A UNIFYING VIEW ON IMPLICIT BIAS IN TRAINING LINEAR NEURAL NETWORKS

Chulhee Yun[*]
MIT
chulheey@mit.edu

Shankar Krishnan
Google Research
skrishnan@google.com

Hossein Mobahi
Google Research
hmobahi@google.com

ABSTRACT

We study the implicit bias of gradient flow (i.e., gradient descent with infinitesimal step size) on linear neural network training. We propose a tensor formulation of neural networks that includes fully-connected, diagonal, and convolutional networks as special cases, and investigate the linear version of the formulation called linear tensor networks. With this formulation, we can characterize the convergence direction of the network parameters as singular vectors of a tensor defined by the network. For L-layer linear tensor networks that are orthogonally decomposable, we show that gradient flow on separable classification finds a stationary point of the $\ell_{2/L}$ max-margin problem in a "transformed" input space defined by the network. For underdetermined regression, we prove that gradient flow finds a global minimum which minimizes a norm-like function that interpolates between weighted ℓ_1 and ℓ_2 norms in the transformed input space. Our theorems subsume existing results in the literature while removing standard convergence assumptions. We also provide experiments that corroborate our analysis.

1 INTRODUCTION

Overparametrized neural networks have infinitely many solutions that achieve zero training error, and such global minima have different generalization performance. Moreover, training a neural network is a high-dimensional nonconvex problem, which is typically intractable to solve. However, the success of deep learning indicates that first-order methods such as gradient descent or stochastic gradient descent (GD/SGD) not only (a) succeed in finding global minima, but also (b) are biased towards solutions that generalize well, which largely has remained a mystery in the literature.

To explain part (a) of the phenomenon, there is a growing literature studying the convergence of GD/SGD on overparametrized neural networks (e.g., Du et al. (2018a;b); Allen-Zhu et al. (2018); Zou et al. (2018); Jacot et al. (2018); Oymak & Soltanolkotabi (2020), and many more). There are also convergence results that focus on linear networks, without nonlinear activations (Bartlett et al., 2018; Arora et al., 2019a; Wu et al., 2019; Du & Hu, 2019; Hu et al., 2020). These results typically focus on the convergence of loss, hence do not address *which* of the many global minima is reached.

Another line of results tackles part (b), by studying the implicit bias or regularization of gradient-based methods on neural networks or related problems (Gunasekar et al., 2017; 2018a;b; Arora et al., 2018; Soudry et al., 2018; Ji & Telgarsky, 2019a; Arora et al., 2019b; Woodworth et al., 2020; Chizat & Bach, 2020; Gissin et al., 2020). These results have shown interesting progress that even without explicit regularization terms in the training objective, algorithms such as GD applied on neural networks have an *implicit bias* towards certain solutions among the many global minima. However, in proving such results, many results rely on *convergence assumptions* such as global convergence of loss to zero and/or directional convergence of parameters and gradients. Ideally, such convergence assumptions should be removed because they cannot be tested *a priori* and there are known examples where GD does not converge to global minima under certain initializations (Bartlett et al., 2018; Arora et al., 2019a).

[*]Based on work performed during internship at Google Research

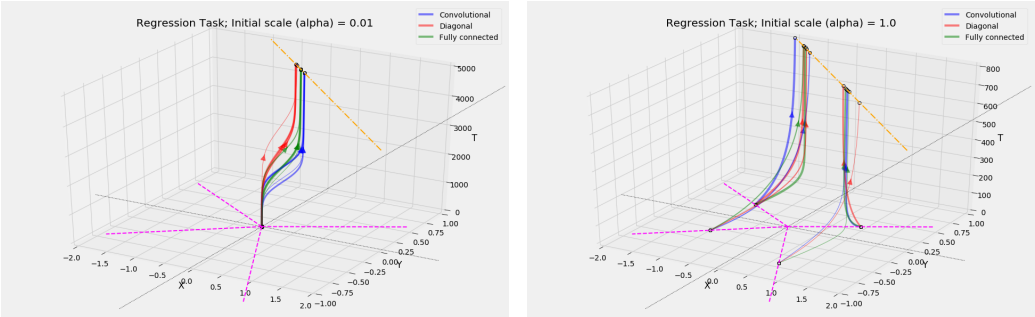

Figure 1: Gradient descent trajectories of linear coefficients of linear fully-connected, diagonal, and convolutional networks on a regression task, initialized with different initial scales $\alpha = 0.01, 1$. Networks are initialized at the same coefficients (circles on purple lines), but follow different trajectories due to implicit biases of networks induced from their architecture. The figures show that our theoretical predictions on limit points (circles on yellow line, the set of global minima) agree with the solution found by GD. For details of the experimental setup, see Section 6.

1.1 Summary of our contributions

We study the implicit bias of gradient flow (GD with infinitesimal step size) on linear neural networks. Following recent progress on this topic, we consider classification and regression problems that have multiple solutions with zero training error. Our analyses apply to a **general class of networks**, and prove **both convergence and implicit bias**, providing a more *complete characterization* of the algorithm trajectory without relying on convergence assumptions.

- We propose a general *tensor formulation* of nonlinear neural networks which includes many network architectures considered in the literature. In this paper, we focus on the linear version of this formulation (i.e., no nonlinear activations), called *linear tensor networks*.
- For linearly separable classification, we prove that linear tensor network parameters converge in direction to *singular vectors* of a tensor defined by the network. As a *corollary*, we show that linear fully-connected networks converge to the ℓ_2 max-margin solution (Ji & Telgarsky, 2020).
- For separable classification, we further show that if the linear tensor network is orthogonally decomposable (Assumption 1), the gradient flow finds the $\ell_{2/\text{depth}}$ max-margin solution in the singular value space, leading the parameters to converge to the *top singular vectors* of the tensor when $\text{depth} = 2$. This theorem subsumes known results on linear convolutional networks and diagonal networks proved in Gunasekar et al. (2018b), *without* using convergence assumptions.
- For underdetermined linear regression, we study the limit points of gradient flow on orthogonally decomposable networks (Assumption 1), and provide a full characterization of the limit points. This theorem covers results on deep matrix sensing (Arora et al., 2019b) as a special case, and extends a similar recent result (Woodworth et al., 2020) to a broader class of networks.
- For underdetermined linear regression with deep linear fully-connected networks, we prove that the network converges to the minimum ℓ_2 norm solutions as we scale the initialization to zero.
- Lastly, we present simple experiments that corroborate our theoretical analysis. Figure 1 shows that our predictions of limit points match with solutions found by GD.

2 Problem settings and related works

We first define notation used in the paper. Given a positive integer a, let $[a] := \{1, \dots, a\}$. We use I_d to denote the $d \times d$ identity matrix. Given a matrix A, we use $\text{vec}(A)$ to denote its vectorization, i.e., the concatenation of all columns of A. For two vectors a and b, let $a \otimes b$ be their tensor product, $a \odot b$ be their element-wise product, and $a^{\odot k}$ be the element-wise k-th power of a. Given an order-L tensor $\mathbf{A} \in \mathbb{R}^{k_1 \times \cdots \times k_L}$, we use $[\mathbf{A}]_{j_1,\dots,j_L}$ to denote the $(j_1, j_2, \dots, j_L)$-th element of $\mathbf{A}$, where $j_l \in [k_l]$ for all $l \in [L]$. In element indexing, we use $:$ to denote all indices in the corresponding dimension, and $a : b$ to denote all indices from a to b. For example, for a matrix A, $[A]_{:,4:6}$ denotes a submatrix that consists of 4th–6th columns of A. The square bracket notation for indexing overloads

with $[a]$ when a 2 N, but they will be distinguishable from the context. Since element indices start from 1, we re-define the modulo operation $a \bmod d := a \quad b\frac{a-1}{d}cd$ 2 $[d]$ for $a > 0$. We use e_j^k to denote the j-th stardard basis vector of the vector space R^k. Lastly, we define the multilinear multiplication between a tensor and linear maps, which can be viewed as a generalization of left- and right-multiplication on a matrix. Given a tensor $\mathbf{A}$ 2 $\mathsf{R}^{k_1 \times \cdots \times k_L}$ and linear maps B_l 2 $\mathsf{R}^{p_l \times k_l}$ for l 2 $[L]$, we define the multilinear multiplication between them as

$$\mathbf{A} \quad (B_1^T, B_2^T, \ldots, B_L^T) = \sum_{j_1,\ldots,j_L} [\mathbf{A}]_{j_1,\ldots,j_L} (e_{j_1}^{k_1} \qquad e_{j_L}^{k_L}) \quad (B_1^T, \ldots, B_L^T)$$
$$:= \sum_{j_1,\ldots,j_L} [\mathbf{A}]_{j_1,\ldots,j_L} (B_1 e_{j_1}^{k_1} \qquad B_L e_{j_L}^{k_L}) \; 2 \; \mathsf{R}^{p_1 \times \cdots \times p_L} .$$

2.1 Problem Settings

We are given a dataset $f(x_i, y_i)g_{i=1}^n$, where x_i 2 R^d and y_i 2 R. We let X 2 $\mathsf{R}^{n \times d}$ and y 2 R^n be the data matrix and the label vector, respectively. We study binary classification and linear regression in this paper, focusing on the settings where there exist *many* global solutions. For binary classification, we assume y_i 2 f 1g and that the data is separable: there exists a unit vector z and a constant $\gamma > 0$ such that $y_i x_i^T z \quad \gamma$ for all i 2 $[n]$. For regression, we consider the underdetermined case ($n \quad d$) where there are many parameters z 2 R^d such that $Xz = y$. Throughout the paper, we assume that X has full row rank.

We use $f(\,;\,) : \mathsf{R}^d \; ! \; \mathsf{R}$ to denote a neural network parametrized by . Given the network and the dataset, we consider minimizing the training loss $L(\,) := \sum_{i=1}^n \ell(f(x_i;\,), y_i)$ over . Following previous results (e.g., Lyu & Li (2020); Ji & Telgarsky (2020)), we use the exponential loss $\ell(\hat{y}, y) = \exp(\hat{y}y)$ for classification problems. For regression, we use the squared error loss $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} \quad y)^2$. On the algorithm side, we minimize L using gradient flow, which can be viewed as GD with infinitesimal step size. The gradient flow dynamics is defined as $\frac{d}{dt} \quad = \quad r \quad L(\,)$.

2.2 Related Works

Gradient flow/descent in separable classification. For linear models $f(x; z) = x^T z$ with separable data, Soudry et al. (2018) show that the GD run on L drives kzk to 1, but z converges in direction to the ℓ_2 max-margin classifier. The limit direction of z is aligned with the solution of

$$\text{minimize}_{z \in \mathsf{R}^d} \quad kzk \quad \text{subject to} \quad y_i x_i^T z \quad 1 \text{ for } i \; 2 \; [n], \tag{1}$$

where the norm in the cost is the ℓ_2 norm. Nacson et al. (2019b;c); Gunasekar et al. (2018a); Ji & Telgarsky (2019b;c) extend these results to other (stochastic) algorithms and non-separable settings.

Gunasekar et al. (2018b) study the same problem on linear neural networks and show that GD exhibits different implicit bias depending on the architecture. The authors show that the linear coefficients of the network converges in direction to the solution of (1) with different norms: ℓ_2 norm for linear fully-connected networks, $\ell_{2/L}$ (quasi-)norm for diagonal networks, and DFT-domain $\ell_{2/L}$ (quasi-)norm for convolutional networks with full-length filters. Here, L denotes the depth. We note that Gunasekar et al. (2018b) assume that GD globally minimizes the loss, and the network parameters and the gradient with respect to the linear coefficients converge in direction. Subsequent results (Ji & Telgarsky, 2019a; 2020) remove such assumptions for linear fully-connected networks.

A recent line of results (Nacson et al., 2019a; Lyu & Li, 2020; Ji & Telgarsky, 2020) studies general homogeneous models and show divergence of parameters to infinity, monotone increase of smoothed margin, directional convergence and alignment of parameters (see Section 4 for details). Lyu & Li (2020) also characterize the limit direction of parameters as the KKT point of a nonconvex max-margin problem similar to (1), but this characterization does not provide useful insights for the functions $f(\,;\,)$ represented by specific architectures, because the formulation is in the parameter space . Also, these results require that gradient flow/descent has already reached 100% training accuracy. Although we study a more restrictive set of networks (i.e., deep linear), we provide a more complete characterization of the implicit bias for the functions $f(\,;\,)$, without assuming 100% training accuracy.

Gradient flow/descent in linear regression. It is known that for linear models $f(x; z) = x^T z$, GD converges to the global minimum that is closest in ℓ_2 distance to the initialization (see e.g.,

Gunasekar et al. (2018a)). However, relatively less is known for deep networks, even for linear networks. This is partly because the parameters do not diverge to infinity, hence making limit points highly dependent on the initialization; this dependency renders analysis difficult. A related problem of matrix sensing aims to minimize $\sum_{i=1}^{n}(y_i - \langle A_i, W_1 \cdots W_L \rangle)^2$ over $W_1, \ldots, W_L \in \mathbb{R}^{d\times d}$. It is shown in Gunasekar et al. (2017); Arora et al. (2019b) that if the sensor matrices A_i commute and we initialize all W_l's to αI, GD finds the minimum nuclear norm solution as $\alpha \to 0$.

Chizat et al. (2019) show that if a network is zero at initialization, and we scale the network output by a factor of $\alpha \to \infty$, then the GD dynamics enters a "lazy regime" where the network behaves like a first-order approximation at its initialization, as also seen in results studying kernel approximations of neural networks and convergence of GD in the corresponding RKHS (e.g., Jacot et al. (2018)).

Woodworth et al. (2020) study linear regression with a diagonal network of the form $f(x; w_+, w_-) = x^T(w_+^{\odot L} - w_-^{\odot L})$, where w_+ and w_- are identically initialized $w_+(0) = w_-(0) = \alpha w$. The authors show that the global minimum reached by GD minimizes a norm-like function which interpolates between (weighted) ℓ_1 norm ($\alpha \to 0$) and ℓ_2 norm ($\alpha \to \infty$). In our paper, we consider a more general class of orthogonally decomposable networks, and obtain similar results interpolating between weighted ℓ_1 and ℓ_2 norms. We also remark that our results include the results in Arora et al. (2019b) as a special case, and we do not assume convergence to global minima, as done in Gunasekar et al. (2017); Arora et al. (2019b); Woodworth et al. (2020).

3 Tensor formulation of neural networks

In this section, we present a general tensor formulation of neural networks. Given an input $x \in \mathbb{R}^d$, the network uses a linear map $\mathbf{M}$ that maps x to an order-L tensor $\mathbf{M}(x) \in \mathbb{R}^{k_1\times\cdots\times k_L}$, where $L \geq 2$. Using parameters $v_l \in \mathbb{R}^{k_l}$ and activation ϕ, the network computes its layers as the following:

$$\begin{aligned} \mathbf{H}_1(x) &= \phi\left(\mathbf{M}(x)(v_1, I_{k_2}, \ldots, I_{k_L})\right) \in \mathbb{R}^{k_2\times\cdots\times k_L}, \\ \mathbf{H}_l(x) &= \phi\left(\mathbf{H}_{l-1}(x)(v_l, I_{k_{l+1}}, \ldots, I_{k_L})\right) \in \mathbb{R}^{k_{l+1}\times\ldots,k_L}, \quad \text{for } l = 2, \ldots, L-1, \\ f(x; \theta) &= \mathbf{H}_{L-1}(x)v_L \in \mathbb{R}. \end{aligned} \tag{2}$$

We use θ to denote the collection of all parameters $(v_1, \ldots, v_L)$. We call $\mathbf{M}(x)$ the *data tensor*. Although this new formulation may look a bit odd in the first glance, it is general enough to capture many network architectures considered in the literature, including fully-connected networks, diagonal networks, and circular convolutional networks. We formally define these architectures below.

Diagonal networks. An L-layer diagonal network is written as

$$f_{\text{diag}}(x; \theta_{\text{diag}}) = \phi(\cdots\phi(\phi(x \odot w_1) \odot w_2)\cdots \odot w_{L-1})^T w_L, \tag{3}$$

where $w_l \in \mathbb{R}^d$ for $l \in [L]$. The representation of f_{diag} as the tensor form (2) is straightforward. Let $\mathbf{M}_{\text{diag}}(x) \in \mathbb{R}^{d\times\cdots\times d}$ have $[\mathbf{M}_{\text{diag}}(x)]_{j,j,\ldots,j} = [x]_j$, while all the remaining entries of $\mathbf{M}_{\text{diag}}(x)$ are set to zero. We can set $v_l = w_l$ for all l, and $\mathbf{M} = \mathbf{M}_{\text{diag}}$ to verify that (2) and (3) are equivalent.

Circular convolutional networks. The tensor formulation (2) includes convolutional networks

$$f_{\text{conv}}(x; \theta_{\text{conv}}) = \phi(\cdots\phi(\phi(x \star w_1) \star w_2)\cdots \star w_{L-1})^T w_L, \tag{4}$$

where $w_l \in \mathbb{R}^{k_l}$ with $k_l \leq d$ and $k_L = d$, and $\star$ defines the circular convolution: for any $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^k$ ($k \leq d$), we have $a \star b \in \mathbb{R}^d$ defined as $[a \star b]_i = \sum_{j=1}^{k}[a]_{(i+j-1) \bmod d}[b]_j$, for $i \in [d]$. Define $\mathbf{M}_{\text{conv}}(x) \in \mathbb{R}^{k_1\times\cdots\times k_L}$ as $[\mathbf{M}_{\text{conv}}(x)]_{j_1,j_2,\ldots,j_L} = [x]_{(\sum_{l=1}^{L} j_l - L + 1) \bmod d}$ for $j_l \in [k_l]$, $l \in [L]$. Setting $v_l = w_l$ and $\mathbf{M} = \mathbf{M}_{\text{conv}}$, we can verify that (2) and (4) are identical.

Fully-connected networks. An L-layer fully-connected network is defined as

$$f_{\text{fc}}(x; \theta_{\text{fc}}) = \phi(\cdots\phi(\phi(x^T W_1)W_2)\cdots W_{L-1})w_L, \tag{5}$$

where $W_l \in \mathbb{R}^{d_l\times d_{l+1}}$ for $l \in [L-1]$ (we use $d_1 = d$) and $w_L \in \mathbb{R}^{d_L}$. One can represent f_{fc} as the tensor form (2) by defining parameters $v_l = \text{vec}(W_l)$ for $l \in [L-1]$ and $v_L = w_L$, and constructing the tensor $\mathbf{M}_{\text{fc}}(x)$ by a recursive "block diagonal" manner. For example, if $L = 2$, we can define $\mathbf{M}_{\text{fc}}(x) \in \mathbb{R}^{d_1 d_2\times d_2}$ to be the Kronecker product of I_{d_2} and x. For deeper networks, we defer the full description of $\mathbf{M}_{\text{fc}}(x)$ to Appendix B.

Our focus: linear tensor networks. Throughout this section, we have used the activation ϕ to motivate our tensor formulation (2) for neural networks with nonlinear activations. For the remaining of the paper, we study the case whose activation is *linear*, i.e., $\phi(t) = t$. In this case,

$$f(x; \quad) = \mathbf{M}(x) \quad (v_1, v_2, \ldots, v_L). \tag{6}$$

We will refer to (6) as *linear tensor networks*, where "linear" is to indicate that the activation is linear. Note that as a function of parameters $v_1, \ldots, v_L$, $f(x; \quad)$ is in fact multilinear. We also remark that when depth $L = 2$, the data tensor $\mathbf{M}(x)$ is a $k_1 \quad k_2$ matrix and the network formulation boils down to $f(x; \quad) = v_1^T \mathbf{M}(x) v_2$.

Since the data tensor $\mathbf{M}(x)$ is a linear function of x, the linear tensor network is also a linear function of x. Thus, the output of the network can also be written as $f(x; \quad) = x^T \quad (\quad)$, where $(\quad) \in \mathbb{R}^d$ denotes the *linear coefficients* computed as a function of the network parameters $\quad$. Since the linear tensor network $f(x; \quad)$ is linear in x, the expressive power of f is at best a linear model $x \mapsto x^T z$. However, even though the models have the same expressive power, their architectural differences lead to different implicit biases in training, which is the focus of our investigation in this paper. Studying separable classification and underdetermined regression is useful for highlighting such biases because there are *infinitely many* coefficients that perfectly classify or fit the dataset.

For our linear tensor network, the evolution of the parameters v_l via gradient flow reads

$$\dot{v}_l = \quad \nabla_{v_l} L(\quad) = \sum_{i=1}^{n} \ell'(f(x_i; \quad), y_i) \mathbf{M}(x_i) \quad (v_1, \ldots, v_{l-1}, I_{k_l}, v_{l+1}, \ldots, v_L)$$
$$= \mathbf{M}(\quad X^T r) \quad (v_1, \ldots, v_{l-1}, I_{k_l}, v_{l+1}, \ldots, v_L), \quad \forall l \in [L],$$

where we initialize $v_l(0) = \alpha v_l$, for $l \in [L]$. We refer to α and v_l as the *initial scale* and *initial direction*, respectively. We note that we do not restrict v_l's to be unit vectors, in order to allow different scaling (at initialization) over different layers. The vector $r \in \mathbb{R}^n$ is the *residual vector*, and each component of r is defined as

$$[r]_i = \ell'(f(x_i; \quad), y_i) = \begin{cases} y_i \exp(\quad y_i f(x_i; \quad)) & \text{for classification,} \\ f(x_i; \quad) \quad y_i & \text{for regression.} \end{cases} \tag{7}$$

4 Implicit Bias of Gradient Flow in Separable Classification

In this section, we present our results on the implicit bias of gradient flow in binary classification with linearly separable data. Recent papers (Lyu & Li, 2020; Ji & Telgarsky, 2020) on this separable classification setup prove that after 100% training accuracy has been achieved by gradient flow (along with other technical conditions), the parameters of L-homogeneous models diverge to infinity, while converging in direction that aligns with the direction of the negative gradient. Mathematically,

$$\lim_{t \to \infty} \| \quad (t) \| = 1, \quad \lim_{t \to \infty} \frac{(t)}{\| \quad (t) \|} = \quad \infty, \quad \lim_{t \to \infty} \frac{(t)^T \nabla_{\Theta} \mathcal{L}(\quad (t))}{\| \quad (t) \| \| \nabla_{\Theta} \mathcal{L}(\quad (t)) \|} = \quad 1.$$

Since the linear tensor network satisfies the technical assumptions in the prior works, we apply these results to our setting and develop a new characterization of the limit directions of the parameters. Here, we present theorems on separable classification with general linear tensor networks. Corollaries for specific networks are deferred to Appendix A.

4.1 Limit Directions of Parameters are Singular Vectors

Consider the singular value decomposition (SVD) of a matrix $A = \sum_{j=1}^{m} s_j (u_j \quad v_j)$, where m is the rank of A. Note that the tuples (u_j, v_j, s_j) are solutions to the system of equations $su = Av$ and $sv = A^T u$. Lim (2005) generalizes this definition of singular vectors and singular values to higher-order tensors: given an order-L tensor $\mathbf{A} \in \mathbb{R}^{k_1 \times \cdots \times k_L}$, we define the singular vectors $u_1, u_2, \ldots, u_L$ and singular value s to be the solution of the following system of equations:

$$s u_l = \mathbf{A} \quad (u_1, \ldots, u_{l-1}, I_{k_l}, u_{l+1}, \ldots, u_L), \text{ for } l \in [L]. \tag{8}$$

Using the definition of the singular vectors of tensors, we can characterize the limit direction of parameters after reaching 100% training accuracy. In Appendix C, we prove the following:

Theorem 1. *Assume that the gradient flow satisfies* $L(\quad(t_0)) < 1$ *for some* $t_0 \quad 0$ *and* $X^T r(t)$ *converges in direction, say* $u^\infty := \lim_{t\to\infty} \frac{X^T r(t)}{\|X^T r(t)\|_2}$. *Then,* $v_1, \ldots, v_L$ *converge to the singular vectors of* $\mathbf{M}(\quad u^\infty)$.

For this theorem, we make some convergence assumptions, because the network is fully general; this is the *only* result where we assume convergence. It fact, for the special case of linear fully-connected networks, the directional convergence assumption is *not* required, and the linear coefficients $_{\text{fc}}(\quad_{\text{fc}})$ converge in direction to the ℓ_2 max-margin classifier. We state this corollary in Appendix A.1; this result also appears in Ji & Telgarsky (2020), but we provide an alternative proof.

4.2 Limit directions for orthogonally decomposable networks

Admittedly, Theorem 1 is not a *full* characterization of the limit directions, because there are usually multiple solutions that satisfy (8). For example, in case of $L = 2$, the data tensor $\mathbf{M}(\quad u^\infty)$ is a matrix and the number of possible limit directions (up to scaling) of (v_1, v_2) is at least the rank of $\mathbf{M}(\quad u^\infty)$. Singular vectors of high order tensors are much less understood than the matrix counterparts, and are much harder to deal with. Although their existence is implied from the variational formulation (Lim, 2005), they are intractable to compute. Testing if a given number is a singular value, approximating the corresponding singular vectors, and computing the best rank-1 approximation are all NP-hard (Hillar & Lim, 2013); let alone orthogonal decompositions.

Given this intractability, it might be reasonable to make some assumptions on the "structure" of the data tensor $\mathbf{M}(x)$, so that they are easier to handle. The following assumption defines a class of *orthogonally decomposable* data tensors, which includes **linear diagonal networks** and **linear full-length convolutional networks** as special cases (for the proof, see Appendix D.2 and D.3).

Assumption 1. *For the data tensor* $\mathbf{M}(x) \; 2 \; \mathbb{R}^{k_1 \times \cdots \times k_L}$ *of a linear tensor network (6), there exist a full column rank matrix* $S \; 2 \; \mathbb{C}^{m \times d}$ $(d \quad m \quad \min_l k_l)$ *and matrices* $U_1 \; 2 \; \mathbb{C}^{k_1 \times m}, \ldots, U_L \; 2 \; \mathbb{C}^{k_L \times m}$ *such that* $U_l^H U_l = I_m$ *for all* $l \; 2 \; [L]$, *and the data tensor* $\mathbf{M}(x)$ *can be written as*

$$\mathbf{M}(x) = \sum_{j=1}^{m} [Sx]_j ([U_1]_{\cdot,j} \quad [U_2]_{\cdot,j} \quad\quad [U_L]_{\cdot,j}). \tag{9}$$

In this assumption, we allow $U_1, \ldots, U_L$ and S to be complex matrices, although $\mathbf{M}(x)$ and parameters v_l stay real, as defined earlier. For a complex matrix A, we use A^* to denote its entry-wise complex conjugate, A^T to denote its transpose (without conjugating), and A^H to denote its conjugate transpose. In case of $L = 2$, Assumption 1 requires that the data tensor $\mathbf{M}(x)$ (now a matrix) has singular value decomposition $\mathbf{M}(x) = U_1 \operatorname{diag}(Sx) U_2^T$; i.e., the left and right singular vectors are independent of x, and the singular values are linear in x. Using Assumption 1, the following theorem characterizes the limit directions.

Theorem 2. *Suppose a linear tensor network satisfies Assumption 1. If there exists* $\lambda > 0$ *such that the initial directions* $v_1, \ldots, v_L$ *of the network parameters satisfy* $|[U_l^T v_l]_j|^2 \quad |[U_L^T v_L]_j|^2 \quad \lambda$ *for all* $l \; 2 \; [L \quad 1]$ *and* $j \; 2 \; [m]$, *then* $(\quad(t))$ *converges in a direction that aligns with* $S^T \quad^\infty$, *where* $^\infty \; 2 \; \mathbb{C}^m$ *denotes a stationary point of the following optimization problem*

$$\text{minimize}_{\in \mathbb{C}^m} \quad \|\| \quad \text{subject to} \quad y_i x_i^T S^T \quad 1, \quad 8i \; 2 \; [n].$$

If S *is invertible, then* $(\quad(t))$ *converges in a direction that aligns with a stationary point* z^∞ *of*

$$\text{minimize}_{z \in \mathbb{R}^d} \quad \|S^{-T} z\|_{2/L} \quad \text{subject to} \quad y_i x_i^T z \quad 1, \quad 8i \; 2 \; [n].$$

Theorem 2 shows that the gradient flow finds sparse $^\infty$ that minimizes the $\ell_{2/L}$ norm in the "singular value space," where the data points x_i are transformed into vectors Sx_i consisting of singular values of $\mathbf{M}(x_i)$. Also, the proof of Theorem 2 reveals that in case of $L = 2$, the parameters $v_l(t)$ in fact converge to the *top* singular vectors of the data tensor $\mathbf{M}(\quad X^T r)$; thus, compared to Theorem 1, we have a more complete characterization of "which" singular vectors to converge to.

The proof of Theorem 2 is in Appendix D. Since the orthogonal decomposition (Assumption 1) of $\mathbf{M}(x)$ tells us that the singular vectors $\mathbf{M}(x)$ in $U_1, \ldots, U_L$ are independent of x, we can transform the network parameters v_l to $U_l^T v_l$ and show that the network behaves similar to a linear diagonal network. This observation comes in handy in the characterization of limit directions.

Remark 1 (Necessity of initialization assumptions)**.** In order to remove the assumption that the loss converges to zero, at least some condition on initialization is *necessary*, because there are examples showing non-convergence of gradient flow for certain initializations (Bartlett et al., 2018; Arora et al., 2019a). In our theorems, we pose assumptions on initial directions v_l that are sufficient conditions for the loss $L(\quad(t))$ to converge to zero. Although such sufficient conditions are "stronger" than assuming $L(\quad(t)) \;!\; 0$, they are useful because they can be easily checked *a priori*, i.e., before running gradient flow. We note an important fact that in Theorems 2 and onwards, the conditions on initialization are used **solely to prove convergence** of the loss to zero, and **our statements on the implicit bias hold whenever the loss converges to zero**, even for initializations that do not satisfy our conditions. In addition, we argue that our assumptions are not too restrictive; λ can be arbitrarily small, so the conditions are satisfied *with probability 1* if we set $v_L = 0$ and randomly sample other v_l's. Setting one layer to zero to prove convergence is also studied in Wu et al. (2019). Lastly, the condition that v_L is "small" can be replaced with any layer; e.g., convergence still holds if $j[U_l^T v_l]_j j^2 \quad j[U_1^T v_1]_j j^2 \quad \lambda$ for all $l = 2, \ldots, L$ and $j \; 2 \; [m]$.

Remark 2 (Comparison to existing results)**.** Theorem 2 leads to corollaries (stated in Appendix A.2) on linear diagonal and full-length convolutional networks, showing that diagonal (or convolutional) networks converge to the stationary point of the max-margin problem with respect to the $\ell_{2/L}$ norm (or DFT-domain $\ell_{2/L}$ norm). Theorem 2 recovers the results in Gunasekar et al. (2018b) without relying on assumptions such as directional convergence of parameters and gradients.

Remark 3 (Implications to architecture design)**.** Theorem 2 shows that the gradient flow finds a solution that is sparse in a "transformed" input space where all data points are transformed with S. This implies something interesting about architecture design: if the sparsity of the solution under a certain linear transformation T is needed, one can design a network using Assumption 1 by setting $S = T$. Training such a network will give us a solution that has the desired sparsity property.

Other than Assumption 1, there is another setting where we can prove a full characterization of limit directions: when there is one data point ($n = 1$) and the network is 2-layer ($L = 2$). This "extremely overparametrized" case is motivated by an experimental paper (Zhang et al., 2019) which studies generalization performance of different architectures when there is only one training data point.

Theorem 3. *Suppose we have a 2-layer linear tensor network* (6) *and a single data point* (x, y)*. Consider the compact SVD* $M(x) = U_1 \, \mathrm{diag}(s) U_2^T$*, where* $U_1 \; 2 \; \mathbb{R}^{k_1 \times m}$*,* $U_2 \; 2 \; \mathbb{R}^{k_2 \times m}$*, and* $s \; 2 \; \mathbb{R}^m$ *for* $m \quad \min\{k_1, k_2\}$*. Let* $\quad^{\infty} \; 2 \; \mathbb{R}^m$ *be a solution of the following optimization problem*

$$\text{minimize}_{\in \mathbb{R}^m} \quad k \; k_1 \quad \text{subject to} \quad y s^T \qquad 1.$$

Assume that there exists $\lambda > 0$ *such that the initial directions* v_1, v_2 *of the network parameters satisfy* $[U_1^T v_1]_j^2 \quad [U_2^T v_2]_j^2 \quad \lambda$ *for all* $j \; 2 \; [m]$*. Then,* v_1 *and* v_2 *converge in direction to* $U_1 \; {}_1^{\infty}$ *and* $U_2 \; {}_2^{\infty}$*, where* $j \; {}_1^{\infty} j = j \; {}_2^{\infty} j = j \; {}^{\infty} j^{\odot 1/2}$*, and* $\mathrm{sign}(\; {}_1^{\infty}) = \mathrm{sign}(y) \quad \mathrm{sign}(\; {}_2^{\infty})$.

The proof of Theorem 3 can be found in Appendix E. Since $^{\infty}$ is the minimum ℓ_1 norm solution in the singular value space, the parameters v_1 and v_2 converge in direction to the top singular vectors. We would like to emphasize that this theorem can be applied to *any* network architecture that can be represented as a linear tensor network. Recall that the previous result (Gunasekar et al., 2018b) only considers full-length filters ($k_1 = d$), hence providing limited insights on networks with small filters, e.g., $k_1 = 2$. In light of this, we present a corollary in Appendix A.3 showing that linear coefficients of convolutional networks converge in direction to a "filtered" version of x.

5 Implicit Bias of Gradient Flow in Underdetermined Regression

In Section 4, the limit directions of parameters we characterized do not depend on initialization. This is due to the fact that the parameters diverge to infinity in separable classification problems, so that the initialization becomes unimportant in the limit. This is not the case in regression setting, because parameters do not diverge to infinity. As we show in this section, the limit points are closely tied to initialization, and our analyses characterize the dependency between them.

5.1 Limit Point Characterization for Orthogonally Decomposable Networks

For the orthogonally decomposable networks satisfying Assumption 1 with real S and U_l's, we consider how limit points of gradient flow change according to initialization. We consider a specific

initialization scheme that, in the special case of diagonal networks, corresponds to setting $w_l(0) = \alpha w$ for $l \in [L-1]$ and $w_L(0) = 0$. We use the following lemma on a relevant system of ODEs:

Lemma 4. *Consider the system of ODEs, where $p, q : \mathbb{R} \to \mathbb{R}$:*

$$\dot{p} = p^{L-2} q, \quad \dot{q} = p^{L-1}, \quad p(0) = 1, \quad q(0) = 0.$$

Then, the solutions $p_L(t)$ and $q_L(t)$ are continuous on their maximal interval of existence of the form $(-c, c)$ where $c \in (0, 1]$. Define $h_L(t) = p_L(t)^{L-1} q_L(t)$; then, $h_L(t)$ is odd and strictly increasing, satisfying $\lim_{t \uparrow c} h_L(t) = 1$ and $\lim_{t \downarrow -c} h_L(t) = -1$.

Using the function $h_L(t)$ from Lemma 4, we can obtain the following theorem that characterizes the limit points as the minimizer of a norm-like function $Q_{L,\alpha,w}$ among the global minima.

Theorem 5. *Suppose a linear tensor network satisfies Assumption 1. Assume further that the matrices $U_1, \dots, U_L$ and S from Assumption 1 are all real matrices. For some $\lambda > 0$, choose any vector $y \in \mathbb{R}^m$ satisfying $|y_j| \neq \lambda$ for all $j \in [m]$, and choose initial directions $v_l = U_l$ for $l \in [L-1]$ and $v_L = 0$. Then, the linear coefficients $w(t)$ converge to $S^T w^\infty$, where w^∞ is the solution of*

$$\text{minimize}_{w \in \mathbb{R}^m} \quad Q_{L,\alpha,w}(w) := \alpha^2 \sum_{j=1}^m |w_j|^2 H_L\left(\frac{w_j}{\alpha^L |w_j|^L}\right) \quad \text{subject to} \quad X S^T w = y,$$

where $Q_{L,\alpha,w} : \mathbb{R}^m \to \mathbb{R}$ is a norm-like function defined using $H_L(t) := \int_0^t h_L^{-1}(\tau) d\tau$. If S is invertible, then $w(t)$ converges to the solution z^∞ of

$$\text{minimize}_{z \in \mathbb{R}^d} \quad Q_{L,\alpha,w}(S^{-T} z) \quad \text{subject to} \quad X z = y.$$

The proofs of Lemma 4 and Theorem 5 are deferred to Appendix F.

Remark 4 (Interpolation between ℓ_1 and ℓ_2)**.** It can be checked that $H_L(t)$ grows like the absolute value function if t is large, and grows like a quadratic function if t is close to zero. This means that

$$\lim_{\alpha \to 0} Q_{L,\alpha,w}(w) / \sum_{j=1}^m \frac{|w_j|}{|w_j|^{L-2}}, \quad \lim_{\alpha \to \infty} Q_{L,\alpha,w}(w) / \sum_{j=1}^m \frac{w_j^2}{w_j^{2L-2}},$$

so $Q_{L,\alpha,w}$ interpolates between the weighted ℓ_1 and weighted ℓ_2 norms of w. Also, the weights in the norm are *dependent* on the initialization direction w unless $L = 2$ and $\alpha \to 0$. In general, $Q_{L,\alpha,w}$ interpolates the standard ℓ_1 and ℓ_2 norms only if $|w_j|$ is the same for all $j \in [m]$. This result is similar to the observations made in Woodworth et al. (2020) which considers a diagonal network with a "differential" structure $f(x; w_+, w_-) = x^T (w_+^{\odot L} - w_-^{\odot L})$. In contrast, our results apply to a more general class of networks, without the need to have the differential structure. In Appendix A.4, we state corollaries of Theorem 5 for linear diagonal networks and linear full-length convolutional networks with even data points. There, we also show that deep matrix sensing with commutative sensor matrices (Arora et al., 2019b) is a special case of our setting.

Next, we present the regression counterpart of Theorem 3, for 2-layer linear tensor networks with a single data point. For this extremely overparametrized setup, we can fully characterize the limit points as functions of initialization $v_1(0) = \alpha v_1$ and $v_2(0) = \alpha v_2$, for *any* linear tensor networks including linear convolutional networks with filter size smaller than input dimension.

Theorem 6. *Suppose we have a 2-layer linear tensor network (6) and a single data point (x, y). Consider the compact SVD $\mathbf{M}(x) = U_1 \operatorname{diag}(s) U_2^T$, where $U_1 \in \mathbb{R}^{k_1 \times m}$, $U_2 \in \mathbb{R}^{k_2 \times m}$, and $s \in \mathbb{R}^m$ for $m = \min\{k_1, k_2\}$. Assume that there exists $\lambda > 0$ such that the initial directions v_1, v_2 of the network parameters satisfy $|[U_1^T v_1]_j^2 - [U_2^T v_2]_j^2| \neq \lambda$ for all $j \in [m]$. Then, gradient flow converges to a global minimizer of the loss L, and $v_1(t)$ and $v_2(t)$ converge to the limit points:*

$$v_1^\infty = \alpha U_1 \left(U_1^T v_1 \cosh\left(g^{-1}\left(\frac{y}{\alpha^2}\right) s\right) + U_2^T v_2 \sinh\left(g^{-1}\left(\frac{y}{\alpha^2}\right) s\right)\right) + \alpha (I_{k_1} - U_1 U_1^T) v_1,$$

$$v_2^\infty = \alpha U_2 \left(U_1^T v_1 \sinh\left(g^{-1}\left(\frac{y}{\alpha^2}\right) s\right) + U_2^T v_2 \cosh\left(g^{-1}\left(\frac{y}{\alpha^2}\right) s\right)\right) + \alpha (I_{k_2} - U_2 U_2^T) v_2,$$

where g^{-1} is the inverse of the following strictly increasing function

$$g(\nu) = \sum_{j=1}^m [s]_j \frac{[U_1^T v_1]_j^2 + [U_2^T v_2]_j^2}{2} \sinh(2[s]_j \nu) + [U_1^T v_1]_j [U_2^T v_2]_j \cosh(2[s]_j \nu).$$

The proof can be found in Appendix G. We can observe that as $\alpha \to 0$, we have $g^{-1}\left(\frac{y}{\alpha^2}\right) \to \infty$, which results in exponentially faster growth of the $\sinh(\cdot)$ and $\cosh(\cdot)$ for the top singular values. As a result, the top singular vectors dominate the limit points v_1^∞ and v_2^∞ as $\alpha \to 0$, and they do not depend on the initial directions v_1, v_2. Experiment results in Section 6 support this observation.

5.2 IMPLICIT BIAS IN FULLY-CONNECTED NETWORKS: THE $\alpha \to 0$ LIMIT

We state our last theoretical element of this paper, which proves that the linear coefficients $_{\text{fc}}(\ _{\text{fc}})$ of deep linear fully-connected networks converge to the minimum ℓ_2 norm solution as $\alpha \to 0$. We assume for simplicity that $d_1 = d_2 = \cdots = d_L = d$ in this section, but we can extend it for $d_l \geq d$ without too much difficulty. Recall $f_{\text{fc}}(x;\ _{\text{fc}}) = x^T W_1 \cdots W_{L-1} w_L$. We minimize the training loss L with initialization $W_l(0) = \alpha W_l$ for $l \in [L-1]$ and $w_L(0) = \alpha w_L$.

Theorem 7. *Assume that initial directions $W_1, \ldots, W_{L-1}, w_L$ satisfy (1) $W_l^T W_l = W_{l+1} W_{l+1}^T$ for $l \in [L-2]$, and (2) there exists $\lambda > 0$ such that $W_{L-1}^T W_{L-1} - w_L w_L^T = \lambda I_d$. Then, the gradient flow converges to a global minimum, and $\lim_{\alpha \to 0} \lim_{t \to \infty} {}_{\text{fc}}(\ _{\text{fc}}(t)) = X^T (X X^T)^{-1} y$.*

The proof is presented in Appendix H. Theorem 7 shows that in the limit $\alpha \to 0$, linear fully-connected networks have bias towards the minimum ℓ_2 norm solution, regardless of the depth. This is consistent with the results shown for classification. We also note that the convergence to a global minimum holds for any $\alpha > 0$, and our sufficient conditions ($W_l^T W_l = W_{l+1} W_{l+1}^T$ and $W_{L-1}^T W_{L-1} - w_L w_L^T = \lambda I_d$) for global convergence is a generalization of the zero-asymmetric initialization scheme ($W_1 = \cdots = W_{L-1} = I_d$ and $w_L = 0$) proposed in Wu et al. (2019).

6 EXPERIMENTS

Regression. To fully visualize the trajectory of linear coefficients, we run simple experiments with 2-layer linear fully-connected/diagonal/convolutional networks with a single 2-dimensional data point $(x, y) = ([1 \ \ 2], 1)$. For this dataset, the minimum ℓ_2 norm solution (corresponding to fully-connected networks) of the regression problem is $[0.2 \ \ 0.4]$, whereas the minimum ℓ_1 norm solution (corresponding to diagonal) is $[0 \ \ 0.5]$ and the minimum DFT-domain ℓ_1 norm solution (corresponding to convolutional) is $[0.33 \ \ 0.33]$. We randomly pick four directions $z_1, \ldots z_4 \in \mathbb{R}^2$, and choose initial directions of the network parameters in a way that their linear coefficients at initialization are exactly $(\ \ (0)) = \alpha^2 z_j$. With varying initial scales $\alpha \in \{0.01, 0.5, 1\}$, we run GD with small step size $\eta = 10^{-3}$ for large enough number of iterations $T = 5 \cdot 10^3$. Figures 1 and 2 plot the trajectories of $(\ \ )$ (appropriately clipped for visual clarity) as well as the predicted limit points (Theorem 6). We observe that even though the networks start at the same linear coefficients $\alpha^2 z_j$, they evolve differently due to different architectures. Note that the prediction of limit points is accurate, and the solution found by GD is less dependent on initial directions when α is small.

Classification. It is shown in the existing works as well as in Section 4 that the limit directions of linear coefficients are independent of the initialization. Is this also true in practice? To see this, we run a set of toy experiments on classification with two data points $(x_1, y_1) = ([1 \ \ 2], +1)$ and $(x_2, y_2) = ([0 \ \ 3], \ -1)$. One can check that the max-margin classifiers for this problem are in the same directions to the corresponding min-norm solutions in the regression problem above. We use the same networks as in regression, and the same set of initial directions satisfying $(\ \ (0)) = \alpha^2 z_j$. With initial scales $\alpha \in \{0.01, 0.5, 1\}$, we run GD with step size $\eta = 5 \cdot 10^{-4}$ for $T = 2 \cdot 10^6$ iterations. All experiments reached $L(\ \ ) \lesssim 10^{-5}$ at the end. The trajectories are plotted in Figure 2 in the Appendix. We find that, in contrast to our theoretical characterization, the actual coefficients are quite dependent on initialization, because we do not train the network all the way to zero loss. This observation is also consistent with a recent analysis (Moroshko et al., 2020) for diagonal networks, and suggests that understanding the behavior of iterates after a finite number of steps is an important future work.

7 CONCLUSION

This paper studies the implicit bias of gradient flow on training linear tensor networks. Under a general tensor formulation of linear networks, we provide theorems characterizing how the network architectures and initializations affect the limit directions/points of gradient flow. Our work provides a unified framework that connects multiple existing results on implicit bias of gradient flow as special cases.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253, 2018.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019a.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 7413–7424, 2019b.

Peter Bartlett, Dave Helmbold, and Philip Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International Conference on Machine Learning*, pp. 521–530, 2018.

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

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2937–2947, 2019.

Simon S Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. *arXiv preprint arXiv:1901.08572*, 2019.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018b.

Christopher J Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations*, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019a.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pp. 1772–1798, 2019b.

Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *arXiv preprint arXiv:1906.04540*, 2019c.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *arXiv preprint arXiv:2006.06657*, 2020.

Lek-Heng Lim. Singular values and eigenvalues of tensors: a variational approach. In *1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005.*, pp. 129–132. IEEE, 2005.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Edward Moroshko, Suriya Gunasekar, Blake Woodworth, Jason D Lee, Nathan Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *arXiv preprint arXiv:2007.06738*, 2020.

Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning*, pp. 4683–4692, 2019a.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019b.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3051–3059. PMLR, 2019c.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference On Learning Theory*, 2020.

Lei Wu, Qingcan Wang, and Chao Ma. Global convergence of gradient descent for deep linear residual networks. In *Advances in Neural Information Processing Systems*, pp. 13389–13398, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C Mozer, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. *arXiv preprint arXiv:1902.04698*, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018.

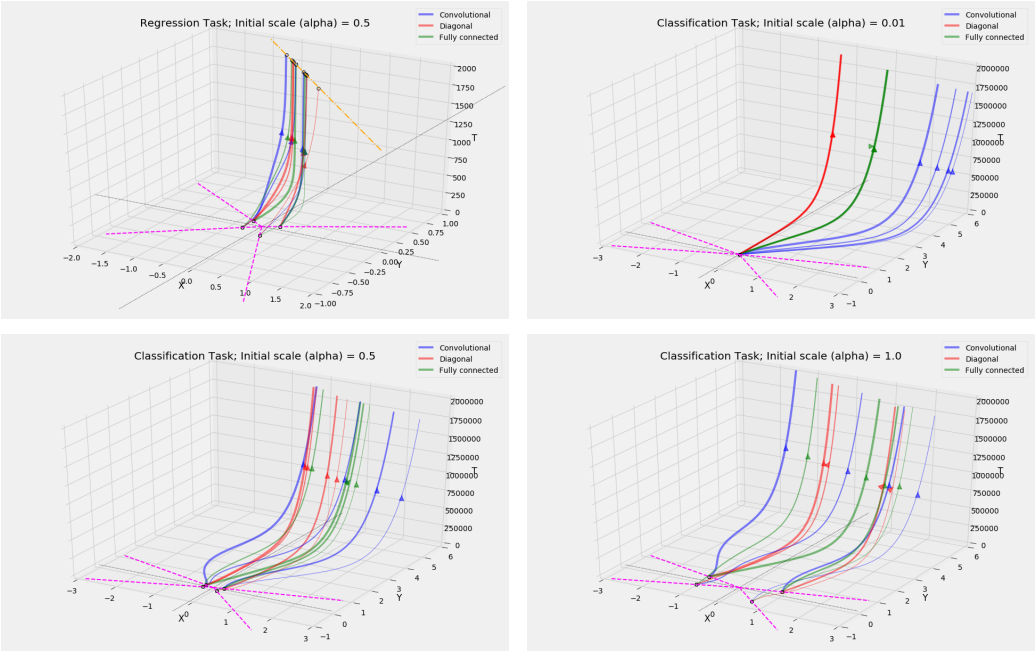

Figure 2: Gradient descent trajectories of linear coefficients of linear fully-connected, diagonal, and convolutional networks on a regression task with initial scale $\alpha = 0.5$ (top left), and networks on a classification task with initial scales $\alpha = 0.01, 0.5, 1$ (rest). Networks are initialized at the same coefficients (circles on purple lines), but follow different trajectories due to different implicit biases of networks induced from their architecture. The top left figure shows that our theoretical predictions on limit points (circles on yellow line, the set of global minima) agree with the solution found by GD. For details of the experimental setup, please refer to Section 6.

A Corollaries on specific network architectures

We present corollaries obtained by specializing the theorems in the main text to specific network architectures. We briefly review the linear neural network architectures studied in this section.

Linear fully-connected networks. An L-layer linear fully-connected network is defined as

$$f_{\text{fc}}(x; \quad _{\text{fc}}) = x^T W_1 \quad W_{L-1} w_L, \tag{10}$$

where $W_l \ 2 \ \mathbb{R}^{d_l \times d_{l+1}}$ for $l \ 2 \ [L \quad 1]$ (we use $d_1 = d$) and $w_L \ 2 \ \mathbb{R}^{d_L}$.

Linear diagonal networks. An L-layer linear diagonal network is written as

$$f_{\text{diag}}(x; \quad _{\text{diag}}) = (x \quad w_1 \quad \quad w_{L-1})^T w_L, \tag{11}$$

where $w_l \ 2 \ \mathbb{R}^d$ for $l \ 2 \ [L]$.

Linear (circular) convolutional networks. An L-layer linear convolutional network is written as

$$f_{\text{conv}}(x; \quad _{\text{conv}}) = (\quad ((x \star w_1) \star w_2) \quad \star w_{L-1})^T w_L, \tag{12}$$

where $w_l \ 2 \ \mathbb{R}^{k_l}$ with $k_l \quad d$ and $k_L = d$, and $\star$ defines the circular convolution: for any $a \ 2 \ \mathbb{R}^d$ and $b \ 2 \ \mathbb{R}^k$ ($k \quad d$), we have $a \star b \ 2 \ \mathbb{R}^d$ defined as $[a \star b]_i = \sum_{j=1}^{k} [a]_{(i+j-1) \bmod d} [b]_j$, for $i \ 2 \ [d]$. In case of $k_l = d$ for all $l \ 2 \ [L]$, we refer to this network as *full-length* convolutional networks.

Deep matrix sensing. The deep matrix sensing problem considered in Gunasekar et al. (2017); Arora et al. (2019b) aims to minimize the following problem

$$\underset{W_1, \ldots, W_L \in \mathbb{R}^d \quad d}{\text{minimize}} \quad L_{\text{ms}}(W_1 \quad W_L) := \sum_{i=1}^{n} (y_i \quad \hbar A_i, W_1 \quad W_L i)^2, \tag{13}$$

where the sensor matrices $A_1, \ldots, A_n \ 2 \ \mathbb{R}^{d \times d}$ are symmetric. Following Gunasekar et al. (2017); Arora et al. (2019b), we consider sensor matrices $A_1, \ldots, A_n \ 2 \ \mathbb{R}^{d \times d}$ that commute. To make the problem underdetermined, we assume that $n \quad d$, and A_i's are linearly independent.

A.1 COROLLARY OF THEOREM 1

Corollary 1. *Consider an L-layer linear fully-connected network (10). If the training loss satisfies $L(\ _{\mathrm{fc}}(t_0)) < 1$ for some $t_0 \quad 0$, then $\ _{\mathrm{fc}}(\ _{\mathrm{fc}}(t))$ converges in a direction that aligns with the solution of the following optimization problem*

$$\text{minimize}_{z\in\mathbb{R}^d} \quad \|z\|_2^2 \quad \text{subject to} \quad y_i x_i^T z \quad 1, \quad 8i \ 2 \ [n].$$

Corollary 1 shows that whenever the network separates the data correctly, the linear coefficients $_{\mathrm{fc}}(\ _{\mathrm{fc}})$ convergence in direction to the ℓ_2 max-margin classifier. Note that this corollary does not require the directional convergence of $X^T r$, which is different from Theorem 1. In fact, this corollary also appears in Ji & Telgarsky (2020), but we provide an alternative proof based on our tensor formulation. The proof of Corollary 1 can be found in Appendix C.

A.2 COROLLARIES OF THEOREM 2

Corollary 2. *Consider an L-layer linear diagonal network (11). If there exists $\lambda > 0$ such that the initial directions $w_1, \dots, w_L$ of the network parameters satisfy $[w_l]_j^2 \quad [w_L]_j^2 \quad \lambda$ for all $l \ 2 \ [L \quad 1]$ and $j \ 2 \ [d]$, then $_{\mathrm{diag}}(\ _{\mathrm{diag}}(t))$ converges in a direction that aligns with a stationary point z^∞ of*

$$\text{minimize}_{z\in\mathbb{R}^d} \quad \|z\|_{2/L} \quad \text{subject to} \quad y_i x_i^T z \quad 1, \quad 8i \ 2 \ [n].$$

For full-length convolutional networks, we define $F \ 2 \ \mathbb{C}^{d\times d}$ to be the matrix of discrete Fourier transform basis $[F]_{j,k} = \frac{1}{\sqrt{d}} \exp(\ \frac{\sqrt{-1}\cdot 2\pi(j-1)(k-1)}{d})$. Note that $F^* = F^{-1}$, and both F and F^* are symmetric, but not Hermitian.

Corollary 3. *Consider an L-layer linear full-length convolutional network (12). If there exists $\lambda > 0$ such that the initial directions $w_1, \dots, w_L$ of the network parameters satisfy $|[F w_l]_j|^2$ $|[F w_L]_j|^2 \quad \lambda$ for all $l \ 2 \ [L \quad 1]$ and $j \ 2 \ [d]$, then $_{\mathrm{conv}}(\ _{\mathrm{conv}}(t))$ converges in a direction that aligns with a stationary point z^∞ of*

$$\text{minimize}_{z\in\mathbb{R}^d} \quad \|Fz\|_{2/L} \quad \text{subject to} \quad y_i x_i^T z \quad 1, \quad 8i \ 2 \ [n].$$

Corollary 2 shows that in the limit, linear diagonal network finds a sparse solution z that is a stationary point of the $\ell_{2/L}$ max-margin classification problem. Corollary 3 has a similar conclusion except that the standard $\ell_{2/L}$ norm is replaced with DFT-domain $\ell_{2/L}$ norm. By specifying mild conditions on initialization (see Remark 1), these corollaries remove the convergence assumptions required in Gunasekar et al. (2018b). The proofs of Corollaries 2 and 3 are in Appendix D.

A.3 COROLLARY OF THEOREM 3

Recall that Theorem 3 can be applied to any 2-layer networks that can be represented as linear tensor networks. Examples include the convolutional networks that are not full-length (i.e., filter size $k_1 < d$), which are not covered by the previous result (Gunasekar et al., 2018b). Here, we present the characterization of convergence directions of $_{\mathrm{conv}}(\ _{\mathrm{conv}}(t))$ for small-filter cases: $k_1 = 1$ and $k_1 = 2$.

Corollary 4. *Consider a 2-layer linear convolutional network (12) with $k_1 = 1$ and a single data point (x, y). If there exists $\lambda > 0$ such that the initial directions w_1 and w_2 of the network parameters satisfy $\|x\|^2 v_1^2 \quad (x^T v_2)^2 \quad \|x\|^2 \lambda$, then $_{\mathrm{conv}}(\ _{\mathrm{conv}}(t))$ converges in direction that aligns with yx.*

Consider a 2-layer linear convolutional network (12) with $k_1 = 2$ and a single data point (x, y). Let $x := [[x]_2 \quad [x]_d \quad [x]_1]$, and $\overleftarrow{x} := [[x]_d \quad [x]_1 \quad [x]_{d-1}]$. If there exists $\lambda > 0$ such that the initial directions w_1 and w_2 of the network parameters satisfy

$$([v_1]_1 + [v_1]_2)^2 \quad \frac{((x + \overleftarrow{x})^T v_2)^2}{\|x\|\|x\|_2^2 + x^T \overleftarrow{x}} \quad \lambda, \quad \textit{and} \quad ([v_1]_1 \quad [v_1]_2)^2 \quad \frac{((x \quad \overleftarrow{x})^T v_2)^2}{\|x\|\|x\|_2^2 \quad x^T \overleftarrow{x}} \quad \lambda,$$

then $_{\mathrm{conv}}(\ _{\mathrm{conv}}(t))$ converges in a direction that aligns with a filtered version of x:

$$\lim_{t\to\infty} \frac{_{\mathrm{conv}}(\ _{\mathrm{conv}}(t))}{\|\ _{\mathrm{conv}}(\ _{\mathrm{conv}}(t))\|_2} \ / \ \begin{cases} 2yx + y\overleftarrow{x} + y\overleftarrow{x} & \text{if } x^T\overleftarrow{x} > 0, \\ 2yx \quad y\overleftarrow{x} \quad y\overleftarrow{x} & \text{if } x^T\overleftarrow{x} < 0. \end{cases}$$

Corollary 4 shows that if the filter size is $k_1 = 1$, then the limit direction of $\beta_{\mathrm{conv}}(\alpha_{\mathrm{conv}})$ is always the ℓ_2 max-margin classifier. Note that this is quite different from the case of $k_1 = d$ which converges to the DFT-domain ℓ_1 max-margin classifier. However, for $1 < k_1 < d$, it is difficult to characterize the limit direction as the max-margin classifier of some common norms. Rather, the limit directions of $\beta_{\mathrm{conv}}(\alpha_{\mathrm{conv}})$ correspond to a "filtered" version of the data point, and the weights of the filter depend on the data point x. For $k_1 = 2$, the filter is a low-pass filter if the autocorrelation $x^T \mathrm{circ}(x)$ of x is positive, and high-pass if the autocorrelation is negative. For $k_1 = 2$, the filter weights are more complicated to characterize in terms of x, and the filter length increases as α increases. We prove Corollary 4 in Appendix E.

A.4 Corollaries of Theorem 5

In this subsection, we apply Theorem 5 to linear diagonal networks, linear full-length convolutional networks with even data, and deep matrix sensing. The proofs of the corollaries can be found in Appendix F.

Corollary 5. Consider an L-layer linear diagonal network (11). For some $\gamma > 0$, choose any vector $w \in \mathbb{R}^d$ satisfying $[w]_j^2 \neq \gamma$ for all $j \in [d]$, and choose initial directions $w_l = w$ for $l \in [L-1]$ and $w_L = 0$. Then, the linear coefficients $\beta_{\mathrm{diag}}(\alpha_{\mathrm{diag}}(t))$ converge to the solution z^1 of

$$\underset{z \in \mathbb{R}^d}{\text{minimize}} \quad Q_{L,\gamma,w}(z) := \sum_{j=1}^{d} [w]_j^2 H_L\left(\frac{[z]_j}{\gamma [w]_j^L}\right) \quad \text{subject to} \quad Xz = y.$$

Recall that the original statement of Assumption 1 allows the matrices $U_1, \ldots, U_L$ to be complex, but Theorem 5 poses another assumption that these matrices are real. In applying Theorem 2 to convolutional networks to get Corollary 3, we used the fact that the data matrix $\mathrm{Mconv}(x)$ of a linear full-length convolutional network satisfies Assumption 1 with $\Phi = d^{\frac{L-1}{2}} F$ and $U_1 = \cdots = U_L = \frac{1}{\sqrt{d}} F$, where $F \in \mathbb{C}^{d \times d}$ is the matrix of discrete Fourier transform basis $[F]_{j,k} = \frac{1}{\sqrt{d}} \exp(\frac{\sqrt{-1}\, 2\pi (j-1)(k-1)}{d})$ and $\bar{F}$ is the complex conjugate of F. Note that these are complex matrices, so one cannot directly apply Theorem 5 to convolutional networks. However, it turns out that if the data and initialization are even, we can derive a corollary for convolutional networks.

We say that a vector is even when it satisfies the even symmetry, as in even functions. More concretely, a vector $x \in \mathbb{R}^d$ is even if $[x]_{j+2} = [x]_{d-j}$ for $j = 0, \ldots, \lfloor \frac{d-3}{2} \rfloor$; i.e., the vector has the even symmetry around its "origin" $[x]_1$. From the definition of the matrix $F \in \mathbb{C}^{d \times d}$, it is straightforward to check that if x is real and even, then its DFT Fx is also real and even (see Appendix F.4 for details).

Corollary 6. Consider an L-layer linear full-length convolutional network (12). Assume that the data points $\{x_i\}_{i=1}^n$ are all even. For some $\gamma > 0$, choose any even vector w satisfying $[Fw]_j^2 \neq \gamma$ for all $j \in [d]$, and choose initial directions $w_l = w$ for $l \in [L-1]$ and $w_L = 0$. Then, the linear coefficients $\beta_{\mathrm{conv}}(\alpha_{\mathrm{conv}}(t))$ converge to the solution z^1 of

$$\underset{z \in \mathbb{R}^d;\ \text{even}}{\text{minimize}} \quad Q_{L,\gamma,Fw}(Fz) := \sum_{j=1}^{d} [Fw]_j^2 H_L\left(\frac{[Fz]_j}{\gamma [Fw]_j^L}\right) \quad \text{subject to} \quad Xz = y.$$

Corollaries 5 and 6 show that the interpolation between minimum weighted ℓ_1 and weighted ℓ_2 solutions occurs for diagonal networks, and also for convolutional networks (in DFT domain, with the restriction of even symmetry). The conclusion of Corollary 5 is similar to the results in Woodworth et al. (2020), but the network architecture (11) considered in our corollary is a slightly different from the "differential" network $f(x; w_+, w_-) = x^T (w_+^L - w_-^L)$ in Woodworth et al. (2020).

As mentioned in the main text, we can actually show that the matrix sensing result in Arora et al. (2019b) is a special case of our Theorem 5. Given any symmetric matrix $M \in \mathbb{R}^{d \times d}$, let $\mathrm{eig}(M) \in \mathbb{R}^d$ be the d-dimensional vector of eigenvalues of M.

Corollary 7. Consider the depth-L deep matrix sensing problem (13). Let A_i's be symmetric, and assume $A_1, \ldots, A_n$ commute. For $\gamma > 0$, choose initialization $W_l(0) = \gamma I_d$ for $l \in [L-1]$ and $W_L(0) = 0$. Then, the product $W_L(t) \cdots W_1(t)$ converge to the solution M^1 of

$$\underset{M \in \mathbb{R}^{d \times d};\ \text{symmetric}}{\text{minimize}} \quad Q_{L,\gamma}(\mathrm{eig}(M)) := \sum_{j=1}^{d} H_L\left(\frac{[\mathrm{eig}(M)]_j}{\gamma^L}\right) \quad \text{subject to} \quad L_{ms}(M) = 0.$$

Under an additional assumption that A_i's are positive semide nite, Theorem 2 in Arora et al. (2019b) studies the initialization $W_l(0) = I_d$ for all $l \in [L]$, and shows that the limit point of $W_L \cdots W_1$ converges to the minimum nuclear norm solution as $\to 0$. We remove the assumption of positive de niteness on A_i's and let $W_L(0) = 0$, to show a complete characterization of the solution found by gradient ow, which interpolates between the minimum nuclear norm (i.e., Schatten 1-norm) solution (when $\to 0$) and the minimum Frobenius norm (i.e., Schatten 2-norm) solution (when $\to \infty$).

B Tensor representation of fully-connected networks

In Section 3, we only de ned the data tensor $M_{fc}(x)$ of fully-connected networks for $L = 2$. Here, we describe an iterative procedure constructing the data tensor for deep fully-connected networks.

We start with $T_1(x) := x \in \mathbb{R}^{d_1}$. Next, de ne a block diagonal matrix $T_2(x) \in \mathbb{R}^{d_1 d_2 \times d_2}$ where the "diagonals" $[T_2(x)]_{d_1(j-1)+1:jd_1,j} = T_1(x)$ for $j \in [d_2]$, while all the other entries are lled with 0. We continue this "block diagonal" procedure, as the following. Having de ned $T_{l-1}(x) \in \mathbb{R}^{d_1 d_2 \cdots d_{l-2} d_{l-1} \times d_{l-1}}$,

1. De ne $T_l(x) \in \mathbb{R}^{d_1 d_2 \cdots d_{l-1} d_l \times d_l}$.
2. Set $[T_l(x)]_{:,\ldots,:,d_{l-1}(j-1)+1:jd_{l-1},j} = T_{l-1}(x), \forall j \in [d_l]$.
3. Set all the remaining entries of $T_l(x)$ to zero.

We repeat this process for $= 2, \ldots, L$, and set $M_{fc}(x) := T_L(x)$. By de ning the parameters of the tensor formulation $v_l = \mathrm{vec}(W_l)$ for $l \in [L-1]$ and $v_L = w_L$, and using the tensor $M(x) = M_{fc}(x)$, we can check the equivalence of (2) and (5).

C Proofs of Theorem 1 and Corollary 1

C.1 Proof of Theorem 1

The proof of Theorem 1 is outlined as follows. First, using the directional convergence and alignment results in Ji & Telgarsky (2020), we prove that each of our network parameters converges in direction, and it aligns with its corresponding negative gradient $-\nabla_{v_l} L$. Then, we prove that the directions of v_l's are actually singular vectors of $M(u^1)$, where $u^1 := \lim_{t\to\infty} \frac{X^\top r(t)}{\|X^\top r(t)\|_2}$.

Since a linear tensor network is an L-homogeneous polynomial of $v_1, \ldots, v_L$, it satis es the assumptions required for Theorems 3.1 and 4.1 in Ji & Telgarsky (2020). These theorems imply that if the gradient ow satis es $L(\theta(t_0)) < 1$ for some $t_0 > 0$, then $\theta(t)$ converges in direction, and the direction aligns with $\nabla L(\theta(t))$; that is,

$$\lim_{t\to\infty} \|\theta(t)\|_2 = \infty, \quad \lim_{t\to\infty} \frac{\theta(t)}{\|\theta(t)\|_2} = \bar{\theta}, \quad \lim_{t\to\infty} \frac{\theta(t)^\top \nabla L(\theta(t))}{\|\theta(t)\|_2 \|\nabla L(\theta(t))\|_2} = 1. \tag{14}$$

For linear tensor networks (6), the parameter θ is the concatenation of all parameter vectors $v_1, \ldots, v_L$, so (14) holds for $\theta = [v_1^\top \cdots v_L^\top]^\top$.

Now, recall that by the de nition of the linear tensor network, we have the following gradient ow

$$\dot{v}_l = M(X^\top r)(v_1, \ldots, v_{l-1}, I_{k_l}, v_{l+1}, \ldots, v_L).$$

Note that we can apply this to calculate the rate of growth of $\|v_l\|_2^2$:

$$\begin{aligned}
\frac{d}{dt}\|v_l\|_2^2 &= 2v_l^\top \dot{v}_l = 2v_l^\top M(X^\top r)(v_1, \ldots, v_{l-1}, I_{k_l}, v_{l+1}, \ldots, v_L) \\
&= 2M(X^\top r)(v_1, \ldots, v_{l-1}, v_l, v_{l+1}, \ldots, v_L) \\
&= \frac{d}{dt}\|v_{l'}\|_2^2 \quad \text{for any } l' \in [L],
\end{aligned}$$

so the rate at which $\|v_l\|_2^2$ grows over time is the same for all layers $l \in [L]$. By the definition of $\dot{v}_l$ and (14), we have

$$\frac{d}{dt}\|v_l\|_2^2 = \sum_{l=1}^{L} \|v_l\|_2^2 \geq \Omega\left(\|v_l\|_2^2\right);$$

which then implies

$$\lim_{t\to\infty} \|v_l(t)\|_2 \to \infty, \quad \lim_{t\to\infty} \frac{\|v_l(t)\|_2}{\|v_{l'}(t)\|_2} = \sqrt{\frac{\|v_l(t)\|_2^2}{\|v_{l'}(t)\|_2^2}} = \sqrt{L};$$

for all $l \in [L]$. Now, let I_l be the set of indices that correspond to the components in v_l. It follows from (14) that

$$\lim_{t\to\infty} \frac{v_l(t)}{\|v_l(t)\|_2} = \lim_{t\to\infty} \frac{v_l(t)}{\|v(t)\|_2} \frac{\|v(t)\|_2}{\|v_l(t)\|_2} = \lim_{t\to\infty} \frac{[v(t)]_{I_l}}{\|v(t)\|_2} \frac{\|v(t)\|_2}{\|v_l(t)\|_2} = \sqrt{L}[\nabla\Phi(v)]_{I_l};$$

thus showing the directional convergence of v_l's.

Next, it follows from directional convergence of v and its alignment with $\nabla L(v)$ (14) that $\nabla L(v)$ also converges in direction, in the opposite direction of v. By comparing the components in I_l's, we get that $\nabla_{v_l} L(v)$ converges in the opposite direction of v_l.

For any $l \in [L]$, now let $v_l^\infty := \lim_{t\to\infty} \frac{v_l(t)}{\|v_l(t)\|_2}$. Also recall the assumption that $X^\top r(t)$ converges in direction; let the unit vector $u^\infty := \lim_{t\to\infty} \frac{X^\top r(t)}{\|X^\top r(t)\|_2}$ be the limit direction. By the gradient flow dynamics of v_l, we have

$$v_l^\infty \propto \nabla_{v_l} L(v^\infty) = M(v^\infty)\left(v_1^\infty, \ldots, v_{l-1}^\infty, I_{k_l}, v_{l+1}^\infty, \ldots, v_L^\infty\right);$$

for all $l \in [L]$. Note that this equation has the same form as (8), the definition of singular vectors in tensors. So this proves that $(v_1^\infty, \ldots, v_L^\infty)$ are singular vectors of $M(v^\infty)$.

C.2 Proof of Corollary 1

The proof proceeds as follows. First, we will show using the structure of the data matrix M_{fc} that the limit direction of linear coefficients $\beta_{fc}(v_{fc}^\infty)$ is proportional to cu^∞, where c is a nonzero scalar and u^∞ is the limit direction of $X^\top r$. Then, through a closer look at v_1^∞ and c, we will prove that $\beta_{fc}(v_{fc}^\infty)$ is in fact a conic combination of the support vectors (i.e., the data points with the minimum margins). Finally, we will compare $\beta_{fc}(v_{fc}^\infty)$ with the KKT conditions of the ℓ_2 max-margin classification problem and conclude that $\beta_{fc}(v_{fc}^\infty)$ must be in the same direction as the ℓ_2 max-margin classifier.

Due to the way how the data tensor M_{fc} is constructed for fully-connected networks (Appendix B), we always have

$$\nabla_{v_1} L(v_{fc}) = M_{fc}(X^\top r)\,(I_{k_1}, v_2, \ldots, v_L) \in \operatorname{span}\left\{ \begin{bmatrix} X^\top r \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ X^\top r \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ X^\top r \end{bmatrix} \right\}.$$

From Theorem 1, we have directional convergence of v_1 and its alignment with $\nabla_{v_1} L(v_{fc})$. This means that the limit direction v_1^∞, which is a fixed vector, must be also in the span of vectors written above. This implies that $X^\top r$ must also converge to some direction, say $u^\infty := \lim_{t\to\infty} \frac{X^\top r(t)}{\|X^\top r(t)\|_2}$.

Now recall the definition of v_1 in case of the fully-connected network: $v_1 = \operatorname{vec}(W_1)$. So, by reshaping v_1^∞ into its original $d_1 \times d_2$ matrix form W_1^∞, we have

$$W_1^\infty \propto u^\infty q^\top;$$

for some $q \in \mathbb{R}^{d_2}$. This implies that the linear coefficients $\beta_{fc}(v_{fc})$ of the network converge in direction to

$$\beta_{fc}(v_{fc}^\infty) = W_1^\infty W_2^\infty \cdots W_{L-1}^\infty w_L^\infty \propto u^\infty q^\top W_2^\infty \cdots W_{L-1}^\infty w_L^\infty = cu^\infty; \tag{15}$$

where c is some nonzero real number.

Let us now take a closer look at the vector u^1, the limit direction of $X^T r$. Recall from Section 2.1 that for any $i \in [n]$,

$$[r]_i = \exp(-y_i f_{\theta_{fc}}(x_i; \theta_{fc})) = \exp(-y_i x_i^T \theta_{fc}(\theta_{fc}));$$

in case of classification. Recall that $\|\theta_{fc}(\theta_{fc}(t))\|_2 \to \infty$ while converging to a certain direction $\theta_{fc}(\theta_{fc}^1)$. This means that if

$$y_j x_j^T \theta_{fc}(\theta_{fc}^1) > y_i x_i^T \theta_{fc}(\theta_{fc}^1)$$

for any $i, j \in [n]$, then

$$\lim_{t\to\infty} \frac{\exp(-y_j x_j^T \theta_{fc}(\theta_{fc}(t)))}{\exp(-y_i x_i^T \theta_{fc}(\theta_{fc}(t)))} = 0. \tag{16}$$

Take i to be the index of any support vector, i.e., any i that attains the minimum $y_i x_i^T \theta_{fc}(\theta_{fc}^1)$ among all data points. Using such i, the observation (16) implies that $\lim_{t\to\infty} [r(t)]_j = 0$ for any x_j that is not a support vector. Thus, by the argument above, u^1 can in fact be written as

$$u^1 = \lim_{t\to\infty} \frac{\sum_{i=1}^n x_i [r(t)]_i}{\|\sum_{i=1}^n x_i [r(t)]_i\|_2} = \sum_{i=1}^n \gamma_i y_i x_i; \tag{17}$$

where $\gamma_i > 0$ for all $i \in [n]$, and $\gamma_j = 0$ for x_j's that are not support vectors. Combining (17) and (15),

$$\theta_{fc}(\theta_{fc}^1) / c \sum_{i=1}^n \gamma_i y_i x_i. \tag{18}$$

Recall that we do not yet know whether c, introduced in (15), is positive or negative; we will now show that c has to be negative. From Lyu & Li (2020), we know that $L(\theta_{fc}(t)) \to 0$, which implies that $y_i x_i^T \theta_{fc}(\theta_{fc}^1) > 0$ for all $i \in [n]$. However, if $c > 0$, then (18) implies that $\theta_{fc}(\theta_{fc}^1)$ is inside a cone K defined as

$$K := \left\{ \sum_{i=1}^n \gamma_i y_i x_i \;\middle|\; \gamma_i \geq 0, \forall i \in [n] \right\}.$$

Note that the polar cone of K, denoted as K°, is

$$K^\circ := \{z \mid z^T v \leq 0, \forall v \in K\} = \{z \mid y_i x_i^T z \leq 0, \forall i \in [n]\}.$$

It is known that $K \cap K^\circ = \{0\}$ for any convex cone K and its polar cone K°. Therefore, having $c > 0$ implies that $\theta_{fc}(\theta_{fc}^1) \in K \cap K^\circ$, which means that there exists some $i \in [n]$ such that $y_i x_i^T \theta_{fc}(\theta_{fc}^1) < 0$; this contradicts the fact that the loss goes to zero as $t \to \infty$. Therefore, c in (15) and (18) must be negative:

$$\theta_{fc}(\theta_{fc}^1) / -\sum_{i=1}^n \gamma_i y_i x_i; \tag{19}$$

for $\gamma_i \geq 0$ for all $i \in [n]$ and $\gamma_j = 0$ for all x_j's that are not suport vectors.

Finally, compare (19) with the KKT conditions of the following optimization problem:

$$\underset{z}{\text{minimize}} \quad \|z\|_2^2 \quad \text{subject to} \quad y_i x_i^T z \geq 1, \; \forall i \in [n].$$

The KKT conditions of this problem are

$$z = \sum_{i=1}^n \gamma_i y_i x_i, \quad \text{and} \quad \gamma_i \geq 0, \;\; \gamma_i (1 - y_i x_i^T z) = 0 \;\text{ for all } i \in [n],$$

where $\gamma_1, \ldots, \gamma_n$ are the dual variables. Note that this is (up to scaling) satisfied by $\theta_{fc}(\theta_{fc}^1)$ (19), if we replace γ_i's with γ_i's. This finishes the proof that $\theta_{fc}(\theta_{fc}^1)$ is aligned with the ℓ_2 max-margin classifier.

D PROOFS OF THEOREM 2 AND COROLLARIES 2 & 3

D.1 PROOF OF THEOREM 2

D.1.1 CONVERGENCE OF LOSS TO ZERO

Since Theorem 2 does not assume the existence of t_0 satisfying $L(\ (t_0)) < 1$, we need to first show that given the conditions on initialization, the training loss $L(\ (t))$ converges to zero. Recall from Section 2.1 that

$$\underline{v}_l = \ r_{\ v_l} L(\ ) = M(\ X^T r)\ \ (v_1; \cdots; v_{l\ 1}; l_{k_l}; v_{l+1}; \cdots; v_L):$$

Applying the structure (9) in Assumption 1, we get

$$\begin{aligned}
\underline{v}_l &= M(\ X^T r)\ \ (v_1; \cdots; v_{l\ 1}; l_{k_l}; v_{l+1}; \cdots; v_L) \\
&= \sum_{j=1}^{m} [SX^T r]_j \, (v_1^T [U_1]_{;j} \quad v_{l\ 1}^T [U_{l\ 1}]_{;j} \quad [U_l]_{;j} \quad v_{l+1}^T [U_{l+1}]_{;j} \quad v_L^T [U_L]_{;j}) \\
&= \sum_{j=1}^{m} [SX^T r]_j \prod_{k \neq l} [U_k^T v_k]_j \ [U_l]_{;j} :
\end{aligned}$$

Left-multiplying U_l^H (the conjugate transpose of U_l) to both sides, we get

$$U_l^H \underline{v}_l = \ SX^T r \prod_{k \neq l} U_k^T v_k; \tag{20}$$

where $\prod$ denotes the product using entry-wise multiplication.

Now consider the rate of growth for the absolute value squared of the j-th component of $U_l^T v_l$:

$$\begin{aligned}
\frac{d}{dt} |[U_l^T v_l]_j|^2 &= \frac{d}{dt} [U_l^T v_l]_j \, [U_l^T v_l]_j = \frac{d}{dt} [U_l^T v_l]_j \, [U_l^H v_l]_j \\
&= [\ U_l^T \underline{v}_l]_j \, [U_l^H v_l]_j + [\ U_l^H \underline{v}_l]_j \, [U_l^T v_l]_j \\
&= 2 \, \mathrm{Re} \ [U_l^H \underline{v}_l]_j \, [U_l^T v_l]_j \\
&= 2 \, \mathrm{Re} \ \ [SX^T r]_j \prod_{k=1}^{L} [U_k^T v_k]_j \\
&= \frac{d}{dt} |[U_{l'}^T v_{l'}]_j|^2 \quad \text{for any } l' \in [L],
\end{aligned}$$

so for any $j \in [m]$, the squared absolute value of the j-th components in $U_l^T v_l$ grow at the same rate for each layer $l \in [L]$. This means that the gap between any two different layers stays constant for all $t \geq 0$. Combining this with our conditions on initial directions, we have

$$\begin{aligned}
|[U_l^T v_l(t)]_j|^2 \ \ |[U_L^T v_L(t)]_j|^2 &= |[U_l^T v_l(0)]_j|^2 \ \ |[U_L^T v_L(0)]_j|^2 \\
&= \ ^2 |[U_l^T v_l]_j|^2 \quad ^2 |[U_L^T v_L]_j|^2 \quad ^2 ;
\end{aligned} \tag{21}$$

for any $j \in [m]$, $l \in [L \quad 1]$, and $t \geq 0$. This inequality also implies

$$|[U_l^T v_l(t)]_j|^2 \ \ |[U_L^T v_L(t)]_j|^2 + \ ^2 \quad ^2 : \tag{22}$$

Let us now consider the time derivative of $L(\ (t))$. We have the following chain of upper bounds on the time derivative:

$$\begin{aligned}
\frac{d}{dt} L(\ (t)) &= r \quad L(\ (t))^T \dot{\ }(t) = \ kr \quad L(\ (t))k_2^2 \\
&\ \ kr \quad _{v_L} L(\ (t))k_2^2 = \ k \ \underline{v}_L(t)k_2^2 \\
&\overset{(a)}{\ } k \ U_L^H \underline{v}_L(t)k_2^2 \overset{(b)}{=} \quad SX^T r(t) \prod_{k \neq L} U_k^T v_k(t) \ ^2_2 \\
&= \sum_{j=1}^{m} |[SX^T r(t)]_j|^2 \prod_{k \neq L} |[U_k^T v_k(t)]_j|^2
\end{aligned}$$

$$\overset{(c)}{=} 2L^{-2}\sum_{j=1}^{m} |[SX^T r(t)]_j|^2$$
$$= 2L^{-2}\|SX^T r(t)\|_2^2$$
$$\overset{(d)}{\geq} 2L^{-2}s_{\min}(S)^2\|X^T r(t)\|_2^2, \tag{23}$$

where (a) used the fact that $\|v_\perp(t)\|_2^2 = \|U_L U_L^H v_\perp(t)\|_2^2$ because it is a projection onto a subspace, and $\|U_L U_L^H v_\perp(t)\|_2^2 = \|U_L^H v_\perp(t)\|_2^2$ because $U_L^H U_L = I_{k_L}$; (b) is due to (20); (c) is due to (22); and (d) used the fact that $S \in \mathbb{C}^{m\times d}$ is a matrix that has full column rank, so for any $z \in \mathbb{C}^d$, we can use $\|Sz\|_2 \geq s_{\min}(S)\|z\|_2$ where $s_{\min}(S)$ is the minimum singular value of S.

We now prove a lower bound on the quantity $\|X^T r(t)\|_2^2$. Recall from Section 2.1 the definition of $[r(t)]_i = -y_i \exp(-y_i f(x_i; \theta(t)))$ for classification problems. Also, recall the assumption that the dataset is linearly separable, which means that there exists a unit vector $z \in \mathbb{R}^d$ such that

$$y_i x_i^T z \geq \gamma > 0$$

holds for all $i \in [n]$, for some $\gamma > 0$. Using these,

$$\|X^T r(t)\|_2^2 = \Big\|\sum_{i=1}^{n} y_i x_i \exp(-y_i f(x_i; \theta(t)))\Big\|_2^2$$
$$\geq \Big[z^T \sum_{i=1}^{n} y_i x_i \exp(-y_i f(x_i; \theta(t)))\Big]^2$$
$$\geq \gamma^2 \Big[\sum_{i=1}^{n} \exp(-y_i f(x_i; \theta(t)))\Big]^2 = \gamma^2 L(\theta(t))^2.$$

Combining this with (23), we get

$$\frac{d}{dt}L(\theta(t)) \leq -2L^{-2}s_{\min}(S)^2\gamma^2 L(\theta(t))^2,$$

which implies

$$L(\theta(t)) \leq \frac{L(\theta(0))}{1 + 2L^{-2}s_{\min}(S)^2\gamma^2 t}.$$

Therefore $L(\theta(t)) \to 0$ as $t \to \infty$.

D.1.2 Characterizing the limit direction

Since we proved that $L(\theta(t)) \to 0$, the argument in the proof of Theorem 1 applies to this case, and shows that the parameters converge in direction and align with $\bar{v}_l = \arg\min_{v_l} L(v_l)$. Let $v_l^\infty := \lim_{t\to\infty} \frac{v_l(t)}{\|v_l(t)\|_2}$ be the limit direction of v_l.

The remaining steps of the proof are as follows. We first prove that $SX^T r(t)$ converges in direction u^∞. Using this u^∞, we derive a number of conditions that has to be satisfied by the limit directions of the parameters. Finally, we compare these conditions with the KKT conditions of the minimization problem, and finish the proof.

By Assumption 1, we have

$$f(x; \theta) = M(x)(v_1, \ldots, v_L) = \sum_{j=1}^{m} [Sx]_j \prod_{l=1}^{L} [U_l^T v_l]_j$$
$$= \sum_{j=1}^{m} \prod_{l=1}^{L} [U_l^T v_l]_j [S]_{j,:} x = x^T S^T \prod_{l\in[L]} U_l^T v_l = x^T S^T \beta,$$

Here, we defined $\beta := \bigodot_{l\in[L]} U_l^T v_l \in \mathbb{C}^m$. Since the linear coefficients must be real, we have $S^T \beta \in \mathbb{R}^d$ for any real v_l's. Since v_l's converge in direction, β also converges in direction, to $\beta^\infty := \bigodot_{l\in[L]} U_l^T v_l^\infty$. So we can express the limit direction of $f(\theta)$ as

$$(\beta^\infty)/\|S^T \beta^\infty\| = S^T \prod_{l\in[L]} U_l^T v_l^\infty / \|S^T \beta^\infty\|. \tag{24}$$

From (20) and alignment of u_l and v_l, we have

$$\lim_{t\to\infty} U_l^H v_l(t) = \lim_{t\to\infty} (U_l^T v_l(t)) \Big/ \lim_{t\to\infty} \|SX^T r(t)\| \prod_{k\neq l} U_k^T v_k(t). \tag{25}$$

Since all vectors $U_l^T v_l(t)$ converge in direction, the term $SX^T r(t)$ should also converge in direction. Let $u^1 := \lim_{t\to\infty} \frac{SX^T r(t)}{\|SX^T r(t)\|_2}$. One can use the same argument as in Appendix C.2, more specifically (16) and (17), to show that u^1 can be written as

$$u^1 = \lim_{t\to\infty} \frac{S\sum_{i=1}^n x_i [r(t)]_i}{\|S\sum_{i=1}^n x_i [r(t)]_i\|_2} = S\sum_{i=1}^n \rho_i y_i x_i; \tag{26}$$

where $\rho_i \geq 0$ for all $i \in [n]$, and $\rho_j = 0$ for x_j's that are not support vectors, i.e., those satisfying $y_j x_j^T S^T u^1 > \min_{i\in[n]} y_i x_i^T S^T u^1$.

Using u^1, we can rewrite (25) as

$$U_l^H v_l^1 \Big/ u^1 \prod_{k\neq l} U_k^T v_k^1;$$

for all $l \in [L]$. Element-wise multiplying $U_l^T v_l^1$ to both sides gives

$$U_l^T v_l^1 \circ U_l^H v_l^1 = |U_l^T v_l^1|^2 \Big/ u^1 \prod_{k\in[L]} U_k^T v_k^1 = u^1 \Big/ u^1, \tag{27}$$

where a^b denotes element-wise bth power of the vector a. Since the LHS of (27) is a positive real number, we have

$$\arg(|[U_l^T v_l]_j|^2) = 0 = \arg([u^1]_j) + \arg([\rho]_j); \tag{28}$$

so using this, (27) becomes

$$|U_l^T v_l^1|^2 / |u^1| |\rho| = |\rho|. \tag{29}$$

Now element-wise multiply (29) for all $l \in [L]$, then we get

$$|\rho|^2 / |u^1| |\rho|^L = |\rho|^L. \tag{30}$$

A close look at (30) reveals that if $\rho \neq 2$, ρ and u^1 must satisfy that

$$|[\rho]_j| \neq 0 \Rightarrow |[u^1]_j| / |[\rho]_j|^{\frac{2}{L}-1}; \tag{31}$$

for all $j \in [m]$. There is another condition that has to be satisfied when $L = 2$:

$$|[\rho]_j| = 0, |[\rho]_{j'}| \neq 0 \Rightarrow |[u^1]_j| \leq |[u^1]_{j'}|; \tag{32}$$

for any $j, j' \in [m]$; let us prove why. First, consider the time derivative of $[\rho]_j = [U_1^T v_1]_j [U_2^T v_2]_j$.

$$\begin{aligned}\frac{d}{dt}[\rho(t)]_j &= [U_1^T v_1(t)]_j \frac{d}{dt}[U_2^T v_2(t)]_j + [U_2^T v_2(t)]_j \frac{d}{dt}[U_1^T v_1(t)]_j \\ &\overset{(a)}{=} [SX^T r(t)]_j (|[U_1^T v_1(t)]_j|^2 + |[U_2^T v_2(t)]_j|^2);\end{aligned} \tag{33}$$

where (a) used (20). Now consider

$$\frac{\frac{d}{dt}[\rho(t)]_j}{\|SX^T r(t)\|_2 |[\rho(t)]_j|} = \frac{|[SX^T r(t)]_j|}{\|SX^T r(t)\|_2} \frac{|[U_1^T v_1(t)]_j|^2 + |[U_2^T v_2(t)]_j|^2}{|[\rho(t)]_j|}. \tag{34}$$

We want to compare this quantity for different $j, j' \in [m]$. Before we do that, we take a look at the last term in the RHS of (34). Recall from (21) that

$$|[U_1^T v_1(t)]_j|^2 = |[U_2^T v_2(t)]_j|^2 + |[U_1^T v_1(0)]_j|^2 - |[U_2^T v_2(0)]_j|^2. \tag{35}$$

For simplicity, let $\gamma_j := |[U_1^T v_1(0)]_j|^2 - |[U_2^T v_2(0)]_j|^2$, which is a positive number due to our assumption on initialization. Then, we can use (35) and $|[\rho(t)]_j| = |[U_1^T v_1(t)]_j| |[U_2^T v_2(t)]_j|$ to show that

$$\frac{|[U_1^T v_1(t)]_j|^2 + |[U_2^T v_2(t)]_j|^2}{|[\rho(t)]_j|} = \frac{2|[U_2^T v_2(t)]_j|^2 + \gamma_j}{|[U_2^T v_2(t)]_j| \sqrt{|[U_2^T v_2(t)]_j|^2 + \gamma_j}} \geq 2;$$

$$\lim_{t\to\infty} \frac{\|[U_1^T v_1(t)]_j\|^2 + \|[U_2^T v_2(t)]_j\|^2}{\|[\ (t)]_j\|} = 2 \quad \text{if} \quad \lim_{t\to\infty} \|[U_2^T v_2(t)]_j\| = 1 \ :$$

Recall that we want to prove that (32) should necessarily hold. For the sake of contradiction, suppose that there exists $j \in [m]$ that satisfies $[\ ^1]_j = 0$ but $[u^1]_j > [u^1]_{j'}$, for some $j' \in [m]$ satisfying $[\ ^1]_{j'} \neq 0$. Note that having $[\ ^1]_j = 0$ and $[\ ^1]_{j'} \neq 0$ implies that $[\ (t)]_{j'} \to 1$ and $\frac{[\ (t)]_j}{[\ (t)]_{j'}} \to 0$. We now want to compare the ratio of (34) for j and j'. First, note that

$$\lim_{t\to\infty} \frac{|[SX^T r(t)]_j|}{|[SX^T r(t)]_{j'}|} \Big/ \frac{\|SX^T r(t)\|_2}{\|SX^T r(t)\|_2} = \frac{[u^1]_j}{[u^1]_{j'}} > 1. \tag{36}$$

Next, using $\frac{[\ (t)]_j}{[\ (t)]_{j'}} \to 0$ and the fact that $x \mapsto \frac{2x^2+}{x^2+}$ is a decreasing function of x for any > 0, we have

$$\frac{(\|[U_1^T v_1(t)]_j\|^2 + \|[U_2^T v_2(t)]_j\|^2)}{(\|[U_1^T v_1(t)]_{j'}\|^2 + \|[U_2^T v_2(t)]_{j'}\|^2))} \Big/ \frac{\|[\ (t)]_j\|}{\|[\ (t)]_{j'}\|} \quad 1, \tag{37}$$

for any $t \quad t_0$, when t_0 is large enough. Combining (36) and (37) to compare the ratio of (34) for j and j', we get that there exists some $t_0' \quad 0$ such that for any $t \quad t_0'$, we have

$$\frac{\frac{d}{dt}|[\ (t)]_j|}{\|[\ (t)]_j\|} \Big/ \frac{\frac{d}{dt}|[\ (t)]_{j'}|}{\|[\ (t)]_{j'}\|} > 1. \tag{38}$$

This implies that the ratio of the absolute value of time derivative of $[\ (t)]_j$ to the absolute value of current value of $[\ (t)]_j$ is strictly bigger than that of $[\ (t)]_{j'}$. Moreover, we saw in (33) that the phase of $\frac{d}{dt}[\ (t)]_j$ converges to that of $[u^1]_j$. Since this holds for all $t \quad t_0'$, (38) results in a growth of $|[\ (t)]_j|$ that is exponentially faster than that of $|[\ (t)]_{j'}|$, so $[\ (t)]_j$ becomes a dominant component in (t) as $t \to 1$. This contradicts that $[\ ^1]_j = 0$, hence the condition (32) has to be satisfied.

So far, we have characterized a number of conditions (26), (28), (31), (32) that have to be satisfied by the limit directions u^1 and ^{1} of $X^T r$ and . We now consider the following optimization problem and prove that these conditions are in fact the KKT conditions of the optimization problem. Consider

$$\underset{\in C^m}{\text{minimize}} \quad \| \|_{2=L} \quad \text{subject to} \quad y_i x_i^T S^T \quad 1, \ \forall i \in [n]. \tag{39}$$

The KKT conditions of this problem are

$$\partial \| \|_{2=L} \ni S \sum_{i=1}^n {}_i y_i x_i, \quad \text{and} \quad {}_i \quad 0, \quad {}_i(1 \quad y_i x_i^T S^T) = 0 \ \text{for all } i \in [n],$$

where ${}_1, \ldots, {}_n$ are the dual variables. The symbol $\partial \| \|_{2=L}$ denotes the (local) subdifferential of the $\ell_{2=L}$ norm[1], which can be written as

$$\partial \| \|_1 = \{ u \in C^m \mid |[u]_j| \quad 1 \text{ for all } j \in [m], \text{ and } [\ ]_j \neq 0 \ ) \quad [u]_j = \exp(\sqrt{-1}\arg([\ ]_j)) \},$$

if $L = 2$ (in this case $\partial \| \|_1$ is the global subdifferential), and

$$\partial \| \|_{2=L} = \{ u \in C^m \mid [\ ]_j \neq 0 \ ) \quad [u]_j = \frac{2}{L}|[\ ]_j|^{\frac{2}{L}\ 1} \exp(\sqrt{-1}\arg([\ ]_j)) \},$$

if $L > 2$. By replacing ${}_i$'s with ${}_i$'s defined in (26), we can check from (26), (28), (31), (32) that the that ^{1} and u^1 satisfy the KKT conditions up to scaling. Therefore, by (24), (t) converges in direction aligned with $S^T\ ^1$, where ^{1} is again aligned with a stationary point (global minimum in case of $L = 2$) of the optimization problem (39).

If S is invertible, we can get $S^T(\ ^1) / \ ^1$. Plugging this into the optimization problem (39) gives the last statement of the theorem.

[1] the definition of subdifferentials used here is taken from Gunasekar et al. (2018b).

D.2 Proof of Corollary 2

It suffices to prove that linear diagonal networks satisfy Assumption 1, with $S = I_d$. The proof is very straightforward, since $M_{diag}(x) \in \mathbb{R}^{d \times d}$ has $[M_{diag}(x)]_{j,j,\ldots,j} = [x]_j$ while all the remaining entries are zero. It is straightforward to verify that $M_{diag}(x)$ satisfies Assumption 1 with $S = U_1 = \cdots = U_L = I_d$. A direct substitution into Theorem 2 gives the corollary.

D.3 Proof of Corollary 3

For full-length convolutional networks ($k_1 = \cdots = k_L = d$), we will prove that they satisfy Assumption 1 with $S = d^{\frac{L-2}{2}} F$ and $U_1 = \cdots = U_L = F$, where $F \in \mathbb{C}^{d \times d}$ is the matrix of discrete Fourier transform basis, $[F]_{j,k} = \frac{1}{\sqrt{d}} \exp(\frac{\sqrt{-1}\, 2\pi (j-1)(k-1)}{d})$ and $\bar{F}$ is the complex conjugate of F.

For simplicity of notation, define $\omega = \exp(\frac{\sqrt{-1}\, 2\pi}{d})$. With such matrices S and $U_1, \ldots, U_L$, we can write $M(x)$ as

$$M(x) = \sum_{j=1}^{d} [Sx]_j \left([U_1]_{:,j} \otimes [U_2]_{:,j} \otimes \cdots \otimes [U_L]_{:,j}\right)$$

$$= \sum_{j=1}^{d} d^{\frac{L-2}{2}} \sum_{k=1}^{d} [x]_k \begin{bmatrix} \omega^{0} \\ \omega^{(j-1)(k-1)} \\ \omega^{2(j-1)(k-1)} \\ \vdots \\ \omega^{(d-1)(j-1)(k-1)} \end{bmatrix}^{\otimes L};$$

where $\otimes^L$ denotes the L-times tensor product of ω. We will show that $M(x) = M_{conv}(x)$.

For any $j_1, \ldots, j_L \in [d]$,

$$[M(x)]_{j_1,\ldots,j_L} = \frac{1}{d} \sum_{l=1}^{d} \left(\sum_{k=1}^{d} [x]_k \, \omega^{(l-1)(k-1)} \right) \omega^{(l-1)(\sum_{q=1}^{L} j_q - L)}$$

$$= \frac{1}{d} \sum_{k=1}^{d} [x]_k \sum_{l=1}^{d} \omega^{(l-1)(k-1+\sum_{q=1}^{L} j_q - L)}.$$

Recall that

$$\sum_{l=1}^{d} \omega^{(l-1)(k-1+\sum_{q=1}^{L} j_q - L)} = \begin{cases} d & \text{if } k-1+\sum_{q=1}^{L} j_q - L \text{ is a multiple of } d; \\ 0 & \text{otherwise} \end{cases}$$

Using this, we have

$$[M(x)]_{j_1,\ldots,j_L} = \frac{1}{d} \sum_{k=1}^{d} [x]_k \sum_{l=1}^{d} \omega^{(l-1)(k-1+\sum_{q=1}^{L} j_q - L)}$$

$$= [x]_{\sum_{q=1}^{L} j_q - L + 1 \bmod d} = [M_{conv}(x)]_{j_1,\ldots,j_L}.$$

Hence, linear full-length convolutional networks satisfy Assumption 1 with $S = d^{\frac{L-2}{2}} F$. A direct substitution into Theorem 2 and then using the fact that $|[Fz]_j| = |[\bar{F}z]_j| = |[Fz]_j|$ for any real vector $z \in \mathbb{R}^d$ gives the corollary.

E Proofs of Theorem 3 and Corollary 4

E.1 Proof of Theorem 3

E.1.1 Convergence of loss to zero

Since Theorem 3 does not assume the existence of t_0 satisfying $L(\beta(t_0)) < 1$, we need to first show that given the conditions on initialization, the training loss $L(\beta(t))$ converges to zero. Since

$L = 2$ and $M(x) = U_1 \operatorname{diag}(s) U_2^T$, we can write the gradient flow dynamics from Section 2.1 as

$$\begin{aligned} \underline{v}_1 &= -M(X^T r) \quad (I_{k_1}; v_2) = -r U_1 \operatorname{diag}(s) U_2^T v_2; \\ \underline{v}_2 &= -M(X^T r) \quad (v_1; I_{k_2}) = -r U_2 \operatorname{diag}(s) U_1^T v_1; \end{aligned} \tag{40}$$

where $r(t) = -y \exp(-yf(x; (t)))$ is the residual of the data point $(x; y)$. From (40) we get

$$U_l^T \underline{v}_1 = -rs \quad U_2^T v_2; \quad U_2^T \underline{v}_2 = -rs \quad U_1^T v_1: \tag{41}$$

Now consider the rate of growth for the jth component of $U_1^T v_1$ squared:

$$\frac{d}{dt}[U_1^T v_1]_j^2 = 2[U_1^T v_1]_j [U_1^T \underline{v}_1]_j = -2r[s]_j [U_1^T v_1]_j [U_2^T v_2]_j = \frac{d}{dt}[U_2^T v_2]_j^2: \tag{42}$$

So for any $j \in [m]$, $[U_1^T v_1]_j^2$ and $[U_2^T v_2]_j^2$ grow at the same rate. This means that the gap between the two layers stays constant for all $t \geq 0$. Combining this with our conditions on initial directions,

$$\begin{aligned} [U_1^T v_1(t)]_j^2 \quad [U_2^T v_2(t)]_j^2 &= [U_1^T v_1(0)]_j^2 \quad [U_2^T v_2(0)]_j^2 \\ &= {}^2[U_1^T v_1]_j^2 \quad {}^2[U_2^T v_2]_j^2 \quad {}^2; \end{aligned} \tag{43}$$

for any $j \in [m]$ and $t \geq 0$. This inequality implies

$$[U_1^T v_1(t)]_j^2 \quad [U_2^T v_2(t)]_j^2 + {}^2 \quad {}^2: \tag{44}$$

Let us now consider the time derivative of $L((t))$. We have the following chain of upper bounds on the time derivative:

$$\begin{aligned} \frac{d}{dt} L((t)) &= -r \quad L((t))^T \dot{(t)} = \|r\| \quad L((t))\|_2^2 \\ &\leq \|r\| \quad {}_{v_2} L((t))\|_2^2 = \| \quad \underline{v}_2(t)\|_2^2 \\ &\overset{(a)}{=} \| \quad U_2^T \underline{v}_2(t)\|_2^2 \overset{(b)}{=} \quad r(t)^2 \quad s \quad U_1^T v_1(t) \quad {}_2^2 \\ &= \quad r(t)^2 \sum_{j=1}^m [s]_j^2 [U_1^T v_1(t)]_j^2 \\ &\overset{(c)}{\leq} \quad {}^2 r(t)^2 \sum_{j=1}^m [s]_j^2 \\ &= \quad {}^2 \|s\|_2^2 L((t))^2; \end{aligned}$$

where (a) used the fact that $\|\underline{v}_2(t)\|_2^2 \quad \| \quad U_2 U_2^T \underline{v}_2(t)\|_2^2$ because it is a projection onto a subspace, and $\| \quad U_2 U_2^T \underline{v}_\perp (t)\|_2^2 = \| \quad U_2^T \underline{v}_2(t)\|_2^2$ because $U_2^T U_2 = I_{k_2}$; (b) is due to (41); (c) is due to (44). From this, we get

$$L((t)) \quad \frac{L((0))}{1 + {}^2 \|s\|_2^2 t}:$$

Therefore $L((t)) \to 0$ as $t \to 1$.

E.1.2 Characterizing the limit direction

Since we proved that $L((t)) \to 0$, the argument in the proof of Theorem 1 applies to this case, and shows that the parameters converge in direction and align with $\underline{v}_l = -r \quad {}_{v_l} L( )$. Let $v_l^1 := \lim_{t \to 1} \frac{v_l(t)}{\|v_l(t)\|_2}$ be the limit direction of v_l. As done in the proof of Theorem 2, define $(t) = U_1^T v_1(t) \quad U_2^T v_2(t)$ and $^1 = U_1^T v_1^1 \quad U_2^T v_2^1$.

It follows from $r(t) = -y \exp(-yf(x; (t)))$ that we have $\operatorname{sign}(r(t)) = -\operatorname{sign}(y)$. Using this, (41), and alignment of v_l and $\underline{v}_l$, we have

$$U_1^T v_1^1 \ / \ ys \quad U_2^T v_2^1 \ ; \ U_2^T v_2^1 \ / \ ys \quad U_1^T v_1^1 \ : \tag{45}$$

Element-wise multiplying $U_l^T v_l^1$ to both sides gives

$$(U_1^T v_1^1)^{\ 2} / \ ys \quad {}^1 \ ; \ (U_2^T v_2^1)^{\ 2} / \ ys \quad {}^1 \ : \tag{46}$$

Since the LHSs are positive and so is positive, the following equations have to be satisfied for all $j \in [m]$:

$$\text{sign}(y) = \text{sign}([\ ^{1}\ ]_j): \tag{47}$$

Now, multiplying both sides of the two equations (46), we get

$$(\ ^{1}\ )^2 / s^2 \quad (\ ^{1}\ )^2: \tag{48}$$

From (48), ^{1} must satisfy that

$$[\ ^{1}\ ]_j \neq 0;\ [\ ^{1}\ ]_{j'} \neq 0 \implies |\ [s]_j| = |[s]_{j'}|; \tag{49}$$

for all $j, j' \in [m]$. As in the proof of Theorem 2, there is another condition that has to be satisfied:

$$[\ ^{1}\ ]_j = 0;\ [\ ^{1}\ ]_{j'} \neq 0 \implies |\ [s]_j| \geq |\ [s]_{j'}|; \tag{50}$$

for any $j, j' \in [m]$; let us prove why. First, consider the time derivative of $[\ ]_j = [U_1^T v_1]_j [U_2^T v_2]_j$.

$$\begin{aligned}\frac{d}{dt}[\ (t)]_j &= [U_1^T v_1(t)]_j \frac{d}{dt}[U_2^T v_2(t)]_j + [U_2^T v_2(t)]_j \frac{d}{dt}[U_1^T v_1(t)]_j \\ &\overset{(a)}{=} r(t)[s]_j ([U_1^T v_1(t)]_j^2 + [U_2^T v_2(t)]_j^2);\end{aligned}$$

where (a) used (41). Now consider

$$\frac{\frac{d}{dt}[\ (t)]_j}{|r(t)|\,|[\ (t)]_j|} = |[s]_j| \frac{[U_1^T v_1(t)]_j^2 + [U_2^T v_2(t)]_j^2}{|[\ (t)]_j|}. \tag{51}$$

We want to compare this quantity for different $j \in [m]$. Before we do that, we take a look at the last term in the RHS of (51). Recall from (43) that

$$[U_1^T v_1(t)]_j^2 = [U_2^T v_2(t)]_j^2 + [U_1^T v_1(0)]_j^2 \quad [U_2^T v_2(0)]_j^2. \tag{52}$$

For simplicity, let $_j := [U_1^T v_1(0)]_j^2 \quad [U_2^T v_2(0)]_j^2$, which is a positive number due to our assumption on initialization. Then, we can use (52) and $|[d(t)]_j| = |[U_1^T v_1(t)]_j|\,|[U_2^T v_2(t)]_j|$ to show that

$$\frac{[U_1^T v_1(t)]_j^2 + [U_2^T v_2(t)]_j^2}{|[\ (t)]_j|} = \frac{2[U_2^T v_2(t)]_j^2 + \ _j}{|[U_2^T v_2(t)]_j| \sqrt{[U_2^T v_2(t)]_j^2 + \ _j}} \quad 2;$$

$$\lim_{t\to 1} \frac{[U_1^T v_1(t)]_j^2 + [U_2^T v_2(t)]_j^2}{|[\ (t)]_j|} = 2 \quad \text{if} \quad \lim_{t\to 1} |[U_2^T v_2(t)]_j| = 1 :$$

Recall that we want to prove that (50) should necessarily hold. For the sake of contradiction, suppose that there exists $j \in [m]$ that satisfies $[\ ^{1}\ ]_j = 0$ but $|[s]_j| > |[s]_{j'}|$, for some $j' \in [m]$ satisfying $[\ ^{1}\ ]_{j'} \neq 0$. Note that having $[\ ^{1}\ ]_j = 0$ and $[\ ^{1}\ ]_{j'} \neq 0$ implies that $|[\ (t)]_{j'}| \to 1$ and $\frac{|[\ (t)]_j|}{|[\ (t)]_{j'}|} \to 0$. We now want to compare the ratio of (51) for j and j'. Using $\frac{|[\ (t)]_j|}{|[\ (t)]_{j'}|} \to 0$ and the fact that $x \mapsto \frac{2x^2+}{x\sqrt{x^2+}}$ is a decreasing function of x for 0 for any > 0, we have

$$\frac{([U_1^T v_1(t)]_j^2 + [U_2^T v_2(t)]_j^2)/|[\ (t)]_j|}{([U_1^T v_1(t)]_{j'}^2 + [U_2^T v_2(t)]_{j'}^2))/|[\ (t)]_{j'}|} \quad 1; \tag{53}$$

for any $t \quad t_0$, when t_0 is large enough. Combining $\frac{|[s]_j|}{|[s]_{j'}|} > 1$ and (53) to compare the ratio of (51) for j and j', there exists some $t_0 \quad 0$ such that for any $t \quad t_0$, we have

$$\frac{\frac{d}{dt}[\ (t)]_j}{|\frac{d}{dt}[\ (t)]_{j'}|} = \frac{|[\ (t)]_j|}{|[\ (t)]_{j'}|} > 1. \tag{54}$$

This implies that the ratio of the absolute value of time derivative $\frac{d}{dt}[\ (t)]_j$ to the absolute value of current value of $[\ (t)]_j$ is strictly bigger than that of $[\ (t)]_{j'}$. Moreover, by the definition of $r(t)$, $\frac{d}{dt}[\ (t)]_j$ does not change sign over time. Since this holds for all t_0, (54) results in a growth of $|[\ (t)]_j|$ that is exponentially faster than that of $|[\ (t)]_{j'}|$, so $[\ (t)]_j$ becomes a dominant component in (t) as $t \to 1$. This contradicts that $[\ ^{1}\ ]_j = 0$, hence the condition (50) has to be satisfied.

So far, we have characterized some conditions (47), (49), (50) that have to be satisfied by the limit direction w^1 of w. We now consider the following optimization problem and prove that these conditions are in fact the KKT conditions of the optimization problem. Consider

$$\underset{z \in \mathbb{R}^m}{\text{minimize}} \quad \|z\|_{k_1} \quad \text{subject to} \quad y z^T x \geq 1. \tag{55}$$

The KKT condition of this problem is

$$0 \in \partial \|z\|_{k_1} - y x,$$

where the global subdifferential $\partial \|z\|_{k_1}$ is defined as

$$\partial \|z\|_{k_1} = \{u \in \mathbb{R}^m \mid |[u]_j| \leq 1 \text{ for all } j \in [m], \text{ and } [z]_j \neq 0 \Rightarrow [u]_j = \operatorname{sign}([z]_j)\}.$$

We can check from (47), (49), (50) that the that w^1 satisfies the KKT condition up to scaling.

Now, how do we characterize v_1^1 and v_2^1 in terms of w^1? Let $w_1^1 := U_1^T v_1^1$ and $w_2^1 := U_2^T v_2^1$. Then, $v_l^1 = U_l w_l^1 = U_l U_l^T v_l^1$ holds because any component orthogonal to the column space of U_l stays unchanged while the component in the column space of W_l diverges to infinity. By (42), $|w_1^1| = |w_2^1| = |w^1|^{1/2}$. By (45), we have $\operatorname{sign}(w_1^1) = \operatorname{sign}(y) \operatorname{sign}(w_2^1)$.

E.2 Proof of Corollary 4

The proof of Corollary 4 boils down to characterizing the SVD of $M_{conv}(x)$.

E.2.1 The $k_1 = 1$ case

First, it is straightforward to check that for $m = 2$ and $k_1 = 1$, we have

$$w_{conv}(w_{conv}) = v_1 v_2.$$

For $k_1 = 1$, the data tensor is simply $M_{conv}(x) = x^T$. Thus, we have $U_1 = 1$, $U_2 = \frac{x}{\|x\|_{k_2}}$, and $s = \|x\|_{k_2}$. Substituting U_1 and U_2 to the theorem gives the condition on initial directions in Corollary 4. Also, the theorem implies us that the limit direction of v_2 satisfies $v_2^1 \propto y v_1^1 x$. Using this, it is easy to check that

$$w_{conv}(w_{conv}^1) \propto v_1^1 v_2^1 \propto yx.$$

E.2.2 The $k_1 = 2$ case

First, it is straightforward to check that for $m = 2$ and $k_1 = 2$, we have

$$w_{conv}(w_{conv}) = \begin{bmatrix} [v_1]_1 & 0 & 0 & 0 & [v_1]_2 \\ [v_1]_2 & [v_1]_1 & 0 & 0 & 0 \\ 0 & [v_1]_2 & [v_1]_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & [v_1]_1 & 0 \\ 0 & 0 & 0 & [v_1]_2 & [v_1]_1 \end{bmatrix} v_2. \tag{56}$$

For $k_1 = 2$, by definition, the data tensor is

$$M_{conv}(x) = \begin{bmatrix} x^T \\ x^T \end{bmatrix},$$

and it is straightforward to check that the SVD of this matrix is

$$M_{conv}(x) = \begin{bmatrix} x^T \\ x^T \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{\|x\|_{k_2}^2 + x^T x} & 0 \\ 0 & \sqrt{\|x\|_{k_2}^2 - x^T x} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \frac{x^T + x^T}{\|x\|_{k_2}^2 + x^T x} & \frac{1}{\sqrt{2}} \frac{x^T - x^T}{\|x\|_{k_2}^2 + x^T x} \\ \frac{1}{\sqrt{2}} \frac{x^T - x^T}{\|x\|_{k_2}^2 - x^T x} & \frac{1}{\sqrt{2}} \frac{x^T + x^T}{\|x\|_{k_2}^2 - x^T x} \end{bmatrix};$$

so

$$U_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}; U_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \frac{x + x}{\|x\|_{k_2}^2 + x^T x} & \frac{1}{\sqrt{2}} \frac{x - x}{\|x\|_{k_2}^2 - x^T x} \end{bmatrix}; s = \begin{bmatrix} \sqrt{\|x\|_{k_2}^2 + x^T x} \\ \sqrt{\|x\|_{k_2}^2 - x^T x} \end{bmatrix}.$$

Substituting U_1 and U_2 to the theorem gives the conditions on initial directions. Also, note that the maximum singular value depends on the sign of $x^T x$. Consider the optimization problem in the theorem statement:

$$\text{minimize}_{z \in \mathbb{R}^m} \quad \|z\|_1 \quad \text{subject to} \quad y z^T x \geq 1.$$

If $x^T x > 0$, then the solution z^* to this problem is in the direction of $[y \; 0]$. Therefore, the limit directions v_1^1 and v_2^1 will be of the form

$$v_1^1 \propto c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad v_2^1 \propto c_2 (x + \Delta x),$$

where $\text{sign}(c_1)\,\text{sign}(c_2) = \text{sign}(y)$. Using (56), it is straightforward to check that

$$\text{conv}\left(\begin{bmatrix} 1 \\ \text{conv} \end{bmatrix} \right) \propto y \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix} (x + \Delta x) = y(2x + \Delta x + \Delta x).$$

Similarly, if $x^T x < 0$, then the solution z^* is in the direction of $[0 \; y]$. Using (56), we have

$$\text{conv}\left(\begin{bmatrix} 1 \\ \text{conv} \end{bmatrix} \right) \propto y \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix} (x - \Delta x) = y(2x - \Delta x - \Delta x).$$

F Proofs of Theorem 5, Corollaries 5, 6 & 7, and Lemma 4

F.1 Proof of Lemma 4

In this subsection, we restate Lemma 4 and prove it.

Lemma 4. Consider the system of ODEs, where $p, q : \mathbb{R} \to \mathbb{R}$:

$$\dot{p} = p^{L-2} q, \quad \dot{q} = -p^{L-1}, \quad p(0) = 1, \quad q(0) = 0.$$

Then, the solutions $p_L(t)$ and $q_L(t)$ are continuous on their maximal interval of existence of the form $(-c, c) \subseteq \mathbb{R}$ for some $c \in (0, 1]$. Define $h_L(t) = p_L(t)^{L-1} q_L(t)$; then, $h_L(t)$ is odd and strictly increasing, satisfying $\lim_{t \to c} h_L(t) = 1$ and $\lim_{t \to -c} h_L(t) = -1$.

Proof First, continuity (and also continuous differentiability) of $p(t)$ and $q(t)$ is straightforward because the RHSs of the ODEs are differentiable in p and q. Next, define $\tilde{p}(t) = p(-t)$ and $\tilde{q}(t) = -q(-t)$. Then, one can show that $\tilde{p}$ and $\tilde{q}$ are also the solution of the ODE because

$$\frac{d}{dt}\tilde{p}(t) = \frac{d}{dt} p(-t) = -\dot{p}(-t) = -p(-t)^{L-2} q(-t) = -\tilde{p}(t)^{L-2} \tilde{q}(t),$$
$$\frac{d}{dt}\tilde{q}(t) = -\frac{d}{dt} q(-t) = -\dot{q}(-t) = p(-t)^{L-1} = \tilde{p}(t)^{L-1}.$$

However, by the Picard-Lindelöf theorem, the solution has to be unique; this means $p(t) = \tilde{p}(t) = p(-t)$ and $q(t) = \tilde{q}(t) = -q(-t)$, which proves that p is even and q is odd and also implies that the domain of p and q has to be of the form $(-c, c)$ (i.e. symmetric around the origin) and $h = p^{L-1} q$ is odd.

To show that h is strictly increasing, it suffices to show that p and q are both strictly increasing on $[0, c)$. To this end, we show that $\dot{p}(t) \geq 0$ for all $t \in [0, c)$. First, due to the initial condition $p(0) = 1$ and continuity of p, there exists $\varepsilon_1 > 0$ such that $p(t) > 0$ for all $t \in [0, \varepsilon_1) =: I_1$. This implies that $\dot{q}(t) = -p(t)^{L-1} < 0$ for $t \in I_1 \setminus \{0\}$, so q is strictly increasing on I_1. Since $q(0) = 0$, we have $q(t) < 0$ for $t \in I_1 \setminus \{0\}$, which then implies that $\dot{p}(t) = p(t)^{L-2} q(t) < 0$. Therefore, p is

also strictly increasing on $[0, \tau_1]$; this then means $p(t) \geq 1$ for $t \in [0, \tau_1]$ because $p(0) = 1$. Now, due to $p(\tau_1) \geq 1$ and continuity of p, there exists $\tau_2 > \tau_1$ such that $p(t) > 0$ for all $t \in [\tau_1, \tau_2) =: I_2$. Using the argument above for I_2 results in $p(t) \geq 1$ for $t \in [0, \tau_2]$. Repeating this until the end of the domain, we can show that $p(t) \geq 1$ holds for all $t \in [0, c)$. By $p \geq 1$, we have $q = p^{L-1} \geq 1$ on $[0, c)$, so q is strictly increasing on $[0, c)$. Also, $q(t) > 0$ on $(0, c)$, so $p = p^{L-2} q > 0$ on $(0, c)$ and p is also strictly increasing on $[0, c)$. This proves that h is strictly increasing on $[0, c)$, and also on $(-c, c)$ by oddity of h.

Finally, it is left to show $\lim_{t \to c} h(t) = 1$ and $\lim_{t \to -c} h(t) = -1$. If $c < \infty$, then this together with monotonicity implies that the limits hold. To see why, suppose $c < \infty$ and $\lim_{t \to c} h(t) < 1$. Then, p and q can be extended beyond c, which contradicts the fact that $(-c, c)$ is the maximal interval of existence of the solution. Next, consider the case $c = \infty$. From $p(t) \geq 1$, we have $q(t) \geq 1$ for $t \geq 0$. This implies that $q(t) \geq t$ for $t \geq 0$. Now, $\dot{p}(t) = p(t)^{L-2} q(t) \geq t$, which gives $p(t) \geq \frac{t^2}{2} + 1$ for $t \geq 0$. Therefore, we have

$$\lim_{t \to \infty} h(t) = \lim_{t \to \infty} p(t)^{L-1} q(t) \geq \lim_{t \to \infty} \left(\frac{t^2}{2} + 1 \right)^{L-1} t = \infty,$$

hence finishing the proof. □

F.2 Proof of Theorem 5

F.2.1 Convergence of loss to zero

We first show that given the conditions on initialization, the training loss $\Phi(t)$ converges to zero. Recall from Section 2.1 that

$$\dot{v}_l = -\nabla_{v_l} L(\Theta) = M(-X^T r) \cdot (v_1, \ldots, v_{l-1}, I_{k_l}, v_{l+1}, \ldots, v_L).$$

Applying the structure (9) in Assumption 1, we get

$$\begin{aligned}
\dot{v}_l &= M(-X^T r) \cdot (v_1, \ldots, v_{l-1}, I_{k_l}, v_{l+1}, \ldots, v_L) \\
&= \sum_{j=1}^{m} [-X^T r]_j \left(v_1^T [U_1]_{:,j} \cdots v_{l-1}^T [U_{l-1}]_{:,j} \cdot [U_l]_{:,j} \cdot v_{l+1}^T [U_{l+1}]_{:,j} \cdots v_L^T [U_L]_{:,j} \right) \\
&= \sum_{j=1}^{m} [-X^T r]_j \prod_{k \neq l} [U_k^T v_k]_j \cdot [U_l]_{:,j}.
\end{aligned}$$

Left-multiplying U_l^T to both sides, we get

$$U_l^T \dot{v}_l = -X^T r \odot \prod_{k \neq l} U_k^T v_k, \tag{57}$$

where $\prod^{\odot}$ denotes the product using entry-wise multiplication.

Now consider the rate of growth for the second power of the jth component of $U_l^T v_l$:

$$\frac{d}{dt} [U_l^T v_l]_j^2 = 2 [U_l^T \dot{v}_l]_j [U_l^T v_l]_j = -2 [X^T r]_j \prod_{k=1}^{L} [U_k^T v_k]_j = \frac{d}{dt} |[U_{l'}^T v_{l'}]_j|^2$$

for any $l' \in [L]$. Thus, for any $j \in [m]$, the second power of the jth components in $U_l^T v_l$ grow at the same rate for each layer $l \in [L]$. This means that the gap between any two different layers stays constant for all $t \geq 0$. Combining this with our conditions on initial directions, we have

$$[U_l^T v_l(t)]_j^2 - [U_L^T v_L(t)]_j^2 = [U_l^T v_l(0)]_j^2 - [U_L^T v_L(0)]_j^2 = \alpha^2 [\beta]_j^2 \leq \alpha^2;$$

for any $j \in [m]$, $l \in [L-1]$, and $t \geq 0$. This inequality also implies

$$[U_l^T v_l(t)]_j^2 \leq [U_L^T v_L(t)]_j^2 + \alpha^2 \leq \alpha^2. \tag{58}$$

Let us now consider the time derivative of $L(\ (t))$. We have the following chain of upper bounds on the time derivative:

$$\begin{aligned}
\frac{d}{dt}L(\ (t)) &= r\ \ L(\ (t))^T \dot{\ }(t) = \ kr\ \ L(\ (t))k_2^2 \\
&\quad kr\ \ _{v_L} L(\ (t))k_2^2 = k\ v_\perp(t)k_2^2 \\
&\overset{(a)}{=} k\ U_L^T v_\perp(t)k_2^2 \overset{(b)}{=} \ \ SX^T r(t) \ \ \prod_{k=L}^{1} U_k^T v_k(t) \ \ _2^2 \\
&= \sum_{j=1}^{m} [SX^T r(t)]_j^2 \prod_{k=L}^{1} [U_k^T v_k(t)]_j^2 \\
&\overset{(c)}{\ } {}^{2L\ 2\ L\ 1} \sum_{j=1}^{m} [SX^T r(t)]_j^2 \\
&= \ {}^{2L\ 2\ L\ 1} kSX^T r(t)k_2^2 \\
&\overset{(d)}{\ } {}^{2L\ 2\ L\ 1} s_{min}(S)^2 s_{min}(X)^2 kr(t)k_2^2; \\
&= \ 2\ {}^{2L\ 2\ L\ 1} s_{min}(S)^2 s_{min}(X)^2 L(\ (t)); \qquad (59)
\end{aligned}$$

where (a) used the fact that $kv_\perp(t)k_2^2 \ \ k\ U_L U_L^T v_\perp(t)k_2^2$ because it is a projection onto a subspace, and $kU_L U_L^T v_\perp(t)k_2^2 = kU_L^T v_\perp(t)k_2^2$ because $U_L^T U_L = I_{k_L}$; (b) is due to (57); (c) is due to (58); and (d) used the fact that $S \ 2\ R^{m\ \ d}$ and $X^T \ 2\ R^{d\ \ n}$ are matrices that have full column rank, so for any $z \ 2\ C^n$, we can use $kSX^T zk_2 \ \ s_{min}(S)s_{min}(X)kzk_2$ where $s_{min}(\ )$ denotes the minimum singular value of a matrix.

From (59), we get

$$L(\ (t)) \ \ L\ (\ (0)) \exp(\ \ 2\ {}^{2L\ 2\ L\ 1} s_{min}(S)^2 s_{min}(X)^2 t); \qquad (60)$$

so that $L(\ (t)) \ ! \ \ 0$ as $t \ ! \ 1$.

F.2.2 Characterizing the limit point

Now, we move on to characterize the limit points of the gradient ow. First, by de ning a "transformed" version of the parameters $_l(t) := U_l^T v_l(t)$ and using (57), one can de ne an equivalent system of ODEs:

$$\begin{aligned}
\dot{\ }_\perp &= \ \ SX^T r \prod_{k=l}^{1} \ _k \ \text{for } l \ 2\ [L]; \\
\ _l(0) &= \ \ \text{for } l \ 2\ [L \ \ 1]; \quad \ _L(0) = \ 0: \qquad (61)
\end{aligned}$$

Using Lemma 4, it is straightforward to verify that the solution to (61) has the following form. For odd L, we have

$$\begin{aligned}
\ _l(t) &= \ \ p_L \ \ {}^{L\ 2} j\ j^{\ L\ 2} \ \ SX^T \int_0^t r(\ )d \ \ \text{for } l \ 2\ [L \ \ 1]; \\
\ _L(t) &= \ \ j\ j\ \ q_L \ \ {}^{L\ 2} j\ j^{\ L\ 2} \ \ SX^T \int_0^t r(\ )d \ \ : \qquad (62)
\end{aligned}$$

Similarly, for even L, the solution for (61) satis es

$$\begin{aligned}
\ _l(t) &= \ \ p_L \ \ {}^{L\ 2\ \ L\ 2} \ \ SX^T \int_0^t r(\ )d \ \ \text{for } l \ 2\ [L \ \ 1]; \\
\ _L(t) &= \ \ q_L \ \ {}^{L\ 2\ \ L\ 2} \ \ SX^T \int_0^t r(\ )d \ \ : \qquad (63)
\end{aligned}$$

Now that we know how the solutions look like, let us see how these relate to the linear coef cients of the network. By Assumption 1, we have

$$f(x; \ ) = M(x) \ \ (v_1; : : : ; v_L) = \sum_{j=1}^{m} [Sx]_j \prod_{l=1}^{L} [U_l^T v_l]_j$$

$$= \sum_{j=1}^{m} \Big(\prod_{l=1}^{L} [\ _l]_j \Big) [S]_{j;}, \quad x = x^T S^T \prod_{l \in [L]} \ _l = x^T S^T :$$

Here, we de ned $:= \prod_{l \in [L]} \ _l \ 2 \ R^m$. Therefore, the linear coef cients of the network can be written as $(\ (t)) = S^T \ (t)$. From the solutions (62) and (63), we can write

$$(t) = \prod_{i=1}^{L} \ _l(t) = \ ^{L} j \ j^{\ L} \ h_L \ \ ^{L-2} j \ j^{\ L-2} \ SX^T \int_0^t r(\ )d \ ;$$

where $h_L := p_L^{L-1} q_L$, de ned in Lemma 4. By the convergence of the loss to zero (60), we have $\lim_{t!1} X \ (\ (t)) = y$. Therefore,

$$\underbrace{XS^T \ ^{L} j \ j^{\ L} \ h_L \ \ ^{L-2} j \ j^{\ L-2} \ SX^T}_{=: \ ^1} \int_0^1 r(\ )d \ = y: \tag{64}$$

Next, we will show that ^{1} is in fact the solution of the following optimization problem

$$\underset{2 R^m}{\text{minimize}} \quad Q_{L; \ ;} \ (\ ) \quad \text{subject to} \quad XS^T \ = y; \tag{65}$$

where $Q_{L; \ ;} \ : R^m \ ! \ R$ is a norm-like function de ned using $_L(t) := \int_0^t h_L^{\ 1}(\ )d \ :$

$$Q_{L; \ ;} \ (\ ) = \ ^2 \sum_{j=1}^{m} [\ ]_j^2 H_L \ \frac{[\ ]_j}{^{L} j[\ ]_j j^{L}} \ :$$

Note that the KKT conditions for (65) are

$$XS^T \ = y; \quad r \ Q_{L; \ ;} \ (\ ) = SX^T \ ;$$

for some $2 \ R^n$. It is clear from (64) that ^{1} satis es the rst condition (primal feasibility), so let us check the other one. Through a straightforward calculation, we get

$$r \ Q_{L; \ ;} \ (\ ) = \ ^{2 \ L} j \ j^{\ 2 \ L} \ h_L^{\ 1} \ \ ^{L} j \ j^{\ (\ L)} \ :$$

Equating this with SX^T gives

$$\begin{aligned} & ^{2 \ L} j \ j^{\ 2 \ L} \ h_L^{\ 1} \ \ ^{L} j \ j^{\ (\ L)} \ = SX^T \\ , \ & h_L^{\ 1} \ \ ^{L} j \ j^{\ (\ L)} \ = \ ^{L-2} j \ j^{\ L-2} \ SX^T \\ , \ & \ = \ ^{L} j \ j^{\ L} \ h_L \ \ ^{L-2} j \ j^{\ L-2} \ SX^T \ : \end{aligned}$$

Hence, by setting $= \int_0^1 r(\ )d \ , \ ^1$ satis es this condition as well. Also, since S is invertible, we can substitute $= S^T z$ to (65) to get the last statement of the theorem. This nishes the proof.

F.3 Proof of Corollary 5

The proof is a direct consequence of the fact that Assumption 1 holds with $U_1 = \ = U_L = I_d$ for linear diagonal networks. Hence, the proof is the same as Corollary 2, proved in Appendix D.2.

F.4 Proof of Corollary 6

We start by showing the DFT of a real and even vector is also real and even. Suppose that $x 2 R^d$ is real and even. First,

$$[F x]_j = p\frac{1}{d} \sum_{k=1}^{d} [x]_k \exp \ \frac{p\overline{\ 1} \ 2 \ (j \ \ 1)(k \ \ 1)}{d}$$

$$= \frac{1}{\sqrt{d}} \sum_{k=1}^{d} [x]_k \cos\left(\frac{2\pi(j-1)(k-1)}{d}\right) + \frac{\sqrt{-1}}{\sqrt{d}} \sum_{k=1}^{d} [x]_k \sin\left(\frac{2\pi(j-1)(k-1)}{d}\right)$$
$$= \frac{1}{\sqrt{d}} \sum_{k=1}^{d} [x]_k \cos\left(\frac{2\pi(j-1)(k-1)}{d}\right) \quad \forall j \in [d].$$

To prove that Fx is even, for $j = 0, \ldots, \lfloor \frac{d}{2} \rfloor$, we have

$$\begin{aligned}
[Fx]_{j+2} &= \frac{1}{\sqrt{d}} \sum_{k=1}^{d} [x]_k \cos\left(\frac{2\pi(j+1)(k-1)}{d}\right) \\
&= \frac{1}{\sqrt{d}} \sum_{k=1}^{d} [x]_k \cos\left(2\pi(k-1) - \frac{2\pi(j+1)(k-1)}{d}\right) \\
&= \frac{1}{\sqrt{d}} \sum_{k=1}^{d} [x]_k \cos\left(\frac{2\pi(d-j-1)(k-1)}{d}\right) \\
&= \frac{1}{\sqrt{d}} \sum_{k=1}^{d} [x]_k \cos\left(\frac{2\pi(d-j-1)(k-1)}{d}\right) \\
&= [Fx]_{d-j}.
\end{aligned}$$

It is proved in Appendix D.3 that linear full-length convolutional networks ($k_1 = \cdots = k_L = d$) satisfy Assumption 1 with $S = d^{\frac{L-1}{2}} F$ and $U_1 = \cdots = U_{L-1} = F$, where $F \in \mathbb{C}^{d \times d}$ is the matrix of discrete Fourier transform basis, $[F]_{j,k} = \frac{1}{\sqrt{d}} \exp(\frac{\sqrt{-1}\, 2\pi(j-1)(k-1)}{d})$ and $\bar{F}$ is the complex conjugate of F.

The proof of convergence of loss to zero in Appendix F.2.1 is written for real matrices $S, U_1, \ldots, U_L$, but we can actually apply the same argument as in Appendix D.1.1 and prove that the loss converges to zero, even in the case where $S, U_1, \ldots, U_L$ are complex.

Next, since U_l's are complex, we can write the system of ODE as (see (20) for its derivation)

$$\dot{w}_l = d^{\frac{L-1}{2}} \bar{F} X^T r \prod_{k=l+1}^{L} \bar{F} w_k, \tag{66}$$

Since all data points x_i and initialization $w_l(0)$ are real and even, we have that $\bar{F} X^T r$ is real and even, and $\bar{F} w_l(0) = F w_l(0)$'s are real and even. By (66), we see that the time derivatives of w_l are also real and even. Thus, the parameters $w_l(t)$ are all real and even for all $t \geq 0$. From this observation, we can define $w_l(t) := \bar{F} w_l(t)$, $\bar{X} := \bar{F} X$, and $S := d^{\frac{L-1}{2}} \mathrm{Re}(F)$, which are all real by the even symmetry. Then, starting from (61), the proof goes through.

F.5 Proof of Corollary 7

Since the sensor matrices $A_1, \ldots, A_n$ commute, they are simultaneously diagonalizable with a real unitary matrix $U \in \mathbb{R}^{d \times d}$, i.e., $U^T A_i U$'s are diagonal matrices. From the deep matrix sensing problem (13), we can compute $\nabla_{W_l} L_{ms}$, which gives the gradient flow dynamics of W_l:

$$\dot{W}_l = -\nabla_{W_l} L_{ms} = -W_{l-1}^T \cdots W_1^T \Big(\sum_{i=1}^{n} r_i A_i\Big) W_L^T \cdots W_{l+1}^T,$$

where $r_i = \langle A_i, W_1 \cdots W_L \rangle - y_i$ is the residual for the i-th sensor matrix. If we left-multiply U^T and right-multiply U to both sides, we get

$$U^T \dot{W}_l U = -U^T W_{l-1}^T \cdots U^T W_1^T U \Big(\sum_{i=1}^{n} r_i U^T A_i U\Big) U^T W_L^T U \cdots U^T W_{l+1}^T U. \tag{67}$$

If $U^T W_k^T U$ is a diagonal matrix for all $k \neq l$, then $U^T \dot{W}_l U$ is also a diagonal matrix. Note also that, since $W_l(0) = \alpha I_d = \alpha U U^T$ for $l \in [L-1]$, the product $U^T W_l U$ is a diagonal matrix at initialization. These observations imply that $W_l(t)$'s are all diagonalizable with U for all $t \geq 0$.

Now, define $v_l(t) = \mathrm{eig}(W_l(t))$, i.e., $U^T W_l U = \mathrm{diag}(v_l)$. Also, let $x_i = \mathrm{eig}(A_i)$. Then, (67) can be written as

$$v_l = \Big(\sum_{i=1}^{n} r_i x_i\Big) \prod_{k \neq l} v_k.$$

Therefore, this is equivalent to the regression problem with linear diagonal networks, initialized at $v_l(0) = \alpha^{-1}$ for $l \in [L-1]$ and $v_L(0) = 0$. Given this equivalence, Corollary 7 can be implied from Corollary 5.

G Proof of Theorem 6

G.1 Convergence of loss to zero

We first show that given the conditions on initialization, the training loss $L(\beta(t))$ converges to zero. Since $L = 2$ and $M(x) = U_1 \mathrm{diag}(s) U_2^T$, we can write the gradient flow dynamics from Section 2.1 as

$$\begin{aligned} \dot{v}_1 &= -M(X^T r) \cdot (I_{k_1}; v_2) = -r U_1 \mathrm{diag}(s) U_2^T v_2; \\ \dot{v}_2 &= -M(X^T r) \cdot (v_1; I_{k_2}) = -r U_2 \mathrm{diag}(s) U_1^T v_1; \end{aligned} \tag{68}$$

where $r(t) = f(x; \beta(t)) - y$ is the residual of the data point (x, y). From (68) we get

$$U_l^T \dot{v}_1 = -r s \cdot U_2^T v_2; \quad U_2^T \dot{v}_2 = -r s \cdot U_1^T v_1. \tag{69}$$

Now consider the rate of growth for the jth component of $U_1^T v_1$ squared:

$$\frac{d}{dt}[U_1^T v_1]_j^2 = 2[U_1^T v_1]_j [U_1^T \dot{v}_1]_j = -2r[s]_j [U_1^T v_1]_j [U_2^T v_2]_j = \frac{d}{dt}[U_2^T v_2]_j^2.$$

So for any $j \in [m]$, $[U_1^T v_1]_j^2$ and $[U_2^T v_2]_j^2$ grow at the same rate. This means that the gap between the two layers stays constant for all $t \geq 0$. Combining this with our conditions on initial directions,

$$\begin{aligned} [U_1^T v_1(t)]_j^2 - [U_2^T v_2(t)]_j^2 &= [U_1^T v_1(0)]_j^2 - [U_2^T v_2(0)]_j^2 \\ &= \alpha^{-2}[U_1^T v_1]_j^2 - \alpha^{-2}[U_2^T v_2]_j^2 = 0; \end{aligned}$$

for any $j \in [m]$ and $t \geq 0$. This inequality implies

$$[U_1^T v_1(t)]_j^2 - [U_2^T v_2(t)]_j^2 \leq \alpha^{-2} - \alpha^{-2} = 0. \tag{70}$$

Let us now consider the time derivative of $L(\beta(t))$. We have the following chain of upper bounds on the time derivative:

$$\begin{aligned} \frac{d}{dt} L(\beta(t)) &= -r \cdot \nabla L(\beta(t))^T \dot{\beta}(t) = -\|r \cdot \nabla L(\beta(t))\|_2^2 \\ &\leq -\|r \cdot \nabla_{v_2} L(\beta(t))\|_2^2 = -\|r \dot{v}_2(t)\|_2^2 \\ &\overset{(a)}{=} -\|r U_2^T \dot{v}_2(t)\|_2^2 \overset{(b)}{=} -\|r(t)^2 \cdot s \cdot U_1^T v_1(t)\|_2^2 \\ &= -r(t)^2 \sum_{j=1}^{m} [s]_j^2 [U_1^T v_1(t)]_j^2 \\ &\overset{(c)}{\leq} -\alpha^2 r(t)^2 \sum_{j=1}^{m} [s]_j^2 \\ &= -2\alpha^2 \|s\|_2^2 L(\beta(t)); \end{aligned}$$

where (a) used the fact that $\|\dot{v}_2(t)\|_2^2 = \|U_2 U_2^T \dot{v}_2(t)\|_2^2$ because it is a projection onto a subspace, and $\|U_2 U_2^T \dot{v}_1(t)\|_2^2 = \|U_2^T \dot{v}_2(t)\|_2^2$ because $U_2^T U_2 = I_{k_2}$; (b) is due to (69); (c) is due to (70). From this, we get

$$L(\beta(t)) \leq L(\beta(0)) \exp(-2\alpha^2 \|s\|_2^2 t). \tag{71}$$

Therefore, $L(\beta(t)) \to 0$ as $t \to \infty$.

G.2 CHARACTERIZING THE LIMIT POINT

Now, we move on to characterize the limit points of the gradient flow. First, note that any changes made in v_l over time are in the subspace spanned by the columns of U_l. Therefore, any component in the initialization $v_l(0) = v_l$ that is orthogonal to the column space of U_l stays constant.

So, we can focus on the evolution of v_l in the column space of U_l; this can be done by defining a "transformed" version of the parameters $\alpha_l(t) := U_l^T v_l(t)$ and using (69), one can define an equivalent system of ODEs:

$$\begin{aligned} \dot{\alpha}_1 &= -r s_1 \alpha_2; \quad \dot{\alpha}_2 = -r s_1 \alpha_1; \\ \alpha_1(0) &= \alpha_{1,0}; \quad \alpha_2(0) = \alpha_{2,0}; \end{aligned} \tag{72}$$

where $\alpha_{1,0} := U_1^T v_{1,0}$, $\alpha_{2,0} := U_2^T v_{2,0}$. It is straightforward to verify that the solution to (72) has the following form.

$$\begin{aligned} \alpha_1(t) &= \alpha_{1,0} \cosh\left(s \int_0^t r(\tau)d\tau\right) + \alpha_{2,0} \sinh\left(s \int_0^t r(\tau)d\tau\right); \\ \alpha_2(t) &= \alpha_{1,0} \sinh\left(s \int_0^t r(\tau)d\tau\right) + \alpha_{2,0} \cosh\left(s \int_0^t r(\tau)d\tau\right): \end{aligned} \tag{73}$$

By the convergence of the loss to zero (71), we have $\lim_{t\to\infty} f(x; \theta(t)) = y$. Note that $f(x; \theta(t))$ can be written as

$$\begin{aligned} f(x; \theta(t)) &= M(x)(v_1(t), v_2(t)) = v_1(t)^T M(x) v_2(t) \\ &= v_1(t)^T U_1 \operatorname{diag}(s) U_2^T v_2(t) = s^T(\alpha_1(t) \odot \alpha_2(t)): \end{aligned}$$

Therefore,

$$\begin{aligned} &\lim_{t\to\infty} f(x; \theta(t)) = \lim_{t\to\infty} s^T(\alpha_1(t) \odot \alpha_2(t)) \\ &= \tfrac{1}{2} s^T (\alpha_{1,0}^2 + \alpha_{2,0}^2) \cosh\left(s \int_0^\infty r(\tau)d\tau\right) \sinh\left(s \int_0^\infty r(\tau)d\tau\right) \\ &\quad + (\alpha_{1,0} \odot \alpha_{2,0}) \cosh^2\left(s \int_0^\infty r(\tau)d\tau\right) + \sinh^2\left(s \int_0^\infty r(\tau)d\tau\right) \\ &= \tfrac{1}{2} s^T \frac{\alpha_{1,0}^2 + \alpha_{2,0}^2}{2} \sinh\left(2s \int_0^\infty r(\tau)d\tau\right) + (\alpha_{1,0} \odot \alpha_{2,0}) \cosh\left(2s \int_0^\infty r(\tau)d\tau\right) \\ &= \tfrac{1}{2} \sum_{j=1}^m [s]_j \left(\frac{[\alpha_{1,0}]_j^2 + [\alpha_{2,0}]_j^2}{2} \sinh(2[s]_j \infty) + [\alpha_{1,0}]_j [\alpha_{2,0}]_j \cosh(2[s]_j \infty) \right) \\ &= y; \end{aligned} \tag{74}$$

where we defined $\infty := \int_0^\infty r(\tau)d\tau$. Consider the function $\gamma \mapsto a\sinh(\gamma) + b\cosh(\gamma)$. This is a strictly increasing function if $a > |b|$. Note also that

$$\frac{[\alpha_{1,0}]_j^2 + [\alpha_{2,0}]_j^2}{2} > |[\alpha_{1,0}]_j [\alpha_{2,0}]_j|; \tag{75}$$

which holds with equality if and only if $|[\alpha_{1,0}]_j| = |[\alpha_{2,0}]_j|$. However, recall from our assumptions on initialization that $[\alpha_{1,0}]_j^2 - [\alpha_{2,0}]_j^2 > 0$, so (75) can only hold with strict inequality. Therefore,

$$g(\gamma) := \sum_{j=1}^m [s]_j \left(\frac{[\alpha_{1,0}]_j^2 + [\alpha_{2,0}]_j^2}{2} \sinh(2[s]_j \gamma) + [\alpha_{1,0}]_j [\alpha_{2,0}]_j \cosh(2[s]_j \gamma) \right)$$

is a strictly increasing (hence invertible) function because it is a sum of strictly increasing functions. Using this $g(\gamma)$, (74) can be written as $\frac{1}{2} g(\gamma) = y$, and by using the inverse of g, we have

$$\gamma = \int_0^\infty r(\tau)d\tau = g^{-1}\left(\frac{y}{2}\right): \tag{76}$$

Plugging (76) into (73), we get

$$\begin{aligned}
&\lim_{t\to\infty} v_1(t) \\
&= U_1 \lim_{t\to\infty} \alpha_1(t) + (I_{k_1} - U_1 U_1^T) v_1 \\
&= U_1 \alpha_1 \cosh\left(\gamma^{-1} \frac{y}{2} s\right) + \alpha_2 \sinh\left(\gamma^{-1} \frac{y}{2} s\right) + (I_{k_1} - U_1 U_1^T) v_1; \\
&\lim_{t\to\infty} v_2(t) \\
&= U_2 \lim_{t\to\infty} \alpha_2(t) + (I_{k_2} - U_2 U_2^T) v_2 \\
&= U_2 \alpha_1 \sinh\left(\gamma^{-1} \frac{y}{2} s\right) + \alpha_2 \cosh\left(\gamma^{-1} \frac{y}{2} s\right) + (I_{k_2} - U_2 U_2^T) v_2.
\end{aligned}$$

This finishes the proof.

H Proof of Theorem 7

H.1 Convergence of loss to zero

We first show that given the conditions on initialization, the training loss $L(t)$ converges to zero. Recall from (10) that the linear fully-connected network can be written as

$$f_{fc}(x; \theta_{fc}) = x^T W_1 W_2 \cdots W_{L-1} w_L.$$

From the definition of the training loss, it is straightforward to check that the gradient flow dynamics read

$$\begin{aligned}
\dot{W}_l &= -\nabla_{W_l} L(\theta_{fc}) = -W_{l-1}^T \cdots W_1^T X^T r w_L^T W_{L-1}^T \cdots W_{l+1}^T \quad \text{for } l \in [L-1]; \\
\dot{w}_L &= -\nabla_{w_L} L(\theta_{fc}) = -W_{L-1}^T \cdots W_1^T X^T r; \\
W_l(0) &= W_l \quad \text{for } l \in [L-1]; \\
w_L(0) &= w_L;
\end{aligned} \tag{77}$$

where $r \in \mathbb{R}^n$ is the residual vector satisfying $[r]_i = f_{fc}(x_i; \theta_{fc}) - y_i$, as defined in Section 2.1. From (77), we have

$$\begin{aligned}
W_l^T \dot{W}_l &= \dot{W}_{l+1} W_{l+1}^T = -W_l^T \cdots W_1^T X^T r w_L^T W_{L-1}^T \cdots W_{l+1}^T; \\
\dot{W}_l^T W_l &= W_{l+1} \dot{W}_{l+1}^T = -W_{l+1} \cdots W_{L-1} w_L r^T X W_{l-1} \cdots W_l;
\end{aligned}$$

for any $l \in [L-2]$. From this, we have

$$\frac{d}{dt} W_l^T W_l = \frac{d}{dt} W_{l+1} W_{l+1}^T;$$

and thus

$$\begin{aligned}
W_l(t)^T W_l(t) - W_{l+1}(t) W_{l+1}(t)^T &= W_l(0)^T W_l(0) - W_{l+1}(0) W_{l+1}(0)^T \\
&= \alpha^2 W_l^T W_l - \alpha^2 W_{l+1} W_{l+1}^T;
\end{aligned} \tag{78}$$

for any $l \in [L-2]$. Similarly, we have

$$\begin{aligned}
W_{L-1}(t)^T W_{L-1}(t) - w_L(t) w_L(t)^T &= W_{L-1}(0)^T W_{L-1}(0) - w_L(0) w_L(0)^T \\
&= \alpha^2 W_{L-1}^T W_{L-1} - \alpha^2 w_L w_L^T.
\end{aligned} \tag{79}$$

Let us now consider the time derivative of $L(\theta_{fc}(t))$. We have the following chain of upper bounds on the time derivative:

$$\begin{aligned}
\frac{d}{dt} L(\theta_{fc}(t)) &= -\nabla_{\theta_{fc}} L(\theta_{fc}(t))^T \dot{\theta}_{fc}(t) = -\gamma \nabla_{\theta_{fc}} L(\theta_{fc}(t)) \|^2 \\
&\le -\gamma \nabla_{W_L} L(\theta_{fc}(t)) \|_2^2 = -\gamma \|\dot{w}_L(t)\|_2^2 \\
&= -\gamma \|W_{L-1}^T \cdots W_1^T X^T r\|_2^2.
\end{aligned} \tag{80}$$

Note from (80) that if $W_{L-1}^T \cdots W_1^T$ is full-rank, its minimum singular value is positive, and one can bound

$$\|W_{L-1}^T \cdots W_1^T X^T r\|_2 \geq \min(W_{L-1}^T \cdots W_1^T)\|X^T r\|_2. \tag{81}$$

We now prove that the matrix $W_{L-1}^T \cdots W_1^T$ is full-rank, and its minimum singular value is bounded from below by $\alpha^{L-1}(L-1)^{-(L-2)}$ for any $t > 0$. To show this, it suffices to show that

$$W_{L-1}^T \cdots W_1^T W_1 \cdots W_{L-1} \geq \alpha^{2L-2} L^{-(L-1)} I_d. \tag{82}$$

Now,

$$\begin{aligned}
&W_{L-1}^T \cdots W_1^T W_1 W_2 \cdots W_{L-1} \\
\overset{(a)}{=}\ &W_{L-1}^T \cdots W_2^T (W_2 W_2^T + \alpha^2 W_1^T W_1 - \alpha^2 W_2 W_2^T) W_2 \cdots W_{L-1} \\
\overset{(b)}{\geq}\ &W_{L-1}^T \cdots W_3^T W_2^T W_2 W_2^T W_2 W_3 \cdots W_{L-1} \\
\overset{(a)}{=}\ &W_{L-1}^T \cdots W_3^T (W_3 W_3^T + \alpha^2 W_2^T W_2 - \alpha^2 W_3 W_3^T)^2 W_3 \cdots W_{L-1} \\
\overset{(b)}{\geq}\ &W_{L-1}^T \cdots W_3^T (W_3 W_3^T)^2 W_3 \cdots W_{L-1} \\
=\ &(W_{L-1}^T \cdots W_1^T W_1 \cdots W_{L-1})^{L-1};
\end{aligned}$$

where equalities marked in (a) used (78), and inequalities marked in (b) used the initialization conditions $W_l^T W_l \geq W_{l+1} W_{l+1}^T$. Next, it follows from (79) that

$$\begin{aligned}
(W_{L-1}^T \cdots W_1^T W_1 \cdots W_{L-1})^{L-1} &= (W_{L-1} \cdots W_1 W_1^T + \alpha^2 W_{L-1}^T \cdots W_1^T W_1 \cdots W_{L-1} - \alpha^2 W_{L-1} \cdots W_1 W_1^T \cdots W_{L-1}^T)^{L-1} \\
&\geq \alpha^{2L-2}(W_{L-1}^T \cdots W_1^T W_1 \cdots W_{L-1} - W_{L-1} \cdots W_1 W_1^T \cdots W_{L-1}^T)^{L-1} \\
&\overset{(c)}{\geq} \alpha^{2L-2} L^{-(L-1)} I_d.
\end{aligned}$$

where (c) used the assumption that $W_{L-1}^T \cdots W_1^T W_1 \cdots W_{L-1} - W_{L-1} \cdots W_1 W_1^T \cdots W_{L-1}^T \succeq -I_d$. This proves (82). Applying (82) to (80) then gives

$$\begin{aligned}
\frac{d}{dt} L(\beta_{fc}(t)) &\geq \|W_{L-1}^T \cdots W_1^T X^T r\|_2^2 \\
&\geq \min(W_{L-1}^T \cdots W_1^T)^2 \|X^T r\|_2^2 \\
&\geq \alpha^{2L-2} L^{-(L-1)} \|X^T r\|_2^2 \\
&\overset{(d)}{\geq} \alpha^{2L-2} L^{-(L-1)} \sigma_{\min}(X)^2 \|r\|_2^2 \\
&= \alpha^{2L-2} L^{-(L-1)} \sigma_{\min}(X)^2 L(\beta_{fc}(t));
\end{aligned}$$

where (d) used the fact that X^T is a full column rank matrix to apply a bound similar to (81). From this, we get

$$L(\beta_{fc}(t)) \leq L(\beta_{fc}(0)) \exp(-\alpha^{2L-2} L^{-(L-1)} \sigma_{\min}(X)^2 t);$$

hence proving $L(\beta_{fc}(t)) \to 0$ as $t \to \infty$.

H.2 Characterizing the limit point: $\alpha \to 0$ case

Now, we move on to characterize the limit points of the gradient flow, for the "active regime" case $\alpha \to 0$. This part of the proof is motivated from the analysis in Ji & Telgarsky (2019a).

Let u_l and v_l be the top left and right singular vectors of W_l, for $l \in [L-1]$. Note that since W_l varies over time, the singular vectors and singular value also vary over time. Similarly, let s be the largest singular value of W_l. We will show that the linear coefficients $\beta_{fc}(\beta_{fc}) = W_1 \cdots W_{L-1} w_L$ align with u_1 as $\alpha \to 0$, and u_1 is in the subspace of row(X) in the limit $\alpha \to 0$, hence proving that $\beta_{fc}(\beta_{fc})$ is the minimum ℓ_2 norm solution in the limit $\alpha \to 0$.

First, note from (78) and (79) that if we take trace of both sides, we get

$$\|W_l\|_F^2 - \|W_{l+1}\|_F^2 = \alpha^2(\|W_l\|_F^2 - \|W_{l+1}\|_F^2) \quad \text{for } l \in [L-2];$$

$$\|W_{L-1}\|_F^2 - \|w_L\|_2^2 = \alpha^2(\|W_{L-1}\|_F^2 - \|w_L\|_2^2).$$

Summing the equations above for $l, l+1, \ldots, L-1$, we get

$$\|W_l\|_F^2 - \|w_L\|_2^2 = \alpha^2(\|W_l\|_F^2 - \|w_L\|_2^2). \tag{83}$$

Next, consider the operator norms (i.e., the maximum singular values), denoted as $\|\cdot\|_2$, of the matrices.

$$\begin{aligned}
\|W_l\|_2^2 &= u_{l+1}^T W_l^T W_l u_{l+1} \\
&\overset{(e)}{=} u_{l+1}^T W_{l+1} W_{l+1}^T u_{l+1} + \alpha^2 u_{l+1}^T (W_l^T W_l - W_{l+1} W_{l+1}^T) u_{l+1} \\
&= \|W_{l+1}\|_2^2 + \alpha^2 u_{l+1}^T (W_l^T W_l - W_{l+1} W_{l+1}^T) u_{l+1} \\
&\geq \|W_{l+1}\|_2^2 - \alpha^2 \|W_l^T W_l - W_{l+1} W_{l+1}^T\|_2 \quad \text{for } l = 2, \ldots, L-1, \\
\|W_{L-1}\|_2^2 &= \frac{w_L}{\|w_L\|_2} W_{L-1}^T W_{L-1} \frac{w_L}{\|w_L\|_2} \\
&\overset{(f)}{=} \frac{w_L}{\|w_L\|_2} w_L w_L^T \frac{w_L}{\|w_L\|_2} + \alpha^2 \frac{w_L}{\|w_L\|_2} (W_{L-1}^T W_{L-1} - w_L w_L^T) \frac{w_L}{\|w_L\|_2} \\
&\geq \|w_L\|_2^2 - \alpha^2 \|W_{L-1}^T W_{L-1} - w_L w_L^T\|_2.
\end{aligned}$$

where (e) used (78) and (f) used (79). Summing the inequalities gives

$$\|W_l\|_2^2 - \|w_L\|_2^2 \geq -\alpha^2 \sum_{k=1}^{L-1} \|W_k^T W_k - W_{k+1} W_{k+1}^T\|_2. \tag{84}$$

From (83) and (84), we get a bound on the gap between the second powers of the Frobenius norm (or the ℓ_2 norm of singular values) and operator norm (or the maximum singular value s_l) of W_l:

$$\|W_l(t)\|_F^2 - \|W_l(t)\|_2^2 \leq \alpha^2(\|W_l\|_F^2 - \|w_L\|_2^2) + \alpha^2 \sum_{k=l}^{L-1} \|W_k^T W_k - W_{k+1} W_{k+1}^T\|_2, \tag{85}$$

which holds for any $t \geq 0$. The gap (85) implies that each W_l, for $l = 2, \ldots, L-1$, can be written as

$$W_l(t) = s_l(t) u_l(t) v_l(t)^T + O(\alpha^2). \tag{86}$$

Next, we show that the "adjacent" singular vectors v_l and u_{l+1} align with each other as $\alpha \to 0$. To this end, we will get lower and upper bounds for a quantity $v_l^T W_{l+1} W_{l+1}^T v_l$.

$$\begin{aligned}
v_l^T W_{l+1} W_{l+1}^T v_l &= v_l^T W_l^T W_l v_l - \alpha^2 v_l^T W_l^T W_l v_l + \alpha^2 v_l^T W_{l+1} W_{l+1}^T v_l \\
&\geq \|W_l\|_2^2 - \alpha^2 \|W_l^T W_l - W_{l+1} W_{l+1}^T\|_2 \\
&= s_l^2 - \alpha^2 \|W_l^T W_l - W_{l+1} W_{l+1}^T\|_2, \qquad (87) \\
v_l^T W_{l+1} W_{l+1}^T v_l &= v_l^T (s_{l+1}^2 u_{l+1} u_{l+1}^T + W_{l+1} W_{l+1}^T - s_{l+1}^2 u_{l+1} u_{l+1}^T) v_l \\
&= s_{l+1}^2 (v_l^T u_{l+1})^2 + v_l^T (W_{l+1} W_{l+1}^T - s_{l+1}^2 u_{l+1} u_{l+1}^T) v_l \\
&\leq s_{l+1}^2 (v_l^T u_{l+1})^2 + \|W_{l+1}\|_F^2 - \|W_{l+1}\|_2^2. \qquad (88)
\end{aligned}$$

Combining (87), (88), and (85), we get

$$\begin{aligned}
s_l^2 &\leq s_{l+1}^2 (v_l^T u_{l+1})^2 + \alpha^2 \|W_l^T W_l - W_{l+1} W_{l+1}^T\|_2 + \|W_{l+1}\|_F^2 - \|W_{l+1}\|_2^2 \\
&\leq s_{l+1}^2 (v_l^T u_{l+1})^2 + \alpha^2(\|W_{l+1}\|_F^2 - \|w_L\|_2^2) + \alpha^2 \sum_{k=l}^{L-1} \|W_k^T W_k - W_{k+1} W_{k+1}^T\|_2. \qquad (89)
\end{aligned}$$

Next, by a similar reasoning as (87), we have

$$s_l^2 \geq u_{l+1}^T W_l^T W_l u_{l+1} - s_{l+1}^2 - \alpha^2 \|W_l^T W_l - W_{l+1} W_{l+1}^T\|_2. \tag{90}$$

Combining (89) and (90) and dividing both sides by s_{l+1}^2, we get

$$(v_l(t)^T u_{l+1}(t))^2 \geq 1 - \alpha^2 \frac{G_l}{s_{l+1}(t)^2} \tag{91}$$

for t 0, where

$$G_l := W_l^T W_l \quad W_{l+1} W_{l+1}^T \;_2 + (W_{l+1} \;_{\mathrm{F}}^2 \quad k w_L k_2^2) + \sum_{k=l}^{L-1} k W_k^T W_k \quad W_{k+1} W_{k+1}^T k_2.$$

By a similar argument, we can also get

$$\frac{(v_{L-1}(t)^T w_L(t))^2}{k w_L(t) k_2^2} \quad 1 \quad \alpha^2 \frac{G_{L-1}}{k w_L(t) k_2^2}, \tag{92}$$

where

$$G_{L-1} := 2 \; W_{L-1}^T W_{L-1} \quad w_L w_L^T \;_2 .$$

From (91) and (92), we can note that as α ! 0, the inner product between the adjacent singular vectors converges to 1, unless $s_2, \ldots, s_{L-1}, k w_L k_2$ also diminish to zero. So it is left to show that the singular values do not diminish to zero as α ! 0. To this end, recall that we proved in the previous subsection that

$$\lim_{t\to\infty} X W_1(t) \quad W_{L-1}(t) w_L(t) = y.$$

A necessary condition for this to hold is that

$$\frac{k y k_2}{k X k_2} \quad \lim_{t\to\infty} k W_1(t) \quad W_{L-1}(t) w_L(t) k_2 \quad \lim_{t\to\infty} \prod_{l=1}^{L-1} s_l(t) \, k w_L(t) k_2 .$$

This means that after converging to the global minimum solution of the problem (i.e., t ! 1), the product of the singular values must be at least greater than some constant independent of α. Moreover, we can see from (87) and (90) that the gap between singular values squared of adjacent layers is bounded by $O(\alpha^2)$, for all t 0; so the maximum singular values become closer and closer to each other as α diminishes. This implies that

$$\lim_{\alpha\to 0} \lim_{t\to\infty} s_l(t) \quad \frac{k y k_2^{1/L}}{k X k_2^{1/L}} \quad \text{for } l \in [L \quad 1], \quad \lim_{\alpha\to 0} \lim_{t\to\infty} k w_L(t) k_2 \quad \frac{k y k_2^{1/L}}{k X k_2^{1/L}}.$$

Therefore, we have the alignment of singular vectors at convergence as α ! 0:

$$\lim_{\alpha\to 0} \lim_{t\to\infty} (v_l(t)^T u_{l+1}(t))^2 = 1, \quad \text{for } l \in [L \quad 2], \quad \lim_{\alpha\to 0} \lim_{t\to\infty} \frac{(v_{L-1}(t)^T w_L(t))^2}{k w_L(t) k_2^2} = 1. \tag{93}$$

So far, we saw from (86) that $W_l(t)$'s become rank-1 matrices as α ! 0, and from (93) that the top singular vectors align with each other as t ! 1 and α ! 0. These imply that, as t ! 1 and α ! 0, $_{\mathrm{fc}}(\;_{\mathrm{fc}})$ is a scalar multiple of the u_1, the top left singular vector of W_1:

$$\lim_{\alpha\to 0} \lim_{t\to\infty} \;_{\mathrm{fc}}(\;_{\mathrm{fc}}(t)) = c \; \lim_{\alpha\to 0} \lim_{t\to\infty} u_1(t), \tag{94}$$

for some $c \in \mathbb{R}$.

In light of (94), it remains to take a close look at $u_1(t)$. Note from the gradient flow dynamics of W_1 that W_1 is always a rank-1 matrix whose columns are in the row space of X, since $X^T r \in \mathrm{row}(X)$. This implies that, if we decompose W_1 into two orthogonal components $W_1^{\perp}$ and $W_1^{\parallel}$ so that the columns in $W_1^{\parallel}$ are in $\mathrm{row}(X)$ and the columns in $W_1^{\perp}$ are in the orthogonal subspace $\mathrm{row}(X)^{\perp}$, we have

$$W_1^{\perp} = 0, \quad W_1^{\parallel} = W_1.$$

That is, any component $W_1^{\perp}(0)$ orthogonal to $\mathrm{row}(X)$ remains unchanged for all t 0, while the component $W_1^{\parallel}$ changes by the gradient flow. Since we have

$$W_1^{\perp}(t) \;_{\mathrm{F}} = \quad W_1^{\perp}(0) \;_{\mathrm{F}} \quad \alpha \quad W_l \;_{\mathrm{F}},$$

the component in W_1 that is orthogonal to $\mathrm{row}(X)$ diminishes to zero as α ! 0. This means that at the limit α ! 0, the columns of W_1 are entirely from $\mathrm{row}(X)$, which also means that

$$\lim_{\alpha\to 0} \lim_{t\to\infty} \;_{\mathrm{fc}}(\;_{\mathrm{fc}}(t)) \in \mathrm{row}(X).$$

However, recall that there is only one unique global minimum of $Xz = y$ in $\mathrm{row}(X)$: namely, $z = X^T(X X^T)^{-1} y$, the minimum ℓ_2 norm solution. This finishes the proof.