2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Dynamics and Perturbations of Overparameterized Linear Neural Networks

Arthur Castello B. de Oliveira[1], Milad Siami[1], and Eduardo D. Sontag [1,2]

Abstract— **Recent research in neural networks and machine learning suggests that using many more parameters than strictly required by the initial complexity of a regression problem can result in more accurate or faster-converging models – contrary to classical statistical belief. This phenomenon, sometimes referred to as "benign overfitting", raises questions regarding in what other ways might overparameterization affect the properties of a learning problem. In this work, we investigate the effects of overfitting on the robustness of gradient-descent training when subject to uncertainty on the gradient estimation, which arises naturally if the gradient is estimated from noisy data or directly measured. Our object of study is a linear neural network with a single, arbitrarily wide, hidden layer and an arbitrary number of inputs and outputs, which can be equivalently written as an overparameterized matrix factorization problem. In this paper we solve the problem for the case where the input and output of our neural network are one-dimensional, deriving sufficient conditions for the robustness of our system based on necessary and sufficient conditions for convergence in the undisturbed case. We then show that the general overparameterized formulation introduces a set of spurious equilibria that lay outside the set where the loss function is minimized, and discuss directions of future work that might extend our current results for more general settings.**

I. Introduction

Benign overfitting is a phenomenon observed when training large/deep neural networks [1]. This observation, when first made, was disruptive because it challenged the traditional notion that overfitting a model always decreased its performance. Since then, many works have attempted to explain or understand this phenomenon [1]–[10] for different classes of systems.

Two common simplifying assumptions made when analyzing benign overfitting on neural networks are that of linear activation functions and a single hidden layer, which together make the problem equivalent to an overparameterized matrix factorization problem, and closely related to linear regressions. In [5] the authors derive conditions for benign overfitting to happen in linear regression problems, which relate closely to the simplified neural network in question. In [8], [9] the authors discuss how the gradient descent on matrix factorization problems tends to prefer solutions with good generalization properties. These works indicate that this simplified version of the problem is not only an important first step for gaining insights about the general case but also is by itself still a rich and complex problem, with works beyond just the scope of overparameterization [11].

Moreover, overparameterization in linear regression and matrix factorization problems is also known to potentially accelerate the training process [12], and in recent works [13], [14] the authors characterize the convergence rate of the system as a function of the initialization of the gradient flow dynamics. Their work shows that the more imbalanced (in a sense formally defined in the papers) the initial conditions are, the quicker the system converges to an equilibrium. Furthermore, they show in [14] that the gradient flow for overparameterized linear regressions converges at least as quickly as the non-overparameterized case. This is a very interesting result and naturally raises the question of what is the disadvantage of using overparameterized formulations, if not only the accuracy might be increased, but the training time is potentially quicker as well, despite the higher number of parameters.

We then look at the robustness trade-off from adopting overparameterized formulations. Many works [15]–[20] evaluate the post-training performance when the input is subject to adversarial disturbances. This became a very active area of research once it was noticed in [21] that by applying visually imperceptible noises, one could completely fool image recognition networks, despite their high training accuracy. The works in the area focus on adapting the training process to maximize the robustness of the network to adversarial attacks while still maintaining satisfying performance.

Other papers analyze the robustness of the gradient flow as a tool for minimizing arbitrary functions. In [22] the authors analyze the convergence of the stochastic gradient descent as a function of the probability distribution of the gradient noise. In [23] one of the coauthors showed that the gradient flow is ISS when the estimation of the gradient is uncertain as long as the loss function satisfies some conditions. This establishes a sense of robustness for this class of systems when no overparameterization is considered, and is a motivation for this paper to study whether this property is maintained, and to what degree, once an overparameterized formulation is considered.

Multiple works in the literature analyze the behavior of linear neural networks when submitted to some gradient dynamics for training [24]–[31], and many of those results, presented for different assumptions on the system, allow us to conclude that for linear neural networks with a single hidden layer: all local minima of the cost function are global minima; all non-local minima critical points are strict

This work was partially supported by ONR Grant N00014-21-1-2431, NSF Grant 2121121, and AFOSR Grant FA9550-21-1-0289.

[1]Department of Electrical & Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mails: {castello.a, m.siami, e.sontag}@northeastern.edu).

[2]Departments of Bioengineering, Northeastern University, Boston, MA 02115 USA.

saddles (the Hessian has at least one negative eigenvalue); all solutions converge to a critical point of the cost function; and for almost every initial condition the solutions converge to the global minimum. All of these results give a complete qualitative understanding of the behavior of our solutions and allow us to conclude stochastic guarantees for the problem under consideration.

In this work, we study the robustness of gradient descent as a training tool when our problem formulation is overparameterized. We will formulate the problem for the general case, but in this conference paper, we work out in detail only the scalar/vector case. Understanding in depth this simplified instance of the problem gives important insight into what to look for, and how to understand the general case, and we discuss similarities and differences between the scalar/vector and general cases.

Our analysis shows that even for the simplified vector case, considering an overparameterized formulation incurs a loss of robustness when compared to the global ISS property of the non-overparameterized formulation demonstrated in [23]. For this case, however, one can characterize the "bad" set of initializations which result in a solution that does not converge to the target set as a zero measure set, and define forward invariant sets of our state space in which the system is guaranteed to be ISS for disturbances with bounded magnitudes. Such a workaround, however, is not easily extendable for the general case, as indicated by our preliminary analysis presented in this paper. We also discuss existing results in the literature and how they can help us understand the problem of robustness when disturbances are present in the estimation of the gradient.

The paper is organized as follows: Section I consists of this introduction, in which we provide a brief overview of the relevance and progress done regarding overparameterized formulations for optimization problems; Section II presents our problem formulation and performs a preliminary analysis of the convergence of our solutions based on considering our cost function as a candidate Lyapunov Function; Section III completely deals with the problem of characterizing the robustness of the vector case ($n = m = 1$ and k arbitrary) demonstrating that the ISS property can be recovered for this case by choosing a proper invariant manifold in which to constrain our dynamics; Section IV presents some preliminary analysis of the general case, focusing on the existence of an enlarged equilibrium set, which hinder the definition of a forward-invariant manifold to recover the ISS property; Section V finally concludes this paper and provides insights on the steps being taken to solve the problem of robustness for the general case.

II. Motivation and Formulation

A. Preliminary Definitions

In this paper, we use I to denote the identity matrix, that is, a square matrix whose all elements are zero except for the ones in its main diagonal, which are one. If we want to emphasize the dimension of the identity matrix we write I_n where $I \in \mathbb{R}^n$; otherwise if the dimension is not indicated, it is clear from the context. Similarly, we define e_i as the i-th elementary vector which is the i-th column of I for the dimension implicit in the context. The matrix $E_{ij} = e_i e_j^\top$ is called an elementary matrix and has all elements zero except for the one in row i and column j, which is one (notice that E_{ij} does not need to be square). Let $\mathbb{R}_+$ be the set of non-negative real numbers, then $\|\cdot\|_F : \mathbb{R}^{m\times n} \to \mathbb{R}_+$ and $\|\cdot\|_2 : \mathbb{R}^n \to \mathbb{R}_+$ denote the Frobenius and ℓ_2 norms for arbitrary matrix and vector spaces respectively.

Let $\mathrm{vec}(\cdot) : \mathbb{R}^{n\times m} \to \mathbb{R}^{nm}$ be the vectorization operator which concatenates the columns of its input matrix into a single column vector. Notice that $\mathrm{vec}(\cdot)$ is bijective and thus admits an inverse, denoted in this paper as $\mathrm{vec}^{-1}(\cdot)$. Furthermore, let $\otimes$ denote the Kronecker product (which is commutative and bilinear), then for three arbitrary matrices A, B, and C of matching dimensions, the following well-known identity is used freely during some derivations of this paper

$$\mathrm{vec}(ABC) = \left(C^\top \otimes A\right) \cdot \mathrm{vec}(B). \tag{1}$$

B. Linear Neural Networks with One Hidden Layer

Given a set of ℓ paired sampled inputs $x = \{x_i\}_{i=1}^{\ell}$ and outputs $y = \{y_i\}_{i=1}^{\ell}$, where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}^m$, the *linear regression problem* can be expressed as the following optimization problem

$$\min_{\Theta\in\mathbb{R}^{n\times m}} \quad \frac{1}{2}\|Y - \Theta^\top X\|_F^2, \tag{2}$$

where X and Y are $n \times \ell$ and $m \times \ell$ real matrices whose i-th columns are x_i and y_i, respectively.

If we assume a rich enough dataset ($\ell > \max\{m, n\}$ and that X is a full rank matrix) the unique solution to this problem can be obtained as follows:

$$\Theta^* = (YX^\dagger)^\top. \tag{3}$$

Alternatively, for the same set of data points, training a neural network with a single hidden layer and linear activation functions can be written as solving the following optimization problem

$$\underset{P\in\mathbb{R}^{n\times k}, Q\in\mathbb{R}^{m\times k}}{\operatorname{argmin}} \quad \frac{1}{2}\|Y - QP^\top X\|_F^2, \tag{4}$$

where $k \geq n, m$. Not only is this formulation similar to linear regression in many ways (although still fundamentally different), but one can also verify that P and Q solve (4) if and only if they also solve the following matrix factorization problem:

$$\underset{P\in\mathbb{R}^{n\times k}, Q\in\mathbb{R}^{m\times k}}{\operatorname{argmin}} \quad \frac{1}{2}\|\bar{Y} - PQ^\top\|_F^2, \tag{5}$$

where $\bar{Y} = \Theta^* = (YX^\dagger)^\top$, albeit at a different minimum value. Notice that by choosing a factorized representation, the number of free parameters is larger than the number of constraints in the problem even if $k \leq \max(m, n)$, however, due to redundancies resulting from the matrix multiplication, one must guarantee the minimum rank of matrix $PQ^\top$, not the number of free parameters. Furthermore, by asking $k >$

$\max(m,n)$ instead of simply $k = \max(m,n)$, we impose the existence of a target set $\mathcal{T} := \{(P,Q) \mid \bar{Y} = PQ^\top\}$ instead of an isolated optimal equilibrium. The existence of a nontrivial set of optimal solutions is the reason for some of the interesting properties observed in the literature, namely the faster convergence for imbalanced initializations observed in [12], therefore we make a point to guarantee the existence of such set for our robustness analysis.

One possible method for solving the matrix factorization problem (5) is the use of a gradient flow for the dynamics of the parameters, however the resulting dynamics can be shown to have multiple *spurious equilibria*, that is, equilibrium points of the dynamics that do not lie in the target set.

By understanding how this problem behaves when solved by an uncertain gradient flow algorithm we also gain an understanding of the robustness of our original linear neural network problem. This motivates us to look at the dynamics of the parameters during training in search of some kind of robustness guarantee. Specifically, we look for ways to guarantee or recover the ISS property that is present in non-overparameterized gradient flow systems, as shown in [23].

C. The Gradient Flow Dynamics

To impose a gradient flow dynamics for the parameters, let us define the loss function as follows:

$$\mathcal{L}(P,Q) = \frac{1}{2}\|\bar{Y} - PQ^\top\|_F^2. \tag{6}$$

Then, as derived in [14], we impose the following dynamics for the parameters P and Q

$$\begin{aligned} \dot{P} &= -\nabla_P \mathcal{L}(P,Q) = (\bar{Y} - PQ^\top)Q, \\ \dot{Q} &= -\nabla_Q \mathcal{L}(P,Q) = (\bar{Y} - PQ^\top)^\top P, \end{aligned} \tag{7}$$

or equivalently

$$\dot{Z} = \begin{bmatrix} \dot{P} \\ \dot{Q} \end{bmatrix} = \begin{bmatrix} (\bar{Y} - PQ^\top)Q \\ (\bar{Y} - PQ^\top)^\top P \end{bmatrix} = \begin{bmatrix} f_P(P,Q) \\ f_Q(P,Q) \end{bmatrix} = f_Z(Z). \tag{8}$$

Often, however, the gradient value used to enforce the dynamics is an estimation of its true value and has an uncertainty associated with it. To model this uncertainty we add two disturbance terms on the dynamics as below

$$\dot{Z} = \begin{bmatrix} \dot{P} \\ \dot{Q} \end{bmatrix} = \begin{bmatrix} (\bar{Y} - PQ^\top)Q \\ (\bar{Y} - PQ^\top)^\top P \end{bmatrix} + \begin{bmatrix} U \\ V \end{bmatrix} = f_Z(Z) + \begin{bmatrix} U \\ V \end{bmatrix}, \tag{9}$$

where $U : \mathbb{R}_+ \to \mathbb{R}^{n\times k}$ and $V : \mathbb{R}_+ \to \mathbb{R}^{m\times k}$. In the next section, we explore a candidate Lyapunov function as a means to characterize the stability of our system as a function of the magnitude of our disturbances and of our initialization.

D. The Loss Function as a Candidate Lyapunov Function

A natural choice for a candidate Lyapunov function is our loss function. By definition, $\mathcal{L}(P,Q) > 0$ whenever P, Q are not on the target set $\bar{Y} = PQ^\top$. Furthermore, one can compute an upper-bound on the time derivative of the loss function under gradient flow as follows

$$\begin{aligned} \dot{\mathcal{L}}(P,Q,U,V) &= \left\langle \nabla\mathcal{L}, \begin{bmatrix} \dot{P} \\ \dot{Q} \end{bmatrix} \right\rangle \\ &= \left\langle \nabla\mathcal{L}, -\nabla\mathcal{L} + \begin{bmatrix} U \\ V \end{bmatrix} \right\rangle \\ &= -\|\nabla\mathcal{L}\|_F^2 + \left\langle \nabla\mathcal{L}, \begin{bmatrix} U \\ V \end{bmatrix} \right\rangle \\ &\leq -\|\nabla\mathcal{L}\|_F^2 + \|\nabla\mathcal{L}\|_F \left\| \begin{bmatrix} U \\ V \end{bmatrix} \right\|_F. \end{aligned} \tag{10}$$

With this, we can establish the following theorem:

Theorem 1. *The time derivative of the loss function* (6) *can be upper-bounded as follows:*

$$\dot{\mathcal{L}} \leq -\mathcal{L}(P,Q) \cdot \left(\sigma_{\min}^2(Q) + \sigma_{\min}^2(P)\right) + \frac{1}{2}\left\| \begin{bmatrix} U \\ V \end{bmatrix} \right\|_F^2. \tag{11}$$

Proof. This proof is omitted due to space limitations, but it is available in the extended arXiv version of this paper [32]. □

This Theorem gives us sufficient criteria for assuring convergence of our system to the target set, however, it depends on us being able to lower-bound, a priori, the singular values of our parameter matrices along a trajectory. We then look into ways of lower-bounding the quantity $\sigma_{\min}^2(Q) + \sigma_{\min}^2(P)$ as a function of our initialization to properly characterize the ISS property of this system for the simplified case where $n = m = 1$. To obtain that lower bound we study the undisturbed case and identify necessary and sufficient conditions for its convergence to the target set, which we use as guidelines to obtain a bound on the maximum admissible disturbance signal.

III. The Scalar and Vector Cases ($n = m = 1, k \geq 1$)

We assume along this section that $n = m = 1$, that the scalar $\bar{Y}$ is non-negative (all results are still valid if otherwise, but the characterization of stable and unstable sets swap), and that P and Q are row vectors in general (in the particular case where $k = 1$ they are scalars). For this simplified version of the problem, one can verify that the undisturbed dynamics of the parameters of the system is a nonlinear reparameterization of linear dynamics, that is

$$\begin{bmatrix} \dot{P} \\ \dot{Q} \end{bmatrix} = (\bar{Y} - PQ^\top) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} P \\ Q \end{bmatrix}, \tag{12}$$

since $\bar{Y} - PQ^\top$ is a scalar function. This means that the trajectories on the state-space will look the same as the linear system $\dot{P} = Q$ and $\dot{Q} = P$ with an extra stable set whenever $\bar{Y} - PQ^\top = 0$ (our target set) and a change in direction if $\bar{Y} - PQ^\top < 0$. For the scalar case, we can draw the phase plane of our system, as in Fig. 1. To formalize this conjecture, we linearize the system around the origin, which results in

$$\begin{bmatrix} \dot{P} \\ \dot{Q} \end{bmatrix} = A_{\text{lin}}(P,Q), \tag{13}$$

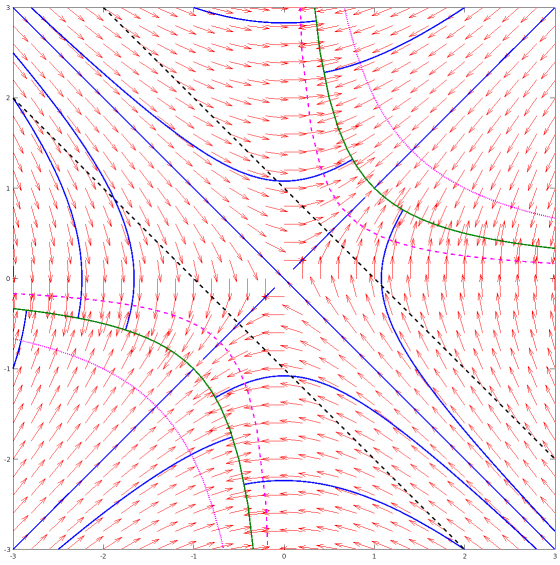

Fig. 1: Phase plane for the scalar case of the gradient flow dynamics. The trajectories followed by our solutions are the same as the 2D saddle, except for the inclusion of a new stable set whenever our nonlinear reparameterization $(\bar{Y} - PQ) = 0$ and a change in the direction of the trajectories when $(\bar{Y} - PQ) < 0$. In the figure there are a couple of different solutions for the system (in blue) as well as the borders of two possible invariant sets (black and pink) that guarantee for initial conditions in them that the system converges to the target set, given sufficiently bound disturbances U and V.

where $A_{\text{lin}} : \mathbb{R}^{2\times k} \to \mathbb{R}^{2\times k}$ is a linear operator on a matrix space. To write this in the familiar state space form we vectorize both sides of the equation, resulting in:

$$\operatorname{vec}\begin{pmatrix}\dot{P}\\ \dot{Q}\end{pmatrix} = \left(I_k \otimes \begin{bmatrix}0 & \bar{Y}\\ \bar{Y} & 0\end{bmatrix}\right)\operatorname{vec}\begin{pmatrix}P\\ Q\end{pmatrix} = \bar{A}_{\text{lin}}\operatorname{vec}\begin{pmatrix}P\\ Q\end{pmatrix}. \quad (14)$$

One can verify that $\bar{A}_{\text{lin}}$ has eigenvalues $+\bar{Y}$ and $-\bar{Y}$ with multiplicity k, and that an orthogonal basis of eigenvectors associated with the positive (resp. negative) eigenvalues is given by $\{e_i \otimes [1,\ 1]^\top\}_{i=1}^k$ (resp, $\{e_i \otimes [-1,\ 1]^\top\}_{i=1}^k$). Then, the following Lemma provides a link between the characterized local (linear) properties of the system around the origin and its global (nonlinear) behavior.

Proposition 1. *For any initial condition* $[P_0; Q_0]$ *such that*

$$\operatorname{vec}\left(\begin{bmatrix}P_0\\ Q_0\end{bmatrix}\right) \in \mathcal{S}^- := \textit{span}\left(\left\{e_i \otimes \begin{bmatrix}-1\\ 1\end{bmatrix}\right\}_{i=1}^{k}\right),$$

the solution of (12) *converges to the saddle point* $[P(t); Q(t)] = 0$. *On the other hand, for every initial condition* $[P_0; Q_0]$ *such that* $\operatorname{vec}([P_0; Q_0]) \notin \mathcal{S}^-$, *the solution of* (12) *converges to the target set* $\mathcal{T}$.

Proof. This proof is omitted due to space limitations, but it is available in the extended arXiv version of this paper [32]. □

Remark 1. For the scalar case, $\mathcal{S}^-$ and $\mathcal{S}^+$ (defined the same as $\mathcal{S}^-$ but for the unstable eigenvectors) correspond to the lines $P - Q = 0$ and $P + Q = 0$ respectively, as highlighted in Fig. 1. The condition that $F(\bar{a}, \bar{b}) < 0$ can be understood as being to the southwest of the lower hyperbola or the northeast of the higher hyperbola, while $F(\bar{a}, \bar{b}) > 0$ is the region in between the two hyperbolas. As mentioned before, the set where $[P; Q]$ is spanned by the eigenvectors of our linearization associated with the negative eigenvalues is the line $P + Q = 0$ and is the only set in the state space that converges to the saddle point $[P; Q] = 0$. In Fig. 1, the dashed black lines are defined, for some $c \in \mathbb{R}_+$, by the equation $|P + Q| = c$ which intuitively measures the magnitude of our projection in $\mathcal{S}^+$, while the dashed pink lines are defined by $PQ = c$ which in some sense measures the distance between a point and our target set along our vector field. While both $|P + Q| > c$ and $PQ > c$ can be shown to be forward invariant, we focus on the black lines for this section, in hopes of being able to more easily generalize this set for the general case in the future.

Considering the minimum distance to $\mathcal{S}^-$ (that is the norm of the projection into $\mathcal{S}^+$) as a possible invariant set for our problem gives

$$\left\|\textbf{vec}^{-1}\left(\left(I_k \otimes \begin{bmatrix}1\\ 1\end{bmatrix}\right)^\top \cdot \operatorname{vec}\left(\begin{bmatrix}P\\ Q\end{bmatrix}\right)\right)\right\|_2^2 = \|P + Q\|_2^2. \quad (15)$$

For some $\alpha > 0$, define the set

$$\mathcal{R}_\alpha = \{P, Q \in \mathbb{R}^k \mid \|P + Q\|_2^2 \geq \alpha^2\}, \quad (16)$$

for which we can state the following theorem:

Theorem 2. *For any* $\alpha \in [0, 2\sqrt{\bar{Y}})$, *the set* $\mathcal{R}_\alpha$ *is forward-invariant under the gradient flow dynamics if*

$$\|U\|_2 + \|V\|_2 \leq \frac{1}{\sqrt{2}}|\alpha|(\bar{Y} - \frac{\alpha^2}{4}). \quad (17)$$

Moreover, if $(P, Q) \in \mathcal{R}_\alpha$, *then* $PP^\top + QQ^\top = \sigma(P)^2 + \sigma(Q)^2 \geq \alpha^2/2$.

Proof. This proof is omitted due to space limitations, but it is available in the extended arXiv version of this paper [32]. □

This theorem allows us to rewrite the previous lower-bound dissipation inequality, for a solution initialized in $\mathcal{R}_\alpha$ for some $\alpha \in [0, 2\sqrt{\bar{Y}})$, as

$$\dot{\mathcal{L}}(P, Q, U, V) \leq -\mathcal{L}(P, Q)\frac{\alpha^2}{2} + \frac{1}{2}\left\|\begin{bmatrix}U\\ V\end{bmatrix}\right\|_2^2, \quad (18)$$

which is strictly negative until we reach the target set, assuming our disturbances and initialization respect the bounds and conditions given by Theorem 2.

The results from Theorem 2 recover the ISS property for the overparameterized gradient flow when restricted to some $\mathcal{R}_\alpha$. Notice that, unlike the non-overparameterized case discussed in [23], the condition for invariance of $\mathcal{R}_\alpha$ imposes a bound on the magnitude of the disturbance. Also, notice

that this is a worst-case analysis, that is we assume that the disturbance will always push our system toward the closest point in the bad set $\mathcal{S}^-$.

Furthermore, any initialization such that $P_0Q_0^\top > \bar{Y}$ is guaranteed to never converge to the bad set. This can be geometrically verified for the scalar case since $P_0Q_0^\top > \bar{Y}$ implies that the system is initialized either to the northeast of the positive hyperbola or to the southwest of the negative one. To guarantee robustness, however, we still require the bound on the magnitude of the disturbance given by Theorem 2 for $\alpha = 2\sqrt{\bar{Y}}$.

Therefore, we effectively impose a bound on the magnitude of the maximum acceptable disturbance for the system, regardless of initialization, indicating a compromise in terms of the robustness of adopting an overparameterized formulation. For the remainder of this paper, we look at the extra complications that arise from generalizing this analysis to the general case.

IV. The General Case

For the general case, where m and n are arbitrary, we have to deal with a matrix ODE for the dynamics of the parameters. One immediate problem from this is the existence of spurious equilibria besides the origin. To formally show this, we first state the following intermediate Lemma 1 followed by Theorem 3 which characterize the system's equilibria.

Lemma 1. *Given two matrices $A \in \mathbb{R}^{p\times o}$ and $B \in \mathbb{R}^{q\times o}$ for $p, q, o \in \mathbb{N}$ with $q \geq o$, the following two statements are equivalent*

1) $AB^\top = 0$;
2) *There exist orthogonal matrices Ψ_A, Φ, and Ψ_B, and rectangular diagonal matrices with non-negative diagonal elements Σ_A and Σ_B, such that*

$$A = \Psi_A \Sigma_A \Phi^\top \tag{19}$$

and

$$B = \Psi_B \Sigma_B \Phi^\top \tag{20}$$

are SVDs of A and B, and $\Sigma_A \Sigma_B^\top = 0$.

Furthermore, in 2) we can write Σ_A and Σ_B as

$$\Sigma_A = \begin{bmatrix} \bar{\Sigma}_A & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$\Sigma_B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{\Sigma}_B & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where $\bar{\Sigma}_A$ and $\bar{\Sigma}_B$ are diagonal matrices whose main diagonal elements are the nonzero singular values of A and B respectively.

Proof. This proof is omitted due to space limitations, but it is available in the extended arXiv version of this paper [32]. □

Theorem 3. *For the dynamics given by (8), and an arbitrary set of parameters P and Q, the following are equivalent*

1) *$Z = [P; Q]$ is an equilibrium point of f_Z, that is $(\bar{Y} - PQ^\top)Q = 0$ and $(\bar{Y} - PQ^\top)^\top P = 0$;*
2) *(a) There exist an SVD of $\bar{Y} - PQ^\top = \Psi\Sigma\Phi^\top$, and orthogonal matrices Γ_P and Γ_Q such that $\Psi\Sigma_P\Gamma_P^\top$ and $\Phi\Sigma_Q\Gamma_Q^\top$ are SVDs of P and Q; and (b) $\Sigma\Sigma_Q = 0$ and $\Sigma^\top\Sigma_P = 0$.*

Proof. This proof is omitted due to space limitations, but it is available in the extended arXiv version of this paper [32]. □

Remark 2. In the full version of the paper, we will use this theorem as a way to analyze the behavior of the system locally around each equilibrium point, by exploring the given structure of the equilibria. This characterization makes it possible to write the Jacobian matrix of the system as a function of the singular value decomposition presented.

Remark 3. In [13] the authors characterize the equilibria of the system for the symmetric case (when $P = Q$ and $\bar{Y}$ is positive definite). Reinterpreting statement 2) of Theorem 3 for this case gives that $\bar{Y} - PP^\top = \Psi\Sigma\Psi^\top$ and $P = \Psi\Sigma_P\Gamma_P$, which in turn imposes that $\bar{Y} = \Psi\Sigma_Y\Psi^\top$. Using these SVDs, the equation $(\bar{Y} - PP^\top)P = 0$ becomes $(\Sigma_Y - \Sigma_P\sigma_P^\top)\Sigma_P = 0$ which holds if for every i either $\sigma_{i,Y} = \sigma_{i,P}^2$ or $\sigma_{i,P} = 0$, which recovers their original result.

The presence of multiple equilibria in our dynamics means that the approach done for the vector case does not immediately translate to the general case. Even so, in [14] the authors present interesting results for the convergence of the system to the target set for the undisturbed case. They guarantee convergence to the target set for any initial condition except for a set of dimension zero, which indicates possible robustness to disturbances, as long as a minimum distance from such set can be guaranteed from the initial conditions. From Theorem 1 of [14] we have an exponential bound on our cost function $\mathcal{L}(P, Q)$ for the undisturbed case, where the exponential constant is a function of the eigenvalues of the imbalance matrix defined as $\Lambda = P^\top P + Q^\top Q$. In the paper the authors explore the invariance of Λ along any trajectory to formulate this bound, however once we allow for disturbances in the computation of the gradient, the invariance property of Λ is lost, as can be seen by computing $\dot{\Lambda}(t)$ for the disturbed dynamics

$$\begin{aligned} \dot{\Lambda} &= \dot{P}^\top P + P^\top \dot{P} - \dot{Q}^\top Q - Q^\top \dot{Q} \\ &= \quad Q^\top(\bar{Y} - PQ^\top)^\top P + U^\top P \\ &\quad + P^\top(\bar{Y} - PQ^\top)Q + P^\top U \\ &\quad - P^\top(\bar{Y} - PQ^\top)Q - V^\top Q \\ &\quad - Q^\top(\bar{Y} - PQ^\top)^\top P - Q^\top V \\ &= U^\top P + P^\top U - V^\top Q - Q^\top V, \end{aligned} \tag{21}$$

and as such, the results derived in [14] are not immediately applicable. Nonetheless, the fact that we can guarantee exponential convergence with some "margin" that depends

only on our initialization for the undisturbed case intuitively indicates our trajectories might accept some level of disturbance while still converging to our desired target set, motivating further works on this subject.

V. Conclusions and Future Works

In this paper, we formulated the overparameterized linear regression problem as a matrix factorization and presented a dissipation-like inequality for the general problem when the parameters are trained through an uncertain gradient flow. The bound obtained, however, does not guarantee convergence to the target set $\mathcal{T}$ for any initial condition, which prompts the search for invariant subsets of the state space in which the system can be shown to be ISS.

This publication focuses on the solution of the problem for the case where the neural network has a single input and a single output. In this situation, the parameter matrices reduce to vectors and the analysis is significantly simplified. We characterize the behavior of the system when training through exact gradient flow and formulate necessary and sufficient conditions for its convergence to the target set. We then use those conditions as a guideline to formulate sufficient conditions on the initialization of our system and on the maximum admissible disturbance on the estimation of the gradient that if satisfied guarantees that the system is ISS.

We finish the paper with a brief discussion about the general case. We show that in general, the dynamics become significantly more complicated with the appearance of multiple sets of spurious equilibria. While there are results in the literature that guarantee convergence to the target set for the general case, their extension to the disturbed case is not straightforward.

Current research being conducted by the authors for the general case indicates that despite its more complex dynamics, we can still predict the behavior of the general case based on its linearization around the origin, similar to how we solve the problem on the vector case. We are currently looking into how we can use this knowledge to characterize regions of our state space where the ISS property is guaranteed in general.

We can, however, conclude that by opting for an overparameterized formulation, our system ceases to be ISS for the whole state space when subject to gradient flow, contrary to the non-overparameterized case, as shown in [23]. This indicates a trade-off on using an overparameterized formulation for performing linear regression, even if it is eventually shown that it can be circumvented by a knowledgeable choice of initial condition.

References

[1] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[2] M. Belkin, S. Ma, and S. Mandal, "To understand deep learning we need to understand kernel learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 541–549.

[3] A. Sanyal, P. K. Dokania, V. Kanade, and P. H. Torr, "How benign is benign overfitting?" *arXiv preprint arXiv:2007.04028*, 2020.

[4] Y. Cao, Z. Chen, M. Belkin, and Q. Gu, "Benign overfitting in two-layer convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 237–25 250, 2022.

[5] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[6] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[7] S. Frei, N. S. Chatterji, and P. Bartlett, "Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data," in *Conference on Learning Theory*. PMLR, 2022, pp. 2668–2703.

[8] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[11] H. Mohammadi, M. Razaviyayn, and M. R. Jovanović, "On the stability of gradient flow dynamics for a rank-one matrix approximation problem," in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 4533–4538.

[12] P. Chen and H.-H. Chen, "Accelerating matrix factorization by overparameterization." in *DeLTA*, 2020, pp. 89–97.

[13] S. Tarmoun, G. Franca, B. D. Haeffele, and R. Vidal, "Understanding the dynamics of gradient flow in overparameterized linear models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 153–10 161.

[14] H. Min, S. Tarmoun, R. Vidal, and E. Mallada, "On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7760–7768.

[15] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *arXiv preprint arXiv:1805.12152*, 2018.

[16] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[17] A. Javanmard, M. Soltanolkotabi, and H. Hassani, "Precise tradeoffs in adversarial training for linear regression," in *Conference on Learning Theory*. PMLR, 2020, pp. 2034–2078.

[18] A. H. Ribeiro and T. B. Schön, "Overparameterized linear regression under adversarial attacks," *IEEE Transactions on Signal Processing*, vol. 71, pp. 601–614, 2023.

[19] Y. Min, L. Chen, and A. Karbasi, "The curious case of adversarially robust models: More data can help, double descend, or hurt generalization," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 129–139.

[20] D. Yin, R. Kannan, and P. Bartlett, "Rademacher complexity for adversarially robust generalization," in *International conference on machine learning*. PMLR, 2019, pp. 7085–7094.

[21] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[22] K. Scaman and C. Malherbe, "Robustness analysis of non-convex stochastic gradient descent using biased expectations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 377–16 387, 2020.

[23] E. D. Sontag, "Remarks on input to state stability of perturbed gradient flows, motivated by model-free feedback control learning," *Systems & Control Letters*, vol. 161, p. 105138, 2022.

[24] Y. Chitour, Z. Liao, and R. Couillet, "A geometric approach of gradient descent algorithms in linear neural networks," *Mathematical Control and Related Fields*, vol. 13, no. 3, pp. 918–945, 2023. [Online]. Available: /article/id/6269e2592d80b75dc9b8cbd4

[25] K. Kawaguchi, "Deep learning without poor local minima," *Advances in neural information processing systems*, vol. 29, 2016.

[26] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural networks*, vol. 2, no. 1, pp. 53–58, 1989.

[27] P. Monzón and R. Potrie, "Local and global aspects of almost global stability," in *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE, 2006, pp. 5120–5125.

[28] I. Panageas and G. Piliouras, "Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions," *arXiv preprint arXiv:1605.00405*, 2016.

[29] S. S. Du, W. Hu, and J. D. Lee, "Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced," *Advances in neural information processing systems*, vol. 31, 2018.

[30] H. Schaeffer and S. G. McCalla, "Extending the step-size restriction for gradient descent to avoid strict saddle points," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1181–1197, 2020.

[31] A. Eftekhari, "Training linear neural networks: Non-local convergence and complexity results," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2836–2847.

[32] A. C. B. de Oliveira, M. Siami, and E. D. Sontag, "On the ISS property of the gradient flow for single hidden-layer neural networks with linear activations," *arXiv preprint arXiv:2305.09904*, 2023.