GRADIENT DESCENT ALIGNS THE LAYERS OF DEEP LINEAR NETWORKS

Anonymous authors
Paper under double-blind review

Abstract

This paper establishes risk convergence and asymptotic weight matrix alignment — a form of implicit regularization — of gradient flow and gradient descent when applied to deep linear networks on linearly separable data. In more detail, for gradient flow applied to strictly decreasing loss functions (with similar results for gradient descent with particular decreasing step sizes): (i) the risk converges to 0; (ii) the normalized $i$th weight matrix asymptotically equals its rank-1 approximation $u_i v_i^\top$; (iii) these rank-1 matrices are aligned across layers, meaning $|v_{i+1}^\top u_i| \to 1$. In the case of the logistic loss (binary cross entropy), more can be said: the linear function induced by the network — the product of its weight matrices — converges to the same direction as the maximum margin solution. This last property was identified in prior work, but only under assumptions on gradient descent which here are implied by the alignment phenomenon.

1 Introduction

Efforts to explain the effectiveness of gradient descent in deep learning have uncovered an exciting possibility: it not only finds solutions with low error, but also biases the search for low complexity solutions which generalize well (Zhang et al., 2017; Bartlett et al., 2017; Soudry et al., 2017; Gunasekar et al., 2018).

This paper analyzes the implicit regularization of gradient descent and gradient flow on deep linear networks and linearly separable data. For strictly decreasing losses, the optimum is off at infinity, and we establish various alignment phenomena:

- For each weight matrix $W_i$, the corresponding normalized weight matrix $W_i/\|W_i\|_F$ asymptotically equals its rank-1 approximation $u_i v_i^\top$, where the Frobenius norm $\|W_i\|_F$ satisfies $\|W_i\|_F \to \infty$. In other words, $\|W_i\|_F/\|W_i\|_F \to 1$, and asymptotically only the rank-1 approximation of each weight matrix contributes to the final predictor, a form of implicit regularization.

- Adjacent rank-1 weight matrix approximations are aligned: $|v_{i+1}^\top u_i| \to 1$.

- For the logistic loss, the first right singular vector $v_1$ of $W_1$ is aligned with the data, meaning $v_1$ converges to the unique maximum margin predictor $\bar{u}$ defined by the data. Moreover, the linear predictor induced by the network, $w_{\text{prod}} := W_L \cdots W_1$, is also aligned with the data, meaning $w_{\text{prod}}/\|w_{\text{prod}}\| \to \bar{u}$.

Simultaneously, this work proves that the risk is globally optimized: it asymptotes to 0. Alignment and risk convergence are proved simultaneously; the phenomena are coupled within the proofs.

The paper is organized as follows. This introduction continues with related work, notation, and assumptions in Sections 1.1 and 1.2. The analysis of gradient flow is in Section 2, and gradient descent is analyzed in Section 3. The paper closes with future directions in Section 4; a particular highlight is a preliminary experiment on CIFAR-10 which establishes empirically that a form of the alignment phenomenon occurs on the standard nonlinear network AlexNet.
Figure 1: Visualization of main results on synthetic data with a 4-layer linear network compared to a 1-layer network (a linear predictor). Figure 1a shows the convergence of 1-layer and 4-layer networks to the same linear predictor on positive (blue) and negative (red) separable data. Figure 1b shows the alignment phenomenon in the 4-layer network, plotted against the risk. Specifically, for each layer, the ratio of spectral to Frobenius norms is plotted, and converges to 1. As in the theoretical analysis, the convergence in alignment and risk occur simultaneously.

1.1 RELATED WORK

On the implicit regularization of gradient descent, Soudry et al. (2017) show that for linear predictors and linearly separable data, the gradient descent iterates converge to the same direction as the maximum margin solution. Ji & Telgarsky (2018) further characterize such an implicit bias for general nonseparable data. Gunasekar et al. (2018) consider gradient descent on fully connected linear networks and linear convolutional networks. In particular, for the exponential loss, assuming the risk is minimized to 0 and the gradients converge in direction, they show that the whole network converges in direction to the maximum margin solution. These two assumptions are on the gradient descent process itself, and specifically the second one might be hard to interpret and justify. Compared with Gunasekar et al. (2018), this paper proves that the risk converges to 0 and the weight matrices align; moreover the proof here proves the properties simultaneously, rather than assuming one and deriving the other. Lastly, Arora et al. (2018) show for deep linear networks (and later Du et al. (2018) for ReLU networks) that gradient flow does not change the difference between squared Frobenius norms of any two layers. We use a few of these tools in our proofs; please see Sections 2 and 3 for details.

For a smooth (nonconvex) function, Lee et al. (2016) show that any strict saddle can be avoided almost surely with small step sizes. If there are only countably many saddle points and they are all strict, and if gradient descent iterates converge, then this implies (almost surely) they converge to a local minimum. In the present work, since there is no finite local minimum, gradient descent will go to infinity and never converge, and thus these results of Lee et al. (2016) do not show that the risk converges to 0.

There has been a rich literature on linear networks. Saxe et al. (2013) analyze the learning dynamics of deep linear networks, showing that they exhibit some learning patterns similar to nonlinear networks, such as a long plateau followed by a rapid risk drop. Arora et al. (2018) show that depth can help accelerate optimization. On the landscape properties of deep linear networks, Lu & Kawaguchi (2017), Laurent & von Brecht (2017) show that under various structural assumptions, all local optima are global. Zhou & Liang (2018) give a necessary and sufficient characterization of critical points for deep linear networks.

1.2 NOTATION, SETTING, AND ASSUMPTIONS

Consider a data set \( \{(x_i, y_i)\}_{i=1}^n \), where \( x_i \in \mathbb{R}^d, \|x_i\| \leq 1 \), and \( y_i \in \{-1, +1\} \). The data set is assumed to be linearly separable, i.e., there exists a unit vector \( u \) which correctly classifies every data point: for any \( 1 \leq i \leq n \), \( y_i \langle u, x_i \rangle > 0 \). Furthermore, let \( \gamma := \max_{\|u\|=1} \min_{1 \leq i \leq n} y_i \langle u, x_i \rangle > 0 \) denote the maximum margin, and \( \bar{u} := \arg\max_{\|u\|=1} \min_{1 \leq i \leq n} y_i \langle u, x_i \rangle \) denote the maximum margin solution (the solution to the hard-margin SVM).
A linear network of depth $L$ is parameterized by weight matrices $W_L,\ldots,W_1$, where $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$, $d_0 = d$, and $d_L = 1$. Let $W = (W_L, \ldots, W_1)$ denote all parameters of the network. The (empirical) risk induced by the network is given by

$$R(W) = R(W_L, \ldots, W_1) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i W_L \cdots W_1 x_i) = \frac{1}{n} \sum_{i=1}^{n} \ell(w_{\text{prod}} z_i),$$

where $w_{\text{prod}} := (W_L \cdots W_1)^\top$, and $z_i := y_i x_i$.

The loss $\ell$ is assumed to be continuously differentiable, unbounded, and strictly decreasing to 0. Examples include the exponential loss $\ell_{\exp}(x) = e^{-x}$ and the logistic loss $\ell_{\log}(x) = \ln(1 + e^{-x})$.

**Assumption 1.** $\ell' < 0$ is continuous, $\lim_{x \to -\infty} \ell(x) = \infty$ and $\lim_{x \to \infty} \ell(x) = 0$.

This paper considers gradient flow and gradient descent, where gradient flow $\{W(t) | t \geq 0, t \in \mathbb{R}\}$ can be interpreted as gradient descent with infinitesimal step sizes. It starts from some $W(0)$ at $t = 0$, and proceeds as

$$\frac{dW(t)}{dt} = -\nabla R(W(t)).$$

By contrast, gradient descent $\{W(t) | t \geq 0, t \in \mathbb{Z}\}$ is a discrete-time process given by

$$W(t+1) = W(t) - \eta_t \nabla R(W(t)),$$

where $\eta_t$ is the step size at time $t$.

We assume that the initialization of the network is not a critical point and induces a risk no larger than the risk of the trivial linear predictor 0.

**Assumption 2.** The initialization $W(0)$ satisfies $\nabla R(W(0)) \neq 0$ and $R(W(0)) \leq R(0) = \ell(0)$.

It is natural to require that the initialization is not a critical point, since otherwise gradient flow/descent will never make a progress. The requirement $R(W(0)) \leq R(0)$ can be easily satisfied, for example, by making $W_1(0) = 0$ and $W_L(0) \cdots W_2(0) \neq 0$. On the other hand, if $R(W(0)) > R(0)$, gradient flow/descent may never minimize the risk to 0. Proofs of those claims are given in Appendix[A].

## 2 Results for Gradient Flow

In this section, we consider gradient flow. Although impractical when compared with gradient descent, gradient flow can simplify the analysis and highlight proof ideas. For convenience, we usually use $W, W_k$, and $w_{\text{prod}}$, but they all change with (the continuous time) $t$. Only proof sketches are given here; detailed proofs are deferred to Appendix[B].

### 2.1 Risk Convergence

One key property of gradient flow is that it never increases the risk:

$$\frac{dR(W)}{dt} = \left\langle \nabla R(W), \frac{dW}{dt} \right\rangle = -\|\nabla R(W)\|^2 = -\sum_{k=1}^{L} \left\| \frac{\partial R}{\partial W_k} \right\|_F^2 \leq 0. \quad (1)$$

We now state the main result: under Assumptions[1] and [2] gradient flow iterates satisfy the following properties:

**Theorem 1.** Under Assumptions[1] and [2] gradient flow iterates satisfy the following properties:

- $\lim_{t \to \infty} R(W) = 0$.
- For any $1 \leq k \leq L$, $\lim_{t \to \infty} \|W_k\|_F = \infty$.
- For any $1 \leq k \leq L$, letting $(u_k, v_k)$ denote the first left and right singular vectors of $W_k$,

$$\lim_{t \to \infty} \left\| \frac{W_k}{\|W_k\|_F} - u_k v_k^\top \right\|_F = 0.$$
Moreover, for any $1 \leq k < L$, \( \lim_{t \to \infty} \left\| (v_{k+1}, u_k) \right\| = 1 \). As a result,

\[
\lim_{t \to \infty} \left\langle \frac{w_{\text{prod}}}{\prod_{k=1}^L \|W_k\|_F}, v_1 \right\rangle = 1,
\]

and thus \( \lim_{t \to \infty} \|w_{\text{prod}}\| = \infty \).

Theorem 1 is proved using two lemmas, which may be of independent interest. To show the ideas, let us first introduce a little more notation. Recall that \( \mathcal{R}(W) \) denotes the empirical risk induced by the deep linear network \( W \). Abusing the notation a little, for any linear predictor \( w \in \mathbb{R}^d \), we also use \( \mathcal{R}(w) \) to denote the risk induced by \( w \). With this notation, \( \mathcal{R}(W) = \mathcal{R}(w_{\text{prod}}) \), while

\[
\nabla \mathcal{R}(w_{\text{prod}}) = \frac{1}{n} \sum_{i=1}^n \ell'((w_{\text{prod}}, z_i)) z_i = \frac{1}{n} \sum_{i=1}^n \ell'(W_L \cdots W_1 z_i) z_i
\]

is in \( \mathbb{R}^d \) and different from \( \nabla \mathcal{R}(W) \), which has \( \sum_{k=1}^L d_k \) entries, as given below:

\[
\frac{\partial \mathcal{R}}{\partial W_k} = W_{k+1}^T \cdots W_L^T \nabla \mathcal{R}(w_{\text{prod}}) W_1^T \cdots W_{k-1}^T.
\]

Furthermore, for any \( R > 0 \), let

\[
B(R) = \left\{ W \left| \max_{1 \leq k \leq L} \|W_k\|_F \leq R \right. \right\}
\]

The first lemma shows that for any \( R > 0 \), the time spent by gradient flow in \( B(R) \) is finite.

**Lemma 1.** Under Assumption 7 and 2, for any \( R > 0 \), there exists a constant \( \epsilon(R) > 0 \), such that for any \( t \geq 1 \) and any \( W \in B(R) \), \( \|\partial \mathcal{R}/\partial W_1\|_F \geq \epsilon(R) \). As a result, gradient flow spends a finite amount of time in \( B(R) \) for any \( R > 0 \), and \( \max_{1 \leq k \leq L} \|W_k\|_F \) is unbounded.

Here is the proof sketch. If \( \|W_k\|_F \) are bounded, then \( \|\nabla \mathcal{R}(w_{\text{prod}})\|_F \) will be lower bounded by a positive constant, therefore if \( \|\partial \mathcal{R}/\partial W_1\|_F = \|W_L \cdots W_2\|_F \nabla \mathcal{R}(w_{\text{prod}}) \) can be arbitrarily small, then \( \|W_L \cdots W_2\|_F \) and \( \|w_{\text{prod}}\|_F \) can also be arbitrarily small, and thus \( \mathcal{R}(W) \) can be arbitrarily close to \( \mathcal{R}(0) \). This cannot happen after \( t = 1 \), otherwise it will contradict Assumption 2 and eq. (1).

To proceed, we need the following properties of linear networks from prior work (Arora et al. 2018; Du et al. 2018). For any time \( t \geq 0 \) and any \( 1 \leq k < L \),

\[
W_{k+1}^T(t)W_{k+1}(0) - W_{k+1}(0) = W_k(t)W_k^T(t) - W_k(0)W_k^T(0).
\]

To see this, just notice that

\[
W_{k+1}^T(t)W_{k+1}(0) = W_{k+1}^T \cdots W_L^T \nabla \mathcal{R}(w_{\text{prod}}) W_1^T \cdots W_k^T,
\]

Taking the trace on both sides of eq. (2), we have

\[
\|W_{k+1}(t)\|_F^2 - \|W_{k+1}(0)\|_F^2 = \|W_k(t)\|_F^2 - \|W_k(0)\|_F^2.
\]

In other words, the difference between the squares of Frobenius norms of any two layers remains a constant. Together with Lemma 1, it implies that all \( \|W_k\|_F \) are unbounded.

However, even if \( \|W_k\|_F \) are large, it does not follow necessarily that \( \|w_{\text{prod}}\|_F \) is also large. Lemma 2 shows that this is indeed true: for gradient flow, as \( \|W_k\|_F \) get larger, adjacent layers also get more aligned to each other, which ensures that their product also has a large norm.

For \( 1 \leq k \leq L \), let \( \sigma_k, u_k \), and \( v_k \) denote the first singular value (the 2-norm), the first left singular vector, and the first right singular vector of \( W_k \), respectively. Furthermore, define

\[
D := \max_{1 \leq k \leq L} \|W_k(0)\|_F^2 - \|W_L(0)\|_F^2 + \sum_{k=1}^{L-1} \|W_k(0)W_k^T(0) - W_{k+1}(0)W_{k+1}(0)\|_F^2,
\]

which depends only on the initialization. If for any \( 1 \leq k < L \), \( W_k(0)W_k^T(0) = W_{k+1}(0)W_{k+1}(0) \), then \( D = 0 \).
Lemma 2. The gradient flow iterates satisfy the following properties:

- For any \( 1 \leq k \leq L \), \( \|W_k\|_F^2 - \|W_k\|_F^2 \leq D \).
- For any \( 1 \leq k < L \), \( \langle v_{k+1}, u_k \rangle^2 \geq 1 - \left( \frac{\langle W_{k+1} (0) \rangle_+^2 + \|W_{k}(0)\|_F^2}{\|W_{k+1}\|_F^2} \right) \).
- Suppose \( \max_{1 \leq k \leq L} \|W_k\|_F \to \infty \), then \( \left| \langle \prod_{k=1}^L \|W_k\|_F, u_1 \rangle \right| \to 1 \).

The proof is based on eq. (2) and eq. (3). If \( W_k(0)W_k^T(0) = W_{k+1}(0)W_{k+1}(0) \), then eq. (2) gives that \( W_{k+1} \) and \( W_k \) have the same singular values, and \( W_{k+1} \)'s right singular vectors and \( W_k \)'s left singular vectors are the same. If it is true for any two adjacent layers, since \( W_L \) is a row vector, all layers have rank 1. With general initialization, we have similar results when \( \|W_k\|_F \) is large enough so that the initialization is negligible. Careful calculations gives the exact results in Lemma 2.

An interesting point is that the implicit regularization result in Lemma 2 helps establish risk convergence in Theorem 1. Specifically, by Lemma 2 if all layers have large norms, \( \|W_L \cdots W_2\| \) will be large. If the risk is not minimized to 0, \( \|\nabla R(w_{\text{prod}})\|_F \) will be lower bounded by a positive constant, and thus \( \|\partial R/\partial W_1\|_F = \|W_L \cdots W_2\| \|\nabla R(w_{\text{prod}})\|_F \) will be large. Invoking eq. (1), Lemma 1 and eq. (3) gives a contradiction. Since the risk has no finite optimum, \( \|W_k\|_F \to \infty \).

2.2 Convergence to the maximum margin solution

Here we focus on the exponential loss \( \ell_{\exp}(x) = e^{-x} \) and the logistic loss \( \ell_{\log}(x) = \log(1 + e^{-x}) \).

In addition to risk convergence, these two losses also enable gradient descent to find the maximum margin solution.

To get such a strong convergence, we need one more assumption on the data set. Recall that \( \gamma = \max_{\|u\|_1} \min_{1 \leq i \leq n} \langle u, z_i \rangle > 0 \) denotes the maximum margin, and \( \hat{u} \) denotes the unique maximum margin predictor which attains this margin \( \gamma \). Those data points \( z_i \) for which \( \langle \hat{u}, z_i \rangle = \gamma \) are called support vectors.

Assumption 3. The support vectors span the whole space \( \mathbb{R}^d \).

Assumption 3 appears in prior work \cite{Soudry_2017}, and can be satisfied in many cases: for example, it is almost surely true if the number of support vectors is larger than or equal to \( d \) and the data set is sampled from some density w.r.t. the Lebesgue measure. It can also be relaxed to the situation that the support vectors and the whole data set span the same space; in this case \( \nabla R(w_{\text{prod}}) \) will never leave this space, and we can always restrict our attention to this space.

With Assumption 3, we can state the main theorem.

Theorem 2. Under Assumptions 2 and 3 for almost all data and for losses \( \ell_{\exp} \) and \( \ell_{\log} \), then \( \lim_{t \to \infty} \langle v_1, \hat{u} \rangle = 1 \), where \( v_1 \) is the first right singular vector of \( W_1 \). As a result, \( \lim_{t \to \infty} w_{\text{prod}}/\prod_{k=1}^L \|W_k\|_F = \hat{u} \).

Theorem 2 relies on two structural lemmas. The first one is based on a similar almost-all argument due to \cite{Soudry_2017} (Lemma 12). Let \( S \subset \{1, \ldots, n\} \) denote the set of indices of support vectors.

Lemma 3. Under Assumption 3 if the data set is sampled from some density w.r.t. the Lebesgue measure, then with probability 1,

\[ \alpha := \min_{|\xi| = 1, \xi \perp \hat{u}, \xi \in S} \max_{z_i \in S} \langle \xi, z_i \rangle > 0. \]

Let \( \hat{u}^\perp \) denote the orthogonal complement of \( \text{span}(\hat{u}) \), and let \( \Pi_{\perp} \) denote the projection onto \( \hat{u}^\perp \). We prove that if \( \|\Pi_{\perp} w\| \) is large enough, gradient flow starting from \( w \) will tend to decrease \( \|\Pi_{\perp} w\| \).

Lemma 4. Under Assumption 3 for almost all data, \( \ell_{\exp} \) and \( \ell_{\log} \), and any \( w \in \mathbb{R}^d \), if \( \langle w, \hat{u} \rangle \geq 0 \) and \( \|\Pi_{\perp} w\| \) is larger than \( 1 + \ln(n)/\alpha \) for \( \ell_{\exp} \) or \( 2n/\alpha \) for \( \ell_{\log} \), then \( \langle \Pi_{\perp} w, \nabla R(w) \rangle \geq 0 \).

With Lemma 3 and Lemma 4 in hand, we can prove Theorem 2. Let \( \Pi_{\perp} W_1 \) denote the projection of rows of \( W_1 \) onto \( \hat{u}^\perp \). Notice that

\[ \Pi_{\perp} w_{\text{prod}} = (W_L \cdots W_2 (\Pi_{\perp} W_1))^T \quad \text{and} \quad \frac{d}{dt} \|\Pi_{\perp} W_1\|_F^2 = -2 \langle \Pi_{\perp} w_{\text{prod}}, \nabla R(w_{\text{prod}}) \rangle. \]
If \( \|\Pi_{t} W_{1}\|_{F} \) is large compared with \( \|W_{1}\|_{F} \), since layers become aligned, \( \|\Pi_{t} w_{\text{prod}}\| \) will also be large, and then Lemma 4 implies that \( \|\Pi_{t} W_{1}\|_{F} \) will not increase. At the same time, \( \|W_{1}\|_{F} \to \infty \), and thus for large enough \( t \), \( \|\Pi_{t} W_{1}\|_{F} \) must be very small compared with \( \|W_{1}\|_{F} \). Many details need to be handled to make this intuition exact; the proof is given in Appendix B.

\section{Results for Gradient Descent}

One key property of gradient flow which is used in the previous proofs is that it never increases the risk, which is not necessarily true for gradient descent. However, for smooth losses (i.e., with Lipschitz continuous derivatives), we can design some decaying step sizes, with which gradient descent never increases the risk, and basically the same results hold as in the gradient flow case. Deferred proofs are given in Appendix C.

We make the following additional assumption on the loss, which is satisfied by the logistic loss \( \ell_{\text{log}} \).

\textbf{Assumption 4.} \( \ell' \) is \( \beta \)-Lipschitz (i.e., \( \ell \) is \( \beta \)-smooth), and \( |\ell'| \leq G \) (i.e., \( \ell \) is \( G \)-Lipschitz).

Under Assumption 4, the risk is also a smooth function of \( W \), if all layers are bounded.

\textbf{Lemma 5.} Suppose \( \ell \) is \( \beta \)-smooth. If \( R \geq 1 \), then \( \beta(R) = 2L^{2}R^{2L-2}(\beta + G) \), and \( \mathcal{R}(W) \) is a \( \beta(R) \)-smooth function on the set \( B(R) = \{ W \|W_{k}\|_{F} \leq R, 1 \leq k \leq L \} \).

Smoothness ensures that for any \( W, V \in B(R), \mathcal{R}(W) - \mathcal{R}(V) \leq \langle \nabla \mathcal{R}(V), W - V \rangle \). (see Bubeck et al. (2015) Lemma 3.4). In particular, if we choose some \( R \) and set a constant step size \( \eta_{t} = 1/\beta(R) \), then, as long as \( W(t+1) \) and \( W(t) \) are both in \( B(R) \),

\[
\mathcal{R}(W(t+1)) - \mathcal{R}(W(t)) \leq \langle \nabla \mathcal{R}(W(t)), \eta_{t} \nabla \mathcal{R}(W(t)) \rangle + \frac{(\beta(R)\eta_{t})^{2}}{2} \|\nabla \mathcal{R}(W(t))\|^{2} = - \frac{1}{2\beta(R)} \|\nabla \mathcal{R}(W(t))\|^{2} - \frac{\eta_{t}}{2} \|\nabla \mathcal{R}(W(t))\|^{2}.
\]

In other words, the risk does not increase at this step. However, similar to gradient flow, the gradient descent iterate will eventually escape \( B(R) \), which may increase the risk.

\textbf{Lemma 6.} Under Assumption 2, 4, and 7, suppose gradient descent is run with a constant step size \( 1/\beta(R) \). Then there exists a time \( t \) when \( W(t) \notin B(R) \), in other words, \( \max_{1 \leq k \leq L} \|W_{k}(t)\|_{F} > R \).

Fortunately, this issue can be handled by adaptively increasing \( R \) and correspondingly decreasing the step sizes, formalized in the following assumption.

\textbf{Assumption 5.} The step size \( \eta_{t} = \min \{1/\beta(R_{t}), 1\} \), where \( R_{t} \) satisfies \( W(t) \in B(R_{t} - 1), \) and if \( W(t+1) \in B(R_{t} - 1), R_{t+1} = R_{t} \).

Assumption 5 can be satisfied by a line search, which ensures that the gradient descent update is not too aggressive and the boundary \( R \) is increased properly.

With the additional Assumptions 4 and 5, exactly the same theorems can be proved for gradient descent. We restate them briefly here.

\textbf{Theorem 3.} Under Assumptions 2, 4, 5 and 7, gradient descent satisfies

- \( \lim_{t \to \infty} \mathcal{R}(W(t)) = 0. \)
- For any \( 1 \leq k \leq L \), \( \lim_{t \to \infty} \|W_{k}(t)\|_{F} = \infty. \)
- \( \lim_{t \to \infty} \langle u_{\text{prod}(t)/\prod_{k=1}^{t} \|W_{k}(t)\|_{F}}, v_{1}(t) \rangle = 1, \) where \( v_{1}(t) \) is the first right singular vector of \( W_{1}(t) \).

\textbf{Theorem 4.} Under Assumptions 2, 5 and Assumption 8, for the logistic loss \( \ell_{\text{log}} \) and almost all data, \( \lim_{t \to \infty} \langle v_{1}(t), \bar{u} \rangle = 1 \), and \( \lim_{t \to \infty} \|u_{\text{prod}(t)/\prod_{k=1}^{t} \|W_{k}(t)\|_{F}}\|_{F} = \bar{u} \).

Proofs of Theorem 3 and 4 are given in Appendix C and are basically the same as the gradient flow proofs. The key difference is that an error of \( \sum_{t=0}^{\infty} \eta_{t}^{2} \|\nabla \mathcal{R}(W(t))\|^{2} \) will be introduced in many
Figure 2: Risk and alignment of dense layers (the ratio $\|W_i\|_2/\|W_i\|_F$) of (nonlinear!) AlexNet on CIFAR-10. Figure 2a uses default PyTorch initialization, while Figure 2b forces initial Frobenius norms to be equal amongst dense layers.

parts of the proof. However, it is bounded in light of eq. (4):

$$\sum_{t=0}^{\infty} \eta_t^2 \left\| \nabla R(W(t)) \right\|^2 \leq \sum_{t=0}^{\infty} \eta_t \left\| \nabla R(W(t)) \right\|^2 \leq 2R(W(0)) .$$

Since all weight matrices go to infinity, such a bounded error does not matter asymptotically, and thus proofs still go through.

4 SUMMARY AND FUTURE DIRECTIONS

This paper rigorously proves that, for deep linear networks on linearly separable data, gradient flow and gradient descent minimize the risk to 0, align adjacent weight matrices, and align the first right singular vector of the first layer to the maximum margin solution determined by the data. There are many potential future directions; a few are as follows.

Convergence rates and practical step sizes. This paper only proves asymptotic convergence, moreover for adaptive step sizes depending on the current weight matrix norms. A refined analysis with rates for practical step sizes (e.g., constant step sizes) would allow the algorithm to be compared to other methods which also globally optimize this objective, would suggest ways to improve step sizes and initialization, and ideally even exhibit a sensitivity to the network architecture and suggest how it could be improved.

Nonseparable data and nonlinear networks. Real-world data is generally not linearly separable, but nonlinear deep networks can reliably decrease the risk to 0, even with random labels (Zhang et al., 2017). This seems to suggest that a nonlinear notion of separability is at play; is there some way to adapt the present analysis?

The present analysis is crucially tied to the alignment of weight matrices: alignment and risk are analyzed simultaneously. Motivated by this, consider a preliminary experiment, presented in Figure 2, where stochastic gradient descent was used to minimize the risk of a standard AlexNet on CIFAR-10 (Krizhevsky et al., 2012; Krizhevsky & Hinton, 2009).

Even though there are ReLUs, max-pooling layers, and convolutional layers, the alignment phenomenon is occurring in a reduced form on the dense layers (the last three layers of the network). Specifically, despite these weight matrices having shape (1024, 4096), (4096, 4096), and (4096, 10) the key alignment ratios $\|W_i\|_2/\|W_i\|_F$ are much larger than their respective lower bounds $(1024^{-1/2}, 4096^{-1/2}, 10^{-1/2})$. Two initializations were tried: default PyTorch initialization, and a Gaussian initialization forcing all initial Frobenius norms to be just 4, which is suggested by the norm preservation property in the analysis and removes noise in the weights.
REFERENCES


A Regarding Assumption

Suppose $W_1(0) = 0$ while $W_L(0) \cdots W_2(0) \neq 0$. First of all, $W_L(0) \cdots W_1(0) = 0$ and thus $R(W(0)) = R(0)$. Moreover,

$$\langle \nabla R(w_{\text{prod}}(0)), \bar{u} \rangle = \frac{1}{n} \sum_{i=1}^{n} \ell'(0)(z_i, \bar{u}) \leq \ell'(0) \gamma < 0,$$

which implies $\nabla R(w_{\text{prod}}(0)) \neq 0$ and $\partial R/\partial W_1 = (W_L(0) \cdots W_2(0))^\top \nabla R(w_{\text{prod}}(0))^\top \neq 0$.

On the other hand, if $R(W(0)) > R(0)$, gradient flow/descent may never minimize the risk to 0. For example, suppose the network has two layers, and both the input and output have dimension 1; the network just computes the dot product of two vectors $w_1$ and $w_2$. Consider minimizing $R(w_1, w_2) = \exp(-\langle w_1, w_2 \rangle)$. If $w_1(0) = -w_2(0) \neq 0$, then $R(w_1(0), w_2(0)) = \exp\left(||w_1||^2\right) > \exp(0)$. It is easy to verify that for any $t$, $w_1(t) = -w_2(t)$, and $R(w_1(t), w_2(t)) \geq \exp(0) > 0$.

B Omitted proofs from Section

Proof of Lemma 1 Fix an arbitrary $R > 0$. If the claim is not true, then for any $\epsilon > 0$, there exists some $t \geq 1$ such that $||W_k||_F \leq R$ for all $k$ while $||\partial R/\partial W_1||^2 \leq \epsilon^2$, which means

$$\left\| \frac{\partial R}{\partial W_1} \right\|_F^2 = \left\| W_2^\top \cdots W_L^\top \nabla R(w_{\text{prod}})^\top \right\|_F^2 = \left\| W_L \cdots W_2 \right\|^2 \left\| \nabla R(w_{\text{prod}}) \right\|_F^2 \leq \epsilon^2.$$

Since $\|w_{\text{prod}}\| \leq RL$, we have

$$\langle \nabla R(w_{\text{prod}}), \bar{u} \rangle = \frac{1}{n} \sum_{i=1}^{n} \ell' \left(||w_{\text{prod}}, z_i||\right)(z_i, \bar{u}) \leq \frac{1}{n} \sum_{i=1}^{n} \ell' \left(||w_{\text{prod}}, z_i||\right) \gamma \leq -M \gamma,$$

where $-M = \max_{-R \leq x \leq R} \ell'(x)$. Since $\ell'$ is continuous and the domain is bounded, the maximum is attained and negative, and thus $M > 0$. Therefore $\left\| \nabla R(w_{\text{prod}}) \right\| \geq M \gamma$, and thus $\|W_L \cdots W_2\| \leq \epsilon / M \gamma$. Since $||W_1||_F \leq R$, we further have $\|w_{\text{prod}}\| \leq \epsilon R / M \gamma$. In other words, after $t = 1$, $\|w_{\text{prod}}\|$ may be arbitrarily small, which implies $R(w_{\text{prod}})$ can be arbitrarily close to $R(0)$.

On the other hand, by Assumption 2, $dR(W)/dt = -||\nabla R(W)||^2 < 0$ at $t = 0$. This implies that $R(W(1)) < R(W(0))$, and for any $t \geq 1$, $R(W(t)) \leq R(W(1)) < R(W(0)) \leq R(0)$, which is a contradiction.

Since the risk is always positive, we have

$$R(W(0)) \geq \int_{t=0}^{\infty} \sum_{k=1}^{L} \left\| \frac{\partial R}{\partial W_k} \right\|_F^2 dt$$

$$\geq \int_{t=0}^{\infty} \left\| \frac{\partial R}{\partial W_1} \right\|_F^2 dt$$

$$\geq \int_{t=1}^{\infty} \left\| \frac{\partial R}{\partial W_1} \right\|_F^2 \left[ \max_{1 \leq k \leq L} ||W_k||_F \leq R \right] dt$$

$$\geq \int_{t=1}^{\infty} \left\| \frac{\partial R}{\partial W_1} \right\|_F^2 \left[ \max_{1 \leq k \leq L} ||W_k||_F \leq R \right] dt$$

$$\geq \epsilon(R)^2 \int_{t=1}^{\infty} \left[ \max_{1 \leq k \leq L} ||W_k||_F \leq R \right] dt,$$

which implies gradient flow only spends a finite amount of time in $\{W | \max_{1 \leq k \leq L} ||W_k||_F \leq R \}$. This directly implies that $\max_{1 \leq k \leq L} ||W_k||_F$ is unbounded. \square
Proof of Lemma 2. The first claim is true for \( k = L \) since \( W_L \) is a row vector. For any \( 1 \leq k < L \), recall that \cite{arora2018structured,du2018gradient} give the following relation:
\[
W_{k+1}^T(0)W_k(0) = W_k(t)W_k^T(t) - W_k(0)W_k^T(0). \tag{5}
\]

Let \( A_{k,k+1} = W_k(0)W_k^T(0) - W_{k+1}^T(0)W_{k+1}(0) \). By eq. (5) and the definition of singular vectors and singular values, we have
\[
\sigma_k^2 \geq v_{k+1}^T W_k W_k^T v_{k+1} \\
= v_{k+1}^T W_{k+1} W_{k+1} v_{k+1} + v_{k+1}^T A_{k,k+1} v_{k+1} \\
= \sigma_{k+1}^2 + v_{k+1}^T A_{k,k+1} v_{k+1} \\
\geq \sigma_{k+1}^2 - \| A_{k,k+1} \|_2. \tag{6}
\]

Moreover, by taking the trace on both sides of eq. (5), we have
\[
\| W_k \|_F^2 = \text{tr} (W_k W_k^T) = \text{tr} (W_{k+1} W_{k+1}^T) - \text{tr} (W_{k+1}^T(0)W_{k+1}(0)) \\
= \| W_{k+1} \|_2^2 + \| W_k(0) \|_F^2 - \| W_{k+1}(0) \|_F^2. \tag{7}
\]

Summing eq. (6) and eq. (7) from \( k \) to \( L - 1 \), we get
\[
\| W_k \|_F^2 - \| W_k \|_2^2 \leq \| W_k(0) \|_F^2 - \| W_L(0) \|_F^2 + \sum_{k' = k}^{L-1} \| A_{k',k+1} \|_2 \leq D. \tag{8}
\]

Next we prove that singular vectors get aligned. Consider \( u_k^T W_{k+1}^T W_{k+1} u_k \). On one hand, similar to eq. (6), we can get that
\[
u_k^T W_{k+1}^T W_{k+1} u_k = u_k^T W_k W_k^T u_k - u_k^T W_k(0) W_k^T(0) u_k + u_k^T W_{k+1}^T(0) W_{k+1}(0) u_k \\
\geq u_k^T W_k W_k^T u_k - u_k^T W_k(0) W_k^T(0) u_k \\
\geq \sigma_k^2 - \| W_k(0) \|_2^2. \tag{9}
\]

On the other hand, it follows from the definition of singular vectors and eq. (9) that
\[
u_k^T W_{k+1}^T W_{k+1} u_k = \langle u_k, v_{k+1} \rangle^2 \sigma_{k+1}^2 + u_k^T (W_{k+1} W_{k+1} - v_{k+1} \sigma_{k+1}^2 v_{k+1}^T) u_k \\
\leq \langle u_k, v_{k+1} \rangle^2 \sigma_{k+1}^2 + \| W_{k+1} \|_2^2 - \| W_{k+1} \|_2^2 \\
\leq \langle u_k, v_{k+1} \rangle^2 \sigma_{k+1}^2 + D. \tag{10}
\]

Combining eq. (9) and eq. (10), we get
\[
\sigma_k^2 \leq \langle u_k, v_{k+1} \rangle^2 \sigma_{k+1}^2 + D + \| W_k(0) \|_2^2. \tag{11}
\]

Similar to eq. (9), we can get
\[
\sigma_k^2 \geq v_{k+1}^T W_k W_k^T v_{k+1} \geq \sigma_{k+1}^2 - \| W_{k+1}(0) \|_2^2.
\]

Therefore
\[
\frac{\sigma_k^2}{\sigma_{k+1}^2} \geq 1 - \frac{\| W_{k+1}(0) \|_2^2}{\sigma_{k+1}^2}. \tag{12}
\]

Combining eq. (11) and eq. (12), we finally get
\[
\langle u_k, v_{k+1} \rangle^2 \geq 1 - \frac{D + \| W_k(0) \|_2^2 + \| W_{k+1}(0) \|_2^2}{\sigma_{k+1}^2}.
\]

Regarding the last claim, first recall that since the difference between the squares of Frobenius norms of any two layers is a constant, \( \max_{1 \leq k \leq L} \| W_k \|_F \to \infty \) implies \( \| W_k \|_F \to \infty \) for any \( k \). We further have the following.

- Since \( \| W_k \|_F^2 - \| W_k \|_2^2 \leq D \), \( \| W_k \|_2 \to \infty \) for any \( k \), and \( W_k/\| W_k \|_F \to u_k v_k^T \).
• Since $\|W_k\|_2 \to \infty$, $\|u_k, v_{k+1}\| \to 1$.

As a result,

$$
\left| \left\langle \frac{w_{\text{prod}}}{\prod_{k=1}^L \|W_k\|_F}, v_1 \right\rangle \right| = \left| \left\langle \prod_{k=1}^L \frac{W_k}{\|W_k\|_F}, v_1 \right\rangle \right| \\
\to \left| \left\langle \prod_{k=1}^L u_kv_k^\top, v_1 \right\rangle \right| \\
\to 1.
$$

Proof of Theorem 1] Suppose for some $\epsilon > 0$, $\mathcal{R}(W) \geq \epsilon$ for any $t$. Then there exists some $1 \leq j \leq n$ such that $\ell' \left( \langle w_{\text{prod}}, z_j \rangle \right) \geq \epsilon$, and thus $\langle w_{\text{prod}}, z_j \rangle \leq \ell^{-1}(\epsilon)$. On the other hand, since $\mathcal{R}(W) \leq \mathcal{R}(0) = \ell(0)$, $\ell' \left( \langle w_{\text{prod}}, z_j \rangle \right) \leq n\ell(0)$, and thus $\langle w_{\text{prod}}, z_j \rangle \geq \ell^{-1} \left( n\ell(0) \right)$. Let $-M = \max_{x \leq t \leq \ell^{-1}(\epsilon/n)} \ell'(x) < 0$, we have for any $t$,

$$
\langle \nabla \mathcal{R}(w_{\text{prod}}), \bar{u} \rangle = \frac{1}{n} \sum_{i=1}^n \ell' \left( \langle w_{\text{prod}}, z_i \rangle \right) \langle z_i, \bar{u} \rangle \\
\leq \frac{1}{n} \sum_{i=1}^n \ell' \left( \langle w_{\text{prod}}, z_i \rangle \right) \gamma \\
\leq \frac{1}{n} \ell' \left( \langle w_{\text{prod}}, z_j \rangle \right) \gamma \\
\leq -\frac{M\gamma}{n} < 0,
$$

and thus $\|\nabla \mathcal{R}(w_{\text{prod}})\| \geq M\gamma/n$.

Similar to the proof of Lemma 2, we can show that if $\|W_k\|_F \to \infty$,

$$
\left| \left\langle \left( W_L \cdots W_2 \right)^\top \frac{\partial \mathcal{R}}{\partial W_k}, v_2 \right\rangle \right| \to 1.
$$

In other words, there exists some $C > 0$, such that when $\min_{1 \leq k \leq L} \|W_k\|_F > C$, $\|W_L \cdots W_2\| \geq \|W_k\|_F \cdots \|W_2\|_F / 2 > C^L / 2$.

Lemma 1 shows that gradient flow spends a finite amount of time in $\{W \mid \max_{1 \leq k \leq L} \|W_k\|_F \leq R\}$ for any $R > 0$. Since the difference between the squares of Frobenius norms of any two layers is a constant, gradient flow also spends a finite amount of time in $\{W \mid \min_{1 \leq k \leq L} \|W_k\|_F \leq C\}$. Now we have

$$
\mathcal{R}(W(0)) \geq \int_0^\infty \sum_{k=1}^L \left| \left\langle \frac{\partial \mathcal{R}}{\partial W_k}, v \right\rangle \right|^2 \, dt \\
\geq \int_0^\infty \left| \left\langle \frac{\partial \mathcal{R}}{\partial W_1}, v \right\rangle \right|^2 \, dt \\
= \int_0^\infty \|W_L \cdots W_2\|^2 \|\nabla \mathcal{R}(w_{\text{prod}})\|^2 \, dt \\
\geq \int_0^\infty \|W_L \cdots W_2\|^2 \|\nabla \mathcal{R}(w_{\text{prod}})\|^2 \mathbb{1} \left[ W \mid \min_{1 \leq k \leq L} \|W_k\|_F > C \right] \, dt \\
\geq \left( \frac{M\gamma}{n} \right)^2 \left( \frac{C^L}{2} \right)^2 \int_0^\infty \mathbb{1} \left[ W \mid \min_{1 \leq k \leq L} \|W_k\|_F > C \right] \, dt \\
= \infty,
$$

Under review as a conference paper at ICLR 2019
which is a contradiction. Therefore $R(e) \to 0$. This further implies $\|W_k\|_F \to \infty$, since $R(W)$ has no finite optimum. Finally, invoking Lemma 2 proves the final claim of Theorem 1.

Proof of Lemma 3. [Soudry et al., 2017] Lemma 12 proves that, with probability 1, there are at most $d$ support vectors, and moreover, the $i$-th support vector $z_i$ has a positive dual variable $\alpha_i$, such that $\sum_{i \in S} \alpha_i z_i = \bar{u}$.

Suppose there exists some $\xi \perp \bar{u}$, such that $\max_{i \in S} \langle \xi, z_i \rangle \leq 0$. Since

$$
\sum_{i \in S} \alpha_i \langle \xi, z_i \rangle = \left( \sum_{i \in S} \alpha_i \right) \langle \xi, \bar{u} \rangle = 0,
$$

we actually have $\langle \xi, z_i \rangle = 0$ for all $i \in S$. This is impossible under Assumption 3, since the support vectors span the whole space.

Proof of Lemma 4. For the sake of presentation, we leave out the subscript in $z_i$ and denote a data point by $z$ generally. For any data point $z$ and predictor $w$, let $z_\perp$ and $w_\perp$ denote their projection onto $\bar{u}^\perp$. Let $z' \in \arg\max_{i \in S} \langle -w_\perp, z \rangle$, and thus $\langle -w_\perp, z' \rangle \geq \alpha \|w_\perp\|$.

For $\ell_{exp}$, we have

$$
\langle w_\perp, \nabla R(w) \rangle = \frac{1}{n} \sum_z \left[ -\exp \left( -\langle w, z \rangle \right) \right] \langle w_\perp, z_\perp \rangle
$$

$$
= \frac{1}{n} \sum_z \left[ \exp \left( -\langle w, z \rangle \right) \right] \langle -w_\perp, z_\perp \rangle
$$

$$
\geq \frac{1}{n} \exp \left( -\langle w, z' \rangle \right) \langle -w_\perp, z'_\perp \rangle + \frac{1}{n} \sum_{(z_\perp, w_\perp) \geq 0} \exp \left( -\langle w, z \rangle \right) \langle -w_\perp, z_\perp \rangle. \quad (13)
$$

The first part can be lower bounded as below (recall that $\langle -w_\perp, z'_\perp \rangle = \langle -w_\perp, z_\perp \rangle \geq \alpha \|w_\perp\|$)

$$
\frac{1}{n} \exp \left( -\langle w, z' \rangle \right) \langle -w_\perp, z'_\perp \rangle = \frac{1}{n} \exp \left( -\langle w, \gamma \bar{u} \rangle \right) \exp \left( -\langle w, z'_\perp \rangle \right) \langle -w_\perp, z'_\perp \rangle
$$

$$
\geq \frac{1}{n} \exp \left( -\langle w, \gamma \bar{u} \rangle \right) \exp \left( \alpha \|w_\perp\| \right) \alpha \|w_\perp\|. \quad (14)
$$

To bound the second part, first notice that since we assume $\langle w, \bar{u} \rangle \geq 0$, for any $z$,

$$
\langle w, z - \gamma \bar{u} \rangle = \langle w, z_\perp \rangle + \langle w, z - \gamma \bar{u} - z_\perp \rangle \geq \langle w, z_\perp \rangle = \langle w_\perp, z_\perp \rangle. \quad (15)
$$

The reason is that every data point has margin at least $\gamma$, and thus $z - \gamma \bar{u} - z_\perp = c\bar{u}$ for some $c \geq 0$. Using eq. (15), we can bound the second part of eq. (13).

$$
\sum_{(z_\perp, w_\perp) \geq 0} \frac{1}{n} \exp \left( -\langle w, z \rangle \right) \langle -w_\perp, z_\perp \rangle
$$

$$
= \sum_{(z_\perp, w_\perp) \geq 0} \frac{1}{n} \exp \left( -\langle w, \gamma \bar{u} \rangle \right) \exp \left( -\langle w, z - \gamma \bar{u} \rangle \right) \langle -w_\perp, z_\perp \rangle
$$

$$
\geq \sum_{(z_\perp, w_\perp) \geq 0} \frac{1}{n} \exp \left( -\langle w, \gamma \bar{u} \rangle \right) \exp \left( -\langle w_\perp, z_\perp \rangle \right) \langle -w_\perp, z_\perp \rangle
$$

$$
\geq \sum_{(z_\perp, w_\perp) \geq 0} \frac{1}{n} \exp \left( -\langle w, \gamma \bar{u} \rangle \right) \left( -\frac{1}{e} \right)
$$

$$
\geq \exp \left( -\langle w, \gamma \bar{u} \rangle \right) \left( -\frac{1}{e} \right). \quad (16)
$$

On the third line eq. (15) is applied. The fourth line applies the property that $f(x) = -xe^{-x} \geq -1/e$ when $x \geq 0$. 


Combining eq. (13), eq. (14) and eq. (16), we get
\[
\langle w_\perp, \nabla R(w) \rangle \geq \frac{1}{n} \exp \left( -\frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) \right) \left( \frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) - \frac{1}{\epsilon} \right).
\]
As long as \( \|w_\perp\| \geq (1 + \ln(n))/\alpha, \langle w_\perp, \nabla R(w) \rangle \geq 0. \)

For \( \ell_{\log} \), similar to eq. (13), we have
\[
\langle w_\perp, \nabla R(w) \rangle \geq \frac{1}{n} \exp \left( -\frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) \right) \left( \frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) - \frac{1}{\epsilon} \right).
\]
The second part of eq. (17) can be bounded again by eq. (16). To bound the first part of eq. (17), first notice that (recall \( \langle w, \bar{u} \rangle \geq 0 \))
\[
\exp \left( -\frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) \right) \leq \exp \left( -\frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) \right).
\]
Using eq. (18), and recall that \( \langle w_\perp, z'_1 \rangle = \langle w, z' \rangle \geq 0 \), we can bound the first part of eq. (17) as below.
\[
\frac{1}{n} \exp \left( -\frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) \right) \left( \frac{1}{n} \exp \left( \frac{\alpha \|w_\perp\|}{\epsilon} \right) - \frac{1}{\epsilon} \right).
\]
Combining eq. (17), eq. (19) and eq. (16), we get
\[
\langle w_\perp, \nabla R(w) \rangle \geq \exp \left( -\frac{1}{n} \alpha \|w_\perp\| - \frac{1}{\epsilon} \right) \left( \frac{1}{n} \alpha \|w_\perp\| - \frac{1}{\epsilon} \right).
\]
As long as \( \|w_\perp\| \geq 2n/\epsilon \alpha, \langle w_\perp, \nabla R(w) \rangle \geq 0. \)

\textbf{Proof of Theorem 2} Recall that
\[
\frac{dW_t}{dt} = -\frac{\partial R}{\partial W_t} = -W_t^T \cdots W_2^T \nabla R(w_{\text{prod}})^T,
\]
and thus
\[
\frac{d\|W_t\|_F^2}{dt} = \left( W_t, \frac{dW_t}{dt} \right) = -2\langle w_{\text{prod}}, \nabla R(w_{\text{prod}}) \rangle.
\]
Let \( \Pi_{\bar{u}} \) denote the projection onto span(\( \bar{u} \)), and let \( \Pi_\perp \) denote the projection onto \( \bar{u}^\perp \). Also let \( \Pi_{\bar{u}} W_1 \) and \( \Pi_\perp W_1 \) denote the projection of rows of \( W_1 \) onto span(\( \bar{u} \)) and \( \bar{u}^\perp \), respectively. Notice that
\[
\Pi_{\bar{u}} w_{\text{prod}} = (W_L \cdots W_2 (\Pi_{\bar{u}} W_1))^T, \quad \text{and} \quad \Pi_\perp w_{\text{prod}} = (W_L \cdots W_2 (\Pi_\perp W_1))^T.
\]
We further have
\[
\frac{d\|\Pi_{\perp} W_1\|_F^2}{dt} = -2\langle \Pi_{\perp} w_{\text{prod}}, \nabla R(w_{\text{prod}}) \rangle.
\]
Let \( W_1 = u_1 \sigma_1 v_1^T + S \). We have \( \|S\|_2 \leq \sigma_{1,2} \leq \sqrt{\sigma_{1,2}^2 - \|W_1\|_F^2} \leq \sqrt{D} \), where \( \sigma_{1,2} \) is the second singular value of \( W_1 \) and \( D \) is the constant introduced in Lemma 2. Then
\[
\Pi_{\perp} W_1 = u_1 \sigma_1 (\Pi_{\perp} v_1)^T + \Pi_{\perp} S,
\]
and

$$
\|\Pi_{\perp}W_1\|_F \leq \left\|u_1\sigma_1 (\Pi_{\perp}v_1)^{T}\right\|_F + \|\Pi_{\perp}S\|_F = \sigma_1\|\Pi_{\perp}v_1\| + \|\Pi_{\perp}S\|_F \leq \sigma_1\|\Pi_{\perp}v_1\| + \sqrt{dD}.
$$

It follows that

$$
\|\Pi_{\perp}v_1\| \geq \frac{\|\Pi_{\perp}W_1\|_F}{\sigma_1} - \frac{\sqrt{dD}}{\sigma_1} \geq \frac{\|\Pi_{\perp}W_1\|_F}{\|W_1\|_F} - \frac{\sqrt{dD}}{\|W_1\|_2}.
$$

(22)

Fix an arbitrary $\epsilon > 0$. By Theorem 1, we can find some $t_0$ large enough such that for any $t \geq t_0$:

1. $\sqrt{dD}/\|W_1\|_2 \leq \epsilon/3$.
2. $\|w_{\text{prod}}/\|W_L\|_F \cdots \|W_1\|_F - v_1\| \leq \epsilon/3$, or $\|w_{\text{prod}}/\|W_L\|_F \cdots \|W_1\|_F + v_1\| \leq \epsilon/3$.
3. $\|W_L\|_F \cdots \|W_1\|_F \geq K/\epsilon$, where $K$ is the threshold given in Lemma 3. Since also $\|w_{\text{prod}}, \bar{u}\| \geq 0$, we can apply Lemma 4 and get that $\langle \Pi_{\perp}w_{\text{prod}}, \nabla R(w_{\text{prod}}) \rangle \geq 0$. In light of eq. (21), $d\|W_1\|_2^2/ dt \leq 0$.

On the other hand, since after $t \geq t_0$, $\langle w_{\text{prod}}, z_i \rangle \geq 0$, we have $d\|W_1\|_F/ dt \geq 0$ by eq. (20). Therefore $\|\Pi_{\perp}W_1\|_F/\|W_1\|_F$ will not increase, and since $\|W_1\|_F \rightarrow \infty$, it will eventually drop below $\epsilon$, and will never exceed $\epsilon$ again. Therefore,

$$
\limsup_{t \rightarrow \infty} \frac{\|\Pi_{\perp}W_1\|_F}{\|W_1\|_F} \leq \epsilon.
$$

Since $\epsilon$ is arbitrary, we have

$$
\limsup_{t \rightarrow \infty} \frac{\|\Pi_{\perp}W_1\|_F}{\|W_1\|_F} = 0,
$$

and thus $\lim_{t \rightarrow \infty} \langle v_1, \bar{u} \rangle = 1$. An application of Theorem 1 gives the other part of Theorem 2.

\[\square\]

C Omitted proofs from Section 3

\textit{Proof of Lemma 3} Given $W, V \in B(R)$, we need to show that $\|\nabla R(W) - \nabla R(V)\| \leq \beta(R)\|W - V\|$ for some $\beta(R)$.

Consider $k = 1$ first. Let $w = (W_L \cdots W_1)^T$, and $v = (V_L \cdots V_1)^T$. Since $|\ell'| \leq G$, $\|\nabla R(w)\|, \|\nabla R(v)\| \leq G$. We have

$$
\left\| \frac{\partial R}{\partial W_1} - \frac{\partial R}{\partial V_1} \right\| = \left\|W_2^T \cdots W_L^T \nabla R(w) - V_2^T \cdots V_L^T \nabla R(v)\right\|
\leq \left\|W_2^T \cdots W_L^T \nabla R(w) - V_2^T W_3^T \cdots W_L^T \nabla R(w)\right\|
+ \left\|V_2^T W_3^T \cdots W_L^T \nabla R(w) - V_2^T \cdots V_L^T \nabla R(v)\right\|
\leq R^{k-2}G\|W_2 - V_2\|
+ \left\|V_2^T W_3^T \cdots W_L^T \nabla R(w) - V_2^T \cdots V_L^T \nabla R(v)\right\|
\leq R^{k-2}G\|W - V\|
+ \left\|V_2^T W_3^T \cdots W_L^T \nabla R(w) - V_2^T \cdots V_L^T \nabla R(v)\right\|. (23)
$$

14
Proof. Let $W_t$ be the solution to the gradient flow of $\mathcal{R}(W)$. The following Lemma 7 is similar to Lemma 1.

Next we prove Theorem 3 and 4. The proofs depend on several lemmas which are similar to the gradient flow ones. The following Lemma 7 is similar to Lemma 1.

Recall that if $W_t$, $W_{t+1} \in B(R)$ and $\eta_t = 1/\beta(R)$,

$$\mathcal{R}(W(t+1)) - \mathcal{R}(W(t)) \leq \langle \nabla \mathcal{R}(W(t)), -\eta_t \nabla \mathcal{R}(W(t)) \rangle + \frac{\beta(R)\eta_t^2}{2} \left\| \nabla \mathcal{R}(W(t)) \right\|^2$$

$$= -\frac{1}{2\beta(R)} \left\| \nabla \mathcal{R}(W(t)) \right\|^2$$

$$= -\frac{\eta_t}{2} \left\| \nabla \mathcal{R}(W(t)) \right\|^2. \quad (26)$$

Suppose $W(t) \in B(R)$ for all $t$. By Assumption 2 and eq. (26),

$$\mathcal{R}(W(1)) \leq \mathcal{R}(W(0)) - \frac{1}{2\beta(R)} \left\| \nabla \mathcal{R}(W(0)) \right\|^2 < \mathcal{R}(W(0)).$$

By eq. (26), gradient descent never increases the risk, and thus for all $t \geq 1$, $\mathcal{R}(W(t)) \leq \mathcal{R}(W(1)) < \mathcal{R}(W(0))$. In exactly the same way as in the proof of Lemma 1, one can show that there exists some constant $\epsilon(R) > 0$, so that $\left\| \partial \mathcal{R}/\partial W_1(t) \right\|_F \geq \epsilon(R)$ for all $t$. Invoking eq. (26) again, we will get

$$\mathcal{R}(W(0)) \geq \sum_{t=0}^{\infty} \frac{1}{2\beta(R)} \epsilon(R)^2 = \infty,$$

which is a contradiction. Therefore $W(t)$ must go out of $B(R)$ at some time.

Next we prove Theorem 3 and 4. The proofs depend on several lemmas which are similar to the gradient flow ones. The following Lemma 7 is similar to Lemma 1.

Lemma 7. Under Assumption 7, gradient descent ensures that

- $\max_{1 \leq k \leq L} \left\| W_k(t) \right\|_F$ is unbounded.
- $\sum_{t=0}^{\infty} \eta_t = \infty.$
- For any $R > 0$, $\sum_{t: W(t) \in B(R)} \eta_t < \infty.$

Proof. By Assumption 5 we always have that $W(t) \in B(R_t)$. Since $\beta(R_t) = 2L^2R_t^{2L-2}(\beta+G) \geq R_t^{L-1}G$, we have for any $1 \leq k \leq L$,

$$\left\| W_k(t+1) \right\|_F \leq \left\| W_k(t) \right\|_F + \eta_t \left\| \frac{\partial \mathcal{R}}{\partial W_k(t)} \right\|_F$$

$$\leq \left\| W_k(t) \right\|_F + \frac{1}{\beta(R_t)} \left\| \frac{\partial \mathcal{R}}{\partial W_k(t)} \right\|_F$$

$$\leq \left\| W_k(t) \right\|_F + \frac{1}{\beta(R_t)} R_t^{L-1}G$$

$$\leq \left\| W_k(t) \right\|_F + 1. \quad (27)$$
Moreover, Lemma 6 shows that \( R_t \to \infty \). Since \( R_{t+1} = R_t \) as long as \( W(t + 1) \in B(R_t - 1) \), \( \max_{1 \leq k \leq L} \left\| W_k(t) \right\|_F \) is unbounded.

It then follows that for any \( t \), by Cauchy-Schwarz,
\[
\left( \sum_{\tau=0}^{t-1} \eta_{\tau} \right) \left( \sum_{\tau=0}^{t-1} \eta_{\tau} \left\| \nabla R (W(\tau)) \right\|^2 \right) \geq \left( \sum_{\tau=0}^{t-1} \eta_{\tau} \left\| \nabla R (W(\tau)) \right\| \right)^2 \to \infty.
\]
since by eq. (26),
\[
\sum_{\tau=0}^{t-1} \eta_{\tau} \left\| \nabla R (W(\tau)) \right\|^2 \leq 2R (W(0)) - 2R (W(t)) \leq 2R (W(0)),
\]
we have \( \sum_{t=0}^{\infty} \eta_t = \infty \).

Since under Assumptions 4 and 5 gradient descent never increases the risk, it can be shown in exactly the same as in the proof of Lemma 1 that, for \( W(t) \in B(R) \), \( \left\| \partial R / \partial W_k(t) \right\|_F \geq \epsilon(R) \) for some constant \( \epsilon(R) > 0 \). Invoking eq. (26) again, we get that \( \sum_{t: W(t) \in \delta(t)} \eta_t < \infty \). \( \square \)

The next lemma is an analogy to Lemma 2.

**Lemma 8.** Under Assumptions 4 and 5, the gradient descent iterates satisfy the following properties:

- For any \( 1 \leq k \leq L \), \( \left\| W_k(t) \right\|^2_F - \left\| W_k(0) \right\|^2_F \leq D + 2R (W(0)) \).
- For any \( 1 \leq k < L \), \( \langle v_{k+1} \cdot u_k \rangle^2 \geq 1 - D + 3R(W(0)) + \left\| W_{k+1} (0) \right\|^2_F + \left\| W_k(0) \right\|^2_F / \left\| W_{k+1} (0) \right\|^2_F \).
- Suppose \( \max_{1 \leq k \leq L} \left\| W_k(t) \right\|_F \to \infty \), then \( \left\| \prod_{k=1}^{t} W_k \right\|_F \cdot \left\| W_k(0) \right\|_F \cdot \left\| v_1 \right\|_F \to 1 \).

**Proof.** Recall that for any \( W \),
\[
W_k^T \frac{\partial R}{\partial W_k} = W_k^T \nabla R (w_{prod})^T W_1^T \cdots W_k^T = \frac{\partial R}{\partial W_k} W_k^T. \tag{28}
\]
For gradient descent iterates, summing eq. (28) from \( 0 \) to \( t - 1 \), we get
\[
W_k(t) - W_k(0) = W_k(t) - W_k(0) + \sum_{\tau=0}^{t-1} \eta_{\tau}^2 \left( \frac{\partial R}{\partial W_k (\tau)} \right) \left( \frac{\partial R}{\partial W_k (\tau)} \right)^T.
\tag{29}
\]
For any \( 1 \leq k \leq L \) and any \( t \), let
\[
P_k(t) = \sum_{\tau=0}^{t-1} \eta_{\tau}^2 \left( \frac{\partial R}{\partial W_k (\tau)} \right) \left( \frac{\partial R}{\partial W_k (\tau)} \right)^T,
\]
and
\[
Q_k(t) = \sum_{\tau=0}^{t-1} \eta_{\tau}^2 \left( \frac{\partial R}{\partial W_k (\tau)} \right) \left( \frac{\partial R}{\partial W_k (\tau)} \right)^T.
\]
We have \( \left\| P_k(t) \right\|_2 = \left\| Q_k(t) \right\|_2 \leq \text{tr} (Q_k(t)) = \text{tr} (P_k(t)) \). Moreover, invoking eq. (26),
\[
\sum_{k=1}^{L} \text{tr} (P_k(t)) = \sum_{k=1}^{L} \sum_{\tau=0}^{t-1} \eta_{\tau}^2 \left\| \nabla R (W(\tau)) \right\|_F^2
\leq \sum_{\tau=0}^{t-1} \eta_{\tau} \left\| \nabla R (W(\tau)) \right\|^2 \leq 2R (W(0)) - 2R (W(t)) \leq 2R (W(0)). \tag{30}
\]
Still let \( \sigma_k(t), u_k(t) \) and \( v_k(t) \) denote the first singular value, left singular vector and right singular vector of \( W_k(t) \). We can then proceed basically in the same way as in the proof of Lemma 2. For example, eq. (9) becomes
\[
\sigma_k(t) \geq \sigma_{k+1}^2(t) - \|A_{k+1}(t)\|_2^2 - \|P_k(t)\|_2^2 \geq \sigma_{k+1}^2(t) - \|A_{k+1}(t)\|_2^2 - \text{tr}(P_k(t)),
\]
while eq. (10) becomes
\[
\|W_k(t)\|_F^2 = \|W_{k+1}(t)\|_F^2 + \|W_k(0)\|_F^2 - \|W_{k+1}(0)\|_F^2 - \text{tr}(P_k(t)) + \text{tr}(Q_k(t)).
\]

Summing eq. (31) and eq. (32) from \( k = 1 \) to \( L - 1 \), and invoke eq. (30), we get
\[
\|W_k(t)\|_F^2 - \|W_k(0)\|_F^2 \leq D - \text{tr}(P_k(t)) + \text{tr}(Q_k(t)) + \sum_{k'=k}^{L-1} \text{tr}(P_{k'}(t)) \leq D + 2R(W(0)).
\]

To prove singular vectors get aligned, we can still proceed in nearly the same way as in the proof of Lemma 2. For example, eq. (6) becomes
\[
\|u_k^T W_{k+1} u_k\|_2^2 \geq \sigma_k^2 - \|W_k(0)\|_2^2 - \|Q_{k+1}(t)\|_2^2,
\]
while eq. (10) becomes
\[
\|u_k^T W_{k+1} u_k\|_2^2 \leq \langle u_k, v_{k+1} \rangle^2 + D + 2R(W(0)).
\]

Combining eq. (33) and eq. (34), we can get
\[
\sigma_k^2 \geq \langle u_k, v_{k+1} \rangle^2 + D + 2R(W(0)) + \|Q_{k+1}(t)\|_2^2 + \|W_k(0)\|_2^2.
\]

Similar to eq. (33), we can get
\[
\sigma_k^2 \geq \sigma_{k+1}^2 + \|W_k(0)\|_2^2 - \|P_k(t)\|_2^2,
\]
and thus eq. (12) becomes
\[
\frac{\sigma_k^2}{\sigma_{k+1}^2} \geq 1 - \frac{\|W_{k+1}(0)\|_2^2 + \|P_k(t)\|_2^2}{\|W_k(0)\|_2^2}.
\]

Combining eq. (35) and eq. (36), we get
\[
\langle u_k, v_{k+1} \rangle^2 \geq 1 - \frac{D + \|W_k(0)\|_2^2 + \|W_{k+1}(0)\|_2^2 + 3R(W(0))}{\sigma_{k+1}^2}.
\]

The final claim of Lemma 8 can be proved in exactly the same way as Lemma 2.

Proof of Theorem 3: Summing eq. (32), we know that for any two different layers \( j > k \),
\[
\|W_{j}(t)\|_F^2 - \|W_j(0)\|_F^2 = \|W_{j+1}(0)\|_F^2 - \|W_j(0)\|_F^2 - \text{tr}(P_{j}(t)) + \text{tr}(Q_{j}(t)).
\]

Recall eq. (30), we know that
\[
\left( \|W_k(t)\|_F^2 - \|W_j(t)\|_F^2 \right) - \left( \|W_k(0)\|_F^2 - \|W_j(0)\|_F^2 \right) \leq 2R(W(0)).
\]

In other words, the difference between the squares of Frobenius norms of any two layers is still bounded.

The proof then goes in the same way as the proof of Theorem 1. Suppose the risk is always above \( \epsilon > 0 \). Then there exists some \( c(\epsilon) > 0 \) such that \( \|\nabla R(w)\|_F \geq c(\epsilon) \). By Lemma 8, there exists some \( C \) such that if \( \min_{k \leq L} \|W_k(0)\|_F > C \), \( \|W_L(t) \cdots W_2(t)\| \geq C^L/2 \). By eq. (37) and Lemma 7, \( \sum_t \|W_k(t)\|_F \leq C \) for some constant \( \eta_t \). On the other hand, by Lemma 7, \( \sum_{t=0}^{\infty} \eta_t = \infty \), and thus \( \sum_t \|W_k(t)\|_F \geq C \). Therefore we have, by invoking eq. (36),
\[
2R(W(0)) \geq \sum_{t=0}^{\infty} \eta_t \left( \|R(W(t))\|_2^2 \right)
\]
\[
\geq \sum_{t=0}^{\infty} \eta_t \left( \frac{\nabla R(W(t))}{\partial W(t)} \right)
\]
\[
\geq c(\epsilon) \frac{C^L}{2} \sum_{t=0}^{\infty} \eta_t \geq c(\epsilon) \frac{C^L}{2}
\]
which is a contradiction. Therefore \( R(W(t)) \to 0 \), and since it has no finite optimum, \( \|W_k\|_F \to \infty \). The other results follow from Lemma 7.
Proof of Theorem 2: Recall that
\[
\frac{\partial R}{\partial W_1} = W_2^\top \cdots W_L^\top \nabla R(w_{\text{prod}}),
\]
and thus
\[
\|W_1(t + 1)\|_F^2 = \|W_1(t)\|_F^2 - 2\eta_t \left\langle \Pi_\perp W_1(t), \frac{\partial R}{\partial W_1(t)} \right\rangle + \eta_t^2 \left\| \frac{\partial R}{\partial W_1(t)} \right\|_F^2
\]
\[
= \|W_1(t)\|_F^2 - 2\eta_t \left\langle \Pi_\perp w_{\text{prod}}(t), \nabla R(w_{\text{prod}}(t)) \right\rangle + \eta_t^2 \left\| \frac{\partial R}{\partial W_1(t)} \right\|_F^2.
\]
If \( \langle w_{\text{prod}}, z_i \rangle \geq 0 \) for all \( i \), then \( \|W_1(t + 1)\|_F \geq \|W_1(t)\|_F \).

Also recall that \( \Pi_\perp W_1(t) \) denote the projection of rows of \( W_1(t) \) onto \( \bar{u}^\perp \), the orthogonal complement of span(\( \bar{u} \)). We have
\[
\|\Pi_\perp W_1(t + 1)\|_F^2 \leq \|\Pi_\perp W_1(t)\|_F^2 - 2\eta_t \left\langle \Pi_\perp W_1(t), \frac{\partial R}{\partial W_1(t)} \right\rangle + \eta_t^2 \left\| \frac{\partial R}{\partial W_1(t)} \right\|_F^2
\]
\[
= \|\Pi_\perp W_1(t)\|_F^2 - 2\eta_t \left\langle \Pi_\perp w_{\text{prod}}(t), \nabla R(w_{\text{prod}}(t)) \right\rangle + \eta_t^2 \left\| \frac{\partial R}{\partial W_1(t)} \right\|_F^2.
\]
(38)

Invoking eq. (26) again gives
\[
\eta_t^2 \left\| \frac{\partial R}{\partial W_1(t)} \right\|_F^2 \leq \eta_t \left\| \nabla R(W(t)) \right\|_F^2 \leq 2 \left( R(W(t)) - R(W(t + 1)) \right).
\]
(39)

The proof then goes in almost the same way as the proof of Theorem 2: For any \( \epsilon > 0 \), we can find some large enough time \( t_0 \), such that for any \( t \geq t_0 \),

1. \( \|\Pi_\perp W_1(t)\|_F / \|W_1(t)\|_F \geq \epsilon \) implies that \( \left\langle \Pi_\perp w_{\text{prod}}(t), \nabla R(w_{\text{prod}}(t)) \right\rangle \geq 0 \).
2. \( \langle w_{\text{prod}}(t), z_i \rangle \geq 0 \) for all \( i \), and thus \( \|W_1(t + 1)\|_F \geq \|W_1(t)\|_F \).
3. \( \|W_1(t)\|_F \geq 1 + \sqrt{2R(W(0))}/\epsilon \).

Suppose at some time \( t_1 \geq t_0 \), \( \|\Pi_\perp W_1(t_1)\|_F / \|W_1(t_1)\|_F \geq \epsilon \). As long as this still holds, in light of bullet (1) above, eq. (38) and eq. (39), \( \|\Pi_\perp W_1\|_F^2 \) will increase by at most \( 2R(W(t_1)) \leq 2R(W(0)) \). On the other hand, \( \|W_1\|_F \to \infty \), and thus there exists some \( t_2 > t_1 \) such that \( \|\Pi_\perp W_1(t_2)\|_F / \|W_1(t_2)\|_F < \epsilon \).

Let \( t_3 \) denote the smallest time after \( t_2 \) such that \( \|\Pi_\perp W_1(t_3)\|_F / \|W_1(t_3)\|_F \geq \epsilon \) (if it exists). Recall that \( \|W_1(t + 1)\|_F \leq \|W_1(t)\|_F + 1 \) for any \( t \geq 0 \), and \( \|W_1(t + 1)\|_F \geq \|W_1(t)\|_F \) for any \( t \geq t_0 \), we have
\[
\|\Pi_\perp W_1(t_3)\|_F / \|W_1(t_3)\|_F \leq \|\Pi_\perp W_1(t_3)\|_F / \|W_1(t_3 - 1)\|_F \leq \|\Pi_\perp W_1(t_3 - 1)\|_F / \|W_1(t_3 - 1)\|_F < \epsilon + \frac{1}{\|W_1(t_3 - 1)\|_F}.
\]

After \( t_3 \), \( \|\Pi_\perp W_1\|_F^2 \) will increase by at most \( 2R(W(0)) \), and thus \( \|\Pi_\perp W_1\|_F \) will increase by at most \( \sqrt{2R(W(0))} \). Therefore, for any \( t_4 \geq t_3 \), as long as \( \|\Pi_\perp W_1(t_4)\|_F / \|W_1(t_4)\|_F \geq \epsilon \), we have
\[
\frac{\|\Pi_\perp W_1(t_4)\|_F}{\|W_1(t_4)\|_F} \leq \frac{\|\Pi_\perp W_1(t_4)\|_F}{\|W_1(t_3)\|_F} \leq \frac{\|\Pi_\perp W_1(t_3)\|_F + \sqrt{2R(W(0))}}{\|W_1(t_3)\|_F} \leq \epsilon + \frac{1}{\|W_1(t_3 - 1)\|_F} + \sqrt{2R(W(0))} \leq 2\epsilon,
\]
since $\|W_1(t)\|_F \geq 1 + \sqrt{2R(W(0))}/\epsilon$ after $t_0$. In other words,

$$\limsup_{t \to \infty} \frac{\|\Pi_{\perp} W_1\|_F}{\|W_1\|_F} \leq 2\epsilon.$$ 

Since $\epsilon$ is arbitrary, we have

$$\limsup_{t \to \infty} \frac{\|\Pi_{\perp} W_1\|_F}{\|W_1\|_F} = 0,$$

and thus $\lim_{t \to \infty} |\langle v_1, \bar{u} \rangle| = 1$. 