On the Convergence of Overparameterized Problems: Inherent Properties of the Compositional Structure of Neural Networks

Arthur Castello B. de Oliveira A.CASTELLO@NORTHEASTERN.EDU
Dhruv D. Jatkar JATKAR.D@NORTHEASTERN.EDU
Eduardo D. Sontag E.SONTAG@NORTHEASTERN.EDU
Northeastern University, 805 Columbus Ave, Boston, MA 02120

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

This paper investigates how the compositional structure of neural networks shapes their optimization landscape and training dynamics. We analyze the gradient flow associated with overparameterized optimization problems, which can be interpreted as training a neural network with linear activations. Remarkably, we show that the global convergence properties can be derived for any cost function that is proper and real analytic. We then specialize the analysis to scalar cost functions, where the geometry of the landscape can be fully characterized. In this setting, we demonstrate that key structural features – such as the location and stability of saddle points – are universal across all admissible costs, depending solely on the overparameterized representation rather than on problem-specific details. Moreover, we show that convergence can be arbitrarily accelerated depending on the initialization, as measured by an imbalance metric introduced in this work. Finally, we discuss how these insights may generalize to neural networks with sigmoidal activations, showing through a simple example that certain geometric and dynamical properties persist beyond the linear case.

Keywords: Neural Networks, Optimization, Gradient Methods, Overparameterization.

1. Introduction

The widespread success of artificial intelligence (AI) in solving complex tasks has ignited interest in understanding the theoretical underpinnings of modern learning systems Belkin et al. (2019); Bartlett et al. (2020); Arora et al. (2019); Du et al. (2019).

Among the many approaches to this question, a particularly fruitful direction studies linear neural networks as tractable models for analyzing how the compositional structure of deep networks affects their optimization landscape and training dynamics. The simplicity of linear activations allows one to isolate the geometric and dynamical consequences of composition itself, independent of nonlinearities Bah et al. (2022); Arora et al. (2019); Kawaguchi (2016); Chitour et al. (2022); Min et al. (2021); Menon (2023); Min et al. (2023); de Oliveira et al. (2023); De Oliveira et al. (2024); de Oliveira et al. (2025b).

A rich body of work has characterized the behavior of linear networks in linear regression and related problems, revealing elegant structures in their loss landscapes and convergence dynamics Bah et al. (2022); Kawaguchi (2016); Arora et al. (2019); Chitour et al. (2022); Min et al. (2021). More recent research has extended some of these results to broader classes of cost functions Menon (2023); Min et al. (2023) and to nonconvex, control-related optimization problems de Oliveira et al. (2025b). Notably, Min et al. (2023) demonstrates an initialization-dependent exponential acceleration of gradient-flow solutions for strongly convex objectives, indicating a possible advantage of

adopting an overparameterized formulation over regular gradient flow. In de Oliveira et al. (2025b) we further explored this property in a problem that is non-convex and not gradient dominant – the policy optimization problem for the linear quadratic regulator (LQR). Furthermore, in the same paper we showed that all the good properties previously demonstrated for the specific case of linear regression also hold for this more complex nonconvex problem. Building on these observations, in Wafi et al. (2025) we later demonstrated that overparameterization induces not merely quantitative acceleration but also qualitative improvements in the convergence profile of gradient-flow solutions.

The present paper seeks to determine to what extent these convergence properties are inherent to the compositional structure of overparameterization itself, rather than to specific features of the underlying cost function. We show that for any proper, real-analytic cost, the corresponding overparameterized gradient flow – interpreted as training a deep linear network – admits invariant quantities that partition the parameter space into disjoint invariant manifolds. Through this geometric property we prove that gradient-flow trajectories always converge to a critical point of the problem (despite the overparameterized cost not being proper anymore), and almost everywhere to critical points of the original problem if the linear neural network has a single hidden layer. We emphasize that our results are for a much more general class of functions than what exists in the literature currently, strongly indicating that the aforementioned properties are inherent to the overparameterized structure rather than to the problem under consideration. For scalar costs, we prove that the geometry of the center–stable manifolds introduced by overparameterization is universal, depending only on the compositional structure and not on the specific cost, while the rate of convergence depends on an initialization-imbalance measure introduced here.

We emphasize that the results in this paper are a significant step towards demonstrating the advantages of the compositional structure of neural networks when compared to other forms of overparameterization. This is a point we began to make in de Oliveira et al. (2025b) but take further in this paper by considering a very general set of problems. This point is further demonstrated at the final section where we show how the geometric intuitions derived for linear activations can be translated to sigmoidal neural networks through a simple example. Furthermore, the advantages of overparameterized formulations proven here for scalar costs might prove beneficial in practice for problems whose optimal solutions are known to be scalar mappings – such as the LQR explored in de Oliveira et al. (2025b).

This paper is organized as follows. We begin in Section 2 by formally defining the class of optimization problems under scrutiny, and the exact structure of an overparameterized formulation. We then discuss the invariant measure, how it is a consequence of the compositional structure, and why it is important for studying the convergence of gradient-flow solutions. This result allows us to state the two main results regarding convergence of solutions for overparameterized problems. Next we move to scalar optimization problems in Section 3 proving the aforementioned problem-independent structure of the parameter space and the initialization-based acceleration of solutions. We then explore in Section 4 how the results of this paper might be extended to nonlinear neural networks, proving, for one example, the existence of an invariant quantity even for this nonlinear case. Finally, in Section 5 we review the results of the paper and their importance for understanding the training behavior of neural networks. Most proofs are deferred to the Appendix, but each result is followed by a sketch of its proof.

2. Overparameterization and neural networks

Let $f : \mathbb{W} \to \mathbb{R}$ be a proper (i.e. the preimage of each compact set is compact), bounded below, continuously differentiable function with a Lipschitz gradient, where $\mathbb{W} \subseteq \mathbb{R}^{n\times n}$ is a simply connected set. For such f, define an optimization problem as

$$\underset{W\in\mathbb{W}}{\text{minimize}} \quad f(W). \tag{1}$$

Assume further that the minimum value of the function exists and is given by $\underline{f} := \inf_{W\in\mathbb{W}} f(W)$, and solving equation 1 means finding any point $W^* \in \mathscr{T} := \{W \in \mathbb{W} \mid f(W) = \underline{f}\}$.

Gradient methods are a popular approach to finding a candidate solution to equation 1. Specifically, a gradient-flow solution consists of solving the following initial value problem

$$\dot{W} = -\nabla f(W), \quad W(0) = W_0 \tag{2}$$

for some $W_0 \in \mathbb{W}$. It is easy to verify that properness of f implies pre-compactness of solutions of equation 2, which by the Krasovskii-LaSalle's principle implies convergence of solutions to $\mathscr{G}_f := \{W \in \mathbb{W} \mid \nabla f(W) = 0\}$. Furthermore, one can verify that $\mathscr{T} \subseteq \mathscr{G}_f$, but $\mathscr{G}_f \not\subseteq \mathscr{T}$ in general, and thus further assumptions are often required in practice. To simplify the analysis in this paper, we also assume that for all $W \in \mathscr{G}_f$, $\text{rank}(W) = n$. This assumption simplifies a few steps in the proofs and can be circumvented by using singular value decomposition and by tracking the relevant base of singular vectors, albeit at significant cost of complexity in the algebraic steps of the proof.

For a given width k and depth N, let $k > n$, $W_1 \in \mathbb{R}^{k\times n}$, $W_N \in \mathbb{R}^{n\times k}$ and $W_i \in \mathbb{R}^{k\times k}$. Then we define an overparameterized formulation of equation 1 as

$$\underset{(W_1,\dots,W_N)\in\mathbb{W}}{\text{minimize}} \quad g(W_1,\dots,W_N), \tag{3}$$

where $g(W_1,\dots,W_N) := f(W_N W_{N-1}\cdots W_2 W_1)$ and $(W_1,\dots,W_N) \in \mathbb{W}$ is an abuse of notation to mean

$$(W_1,\dots,W_N) \in \left\{(W_1,\dots,W_N) \in \mathbb{R}^{k\times n} \times \left(\mathbb{R}^{k\times k}\right)^{N-2} \times \mathbb{R}^{n\times k} \;\middle|\; W_N W_{N-1}\cdots W_2 W_1 \in \mathbb{W}\right\}.$$

We highlight that this is a very general framework, covering a broad class of optimization problems, with the assumptions made to f being standard for "well-behaved" optimization problems. Notice, however, that despite f being a proper function, g is not. This happens because for any $\eta \in \mathbb{R}$, $g(W_1, W_2) = g(\eta W_1, (1/\eta)W_2)$ which introduces unbounded directions in the state space for which the cost remains bounded despite the parameter norms becoming unbounded.

For simplicity of notation define $\mathbf{W}(W) := W_N W_{N-1}\cdots W_2 W_1$ for any given $W = (W_1,\dots,W_N)$, where the dependency of $\mathbf{W}$ on W can be omitted if it is clear from context.

Notice that $\mathbf{W}$ is the resulting expression of a feedforward neural network with weight matrices W, linear activations, $N-1$ hidden layers, and width of k. Fig. 1 illustrates this interpretation, showing how the compositional structure of feedforward neural networks result in the proposed overparameterized formulation.

This is a well-studied formulation in the literature Bah et al. (2022); Chitour et al. (2022); de Oliveira et al. (2023); De Oliveira et al. (2024); Eftekhari (2020); Kawaguchi (2016); Min et al.

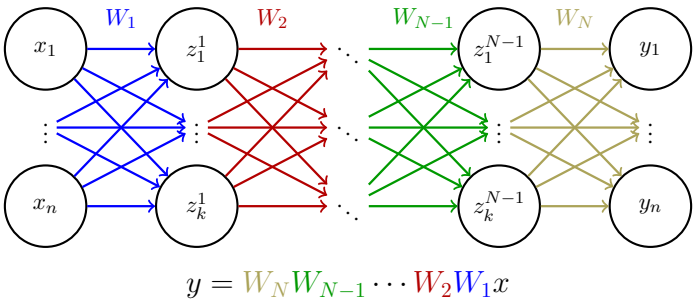

Figure 1: Depiction of a linear neural network.

(2021, 2023), with many results characterizing the optimization landscape for different applications, as well as providing guarantees for gradient methods. An *overparameterized gradient flow* is defined as a set of ODEs given by

$$\dot{W}_i := -\nabla_{W_i} g(W_1, \ldots, W_N) = -(W_N \ldots W_{i+1})^\top \nabla f(\mathbf{W})(W_{i-1} \ldots W_1)^\top, \tag{4}$$

for $i = 1 \ldots N$. As an immediate consequence of this structure for the gradient flow, we can state the following key result:

Definition 1 *For any state $W \in \mathbb{W}$ of the overparameterized gradient flow equation 4, the* invariant *is defined as $\mathscr{C} := (\mathcal{C}_1, \ldots, \mathcal{C}_{N-1}) \in \left(\mathbb{R}^{k \times k}\right)^{N-1}$ where*

$$\mathcal{C}_i := W_i W_i^\top - W_{i+1}^\top W_{i+1} \tag{5}$$

Proposition 2 *The value of the invariant $\mathcal{C}$ along any solution of the overparameterized gradient flow equation 4 is invariant, i.e.*

$$\frac{d}{dt}\mathscr{C} = 0. \tag{6}$$

This result is independent of the cost function f and is a consequence of the compositional structure assumed for overparameterization. As such, it has been noted in the literature Menon (2023) and it has been a key result for proving the convergence of overparameterized gradient methods for different applications Kawaguchi (2016); Chitour et al. (2022); Min et al. (2023); de Oliveira et al. (2025b); Arora et al. (2019). This is a powerful result because it shows that the optimization landscape is "foliated" by invariant disjoint manifolds, as illustrated in Fig. 2. Through this result, we obtain global convergence properties of solutions by characterizing the same property for solutions in each of these manifolds, greatly simplifying the analysis.

Once we have established the invariant for any solution, we can leverage it to prove the following result:

Theorem 3 *Let $f : \mathbb{W} \to \mathbb{R}$ be any real analytic and proper cost function that attains its minimum for some point in the interior of $\mathbb{W}$. Consider its overparameterized optimization problem as given in equation 3. For any $W = (W_1, \ldots, W_N) \in \mathbb{W}$, a solution of equation 4 initialized at W exists for all time, remains in $\mathbb{W}$, and converges to a critical point of the cost $g(W_1, \ldots, W_N)$.*

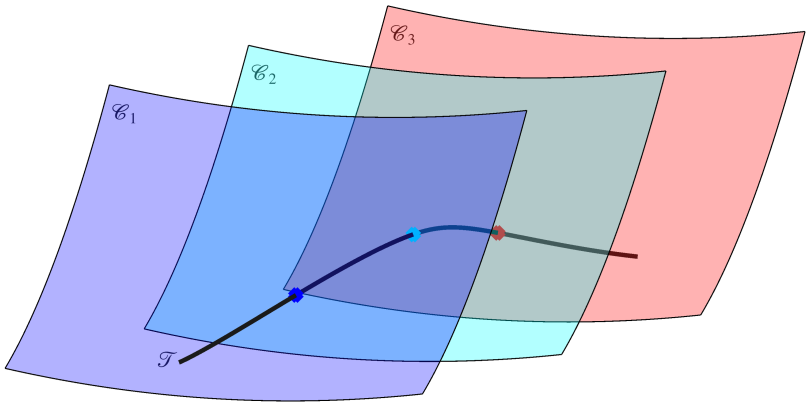

Figure 2: Illustration of the foliation of the state space of overparameterized optimization problems. The black curve illustrates a branch of the set of critical points given by $\mathbf{W}(W_1, \ldots, W_N) = W^*$, and the sets $\mathscr{C}_i$ are the invariant manifold of the dynamics. Notice that the manifolds are invariant and do not intersect with each other, but every point in the parameter space is within one of such manifolds, resulting in a "foliated" optimization landscape. Global properties of the training can, then, be shown by simply showing that local properties in the manifold hold for all manifolds.

We provide the proof in the appendix for completeness. However, it has been mentioned in the literature Menon (2023); Bah et al. (2022) that pre-compactness of $\mathbf{W}(W(t))$ is enough to guarantee existence and convergence of solutions – in this setup, properness of f is enough to guarantee this condition, despite g not being proper after reparameterization.

Despite this, notice that critical points of g are not necessarily critical points of f, as g introduces multiple saddle points via orthogonality conditions of the matrix products in equation 4. A better condition can be stated for the case where $N = 2$ as follows:

Theorem 4 *Let $f : \mathbb{W} \to \mathbb{R}$ be any real analytic proper cost function that attains its minimum for some point in the interior of $\mathbb{W}$. Consider its overparameterized optimization problem, as given in equation 3, with $N = 2$ (single hidden-layer case). Then, the overparameterized gradient-flow of equation 4 will converge to a point in $\mathscr{G}_f := \{(W_1, W_2) \in \mathbb{W} \mid \nabla f(\mathbf{W}(W_1, W_2)) = 0\}$ for all but a measure-zero set of initializations.*

The proof of this result is given in the appendix for completeness, but it follows from a previously published result of ours de Oliveira et al. (2025b) once it is established that for the single hidden-layer case all critical points introduced by overparameterization are *strict saddles* (i.e. a point at which the gradient of the cost is zero, but the Hessian has one strictly negative eigenvalue). Being able to extend this result to a general class of functions indicates that these properties are a consequence of the overparameterized structure choice and, to some degree, independent of the specific problem being solved.

The results above establish that overparameterization introduces a geometric structure that guarantees the convergence of gradient-flow solutions for any proper real-analytic cost function, despite the overparameterized cost no longer being proper. However, these results remain somewhat ab-

stract, as they describe convergence in terms of invariant manifolds and precompactness rather than explicit trajectories. To gain deeper intuition, it is instructive to examine a simpler setting in which all quantities can be written in closed form.

In the following section, we restrict attention to scalar cost functions overparameterized through linear neural networks with a single hidden layer. We call this the "vector" case, since it forces W_1 and W_2 to be vectors instead of matrices. This reduction preserves the essential compositional structure while allowing a full analytical characterization of the dynamics. Within this framework, we will explicitly describe (i) how convergence arises from the invariant geometry, (ii) how initialization imbalance affects the rate of convergence, and (iii) why these behaviors are universal across all proper analytic cost functions.

3. Analysis of the vector case

We consider a simplified version of the problem, where $\mathbb{W} \subseteq \mathbb{R}$ and $N = 2$ – this setup is equivalent to a single arbitrarily wide hidden layer linear neural network being trained to minimize a scalar cost. Along this section, the cost *is still assumed to be proper and real analytic*, but no further assumptions are necessary at this point. The simpler setup allows a more complete understanding of the gradient flow behavior, as we hope to explain next.

The overparameterized parameters are written as $w = (w_1, w_2) \in \mathbb{R}^{k\times 1} \times \mathbb{R}^{1\times k}$ and the parameter dynamics under gradient flow are given by

$$\begin{bmatrix} \dot{w}_1 \\ \dot{w}_2^\top \end{bmatrix} = -f'(w_2 w_1) \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2^\top \end{bmatrix}. \tag{7}$$

For this setup, one can identify that the dynamics are a simple nonlinear reparameterization of a linear saddle. This fact, together with the fact that f is still assumed to be proper and bounded below, allows for the following result:

Proposition 5 *Let $f : \mathbb{R} \to \mathbb{R}$ be proper, continuously differentiable and bounded below. Let $\mathscr{G}_f := \{(w_1, w_2) \in \mathbb{W} \mid f'(w_2 w_1) = 0\}$, and let $d(w_1, w_2) := \|w_1 - s w_2^\top\|$, with $s = \operatorname{sign}(f'(0))$. Then, any solution of the gradient flow initialized at a point $w^0 = (w_1^0, w_2^0)$ that satisfies $d(w_1^0, w_2^0) > 0$ will converge to a point in $\mathscr{G}_f$. In particular, $\lim_{t\to\infty} f'(w_2(t, w^0) w_1(t, w^0)) = 0$.*

The proof is provided in the appendix, but it leverages properness of f to show that solutions are pre-compact, and then argues that if $d(w_1^0, w_2^0) > 0$ then the solutions are initialized outside the center-stable manifold of the saddle at the origin, and thus can only converge to a point where the nonlinear reparameterization $f'(\cdot)$ is zero.

This is an important result because it indicates that the overparametrized problem can be seen as almost equivalent to the non-overparametrized problem in terms of convergence of gradient methods, except for a set of measure zero of initializations. Furthermore, the shape of this set is *independent of f* except for the sign of its derivative at the saddle-point, indicating we can predict the bad initializations independently from the problem being solved. This indicates a high transferability of skills between problems when training a neural network, providing another important clue to justify the widespread success of neural networks.

Despite that, the current results do not guarantee optimality of the solution, neither for the overparameterized formulation, nor for the original problem. Still, we will show next that ensuring

optimality of the gradient-flow of f is enough to do the same for g. For that, assume there exists a positive definite function $\alpha : \mathbb{R} \to \mathbb{R}_+$ for which f satisfies

$$|f'(w)| \geq \alpha(f(w) - \underline{f}), \qquad \forall w \in \mathbb{W}.$$

A function that satisfies such a condition is said to satisfy a positive-definite ($\mathcal{PD}$) Polyak-Łojasiewicz inequality ($\mathcal{PD}$-PŁI), which is a weaker version of the Polyak-Łojasiewicz inequality (PŁI) de Oliveira et al. (2025a). For the standard gradient flow, this condition guarantees asymptotic convergence of the cost to a global minimizer of f. For the overparameterized gradient flow, we can state the following corollary of Proposition 5:

Corollary 6 *If f satisfies a $\mathcal{PD}$-PŁI and $f(0) > \underline{f}$, then $\lim_{t\to\infty} f(w_2(t, w^0)w_1(t, w^0)) = \underline{f}$ if and only if $d(w_1^0, w_2^0) > 0$.*

As mentioned, this result is an immediate consequence of Proposition 5 and the fact that satisfying a $\mathcal{PD}$-PŁI implies that all critical points of f are global minimizers.

3.1. Accelerated Convergence

Another useful property of overparameterization is the imbalance-based acceleration of solutions. This can be characterized in general for the vector case as follows

Proposition 7 *Assume $f : \mathbb{W} \to \mathbb{R}$ is a proper, real analytic, bounded below, scalar function with minimum given by $\underline{f} = \min_{w\in\mathbb{W}} f(w)$. Assume further that $f(0) > \underline{f}$. Let $\overline{w}, \widetilde{w} \in \mathbb{R}^{k\times 1} \times \mathbb{R}^{1\times k}$ be two points $\overline{w} = (\overline{w}_1, \overline{w}_2)$ and $\widetilde{w} = (\widetilde{w}_1, \widetilde{w}_2)$ such that*

- $\mathbf{w}(\overline{w}) := \overline{w}_2\overline{w}_1 = \widetilde{w}_2\widetilde{w}_1 =: \mathbf{w}(\widetilde{w}) \in \mathbb{W}$;
- $c(\overline{w}) := 2\mathit{trace}(\mathscr{C}(\overline{w})^2) - \mathit{trace}(\mathscr{C}(\overline{w}))^2 > 2\mathit{trace}(\mathscr{C}(\widetilde{w})^2) - \mathit{trace}(\mathscr{C}(\widetilde{w}))^2 =: c(\widetilde{w})$;

then, for all $t > 0$ it holds that $g(w_1(t, \overline{w}), w_2(t, \overline{w})) \leq g(w_1(t, \widetilde{w}), w_2(t, \widetilde{w}))$. If additionally we impose that $f'(\mathbf{w}(\overline{w})) = f'(\mathbf{w}(\widetilde{w})) \neq 0$, then we can strengthen the result to $g(w_1(t, \overline{w}), w_2(t, \overline{w})) < g(w_1(t, \widetilde{w}), w_2(t, \widetilde{w}))$ for all $t > 0$.

This result is an extension of a similar result given in de Oliveira et al. (2025b) and shows that two solutions initialized at the same "point" as measured by $\mathbf{w}(\cdot)$, will converge at different rates, with one being strictly faster than the other if it has a larger value of "imbalance" as measured by $c(\cdot)$.

This is a known result in different overparameterized scenarios, with Min et al. (2023) characterizing it for any overparameterized problem whose original cost is gradient dominant. The result we present, however, is true for any proper real-analytic scalar cost, illustrating how the imbalance-based acceleration is a property of the overparameterized structure rather than of the optimization problem itself.

4. Possible extension to sigmoidal neural networks – A proof-of-concept

We finish our discussion with a brief overview of how these results can be extended to sigmoidal neural networks. As a motivating example we study a scalar factorization problem. Let $f(w) =$

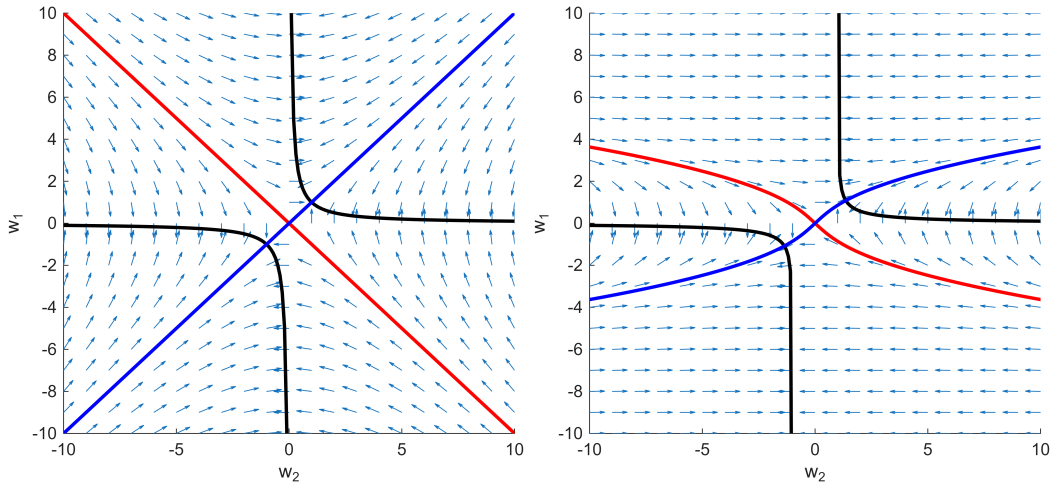

Figure 3: Illustration of the effect of sigmoidal activations to the optimization landscape of neural network training for factorization problems. The left figure displays the optimization landscape for the scalar factorization problem trained with linear neural networks, while the right one considers the same problem, but with sigmoidal networks. In solid black are displayed the target sets (global optima) for each problem; in red the center-stable manifolds of the saddle at the origin; and in blue its unstable manifold. Notice that the qualitative description of the parameter space remains unchanged: measure zero center-stable manifold for the saddle and almost everywhere convergence to the target. Despite that, notice that the center stable and unstable manifolds on the left figure segment the parameter space into four "equivalent" invariant subspaces, while on the right figure they segment the parameter space into four regions but of two different "types".

$(1-w)^2$, and let the overparameterized parameters be $w_1, w_2 \in \mathbb{R}$ (i.e. $k = 1$). Consider the following sigmoidal activation function

$$\sigma(z) = \frac{z}{\sqrt{1+z^2}}. \tag{8}$$

The overparameterized cost function is, then, written as $g(w_1, w_2) = (1 - w_2\sigma(w_1))^2$, and the overparameterized gradient-flow becomes

$$\dot{w}_1 := -\frac{\partial g}{\partial w_1} = 2(1 - w_2\sigma(w_1))\sigma'(w_1)w_2 = 2\frac{w_2\sqrt{1+w_1^2} - w_2^2 w_1}{(1+w_1^2)^2} \tag{9a}$$

$$\dot{w}_2 := -\frac{\partial g}{\partial w_2} = 2(1 - w_2\sigma(w_1))\sigma(w_1) = 2\frac{w_1\sqrt{1+w_1^2} - w_2 w_1^2}{1+w_1^2}. \tag{9b}$$

For this system, we can characterize an invariant quantity as follows:

Proposition 8 *For the system given in equation 9, the following quantity is invariant along any solutions:*

$$\mathscr{C}(w_1, w_2) := w_2^2 - \frac{1}{2}(1+w_1^2)^2 \tag{10}$$

The proof of this proposition is presented in the appendix, but it is purely algebraic. The more interesting consequence of this fact is that by characterizing this invariant quantity, many of the qualitative conclusions presented for linear neural networks will still hold. For example, since $\mathcal{C}$ is invariant, then one can compute $\mathcal{C}(0,0) = -1$ and find all (w_1, w_2) that satisfy $\mathcal{C}(w_1, w_2) = -1$. This results in the following two curves

$$w_2 = \pm w_1 \sqrt{\frac{2 + w_1^2}{2}} \tag{11}$$

which characterize the center-stable and unstable manifolds of the saddle at the origin. The parallel becomes even more evident when comparing the phase portrait of this problem with and without the sigmoidal activation, both of which we present in Fig. 3.

Notice from Fig. 3 that we can also characterize necessary and sufficient conditions for optimality of the sigmoidal gradient flow solution, but also that the two phase portraits are significantly distinct, with the sigmoidal phase portrait having two distinct "types" of regions defined by the center-stable and unstable manifolds of the saddle, while the linear phase portrait has four equivalent regions.

5. Conclusion

In this paper, we analyzed how the compositional structure of neural networks influences their optimization landscape and convergence behavior in a broad class of problems. To isolate the effects of composition from those of nonlinear activations, we focused on linear neural networks as a canonical model. We showed that the convergence guarantees traditionally associated with overparameterized linear regression/factorization extend, perhaps surprisingly, to any optimization problem whose cost function is proper and real analytic. This finding reveals that the favorable convergence behavior of deep linear models is not tied to specific data structures but is instead a structural consequence of composition itself, indicating favorable properties of the specific structure of feed-forward neural networks.

We then specialized our analysis to scalar optimization problems under overparameterization, where the dynamics can be fully characterized. In this setting, we demonstrated that the center–stable manifold associated with the saddle introduced by overparameterization is universal across all proper real-analytic cost functions. Its geometry depends solely on the overparameterized structure of the parameter space and not on the particular problem instance. This universality highlights a remarkable degree of generality in neural network training: the qualitative features of the learning dynamics arise from the compositional architecture, rather than from problem-specific properties, further indicating favorable properties of the structure naturally adopted by feedforward neural networks. Furthermore, we showed that the convergence rate of the overparameterized gradient flow is strongly affected by the initialization. Specifically, solutions can be significantly accelerated or slowed down depending on the level of imbalance of the initialization, as quantified by a metric introduced in this work.

Finally, we explored potential extensions to nonlinear (sigmoidal) neural networks. In the simplest nontrivial case, we showed that the qualitative structure of the parameter space persists: a strict saddle at the origin and a center–stable manifold of measure zero. We also derived an invariant quantity that remains conserved along trajectories, suggesting that the geometric and dynamical properties uncovered in the linear case may extend, to some degree, to nonlinear settings. These

results demonstrate how the conclusions of this work can be extended to less trivial neural network architectures, and provide a foundation for future work aimed at understanding how compositional structures shape not only convergence and stability but also emergent properties such as feature learning.

Acknowledgments

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

The authors would also like to acknowledge Prof. Shahriar Talebi's contributions to discussions regarding sigmoidal neural networks, and our ongoing efforts to extend the results of this paper to nonlinear activations in the near future.

References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, March 2022. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa039. URL https://academic.oup.com/imaiai/article/11/1/307/6127129.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Yacine Chitour, Zhenyu Liao, and Romain Couillet. A geometric approach of gradient descent algorithms in linear neural networks. *Mathematical Control and Related Fields DOI:10.1007/s10107-023-01937-5*, 2022.

Arthur Castello B de Oliveira, Milad Siami, and Eduardo D Sontag. Dynamics and perturbations of overparameterized linear neural networks. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 7356–7361. IEEE, 2023.

Arthur Castello B De Oliveira, Milad Siami, and Eduardo D Sontag. Remarks on the gradient training of linear neural network based feedback for the LQR problem. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 7846–7852. IEEE, 2024.

Arthur Castello B. de Oliveira, Leilei Cui, and Eduardo D. Sontag. Remarks on the polyak-lojasiewicz inequality and the convergence of gradient systems, 2025a. URL https://arxiv.org/abs/2503.23641.

Arthur Castello B de Oliveira, Milad Siami, and Eduardo D Sontag. Convergence analysis of gradient flow for overparameterized LQR formulations. *Automatica*, 182:112504, 2025b.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2836–2847. PMLR, November 2020. URL https://proceedings.mlr.press/v119/eftekhari20a.html. ISSN: 2640-3498.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/f2fc990265c712c49d51a18a32b39f0c-Abstract.html.

Govind Menon. The geometry of the deep linear network. In *Symposium on Probability and Stochastic Processes*, pages 1–47. Springer, 2023.

Hancheng Min, Salma Tarmoun, René Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *International Conference on Machine Learning*, pages 7760–7768. PMLR, 2021.

Hancheng Min, René Vidal, and Enrique Mallada. On the convergence of gradient flow on multi-layer linear models. In *International Conference on Machine Learning*, pages 24850–24887. PMLR, 2023.

Moh Kamalul Wafi, Arthur Castello B de Oliveira, and Eduardo D Sontag. On the (almost) global exponential convergence of the overparameterized policy optimization for the LQR problem. *arXiv preprint arXiv:2510.02140*, 2025.

Appendix A. Proofs

A.1. Proof of Proposition 2

This is a well-known result in overparameterized problems, having been proven in multiple other previous papers Bah et al. (2022); Menon (2023); Arora et al. (2019). We include the proof here for completeness. Notice that $\dot{\mathscr{C}} = (\dot{\mathcal{C}}_1, \dots \dot{\mathcal{C}}_{N-1})$. Then, compute the time derivative of each invariant as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{C}_i = \dot{W}_i W_i^\top + W_i \dot{W}_i^\top - \dot{W}_{i+1}^\top W_{i+1} - W_{i+1}^\top \dot{W}_{i+1}$$

and notice that

$$\begin{aligned} \dot{W}_i W_i^\top &= -(W_N \dots W_{i+1})^\top \nabla f(\mathbf{W})(W_{i-1} \dots W_1)^\top W_i^\top && (12)\\ &= -W_{i+1}^\top (W_N \dots W_{i+2})^\top \nabla f(\mathbf{W})(W_i \dots W_1)^\top && (13)\\ &= W_{i+1}^\top \dot{W}_{i+1} && (14) \end{aligned}$$

and similarly that $W_i \dot{W}_i^\top = \dot{W}_{i+1}^\top W_{i+1}$. This implies that $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{C} = (0, \dots, 0)$, completing the proof. □

A.2. Proof of Theorem 3

For clarity of presentation, we break this proof into a few smaller Lemmas to be proven in order. We first show that along any solution of the overparameterized gradient flow, the product of the parameter matrices is precompact. Then we show that the norm of all parameter matrices along a solution can be upper-bounded by an affine expression of the maximum singular value of the N-th parameter matrix. Then, using these two results we prove that the trajectory of each parameter matrix is precompact which finally allows the use of Łojasiewicz's Theorem to guarantee convergence of solutions to critical points.

Lemma 9 *For any $W^0 := (W_1^0, \dots, W_N^0) \in \mathbb{W}$, let $W(t, W^0) := (W_1(t, W^0), \dots, W_N(t, W^0))$ be a solution to the gradient flow in equation 4 initialized at W^0. Then, $\mathbf{W}(W(t, W^0))$ is pre-compact.*

Proof First, notice that along any solution of the gradient flow initialized in $\mathbb{W}$, the value of the overparameterized cost is non-increasing, that is

$$\frac{\mathrm{d}}{\mathrm{d}t} g(W(t, W^0)) = -\sum_i \left\langle \nabla_{W_i} g, \dot{W}_i \right\rangle = -\sum_i \|\nabla_{W_i} g\|_F^2 \leq 0, \tag{15}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner-product. Therefore, along any trajectory, $g(W(t, W^0)) \leq g(W^0) =: c$, which together with the fact that g is bounded below imply that $g(W(t, W^0))$ is bounded, which tautologically implies that $f(\mathbf{W}(W(t, W^0)))$ is bounded. From this we conclude that $\mathbf{W}(W(t, W^0))$ lies in $\mathcal{W}_c := \{\mathbf{W} \in \mathbb{W} \mid f(\mathbf{W}) \leq c\}$, which is a compact set because f is assumed to be proper, implying that $\mathbf{W}(W(t, W^0))$ is pre-compact. ■

Lemma 10 *For any i, j between 1 and N, there exists a constant c_{ij} such that*

$$\|W_i\|_F^2 = \|W_j\|_F^2 + c_{ij}. \tag{16}$$

Furthermore, let σ_N be the maximum singular value of W_N, then it follows that one can always find a_i and b_i such that

$$\|W_i\|_F \leq a_i \sigma_N + b_i \tag{17}$$

Proof To begin the proof, assume $i < j$ and take the trace of both sides of equation 5 to obtain

$$\|W_i\|_F^2 = \text{trace}\,(\mathcal{C}_i) + \|W_{i+1}\|_F^2,$$

which when performed iteratively results in

$$\|W_i\|_F^2 = \|W_j\|_F^2 + \sum_{k=i}^{j-1} \text{trace}\,(\mathcal{C}_k)\,.$$

The proof for $i > j$ is very similar except with a minus sign on the trace of the imbalance and appropriate summation limits. The statement is trivially true for $i = j$ and $c_{ii} = 0$.

At this point, equation 16 is proven. To prove equation 17 we take $i = N$ and use equivalence between the Frobenius and matrix 2 norm to justify the existence of an a_N such that $\|W_N\|_F \leq a_N \sigma_N$.

With this, first notice that if $\|W_i\|_F \leq \|W_N\|_F$, then *equation* 17 follows with $a_i = a_N$ and $b_i = 0$.

If, however, $\|W_i\|_F > \|W_N\|$, then we use *equation* 16 with $j = N$ to obtain that

$$\|W_i\|_F = \sqrt{\|W_N\|_F^2 + c_{iN}}.$$

Because $\|W_i\|_F > \|W_N\|_F$, we can conclude that $c_{iN} > 0$ which allows us to use the subadditive property of the square root to obtain that

$$\|W_i\|_F \leq \|W_N\|_F + \sqrt{c_{iN}} \leq a_N \sigma_N + \sqrt{c_{iN}},$$

concluding the proof. ■

Before proceeding to the next step of the proof, we need to formally define matrix polynomials in the context of this paper.

Definition 11 *For arbitrary (possibly repeating) integers $k > 0$, $i_1, \ldots, i_k \leq N$ and $j_0, \ldots j_k$, let $A_{j_0}, \ldots, A_{j_k}$ be constant matrices and $X_{i_1} \ldots X_{i_k}$ be variable matrices. A matrix monomial M is any term of the form*

$$M(X_1, \ldots, X_N) = A_{j_0} X_{i_1} A_{j_1} \ldots A_{j_{k-1}} X_{i_k} A_{j_k},$$

with k being the number of variable blocks X_{i_ℓ}, and it is called the degree of the monomial. Notice that the matrices A_i and X_j are assumed compatible so the monomial is well-defined. A matrix polynomial is, then, written as

$$\mathcal{P}(X) = \sum_{\lambda \in \Lambda} a_\lambda M_\lambda(X_1, \ldots, X_N)$$

and is a sum of matrix monomials as defined above, with the degree of the polynomial being the largest monomial degree of the sum.

Lemma 12 *Along any fixed trajectory, the following equality holds*

$$\mathbf{W}\mathbf{W}^\top = \left(W_N W_N^\top\right)^N + \mathcal{P}_N\left(W_N, \ldots, W_2\right), \tag{18}$$

where $\mathcal{P}_N$ is a polynomial of degree at most $2N-2$ on the matrix variables W_is and their transposes for $i \geq 2$.

Proof This lemma is proven inductively. For the first step, notice that using equation 5 for $i = 1$ one can write

$$\begin{aligned}
\mathbf{W}\mathbf{W}^\top &= W_N \ldots W_1 W_1^\top \ldots W_N^\top \\
&= W_N \ldots W_2 \left(\mathcal{C}_1 + W_2^\top W_2\right) W_2^\top \ldots W_N^\top \\
&= W_N \ldots W_3 \left(W_2 W_2^\top\right)^2 W_3^\top \ldots W_N^\top \; + \; \mathcal{P}_2(W_2, \ldots, W_N),
\end{aligned}$$

where the coefficients of $\mathcal{P}_2$ depend only on $\mathcal{C}_1$, and is of degree $2N-2$ on the variables $W_N, \ldots, W_2$. Now for the induction step, assume that for some i it was shown that one can write

$$\mathbf{W}\mathbf{W}^\top = W_N \ldots W_{i+1} \left(W_i W_i^\top\right)^i W_{i+1}^\top \ldots W_N^\top + \mathcal{P}_i(W_2, \ldots, W_N),$$

with $\mathcal{P}_i$ of degree at most $2N-2$ and whose coefficients depend only on $(\mathcal{C}_1, \ldots, \mathcal{C}_{i-1})$. Then, apply equation 5 to obtain

$$\mathbf{W}\mathbf{W}^\top = W_N \ldots W_{i+1} \left(W_{i+1}^\top W_{i+1} + \mathcal{C}_i\right)^i W_{i+1}^\top \ldots W_N^\top + \mathcal{P}_i(W_2, \ldots, W_N).$$

Notice that $(W_{i+1}^\top W_{i+1} + \mathcal{C}_i)^i$ can be expanded to $(W_{i+1}^\top W_{i+1})^i + \mathcal{R}_i(W_{i+1})$, where $\mathcal{R}_i$ collects all terms of the expansion of degree $2i-2$ or less on W_{i+1} and its transpose. From this, write

$$\begin{aligned}
\mathbf{W}\mathbf{W}^\top &= W_N \ldots W_{i+1} \left(W_{i+1}^\top W_{i+1}\right)^i W_{i+1}^\top \ldots W_N^\top \\
&\quad + W_N \ldots W_{i+1} \mathcal{R}_i(W_{i+1}) W_{i+1}^\top \ldots W_N^\top + \mathcal{P}_i(W_2, \ldots, W_N) \\
&= W_N \ldots W_{i+2} \left(W_{i+1} W_{i+1}^\top\right)^{i+1} W_{i+2}^\top \ldots W_N^\top + \mathcal{P}_{i+1}(W_N, \ldots, W_2),
\end{aligned}$$

where $\mathcal{P}_{i+1}$ is of degree at most $2N-2$ because it is a sum of two polynomials of degree at most $2N-2$, and has terms depending only on $(\mathcal{C}_1, \ldots, \mathcal{C}_i)$. Writing the induction step expression for $i = N-1$ completes the proof. ■

Lemma 13 *There exists a polynomial $p_N : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ of degree at most $2N-2$ such that*

$$\|P_N(W_N, \ldots, W_2)\|_F \leq p_N(\sigma_N),$$

where σ_N denotes the largest singular value of W_N.

Proof By Lemma 12, the matrix P_N can be expressed as a finite sum of matrix monomials of total degree at most $2N-2$ in the variables W_i and $W_i^\top$ ($i \geq 2$), that is,

$$P_N = \sum_{\lambda \in \Lambda} a_\lambda \, M_\lambda(W_2, \ldots, W_N),$$

where each coefficient $a_\lambda \in \mathbb{R}$ and each monomial M_λ has the form

$$M_\lambda = A_{0,\lambda} X_{i_1} A_{1,\lambda} X_{i_2} \cdots A_{k-1,\lambda} X_{i_k} A_{k,\lambda},$$

for some $k \leq 2N-2$, with constant matrices $A_{\ell,\lambda}$ depending only on the initialization (through the invariants $\mathcal{C}_i$), and factors $X_{i_j} \in \{W_2, \ldots, W_N, W_2^\top, \ldots, W_N^\top\}$.

Applying the triangle inequality and submultiplicativity of the Frobenius norm yields

$$\|P_N\|_F \leq \sum_{\lambda \in \Lambda} |a_\lambda| \, \|M_\lambda\|_F, \qquad \|M_\lambda\|_F \leq \Big(\prod_{\ell=0}^{k} \|A_{\ell,\lambda}\|_2\Big) \prod_{j=1}^{k} \|X_{i_j}\|_F =: c_\lambda \prod_{j=1}^{k} \|X_{i_j}\|_F.$$

By Lemma 10, for each $i \in \{2, \dots, N\}$ there exist constants $a_i, b_i \geq 0$ such that

$$\|W_i\|_F = \|W_i^\top\|_F \leq a_i\, \sigma_N + b_i.$$

Hence, for every monomial M_λ we obtain

$$\|M_\lambda\|_F \leq c_\lambda \prod_{m=2}^{N} (a_m \sigma_N + b_m)^{d_{m,\lambda}},$$

where $d_{m,\lambda}$ counts how many times W_m or $W_m^\top$ appears in M_λ. Since each monomial has total degree $\sum_m d_{m,\lambda} \leq 2N-2$, the right-hand side is a polynomial in σ_N of degree at most $2N-2$.

Define

$$p_N(x) := \sum_{\lambda \in \Lambda} |a_\lambda|\, c_\lambda \prod_{m=2}^{N} (a_m x + b_m)^{d_{m,\lambda}}.$$

This is a nonnegative polynomial of degree at most $2N-2$, depending only on the constants a_i, b_i and c_λ (which in turn depend only on the initialization through the invariants). Combining the inequalities above gives

$$\|P_N(W_N, \dots, W_2)\|_F \leq p_N(\sigma_N),$$

which completes the proof. ■

Remark 14 *It is evident that for a polynomial $p_1(x)$ of even degree $2k$, there exists a constant K such that $p_1(x) \leq 0.5x^{2(k+1)} + K$ for all x (the higher order term in x dominates for large values of x and the constant compensates for smaller values). We do not prove, but use this fact to prove the following lemma.*

Lemma 15 *If $\mathbf{W}(t)$ is precompact, then all $W_i(t)$ are precompact.*

Proof Notice that since the Frobenius norm of a matrix upper-bounds its spectral norm, we can write

$$\sigma_N^{2N} \leq \|(W_N W_N^\top)^N\|_F,$$

which, by using Lemma 12 becomes

$$\sigma_N^{2N} \leq \|\mathbf{W}\mathbf{W}^\top\|_F + \|\mathcal{P}_N\|_F.$$

Then, by applying Lemma 13 it becomes

$$\sigma_N^{2N} \leq \|\mathbf{W}\|_F^2 + p_N(\sigma_N).$$

Then, through Remark 14 we know that there exists some K such that $p_N(\sigma_N) \leq 0.5\sigma_N^{2N} + K$, we can write

$$\sigma_N^{2N} \leq \|\mathbf{W}\|_F^2 + 0.5\sigma_N^{2N} + K$$

which in turn implies that there exist some η and γ such that

$$\sigma_N \leq \eta \|\mathbf{W}\|_F^{1/N} + \gamma.$$

however, since from Lemma 10 there exist a_i and b_i such that $\|W_i\|_F \leq a_i \sigma_N + b_i$, it follows that there exist η_i and γ_i such that

$$\|W_i\|_F \leq \eta_i \|\mathbf{W}\|_F^{1/N} + \gamma_i.$$

Finally, since $\mathbf{W}(t)$ is precompact, then $\|\mathbf{W}(t)\|_F$ is bounded above, which in turn implies that $\|W_i(t)\|_F$ is bounded above, proving that $W_i(t)$ is precompact. ■

We finally have all results to prove Theorem 3. From Lemma 9, we know that $\mathbf{W}(t)$ is precompact. Then, from Lemma 15, we know that $\mathbf{W}(t)$ being precompact implies that for all i between 1 and N, $W_i(t)$ is precompact. From here, we can conclude the statement of the Theorem from applying Lojasiewicz's Theorem to this gradient system, since f (and thus g) is assumed real analytic. □

A.3. Proof of Theorem 4

To prove Theorem 4, we first state the following definition and lemma:

Definition 16 *Given a function $f : \mathbb{W} \to \mathbb{R}$, a point W is a* strict saddle *of the gradient flow of f if the Hessian of f at W has a direction of strictly negative curvature.*

Lemma 17 *If (W_1, W_2) is a critical point of g but W_2W_1 is not a critical point of f, then (W_1, W_2) is a strict saddle of the overparameterized gradient flow dynamics.*

Proof To begin this proof, we first show that if (W_1, W_2) is a critical point of g, but W_2W_1 is not of f, then both W_1 and W_2 must be rank deficient. To see this, first notice that if W_2W_1 is not a critical point of f, then $\nabla f(W_2W_1) \neq 0$ by definition. However, if (W_1, W_2) is a critical point of g, then

$$\begin{aligned}\nabla_{W_1} g(W_1, W_2) :=& W_2^\top \nabla f(W_2W_1) = 0 \Rightarrow \text{Im}(\nabla f(W_2W_1)^\top) \subseteq \ker(W_2^\top)\\ \nabla_{W_2} g(W_1, W_2) :=& \nabla f(W_2W_1) W_1^\top = 0 \Rightarrow \text{Im}(\nabla f(W_2W_1)) \subseteq \ker(W_1).\end{aligned}$$

Since $\nabla f(W_2W_1) \neq 0$ then $\text{Im}(\nabla f(W_2W_1)) \neq \{0\}$ implying that the kernel of W_2 must also be non-empty (and similarly for W_1). This proves W_1 and W_2 must be rank deficient.

Next we state the Taylor expansion of f around a point W in a direction M, which is well defined since f is assumed real-analytic.

$$f(W + M) = f(W) + f'(W)[M] + f''(W)[M, M] + o(\|M\|^2)$$

where

$$\begin{aligned}f'(W)[M] :=& \sum_{(i,j)=1}^{(n,m)} \frac{\partial f}{\partial w_{ij}}(W) m_{ij} = \langle \nabla f(W), M \rangle\\ f''(W)[M, M] :=& \frac{1}{2} \sum_{(i,j),(k,l)}^{(n,m),(n,m)} \frac{\partial f}{\partial w_{ij} w_{kl}}(W) m_{ij} m_{kl},\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product. From here we will build the Taylor expansion of g around a point (W_1, W_2) in a direction (M_1, M_2). First notice that

$$(W_2 + M_2)(W_1 + M_1) = W_2W_1 + \underbrace{W_2M_1 + M_2W_1}_{A} + \underbrace{M_2M_1}_{B}$$

Where A is first order in (M_1, M_2) and B is second order. From here we are ready to define

$$\begin{aligned} g(W_1+M_1, W_2 + M_2) := g(W_1, W_2) + g'(W_1, W_2)[(M_1, M_2)] \\ + g''(W_1, W_2)[(M_1, M_2), (M_1, M_2)] + o(\|(M_1, M_2)\|^2), \end{aligned}$$

where

$$\begin{aligned} g'(W_1, W_2)[(M_1, M_2)] &:= f'(W_2W_1)[A] \\ g''(W_1, W_2)[(M_1, M_2), (M_1, M_2)] &:= f'(W_2W_1)[B] \\ &\quad + f''(W_2W_1)[A, A]. \end{aligned}$$

We skip the full derivation of the expression above, however notice that the $f'(W_2W_1)[B]$ term in the Hessian expression of g originates from the second order chain rule. To prove that (W_1, W_2) is a strict saddle of the overparameterized gradient flow, it is enough to prove that there exist (M_1, M_2) such that

$$g''(W_1, W_2)[(M_1, M_2), (M_1, M_2)] < 0.$$

From here, we are ready to prove the claim.

Let $\psi \in \mathbb{R}^n$ and $\phi \in \mathbb{R}^m$ be a left and right singular vector pair of $\nabla f(W_2W_1)$ associated with a nonzero singular value $\sigma > 0$. Next, let γ be a unit vector and be such that $\gamma^\top W_1 = 0$ (no assumption is made about the value of $W_2\gamma$). Such a γ always exists because we established that W_1 and W_2 must be rank deficient. Pick $M_1 = -\gamma\phi^\top p = \overline{M}_1 p$ and $M_2 = \psi\gamma^\top q = \overline{M}_2 q$ for some positive scalars p, q. For this choice of M_1 and M_2 notice that $M_2W_1 = 0$ and thus $A = W_2M_1 = W_2\overline{M}_1 p$ and $B = M_2M_1 = \overline{M}_2\overline{M}_1 pq$. For this choice of M_1 and M_2 we can write

$$\begin{aligned} g''(W_1, W_2)[(M_1, M_2), (M_1, M_2)] &= f'(W_2W_1)[B] + f''(W_2W_1)[A, A] \\ &= pqf'(W_2W_1)[\overline{M}_2\overline{M}_1] + p^2 f''(W_2W_1)[W_2\overline{M}_1, W_2\overline{M}_1] \\ &= apq + bp^2. \end{aligned}$$

From the chosen M_1 and M_2 notice that

$$\begin{aligned} a &= f'(W_2W_1)[\overline{M}_2\overline{M}_1] \\ &= \text{trace}\left(\nabla f(W_2W_1)^\top \overline{M}_2\overline{M}_1\right) \\ &= \text{trace}\left(\nabla f(W_2W_1)^\top(-\psi\phi^\top)\right) \\ &= -\psi^\top \nabla f(W_2W_1)\phi \\ &= -\sigma, \end{aligned}$$

however the sign of b is undefined (and if $W_2\gamma = 0$, then $b = 0$) although we know it is finite due to continuity of f''. Despite that, if we pick the particular case where $p = q^2$ then we can write

$$g''(W_2W_1)[(M_1, M_2), (M_1, M_2)] = -\sigma q^3 + bq^4.$$

From the expression above, let $\overline{q} = \sigma/(2|b|)$ if $b \neq 0$ or any positive value if $b = 0$, then for all $0 < q < \overline{q}$, $-\sigma q^3 + bq^4 < 0$.

Hence, there exists a direction (M_1, M_2) along which the Hessian of g is strictly negative, proving that every (W_1, W_2) that consists on a critical point of g but that is such that W_2W_1 is not a critical point of f is a strict saddle. ■

We also state the following Theorem, and refer the reader to de Oliveira et al. (2025b) for the proof.

Theorem 18 *Consider a nonlinear system of the form $\dot{x} = f(x)$, $f : \mathcal{X} \to \mathcal{TX}$. Suppose that $E \subseteq \mathcal{X}$ is a set consisting of strict saddle equilibria of the system. Then the set $\mathcal{C}_E$ of points $x_0 \in \mathcal{X}$ whose trajectories converge to points in E has measure zero.*

Finally, we know that all solutions converge to a critical point due to Theorem 3. Then, Lemma 17 tells us that all points in

$$E = \{(W_1, W_2) \mid \nabla_{W_{1,2}} g(W_1, W_2) = 0,\ \nabla f(W_2W_1) \neq 0\},$$

are strict saddles, which allows us to invoke Theorem 18 to prove Theorem 4 □

A.4. Proof of Proposition 5

Let $z := w_2w_1$ and $g(w) := f(z)$ for $w = (w_1, w_2) \in \mathbb{R}^k \times \mathbb{R}^k$. For the gradient flow $\dot{w} = -\nabla g(w)$ we have

$$\nabla g(w) = f'(z)\begin{bmatrix} w_2 \\ w_1 \end{bmatrix}, \qquad \dot{w}_1 = -f'(z)\, w_2, \quad \dot{w}_2 = -f'(z)\, w_1.$$

Define

$$S(t) = \|w_1\|^2 + \|w_2\|^2, \qquad z(t) = w_2w_1, \qquad V(t) = f(z(t)).$$

Then

$$\dot{z} = -S\, f'(z), \qquad \dot{S} = -4z\, f'(z), \qquad \dot{V} = f'(z)\dot{z} = -S\,[f'(z)]^2 \leq 0.$$

The quantity $D := S^2 - 4z^2$ is conserved: $\dot{D} = 0$ and thus $D(t) \equiv D(0) \geq 0$. Introduce

$$u := w_1 - s\, w_2, \qquad v := w_1 + s\, w_2,$$

In these coordinates,

$$\dot{u} = s\, f'(z)\, u, \qquad \dot{v} = -\, s\, f'(z)\, v,$$

and

$$S = \tfrac{1}{2}\big(\|u\|^2 + \|v\|^2\big), \qquad z = \tfrac{s}{4}\big(\|v\|^2 - \|u\|^2\big), \qquad D = \|u\|^2\, \|v\|^2$$

Let $\mathcal{S}^+ = \{f'(z) > 0\}$ and $\mathcal{S}^- = \{f'(z) < 0\}$; their common boundary is $Z = \{f'(z) = 0\}$, we show they are forward invariant. Suppose towards contradiction a trajectory starting in $\mathcal{S}^+$ were to enter $\mathcal{S}^-$ at the first time $t_* > 0$. Then $f'(z(t_*)) = 0$ and hence $\dot{w}(t_*) = 0$; by uniqueness, the solution with initial condition $w(t_*)$ is constant, thus the trajectory cannot cross the boundary. The same holds for the roles of $\mathcal{S}^{\pm}$ reversed. Thus $\mathcal{S}^{\pm}$ are forward invariant. In particular, while $f'(z(t)) \neq 0$ the sign of $\dot{z} = -Sf'(z)$ is fixed and $z(t)$ is strictly monotone. Similarly, the sets $\{u = 0\}$ and $\{v = 0\}$ are forward invariant because $\dot{u} = sf'(z)\, u$ and $\dot{v} = -sf'(z)\, v$. Note

$\{u = 0\}$ coincides with $\Delta_+ = \{w_1 = w_2\}$ if $s = +1$ and with $\Delta_- = \{w_1 = -w_2\}$ if $s = -1$; the correspondence is reversed for $\{v = 0\}$. Since $\dot{V} \leq 0$, we have $V(t) \leq V(0)$ for all t. By the properness of f, the set $\{\, \xi : \ f(\xi) \leq V(0) \,\}$ is compact, so $z(t)$ remains bounded. Then $S^2(t) = 4z^2(t) + D(0)$ bounds $S(t)$, and hence $(w_1(t), w_2(t))$ is bounded. Thus all trajectories are precompact.

By the real analyticity of $g(w) = f(w_2^\top w_1))$, Łojasiewicz's theorem says that every precompact trajectory of $\dot{w} = -\nabla g(w)$ converges to a single critical point of g. The critical points satisfy $\nabla g(w) = 0$, i.e.,

$$f'(z) = 0 \quad \text{or} \quad (w_1, w_2) = (0, 0).$$

Recall $d(w_1, w_2) = \|u\|$. Consider the following cases. First, suppose $D(0) > 0$. Then $\|u(0)\| > 0$ and $\|v(0)\| > 0$, and since $D(t) = \|u(t)\|^2\|v(t)\|^2 \equiv D(0)$, both $\|u(t)\|$ and $\|v(t)\|$ stay bounded away from zero for all t. Hence the limit cannot be $(0, 0)$, so the trajectory converges to a point in Z. Then suppose $D(0) = 0$ and $d(w_1^0, w_2^0) > 0$. Here $\|u(0)\| > 0$ forces $\|v(0)\| = 0$, and by invariance $v(t) \equiv 0$ for all t. Along this invariant line we have $z = -\frac{s}{4}\|u\|^2$ and $\dot{u} = s\, f'(z)\, u$. If the trajectory were to converge to $(0, 0)$, then $u(t) \to 0$ and hence $z(t) \to 0$, so for t large we would have $s\, f'(z(t)) > c > 0$ (because $f'(0) = s$). But then

$$\frac{d}{dt}\|u(t)\|^2 \ = \ 2\, s\, f'\big(z(t)\big)\, \|u(t)\|^2 \ \geq \ 2c\, \|u(t)\|^2,$$

which does not allow $\|u(t)\| \to 0$ forward in time. Thus the origin cannot be the limit. By Łojasiewicz's theorem, the trajectory must converge to a point with $f'(z) = 0$, i.e., to an element of Z. In all cases compatible with $d(w_1^0, w_2^0) > 0$, the trajectory converges to a point in Z. Consequently,

$$\lim_{t\to\infty} f'\big(w_2(t)^\top w_1(t)\big) = 0.$$

□

A.5. Proof of Corollary 6

By assumption, $|f'(0)| \geq \alpha\big(f(0) - f_{\min}\big) > 0$, so $f'(0) \neq 0$. Set $s = \operatorname{sign}(f'(0))$, and write $z(t) = w_2(t)^\top w_1(t)$ and $V(t) = f(z(t))$. If $d(w_1^0, w_2^0) > 0$, Proposition 5 yields $(w_1(t), w_2(t)) \to (w_1^*, w_2^*)$ with $f'\big(w_2^{*\top} w_1^*\big) = 0$. From assumption of $\mathcal{PD}$-PŁI , $f'(z) = 0$ implies $f(z) = f_{\min}$, hence $V(t) \to f_{\min}$ by continuity. For the converse, argue by contrapositive and suppose $d(w_1^0, w_2^0) = 0$. Let $u = w_1 - s\, w_2$. From $\dot{w}_1 = -f'(z)\, w_2$ and $\dot{w}_2 = -f'(z)\, w_1$ we obtain $\dot{u} = s\, f'(z)\, u$, hence $u(0) = 0$ implies $u(t) \equiv 0$ and therefore $w_1(t) = s\, w_2(t)$ for all t. As in Proposition 5, the trajectory is precompact (since $\dot{V} = -S[f'(z)]^2 \leq 0$ and f is proper), and $g(w) = f(w_2^\top w_1)$ is real-analytic; therefore, by Łojasiewicz's theorem, the trajectory converges to a single critical point w^* of g. Critical points of g satisfy either $f'(z^*) = 0$ or $w^* = (0, 0)$. On the invariant line $\{u = 0\}$ we have $z = s\, \|w_1\|^2$ and thus $s\, z \geq 0$. We claim there is no point on the ray $\{s\, z > 0\}$ with $f'(z) = 0$. Indeed, define $g_1(\rho) = f(s\rho)$ for $\rho \geq 0$. Then $g_1'(0) = s f'(0) > 0$. If there were $\rho_* > 0$ with $g_1'(\rho_*) = 0$, the inequality would force $g_1(\rho_*) = f_{\min} < g_1(0)$. By the mean value theorem there exists $\hat{\rho} \in (0, \rho_*)$ with $g_1'(\hat{\rho}) < 0$, so by continuity g_1' would vanish first at some $\rho_0 \in (0, \rho_*)$ with $g_1(\rho_0) > f_{\min}$, contradicting the $\mathcal{PD}$-PŁI assumption (which gives $g_1' = 0 \Rightarrow g_1 = f_{\min}$). Hence $f'(z) \neq 0$ for all $s\, z > 0$, which excludes critical points of the form $f'(z) = 0$ on the stable manifold of the saddle. Therefore the only critical point accessible on

$\{u = 0\}$ is $w^* = (0, 0)$, so the Łojasiewicz theorem yields $(w_1(t), w_2(t)) \to (0, 0)$; in particular, $z(t) \to 0$ and $V(t) = f(z(t)) \to f(0) > f_{\min}$. Therefore, $\lim_{t\to\infty} f\big(w_2(t)^\top w_1(t)\big) = f_{\min}$ holds if and only if $d(w_1^0, w_2^0) > 0$. □

A.6. Proof of Proposition 7

Remember that $\mathbf{w}(w_1, w_2) = w_2 w_1$, and drop the explicit dependencies when it is obvious by context. Notice from the gradient flow that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{w} &= -f'(\mathbf{w})(w_2 w_2^\top + w_1^\top w_1) \\ &= -f'(\mathbf{w})(2w_2 w_2^\top + \text{trace}\,(\mathscr{C})) \\ &= -f'(\mathbf{w})\sqrt{c + 4\mathbf{w}^2} \qquad (19)\end{aligned}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{w}) = -[f'(\mathbf{w})]^2\sqrt{c + 4\mathbf{w}^2} \qquad (20)$$

where $c := 2\text{trace}\left(\mathscr{C}^2\right) - \text{trace}\,(\mathscr{C})^2$. From the expressions above we can see that the imbalance constant c merely rescales the rate of $\mathbf{w}$ and g.

Let $\tau(t)$ be a time reparameterization such that $\frac{\mathrm{d}\tau}{\mathrm{d}t} = \sqrt{c + 4\mathbf{w}(\tau)^2}$ with $\tau(0) = 0$. Then $\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}\tau} = -f'(\mathbf{w})$ and $\frac{\mathrm{d}f}{\mathrm{d}\tau} = -[f'(\mathbf{w})]^2$. Therefore, $\mathbf{w}(\tau)$ and $f(\mathbf{w}(\tau))$ will have the same trajectory for the same values of τ if initialized at the same point, independently of the value of the imbalance c.

Then, to prove the statement it is enough to show that for two points $\overline{w}$ and $\widetilde{w}$ that satisfy the conditions in the proposition, $\overline{\tau} > \widetilde{\tau}$ for all $t > 0$, since $\frac{\mathrm{d}}{\mathrm{d}t} f \leq 0$.

To prove that, let $\Delta(t) := \overline{\tau}(t) - \widetilde{\tau}(t)$. Notice that $\Delta(0) = 0$ and $\dot{\Delta}(0) = \sqrt{\overline{c} + 4\overline{\mathbf{w}}^2} - \sqrt{\widetilde{c} + 4\widetilde{\mathbf{w}}^2} > 0$, and furthermore notice that for any $\bar{t}$ such that $\Delta(\bar{t}) = 0$ then $\dot{\Delta}(\bar{t}) > 0$ necessarily. We argue that this implies that $\Delta(t) > 0$ for all $t > 0$. To see that, first notice that $\Delta(\epsilon) > 0$ for some $\epsilon > 0$ since $\dot{\Delta}(0) > 0$. Then, assume for contradiction that at some time $\bar{t} > 0$, $\Delta(\bar{t}) = 0$. In particular, let $\bar{t}$ be the smallest $t > 0$ for which this condition is satisfied. Since $\dot{\Delta}(\bar{t}) > 0$, then there must exist some $h \in (\epsilon, \bar{t})$ for which $\Delta(h) < 0$, however if that is the case, then $\Delta(\epsilon) > 0$ and $\Delta(h) < 0$ which by the mean value theorem implies that there must exist some $\delta \in (\epsilon, h)$ such that $\Delta(\delta) = 0$, which breaks the condition that $\bar{t}$ is the smallest time for which $\Delta(\bar{t}) = 0$, reaching contradiction.

This proves that $\Delta(t) > 0$ for all $t > 0$ which in turn implies that $\overline{\tau}(t) > \widetilde{\tau}(t)$, which implies that $f(t, \overline{w}) < f(t, \widetilde{w})$ for all $t > 0$. □

A.7. Proof of Proposition 8

To prove this proposition, all one needs to do is to algebraically compute $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{C}$ as follows

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{C} &= 2w_2\dot{w}_2 - 2(1 + w_1^2)w_1\dot{w}_1 \\ &= 2w_2 w_1\left(\frac{\sqrt{1 + w_1^2} - w_2 w_1}{1 + w_1^2} - (1 + w_1^2)\frac{\sqrt{1 + w_1^2} - w_2 w_1}{(1 + w_1^2)^2}\right) = 0.\end{aligned}$$

□