Abstract

This paper studies the behaviour of two-layer fully connected networks with linear activations trained with gradient flow on the square loss. We show how the optimization process carries an implicit bias on the parameters that depends on the scale of its initialization. The main result of the paper is a variational characterization of the loss minimizers retrieved by the gradient flow for a specific initialization shape. This characterization reveals that, in the small scale initialization regime, the linear neural network’s hidden layer is biased toward having a low-rank structure. To complement our results, we showcase a hidden mirror flow that tracks the dynamics of the singular values of the weights matrices and describe their time evolution. We support our findings with numerical experiments illustrating the phenomena.

1 Introduction

The most forceful driver of advancements in the field of Machine Learning over the past decades has been the success of deep neural networks. Amongst the striking qualities of these models is the fact that, despite being heavily overparametrized, their optimization consistently yields minima with good generalization properties. A beckoning research direction is thus to unravel the process through which neural networks learn internal representations for a given task [Bengio et al., 2013]. Understanding such phenomena is crucial for lowering the interpretability barrier of these models and developing a principled approach to their training and deployment in practice.

Recent experimental evidence identified one of the likely paths towards achieving these goals as the study of the inherent regularization properties (or implicit biases) of training algorithms [Neyshabur et al., 2014, Zhang et al., 2016]. These observations laid the foundation for a new line of work [see, e.g., Vardi, 2022] whose driving question is which minimum, amongst the many, awaits at the tail end of optimization.

One of the determining factors for the implicit bias of gradient methods is the initialization scale, which controls their operational regime as shown by empirical studies [Chizat et al., 2019]. More precisely, gradient descent with a low-scale initialization is capable of learning rich feature representations from the data. Strikingly, despite overparameterization, the hidden-layer neurons align in the direction of the features [Chizat et al., 2019, Atanasov et al., 2022] and learning of representations reflects in the low-rank structure of the hidden layers. Our work aims to precisely explain this phenomenon and quantify the impact of the initialization scale on feature learning.

Unfortunately, studying such phenomena for the types of neural networks used in practice is mathematically challenging at present due to the non-linearity of their activations. Their less expressive linear counterparts, however, are more tractable and represent a good proxy due to their non-convex
loss landscape and non-linear learning dynamics. Consequently, the study of deep linear networks has received significant amounts of attention over the past years, and spans several important directions, including convergence [Arora et al., 2019a, 2018, Min et al., 2021], learning dynamics [Saxe et al., 2014, Braun et al., 2022] and the implicit bias of optimization algorithms [Azulay et al., 2021]. This work complements these previous approaches by mathematically describing the properties of their parameters at convergence, highlighting the implicit bias phenomenon, and further analyzing the evolution of weight matrices throughout the optimization process.

Specifically, this paper studies overparameterized vector regression problems on two-layer fully-connected linear neural networks. We show the following results when the network is trained with gradient flow (GF).

(i) In Theorem 3.1, we prove that the zero-loss solutions retrieved by the gradient flow are the minimizers of a potential that depends on the initialization scale. Additionally, we provide explicit expressions for the singular values of the hidden layer weights, also as a function of the initialization scale. These characterizations reveal how low-magnitude initialization induces a low-rank structure of the hidden layer.

(ii) In Theorem 3.2, we show that gradient flow on the parameters induces a mirror flow on the singular values. In the specific case of scalar regression, we show that the gradient flow on the weights is equivalent to a mirror flow on the linear predictor. These characterizations give the geometrical structure of the training dynamics of linear neural networks.

(iii) In Proposition 4.1, we design a simple process to analytically describe how stochastic noise in the training algorithm can likewise induce low-rank structures in the weights regardless of the initialization scale.

We proceed by presenting related work in Section 1.1, formalizing the problem setup and assumptions in Section 2, stating and discussing our results in Section 3, and finally, we provide supporting numerical evidence in Section 4.

1.1 Related Work

The first pillar of our work addresses the implicit bias of GF and its stochastic variant in regression problems. One of the hallmarks of bias in this setting is the impact of initialization scale: large initial weights induce a learning regime in which the parameters travel a short distance to convergence and feature learning fails to happen (lazy regime), while small initialization effects a polar opposite behaviour of the system (rich regime) Chizat et al. [2019], Woodworth et al. [2020]. Training and generalization in the lazy regime are well-studied [Jacot et al., 2018, Du et al., 2019a, Arora et al., 2019c, Soltanolkotabi et al., 2017], however this scenario fails to capture the observed behaviour of neural networks in practice [Ghorbani et al., 2019]. While the rich regime more faithfully approximates the feature learning abilities of these models, it is comparatively more challenging to analyze and few results are known. Amongst them are those concerning diagonal linear networks, where a preference towards sparse representations is shown [Woodworth et al., 2020], and a restricted setting of the matrix factorization problem, where the implicit bias leads to low-rank representations [Gunasekar et al., 2017, Arora et al., 2019b, Li et al., 2018]. We similarly study the rich representation learning regime and provide initialization scale-dependent statements on implicit bias for two-layer fully connected linear networks.

Most theoretical results on the implicit bias of GF in overparametrized models rely on the identification of a related mirror flow in a reparametrized space [Gunasekar et al., 2018]. Diagonal linear networks are amenable to this technique and therefore well-studied [Woodworth et al., 2020, Pesme et al., 2021]. For linear fully-connected networks, however, the existence of a mirror flow is not always guaranteed [Li et al., 2022]. To partly alleviate this issue, Azulay et al. [2021] introduce a nonlinear time-rescaling technique and show that for scalar least-squares regression on a two-layer fully connected network with zero-balance initialization, the implicit bias selects low \( \ell_2 \)-norm predictors. We prove a similar result under imbalanced initialization controlled by a scale parameter, and characterize the weight matrices independently at convergence, thus presenting a higher-resolution view of the problem.

Other works on linear networks include [Min et al., 2021] where convergence is studied in the presence of weight imbalance and implicit bias results are provided in the functional space; and [Yun
where tensor networks are studied with the goal of unifying the implicit bias results for linear parameterization. In the case of linear networks, Yun et al. [2021] further show an implicit bias towards minimum ℓ₂ linear predictors for vanishingly small initializations. For classification in the case of linear networks, Ji and Telgarsky [2019] show that the weights grow to infinity and the layers of the deep linear network align during the course of optimization. Timor et al. [2023] show a similar phenomenon happens for two-layer ReLU networks.

The second pillar concerns the learning dynamics of linear neural networks. The two-layer case optimized with GF on the square loss has been studied by Fukumizu [1998], Saxe et al. [2014, 2019], Braun et al. [2022]. The common setup of these works is that of zero-balance initialization and whitened data. First, Saxe et al. [2014, 2019] provide expressions for the temporal evolution of singular values of the weight matrices. Our work removes the requirement of zero-balanced initialization and full-rank network initialization, and gives formulas for the weights’ time-evolution as a function of the initialization scale. We further provide mirror flows on the weights’ singular values and show that components are learned in a hierarchical manner for the case of whitened data.

Related work is further addressed in the following sections, as part of the discussion of results.

2 Preliminaries and problem setup

Notation. Time-dependent variables are written in bold fonts: we drop the $t$ in $A(t)$ and simply denote it as $A$. The time derivative of such variables is denoted $\frac{d}{dt}A(t)$ as $\dot{A}$.

Vector Regression. The set-up is that of standard vector regression problems with inputs $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ and outputs $(y_1, \ldots, y_n) \in (\mathbb{R}^k)^n$ in the so-called overparametrized regime where $d \geq n$. Regarding the output dimension, the reader may keep in mind throughout the article that $k \ll d$, though the analysis holds for any $k, d$ pair. In order to learn the input/output rule, we minimize the square loss over a class of parametric models $\mathcal{H} = \{f_\theta(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^k \mid \theta \in \mathbb{R}^p\}$ which we specify in the next paragraph. The train loss therefore can be written as

$$
\mathcal{L}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2. \tag{2.1}
$$

Parameterization with a Linear Network. We consider the parametric model of two-layer linear neural networks of width $l \in \mathbb{N}^+$: this corresponds to the parametrization $\theta = (W_1, W_2)$, $W_1 \in \mathbb{R}^{d \times l}$, $W_2 \in \mathbb{R}^{l \times k}$ and $f_\theta(x) = W_2^\top W_1^\top x$. The model is linear in the input $x$, and in terms of expressivity, it is equivalent to the linear class of predictors given by $f_\beta(x) = \beta^\top x$, with $\beta = W_1 W_2$. We henceforth use the symbol $\beta$ to denote the associated linear predictor of the network. An important consequence of this reparametrization is that the prediction function $f_\theta$ is positively homogeneous of degree 2 in $\theta$: $\forall \lambda \in \mathbb{R}$, it holds that $f_{\lambda \theta} = \lambda^2 f_\theta$, as it is the case for two-layer ReLU networks. This property has important consequences in the loss landscape through which $\theta$ goes.

Train loss. Assume momentarily that $k = 1$ and denote $\phi(x) = W_1^\top x \in \mathbb{R}^l$. It is then clear that the predictor rewrites as $f_\theta(x) = \langle \phi(x), W_2 \rangle$. For this reason, we call the hidden layer $W_1$ the feature layer and the last layer $W_2$ the weight matrix. We study the overparametrized setting where $l \gg d$. Letting $X^\top := [x_1, \ldots, x_n]$ and $Y^\top := [y_1, \ldots, y_n]$, the loss becomes

$$
\mathcal{L}(W_1, W_2) = \frac{1}{2N} \|X W_1 W_2 - Y\|^2. \tag{2.2}
$$

For brevity, we ignore the $N$ in Eq. (2.2) by implicitly rescaling the data as $(X,Y) \leftarrow (X/\sqrt{N}, Y/\sqrt{N})$. 


We emphasize a crucial point: even if the function \( f \) is non-convex, there always exists a solution which attains zero loss, i.e., \( \beta^* \in \mathbb{R}^{d \times k} \) such that \( X\beta^* = Y \). We emphasize the fact that there are two levels of overparameterization here: on one hand, when \( d > n \), the set of zero loss linear predictors \( \mathcal{I}_\beta := \{ \beta \in \mathbb{R}^{d \times k} \mid X\beta = Y \} \) is typically an affine set of dimension \((d-n)k\). On the other hand, since we also reparametrize \( \beta \) as a linear network of width \( l \gg d \), the manifold of interpolators in the reparametrized space of \( \theta \), defined by \( \mathcal{I}_\theta := \{ \theta = (W_1, W_2) \mid W_1W_2 \in \mathcal{I}_\beta \} \), is of dimension \( l(d+k) - nk \). A natural question, therefore, is to which of these interpolators \( \theta^* \in \mathcal{I}_\theta \) does a given optimization algorithm converge. This concept is referred to as the implicit bias of an algorithm. The aim of this work is to study that of gradient flow.

### Gradient Flow

The dynamics induced in parameter space by running gradient flow on (2.2) is given by

\[
\dot{\theta} = -\nabla_{\theta} \mathcal{L}(\theta).
\]  

We wish to describe the implicit regularization properties of this continuous-time process, which is the vanishing stepsize limit of (stochastic) gradient descent. While the latter algorithms incur additional regularization properties from using non-zero stepsizes [Keskar et al., 2017], the study of GF is an important stepping stone to understanding the implicit bias of gradient-based methods in practice. In terms of \( W_1, W_2 \) the dynamics translates to

\[
\begin{align*}
\dot{W}_1 &= X^T (Y - XW_2W_2^T) W_2^T, \\
\dot{W}_2 &= W_1^T X^T (Y - XW_2W_2^T).
\end{align*}
\]

We emphasize a crucial point: even if the function \( \beta \rightarrow \|X\beta - Y\|^2 \) is convex, its reparametrization in terms of \( W_1, W_2 \) is not. Non-convexity and non-linearity makes the analysis challenging and a priori it is not even clear whether the time evolution of \( \beta \) can be expressed as a closed system.

### Initialization

One of our primary objects of study is the impact of initialization on the behaviour of GF. We describe here our initialization choice, to which we henceforth refer as \( I_\gamma \).

(a) **Orthogonal feature layer:** We initialize the inner layer such that the rows of \( W_1 \) are orthogonal and scale with parameter \( \gamma > 0 \). Mathematically, this translates to \( W_1(0) = \sqrt{2\gamma}P \) for \( P \in \mathbb{R}^{d \times l} \) in the Stiefel manifold \( V_d(\mathbb{R}^l) := \{ P \in \mathbb{R}^{d \times l} \mid PP^T = I_d \} \). Initializing with an orthogonal matrix is studied by Pennington et al. [2018], Hu et al. [2020], however from an optimization perspective. Note that when \( l \) is very large, this setting approximates the real-world scenario of initializing the hidden neurons with \( d \) i.i.d. Gaussian vectors in \( \mathbb{R}^l \), which are known to be almost orthogonal.

(b) **Zero weight layer:** In order to remove any initialization bias from the linear layer, we initialize it at \( W_2(0) = 0 \). This can be seen as the limiting case of initializing the weight layer with a very small relative scale \( \gamma \ll \gamma \).

As already mentioned in Section 1.1, existing studies on linear networks assume a “zero-balance initialization”, namely that \( W_1^T(0)W_1(0) = W_2(0)W_2^T(0) \) [Saxe et al., 2014, Arora et al., 2019b, Azulay et al., 2021]. This condition introduces the invariant \( W_1^TW_1 = W_2^TW_2 \) [Du et al., 2018], which holds for all times \( t \geq 0 \). This balancedness can be seen as a degeneracy assumption on the flow, since it implies that \( W_1 \) has at most rank \( k \) during the entire process, irrespective of the scale of initialization \( \gamma \). In contrast, we show that depending on \( \gamma \) the feature layer \( W_1 \) is biased (or not) toward a low-rank predictor, thus unveiling a truly rich representation learning regime.

\section{Main result: implicit bias and dynamics description}

### 3.1 Implicit bias on the parameters

Non-convex gradient flows are generally not guaranteed to reach global minimizers of the objective and even when they do, such results are difficult to formally prove. Moreover, the existence of many zero-loss solutions with different generalization properties raises the question of which interpolating
network is yielded by training. An elegant answer to such questions is to express the resulting predictor as the optimum among all the possible interpolators of some new, a priori unspecified cost. In addition to the descriptive power of such variational formulations, they express a form of capacity control over the estimator which can be further used to describe its generalization abilities [Bartlett et al., 2020]. The following theorem adds to this series of works, by precisely deriving such a characterization for GF in the setting of linear networks.

**Theorem 3.1.** Let \((W_1, W_2)\) be the process that follows the GF equations (2.4a)-(2.4b), initialized according to condition \(I_\gamma\), for some \(\gamma > 0\). Then

(i) The parameters converge to a global optimum of the loss

\[
\lim_{t \to \infty} (W_1(t), W_2(t)) = (W_1^\infty, W_2^\infty) \in \mathcal{I}_\beta.
\]

(ii) The linear predictor \(\beta\) converges to the minimum \(\ell_2\)-norm interpolator

\[
\lim_{t \to \infty} \beta(t) = \arg\min_{X\beta=Y} \|\beta\|_F \overset{\text{def}}{=} \beta^*.
\]

(iii) We have the following variational characterization of the limiting parameters

\[
(W_1^\infty, W_2^\infty) \in \arg\min_{XW_1=Y} \frac{1}{2} \|W_2\|_F^2 + \frac{1}{2} \|W_1\|_F^2 - \gamma \log \left( \det \left( W_1W_1^\top \right) \right). \quad (3.1)
\]

**Interpretation of the theorem.** The theorem is divided into three parts which state that (i) the matrices converge to a zero loss solution, which is a priori non-trivial since the loss in non-convex; (ii) among all the interpolators in \(\mathcal{I}_\beta\), \(\beta\) converges to the minimum \(\ell_2\)-norm interpolator for all \(\gamma > 0\); and (iii) among all the interpolators in \(\mathcal{I}_\beta\), \((W_1, W_2)\) converge to the ones that minimize a \(\gamma\)-dependent potential. To fully capture the richness of this result, we observe that in the limit of \(\gamma \to 0\), problem (3.1) formally translates to [Attouch, 1996]

\[
\lim_{\gamma \to 0} (W_1^\infty, W_2^\infty) \in \arg\min_{XW_1=Y} \frac{1}{2} \|W_2\|_F^2 + \frac{1}{2} \|W_1\|_F^2.
\]

This is equivalent, in the space of linear predictors \(\beta\) to the minimum nuclear norm solution \(\beta \in \arg\min_{X\beta=Y} \|\beta\|_*\) (which is also the minimum \(\ell_2\)-norm interpolator for the problem we study). We informally derived this interpretation by taking first the limit \(t \to \infty\) and only after \(\gamma \to 0\). The theorem naturally does not hold if the two limits are reversed, since \(\gamma = 0\) places the initialization at a saddle point of the loss, which is a stationary point of the flow.

With increasing \(\gamma\), we move towards solutions with a large \(\log \det \left( W_1W_1^\top \right)\), which is a smooth approximation of the rank [Fazel et al., 2003]. Intuitively, this means that solutions with increasing rank are preferred as \(\gamma\) grows. This scale-induced implicit bias is reminiscent of the rich and lazy regimes [Chizat et al., 2019, Woodworth et al., 2020], albeit visible in the space of representations rather than in that of predictors. As such, our result for linear networks is akin to Woodworth et al. [2020]'s, which characterizes the rich and lazy regimes for simpler diagonal linear networks.

**Comparison with works on implicit bias of \(\beta\).** Azulay et al. [2021], Min et al. [2021] also study the implicit bias phenomenon in linear networks, however, these results only address the structure of the final predictor \(\beta\) and not that of the factorized problem \((W_1, W_2)\). As shown in Theorem 3.1, these works fall short of unveiling all the nuances of the implicit regularization induced by GF in the case of linear networks.

To give a precise example, consider the simplest case of scalar regression \((k = 1)\) for which both Theorem 3.1 and [Azulay et al., 2021] show that \(\beta\) is biased towards low-\(\ell_2\) interpolators. This view is not complete, since there exist many pairs \((W_1, W_2)\) such that \(W_1W_2 = \beta\). Theorem 3.1 goes one step further and provides variational characterization of \((W_1, W_2)\) at convergence. Moreover, it shows that when \(\gamma \to 0\) all columns of \(W_1\) align in the direction of \(\beta\), thus creating a rank one hidden layer. This is an example of rich representation learning, where \(W_1\) is learning the only feature needed to make a prediction.
**Implicit bias of the singular values.** Proceeding with the description of the spectral bias, we provide a characterization of $W_1$’s limiting singular values which highlights their dependence on $\gamma$.

**Corollary 3.1.** [Singular values at the limit] Using the same quantities as in Theorem 3.1, denote $(\sigma_1(W_1^\infty), \ldots, \sigma_d(W_1^\infty))$ and $(\sigma_1(\beta_\ast), \ldots, \sigma_k(\beta_\ast))$ the singular values of $W_1^\infty$, $\beta_\ast$ are

$$
\sigma_i(W_1^\infty) = \left(\sqrt{\sigma_1(\beta_\ast)^2 + \gamma^2 + \gamma}\right)^{1/2}, \text{ for } 1 \leq i \leq k.
$$

$$
\sigma_i(W_1^\infty) = (2\gamma)^{1/2}, \text{ for } k < i \leq d.
$$

**Discussion.** Similar expressions for the singular values of $W_2$ can be derived and are deferred to the Appendix B due to lack of space. In the vanishing initialization limit, only the first $k$ singular values are activated and $W_1$ resembles a rank $k$ matrix. Conversely, for large $\gamma$ all the singular values are approximately equal in scale and $W_1$ resembles an isotropic full-rank matrix.

To ease interpretation, we again focus on the case of scalar regression. Corollary 3.1 shows that, for small $\gamma$, only one singular value grows while the others remain small constants. More precisely, when $\gamma \sim o(\|\beta_\ast\|)$, the training model is approximately rank one, with only one spiked singular value. Conversely, when $\gamma \sim \Omega(\|\beta_\ast\|)$ the low-rank structure disappears. This result perfectly captures the rich learning at low initialization where the hidden layer learns the defining feature of the problem, whereas in the lazy regime (large $\gamma$) the singular values of the matrix hardly move and no structure is present in $W_1$.

Our analysis removes the assumptions of balanced/spectral initialization and whitened data of previous works studying the evolution of singular values [Saxe et al., 2014, Gidel et al., 2019], thus allowing us to reveal the dependence on the scale of initialization.

### 3.2 Description of the dynamics

So far we have described the structure of the parameters of the neural network at convergence. Here, we show that the dynamics of the singular values of $\beta$ enjoy a very particular property: it satisfies a mirror flow [Alvarez et al., 2004] with a mirror potential that can be written explicitly.

**Theorem 3.2.** [Dynamics of the flow] With the same notations as in Theorem 3.1,

(a) **Mirror on singular values:** The singular values of $\beta$, denoted by $D_{\beta}$, follow the mirror flow

$$
d\nabla_{\gamma} (D_\beta) = -\nabla_{D_\beta} \mathcal{L} \ dt,
$$

where the potential writes $\Psi_{\gamma} (D_\beta) := \text{tr} \left( \frac{1}{2} D_\beta \sinh^{-1} (D_\beta/\gamma) - \sqrt{D_\beta^2 + \gamma^2} \right)$.

(b) **Mirror on $\beta$.** If $k = 1$, the dynamics of $\beta$ can be characterized as a mirror flow

$$
d\nabla_{\psi_{\gamma}} (\beta) = -\left[ \gamma + \sqrt{||\beta||^2 + \gamma^2} \right]^{1/2} \nabla \mathcal{L} (\beta) \ dt, \quad (3.3)
$$

where the potential writes $\psi_{\gamma} (\beta) := \frac{2}{\gamma} \left( \sqrt{||\beta||^2 + \gamma^2} + \gamma \right)^{3/2} - 2\gamma \left( \sqrt{||\beta||^2 + \gamma^2} \right)^{1/2}.

**Mirror on Singular Values.** For vector regression, GF on the parameters induces a continuous-time mirror descent (which we also refer as mirror flow) with the hyperbolic entropy function [Ghai et al., 2020]. This extends Arora et al. [2019b]’s characterization of the evolution of singular values when the initialization is balanced. Note that our result does not fully characterize the evolution of the system, since the characterization of the singular vectors is absent. Still, some interesting comments can be made. In the rich regime in which $\gamma \to 0$, the hyperbolic entropy $\Psi_{\gamma} \sim (-\ln \gamma) \|\beta\|_{\ast}$. Thus, informally, for small $\gamma$ the gradient flow on parameters is approximately equivalent to a mirror flow on the nuclear norm. This is reminiscent of the case of diagonal linear networks where such an equivalence is proven rigorously [Pesme and Flammarion, 2023] and is known to lead to an incremental saddle-to-saddle dynamics [Li et al., 2021, Jacot et al., 2022, Berthier, 2022].

In the case of whitened data, i.e., $X^T X = I$, we show that the singular vectors of $\beta$ are stationary (see Appendix C.11 for details). Therefore, the mirror flow on the singular values characterizes the entire system. In this case, we can even provide an exact expression for the evolution of the singular values.
Furthermore, the dynamics of gradient flow with balanced initializations. In contrast, with a specific initialization shape, we show the existence of a mirror flow is surprising since the reparametrization defining linear networks is not commutative in general [Li et al., 2022]. However, due to the specific initialization we use, this problem can be circumvented by preserving certain commutative properties.

The equivalence with mirror descent enables us to show that \( \mathcal{L}(\beta(t)) = \mathcal{O}(1/\gamma t) \) (see Appendix C.8), thus providing a convergence rate for the training loss independent of the conditioning of data, in contrast to Min et al. [2021], Du et al. [2019b]. Note that with decreasing initialization scale \( \gamma \), the convergence speed diminishes, while according to the results in Theorems 3.1 better implicit bias is achieved. This suggests the existence of a trade-off between optimization and implicit bias already observed in several works [Woodworth et al., 2020], where achieving better quality solutions is linked to slower optimization. In contrast to this behaviour, for the case of balanced initialization [Braun et al., 2022] emphasizes a decoupling between the learning speed and the quality of solutions. Conversely, we stress that in the general setting (e.g., under imbalance) such a decoupling is absent.

### 3.3 Sketch of the proofs

In this section, we give a short description of the proofs of the main results from the previous sections. The common theme of the following intermediate results is to identify natural invariants of the dynamics, which can be leveraged to understand the hidden mirror structure of the flow.

**Lemma 3.1.** Consider the dynamics of the gradient flow (2.4) initialized at \((W_1(0), W_2(0)) = (\sqrt{2}P, 0)\). Let \(Z_1 := W_1P^\top, Z_2 := PW_2\) and the residual \(R := X^\top(Y - XZ_1Z_2)\), then the evolution of \((Z_1, Z_2)\) is governed by the following ODE

\[
\dot{Z}_1 = RZ_2^\top, \quad \dot{Z}_2 = Z_1^\top R. \tag{3.4}
\]

Furthermore, the dynamics of gradient flow (2.4) is equivalent to (3.4), i.e., \((W_1(t), W_2(t)) = (Z_1(t)P, P^\top Z_2(t))\) at any time \(t\).

Lemma 3.1 derives an equivalent dynamics to equations (2.4). It shows that weights \(W_1, W_2\) always stay in the column span of the initialization \(P\), thus restricting their evolution to a subspace. Going forward, we derive the invariants of the dynamics (3.4).

**Lemma 3.2.** For the projected matrices given in (3.4), we have the following invariant,

\[Z_1^\top Z_1 - Z_2Z_2^\top = 2\gamma I.\]

This invariant ensures that \(Z_1^\top Z_1\) and \(Z_2Z_2^\top\) commute which is a crucial ingredient in the proofs of Theorems 3.1, 3.2. Now, we derive the evolution of \(\alpha := Z_1^\top Z_2\), which turns out to be the central quantity enabling our result. The lemma below describes certain properties of the evolution of \(\alpha\).

**Lemma 3.3.** Let \(\alpha := Z_1^\top Z_2\), we have the following time evolution of parameters:

\[
\dot{\alpha} = R - \alpha R^\top \alpha, \quad \text{and} \quad \beta = (1 - \alpha \alpha^\top)^{-1} \alpha.
\]
An outline of the proof of Theorem 3.1. With an ansatz on the potential that it is decomposable in terms of $Z_1, Z_2$, we derive KKT conditions for the constrained optimization problem

$$\arg\min_{\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2 = Y} \psi_1(Z_1) + \psi_2(Z_2).$$

Using Lemmas 3.3, we show that $\alpha$ stays in $\text{span}(X)$. We use the isotropic property of the imbalance from Lemma 3.2 to find appropriate functions $\psi_1, \psi_2$ and finally prove Theorem 3.1. The proofs for theorem 3.1, 3.2 and corollary 3.1 can be found in Appendix B.

4 Further thoughts and perspectives

The previous section provided a deep-dive into the dynamics of the gradient flow, which we complement here with a few steps in the direction of understanding the dynamics with stochastic gradients. We investigate stochastic gradient descent (SGD) by studying its simpler counterpart, label noise gradient descent (LNGD) [Blanc et al., 2020].

4.1 The role of noise

It was observed that the noise in stochastic gradient descent has a parameter-dependent shape that induces certain regularization properties [HaoChen et al., 2021, Blanc et al., 2020]. Here, we study the properties of the noise shape induced in the case of parameterization with linear neural networks.

Inspired from the analysis of HaoChen et al. [2021] and the large noise regime described by Pillaud-Vivien et al. [2022] in the context of diagonal neural networks, we design a process driven purely by noise and which carries the same geometric properties as SGD’s noise. We consider the scalar ($k = 1$) regression problem with $l = d$ and, through an abuse of notation, denote $W_1 = W$, and $W_2 = a$. The noise-driven process which we consider is:

$$dW = (dB_t) a^\top, \quad da = W^\top dB_t, \quad (4.1)$$

where $B_t$ is a $d$-dimensional Brownian motion. Details on how this SDE captures the noise of SGD are deferred to Appendix D. We show that, similarly to the rich regime of the gradient flow ($\gamma \to 0$), this noise also carries a rich spectral bias but for any initialization. Indeed, we have the following result on the SDE dynamics. The proof can be found in Appendix D.

**Proposition 4.1.** The dynamics (4.1) has the following convergence properties

(a) Variance explosion. The variance of the norms of $W$, $a$ explode, i.e.,

$$\lim_{t \to \infty} E \left[ \|W(t)\|^2 \right] \to \infty, \quad \lim_{t \to \infty} E \left[ \|a(t)\|^2 \right] \to \infty.$$

(b) Scale divergence. For $d \geq 5$, for any $\alpha > 0$, we have that,

$$\lim_{t \to \infty} E \left[ \|W(t)\|^\alpha \right] \to \infty, \quad \lim_{t \to \infty} E \left[ \|a(t)\|^\alpha + \|\bar{a}(t)\|^\alpha \right] \to \infty.$$

where $\bar{a} := e^{-t} \int_0^t e^s a(s) ds$ is the exponential moving average of $a$.

(c) Alignment - spectral bias. Denote the $i^{th}$ row of $W$ as $w_i$. Using $[w_i, a] \defeq w_i a^\top - aw_i^\top$, we have

$$\lim_{t \to \infty} E \left[ \|[w_i, a]\| \right] \to 0.$$

For any two vectors $u, v$, $[u, v]$ denotes the commutator of the vectors: remark that if $[u, v] = 0$, then $u, v$ are aligned, i.e., $u = cv$, for some scalar $c \in \mathbb{R}$. First, notice that for $d = 1$, the SDE in fact corresponds to the geometric Brownian motion with no drift and the dynamics collapses to zero [Oksendal, 2013]. For dimension $d \geq 2$, the proposition states that the system diverges and the weights grow towards infinity. However, despite the fact that the norm grows, the commutator $[w_i, a]$ goes to zero, indicating that all the rows of $w_i$ align towards $a$. Overall, similarly to the gradient flow in the rich regime, this induces a low rank structure in $W$. This phenomenon can be further seen through the evolution of singular values, where the top singular value of $W$ grows unboundedly, whereas the remaining singular values decay to 0 as depicted in Figure 3a in the Appendix. This sheds some light on how SGD induces a particular parameter-dependent noise which implicitly biases the solutions towards having a low-rank structure of the hidden layer [Andriushchenko et al., 2022].
Intricate dynamics in presence of drift. Proposition 4.1 focuses on the process that is solely driven by noise. However, in general SGD also encompasses a drift term which corresponds to the dynamics studied in Section 3. The continuous-time SDE describing the process is

\[ dW = -\nabla_W \mathcal{L}(W, \mathbf{a}) \, dt + \delta (dB_t) \mathbf{a}^\top, \quad da = -\nabla_a \mathcal{L}(W, \mathbf{a}) \, dt + \delta W^\top dB_t, \]

where \( \delta > 0 \) indicates the scale of the noise. The presence of drift quickly complicates the analysis, but intuitively, the noise simplifies the model by inducing a rank reduction, whereas the drift terms prevent the weights from growing unbounded. This noise-driven mechanism relaxes the role of initialization. Empirically this is illustrated in Figure 1b. Gradient descent already exhibits a regularization effect as it increases only one singular value while keeping the others constant. However, gradient descent with label noise [Blanc et al., 2020] enhances this regularization effect by decaying the singular values and promoting low-rank representations. As a result, even for larger initialization scales, we observe the presence of low-rank structures in the hidden layer, unlike in gradient descent. The precise characterization of this phenomenon is left for future research.

Experiments. We consider a regression problem on synthetic data, with \( n = 5 \) samples of Gaussian data in \( \mathbb{R}^{10} \) \((d = 10)\) and the labels in \( \mathbb{R}^5 \) \((k = 3)\) generated by a ground truth \( \mathbf{\beta}_* \in \mathbb{R}^{d \times k} \). We consider a network with width \( l = 200 \). In Figure 1a, we show the evolution of the top-4 singular values of the hidden layer \( W_1 \). We use orthogonal initialization for the network with the two scales of initialization \( \gamma = 1, 10^{-4} \). Note that, as depicted by Corollary 3.1, for the smaller scale only the first \( k = 3 \) singular values are significant in comparison to the remaining \( d - k \) singular values. This shows that the matrix is approximately rank \( k \) and the neurons align along three directions. In contrast, for the larger scale \( \gamma = 1 \), the final weight matrix has rank \( d \). To complement this, we also consider a Gaussian initialization with variance 0.01 – specifically, we initialize the inner layer with \( d = 10 \) Gaussian random vectors in \( \mathbb{R}^d \). As described when \( l \gg d \), the initialization is close to the orthogonal initialization. Hence, in this case, we can see that only \( k \) singular values grow and the final model has an approximately rank \( k \) hidden layer. In figure 1b, we depict the time evolution of singular values for GD and LNGD on a scalar regression problem with orthogonal data in \( \mathbb{R}^5 \) \((n, d = 5)\) and a network with \( l = 200 \). Further details on hyper-parameters can be found in the Appendix.

Extension to non-linear activations. Huh et al. [2023], Andriushchenko et al. [2022] empirically demonstrate a low-rank phenomenon through extensive experiments on deep networks with non-linear activations. However, a comprehensive theoretical comprehension of this behavior remains elusive, despite some efforts addressing these issues [Boursier et al., 2022]. To show that our analysis extends beyond linear activations, we present a toy experiment for ReLU networks (see Figure 2 and further details in Figure 5). Consider a scalar regression problem in a ReLU teacher-student setup. We generate a training set of size 10 sampled from a random Gaussian distribution in \( \mathbb{R}^5 \). The training data \((x_i, y_i)_{i=1}^{10} \in \mathbb{R}^5 \times \mathbb{R} \) is generated by a teacher ReLU network with 2 neurons \((w_0, w_1)\), i.e.,

\[ y_i = a_0 \sigma(w_0^\top x_0) + a_1 \sigma(w_1^\top x_1), \]
where $\sigma$ is the ReLU non-linearity. We train a student network with 20 hidden neurons. Note that there are two relevant directions $w_0, w_1$ for the student network to learn, therefore we expect the hidden layer to represent these two directions (i.e., a rank 2 hidden layer, and a singular value decomposition with two non-zero singular values). This property is empirically verified in Figure 2. We plot the time evolution of singular values and when initialized at low-scale the network converges to an approximately rank-2 matrix. When initialized at a larger scale, the network weight matrix is high rank and the neurons do not learn the teacher directions.

**Perspectives.** Learning representations which can be transferred to downstream tasks is a key attribute for the success of deep learning [Bengio et al., 2013, LeCun et al., 2015]. In this work, we present an archetypal problem where for the same predictor in functional space, there exist multiple representations in parameter space, some of which can exhibit a rich structure. This scenario presents a case for going beyond the characterization of implicit bias in the functional space [Parhi and Nowak, 2022] and further studying the implicit bias in the parameter space. Such characterizations facilitate the identification of crucial ingredients in training algorithms that enable effective feature learning.

**Limitations and Future Work.** This paper tackles the phenomenon of implicit bias, with the aim of furthering the understanding of how neural networks learn in practice. Unfortunately, practical models are highly nonlinear due to their activations and rely on various heuristics to achieve state-of-the-art performance, thus being difficult to grasp mathematically. This work therefore studies the simplified setting of two-layer linear neural networks. In terms of the assumptions we make, the orthogonality of initialization is only approximately faithful to practical settings where small random weights are used. Nevertheless, we are confident that this requirement can be loosened through a perturbation analysis in the vein of Gidel et al. [2019]. Finally, our dynamical description of the system is yet to be completed in the vector regression case with non-whitened data. A careful set of assumptions is necessary here, and hopefully ones that are weaker than the restricted isometry property used in related works [Li et al., 2018]. Finally, we only partially describe the dynamics in the presence of stochastic noise and giving a full characterization remains a desired objective of future investigations. Further discussion on these aspects is presented in Appendix C.1.

**Acknowledgments and Disclosure of Funding**

AV is supported by Swiss data science fellowship.

**References**


Organization

The appendix is organized as follows,

- In section A, we present the experiment details.
- In section B, we present the proof of Theorems 3.1, 3.2 and Corollary 3.1.
- section C contains the proofs of the supporting lemmas.
- In the final section D, we discuss our choice on the noise model and present the proof of Proposition 4.1.
- For the results referenced in the main section, the convergence rate of mirror flow can be found at C.8, the time evolution of singular values and their limiting jump process is available at C.12, the stationarity of singular vectors in the orthogonal case at C.11, the discussion on the noise model at D.

Notations. For matrices of appropriate dimensions, we use \([A, B]\) to denote \(AB - BA\).

A Experiment Details

Experiments. We discretize the SDE (4.1) with a step-size \(\sim \frac{1}{\sqrt{t}}\). We simulate three parallel runs and track the evolution of singular values and the evolution of alignment using the commutator of the row \(w_i, a\), i.e., \([w_i, a] := (w_i a^\top - aw_i^\top)\). We consider the SDE for dimension \(d = 2\). The evolution is initialized at \(W(0) = I_2\) and \(a(0) = 0\). As seen in Figure 3a, the noise shape inherently induces a low-rank structure where it intensifies a singular value and significantly diminishes the other singular value. As predicted by our proposition (4.1), figure 3b shows that the rows of \(W\) align with \(a\), thus giving a rank 1 structure. The experiments were run on a 16-GB RAM Apple M1 mac with OS Ventura 13.3.1.

![Figure 3: Three parallel runs of the noise dynamics Eq. (4.1) for \(d = 2\). (a) The evolution of singular values with \(\sigma_0\) increasing and \(\sigma_1\) decaying. (b) Measuring the norm of the commutator again as predicted by Proposition 4.1.](image)

B Main Proofs

In this section, we present the proofs of the theorem discussed in Section 3.

Theorem 3.1. Let \((W_1, W_2)\) be the process that follows the GF equations (2.4a)-(2.4b), initialized according to condition \(I_\gamma\), for some \(\gamma > 0\). Then

(i) The parameters converge to a global optimum of the loss

\[ \lim_{t \to \infty} (W_1(t), W_2(t)) = (W_1^\infty, W_2^\infty) \in I_\theta. \]
(ii) The linear predictor $\beta$ converges to the minimum $\ell_2$-norm interpolator

$$\lim_{t \to \infty} \beta(t) = \arg \min_{X: Y = X} \frac{1}{2} \|\beta\|_F^2 \overset{\text{def}}{=} \beta_*.$$

(iii) We have the following variational characterization of the limiting parameters

$$(W_1^\infty, W_2^\infty) \in \arg \min_{X: W_1^\infty, W_2^\infty = Y} \frac{1}{2} \|W_2\|_F^2 + \frac{1}{2} \|W_1\|_F^2 - \gamma \log \left( \det (W_1 W_1^\top) \right). \quad (3.1)$$

Proof. We initialize such that $W_1 = \sqrt{2\gamma} P, W_2 = 0$. Lemma 3.1 states that the dynamics of gradient flow is restricted to a subspace and can be equivalently described by,

$$\dot{Z}_1 = RZ_2^\top, \quad \dot{Z}_2 = Z_1^\top R. \quad (B.1)$$

where $R := X^\top (Y - XZ_1 Z_2)$.

(i) To show the convergence, we track the evolution of $\text{tr} \left( R^\top R \right)$ and use the following descent inequality to show that it converges to 0. With $\lambda_{\min}(X^\top X)$ denoting the smallest eigenvalue of the $X^\top X$, the descent inequality (C.13) is as follows,

$$\text{tr} \left( R^\top R \right) \leq -2\gamma \lambda_{\min}(X^\top X) \text{tr} \left( R^\top R \right).$$

Refer to Lemma C.6 for the detailed proof.

(ii) To show that $\beta \to \beta_*$ in the limit, we show that $\beta \in \text{span} \left( X^\top \right)$, i.e., $\beta = X^\top \lambda$, for some $\lambda$. This satisfies the KKT conditions required for the following minimization problem.

$$\beta_* \in \arg \min_{X: Y = X} \frac{1}{2} \|\beta\|_F^2 = \beta_*.$$

The complete proof can be found at Lemma C.6.

(iii) For the limit of the projected dynamics $(Z_1^\infty, Z_2^\infty) := \lim_{t \to \infty} (Z_1(t), Z_2(t))$, Lemma C.4 shows the following,

$$(Z_1^\infty, Z_2^\infty) \in \arg \min_{X: Z_1^\infty, Z_2^\infty = y} \frac{1}{2} \|Z_2\|_F^2 + \frac{1}{2} \|Z_1\|_F^2 - \gamma \log \left( \det (Z_1 Z_1^\top) \right).$$

Using the transformation from Eq. C.7, we have,

$$W_1 = Z_1 P, \quad W_2 = P^\top Z_2.$$

Thus,

$$\|W_1\|_F = \|Z_1\|_F, \quad \|W_2\|_F = \|Z_2\|_F, \quad Z_1 Z_1^\top = W_1 W_1^\top, \quad Z_1 Z_2 = W_1 W_2.$$

Therefore,

$$(W_1^\infty, W_2^\infty) \in \arg \min_{X: W_1^\infty, W_2^\infty = Y} \frac{1}{2} \|W_2\|_F^2 + \frac{1}{2} \|W_1\|_F^2 - \gamma \log \left( \det (W_1 W_1^\top) \right).$$

This hold on the set $\{(W_1, W_2) : W_1 P_+ = 0, P_+ W_2 = 0\}$ which is ensured from gradient flow from Lemma 3.1.

\[\Box\]

**Corollary 3.1.** [Singular values at the limit] Using the same quantities as in Theorem 3.1, denote $(\sigma_1(W_1^\infty), \ldots, \sigma_d(W_1^\infty))$ and $(\sigma_1(\beta_*), \ldots, \sigma_k(\beta_*))$ the singular values of $W_1^\infty, \beta_*$ are

$$\sigma_i(W_1^\infty) = \left( \sqrt{\sigma_i(\beta_*)^2 + \gamma^2 + \gamma} \right)^{1/2}, \text{ for } 1 \leq i \leq k.$$  

$$\sigma_i(W_1^\infty) = (2\gamma)^{1/2}, \text{ for } k < i \leq d.$$
Proof. Using the transformation $W_1 = Z_1 P$ from Lemma 3.1, we obtain,

$$W_1 W_1^\top = Z_1 Z_1^\top.$$  \hfill (B.2)

Thus, $W_1$ and $Z_1$ share the same singular values, similarly $W_2$ and $Z_2$ share the same singular values. The expressions for the singular values of $Z_1$ (similarly $Z_2$) can be found at Lemma C.10. This along with the fact that $\beta \to \beta_*$ proves the Corollary 3.1

\qed

**Theorem 3.2.** [Dynamics of the flow] With the same notations as in Theorem 3.1,

(a) **Mirror on singular values:** The singular values of $\beta$, denoted by $D_\beta$, follow the mirror flow

$$d\nabla \psi_\gamma(D_\beta) = -\nabla D_\beta \nabla L \, dt,$$

where the potential writes $\psi_\gamma(D_\beta) := \text{tr}\left(\frac{1}{2}D_\beta \sinh^{-1}(D_\beta/\gamma) - \sqrt{D_\beta^2 + \gamma^2}\right)$.  

(b) **Mirror on $\beta$:** If $k = 1$, the dynamics of $\beta$ can be characterized as a mirror flow

$$d\nabla \psi_\gamma(\beta) = -\left[\gamma + \sqrt{\|\beta\|^2 + \gamma^2}\right]^{1/2} \nabla L(\beta) \, dt,$$

where the potential writes $\psi_\gamma(\beta) := \frac{3}{2} \left[\sqrt{\|\beta\|^2 + \gamma^2} + \gamma\right]^{3/2} - 2\gamma \left[\sqrt{\|\beta\|^2 + \gamma^2 + \gamma}\right]^{1/2}$.

Proof. The equivalence with mirror flow for scalar regression is shown in Lemma C.7. The continuous time mirror descent for singular values of $\beta$ is derived in Lemma C.13. \hfill \square

\section{Supporting Lemmas}

This section contains all the technical lemmas and definitions used in the proofs in the section before.

**Lemma 3.1.** Consider the dynamics of the gradient flow (2.4) initialized at $(W_1(0), W_2(0)) = (\sqrt{2} P, 0)$. Let $Z_1 := W_1 P^\top$, $Z_2 := P W_2$ and the residual $R := X^\top (Y - X Z_1 Z_2)$, then the evolution of $(Z_1, Z_2)$ is governed by the following ODE

$$\dot{Z}_1 = R Z_2^\top, \quad \dot{Z}_2 = Z_1^\top R. \tag{C.1}$$

Furthermore, the dynamics of gradient flow (2.4) is equivalent to (3.4), i.e., $(W_1(t), W_2(t)) = (Z_1(t) P, P^\top Z_2(t))$ at any time $t$.

Proof. We choose $P_\perp \in \mathbb{R}^{(l-d)\times l}$ such that $P_\perp P_\perp = 0$. Using the fact that $P, P_\perp$ orthogonal and span the entire $\mathbb{R}^{l\times l}$,

$$P^\top P + P_\perp^\top P_\perp = I_l.$$

Denoting $(Z_1)_\perp := W_1 P_\perp^\top, (Z_2)_\perp := P_\perp W_2$, we have,

$$W_1 = W_1 \left[P^\top P + P_\perp^\top P_\perp\right] = W_1 P^\top P + W_1 P_\perp^\top P_\perp = Z_1 P + (Z_1)_\perp P_\perp, \tag{C.2}$$

$$W_2 = \left[P^\top P + P_\perp^\top P_\perp\right] W_2 = P^\top P W_2 + P_\perp^\top P_\perp W_2 = P^\top Z_2 + P_\perp^\top (Z_2)_\perp. \tag{C.3}$$

Recalling the evolution of gradient flow (2.4) on the loss

$$\dot{W}_1 = X^\top (Y - X W_1 W_2) W_2^\top,$$

$$\dot{W}_2 = W_1^\top X^\top (Y - X W_1 W_2).$$

Multiplying the gradient flow updates with $P^\top, P$ from the right, left (resp.) for the above equations

$$\dot{W}_1 P^\top = X^\top (Y - X W_1 W_2) W_2^\top P^\top, \quad P W_2 = P W_1^\top X^\top (Y - X W_1 W_2).$$

Similarly multiplying with $P_\perp^\top, P_\perp$, we have,

$$\dot{W}_1 P_\perp^\top = X^\top (Y - X W_1 W_2) W_2^\top P_\perp^\top, \quad P_\perp \dot{W}_2 = P_\perp^\top W_1^\top X^\top (Y - X W_1 W_2).$$

16
Using the above, we have,
\[
W_1 W_2 = W_1 \left[ P^T P + P_\perp^T P_\perp \right] W_2,
\]
\[
= W_1 P^T P W_2 + W_1 P_\perp^T P_\perp W_2.
\]
\[
W_1 W_2 = Z_1 Z_2 + (Z_1)_\perp (Z_2)_\perp. \tag{C.5}
\]
Rewriting the evolution in terms of \( Z_1, Z_2, (Z_1)_\perp, (Z_2)_\perp \),
\[
\dot{Z}_1 = X^T (Y - X [Z_1 Z_2 + (Z_1)_\perp (Z_2)_\perp]) Z_2^T, \quad \dot{Z}_2 = Z_1^T X^T (Y - X [Z_1 Z_2 + (Z_1)_\perp (Z_2)_\perp]).
\]
\[(Z_1)_\perp = X^T (Y - X [Z_1 Z_2 + (Z_1)_\perp (Z_2)_\perp]) (Z_2)_\perp, \quad (Z_2)_\perp = (Z_1)_\perp X^T (Y - X [Z_1 Z_2 + (Z_1)_\perp (Z_2)_\perp]). \tag{C.6}
\]
This is exactly equivalent to the gradient flow under the linear transformation which maps \( W_1, W_2 \) to \((Z_1, (Z_1)_\perp), (Z_2, (Z_2)_\perp)\). Now taking into consideration the evolution of \((Z_1)_\perp, (Z_2)_\perp\), we have that \((0, 0)\) is an equilibrium point for the dynamics. From our initialization \( I_\gamma \),
\[
(Z_1)_\perp \bigg|_{t=0} = W_1 \bigg|_{t=0} = P_\perp = \sqrt{2\gamma} P P_\perp = 0,
\]
\[
(Z_2)_\perp \bigg|_{t=0} = P_\perp W_2 \bigg|_{t=0} = 0.
\]
As we initialized at the equilibrium of the dynamics we have \(((Z_1)_\perp, (Z_2)_\perp) = (0, 0)\) for any time \( t \). From Eq. (C.5), we have,
\[
W_1 W_2 = Z_1 Z_2. \tag{C.7}
\]
This finishes the proof of the lemma.

\textbf{Lemma C.1.} \textit{For the projected matrices given in (C.1), we have the following invariant,}
\[
Z_1^T Z_1 - Z_2 Z_2^T = 2\gamma I. \tag{C.8}
\]

\textit{Proof.} Recalling the dynamics (C.1)
\[
\dot{Z}_1 = R Z_2^T, \quad \dot{Z}_2 = Z_1^T R.
\]
\[
\dot{Z}_1^T Z_1 = (\dot{Z}_1)^T Z_1 + Z_1^T (\dot{Z}_1), = Z_2 R^T Z_1 + Z_1^T R Z_2^T.
\]
Similarly,
\[
\dot{Z}_2 Z_2^T = Z_2 R^T Z_1 + Z_1^T R Z_2^T.
\]
Hence, \( Z_1^T Z_1 - Z_2 Z_2^T = 0 \). This implies,
\[
Z_1^T Z_1 - Z_2 Z_2^T \bigg|_{t=0} = Z_1^T Z_1 - Z_2 Z_2^T = 2\gamma P P^T = 2\gamma I_d.
\]

\textbf{Lemma C.2.} \textit{Let} \( \alpha := Z_1^T Z_2 \), \textit{we have the following time evolution of parameters:}
\[
\dot{\alpha} = R - \alpha R^T \alpha, \quad \text{and} \quad \dot{\beta} = 2\gamma (I - \alpha \alpha^T)^{-1} \alpha.
\]

17
Proof. Taking the time derivative of $\alpha$,

$$\dot{\alpha} = Z_i^{-1}Z_2 + Z_i^{-1}\dot{Z}_2 = -Z_i^{-1}\dot{Z}_1 Z_i^{-1} Z_2 + Z_i^{-1} Z_1 R,$$

$$= -\alpha R^\top \alpha + R.$$

The evolution of $Z_1 Z_1^\top$, $\alpha \alpha^\top$,

$$\dot{(Z_1 Z_1^\top)} = Z_1 Z_1^\top \alpha R^\top + R \alpha^\top Z_1 Z_1^\top, \quad \text{(C.9)}$$

$$\dot{\alpha \alpha^\top} = (R - \alpha R^\top \alpha) \alpha^\top + \alpha (R - \alpha R^\top \alpha)^\top,$$

$$= R \alpha^\top - \alpha R^\top \alpha \alpha^\top + \alpha R^\top - \alpha \alpha^\top R \alpha^\top,$$

$$= (I - \alpha \alpha^\top) R \alpha^\top + \alpha R^\top (I - \alpha \alpha^\top).$$

Computing the evolution of $(I - \alpha \alpha^\top)^{-1}$,

$$\dot{(I - \alpha \alpha^\top)^{-1}} = (I - \alpha \alpha^\top)^{-1} \dot{\alpha \alpha^\top} (I - \alpha \alpha^\top)^{-1}, \quad \text{(C.10)}$$

$$= R \alpha^\top (I - \alpha \alpha^\top)^{-1} + (I - \alpha \alpha^\top)^{-1} \alpha R^\top. \quad \text{(C.11)}$$

Let $C_\alpha := 2\gamma (I - \alpha \alpha^\top)^{-1}$ and $C_Z := Z_1 Z_1^\top$, so we have,

$$\dot{C_\alpha} = C_\alpha \alpha R^\top + R \alpha^\top C_\alpha,$$

$$\dot{C_Z} = C_Z \alpha R^\top + R \alpha^\top C_Z.$$

Since, at initialization, $C_\alpha(0) = C_Z(0)$, we have $C_\alpha = C_Z$, for any time $t$. Therefore, we have,

$$Z_1 Z_1^\top = 2\gamma (I - \alpha \alpha^\top)^{-1}. \quad \text{(C.12)}$$

Using $\beta = Z_1 Z_2 = Z_1 Z_1^\top \alpha$ and the above invariant, we obtain $\beta = 2\gamma (I - \alpha \alpha^\top)^{-1} \alpha$. $\square$

Lemma C.3. The following property holds for $\alpha$:

$$\alpha_\infty = \lim_{t \to \infty} \alpha(t) \in \text{span}(X^\top).$$

Proof. From the evolution of $\alpha$, we have

$$\dot{\alpha} = R - \alpha R^\top \alpha.$$

Let $U \in \mathbb{R}^{d \times d}$ be the matrix projection each column of $\alpha$ on the column span of $X^\top$, i.e., span\{\(x_1, \ldots, x_n\)$, $U_\perp \in \mathbb{R}^{d \times d}$ be the matrix projection on the orthogonal space, i.e., Ker $(X^\top)$.

So $\alpha = U \alpha + U_\perp \alpha$. Note that $R = X^\top (Y - X Z_1 Z_1^\top \alpha)$. So $U_\perp R = 0$, since $U_\perp X^\top = 0$. The evolution of $U_\perp \alpha$ is that

$$\dot{U_\perp \alpha} = -U_\perp \alpha R^\top (U \alpha + U_\perp \alpha).$$

Again, $U_\perp \alpha = 0$ is the equilibrium point and our initialization $\alpha = 0$ ensures that it stays at this equilibrium. This proves the lemma. $\square$

Lemma C.4. Let $(Z_1^\infty, Z_2^\infty) \overset{\text{def}}{=} \lim_{t \to \infty} (Z_1(t), Z_2(t))$ the limit of the gradient flow dynamics. Then,

$$(Z_1^\infty, Z_2^\infty) \in \arg\min_{X Z_1 Z_2 = Y} \frac{1}{2} \|Z_2\|^2 + \frac{1}{2} \|Z_1\|_F^2 - \gamma \log \left(\det (Z_1 Z_1^\top)\right).$$
Proof. From Lemma C.3, we have that the \((Z_1^\infty)^{-T} Z_2^\infty \in \text{span}(X)\), so the condition (P2) from Proposition C.5 holds. Note that from (P2) of Proposition C.5, we have,

\[
\nabla \Psi_1(Z_1^\infty) = (Z_1^\infty)^{-T} Z_2^\infty (Z_2^\infty)^{-T},
\]

which is satisfied by the potential

\[
\Psi_1(Z_1) = \frac{1}{2} \|Z_1\|_F^2 - \gamma \log (\det (Z_1 Z_1^T)).
\]

When the imbalance is not isotropic, i.e., \(Z\), which is satisfied by the potential \(As\), \((Z_1^\infty)^{-T} Z_1 - Z_2 Z_2^T = D\), where \(D\) is some diagonal matrix \((\neq c I, \text{for any constant } c)\). If this case, \(\nabla \Psi_1(Z_1^\infty) = Z_1^\infty - D (Z_1^\infty)^{-T}\), and there exists no such function \(\Psi_1\) and the proof breaks. \(\square\)

Proposition C.5. Let \((Z_1^*, Z_2^*)\) satisfy the following minimization problem

\[
(Z_1^*, Z_2^*) = \arg \min_{XZ_1, Z_2 = y} \Psi_1(Z_1) + \Psi_2(Z_2),
\]

for some non-negative potential functions \(\Psi_1(Z_1), \Psi_2(Z_2)\). Then, \((Z_1^*, Z_2^*)\) satisfies

\[
\begin{align*}
(P1) & \ (Z_1^*)^{-T} \nabla \Psi_2(Z_2^*) \in \text{span}(X), \\
(P2) & \ \nabla \Psi_1(Z_1^*) = (Z_1^*)^{-T} \nabla \Psi_2(Z_2^*)(Z_2^*)^T.
\end{align*}
\]

Proof. The Lagrangian for the minimization problem above is,

\[
\mathcal{L}(Z_1, Z_2, \lambda) = \Psi_1(Z_1) + \Psi_2(Z_2) + \langle \lambda, XZ_1, Z_2 - y \rangle.
\]

Taking derivatives w.r.t. to \(Z_1, Z_2\), we get,

\[
\begin{align*}
\nabla_{Z_1} \mathcal{L}(Z_1, Z_2, \lambda) = \nabla \Psi_1(Z_1) + (X^T \lambda) Z_2^T, \\
\nabla_{Z_2} \mathcal{L}(Z_1, Z_2, \lambda) = \nabla \Psi_2(Z_2) + Z_1^T (X^T \lambda).
\end{align*}
\]

As \(Z_1^*, Z_2^*\) should satisfy \(\nabla \mathcal{L} = 0\),

\[
\begin{align*}
(Z_1^*)^{-T} \nabla \Psi_2(Z_2^*) &= -X^T \lambda, \\
\nabla \Psi_1(Z_1^*) &= (Z_1^*)^{-T} \nabla \Psi_2(Z_2^*)(Z_2^*)^T
\end{align*}
\]

\(\square\)

Lemma C.6. Let \((Z_1, Z_2)\) be the process that follows the GF equations (3.1), initialized according to condition \(I_\gamma\), for some \(\gamma > 0\). Then

\(\text{(i)}\) The parameters converge to a global optimum of the loss

\[
\lim_{t \to \infty} Y - XZ_1(t)Z_2(t) = 0.
\]

\(\text{(ii)}\) The linear predictor \(\beta\) converges to the minimum \(\ell_2\)-norm interpolator

\[
\lim_{t \to \infty} \beta(t) = \arg \min_{X \beta = Y} \|\beta\|_2 \overset{\text{def}}{=} \beta_*.
\]

Proof. The evolution of \(R\) writes,

\[
\dot{R} = -(X^T X) \left[ Z_1 Z_1^T R + RZ_2^T Z_2 \right],
\]

\[
\text{tr} (R^T R) = -2\text{tr} \left( R^T (X^T X) Z_1 Z_1^T R \right) - 2\text{tr} \left( R^T (X^T X) RZ_2^T Z_2 \right).
\]
Therefore \( \beta \)

Thus, using Gronwall we can show that

Proof. Here we consider the case \( k = 1 \), let \( r := X^\top(Y - X\beta) \). It is denoted lowercase since \( r \) here is a vector. The gradient flow C.1 now in \( Z_1, \alpha \) can be now can be written as,

\[
\dot{\alpha} = (I - \alpha \alpha^\top) r , \quad \dot{Z}_1 = Z_1 \alpha r^\top
\]

\[
\dot{\beta} = (Z_1^\top Z_1 \alpha) = (Z_1^\top Z_1 + \alpha^\top Z_1^\top Z_1 \alpha) r , \quad \dot{r} = - (X^\top X) (Z_1^\top Z_1 + \alpha^\top Z_1^\top Z_1 \alpha) r
\]
Recall from Lemma 3.3 that
\[ \beta = 2\gamma(I - \alpha\alpha^T)^{-1}\alpha. \]

Using Sherman-Morrison,
\[
(I - \alpha\alpha^T)^{-1} = I + \frac{\alpha\alpha^T}{1 - \|\alpha\|^2}.
\]

\[ \beta = 2\gamma [(I - \alpha\alpha^T)^{-1}] \alpha = 2\gamma \left[ I + \frac{\alpha\alpha^T}{1 - \|\alpha\|^2} \right] \alpha = 2\gamma \frac{\alpha}{1 - \|\alpha\|^2}. \]

Taking the norms, we get,
\[ \beta = 2\gamma \alpha \]
\[ \frac{1}{1 - \|\alpha\|^2} \]}
\[ = 2\gamma \frac{\alpha}{\sqrt{\gamma^2 + \|\beta\|^2 + \gamma}}. \]

Now consider the evolution of \( \alpha \)
\[ \dot{\alpha} = [I - \alpha\alpha^T] r, \]
\[ [I - \alpha\alpha^T]^{-1} \dot{\alpha} = r, \]
\[ \left[ I + \frac{\alpha\alpha^T}{1 - \|\alpha\|^2} \right] \dot{\alpha} = r. \]

One cannot write \[ \left[ I + \frac{\alpha\alpha^T}{1 - \|\alpha\|^2} \right] \] as a Hessian of any function of \( \alpha \). However, multiplying on both sides with \( (1 - \|\alpha\|^2)^{-1/2} \), we obtain,
\[
\frac{\alpha}{(1 - \|\alpha\|^2)^{1/2}} = \frac{\beta}{2\gamma} (1 - \|\alpha\|^2)^{1/2} = \frac{\beta}{2\gamma} \left( \frac{2\gamma}{\sqrt{\gamma^2 + \|\beta\|^2 + \gamma}} \right)^{1/2},
\]
\[ = \frac{\beta}{\sqrt{2\gamma \left( \sqrt{\gamma^2 + \|\beta\|^2 + \gamma} \right)^{1/2}}}, \]
Substituting the above expression and also \((1 - \|\alpha\|^2)^{1/2}\) on the RHS gives us,

\[
\frac{d}{dt} \left[ \frac{\beta}{\sqrt{\gamma^2 + \|\beta\|^2 + \gamma}} \right] = \left( \sqrt{\gamma^2 + \|\beta\|^2 + \gamma} \right)^{1/2} r.
\]

Define the mirror potential,

\[
\psi_\gamma (\beta) \overset{\text{def}}{=} \frac{2}{3} \left[ \sqrt{\|\beta\|^2 + \gamma^2 + \gamma} \right]^{3/2} - 2\gamma \left[ \sqrt{\|\beta\|^2 + \gamma^2 + \gamma} \right]^{1/2}.
\]  

(C.15)

Using the fact that,

\[
\frac{d}{dx} \left[ \frac{2}{3} \left( \sqrt{x^2 + p^2 + q} \right)^{3/2} - 2q \left[ \sqrt{x^2 + p^2 + q} \right]^{1/2} \right] = \left[ \sqrt{x^2 + p^2 + q} \right]^{-1/2} x.
\]

We have,

\[
\nabla \psi_\gamma (\beta) = \left( \gamma + \sqrt{\|\beta\|^2 + \gamma^2} \right)^{-1/2} \beta.
\]

Thus, we can write it as a continuous-time mirror descent.

\[
\frac{d}{dt} \nabla \psi_\gamma (\beta) = - \left( \gamma + \sqrt{\|\beta\|^2 + \gamma^2} \right)^{1/2} \nabla \mathcal{L} (\beta).
\]

Lemma C.8 (Convergence rate of mirror flow). For the dynamics of the mirror flow given by Eq. (C.14), we have the following rate of convergence,

\[
\mathcal{L} (\beta (t)) \leq \frac{D_{\psi_\gamma} (\beta_*, \beta_0)}{\sqrt{2\gamma} t}.
\]

Proof. Note that \(\psi_\gamma\) is convex. Let \(D_{\psi_\gamma} (\cdot, \cdot)\) be the Bregman divergence defined with the potential \(\psi_\gamma (\cdot)\). Taking the time derivative of the Bregman divergence, we get,

\[
\frac{d}{dt} D_{\psi_\gamma} (\beta_*, \beta) = \left\langle \frac{\nabla \psi_\gamma (\beta_*)}{\beta - \beta_*} \right\rangle,
\]

\[
= - \left( \gamma + \sqrt{\|\beta\|^2 + \gamma^2} \right)^{1/2} \left\langle \nabla \mathcal{L} (\beta), \beta - \beta_* \right\rangle.
\]

We have,

\[
- \left( \gamma + \sqrt{\|\beta\|^2 + \gamma^2} \right)^{1/2} \leq - \sqrt{2\gamma}
\]

Using the convexity of \(\mathcal{L} (\cdot)\) in \(\beta\), we have

\[
\left\langle \nabla \mathcal{L} (\beta), \beta - \beta_* \right\rangle \leq \mathcal{L} (\beta) - \mathcal{L} (\beta_*) = \mathcal{L} (\beta).
\]

The last step is using the existence of an interpolating solution. Substituting the above two inequalities, we get,

\[
\frac{d}{dt} D_{\psi_\gamma} (\beta_*, \beta) \leq - \sqrt{2\gamma} \mathcal{L} (\beta).
\]

Integrating, we get,

\[
\int_0^t \mathcal{L} (\beta (s)) \, ds \leq \frac{1}{\sqrt{2\gamma}} \left[ D_{\psi_\gamma} (\beta_*, \beta_0) - D_{\psi_\gamma} (\beta_*, \beta_t) \right] \leq \frac{D_{\psi_\gamma} (\beta_*, \beta_0)}{\sqrt{2\gamma}}.
\]
The last step is using the fact that Bregman divergence is positive for convex functions. We will now show that the loss is decreasing along the trajectory,

\[ \mathcal{L}(\beta) = \left\langle \nabla \mathcal{L}(\beta), \dot{\beta} \right\rangle = -\frac{1}{\left(\sqrt{\gamma^2 + \|\beta\|^2} + \gamma\right)^{1/2}} \left\langle \nabla \psi_\gamma(\beta), \dot{\beta} \right\rangle. \]

\[ \nabla \psi_\gamma(\beta) = \nabla^2 \psi_\gamma(\beta) \dot{\beta}, \quad \left\langle \nabla \psi_\gamma(\beta), \dot{\beta} \right\rangle = \left\langle \nabla^2 \psi_\gamma(\beta) \dot{\beta}, \dot{\beta} \right\rangle \geq 0, \quad \text{(Using convexity).} \]

Thus \( \mathcal{L}(\beta) \leq 0 \), using this,

\[ t \mathcal{L}(\beta(t)) \leq \int_0^t \mathcal{L}(\beta(s)) \, ds \leq \frac{D_{\psi_\gamma}(\beta_\alpha, \beta_0)}{\sqrt{2\gamma}}. \]

This completes the proof. \( \square \)

**Definition C.9.** [Singular value decomposition and notation.] The singular value decomposition of a given matrix \( A \in \mathbb{R}^{d \times k} \) be denoted as \( U_A D_A V_A^\top \) where \( U_A \in \mathbb{R}^{d \times d}, D_A \in \mathbb{R}^{d \times k}, V_A \in \mathbb{R}^{k \times k}. \)

Let \( k < d \) and consider the products \( AA^\top \) and \( A^\top A \). We will refer to the diagonal matrices \( D_{A^\top A} \in \mathbb{R}^{k \times k} \) and \( D_{AA^\top} \in \mathbb{R}^{d \times d} \) as \( D_A^2 \) and \( \tilde{D}_A^2 \), in order to ease notation. Note that \( D_A^2 \) is a block-diagonal matrix containing \( D_A^2 \) as the first diagonal block and \( 0 \) as the second diagonal block.

Finally, if \( k = d \) then \( D_A^2 = \tilde{D}_A^2 \).

**Lemma C.10 (Singular values of \( Z_1 \) and \( \beta \)).** Using the notations in Definition C.9, the singular values of \( Z_1, Z_2 \) and \( \beta \) are related as the following,

\[ D_{Z_2}^2 = -\gamma I_k + \sqrt{\gamma^2 I_k + D_{\beta}^2} \quad \text{and} \quad D_{Z_1}^2 = \gamma I_d + \sqrt{\gamma^2 I_d + D_{\beta}^2}. \]

**Proof.** From the invariance (C.8) we further deduce (by appropriately multiplying left and right with \( Z_1, Z_2 \) and their transposes):

\[ \beta^\top \beta - (Z_2^\top Z_2)^2 = 2\gamma Z_2^\top Z_2 \quad \text{and} \quad (Z_1^\top Z_1)^2 = 2\gamma Z_1^\top Z_1, \]

which implies that \( [\beta^\top \beta, Z_1^\top Z_2] = 0 \) and \( [\beta^\top \beta, Z_1^\top Z_1] = 0 \). Hence \( \beta^\top \beta \) and \( Z_1^\top Z_1 \) commute and can be simultaneously diagonalizable, same is the case with \( \beta^\top \beta \), \( Z_1^\top Z_1 \). Therefore, the following relation holds (elementwise):

\[ D_{Z_2}^2 = -\gamma I_k + \sqrt{\gamma^2 I_k + D_{\beta}^2} \quad \text{and} \quad D_{Z_1}^2 = \gamma I_d + \sqrt{\gamma^2 I_d + D_{\beta}^2}; \quad (C.16) \]

where the appropriate dimensionality of the diagonal matrices is evident from the indexing of the identity matrices. \( \square \)

**Theorem C.11 (Singular vectors are static under orthogonal data).** Let \( \alpha := Z_1^\top Z_2 \). Then, it holds that

\[ U_\beta = U_{Z_1} = U_\alpha, \]
\[ V_\beta = V_{Z_2} = V_\alpha. \]

Furthermore, if \( X^\top X = I_d \), it additionally holds that

\[ U_\alpha = U_R = U_\beta, \]
\[ V_\alpha = V_R = V_\beta. \]

**Proof.** From invariance (C.8), by multiplying left and right with \( V_{Z_1}^\top \) and \( V_{Z_2}^\top \), respectively, we get that:

\[ D_{Z_1}^2 - V_{Z_1}^\top Z_2 Z_2^\top V_{Z_1} = 2\gamma I_d, \]

\[ 23 \]
which implies that $V_{Z_2}$ also diagonalizes $Z_2Z_2^\top$ (i.e., the left singular vectors of $Z_1$ are the same as the right singular vectors of $Z_2$ for all $t > 0$). As a result, $\beta = U_{Z_1}D_\beta V_{Z_2}$, and thus we have shown that $U_{\beta} = U_{Z_1}$ and $V_{\beta} = V_{Z_2}$.

Next, by definition $\alpha = Z_1^{-\top}Z_2$, and $\alpha$ and $\alpha_{\beta}$ have the same left and right singular vectors (since $Z_1^{-1} := V_{Z_2}D_{Z_2}^\top U_{Z_1}$). This, in conjunction with the above proves $U_{Z_1} = U_\alpha$ and $V_{Z_2} = V_\alpha$.

To show that $U_\alpha = U_R$ and $V_\alpha = V_R$, we need to do a bit more work. We will show that:

$$\alpha^\top R = R^\top \alpha,$$  \hfill (C.17)

$$\alpha R^\top = R \alpha^\top. \hfill (C.18)$$

Once we prove (C.17) and (C.18), it directly follows that

$$\begin{align*}
(\alpha \alpha^\top) (R R^\top) &= (R R^\top) (\alpha \alpha^\top) \quad \text{and} \\
(\alpha^\top \alpha) (R^\top R) &= (R^\top R) (\alpha^\top \alpha),
\end{align*}$$

which gives the desired result.

To prove (C.17) we will show that $R^\top \alpha - \alpha^\top R = 0, \forall t > 0$. First, we recall some relations that will be needed. Under orthogonal data, it holds that

$$R = (\beta^* - \beta) \quad \text{and} \quad \dot{R} = -\beta.$$  \hfill (C.20)

Furthermore, using the reparametrization in terms of $\alpha$ which induces identity (C.12) and the invariance (C.8), the original dynamics of $\beta$ given by $\dot{\beta} = RZ_2^\top Z_2 + Z_1Z_1^\top R$ rewrites as

$$\dot{\beta} = 2\gamma \left[ R \alpha^\top (I_d - \alpha \alpha^\top)^{-1} \alpha + (I_d - \alpha \alpha^\top)^{-1} R \right]. \hfill (C.21)$$

Now, using (C.21), it holds that

$$\begin{align*}
\dot{R} \alpha &= R^\top R - (R^\top \alpha)^2 - 2\gamma R^\top (I_d - \alpha \alpha^\top)^{-1} \alpha - 2\gamma (I_d - \alpha \alpha^\top)^{-1} \alpha R^\top \alpha \\
&= R^\top R - (R^\top \alpha)^2 - 2\gamma R^\top \alpha (I_k - \alpha^\top \alpha)^{-1} - 2\gamma \alpha^\top \alpha (I_k - \alpha^\top \alpha)^{-1} R^\top \alpha \\
&= R^\top R - (R^\top \alpha)^2 - 2\gamma R^\top \alpha (I_k - \alpha^\top \alpha)^{-1} - 2\gamma (I_k - \alpha^\top \alpha)^{-1} R^\top \alpha + R^\top \alpha,
\end{align*}$$

were we used the push-through identity $(I + UV)^{-1}U = U(I + UV)^{-1}$.

Denoting $C := R^\top \alpha$, computing the transpose version of the above ODE and using the fact that $C^2 - (C^\top)^2 = \frac{1}{2} (C + C^\top) (C - C^\top) + \frac{1}{2} (C - C^\top) (C + C^\top)$, the following holds:

$$\begin{align*}
\dot{C} - C^\top &= -C^2 - 2\gamma C(I_k - \alpha^\top \alpha)^{-1} - 2\gamma (I_k - \alpha^\top \alpha)^{-1} C + C + (C^\top)^2 \\
&= -\left[ C^2 - (C^\top)^2 \right] + (C - C^\top) - 2\gamma (I_k - \alpha^\top \alpha)^{-1} (C - C^\top) - 2\gamma (C - C^\top) (I_k - \alpha^\top \alpha)^{-1} \\
&= \left[ \frac{1}{2} (I - C - C^\top) - 2\gamma (I_k - \alpha^\top \alpha)^{-1} \right] (C - C^\top) \\
&\quad - (C - C^\top) \left[ \frac{1}{2} (I - C - C^\top) - 2\gamma (I_k - \alpha^\top \alpha)^{-1} \right].
\end{align*}$$

Finally, since $C(0) - C^\top(0) = 0$, it is a fixed point of the above equation and it implies that $C = C^\top, \forall t$. Thus, identity (C.17) is proven.

We proceed similarly for identity (C.18).

$$\dot{R} \alpha^\top = -2\gamma R \alpha^\top (I_d - \alpha \alpha^\top)^{-1} - 2\gamma (I_d - \alpha \alpha^\top)^{-1} R \alpha^\top + RR^\top + R \alpha^\top - (R \alpha^\top)^2.$$

Denoting $B := R \alpha^\top$ and computing $\dot{B} - B^\top$ we get:

$$\dot{B} - B^\top = \left[ \frac{1}{2} (I_d - B - B^\top) - 2\gamma (I_d - \alpha \alpha^\top)^{-1} \right] (B - B^\top)$$

$$+ (B - B^\top) \left[ \frac{1}{2} (I_d - B - B^\top) - 2\gamma (I_d - \alpha \alpha^\top)^{-1} \right].$$
Since \((B(0) - B^T(0)) = 0\) is a fixed point of the above equation, identity (C.18) is proven.

Finally, we show \(U_R = U_{\beta^*}\) and \(V_R = V_{\beta^*}\) using yet another invariance, namely
\[
\dot{Z}_1^T R Z_2^T = Z_2 R^T Z_1, \forall t > 0.
\]
(C.22)

We proceed by computing the associated time derivatives and showing that their difference is null:
\[
\dot{Z}_1^T R Z_2^T - Z_2 R^T Z_1 = Z_1^T \dot{R} Z_2^T - \dot{Z}_2^T R Z_1
\]
(C.23)
\[
= -Z_1^T R Z_2^T Z_2 Z_1^T - 2\gamma Z_1^T R Z_2^T Z_1^T + 2\gamma Z_2^T R^T Z_1 + Z_2 R^T Z_1 Z_2 Z_1^T
\]
(C.24)
\[
= (Z_2 R^T Z_1 - Z_1^T R Z_2^T) Z_2 Z_1^T + Z_2 Z_2^T (Z_2 R^T Z_1 - Z_1^T R Z_2^T)
\]
(C.25)
\[
- 2\gamma (Z_1^T R Z_2^T - Z_2^T R Z_1),
\]
(C.27)
where we used (C.20) and the invariance (C.8). Since \(Z_2(0) R^T(0) Z_1(0) - Z_1(0)^T R(0) Z_2(0)^T = 0\), it is a fixed point of the above equation and we have shown (C.22).

Finally, it holds that
\[
\dot{Z}_1^T R Z_2^T = Z_2 R^T Z_1 \iff Z_1^T \beta^* Z_2^T - Z_1^T \beta Z_2^T = Z_2^T \beta^* Z_1 - Z_2 \beta^T Z_1
\]
\[
\iff Z_1^T \beta^* Z_2^T = Z_2^T \beta^* Z_1
\]
\[
\iff \beta^* \alpha = \alpha^T \beta^*.
\]
(C.28)
where the second equivalence comes from the fact that \(Z_1^T Z_1\) and \(Z_2 Z_2^T\) are simultaneously diagonalizable and thus commute (from invariance (C.8)), and the second equivalence comes from multiplying left and right with \(Z_1^{-T}\) and \(Z_1^{-1}\), respectively and the definition of \(\alpha\). Furthermore,
\[
R^T \alpha = \alpha^T R \iff \beta^T \alpha - \beta^T \alpha = \alpha^T \beta^* - \alpha^T \beta
\]
\[
\iff \beta^T \alpha = \alpha^T \beta^*,
\]
(C.29)

From repeated applications of (C.28) and (C.29) we obtain that
\[
(\alpha^T \alpha) (\beta^T \beta^*) = (\beta^* \beta^*) (\alpha^T \alpha) \quad \text{and} \quad (\alpha \alpha^T) (\beta^* \beta^*) = (\beta^* \beta^*) (\alpha \alpha^T),
\]
(C.30)
and the final identity follows.

**Theorem C.12 (Time evolution of singular values - orthogonal data).** Assume that \(X^T X = I_d\). Then, the \(i^{th}\) singular value of \(\beta\) evolves over time as
\[
\sigma_{\beta,i}(t) = \frac{2\gamma}{1 - \left( \frac{\sqrt{\gamma + \sigma_{\beta,i}^2 + \gamma^2 - \exp(-2t \sqrt{\sigma_{\beta,i}^2 + \gamma^2})(\sqrt{\sigma_{\beta,i}^2 + \gamma^2})}}{\sigma_{\beta,i}} \right)^2}.
\]
As a consequence, under vanishing initialization \(\gamma \to 0\) and with a rescaling of time as \(t \to \ln(1/\gamma) t\), the \(i^{th}\) singular value is learned at time \(T_1 = 1/2\sigma_{\beta,i}\):
\[
\lim_{\gamma \to 0} \sigma_{\beta,i}(\ln(1/\gamma) t) = \begin{cases} 0, & \text{if } t < 1/2\sigma_{\beta,i} \\ \sigma_{\beta,i}, & \text{otherwise.} \end{cases}
\]

**Proof.** From Lemma 3.3 we have that \(\dot{\alpha} = R - \alpha R \alpha^T\). In light of Theorem C.11, identity (C.18) and it holds that:
\[
\dot{D}_\alpha = (I_d - \tilde{D}_\beta^2) D_{\beta^*} - 2\gamma D_\alpha.
\]

25
For the $i^{th}$ singular value of $\alpha$ it holds that:

$$\dot{\sigma}_{\alpha,i}(t) = -\sigma_{*,i}\sigma^2_{\alpha,i}(t) - 2\gamma\sigma_{\alpha,i}(t) + \sigma_{*,i}$$

$$= -\sigma_{*,i} \left[ \left(\sigma_{\alpha,i}(t) + \frac{\gamma}{\sigma_{*,i}}\right)^2 - \left(1 + \frac{\gamma^2}{\sigma_{*,i}^2}\right) \right], \quad (C.31)$$

where $\sigma_{*,i}$ is the $i^{th}$ singular value of $\beta^*$. Equation (C.31) is a Ricatti ODE and is separable. For ease of notation, let $p := \frac{\gamma}{\sigma_{*,i}}$, $q := 1 + \frac{\gamma^2}{\sigma_{*,i}^2}$ and $r := -\sigma_{*,i}$. Then, we need to solve the IVP:

$$\begin{cases}
\dot{\sigma}_{\alpha,i}(t) = r \left(\sigma_{\alpha,i}(t) + p\right)^2 - q \\
\sigma_{\alpha,i}(0) = 0,
\end{cases} \quad (C.32)$$

for which we get $\sigma_{\alpha,i}(t) = \sqrt{q} \left[ \frac{1 - \exp\left(2\sqrt{q}(t+c_1)\right)}{1 + \exp\left(2\sqrt{q}(t+c_1)\right)} \right] - p$ and solving for the initial value gives us $c_1 = \frac{1}{2\sqrt{q}} \log\left(\frac{\sqrt{q} - p}{\sqrt{q} + p}\right)$. Replacing $p, q, r$ and rearranging we finally obtain:

$$\sigma_{\alpha,i}(t) = \sqrt{1 + \frac{\gamma^2}{\sigma_{*,i}^2}} \left[ \frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} - \exp\left(-2t\sqrt{\sigma_{*,i}^2 + \gamma^2}\right)}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} + \exp\left(-2t\sqrt{\sigma_{*,i}^2 + \gamma^2}\right)}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2}}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} + \exp\left(-2t\sqrt{\sigma_{*,i}^2 + \gamma^2}\right)}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} - \gamma}{\sigma_{*,i}}} - \gamma}{\sigma_{*,i}}} - \gamma \right], \quad (C.33)$$

In order to obtain the dynamics for the singular values of $\beta = Z_i Z_i^\top \alpha$ we recall the relation given in Lemma C.10:

$$\dot{D}_{Z_i}^2 = \gamma I_d + \sqrt{\gamma^2 I_d + D_{\beta}^2},$$

where the appropriate dimensionality of the diagonal matrices is evident from the indexing of the identity matrices. Therefore, the singular values $\sigma_{\beta,i}(t)$ are the solutions to the equation

$$\sigma_{\beta,i}(t) = \sigma_{Z_{*,i}}^2(t) \sigma_{\alpha,i}(t) = \sigma_{\alpha,i}(t) \left[ \gamma + \sqrt{\gamma^2 + \sigma_{\beta,i}^2(t)} \right],$$

and have the following expression:

$$\sigma_{\beta,i}(t) = \frac{2\gamma \sigma_{\alpha,i}(t)}{1 - \sigma_{\alpha,i}^2(t)}$$

$$= \frac{2\gamma}{\left(1 + \frac{\gamma^2}{\sigma_{*,i}^2}\right) \left(\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} - \exp\left(-2t\sqrt{\sigma_{*,i}^2 + \gamma^2}\right)}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} + \exp\left(-2t\sqrt{\sigma_{*,i}^2 + \gamma^2}\right)}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2}}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} + \exp\left(-2t\sqrt{\sigma_{*,i}^2 + \gamma^2}\right)}{\frac{\gamma + \sqrt{\sigma_{*,i}^2 + \gamma^2} - \gamma}{\sigma_{*,i}}} - \gamma}{\sigma_{*,i}}} - \gamma}{\sigma_{*,i}}} - \gamma}{\sigma_{*,i}} \right),$$

Note that $\lim_{t \to \infty} \sigma_{\alpha,i}(t) = \frac{1}{\sigma_{*,i}} \left(\sqrt{\sigma_{*,i}^2 + \gamma^2} - \gamma\right)$ and therefore we can verify that $\beta \overset{t \to \infty}{\rightarrow} \beta^*$ by looking at the singular values, since the singular vectors are static (Lemma C.11).

$$\lim_{t \to \infty} \sigma_{\beta,i}(t) = \frac{2\gamma}{\sigma_{*,i}} \left(\sqrt{\sigma_{*,i}^2 + \gamma^2} - \gamma\right)$$

$$= \sigma_{*,i}.$$

We can now derive asymptotic transition times at which the singular values are learned. Now, we further process expression (C.33). Let $v := \sqrt{\sigma_{*,i}^2 + \gamma^2} - \gamma$, $w := \sqrt{\sigma_{*,i}^2 + \gamma^2 + \gamma}$, then
\[ v + w = 2\sqrt{\sigma^2_{*,i} + \gamma^2} \text{ and } w - v = 2\gamma. \] We re-write:

\[
\sigma_{\alpha,i}(t) = \frac{1}{2\sigma_{*,i}} \left[ (w + v) \left( \frac{w - \exp(-t(v + w)) v}{w + \exp(-t(v + w))v} \right) - (w - v) \right]
\]

\[
= \frac{1}{2\sigma_{*,i}} \left[ w \left( \frac{w - \exp(-t(v + w)) v}{w + \exp(-t(v + w))v} - 1 \right) + v \left( \frac{w - \exp(-t(v + w)) v}{w + \exp(-t(v + w))v} + 1 \right) \right]
\]

\[
= \frac{1}{\sigma_{*,i}} \left[ \frac{-vw \exp(-t(v + w))}{w + \exp(-t(v + w))v} + \frac{vw}{w + \exp(-t(v + w))v} \right]
\]

\[
= \frac{v}{\sigma_{*,i}} \left[ 1 - \frac{(v + w) \exp(-t(v + w))}{w + v \exp(-t(v + w))} \right].
\]

Since from before we have that \( \sigma_{\beta,i}(t) = \frac{2\gamma \sigma_{\alpha,i}(t)}{1 - \sigma^2_{\alpha,i}(t)} \) and we can write \( 1 - \sigma^2_{\alpha,i}(t) = \frac{2\gamma v}{\sigma^2_{*,i}} + \frac{v^2 h(t) (2 - h(t))}{\sigma^2_{*,i}} \), the singular values of \( \beta \) become:

\[
\sigma_{\beta,i}(t) = \frac{2\gamma \sigma_{\alpha,i}(t)}{1 - \sigma^2_{\alpha,i}(t)} = \frac{2\gamma}{\sigma^2_{*,i}} \frac{[1 - h(t)]}{1 + \frac{v^2 h(t) (2 - h(t))}{2\gamma}} = \frac{\sigma_{*,i} [1 - h(t)]}{1 + \frac{v^2 h(t) (2 - h(t))}{2\gamma}}.
\]

We wish to study the limit of infinitesimal initialization \( \gamma \to 0 \). We introduce a constant \( c > 0 \) and rewrite \( \gamma = \exp(-c) \) and have \( c = \ln(1/\gamma) \). Rescaling time \( t \to ct \) and taking the limit \( c \to \infty \) we have

\[
\lim_{c \to \infty} \sigma_{\beta,i}(ct) = \lim_{c \to \infty} \frac{2\sigma_{*,i} [1 - h(ct)]}{2 + v h(ct) \exp(c) (2 - h(ct))}
\]

\[
= \frac{\sigma_{*,i}}{1 + \lim_{c \to \infty} v h(ct) \exp(c)}
\]

\[
= \begin{cases} 
0, & \text{if } t < 1/2\sigma_{*,i}, \\
\frac{\sigma_{*,i}}{1 + 2\sigma_{*,i}}, & \text{if } t = 1/2\sigma_{*,i} \\
\sigma_{*,i}, & \text{if } t > 1/2\sigma_{*,i}, 
\end{cases}
\]

Since \( \lim_{c \to \infty} \exp(c) h(ct) = \lim_{c \to \infty} \frac{(v + w) \exp(c[1 - t(v + w)])}{w + v \exp(-ct(v + w))} = \lim_{c \to \infty} \frac{2\sqrt{\sigma^2_{*,i} + \exp(-2c)} \exp(c(1 - 2t\sigma_{*,i})) \exp\left(\frac{-2ct \exp(-2c)}{\sigma_{*,i} + \sqrt{\sigma^2_{*,i} + \exp(-2c)}}\right)}{\sqrt{\sigma^2_{*,i} + \exp(-2c)} + \exp(-c) + v \exp(-ct(v + w))}
\]

\[
= \begin{cases} 
\infty, & \text{if } t < 1/2\sigma_{*,i} \\
2, & \text{if } t = 1/2\sigma_{*,i} \\
0, & \text{if } t > 1/2\sigma_{*,i}. 
\end{cases}
\]

**Lemma C.13 (Mirror on singular values).** With the same notations as in Theorem 3.1, The singular values of \( \beta \), denoted by \( D_{\beta} \), follow the mirror flow

\[
d\nabla \Psi_\gamma (D_{\beta}) = -\nabla_{D_{\beta}} \mathcal{L}(\beta) \, dt,
\]

where the potential is \( \Psi_\gamma (D_{\beta}) := \text{tr} \left( \frac{1}{2} D_{\beta} \sinh^{-1} \left( \frac{D_{\beta}}{\gamma} \right) - \sqrt{D_{\beta}^2 + \gamma^2} \right) \).
Proof. We recall the dynamics induces on \( \alpha = Z_1^{-\top} Z_2 \) given in Lemma 3.3:
\[
\dot{\alpha} = R - \alpha R^\top \alpha. \tag{C.35}
\]
Writing the SVD decomposition of \( \alpha \) as \( U_\alpha D_\alpha V_\alpha^\top = U Z_1 D_\alpha V Z_2 \), we have that:
\[
\dot{U}_Z \dot{V} Z_2 = U_{\alpha} \dot{D}_\alpha V_{Z_2}^\top V Z_2
\]
\[
= U_{Z_1} \dot{U}_{Z_1} D_\alpha V_{Z_2}^\top V Z_2 + U_{Z_1} U_{Z_1} \dot{D}_\alpha V_{Z_2}^\top V Z_2 + U_{Z_1} U_{Z_1} D_\alpha \dot{V}_{Z_2}^\top V Z_2
\]
\[
= U_{Z_1} \dot{U}_{Z_1} D_\alpha + \dot{D}_\alpha + D_\alpha U_{Z_1} \dot{V}_{Z_2}^\top V Z_2.
\]
Since \( U_{Z_1} \dot{U}_{Z_1} + U_{Z_1} U_{Z_1} = 0 \) and \( V_{Z_2}^\top V Z_2 + \dot{V}_{Z_2}^\top V Z_2 = 0 \), the matrices \( U_{Z_1} \dot{U}_{Z_1} \) and \( V_{Z_2}^\top \dot{V}_{Z_2} \) are skew-symmetric (have 0 diagonal), and therefore the principal diagonals of the products \( U_{Z_1} \dot{U}_{Z_1} D_\alpha \) and \( D_\alpha V_{Z_2}^\top V Z_2 \) are also 0.

We define the linear operator \( \text{diag} : \mathbb{R}^{d \times k} \to \mathbb{R}^{d \times k} \) which maps any \( A \in \mathbb{R}^{d \times k} \) to a matrix of the same dimensions whose principal diagonal contains the elements on the principal diagonal of \( A \), and zeros otherwise. With this notation, we have that \( \text{diag}(U_{Z_1} \dot{V} Z_2) = \dot{D}_\alpha \).

We similarly apply the orthogonal matrices to the left and right of the RHS of (C.35), letting \( R' := U_{Z_1} R V Z_2 \in \mathbb{R}^{d \times k} \):
\[
\text{diag} (U_{Z_1} (R - \alpha R^\top \alpha) V Z_2) = \text{diag} (R' - D_\alpha R'^\top D_\alpha)
\]
\[
= \text{diag} (R') - \text{diag} (D_\alpha R'^\top D_\alpha)
\]
\[
= \text{diag} (R') - D_\alpha \text{diag} (R'^\top D_\alpha)
\]
\[
= (I_d - \dot{D}_\alpha) \text{diag} (R').
\]
Therefore it holds that:
\[
\dot{D}_\alpha = (I_d - \dot{D}_\alpha) \text{diag} (R'). \tag{C.36}
\]
We now wish to arrive at an expression in \( \beta \). From the definition of \( \alpha \) it holds that \( \alpha = (Z_1 Z_1^\top)^{-1} \beta \) and, in conjunction with the first part of Theorem C.11 it holds that \( D_\alpha = (\dot{D}_{Z_1})^{-1} D_{\beta} \). Therefore, the LHS of (C.36) becomes:
\[
\dot{D}_\alpha = (\dot{D}_{Z_1})^{-1} D_{\beta} + (\dot{D}_{Z_1}^2)^{-1} \dot{D}_{\beta}.
\tag{C.37}
\]
For computing \( (\dot{D}_{Z_1})^{-1} \) it is perhaps easiest to proceed as we did with \( \alpha \).
\[
U_{Z_1}^\top (Z_1 Z_1^\top)^{-1} U_{Z_1} = -U_{Z_1}^\top (Z_1 Z_1^\top)^{-1} (Z_1 Z_1^\top)^{-1} U_{Z_1}
\]
\[
= -(\dot{D}_{Z_1})^{-1} (D_{\beta} R'^\top + R' D_{\beta}^\top) (\dot{D}_{Z_1})^{-1},
\]
and therefore
\[
(\dot{D}_{Z_1})^{-1} = \text{diag} \left( U_{Z_1}^\top (Z_1 Z_1^\top)^{-1} U_{Z_1} \right)
\]
\[
= -(\dot{D}_{Z_1})^{-1} (D_{\beta} \text{diag} (R'^\top) + \text{diag} (R'^\top) D_{\beta}) (\dot{D}_{Z_1})^{-1}
\]
\[
= -2(\dot{D}_{Z_1})^{-2} \text{diag} (R'^\top).
\]
Putting everything back together in (C.37) we have that
\[
\dot{D}_\alpha = -2(\dot{D}_{Z_1})^{-2} \text{diag} (R'^\top) D_{\beta} + (\dot{D}_{Z_1})^{-1} \dot{D}_{\beta}
\tag{C.38}
\]
\[
= -2(\dot{D}_{Z_1})^{-2} \dot{D}_{\beta} \text{diag} (R') + (\dot{D}_{Z_1})^{-1} \dot{D}_{\beta}.
\tag{C.39}
\]
Finally, for dealing with the RHS of (C.36), we recall equation (C.12) which is implies that
\[ 2\gamma(D_{Z_1}^2)^{-1} = I_d - D_\alpha^2. \] Putting together (C.39) and (C.36), we get that:

\[ -2(D_{Z_1}^2)^{-2}D_\beta \text{diag}(R') + (D_{Z_1}^2)^{-1}D_\beta = 2\gamma(D_{Z_1}^2)^{-1} \text{diag}(R'), \tag{C.40} \]

Therefore, we have the following string of implications:

\[
\begin{align*}
(C.40) \iff & \frac{1}{2} [\gamma I + D_\beta^2 (D_{Z_1}^2)^{-1}]^{-1} \dot{D}_\beta = \text{diag}(R') \\
\iff & \frac{1}{2} D_{Z_1} [\gamma (D_{Z_1}^2) + D_\beta^2]^{-1} \dot{\gamma}_\beta = \text{diag}(R') \\
\iff & \frac{1}{2} \left( \sqrt{\gamma^2 I_d + D_\beta^2} \right)^{-1} \dot{D}_\beta = \text{diag}(R') \\
\iff & \frac{1}{2} \left( \frac{1}{2} \gamma D_\beta \sinh^{-1}(\frac{1}{\gamma} D_\beta) \right) = \text{diag}(R') .
\end{align*}
\]

As a final step, we remark that \( \text{diag}(R') = -\nabla_{D_\beta} \mathcal{L}(\beta) \) and that
\[
\nabla_{D_\beta} \text{tr} \left( \frac{1}{2} D_\beta \sinh^{-1}(D_\beta/\gamma) - \sqrt{D_\beta^2 + \gamma^2} \right) = \sinh^{-1}(\frac{1}{\gamma} D_\beta).
\]

C.1 Extensions and further experiments

In this subsection, we describe how the insights from our analysis can be extend to relaxed assumptions on initialization, discrete gradient descent and non-linear activations.

**Perturbations from assumption on initialization.** Our analysis is contingent upon two assumptions regarding the initialization shape. The first assumption concerns the orthogonality of the initial feature layer \( W_1(0) \). In Figure 1a, we explored a scenario where the feature layer is initialized with a random Gaussian matrix, yet the evolution of singular values closely aligns with our theoretical analysis for orthogonal initialization. In Figure 4, we demonstrate the impact of zero initialization for the weight layer \( W_2 \). We maintain the same experimental setup as in Figure 1a, but employ a random Gaussian initialization for \( W_2 \) with variance scales of \( 10^{-2} \) and \( 10^{-3} \), while initializing \( W_1 \) with a variance of \( 10^{-3} \). In this context, the evolution of singular values continues to adhere to the predicted trend from our analysis.

![Figure 4: The time evolution of singular values of the hidden layer weights of a 2-layer linear network when trained with gradient flow initialized with Gaussian random variables and non-zero \( W_2 \).](image)

**Discrete step size.** Here we present a simpler problem to show how we can go beyond continuous time analysis. Consider the problem with \( l = d, k = 1, W_1, W_2 = W, a \) and \( W \) is initialized with
The evolution of $W, a$ with a learning rate $\eta$ can be written as

$$W_{t+1} = W_t + \eta R_t a_t^\top,$$
$$a_{t+1} = a_t + \eta W_t^\top a_t,$$
$$R_{t+1} = R_t - \eta (X^\top X) (W_t W_t^\top + \|a_t\|^2) R_t - \eta^2 (a_t^\top W_t R_t) (X^\top X) R_t.$$

If we further assume that $X^\top X = I$, then it can be shown that $R_t, a_t$ only grow in norm and do not change in direction. So the update of $W_t$ is always aligned with the rank-1 matrix $R_0 a_0^\top$. Hence for small initialization, the final $W_\infty$ is approximately a rank-1 matrix. This presents a way forward for the discrete step-size case and orthogonal data. With further analysis, we think this can be generalized to any data matrix satisfying the RIP conditions. This is not included in the main paper due to restrictive assumptions on the data but we will include these comments in the appendix.

**Non-linear activations.** We consider the same teacher-student setup as in the Figure 2. To completely characterize the dynamics and the final weight parameters is a challenging problem. The characterization of the dynamics at the small scale of initialization is also absent (for any general data matrix), Boursier et al. [2022] solves this in the case of orthogonal data. To study even this simple case of two neurons, one has to study the dynamics in two phases where one jumps from zero initialization (a saddle point) to another saddle and then further converge to zero train loss as seen in the Figure 5a. It is difficult to precisely characterize this intermediate saddle. With some careful additional work, we believe that our analysis can capture Phase 1 (where you jump to the first saddle) of the dynamics where you approximately learn rank 1 matrix, see Figure 5b, for general data matrices extending the current understanding. We hope this briefly sketches a way forward for ReLU networks.

![Figure 5: (a) The training curve of the teacher-student network which follows a saddle-to-saddle dynamics. (b) The time evolution of singular values of the hidden layer weights of a 2-layer ReLU network when trained with gradient flow. The plot represents Phase 1 of the training where you first learn a (approximately) rank-1 hidden layer.](image)

## D Noise Dynamics

**Noise Model.** Here we consider the scalar case ($k = 1$), abusing the notation $W = W_1, a = W_2$. The gradients of the loss $\mathcal{L}(W, a)$ are

$$\nabla_W \mathcal{L} = -X^\top (Y - XW a) a^\top, \quad \nabla_a \mathcal{L} = -W^\top X^\top (Y - XW a).$$

When the labels or outputs are doped with a noise of magnitude $\delta > 0$, i.e., adding $\varepsilon \sim \delta N(0, I_n)$. Now the gradients computed after doping with this label noise are

$$\nabla_W \hat{\mathcal{L}} = -X^\top (Y + \varepsilon - XW a) a^\top, \quad \nabla_a \hat{\mathcal{L}} = -W^\top X^\top (Y + \varepsilon - XW a),$$

$$\nabla_W \hat{\mathcal{L}} = \nabla_W \mathcal{L} - X^\top \varepsilon a^\top, \quad \nabla_a \hat{\mathcal{L}} = \nabla_a \mathcal{L} - W^\top X^\top \varepsilon.$$
Thus label noise gradient descent with step size $\eta$ and with added label noise $\varepsilon_t$ at each iteration writes

$$W_{t+1} = W_t - \eta \left( \nabla_{W} \mathcal{L}(W_t, a_t) - X^T \varepsilon_t a_t^T \right).$$

The continuous time version of this SDE writes,

$$dW = -\eta \nabla_{W} \mathcal{L}(W, a) \, dt + \eta \delta X^T dB_t a^T.$$

Now, we consider the large noise regime, where the dominating term in the SDE is the diffusion term. Therefore we consider the SDE,

$$dW = \eta \delta X^T dB_t a^T,$$

where $B_t$ is a $n$-dimensional Brownian motion. Note that we can get rid of the term $\eta \delta$ by re-scaling time by a constant factor. Similar steps for the evolution of $a$ gives the SDE,

$$dW = X^T dB_t a^T, \quad da = W^T X^T dB_t.$$

Now consider the compact SVD decomposition of $X$, i.e., $X = UDV^T$, where $U, D \in \mathbb{R}^{d \times n}$ and $V \in \mathbb{R}^{d \times d}$, and $U^T U = V^T V = I_n$.

$$dW = VDU^T dB_t a^T, \quad da = W^T VDU^T dB_t,$$

$$dV^T W = D (U^T dB_t) a^T, \quad da = (V^T W)^T D (U^T dB_t).$$

Using Levy’s characterization $Y_t = U^T B_t$ is a Brownian motion, since $U$ is an orthogonal matrix. Let $W = V^T W$, then

$$dW = D dY_t a^T, \quad da = W^T D dY_t.$$

Here we consider $D = I$ and change the notation to get the SDE (4.1) below. Our results can be extended to any diagonal matrix $D$.

$$dW = (dB_t) a^T, \quad da = W^T dB_t.$$

**Proposition 4.1.** The dynamics (4.1) has the following convergence properties

(a) **Variance explosion.** The variance of the norms of $W, a$ explode, i.e.,

$$\lim_{t \to \infty} E \left[ \|W(t)\|^2 \right] \to \infty, \quad \lim_{t \to \infty} E \left[ \|a(t)\|^2 \right] \to \infty.$$

(b) **Scale divergence.** For $d \geq 5$, for any $\alpha > 0$, we have that,

$$\lim_{t \to \infty} E \left[ \|W(t)\|^\alpha \right] \to \infty, \quad \lim_{t \to \infty} E \left[ \|a(t)\|^\alpha + \tilde{a}(t)\|^\alpha \right] \to \infty,$$

where $\tilde{a} := e^{-t} \int_0^t e^s a(s) ds$ is the exponential moving average of $a$.

(c) **Alignment - spectral bias.** Denote the $i^{th}$ row of $W$ as $w_i$. Using $[w_i, a] \overset{\text{def}}{=} w_i a^T - aw_i^T$,

$$\lim_{t \to \infty} E \left[ [[w_i, a]] \right] \to 0.$$

**Proof.** Consider the noise model Eq. (4.1),

$$dW = (dB_t) a^T, \quad da = W^T dB_t,$$

**Variance.** Let $w_i$ be the $i^{th}$ column of $W$ and $a_i$ be the $i^{th}$ coordinate of $a$. For any $i \in [n]$, the diffusion of the quantities can be separately written as

$$dw_i = a_i dB_t, \quad da_i = \langle w_i, dB_t \rangle.$$

Using the Itô chain rule,

$$d\|w_i\|^2 = 2 \langle w_i, dw_i \rangle + \langle dw_i, dw_i \rangle = d\|a_i\|^2 dt + 2a_i \langle w_i, dB_i \rangle.$$
Similarly,
\[ da_i^2 = 2a_i da_i + da_i da_i = 2a_i \langle w_i, dB_i \rangle + \|w_i\|^2 dt. \]

Note that \( \langle w_i, dB_i \rangle \sim \|w_i\| dB_i \) for some one-dimensional Brownian motion \( (B_i)_{t \geq 0} \).
\[ da_i^2 = \|w_i\|^2 dt + 2a_i \|w_i\| dB_i, \quad d\|w_i\|^2 = d\|a_i\|^2 dt + 2a_i \|w_i\| dB_i. \]

Let \( v := a_i, u := \|w_i\| \), using this notation we get,
\[ du^2 = dv^2 dt + 2uv dB_i, \quad dv^2 = u^2 dt + 2uv dB_i. \quad (D.2) \]

Let \( u_0 := \mathbb{E} [u^2], v_0 := \mathbb{E} [v^2] \). Using the Dynkin's formula, taking the expectation, we get,
\[ d\mathbb{E} [u^2] = d\mathbb{E} [v^2] dt, \quad d\mathbb{E} [v^2] = \mathbb{E} [u^2], \quad du_0 = dv_0 dt, \quad dv_0 = u_0 dt. \]

This system can be transformed into,
\[ d(u_0 + \sqrt{v_0}) = \sqrt{d} (u_0 + \sqrt{v_0}) dt. \]

Solving the above ODE we get,
\[ (u_0 + \sqrt{v_0}) = c_0 e^{\sqrt{d} t}, \quad \text{where} \quad c_0 = u_0(0) + \sqrt{v_0}(0). \]

Similarly,
\[ d(u_0 - \sqrt{v_0}) = -\sqrt{d} (u_0 - \sqrt{v_0}) dt, \quad (u_0 - \sqrt{v_0}) = c_1 e^{-\sqrt{d} t}, \quad \text{where} \quad c_1 = u_0(0) - \sqrt{v_0}(0). \]

\[ u_0 = \frac{1}{2} \left[ 2c_0 e^{\sqrt{d} t} + c_1 e^{-\sqrt{d} t} \right], \quad v_0 = \frac{1}{2\sqrt{d}} \left[ 2c_0 e^{\sqrt{d} t} - c_1 e^{-\sqrt{d} t} \right]. \]

Taking the limit proves the first part of the result.

**Scale.** From Eq. D.2, we have,
\[ d(u^2 + \sqrt{v}v^2) = \sqrt{n}(u^2 + \sqrt{v}v^2) + 2(\sqrt{d} + 1) uv dB_i. \]

Again using the Itô chain rule, we get,
\[
\begin{align*}
\text{drift term is} & \quad \alpha \sqrt{n} \left( u^2 + \sqrt{v}v^2 \right) \alpha + 2\alpha (\alpha - 1) \left( \sqrt{d} + 1 \right)^2 \left( u^2 + \sqrt{v}v^2 \right) \alpha - u^2 v^2 \\
&= \alpha \left( u^2 + \sqrt{v}v^2 \right) ^\alpha \left[ \sqrt{d} + 2(\alpha - 1) \left( \sqrt{d} + 1 \right)^2 \frac{u^2 v^2}{(u^2 + \sqrt{v}v^2)^2} \right].
\end{align*}
\]

Again using the Dynkin's formula for the evolution of expectation, we have,
\[
\begin{align*}
\text{scale.} \quad d\mathbb{E} \left[ \left( u^2 + \sqrt{v}v^2 \right) ^\alpha \right] &= \mathbb{E} \left[ \alpha \left( u^2 + \sqrt{v}v^2 \right) ^\alpha \left[ \sqrt{d} + 2(\alpha - 1) \left( \sqrt{d} + 1 \right)^2 \frac{u^2 v^2}{(u^2 + \sqrt{v}v^2)^2} \right] \right] dt. \quad \text{(D.3)}
\end{align*}
\]
For any function,

\[ g(y) = \frac{y}{(y + \sqrt{d})^2}, \]

attains its maximum value at \( y = \sqrt{d} \), i.e., \( g(\sqrt{n}) = 1/(4\sqrt{d}) \). Note that \( \alpha < 1 \) and we have

\[ 2(\alpha - 1) \left( \sqrt{\alpha} + 1 \right)^2 \left( \frac{u^2v^2}{(u^2 + \sqrt{\alpha}v^2)} \right) \geq 2(\alpha - 1) \left( \sqrt{\alpha} + 1 \right)^2 g(\sqrt{n}) = (\alpha - 1) \frac{(\sqrt{\alpha} + 1)^2}{2\sqrt{n}}. \]

The drift can be lower bounded as the following,

\[ \alpha \sqrt{n} \left( u^2 + \sqrt{\alpha}v^2 \right)^\alpha + 2\alpha(\alpha - 1) \left( \sqrt{\alpha} + 1 \right)^2 \left( u^2 + \sqrt{\alpha}v^2 \right)^{\alpha-2} u^2 v^2 \geq \alpha \left( u^2 + \sqrt{\alpha}v^2 \right)^\alpha \left[ \sqrt{\alpha} + (\alpha - 1) \frac{(\sqrt{\alpha} + 1)^2}{2\sqrt{d}} \right]. \]

For \( d \geq 5 \) and \( 0 < \alpha < 1 \), we have \( c_0 > 0 \),

\[ \left[ \sqrt{\alpha} + (\alpha - 1) \frac{(\sqrt{\alpha} + 1)^2}{2\sqrt{d}} \right] > c_0. \]

Using the above expression in Eq. D.3,

\[ dE \left[ (u^2 + \sqrt{\alpha}v^2)^\alpha \right] \geq \alpha c_0 E \left[ (u^2 + \sqrt{\alpha}v^2)^\alpha \right] dt. \]

Taking the limit,

\[ \lim_{t \to \infty} E \left[ (u^2 + \sqrt{\alpha}v^2)^\alpha \right] \to \infty. \]

From the SDE (D.2), we obtain the following process with only diffusion,

\[ d(u^2 - v^2) = (du^2 - u^2)dt, \]
\[ d(u^2 - v^2) + (u^2 - v^2) dt = (n-1)v^2, \]
\[ e^t d(u^2 - v^2) + e^t (u^2 - v^2) dt = (n-1)e^t v^2 dt, \]
\[ de^t (u^2 - v^2) = (n-1)e^t v^2 dt, \]
\[ (u^2(t_2) - v^2(t_2)) = e^{-(t_2-t_1)} (u^2(t_1) - v^2(t_1)) + (n-1) \int_{t_1}^{t_2} e^{-(t_2-t)} v^2(t) dt, \]
\[ (u^2(t_2) + \sqrt{\alpha}v^2(t_2)) = e^{-(t_2-t_1)} (u^2(t_1) - v^2(t_1)) + \sqrt{\alpha}v^2(t_2) + (n-1)e^{-t_2} \int_{t_1}^{t_2} e^t v^2(t) dt, \]

Denote \( C(t) := u^2(t) + \sqrt{\alpha}v^2(t) \), using the fact that \( u^2(t_1) - v^2(t_1) \leq C(t_1) \),

\[ C(t_2) \leq e^{-(t_2-t_1)} C(t_1) + \sqrt{\alpha}v^2(t_2) + (d-1)e^{-t_2} \int_{t_1}^{t_2} e^t v^2(t) dt, \]

With \( t_1 = 0 \), using the notation \( \tilde{v}^2(t) := e^{-t} \int_0^t e^s v^2(s) ds \), to denote the exponential moving average.

For any time \( t > 0 \), we have,

\[ C(t) \leq e^{-t} C(0) + d (u^2(t) + \tilde{v}^2(t)). \]

Raising to the power of \( \alpha \),

\[ C(t)^\alpha \leq \left[ e^{-t} C(0) + d (u^2(t) + \tilde{v}^2(t)) \right]^\alpha. \]
For $a, b > 0$ and $0 < p < 1$, we have $(a + b)^p \leq a^p + b^p$. Using the inequality,

$$C(t)^\alpha \leq e^{-\alpha t} C(0)^\alpha + d^\alpha \left( (w^2(t))^\alpha \right) + d^\alpha \left( (\tilde{w}^2(t))^\alpha \right).$$

Taking the expectation,

$$\mathbb{E} \left[ C(t)^\alpha \right] \leq e^{-\alpha t} \mathbb{E} \left[ C(0)^\alpha \right] + d^\alpha \mathbb{E} \left[ (w^2(t))^\alpha \right] + d^\alpha \mathbb{E} \left[ (\tilde{w}^2(t))^\alpha \right].$$

Now, we proceed by taking the limit,

$$\lim_{t \to \infty} \mathbb{E} \left[ C(t)^\alpha \right] \leq e^{-\alpha t} \lim_{t \to \infty} \mathbb{E} \left[ C(0)^\alpha \right] + d^\alpha \lim_{t \to \infty} \mathbb{E} \left[ (w^2(t))^\alpha \right] + d^\alpha \mathbb{E} \left[ (\tilde{w}^2(t))^\alpha \right].$$

Thus, we obtain,

$$\lim_{t \to \infty} \left( \mathbb{E} \left[ (w^2(t))^\alpha \right] + \mathbb{E} \left[ (\tilde{w}^2(t))^\alpha \right] \right) \to \infty,$$

where $\tilde{w}^2(t) := e^{-t} \int_0^t e^s w^2(s) ds$ is the exponential moving average.

A similar computation for $u$ will yield,

$$\lim_{t \to \infty} \mathbb{E} \left[ (u^2(t))^\alpha \right] \to \infty.$$

Therefore for the limit of $a_i$, $\|w_i\|$, we have,

$$\lim_{t \to \infty} \mathbb{E} \left[ \|w_i(t)\|^{2\alpha} \right] \to \infty, \quad \lim_{t \to \infty} \left( \mathbb{E} \left[ (a_i^2(t))^{\alpha} \right] + \mathbb{E} \left[ (\tilde{a}_i^2(t))^{\alpha} \right] \right) \to \infty.$$

Now we proceed to combine the above results and obtain the result on $a, W$. Note that for $0 < \alpha < 1$, $x^\alpha$ is concave. Further using the Jensen’s inequality, we have,

$$\left( \|W\|^2 \right)^\alpha = \left( \sum_{i=1}^d \|w_i\|^2 \right)^\alpha \geq d^{1-\alpha} \sum_{i=1}^d \|w_i(t)\|^{2\alpha}.$$

Similar expression for $a$ and taking the limit, we obtain the result

$$\lim_{t \to \infty} \mathbb{E} \left[ \|W(t)\|^\alpha \right] \to \infty, \quad \lim_{t \to \infty} \mathbb{E} \left[ \|a(t)\|^\alpha + \|\tilde{a}(t)\|^\alpha \right] \to \infty.$$

where $\tilde{a} := e^{-t} \int_0^t e^s a(s) ds$ is the exponential moving average of $a$.

**Alignment.** Let $z_1, z_2 \ldots z_d$ be the rows of the matrix $W$. Now the evolution of the rows can be written as

$$z'_i = a(dB_i^t),$$

$$da = \sum_{j=1}^d z_j dB_i^t.$$

For any two matrices, with same dimensions define $[u, v] \overset{\text{def}}{=} uu^\top - vv^\top$.

$$d[z_i, a] = d(z_i a^\top) - d(az_i^\top),$$
Using the Itô chain rule,

\[ d(z_i^a) = dz_i^a + z_i^a db^T + dz_i^a da, \]

\[
= aa^T d(B_i^j) + z_i \left( \sum_{j=1}^{d} z_j^a d(B_i^j) \right) + (a(dB_i^j)) \left( \sum_{j=1}^{d} z_j^a d(B_i^j) \right), 
\]

\[
= aa^T d(B_i^j) + \sum_{j=1}^{d} z_i z_j^a d(B_i^j) + az_i^a dt, 
\]

\[ d(az_i^a) = aa^T d(B_i^j) + z_i a^T dt + \sum_{j=1}^{d} z_j z_i^a d(B_i^j), \]

\[ d[z_i, a] = -[z_i, a] dt + \sum_{i=1}^{d} [z_i, z_j] dB_i^j, \]

From the above evolution, we have that,

\[ d\mathbb{E}([z_i, a]) = -\mathbb{E}([z_i, a]) dt. \]

Hence, we have,

\[ \mathbb{E}([z_i, a]) = [z_i(0), a(0)] e^{-t}. \]

Let \( e_i \overset{\text{def}}{=} [z_i, a] [k, l] \), be any \((kl)^{th}\) entry of the matrix. Similarly, let \( c_{ij} \overset{\text{def}}{=} [z_i, z_j] [k, l]. \)

\[ de_i = -e_i dt + \sum_j c_{ij} dB_i^j, \]

\[ de_i^2 = -2e_i \left( e_i dt + \sum_j c_{ij} dB_i^j \right) + \sum_j c_{ij}^2 dt, \]

\[ de_i^2 = -2e_i^2 dt + \sum_j c_{ij}^2 dt - 2e_i \sum_j c_{ij} dB_i^j. \]

Again using the Ito formula and computing \((e_i^2)^\alpha\) for some \(\alpha \in (0, 1)\), we get,

\[ d(e_i^2)^\alpha = \alpha (e_i^2)^{\alpha - 1} de_i^2 + 1/2 \alpha (\alpha - 1) (e_i^2)^{\alpha - 2} de_i^2 d(e_i^2), \]

\[ d(e_i^2)^\alpha = \alpha (e_i^2)^{\alpha - 1} \left[ -2e_i^2 dt + \sum_j c_{ij}^2 dt - 2e_i \sum_j c_{ij} dB_i^j \right] + 1/2 \alpha (\alpha - 1) (e_i^2)^{\alpha - 2} 4e_i^2 \left[ \sum_j c_{ij}^2 \right] dt, \]

\[ = -2\alpha (e_i^2)^{\alpha - 1} \sum_j c_{ij}^2 dt - 2\alpha (e_i^2)^{\alpha - 1} e_i \sum_j c_{ij} dB_i^j + 1/2 \alpha (\alpha - 1) (e_i^2)^{\alpha - 2} 4e_i^2 \left[ \sum_j c_{ij}^2 \right] dt, \]

\[ = -2\alpha (e_i^2)^{\alpha - 1} \sum_j c_{ij}^2 dt + \alpha (2\alpha - 1) (e_i^2)^{\alpha - 1} e_i \sum_j c_{ij} dB_i^j + 1/2 \alpha (\alpha - 1) (e_i^2)^{\alpha - 2} 4e_i^2 \left[ \sum_j c_{ij}^2 \right] dt, \]

Taking \(\alpha = 0.5\),

\[ d|e_i| = -|e_i| dt - |e_i|^{-1} e_i \sum_j c_{ij} dB_i^j. \]

Taking expectation, we get,

\[ d\mathbb{E}(|e_i|) = -\mathbb{E}(|e_i|) dt. \]

Hence,

\[ \mathbb{E}(|[z_i, a]|) = |[z_i(0), a(0)]| e^{-t}. \]