Understanding the Dynamics of Gradient Flow in Overparameterized Linear Models

Salma Tarmoun [*1,2] **Guilherme França** [*1,3] **Benjamin Haeffele** [1,4] **René Vidal** [1,4]

Abstract

We provide a detailed analysis of the dynamics of the gradient flow in overparameterized two-layer linear models. A particularly interesting feature of this model is that its nonlinear dynamics can be exactly solved as a consequence of a large number of conservation laws that constrain the system to follow particular trajectories. More precisely, the gradient flow preserves the difference of the Gramian matrices of the input and output weights, and its convergence to equilibrium depends on both the magnitude of that difference (which is fixed at initialization) and the spectrum of the data. In addition, and generalizing prior work, we prove our results without assuming small, balanced or spectral initialization for the weights. Moreover, we establish interesting mathematical connections between matrix factorization problems and differential equations of the Riccati type.

1. Introduction

Understanding *overparameterization* in deep learning is a puzzling question. Contrary to the common belief that it may hurt generalization and optimization, recent work suggests that overparameterization may actually bias the optimization algorithm towards solutions that generalize well—a phenomenon known as *implicit regularization* or *implicit bias*—and may also accelerate convergence—a phenomenon known as *implicit acceleration*. Both phenomena are consequences of the fact that different optimization algorithms correspond to different dynamical systems acting on different models. Understanding these effects thus requires a thorough analysis of the dynamics of optimization methods, such as gradient descent, which despite its simplicity can lead to a complicated nonlinear dynamics.

Recent work on the implicit bias in the overparameterized regime (Chizat & Bach, 2020; Gunasekar et al., 2018a;b; Ji & Telgarsky, 2019b) shows that gradient descent on unregularized problems finds *minimum norm solutions*. For instance, Ji & Telgarsky (2019a); Soudry et al. (2018) analyze linear networks trained for binary classification on linearly separable data, and show that the predictor converges to a max-margin solution. Similar ideas have been developed for matrix factorization, yielding solutions with minimum nuclear norm (Gunasekar et al., 2017; Li et al., 2018) or low-rank (Arora et al., 2019a). The convergence properties of gradient descent on wide neural networks was also analyzed (Arora et al., 2019b; Du & Hu, 2019; Du et al., 2018b), leading to a linear convergence when the initialization is Gaussian or balanced. Recent results regarding the implicit acceleration of gradient descent on matrix factorization and deep linear networks (Arora et al., 2018) show that when the initialization is sufficiently *small* and *balanced* (see Definition 2), overparameterization acts as a preconditioning and can be interpreted as a combination of momentum and an adaptive learning rate, suggesting that acceleration for ℓ_p-regression is possible when $p > 2$.

A parallel line of research has been exploring the continuous-time limit of gradient descent, i.e., the *gradient flow*, which has the advantage of being mathematically more tractable. Furthermore, such a continuous-time analysis can provide a stepping stone towards understanding the (discrete-time) dynamics of gradient descent. For instance, Saxe et al. (2014) analyzed the gradient flow dynamics for deep linear models with least squares loss under the assumptions of whitened data, small, balanced, and spectral initialization (see Definition 3); they obtained a closed-form solution in this setting. For two-layer linear networks, by means of such solution, Gidel et al. (2019); Saxe et al. (2019) highlighted the sequential learning of the hierarchical components as a phenomenon that could improve generalization. Moreover, Gidel et al. (2019) was able to extend some of these results to gradient descent. However, both of these papers limited the analysis to vanishing spectral initialization.

[*]Equal contribution [1]Mathematical Institute for Data Science, Johns Hopkins University, [2]Department of Applied Mathematics and Statistics, Johns Hopkins University, [3]Computer Science Division, University of California, Berkeley, [4]Department of Biomedical Engineering, Johns Hopkins University. Correspondence to: Salma Tarmoun <starmou1@jhu.edu>, Guilherme França <guifranca@gmail.com>, Benjamin Haeffele <bhaeffele@jhu.edu>, René Vidal <rvidal@jhu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

In the present paper, we explore the same setting as Gidel et al. (2019); Saxe et al. (2019) but present a novel and more general analysis of the gradient flow on two-layer linear networks that applies not only to small, balanced, or spectral initializations, but also to *imbalanced and nonspectral* initializations. We show that a key ingredient is the existence of a sufficiently large number of *conservation laws* that constrain the dynamics to follow a particular path.[1] The quantity that is preserved by the gradient flow is the difference of the Gramians of the input and output weight matrices, which in turn implies that the difference of the norm square of the weight matrices is preserved. The particular case where this difference is zero corresponds to *balanced weights*, but the more general case of *imbalanced weights* also emerges as a conserved quantity and plays an important role in the convergence of the system. In particular, we prove convergence of the gradient flow for two-layer models without the assumption of small, balanced or spectral initialization, and we show explicitly the dependency of initialization and data spectrum on the convergence rate. Our work thus extends some of the results previously established by Gidel et al. (2019); Saxe et al. (2014; 2019) which now follow as a particular cases of our analysis. We also establish interesting connections with Riccati differential equations, providing an explicit characterization of the gradient flow dynamics. In short, our work makes the following contributions—see also Table 1:

- In Section 2, we analyze the dynamics of gradient flow for *symmetric* matrix factorization, providing a closed-form solution and a convergence rate that depends on the eigenvalues of the data; this is done without assuming small or spectral initialization.
- In Section 3, we consider an *asymmetric* matrix factorization with spectral initialization. We highlight the role of conservation laws that only appear in the overparameterized setting—as a consequence of an underlying rotational symmetry—and provide a convergence rate under imbalanced initialization.
- In Section 4, we analyze the dynamics of gradient flow for *asymmetric* matrix factorization with *arbitrary initialization*. We make interesting connections with Riccati differential equations, yielding a more general characterization of the convergence rate and an interesting connection with explicit regularization.

[1] A quantity $\mathcal{Q}(x(t))$ is said to be conserved under the flow $\dot{x}(t) = f(x(t))$ if it remains constant through time, i.e., $\frac{d}{dt}\mathcal{Q}(x(t)) = 0$. For example, in mechanics the sum of potential and kinetic energies remains constant for a conservative system. A conservation law is usually a consequence of an underlying symmetry (Noether's theorem). In optimization, this can be seen as a constraint $\mathcal{Q}(x) = \mathcal{Q}_0$ that is automatically satisfied, without having to be explicitly enforced.

2. Symmetric Matrix Factorization

In this section we analyze and compare the dynamics of the gradient flow,[2]

$$\dot{X}(t) = -\nabla_X \ell(X(t)), \tag{1}$$

when applied on two problems. The first is a symmetric one-layer linear model:

$$\min_{X\in\mathbb{R}^{m\times m}} \left\{\ell(X) \equiv \tfrac{1}{2}||Y - X||_F^2\right\} \tag{2}$$

where $Y \in \mathbb{R}^{m\times m}$ is a given data matrix that one wishes to approximate by $X \in \mathbb{R}^{m\times m}$. The second is its overparameterized counterpart:

$$\min_{U\in\mathbb{R}^{m\times k}} \left\{\ell(U) \equiv \tfrac{1}{2}||Y - UU^T||_F^2\right\}. \tag{3}$$

We show that the dynamics of the linear model converge at a rate $O(e^{-t})$, while the overparameterized model has a rate $O(e^{-4t|\sigma_i|})$, where σ_i is the ith eigenvalue of the data matrix Y. Therefore, different spectral components are learned at different rates—this is the sequential learning phenomenon described by Saxe et al. (2019) which we extend to the symmetric factorization case.

Linear model. Let us start with the (trivial) problem of learning the linear model (2).[3] Applying the gradient flow (1) to problem (2) yields $\dot{X}(t) + X(t) = Y$ with $X(0) = X_0$. This is a linear differential equation whose unique solution is

$$X(t) = Y + (X_0 - Y)e^{-t}. \tag{4}$$

Thus, $\|X(t)-Y\|_F = e^{-t}\|X_0-Y\|_F$ and $\lim_{t\to\infty} X(t) = Y$ at an exponential rate of $O(e^{-t})$.

For completeness, it will be interesting to consider the particular case in which X is constrained to be positive semidefinite (PSD), i.e., $X \succeq 0$. In this case, notice that if $X_0 \succeq 0$ and $Y \succeq 0$, then $X(t) \succeq 0$ for all $t > 0$, hence the same dynamics and convergence rate still apply without having to enforce the PSD constraint. Otherwise, if Y is not PSD, gradient flow is not directly applicable.

Symmetric matrix factorization model. Consider now the more interesting case of learning a two-layer linear model with tied weights $U \in \mathbb{R}^{m\times k}$, formulated as the symmetric matrix factorization problem in (3). In classical low-rank matrix factorization one assumes $k < m$. Here, we consider an overparameterized formulation where $k \geq m$ plays the role of the number of hidden units

[2] Gradient descent, $x_{n+1} = x_n - \eta\nabla\ell(x_n)$, is simply an explicit Euler discretization of (1).

[3] In a linear neural network, Y plays the role of the input-output data correlation matrix and X plays the role of the model's input-output map. In this trivial model, the input correlation matrix is assumed to be the identity as is the case when the data is whitened.

Table 1. Comparison between our work and the state-of-the-art.

	Small and Balanced	*Imbalanced*	*Spectral*	*Nonspectral*
Gradient flow	Our work Saxe et al. (2014) Saxe et al. (2019)	Our work	Our work Saxe et al. (2014) Saxe et al. (2019)	Our work
Gradient descent	Arora et al. (2018) Gidel et al. (2019)	None	Gidel et al. (2019)	Arora et al. (2018)

(width). The gradient flow (1) on problem (3), for U, now yields $\dot{U} = 2(Y - UU^T)U$ with $U(0) \equiv U_0$. Letting $X(t) \equiv U(t)U(t)^T \succeq 0$ and $X(0) = U_0U_0^T \succeq 0$, one can easily verify that

$$\dot{X} = \dot{U}U^T + U\dot{U}^T = 2YX + 2XY - 4X^2. \quad (5)$$

We note that problem (3) is nonconvex and the system has multiple stationary points characterized by

$$YX + XY - 2X^2 = (Y - X)X + X(Y - X) = 0. \quad (6)$$

Due to the symmetric and positive semidefinite nature of X and Y, the algebraic equation (6) can be reduced to

$$X(Y - X) = 0. \quad (7)$$

This problem shares the trivial solution $X = Y$ with the one-layer problem. However, any matrix of the form $X = \Phi\,\mathrm{diag}(x_1, \dots x_n)\Phi^T$ where $Y = \Phi\,\mathrm{diag}(\sigma_1, \dots, \sigma_n)\Phi^T$ and $x_i \in \{0, \sigma_i\}$ is also a solution.

Equation (5) is known to be rank preserving, i.e., if $r \equiv \mathrm{rank}(X_0) \leq m$, and $X^* = \lim_{t\to\infty} X(t)$ exists, then $\mathrm{rank}(X^*) \leq r$. Thus, low-rank initializations lead to low-rank solutions and, importantly, it is impossible to recover a solution with higher rank than that of the initialization.

We note that Eq. (5) (resp. Eq. (6)) is a matrix differential equation (resp. algebraic equation) of the *Riccati* type. Such equations often characterize dynamical systems behind least squares problems and have been extensively studied in optimal control. Using results from this literature, we obtain (see Appendix A for the proof):

Proposition 1. *For any $X_0 \in \mathbb{R}^{m\times m}$ the solution to Eq. (5) exists and is given by*

$$X(t) = e^{2tY}X_0\left[I + Y^{-1}(e^{4tY} - I)X_0\right]^{-1}e^{2tY}, \quad (8)$$

provided Y and the matrix inside $[\cdots]^{-1}$ are invertible.

This solution is derived for any X_0, while the overparameterized model requires $X_0 = U_0U_0^T \succeq 0$. Thus, in using (8) as an analysis tool, it is important to keep in mind the set of consistent initializations. In what follows, we consider the spectral initialization (Gidel et al., 2019; Saxe et al., 2019) and show that the eigenspace of the data is preserved throughout the entire evolution of the learning dynamics.

Definition 1 (Symmetric Spectral initialization). *Let $Y = \Phi\Sigma\Phi^T$ be the eigendecomposition of the data. A spectral initialization is defined as $U_0 \equiv \Phi\Sigma_0^{1/2}$ and $X_0 \equiv U_0U_0^T = \Phi\Sigma_0\Phi^T$ where $\Sigma_0 \succeq 0$ is a diagonal matrix.*

From the explicit solution (8) we can readily obtain a *convergence rate* under spectral initializations.

Corollary 1. *If $Y = \Phi\Sigma\Phi^T = \sum_{i=1}^m \sigma_i\phi_i\phi_i^T$ is invertible and $X_0 = \Phi\Sigma_0\Phi^T = \sum_{i=1}^m \sigma_{0,i}\phi_i\phi_i^T$ is a spectral initialization, the solution to Eq. (5) is given by $X(t) = \Phi\Sigma(t)\Phi^T = \sum_{i=1}^m \sigma_i(t)\phi_i\phi_i^T$ with*

$$\sigma_i(t) = \sigma_i + \frac{\sigma_i(\sigma_{0,i} - \sigma_i)}{\sigma_i + \sigma_{0,i}(e^{4t\sigma_i} - 1)}, \quad (9)$$

provided the denominator is nonzero. Moreover, if $\tilde{Y} = \sum_{i=1}^m \max(\sigma_i, 0)\phi_i\phi_i^T = \Phi\tilde{\Sigma}\Phi^T$ is the projection of Y onto the PSD cone, then for all initializations $X_0 \succeq 0$ such that rank$(\Sigma_0\tilde{\Sigma})$ = rank$(\tilde{\Sigma})$, then $X(t)$ converges to $\tilde{Y}$ at a rate $O(e^{-4t\sigma_{\min}(Y)})$ where $\sigma_{\min}(Y) = \min_i |\sigma_i|$.

Proof. The first part follows trivially by substitution into (8) and verifying the invertibility condition. For the second part, note from (9) that if $\sigma_i > 0$ then $\sigma_i(t) \to \sigma_i$ as $t \to \infty$ at a rate $O(e^{-4t\sigma_i})$, and if $\sigma_i < 0$, then $\sigma_i(t) \to 0$ at a rate $O(e^{4t\sigma_i})$. Therefore, $\Sigma(t) \to \max(\Sigma, 0)$ and $X(t) \to \Phi\max(\Sigma, 0)\Phi^T$ at a rate $O(e^{-4t\sigma_{\min}(Y)})$. □

From (9) we see that the ith eigencomponent of $X(t)$ converges at a rate $O(e^{-4t|\sigma_i|})$. This result about different components of the network being learned at different rates is related in spirit to the result of Gidel et al. (2019); Saxe et al. (2019) about sequential learning with spectral balanced initialization. Here the balancedness is enforced by construction.

Next, we derive the same convergence rate with a more general— nonspectral—initialization. The proof is in Appendix B and makes use of several interesting relations for Riccati differential equations.

Proposition 2 (Convergence rate). *Consider the eigenvalue decomposition* $Y = \sum_{i=1}^{m} \sigma_i \phi_i \phi_i^T$. *Let* $\tilde{Y} = \sum_{i=1}^{m} \max(\sigma_i, 0) \phi_i \phi_i^T$ *be the projection of* Y *onto the PSD cone and* $\hat{Y} = \sum_{i=1}^{m} |\sigma_i| \phi_i \phi_i^T$. *For any initialization* $X_0 \succeq 0$, *assume* $I + \hat{Y}^{-1}\big(X_0 - \tilde{Y}\big)$ *and* Y *are nonsingular. Then the solution* $X(t)$ *of* (5) *converges to* $\tilde{Y}$ *as*

$$\big\|X(t) - \tilde{Y}\big\|_F \leq Ce^{-4t\sigma_{min}(Y)}, \tag{10}$$

where $\sigma_{min}(Y)$ *is the smallest eigenvalue of* Y *in absolute value, and* $C > 0$ *is a constant.*

It follows from Proposition 2 that the convergence result for symmetric matrix factorization with spectral initialization can be extended to any positive semidefinite initialization X_0, provided $I + \hat{Y}^{-1}\big(X_0 - \tilde{Y}\big)$ is invertible. This is an extension of the previous assumption on X_0, namely that $\mathrm{rank}(\Sigma_0 \tilde{\Sigma}) = \mathrm{rank}(\tilde{\Sigma})$. The main difference is that, in the spectral initialization case we can derive the convergence rate for each eigenvalue of $X(t)$, while in general we can only obtain a global convergence rate of $X(t)$. We note that conditions on $I + \hat{Y}^{-1}\big(X_0 - \tilde{Y}\big)$ being nonsingular are almost surely satisfied by random initializations. They merely characterize a few pathological initializations that lead to suboptimal solutions. For example, in the case of spectral initialization they suggest that it is impossible to learn a nonzero component starting from zero initialization. Therefore, such conditions impose no practical limitations.

3. Asymmetric Matrix Factorization with Spectral Initialization

In this section we analyze the dynamics of gradient flow for a more general—asymmetric—matrix factorization. We transform the dynamics to a canonical form and show that the solutions under the spectral initialization are diagonal and can be computed in closed form. This solution reveals a convergence rate $O\big(e^{-t\sqrt{4\sigma_i^2 + \lambda_{0,i}^2}}\big)$, where σ_i is the ith singular value of Y and $\lambda_{0,i}$ defines the level of imbalance in the initialization for the ith component. As in the symmetric case, the components of the solution are learned at different rates, however, in the asymmetric formulation the imbalance at initialization also plays a role and changes the rate at which different components are learned.

Asymmetric matrix factorization model. Consider

$$\min_{U,V} \ell(U, V), \qquad \ell(U, V) \equiv \tfrac{1}{2}||Y - UV^T||_F^2, \tag{11}$$

where $U \in \mathbb{R}^{m \times k}$, $V \in \mathbb{R}^{n \times k}$ and $k \geq n \geq m$. The gradient flow thus takes the form

$$\begin{aligned} \dot{U} &= -\nabla_U \ell = (Y - UV^T)V, \\ \dot{V} &= -\nabla_V \ell = (Y - UV^T)^T U. \end{aligned} \tag{12}$$

We will make use of a conservation law for the difference of the Gramian matrices $U^T U$ and $V^T V$ (this has also been identified by Arora et al. (2018); Du et al. (2018a)). Previous works have used this conservation law to ensure balancedness under vanishingly small initialization. In contrast, our analysis highlights the role of imbalance in the dynamics, e.g., in the convergence rate of the gradient flow, which has not been previously considered or even noticed.

Conservation law. A straightforward calculation shows that (12) admits an invariant:

$$\begin{aligned} &\mathcal{Q} \equiv U^T U - V^T V, \\ &\frac{d\mathcal{Q}}{dt} = \dot{U}^T U + U^T \dot{U} - \dot{V}^T V - V^T \dot{V} = 0 \end{aligned} \tag{13}$$

so that $\mathcal{Q}(t) = \mathcal{Q}(0)$. The origin behind this conserved quantity $\mathcal{Q}$ is a global rotational symmetry of (12), i.e., the system is invariant under the orthogonal group $O(k)$. To see this, consider the singular value decomposition $Y = \Phi \Sigma \Psi^T$ and, following Saxe et al. (2019), define $\bar{U}$ and $\bar{V}$ through

$$U \equiv \Phi \bar{U} G^T, \qquad V \equiv \Psi \bar{V} G^T, \tag{14}$$

where G is an arbitrary element of $O(k)$. These transformed variables obey

$$\dot{\bar{U}} = (\Sigma - \bar{U}\bar{V}^T)\bar{V}, \qquad \dot{\bar{V}} = (\Sigma - \bar{U}\bar{V}^T)^T \bar{U}, \tag{15}$$

which have *exactly the same form* as (12) up to a gauge freedom on the choice of G. Since $\mathcal{Q}$ is real and symmetric, it is diagonalizable by an orthogonal matrix. Therefore, we can choose G to be the matrix that diagonalizes $\mathcal{Q}$. Hence, from (13) we have

$$\begin{aligned} \bar{U}^T \bar{U} - \bar{V}^T \bar{V} = G^T \mathcal{Q}(t) G = \Lambda_{\mathcal{Q}} \\ = \Lambda_{\mathcal{Q}_0} = \bar{U}_0^T \bar{U}_0 - \bar{V}_0^T \bar{V}_0, \end{aligned} \tag{16}$$

where $\Lambda_{\mathcal{Q}_0}$ is the (constant) *diagonal matrix* containing the k eigenvalues of $\mathcal{Q}_0 \equiv \mathcal{Q}(0)$ (or $\mathcal{Q}(t)$) which is completely specified by the initial conditions U_0 and V_0 alone. Note that the number of conserved quantities in $\Lambda_{\mathcal{Q}_0}$ depends on k, which is equal to the degree of overparameterization. Although we do not assume balanced initialization in this paper, for further reference and comparison with prior work (Arora et al., 2018; Saxe et al., 2014; 2019), let us state its precise meaning since it relates to the conservation law.

Definition 2 (Balanced initialization). (U_0, V_0) *is said to be balanced if* $\|\mathcal{Q}(t)\|_F = \|\mathcal{Q}_0\|_F \leq \epsilon$ *for sufficiently small* $\epsilon > 0$, *i.e., the conserved quantity in* (13) *is small.*

Under the above transformation, the matrix factorization problem with spectral initialization can be reduced to solving k one-dimensional systems (one for each component). Proposition 3 below provides a closed-form solution and explicitly characterizes the evolution of each component.

Definition 3 (Asymmetric Spectral initialization). *Let $Y = \Phi\Sigma\Psi^T$ be the SVD of the data. The spectral initialization is defined as $U_0 = \Phi\bar{U}_0 G$, $V_0 = \Psi\bar{V}_0 G$, and $X_0 = U_0 V_0^T$, where $\bar{U}_0$ and $\bar{V}_0$ are rectangular diagonal matrices and G is any orthogonal matrix.*

Proposition 3 (Exact solution and convergence rate). *Let $Y = \Phi\Sigma\Psi^T = \sum_{i=1}^m \sigma_i \phi_i \psi_i^T$ be the SVD of the data. The solution to (12) with spectral initialization $X_0 = \Phi\Sigma_0\Psi^T = \sum_{i=1}^m \sigma_{0,i}\phi_i\psi_i^T$ yields $X(t) = U(t)V(t)^T = \Phi\Sigma(t)\Psi^T = \sum_{i=1}^m \sigma_i(t)\phi_i\psi_i^T$ where $\sigma_i(t)$ is given by*

$$\frac{\sigma_i e^{2t\sqrt{4\sigma_i^2+\lambda_{0,i}^2}} - 2C_i\lambda_{0,i}^2 e^{t\sqrt{4\sigma_i^2+\lambda_{0,i}^2}} - 4\sigma_i\lambda_{0,i}^2 C_i^2}{e^{2t\sqrt{4\sigma_i^2+\lambda_{0,i}^2}} + 8\sigma_i C_i e^{t\sqrt{4\sigma_i^2+\lambda_{0,i}^2}} - 4\lambda_{0,i}^2 C_i^2}, \quad (17)$$

$\lambda_0 = \mathrm{diag}(\bar{U}_0^T\bar{U}_0 - \bar{V}_0^T\bar{V}_0)$, and $C_i = C_i(\sigma_i, \lambda_{0,i}, \sigma_{0,i})$ is a constant. Moreover, the ith eigencomponent of $X(t)$ converges to the ith eigencomponent of Y at a rate $O\big(e^{-t\sqrt{4\sigma_i^2+\lambda_{0,i}^2}}\big)$.

Proof. Under the spectral initialization, $\bar{U}_0$ and $\bar{V}_0$ are diagonal, thus so are $\dot{\bar{U}}(0)$ and $\dot{\bar{V}}(0)$. Consequently, $\dot{\bar{U}}(t)$ and $\dot{\bar{V}}(t)$ remain diagonal for all $t \geq 0$ since the components of (15) can be decoupled and the evolution will induce no change in off-diagonal elements. To see this, observe that

$$\dot{\bar{U}}_{ii} = (\sigma_i - \bar{U}_{ii}\bar{V}_{ii})\bar{V}_{ii}, \quad \dot{\bar{V}}_{ii} = (\sigma_i - \bar{U}_{ii}\bar{V}_{ii})\bar{U}_{ii}, \quad (18)$$

for all $1 \leq i \leq m$, while off-diagonal terms obey $\dot{\bar{U}}_{ij} = 0$ and $\dot{\bar{V}}_{ij} = 0$ $(i \neq j)$. Thus (18) describes the evolution of the *singular values of the solution*. This decouples the problem into a set of independent one-dimensional equations. Therefore, it suffices to consider the scalar system

$$\dot{\bar{u}} = (\sigma - \bar{u}\bar{v})\bar{v}, \qquad \dot{\bar{v}} = (\sigma - \bar{u}\bar{v})\bar{u}, \quad (19)$$

where we drop the index $i = 1, \ldots, m$ for simplicity. Saxe et al. (2019) makes the strong assumption $\bar{u}_{ii}(0) = \bar{v}_{ii}(0)$ for all i, which is a *balanced initialization* (Definition 2). Here instead we solve (18) without such an assumption. From (19), it is immediate that the conservation law (13) becomes $\frac{d}{dt}\left(\bar{u}^2 - \bar{v}^2\right) = 0$. Trajectories $(\bar{u}(t), \bar{v}(t))$ are thus constrained to lie on hyperbolas:

$$\bar{u}^2(t) - \bar{v}^2(t) = \bar{u}_0^2 - \bar{v}_0^2 = \lambda_0 = \text{const.} \quad (20)$$

Since we are mostly interested in the behavior of the product $x(t) = \bar{u}(t)\bar{v}(t)$, by making explicit use of the conservation law (20), i.e., $\lambda_0^2 = \bar{u}^4 - 2\bar{u}^2\bar{v}^2 + \bar{v}^4 = \bar{u}^4 + \bar{v}^4 - 2x^2$ and $(\bar{u}^2 + \bar{v}^2)^2 = \lambda_0^2 + 4x^2$, we obtain

$$\begin{aligned} \dot{x} &= (\sigma - \bar{u}\bar{v})\bar{v}^2 + (\sigma - \bar{u}\bar{v})\bar{u}^2 \\ &= (\sigma - \bar{u}\bar{v})\sqrt{(\bar{v}^2 + \bar{u}^2)^2} \\ &= 2(\sigma - x)\sqrt{x^2 + \lambda_0^2/4}. \end{aligned} \quad (21)$$

Even though this is a nonlinear differential equation, it is separable, thus integrating both sides yields precisely (17) (we restore i, and $x \to \sigma_i$ represents the corresponding component associated with singular value σ_i and conserved quantity $\lambda_{0,i}$), where $C > 0$ is a constant given by

$$\frac{\sqrt{4\sigma^2\lambda_0^2 + 16\sigma^2 x_0^2 + \lambda_0^4 + 4x_0^2\lambda_0^2} - 4\sigma x_0 - \lambda_0^2}{4\lambda_0^2(\sigma - x_0)}. \quad (22)$$

□

Above, only m out of $k \geq m$ conserved quantities are used. Hence, there is degeneracy in the solution and only m effective degrees of freedom regardless how large k is. Note that if $k < m$—underparameterized case—then (18) becomes underdetermined.

Moreover, recall from (4) that the convergence rate for the non overparameterized problem in (2) is $O(e^{-t})$, which does not depend on the data or the initialization. It follows from (17) that the asymptotic behavior of the singular values of the overparameterized solution is

$$|\sigma_i(t) - \sigma_i| \simeq 2C_i(4\sigma_i^2 + \lambda_{0,i}^2)e^{-t\sqrt{4\sigma_i^2+\lambda_{0,i}^2}}, \quad (23)$$

which depends on both σ_i (singular values of the data) and $\lambda_{0,i}$ (level of initialization imbalance). When the initialization is balanced, i.e., $\lambda_{0,i} \approx 0$, we recover the results of Gidel et al. (2019); Saxe et al. (2019) about the sequential learning of components. We note that this is similar to the symmetric case, which is not surprising since a symmetric factorization is by construction balanced. However, when the initialization is not balanced, the eigenvalues of $\Lambda_{\mathcal{Q}_0}$ also play a role and can make smaller components converge faster than larger ones.

4. Asymmetric Matrix Factorization without Spectral Initialization

We now relax the assumption of spectral initialization (Definition 3). Defining the quantities

$$R(t) \equiv \begin{bmatrix} \bar{U}(t) \\ \bar{V}(t) \end{bmatrix}, \quad S \equiv \begin{bmatrix} 0 & \Sigma \\ \Sigma^T & 0 \end{bmatrix}, \quad \bar{S} \equiv \frac{1}{2}\begin{bmatrix} I_m & 0 \\ 0 & -I_n \end{bmatrix}, \quad (24)$$

one can immediately obtain from (15) and (16) a Riccati-like differential equation:

$$\dot{R} = SR - \tfrac{1}{2}RR^TR + \bar{S}R\Lambda_{\mathcal{Q}_0}, \quad (25)$$

where from (16) we conclude that $2R_0^T\bar{S}R_0 = \Lambda_{\mathcal{Q}_0}$ with $R(0) \equiv R_0$. However, in general, one cannot go back from (25) to (15) unless the conservation law (16) is explicitly imposed for all times $t \geq 0$. The natural question is then, when are they equivalent? Our next result provides the

answer, and additionally reveals an interesting relation between (25) and a matrix factorization problem with *explicit regularization* (the proof is in Appendix D).

Proposition 4 (Explicit regularization). *The differential equation (25) is equivalent to*

$$\begin{aligned}\dot{\bar{U}} &= (\Sigma - \bar{U}\bar{V}^T)\bar{V} - \tfrac{1}{2}\bar{U}(\bar{U}^T\bar{U} - \bar{V}^T\bar{V} - \Lambda_{\mathcal{Q}_0}),\\ \dot{\bar{V}} &= (\Sigma - \bar{U}\bar{V}^T)^T\bar{U} + \tfrac{1}{2}\bar{V}(\bar{U}^T\bar{U} - \bar{V}^T\bar{V} - \Lambda_{\mathcal{Q}_0}).\end{aligned} \tag{26}$$

This system corresponds to the dynamics of gradient flow applied to the regularized problem

$$\min_{\bar{U},\bar{V}} \left\{ \tfrac{1}{2}||\Sigma - \bar{U}\bar{V}^T||_F^2 + \tfrac{1}{8}||\bar{U}^T\bar{U} - \bar{V}^T\bar{V} - \Lambda_{\mathcal{Q}_0}||_F^2 \right\}. \tag{27}$$

Moreover, if $\bar{\mathcal{Q}}(t) \equiv \bar{U}^T(t)\bar{U}(t) - \bar{V}^T(t)\bar{V}(t)$ obeys $\bar{\mathcal{Q}}(t_0) = \Lambda_{\mathcal{Q}_0}$ at some $t = t_0$, then $\bar{\mathcal{Q}}(t) = \Lambda_{\mathcal{Q}_0}$ for all $t \geq t_0$. In particular, if we initialize (26)—or equivalently (25)—such that $\bar{\mathcal{Q}}(0) = 2R_0^T\bar{S}R_0 = \Lambda_{\mathcal{Q}_0}$, then the conservation law (16) holds true for all t and the dynamics of both (26) and (15) are the same.

Let us stress a few points:

- For any orthogonal matrix $G \in O(k)$, under definitions (14) and with $\mathcal{Q}_0 \equiv G^T\Lambda_{\mathcal{Q}_0}G$, problem (27) is equivalent to

$$\min_{U,V} \left\{ \tfrac{1}{2}||Y - UV^T||_F^2 + \tfrac{1}{8}||U^TU - V^TV - \mathcal{Q}_0||_F^2 \right\}. \tag{28}$$

- A solution to (15) implies a solution to (26): When (13) holds the 2nd terms on the RHS of (26) vanish, while the 1st terms are exactly (15). However, the converse is not necessarily true, unless (26) is initialized in the same way as (15). Proposition 4 relates the *conservation law* to an *explicit regularization* (see Proposition 5 below), namely, one can either select a particular initialization and solve an unregularized problem, or start at an arbitrary initialization and explicitly regularize.

- The dynamics of (26) versus (15) are analogous to solving the explicitly regularized problem (27) versus the unregularized problem (11) subject to $\|U^TU - V^TV - \mathcal{Q}_0\|_F^2 = 0$–which in continuous-time is automatically satisfied thanks to the conservation law (16).

- The specific weight of $1/8$ in (27) is special: If one replaces $1/8$ by some constant $\alpha > 0$, the gradient flow dynamics, i.e., the analog of (26), will not be equivalent to (25). We note that problem (27) also appeared in (Du et al., 2018a) but without any of such connections.

Eq. (25) is reminiscent of a Riccati differential equation due to the cubic term in R (similar to the gradient flow in the symmetric case) but we believe that, in general, it cannot be solved exactly due to the last term. However, it can be solved exactly in a particular case (proof in Appendix C).

Proposition 5 (Exact solution and convergence rate). *If $\Lambda_{\mathcal{Q}_0} = \lambda_0 I_k$ for some constant λ_0, then the differential equation (25) reduces to*

$$\dot{R} = \tilde{S}R - \tfrac{1}{2}RR^TR \tag{29}$$

which yields a close-form solution for $R(t)R^T(t)$ equal to

$$e^{t\tilde{S}}R_0R_0^T\left(I + \tfrac{1}{2}\tilde{S}^{-1}(e^{2t\tilde{S}} - I)R_0R_0^T\right)^{-1}e^{t\tilde{S}} \tag{30}$$

where $\tilde{S} \equiv S + \lambda_0\bar{S} = \Phi\tilde{\Sigma}\Phi^T$, $R_0 \equiv R(0)$. Moreover, if $\tilde{S}$ and $I + \frac{1}{2}\hat{S}^{-1}(R_0R_0^T - R_\star R_\star^T)$ are invertible then $R(t)R^T(t)$ converges exponentially to $R_\star R_\star^T$, defined as the projection of the matrix $2\tilde{S}$ on the PSD cone, $\hat{S} = \Phi|\tilde{\Sigma}|\Phi^T$. More precisely, if Y is a square matrix the convergence rate is $O\big(e^{-t\sqrt{4\sigma_{\min}^2 + \lambda_0^2}}\big)$, where σ_{min} is the smallest eigenvalue of Y, and otherwise the rate is $O(e^{-|\lambda_0|t})$.

Note that Eq. (29) is nothing but the gradient flow for the symmetric factorization problem: $\min_R \left\{ \frac{1}{8}||2\tilde{S} - RR^T||_F^2 \right\}$. The particular case $\Lambda_{\mathcal{Q}_0} = \lambda_0 I_k$ is mathematically interesting because it is amenable to an analytical treatment. However, it may not be realizable in practice because the conserved quantity $\mathcal{Q}_0$ (or $\Lambda_{\mathcal{Q}_0}$) must have low rank, i.e.,

$$\begin{aligned}\text{rank}(\mathcal{Q}_0) &= \text{rank}(U_0^TU_0 - V_0^TV_0)\\ &\leq \text{rank}(U_0^TU_0) + \text{rank}(V_0^TV_0)\\ &\leq m + n.\end{aligned} \tag{31}$$

Since $\text{rank}(\lambda_0 I_k) = k$, choosing $\bar{U}_0$ and $\bar{V}_0$ such that $\bar{U}_0^T\bar{U}_0 - \bar{V}_0^T\bar{V}_0 = \lambda_0 I_k$ is not generally possible in an overparameterized setting with $k > m + n$. On the other hand, the choice $\Lambda_{\mathcal{Q}_0} = \lambda_0 I_k$ does not present a problem if we consider the system (26) where we have the freedom to choose any initialization. The experiments in Section 6 illustrate that Eq. (26), or equivalently Eq. (29), is actually enough to capture the general behaviour of system (15).

5. Discrete- versus Continuous-Time Rates

We provide an explicit example to illustrate why studying the continuous-time dynamics of the gradient flow is expected to reproduce the behaviour of its discretization, i.e., gradient descent. For simplicity, we limit the discussion to the case considered in Section 2 and in the scalar case $Y = \sigma \in \mathbb{R}$. What we would like to do is to compare two different algorithms, namely gradient descent applied to problem (2) versus gradient descent applied to the factorized problem (3). We thus have

$$X_{k+1} = X_k + \eta(\sigma - X_k) \tag{32}$$

versus

$$U_{k+1} = U_k + 2\eta(\sigma - U_k^2)U_k. \tag{33}$$

The respective continuum limits of these algorithms are

$$\dot{X} = (\sigma - X) \tag{34}$$

versus

$$\dot{U} = 2(\sigma - U^2)U. \tag{35}$$

The solution of (34) is $X(t) - \sigma = (X_0 - Y)e^{-t}$, yielding a rate $O(e^{-t})$. Define the perturbed variable

$$\tilde{X}_k \equiv X_k - \sigma. \tag{36}$$

Hence (32) gives $\tilde{X}_{k+1} = (1-\eta)\tilde{X}_k$, i.e., a matching rate of $O(e^{-\eta k})$; this example is trivial because both systems are linear. Now let us consider the more interesting nonlinear case. Consider (34) and let $X(t) \equiv U^2(t)$. Thus $\dot{X} = 4\sigma X - 4X^2$ whose solution is

$$X(t) = \frac{\sigma}{1 - ce^{-4\sigma t}} \approx \sigma - ce^{-4\sigma t}, \tag{37}$$

implying a continuous-time rate $O(e^{-4\sigma t})$—compare this with the last phrase in Corollary 1. Now let us see what happens for (33). This is a complicated nonlinear recurrence relation, but fortunately we can solve it approximately. By introducing $X_k \equiv U_k^2$ it becomes

$$X_{k+1} = X_k + 4\eta(\sigma - X_k)X_k + 4\eta^2(\sigma - X_k)^2 X_k. \tag{38}$$

Consider the perturbed variable (36). For sufficiently small η we can neglect terms of $O(\eta^2)$, hence

$$\tilde{X}_{k+1} \approx \tilde{X}_k - 4\eta\tilde{X}_k(\sigma + \tilde{X}_k) \leq (1 - 4\eta\sigma)\tilde{X}_k. \tag{39}$$

This implies $\tilde{X}_k \to 0$, or $X_k \to \sigma$, at a discrete-time rate of $O(e^{-4\sigma\eta k})$, which matches the continuous-time rate.

It is not hard to see how such nonlinear recurrence relations quickly become intractable for more complicated problems. On the other hand, even though the continuous-time limit provided by the gradient flow consists of a nonlinear ODE, the analysis is much more feasible besides introducing interesting mathematical connections.

6. Numerical Experiments

Imbalanced initialization. Here we provide numerical evidence to our theoretical results. First, we generate a random matrix Y with $Y_{ij} \sim \mathcal{N}(0,1)$ and set $m = 5$, $n = 10$ and $k = 50$. We approximate the dynamics of gradient flow for one-layer and two-layer linear models by using gradient descent with a step size $\eta = 10^{-3}$ (smaller step sizes did not lead to a discernible change). We evaluate the *reconstruction error* $\|Y - X(t)\|_F / \|Y\|_F$, where $X(t) = U(t)V^T(t)$, and compare the evolution of the singular values of $X(t)$.

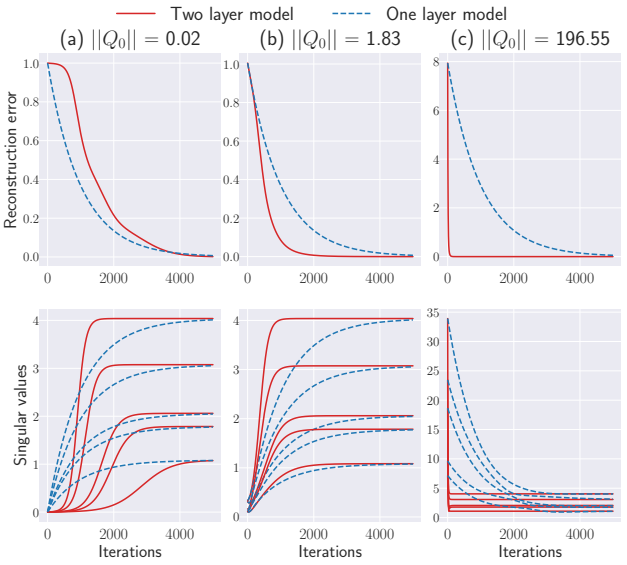

Figure 1. Top row: Reconstruction error for one- vs. two-layer linear models. *Bottom row:* Evolution of singular values. From left to right we use $\sigma = 10^{-2}$, $\sigma = 10^{-1}$, and $\sigma = 1$, respectively.

We consider Gaussian initializations, i.e., U_0 and V_0 have entries $\sim \mathcal{N}(0, \sigma^2)$ where σ is varied to obtain different degrees of imbalance. To start both models in the same state, we choose $X(0) = U_0 V_0^T$ for the one-layer case. The results are shown in Fig. 1. From our theoretical analysis, we expect a different behaviour for the convergence rate depending whether the initialization is balanced or imbalanced, i.e., whether $\|\mathcal{Q}\|_F = \|\mathcal{Q}_0\|_F \equiv \|U_0^T U_0 - V_0^T V_0\|_F$ is small or large, respectively. When it is very small (Fig. 1a) the strength of the singular values dominate and we expect the components to be learned sequentially from the largest to the smallest, in agreement with Gidel et al. (2019); Saxe et al. (2019). As we make the weights more imbalanced (Fig. 1b) the singular values are learned faster, even the smaller ones. Finally, as $\|\mathcal{Q}\|_F$ becomes very large, the imbalance becomes the dominating term in the convergence rate and the solution of the factorized problem converges significantly faster (Fig. 1c). In other words, these numerical results are consistent with Propositions 2 and 5.[4]

$\Lambda_{\mathcal{Q}_0} = \lambda_0 I$ is general enough. Since Proposition 5 contains the case where an exact solution is available, we want to investigate whether this is general enough to capture the qualitative behaviour of system (15). To avoid confusion, we refer to $\bar{U}^I$ and $\bar{V}^I$ as the variables of (26), as well as $\Lambda^I_{\mathcal{Q}_0} \equiv \lambda_0 I_k$; here I stands for "identity." The variables $\bar{U}$ and $\bar{V}$ refer to system (15), with its $\Lambda_{\mathcal{Q}}$ fixed by the initial

[4] We note that such a comparison is meaningful only if the step size is the same for both models, which should be small enough to ensure stability of the discretizations; increasing imbalance does not accelerate convergence in discrete-time because a smaller step size would be required.

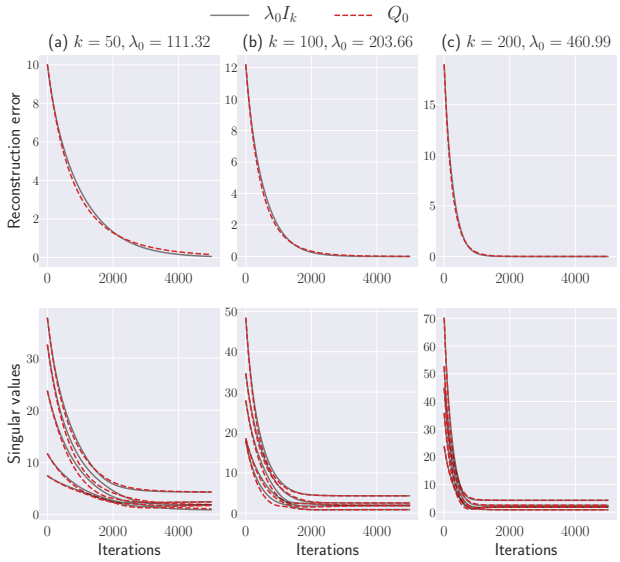

Figure 2. Top row: Reconstruction error for the asymmetric factorization dynamics without regularization in (15) and (16) and general $\mathcal{Q}_0$ (red dashed line), versus the regularized dynamics in (26) with a diagonal $\Lambda_{\mathcal{Q}_0} = \lambda_0 I_k$ (black solid line). *Bottom row:* Evolution of the corresponding singular values. From left to right we set $k = 50$, $k = 100$, and $k = 200$, respectively.

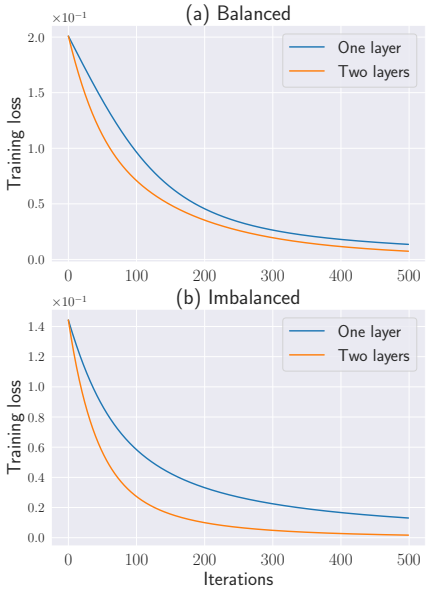

Figure 3. Evolution of the training loss for nonlinear one-layer and two-layer models. *Top row*: $||Q_0||_2 = 0$. *Bottom row*: $||Q_0||_2 = 4.6$. Initial weights are drawn from a normal distribution $\mathcal{N}(0, 10^{-1})$.

conditions; see (16). We want to show that it is possible to find an "optimal" $\lambda_0 \in \mathbb{R}$ such that both cases have similar dynamics. We thus initialize U_0 and V_0 (and equivalently $\bar{U}_0$ and $\bar{V}_0$) with entries $\sim \mathcal{N}(0, 1)$. The same initial condition is used for (26), i.e., $\bar{U}_0^I = \bar{U}_0$ and $\bar{V}_0^I = \bar{V}_0$. We set $\eta = 10^{-5}$, $Y \sim \mathcal{N}(0, 1)$, $m = 5$, $n = 10$ and vary k. We look for λ_0 that minimizes $\|X^I(t) - X(t)\|_F$. In Fig. 2 we illustrate that, indeed, this can be done. Note that the evolution of both systems are nearly indistinguishable. In practice, λ_0 does not need to be chosen. It is implicitly determined by the choice of initialization and typically grows with $\|\mathcal{Q}_0\|$. This is only relevant to theoretically understand the dynamics of the gradient flow on this problem.

Extension to nonlinear networks. Our analysis so far has shown that imbalance affects the convergence rate and is induced by a conservation law. However, its definition should change when introducing nonlinearities. In fact, both the network architecture and the objective function should affect these conserved quantities. As such, conducting a full analysis for more complex nonlinear networks is necessary to characterize the the dynamics and implicit bias in such cases, which we leave open for future work. Nonetheless, we provide some numerical evidence by adding nonlinearity (sigmoid) to the final layer. We train the two networks (one layer vs. two layers) on synthetic data, i.e., we compare the dynamics of gradient descent for the objectives $\ell_1(W) = \frac{1}{2}||Y - \phi(XW)||_F^2$ versus $\ell_2(U, V) = \frac{1}{2}||Y - \phi(XUV^T)||_F^2$, where ϕ is the sigmoid function, $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times d}$ represent the training samples and labels, respectively ($n = 10^3, d = 10$), $W \in \mathbb{R}^{d \times d}, U, V \in \mathbb{R}^{n \times k}$ are the weight matrices and the width is $k = 100$. We generated the matrices W^* and X with entries drawn from $\mathcal{N}(0, 1)$ and $Y = \phi(XW^*) + \epsilon$ where $\epsilon \sim 10^{-3}\mathcal{N}(0, I)$. The results in Fig. 3 interestingly suggest that our conclusions about the role of imbalance still hold in this case as well.

7. Conclusion

We considered the gradient flow dynamics on two-layer linear neural networks, providing an analytical treatment to a great level of detail. Our results establish a precise characterization of the solutions. Importantly, we do not assume balanced or vanishingly small initialization which so far have been present in all prior work in this vein.

Our analysis shows that the dynamics and convergence of the gradient flow is strongly related to an emerging rotational symmetry induced by overparameterization which gives rise to several conservation laws that constrain the dynamics to follow specific trajectories; such conserved quantities are completely fixed by the initialization. Our analysis focused on the simple case of linear networks, however, it reveals a potential key to understand implicit bias which lies in the conservation laws that arise from the symmetries of the problem. Such symmetries depend on the network architecture, objective function, optimization al-

gorithm, and they constrain the dynamics to an invariant manifold that encapsulates the implicit regularization and acceleration effects. Understanding this in more complex models may thus be reduced to finding dynamical invariants, for which our results provide a foundational starting point.

References

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning (ICML)*, 2018.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2019a.

Arora, S., Golowich, N., Cohen, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. In *7th International Conference on Learning Representations, ICLR 2019*, 2019b.

Callier, F. M., Winkin, J., and Willems, J. L. On the exponential convergence of the time-invariant matrix Riccati differential equation. In *Proceedings of the 31st IEEE Conference on Decision and Control*, pp. 1536–1537 vol.2, 1992.

Callier, F. M., Winkin, J., and Willems, J. L. Convergence of the time-invariant Riccati differential equation and lq-problem: mechanisms of attraction. *International Journal of Control*, 59(4):983–1000, 1994.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *arXiv preprint arXiv:2002.04486*, 2020.

Du, S. and Hu, W. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pp. 1655–1664, 2019.

Du, S. S., Hu, W., and Lee, J. D. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018b.

Gidel, G., Bach, F., and Lacoste-Julien, S. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1832–1841, 2018a.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 9461–9471, 2018b.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations (ICLR)*, 2019a.

Ji, Z. and Telgarsky, M. The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, pp. 1772–1798, 2019b.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, 2018.

Molinari, B. P. The time-invariant linear-quadratic optimal control problem. *Automatica*, 13(4):347–357, July 1977.

Sasagawa, T. On the finite escape phenomena for matrix Riccati equations. *IEEE Transactions on Automatic Control*, 1982.

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

Saxe, A. M., McClelland, J. L., and Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1820226116.

Soudry, D., Hoffer, E., and Srebro, N. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations (ICLR)*, 2018.