Assumption 3, $\mathbf{U}$ only takes a finite number of values (the 2^{d_y} permutation matrices) for a given $\boldsymbol{\Xi} \in \operatorname{Crit}_r(L)$, hence the state variable essentially becomes the tuple $(\overline{\mathbf{A}}, \mathbf{B}, \overline{\mathbf{C}})$.

For $\boldsymbol{\Xi} \in \operatorname{Crit}(L)$ with $\mathbf{A}, \mathbf{C}$ of rank r with $0 \le r \le d_y$, rewriting $\mathbf{S_Y}$ in two blocks $\mathbf{S_Y} = \begin{bmatrix} \mathbf{D_Y} & \mathbf{0} \\ \mathbf{0} & \mathbf{E_Y} \end{bmatrix}$, with $\mathbf{D_Y} \in \mathbb{R}^{r \times r}$ and $\mathbf{E_Y} \in \mathbb{R}^{(d_y - r) \times (d_y - r)}$. With Lemma 2, we then simplify (6) as

$$\begin{cases} \overline{\mathbf{C}}\,\overline{\mathbf{A}} = \mathbf{D_Y} \\ \overline{\mathbf{C}}\mathbf{B} = \mathbf{0} \end{cases}, \quad \mathbf{U}^\mathsf{T}\mathbf{MU} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{E_Y} \end{bmatrix} \tag{8}$$

with the fact that $\overline{\mathbf{C}}^\mathsf{T}, \overline{\mathbf{A}}$ are both of full rank (equal to r). The loss $L(\boldsymbol{\Xi})$ (at critical points) can thus be simplified as $L(\boldsymbol{\Xi}) = \frac{1}{2}\|\mathbf{E_Y}\|_F^2$ where $\mathbf{E_Y}$ measures the "quality" of each critical points.

For any $\boldsymbol{\Xi} \in \operatorname{Crit}(L)$, with Lemma 2 we are allowed to "extract" the full rank (sub-)structures of $\mathbf{A}, \mathbf{C}$ with $\mathbf{S_Y}$ unchanged, via a simple change of basis. For $0 \le r \le d_y$, let $\operatorname{Crit}_r(L)$ be the subset of $\operatorname{Crit}(L)$ such that the rank of $\mathbf{A}$ and of $\mathbf{C}$ is equal to r. Then, one has the following disjoint union

$$\operatorname{Crit}(L) = \cup_{r=0}^{d_y} \operatorname{Crit}_r(L).$$

This precise characterization of critical points naturally leads to the following proposition on on the loss function $L(\cdot)$, that can be further "visualized" as in Figure 2.

Proposition 2 (Landscape of Single-hidden-layer Linear Network)**.** Under Assumptions 1-3, the loss function $L(\boldsymbol{\Xi})$ has the following properties:

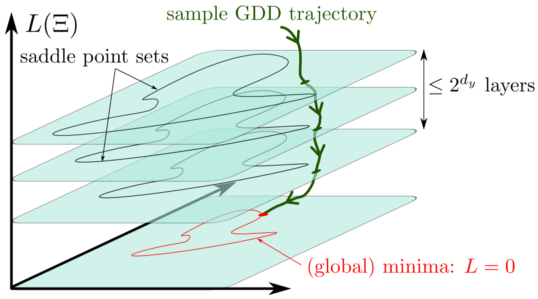

Figure 2: A geometric "vision" of the loss landscape.

$i)$ The set of possible limits of L along the GDD given by (3) is equal to the finite set made of the sum of the squares of any subset of the singular values of $\overline{\mathbf{Y}}$.

$ii)$ The set $\mathrm{Crit}_{d_y}(L)$ is in fact the set of local (and global) minima, with $L = 0$ and $\mathbf{M} = \mathbf{0}$.

$iii)$ Every critical point in $\mathrm{Crit}_r(L)$ with $0 \leq r \leq d_y - 1$ is a saddle point such that the Hessian has at least two negative eigenvalues. In particular, the set of saddle points is an algebraic variety of positive dimension, i.e., (up to a permutation matrix) the zero set of the polynomial functions given in (8), with $\mathbf{E}_\mathbf{Y} \neq \mathbf{0}$.

Proof. Item $i)$ follows directly from the discussion preceding the proposition. As for Item $ii)$, we write the Hessian for a given $\mathbf{\Xi} \in \mathrm{Crit}(L)$ as

$$H_{\mathbf{\Xi}}(\boldsymbol{\xi}) = -\operatorname{tr}(\mathbf{M}^\mathsf{T}\mathbf{c}\mathbf{a}) + \frac{1}{2}\|\mathbf{C}\mathbf{a} + \mathbf{c}\mathbf{A}\|_F^2 + \frac{1}{2}\|\mathbf{C}\mathbf{b} + \mathbf{c}\mathbf{B}\|_F^2$$

with $\mathbf{\Xi} \equiv (\mathbf{A}, \mathbf{B}, \mathbf{C})$ such that (6) is satisfied and arbitrary $\boldsymbol{\xi} \equiv (\mathbf{a}, \mathbf{b}, \mathbf{c})$.

The fact that no critical point is a *local maximum* is easily checked by taking $\boldsymbol{\xi}$ such that $\mathbf{C}\mathbf{b}+\mathbf{c}\mathbf{B} \neq \mathbf{0}$ and $\mathbf{a} = \mathbf{0}$. Therefore, the Hessian has positive eigenvalues and $\mathbf{\Xi}$ is a not local maximum. Such $\boldsymbol{\xi}$ always exists for at least one of $\mathbf{B}, \mathbf{C}$ away from $\mathbf{0}$, while the case $\mathbf{B} = \mathbf{C} = \mathbf{0}$ is essentially the case $r = 0$. Moreover, note from Lemma 2 that, for $\mathbf{\Xi}$ with $\mathbf{A}, \mathbf{C}$ both of full rank ($r = d_y$) we have $\mathbf{M} = \mathbf{0}$, resulting in equivalent global minima.

To show Item $iii)$ it remains to prove that a critical point $\mathbf{\Xi} \in \mathrm{Crit}_r(L)$ with $0 \leq r \leq d_y - 1$ is in fact a saddle point. To this end, with the same change of basis as in Lemma 2 we write $\mathbf{a}\mathbf{U} = \left[\begin{array}{c|c} \mathbf{a}_1 & \mathbf{a}_2 \end{array}\right]$, $\mathbf{c}^\mathsf{T}\mathbf{U} = \left[\begin{array}{c|c} \mathbf{c}_1^\mathsf{T} & \mathbf{c}_2^\mathsf{T} \end{array}\right]$, with $\mathbf{a}_1, \mathbf{c}_1^\mathsf{T} \in \mathbb{R}^{d_1 \times r}$ so that the associated Hessian becomes

$$H_{\mathbf{\Xi}}(\boldsymbol{\xi}) = -\operatorname{tr}(\mathbf{E}_\mathbf{Y}\mathbf{c}_2\mathbf{a}_2) + \frac{1}{2}\|\overline{\mathbf{C}}\mathbf{a}_1 + \mathbf{c}_1\overline{\mathbf{A}}\|_F^2 + \frac{1}{2}\|\overline{\mathbf{C}}\mathbf{a}_2\|_F^2 + \frac{1}{2}\|\mathbf{c}_2\overline{\mathbf{A}}\|_F^2 + \frac{1}{2}\|\overline{\mathbf{C}}\mathbf{b} + \mathbf{c}_1\mathbf{B}\|_F^2 + \frac{1}{2}\|\mathbf{c}_2\mathbf{B}\|_F^2$$

with $\mathbf{E}_\mathbf{Y}$ given in (8).

We next determine a non trivial linear subspace $\mathcal{G}_{\mathbf{\Xi}}$ such that the restriction of $H_{\mathbf{\Xi}}$ to $\mathcal{G}_{\mathbf{\Xi}}$ is negative definite, which implies, as a simple consequence of the min-max theorem for quadratic forms (see for example Theorem 4.2.6 in [11]), that $H_{\mathbf{\Xi}}$ has at least $\dim \mathcal{G}_{\mathbf{\Xi}}$ negative eigenvalues. To this end, we confine our attention to the subspace $\mathcal{F}_{\mathbf{\Xi}}$ defined by $\mathbf{a}_1 = \mathbf{c}_1 = \mathbf{b} = \overline{\mathbf{C}}\mathbf{a}_2 = \mathbf{c}_2\overline{\mathbf{A}} = \mathbf{0}$. Since $\overline{\mathbf{C}}\,\overline{\mathbf{A}} = \mathbf{D}_\mathbf{Y}$ from (8), we have that $\mathbb{R}^{d_1}$ is the direct sum of $\operatorname{im}\overline{\mathbf{A}}$ and $\operatorname{Ker}\overline{\mathbf{C}}$, with $\operatorname{im}\overline{\mathbf{A}}$ the image of the linear map $\mathbf{x} \mapsto \overline{\mathbf{A}}\mathbf{x}$. Hence, up to a change of basis, one can assume that $\overline{\mathbf{C}} = \left[\begin{array}{c|c} \widetilde{\mathbf{C}} & \mathbf{0} \end{array}\right]$ and $\overline{\mathbf{A}}^\mathsf{T} = \left[\begin{array}{c|c} \widetilde{\mathbf{A}}^\mathsf{T} & \mathbf{0} \end{array}\right]$, with nonsingular $\widetilde{\mathbf{C}}, \widetilde{\mathbf{A}} \in \mathbb{R}^{r \times r}$. If we write $\mathbf{a}_2$ and $\mathbf{c}_2$ in accordance to the previous direct sum, $\mathcal{F}_{\mathbf{\Xi}}$ is in fact defined by $\mathbf{a}_1 = \mathbf{c}_1 = \mathbf{b} = \mathbf{0}$, $\mathbf{a}_2^\mathsf{T} = \left[\begin{array}{c|c} \mathbf{0} & \widetilde{\mathbf{a}}_2^\mathsf{T} \end{array}\right]$ and $\mathbf{c}_2 = \left[\begin{array}{c|c} \mathbf{0} & \widetilde{\mathbf{c}}_2 \end{array}\right]$. It is therefore of dimension $2(d_1 - r)(d_y - r)$ and, on it, there exists $c_0 > 0$ such that we have

$$\begin{aligned} H_{\mathbf{\Xi}}(\boldsymbol{\xi}) &= -\operatorname{tr}(\mathbf{E}_\mathbf{Y}\mathbf{c}_2\mathbf{a}_2) + \frac{1}{2}\|\mathbf{c}_2\mathbf{B}\|_F^2 \leq -\operatorname{tr}(\mathbf{E}_\mathbf{Y}\widetilde{\mathbf{c}}_2\widetilde{\mathbf{a}}_2) + c_0^2\|\widetilde{\mathbf{c}}_2\|_F^2 \\ &= \left\| c_0\widetilde{\mathbf{c}}_2^\mathsf{T} - \frac{\widetilde{\mathbf{a}}_2\mathbf{E}_\mathbf{Y}}{2c_0} \right\|_F^2 - \frac{\|\widetilde{\mathbf{a}}_2\mathbf{E}_\mathbf{Y}\|_F^2}{4c_0^2}. \end{aligned}$$

We now define $\mathcal{G}_{\mathbf{\Xi}}$ as the subspace of $\mathcal{F}_{\mathbf{\Xi}}$ such that $\widetilde{\mathbf{c}}_2^\mathsf{T} = \frac{\widetilde{\mathbf{a}}_2 \mathbf{E}_{\mathbf{Y}}}{2c_0^2}$ on which $H_{\mathbf{\Xi}}$ is clearly negative definite by the above equation. We thus deduce that $H_{\mathbf{\Xi}}$ has at least $(d_1 - r)(d_y - r) \geq 2$ negative eigenvalues (since by Assumption 1 we have $d_1 > d_y > r$), which concludes the proof. □

The fact that all local minima are equivalently global minima and all critical points that are not global minima are saddle points is in fact already known for single-hidden-layer linear networks [2] as well as for deep linear networks [13]. Here we provided an alternative and shorter proof.

3.3 Convergence to Global Minima for Almost All Initializations

Having characterized the critical points, we now show that the GDD almost always converges to a local (and thus global) minimum, thereby completing the proof of Theorem 1.

End of proof of Theorem 1. We now complete the proof of Theorem 1 (for $H = 1$). Since $\mathrm{Crit}_r(L)$ is fully characterized by (8), it is an algebraic variety, i.e., the set of zeros of a polynomial of degree two, made of a finite number of two by two disjoint smooth strata [10]. One can therefore attach a dimension to this algebraic variety as the largest dimension of each smooth stratum (the latter integer defined as the standard dimension of a differentiable manifold). A simple but instrumental remark to be made is that, for each $\mathbf{\Xi} \in \mathrm{Crit}_r(L)$, the corresponding stratum at $\mathbf{\Xi}$ does not belong to the basin of attraction of $\mathbf{\Xi}$ (since the corresponding GDD trajectories are converging to *other* critical points rather that $\mathbf{\Xi}$). More concretely, let $\mathcal{C}(\mathbf{\Xi})$ be the tangent space to the corresponding stratum at $\mathbf{\Xi}$ that is of dimension $D_C^r(\mathbf{\Xi})$ (which is locally constant on each stratum since it is the dimension of that stratum). Since the associated Hessian $H_{\mathbf{\Xi}}$ is identically equal to zero on $\mathcal{C}(\mathbf{\Xi})$, the latter is contained in the tangent space of the central manifold associated with the GDD at $\mathbf{\Xi}$. Hence, by standard transversality arguments [10], one deduces that there exists (in an open neighborhood of every $\mathbf{\Xi}$ in every stratum of $\mathrm{Crit}_r(L)$) a differentiable manifold $\mathcal{M}(\mathbf{\Xi})$ containing the basin of attraction of $\mathbf{\Xi}$, which is transverse to the direct sum of $\mathcal{C}(\mathbf{\Xi})$ and the linear span made of the variations corresponding to negative eigenvalues of $H_{\mathbf{\Xi}}$ (the tangent space of the unstable manifold of the GDD at $\mathbf{\Xi}$).

Therefore, by Item $iii)$ of Proposition 2, we have the dimension of the aforementioned differentiable manifold $\dim \mathcal{M}(\mathbf{\Xi}) = \dim \mathcal{X} - 2 - D_C^r(\mathbf{\Xi})$. As a consequence, the union of all the basins of attraction of saddle points in $\mathrm{Crit}_r(L)$ is (locally) contained in a set parameterized by $(\mathbf{\Xi}, \mathcal{M}(\mathbf{\Xi}))$, for $\mathbf{\Xi}$ in a stratum of $\mathrm{Crit}_r(L)$. The dimension of this set $\mathcal{S}_r$ is then clearly upper bounded by the sum of $D_C^r(\mathbf{\Xi})$ (that is locally constant) and $\dim \mathcal{X} - 2 - D_C^r(\mathbf{\Xi})$, hence by $\dim \mathcal{X} - 2$. Since this bound does not depend on (any open neighborhood in $\mathrm{Crit}_r(L)$ of) $\mathbf{\Xi}$, we deduce that the union of *all* the basins of attraction of saddle points in $\mathrm{Crit}_r(L)$ is contained in a subset (of the state space $\mathcal{X}$) of codimension at least 2. Since there is a finite number of $\mathcal{S}_r$, their union for $r = 0, \ldots, d_y - 1$ is also of codimension two, which concludes the proof of Theorem 1. □

As stated in Proposition 1 and Theorem 1, GDD achieves at least a polynomial convergence rate [8] to a global minimum (for almost all initializations). As a side and immediate aftermath, it can be shown that, upon proper initialization, exponential convergence can be achieved (here for $H = 1$).

Remark 1 (Exponential Convergence of GDD)**.** *Let Assumptions 1 and 2 hold. Then, every trajectory of the GDD such that* $\mathbf{C}^0 \equiv (\overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2 - \overline{\mathbf{W}}_1 \overline{\mathbf{W}}_1^\mathsf{T})\big|_{t=0}$ *has at least* d_y *strictly positive eigenvalues, converges to a global minimum at the rate of* $e^{-2\alpha t}$ *with* α *the* d_y*-th smallest eigenvalue of* $\mathbf{C}^0$.

Proof. Recalling that for $H = 1$ we have $L = \frac{1}{2} \left\| \mathbf{\Sigma}_{\mathbf{Y}} - \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1 \right\|_F^2$, with (4) we deduce

$$\frac{d \|\overline{\mathbf{M}}\|_F^2}{dt} = \frac{d \operatorname{tr}(\overline{\mathbf{M}}^\mathsf{T} \overline{\mathbf{M}})}{dt} = -2 \operatorname{tr}(\overline{\mathbf{M}}^\mathsf{T} \overline{\mathbf{M}}\, \overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_1 + \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{M}}\, \overline{\mathbf{M}}^\mathsf{T}) \leq -2c_0 \|\overline{\mathbf{M}}\|_F^2 \tag{9}$$

with $\overline{\mathbf{M}} \equiv \mathbf{\Sigma}_{\mathbf{Y}} - \overline{\mathbf{W}}_2 \overline{\mathbf{W}}_1$ and $c_0 = \lambda_{\min}(\overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_1) + \lambda_{\min}(\overline{\mathbf{W}}_2 \overline{\mathbf{W}}_2^\mathsf{T})$. Since $\overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_1 \in \mathbb{R}^{d_x \times d_x}$ is of maximum rank d_1 (with $d_1 \leq d_x$ from Assumption 1), we have $\lambda_{\min}(\overline{\mathbf{W}}_1^\mathsf{T} \overline{\mathbf{W}}_1) = 0$. Nonetheless, $\overline{\mathbf{W}}_2 \overline{\mathbf{W}}_2^\mathsf{T} \in \mathbb{R}^{d_y \times d_y}$ may be of full rank so that $\lambda_{\min}(\overline{\mathbf{W}}_2 \overline{\mathbf{W}}_2^\mathsf{T}) > 0$. To this end, we decompose $\overline{\mathbf{W}}_2 = \left[\begin{array}{c|c} \overline{\mathbf{W}}_{21} & \overline{\mathbf{W}}_{22} \end{array} \right]$, with $\overline{\mathbf{W}}_{21} \in \mathbb{R}^{d_y \times d_y}$. Then with the inclusion principle of Hermitian matrices (e.g., Theorem 4.3.28 in [11]) we deduce $\lambda_{\min}(\overline{\mathbf{W}}_{21}^\mathsf{T} \overline{\mathbf{W}}_{21}) \geq \lambda_{d_y}(\overline{\mathbf{W}}_2^\mathsf{T} \overline{\mathbf{W}}_2)$. Moreover, since

$\lambda_{\min}(\overline{\mathbf{W}}_{21}^{\mathsf{T}}\overline{\mathbf{W}}_{21}) = \lambda_{\min}(\overline{\mathbf{W}}_{21}\overline{\mathbf{W}}_{21}^{\mathsf{T}})$, by Lemma 1 and Weyl's inequality (e.g., Corollary 4.3.12 in [11]), we have

$$\lambda_{\min}(\overline{\mathbf{W}}_2\overline{\mathbf{W}}_2^{\mathsf{T}}) \geq \lambda_{\min}(\overline{\mathbf{W}}_{21}\overline{\mathbf{W}}_{21}^{\mathsf{T}}) \geq \lambda_{d_y}(\overline{\mathbf{W}}_2^{\mathsf{T}}\overline{\mathbf{W}}_2) \geq \lambda_{d_y}(\mathbf{C}^0).$$

As such, (9) yields $\frac{d\,\mathrm{tr}(\overline{\mathbf{M}}^{\mathsf{T}}\overline{\mathbf{M}})}{dt} \leq -2\lambda_{d_y}(\mathbf{C}^0)\,\mathrm{tr}(\overline{\mathbf{M}}^{\mathsf{T}}\overline{\mathbf{M}})$ which concludes the proof. $\square$

4 Concluding Remarks

To the best of the authors' knowledge, it is the first time that the *global* behavior of the gradient descent dynamics in linear neural networks is fully characterized, in the sense that we show a *global* convergence to critical points of all trajectories of the gradient flow via Lojasiewicz's theorem, which helps eliminate the possibility of divergence. Then with a fine local study of critical points we exclude the (possible) worries concerning the "accumulation" of saddle points together with associated basin of attractions so that they form "disjoint layers" that are of total measure zero in the total weight space. Interestingly, Lojasiewicz's theorem is more powerful than needed here and may enable extensions of the present results to more advanced dynamics than the simple GDD (see Remark 2 in Supplementary Material for more details).

It is interesting to note that the authors in [9, 6], made a strong case to warn against saddle points in deep learning, which is in sharp contrast with our conclusions. Yet, the analysis in [9, 6] is asymptotic in the network dimensions, where the present one is set for fixed network sizes. It would be of interest to conciliate both results to gain a even clearer picture of deep linear learning in practical scenarios.

When nonlinear networks are considered, obtaining an equivalent version of Lemma 1 would be a key enabler to achieve the global convergence to critical points as per Lojasiewicz's theorem and therefore would allows for a better understanding of the *nonlinear* deep networks performance. Exploring a random model setting for $\mathbf{X}, \mathbf{Y}$, the authors in [6] argue that the loss surfaces of these networks loosely recall (yet is formally quite different from) a spin-glass model, familiar to statistical physicists. In this case, as the network gets large, local minima gather in a thin "band" of similar losses isolated from the global minimum. Stating that the number of local minima outside that band diminishes exponentially with the size of the network, the authors argue that the gradient descent dynamics (in their case the stochastic gradient descent dynamics) converges to this band and therefore leads to deep *nonlinear* networks with good generalization performance. Taking advantage of a random nature for $\mathbf{X}, \mathbf{Y}$ in our present setting would allow for a refinement of our proposed geometric vision, likely by means of a "statistical extension" of the key Lemma 1.

References

[1] Pierre-Antoine Absil, Robert Mahony, and Benjamin Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[2] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[3] Avrim Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[4] Joseph-Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

[7] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[8] Didier D'Acunto and Krzysztof Kurdyka. Explicit bounds for the Lojasiewicz exponent in the gradient inequality for polynomials. In *Annales Polonici Mathematici*, volume 1, pages 51–61, 2005.

[9] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

[10] Mark Goresky and Robert MacPherson. Stratified morse theory. In *Stratified Morse Theory*, pages 3–22. Springer, 1988.

[11] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 1990.

[12] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[13] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

[17] S Lojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. *Seminari di geometria*, 1983:115–117, 1982.

[18] Stanislaw Lojasiewicz. Ensembles semi-analytiques. *Lectures Notes IHES (Bures-sur-Yvette)*, 1965.

[19] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, 2012.

[20] Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

[21] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[22] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[23] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[24] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Society Providence, 2012.

[25] Willard I Zangwill. Convergence conditions for nonlinear programming algorithms. *Management Science*, 16(1):1–13, 1969.

Supplementary Material

Almost Global Convergence to Global Minima for Gradient Descent in Deep Linear Networks

A Lojasiewicz's theorem

We first recall Lojasiewicz's theorem for the convergence of real analytic gradient flows, which is essentially the key enabler to prove the global convergence of the GDD trajectories.

Theorem 2 (Lojasiewicz's theorem, [17])**.** *Let L be a real analytic function and let $\Xi(\cdot)$ be a solution trajectory of the gradient system given by Definition 1. Further assume that $\sup_{t\geq 0} \|\Xi(t)\| < \infty$. Then $\Xi(\cdot)$ converges to a critical point of L, as $t \to \infty$.*[4]

Remark 2. Since the fundamental (strict) gradient descent direction (as in Definition 1) in Lojasiewicz's theorem can in fact be relaxed to a (more general) angle condition (see for example Theorem 2.2 in [1]), the line of argument developed in the core of the article may be similarly followed to prove the global convergence of more advanced optimizers (e.g., SGD, SGD-Momentum [21], ADAM [14], etc.), for which the direction of descent is not strictly the opposite of the gradient direction. This constitutes an important direction of future exploration.

[4]This theorem is based on the fundamental Lojasiewicz's inequality of analytic functions [18].