Proceedings of Machine Learning Research vol 125:1–39, 2020 33rd Annual Conference on Learning Theory

Kernel and Rich Regimes in Overparametrized Models

Blake Woodworth BLAKE@TTIC.EDU
Toyota Technological Institute at Chicago

Suriya Gunasekar SURIYA@TTIC.EDU
Microsoft Research

Jason D. Lee JASONLEE@PRINCETON.EDU
Princeton University

Edward Moroshko EDWARD.MOROSHKO@GMAIL.COM
Technion

Pedro Savarese SAVARESE@TTIC.EDU
Toyota Technological Institute at Chicago

Itay Golan ITAYGOLAN@GMAIL.COM
Technion

Daniel Soudry DANIEL.SOUDRY@TECHNION.AC.IL
Technion

Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

A recent line of work studies overparametrized neural networks in the "kernel regime," *i.e.,* when during training the network behaves as a kernelized linear predictor, and thus, training with gradient descent has the effect of finding the corresponding minimum RKHS norm solution. This stands in contrast to other studies which demonstrate how gradient descent on overparametrized networks can induce rich implicit biases that are not RKHS norms. Building on an observation by Chizat et al. (2019), we show how the ***scale of the initialization*** controls the transition between the "kernel" (aka lazy) and "rich" (aka active) regimes and affects generalization properties in multilayer homogeneous models. We provide a complete and detailed analysis for a family of simple depth-D linear networks that exhibit an interesting and meaningful transition between the kernel and rich regimes, and highlight an interesting role for the *width* of the models. We further demonstrate this transition empirically for matrix factorization and multilayer non-linear networks.

1. Introduction

A string of recent papers study neural networks trained with gradient descent in the "kernel regime." They observe that, in a certain regime, networks trained with gradient descent behave as kernel methods (Jacot et al., 2018; Daniely, 2017; Yang, 2019). This allows one to prove convergence to zero error solutions in overparametrized settings (Li and Liang, 2018; Du et al., 2018, 2019; Allen-Zhu et al., 2018; Zou et al., 2018; Allen-Zhu et al., 2019;

Arora et al., 2019b; Chizat et al., 2019). This also implies that the learned function is the the minimum norm solution in the corresponding RKHS (Chizat et al., 2019; Arora et al., 2019b; Mei et al., 2019), and more generally that models inherit the inductive bias and generalization behavior of the RKHS. This suggests that, in a certain regime, deep models can be equivalently replaced by kernel methods with the "right" kernel, and deep learning boils down to a kernel method with a fixed kernel determined by the architecture and initialization, and thus it can only learn problems learnable by appropriate kernel.

This contrasts with other recent results that show how in deep models, including infinitely overparametrized networks, training with gradient descent induces an inductive bias that cannot be represented as an RKHS norm. For example, analytic and/or empirical results suggest that gradient descent on deep linear convolutional networks implicitly biases toward minimizing the L_p bridge penalty, for $p = 2/\text{depth} \leq 1$, in the frequency domain (Gunasekar et al., 2018); on an infinite width single input ReLU network infinitesimal weight decay biases towards minimizing the second order total variations $\int |f''(x)| dx$ of the learned function (Savarese et al., 2019), further, empirically it has been observed that this bias is implicitly induced by gradient descent without explicit weight decay (Savarese et al., 2019; Williams et al., 2019); and gradient descent on a overparametrized matrix factorization, which can be thought of as a two layer linear network, induces nuclear norm minimization of the learned matrix and can ensure low rank matrix recovery (Gunasekar et al., 2017; Li et al., 2018; Arora et al., 2019a). None of these natural inductive biases are Hilbert norms, and therefore they *cannot* be captured by any kernel. This suggests that training deep models with gradient descent can behave very differently from kernel methods, and have richer inductive biases.

So, does the kernel approximation indeed capture the behavior of deep learning in a relevant and interesting regime, or does the success of deep learning come from escaping this regime to have *richer* inductive biases that exploits the multilayer nature of neural networks? In order to understand this, we must first understand when each of these regimes hold, and how the transition between the "kernel regime" and the "rich regime" happens.

Early investigations of the kernel regime emphasize the number of parameters ("width") going to infinity as leading to this regime (see *e.g.,* (Jacot et al., 2018; Daniely, 2017; Yang, 2019)). However, Chizat et al. (2019) identified the *scale* of the model at initialization as a quantity controlling entry into the kernel regime. Their results suggest that for any number of parameters (any width), a homogeneous model can be approximated by a kernel when its scale at initialization goes to infinity (see the survey in Section 3). Considering models with increasing (or infinite) width, the relevant regime (kernel or rich) is determined by how the scaling at initialization behaves as the width goes to infinity. In this paper we elaborate and expand of this view, carefully studying how the scale of initialization affects the model behaviour for D-homogeneous models.

Our Contributions In Section 4 we analyze in detail a simple 2-homogeneous model for which we can exactly characterize the implicit bias of training with gradient descent as a function of the scale, α, of initialization. We show: (a) the implicit bias transitions from the ℓ_2 norm in the $\alpha \to \infty$ limit to ℓ_1 in the $\alpha \to 0$ limit; (b) consequently, for certain problems *e.g.,* high dimensional sparse regression, using a small initialization can be necessary for good generalization; and (c) we highlight how the "shape" of the initialization, *i.e.,* the relative scale of the parameters, affects the $\alpha \to \infty$ bias but *not* the $\alpha \to 0$ bias.

In Section 5 we extend this analysis to analogous D-homogeneous models, showing that the order of homogeneity or the "depth" of the model hastens the transition into the ℓ_1 regime. In Section 6, we analyze asymmetric matrix factorization models, and show that the "width" (*i.e.,* the inner dimension of the factorization) has an interesting role to play in controlling the transition between kernel and rich behavior which is distinct from the scale. In Appendix B, we show qualitatively similar behavior for deep ReLU networks.

2. Setup and preliminaries

We consider models $f : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ which map parameters $\mathbf{w} \in \mathbb{R}^p$ and examples $\mathbf{x} \in \mathcal{X}$ to predictions $f(\mathbf{w}, \mathbf{x}) \in \mathbb{R}$. We denote the predictor implemented by the parameters $\mathbf{w}$ as $F(\mathbf{w}) \in \{f : \mathcal{X} \to \mathbb{R}\}$, such that $F(\mathbf{w})(\mathbf{x}) = f(\mathbf{w}, \mathbf{x})$. Much of our focus will be on models, such a linear networks, which are linear in $\mathbf{x}$ (but not in the parameters $\mathbf{w}$), in which case $F(\mathbf{w})$ is a linear functional in the dual space $\mathcal{X}^*$ and can be represented as a vector $\boldsymbol{\beta}_\mathbf{w}$ with $f(\mathbf{w}, \mathbf{x}) = \langle \boldsymbol{\beta}_\mathbf{w}, \mathbf{x} \rangle$. Such models are essentially alternate parametrizations of linear models, but as we shall see that the specific parametrization is crucial.

We focus on models that are D-positive homogeneous in the parameters $\mathbf{w}$, for some integer $D \geq 1$, meaning that for any $c \in \mathbb{R}_+$, $F(c \cdot \mathbf{w}) = c^D F(\mathbf{w})$. We refer to such models simply as D-homogeneous. Many interesting model classes have this property, including multi-layer ReLU networks with fully connected and convolutional layers, layered linear networks, and matrix factorization, where D corresponds to the depth of the network.

We use $L(\mathbf{w}) = \tilde{L}(F(\mathbf{w})) = \sum_{n=1}^{N} (f(\mathbf{w}, \mathbf{x}_n) - y_n)^2$ to denote the squared loss of the model over a training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. We consider minimizing the loss $L(\mathbf{w})$ using gradient descent with infinitesimally small stepsize, *i.e.,* gradient flow dynamics

$$\dot{\mathbf{w}}(t) = -\nabla L(\mathbf{w}(t)). \tag{1}$$

We are particularly interested in the scale of initialization and capture it through a scalar parameter $\alpha \in \mathbb{R}_+$. For scale α, we will denote by $\mathbf{w}_{\alpha,\mathbf{w}_0}(t)$ the gradient flow path (1) with the initial condition $\mathbf{w}_{\alpha,\mathbf{w}_0}(0) = \alpha \mathbf{w}_0$. We consider underdetermined/overparameterized models (typically $N \ll p$), where there are many global minimizers of $L(\mathbf{w})$ with $L(\mathbf{w}) = 0$. Often, the dynamics of gradient flow converge to global minimizers of $L(\mathbf{w})$ which perfectly fits the data—this is often observed empirically in large neural network learning, though proving this is challenging and is not our focus. Rather, we want to understand *which* of the many minimizers gradient flow converges to, *i.e.,* $\mathbf{w}^\infty_{\alpha,\mathbf{w}_0} := \lim_{t\to\infty} \mathbf{w}_{\alpha,\mathbf{w}_0}(t)$ or, more importantly, the predictor $F(\mathbf{w}^\infty_{\alpha,\mathbf{w}_0})$ reached by gradient flow depending on the scale α.

3. The Kernel Regime

Locally, gradient descent/flow depends solely on the first-order approximation w.r.t. $\mathbf{w}$:

$$f(\mathbf{w}, \mathbf{x}) = f(\mathbf{w}(t), x) + \langle \mathbf{w} - \mathbf{w}(t), \nabla_\mathbf{w} f(\mathbf{w}(t), \mathbf{x}) \rangle + O(\|\mathbf{w} - \mathbf{w}(t)\|^2). \tag{2}$$

That is, gradient flow operates on the model as if it were an affine model $f(\mathbf{w}, \mathbf{x}) \approx f_0(\mathbf{x}) + \left\langle \mathbf{w}, \phi_{\mathbf{w}(t)}(\mathbf{x}) \right\rangle$ with feature map $\phi_{\mathbf{w}(t)}(\mathbf{x}) = \nabla_\mathbf{w} f(\mathbf{w}(t), \mathbf{x})$, corresponding to the *tangent kernel* $K_{\mathbf{w}(t)}(\mathbf{x}, \mathbf{x}') = \langle \nabla_\mathbf{w} f(\mathbf{w}(t), \mathbf{x}), \nabla_\mathbf{w} f(\mathbf{w}(t), \mathbf{x}') \rangle$. Of particular interest is the tangent kernel at initialization, $K_{\mathbf{w}(0)}$ (Jacot et al., 2018; Yang, 2019).

Previous work uses "kernel regime" to describe a situation in which the tangent kernel $K_{\mathbf{w}(t)}$ does not change over the course of optimization or, less formally, where it does not change significantly, *i.e.,* where $\forall t, K_{\mathbf{w}(t)} \approx K_{\mathbf{w}(0)}$. For D homogeneous models with initialization $\mathbf{w}_\alpha(0) = \alpha \mathbf{w}_0$, $K_{\mathbf{w}_\alpha(0)} = \alpha^{2(D-1)} K_0$, where we denote $K_0 = K_{\mathbf{w}_0}$. Thus, in the kernel regime, training the model $f(\mathbf{w}, \mathbf{x})$ is exactly equivalent to training an affine model $f_K(\mathbf{w}, \mathbf{x}) = \alpha^D f(\mathbf{w}(0), \mathbf{x}) + \left\langle \phi_{\mathbf{w}(0)}(\mathbf{x}), \mathbf{w} - \mathbf{w}(0) \right\rangle$ with kernelized gradient descent/flow with the kernel $K_{\mathbf{w}(0)}$ and a "bias term" of $f(\mathbf{w}(0), \mathbf{x})$. Minimizing the loss of this affine model using gradient flow reaches the solution nearest to the initialization where distance is measured with respect to the RKHS norm determined by K_0. That is, $F(\mathbf{w}_\alpha^\infty) = \arg\min_h \|h - F(\alpha \mathbf{w}_0)\|_{K_0}$ *s.t.* $h(X) = \mathbf{y}$. To avoid handling this bias term, and in particular its large scale as α increases, Chizat et al. (2019) suggest using "unbiased" initializations such that $F(\mathbf{w}_0) = 0$, so that the bias term vanishes. This is often achieved by replicating units with opposite signs at initialization (see, *e.g.,* Section 4).

But when does the kernel regime happen? Chizat et al. (2019) showed that for any homogeneous[1] model satisfying some technical conditions, the kernel regime is reached when $\alpha \to \infty$. That is, as we increase the scale of initialization, the dynamics converge to the kernel gradient flow dynamics for the initial kernel K_0. In Sections 4 and 5, for our specific models, we prove this limit as a special case of our more general analysis for all $\alpha > 0$, and we also demonstrate it empirically for matrix factorization and deep networks in Sections 6 and B. In Section 6, we additionally show how increasing the "width" of certain asymmetric matrix factorization models can also lead to the kernel regime, even when the initial scale α goes to zero at an appropriately slow rate.

In contrast to the kernel regime, and as we shall see in later sections, the $\alpha \to 0$ small initialization limit often leads to very different and rich inductive biases, *e.g.,* inducing sparsity or low-rank structure (Gunasekar et al., 2017; Li et al., 2018; Gunasekar et al., 2018), that allow for generalization in settings where kernel methods would not. We will refer to the limit of this distinctly non-kernel behavior as the "rich limit." This regime is also called the "active," "adaptive," or "feature-learning" regime since the tangent kernel $K_{\mathbf{w}(t)}$ changes over the course of training, in a sense adapting to the data. We argue that this rich limit is the one that truly allows us to exploit the power of depth, and thus is the more relevant regime for understanding the success of deep learning.

4. Detailed Study of a Simple Depth-2 Model

Consider the class of linear functions over $\mathcal{X} = \mathbb{R}^d$, with squared parameterization as follows:

$$f(\mathbf{w}, \mathbf{x}) = \sum_{i=1}^{d} (\mathbf{w}_{+,i}^2 - \mathbf{w}_{-,i}^2)\mathbf{x}_i = \langle \boldsymbol{\beta}_{\mathbf{w}}, \mathbf{x} \rangle, \ \mathbf{w} = \left[\begin{smallmatrix} \mathbf{w}_+ \\ \mathbf{w}_- \end{smallmatrix}\right] \in \mathbb{R}^{2d}, \text{ and } \boldsymbol{\beta}_{\mathbf{w}} = \mathbf{w}_+^2 - \mathbf{w}_-^2 \quad (3)$$

where $\mathbf{z}^2$ for $\mathbf{z} \in \mathbb{R}^d$ denotes elementwise squaring. The model can be thought of as a "diagonal" linear neural network (*i.e.,* where the weight matrices have diagonal structure) with $2d$ units. A "standard" diagonal linear network would have d units, with each unit

1. Chizat et al. did not consider only homogeneous models, and instead of studying the scale of initialization they studied scaling the output of the model. For homogeneous models, the dynamics obtained by scaling the initialization are equivalent to those obtained by scaling the output, and so here we focus on homogeneous models and on scaling the initialization.

connected to just a single input unit with weight $\boldsymbol{u}_i$ and the output with weight $\boldsymbol{v}_i$, thus implementing the model $f((\boldsymbol{u},\boldsymbol{v}),\mathbf{x}) = \sum_i \boldsymbol{u}_i\boldsymbol{v}_i\mathbf{x}_i$ which is illustrated in Figure 9(*a*) in Appendix C. However, we also show in Appendix C that if $|\boldsymbol{u}_i| = |\boldsymbol{v}_i|$ at initialization, then their magnitudes will remain equal and their signs will not flip throughout training. Therefore, we can equivalently parametrize the model in terms of a single shared input and output weight $\mathbf{w}_i$ for each hidden unit, yielding the model $f(\mathbf{w},\mathbf{x}) = \langle \mathbf{w}^2, \mathbf{x}\rangle$.

The reason for using an "unbiased model" with two weights $\mathbf{w}_+$ and $\mathbf{w}_-$ (*i.e.,*. $2d$ units, see illustration in Figure 9(*b*) in Appendix C) is two-fold. First, it ensures that the image of $F(\mathbf{w})$ is all (signed) linear functions, and thus the model is truly equivalent to standard linear regression. Second, it allows for initialization at $F(\alpha\mathbf{w}_0) = 0$ (by choosing $\mathbf{w}_+(0) = \mathbf{w}_-(0)$) without this being a saddle point from which gradient flow will never escape.[2]

The model (3) is perhaps the simplest non-trivial D-homogeneous model for $D > 1$, and we chose it for studying the role of scale of initialization because it already exhibits distinct and interesting kernel and rich behaviors, and we can also completely understand both the implicit regularization and the transition between regimes analytically.

We study the underdetermined $N \ll d$ case where there are many possible solutions $X\boldsymbol{\beta} = \mathbf{y}$. We will use $\boldsymbol{\beta}^\infty_{\alpha,\mathbf{w}_0}$ to denote the solution reached by gradient flow when initialized at $\mathbf{w}_+(0) = \mathbf{w}_-(0) = \alpha\mathbf{w}_0$. We will start by focusing on the special case where $\mathbf{w}_0 = \mathbf{1}$. In this case, the tangent kernel at initialization is $K_{\mathbf{w}(0)}(\mathbf{x},\mathbf{x}') = 8\alpha^2\langle \mathbf{x}, \mathbf{x}'\rangle$, which is just a scaling of the standard inner product kernel, so $\|\boldsymbol{\beta}\|_{K_{\mathbf{w}(0)}} \propto \|\boldsymbol{\beta}\|_2$. Thus, in the kernel regime, $\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}$ will be the minimum ℓ_2 norm solution, $\boldsymbol{\beta}^*_{\ell_2} := \arg\min_{X\boldsymbol{\beta}=y}\|\boldsymbol{\beta}\|_2$. Following Chizat et al. (2019) and the discussion in Section 3, we thus expect that $\lim_{\alpha\to\infty}\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}} = \boldsymbol{\beta}^*_{\ell_2}$.

In contrast, from Corollary 2 in Gunasekar et al. (2017), as $\alpha \to 0$, gradient flow leads instead to a rich limit of ℓ_1 minimization, *i.e.,* $\lim_{\alpha\to 0}\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}} = \boldsymbol{\beta}^*_{\ell_1} := \arg\min_{X\boldsymbol{\beta}=y}\|\boldsymbol{\beta}\|_1$. Comparing this with the kernel regime, we already see two distinct behaviors and, in high dimensions, two very different inductive biases. In particular, the rich limit ℓ_1 bias is *not* an RKHS norm for any choice of kernel. We have now described the asymptotic regimes where $\alpha \to 0$ or $\alpha \to \infty$, but can we characterize and understand the transition between the two regimes as α scales from very small to very large? The following theorem does just that.

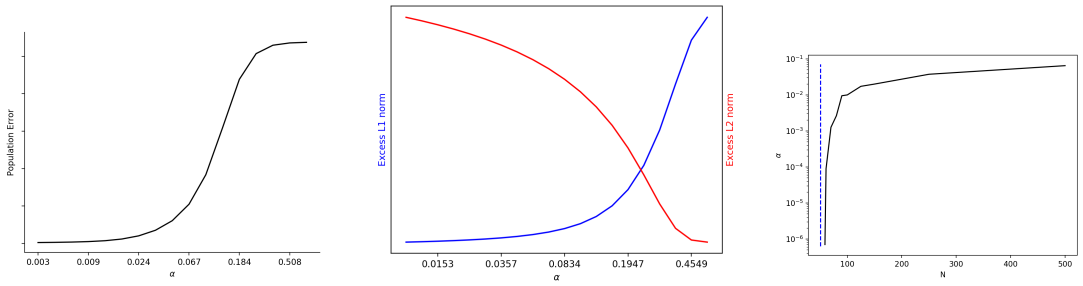
Theorem 1 (Special case: $\mathbf{w}_0 = \mathbf{1}$) *For any $0 < \alpha < \infty$, if the gradient flow solution $\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}$ for the squared parameterization model in eq. (3) satisfies $X\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}} = \mathbf{y}$, then*

$$\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}} = \arg\min_{\boldsymbol{\beta}} Q_\alpha(\boldsymbol{\beta}) \;\; s.t.\; X\boldsymbol{\beta} = \mathbf{y}, \tag{4}$$

where $Q_\alpha(\boldsymbol{\beta}) = \alpha^2\sum_{i=1}^d q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2}\right)$ and $q(z) = \int_0^z \operatorname{arcsinh}\left(\frac{u}{2}\right)du = 2 - \sqrt{4+z^2} + z\operatorname{arcsinh}\left(\frac{z}{2}\right)$.

A General Approach for Deriving the Implicit Bias Once given an expression for Q_α, it is straightforward to analyze the dynamics of $\boldsymbol{\beta}_{\alpha,\mathbf{1}}$ and show that it is the minimum Q_α solution to $X\boldsymbol{\beta} = \mathbf{y}$. However, a key contribution of this work is in developing a method for determining what the implicit bias is *when we do not already have a good guess.* First, we analyze the gradient flow dynamics and show that if $X\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}} = \mathbf{y}$ then $\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}} = b_\alpha(X^\top\nu)$

2. Our results can be generalized to "biased" initialization (*i.e.,* where $\mathbf{w}_- \neq \mathbf{w}_+$ at initialization), or the asymmetric parametrization $f((u,v),\mathbf{x}) = \sum_i u_iv_ix_i$, however this complicates the presentation without adding much insight.



(a) Generalization (b) Norms of solution (c) Sample complexity

Figure 1: In (a) the population error of the gradient flow solution vs. α in the sparse regression problem described in Section 4. In (b), we plot $\|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_1 - \|\boldsymbol{\beta}^*_{\ell_1}\|_1$ in blue and $\|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_2 - \|\boldsymbol{\beta}^*_{\ell_2}\|_2$ in red vs. α. In (c), the largest α such that $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}$ achieves population error at most 0.025 is shown. The dashed line indicates the number of samples needed by $\boldsymbol{\beta}^*_{\ell_1}$.

for a certain function b_α and vector ν. It is *not* necessary to be able to calculate ν, which would be very difficult, even for our simple examples. Next, we suppose that there is some function Q_α such that (4) holds. The KKT optimality conditions for (4) are $X\boldsymbol{\beta}^* = \mathbf{y}$ and $\exists \nu$ s.t. $\nabla Q_\alpha(\boldsymbol{\beta}^*) = X^\top \nu$. Therefore, if indeed $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}} = \boldsymbol{\beta}^*$ and $X\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}} = \mathbf{y}$ then $\nabla Q_\alpha\left(\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\right) = \nabla Q_\alpha\left(b_\alpha(X^\top \nu)\right) = X^\top \nu$. We solve the differential equation $\nabla Q_\alpha = b_\alpha^{-1}$ to yield Q_α. Theorem 1 in Appendix D is proven using this method.

In light of Theorem 1, the function Q_α (referred to as the "hypentropy" function in Ghai et al. (2019)) can be understood as an implicit regularizer which biases the gradient flow solution towards one particular zero-error solution out of the many possibilities. As α ranges from 0 to ∞, the Q_α regularizer interpolates between the ℓ_1 and ℓ_2 norms, as illustrated in Figure 3(a) (the line labelled $D = 2$ depicts the coordinate function q). As $\alpha \to \infty$ we have that $\boldsymbol{\beta}_i/\alpha^2 \to 0$, and so the behaviour of $Q_\alpha(\boldsymbol{\beta})$ is governed by $q(z) = \Theta(z^2)$ around $z = 0$, thus $Q_\alpha(\boldsymbol{\beta}) \propto \sum_i \boldsymbol{\beta}_i^2$. On the other hand when $\alpha \to 0$, $|\boldsymbol{\beta}_i/\alpha^2| \to \infty$ is determined by $q(z) = \Theta(|z|\log|z|)$ as $|z| \to \infty$. In this regime $\frac{1}{\log(1/\alpha^2)} Q_\alpha(\boldsymbol{\beta}) \propto \frac{1}{\log(1/\alpha^2)} \sum_i |\boldsymbol{\beta}_i| \log|\frac{\boldsymbol{\beta}_i}{\alpha^2}| = \|\boldsymbol{\beta}\|_1 + O(1/\log(1/\alpha^2))$. The following Theorem, proven in Appendix E, quantifies the scale of α which guarantees that $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}$ approximates the minimum ℓ_1 or ℓ_2 norm solution:

Theorem 2 *For any $0 < \epsilon < d$, under the setting of Theorem 1 with $\mathbf{w}_0 = \mathbf{1}$,*

$$\alpha \le \min\left\{ \left(2(1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1\right)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-d/(\epsilon\|\boldsymbol{\beta}^*_{\ell_1}\|_1)\right) \right\} \implies \|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_1 \le (1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$$

$$\alpha \ge \sqrt{2(1+\epsilon)(1+2/\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2} \implies \|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_2^2 \le (1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2^2$$

Looking carefully at Theorem 2, we notice a certain asymmetry between reaching the kernel regime versus the rich limit: polynomially large α suffices to approximate $\boldsymbol{\beta}^*_{\ell_2}$ to a very high degree of accuracy, but *exponentially* small α is needed to approximate $\boldsymbol{\beta}^*_{\ell_1}$.[3] This suggests an explanation for the difficulty of empirically demonstrating rich limit behavior

3. Theorem 2 only shows that exponentially small α is *sufficient* for approximating $\boldsymbol{\beta}^*_{\ell_1}$ and is not a proof that it is necessary. However, Lemma 6 in Appendix E proves that $\alpha \le d^{-\Omega(1/\epsilon)}$ is indeed *necessary* for Q_α to be proportional to the ℓ_1 norm for every unit vector simultaneously. This indicates that α must be exponentially small to approximate $\boldsymbol{\beta}^*_{\ell_1}$ for certain problems.

in matrix factorization problems (Gunasekar et al., 2017; Arora et al., 2019a): since the initialization may need to be exceedingly small, conducting experiments in the truly rich limit may be infeasible for computational reasons.

Generalization In order to understand the effect of the initialization on generalization, consider a simple sparse regression problem, where $\mathbf{x}_1, \ldots, \mathbf{x}_N \sim \mathcal{N}(0, I)$ and $y_n \sim \mathcal{N}(\langle \boldsymbol{\beta}^*, \mathbf{x}_n \rangle, 0.01)$ where $\boldsymbol{\beta}^*$ is r^*-sparse with non-zero entries equal to $1/\sqrt{r^*}$. When $N \leq d$, gradient flow will generally reach a zero training error solution, however, not all of these solutions will generalize the same. In the rich limit, $N = \Omega(r^* \log d)$ samples suffices for $\boldsymbol{\beta}^*_{\ell_1}$ to generalize well. On the other hand, even though we can fit the training data perfectly well, the kernel regime solution $\boldsymbol{\beta}^*_{\ell_2}$ would not generalize at all with this sample size ($N = \Omega(d)$ samples would be needed), see Figure 1(*c*). Thus, in this case good generalization requires using very small initialization, and generalization will tend to improve as α decreases. From an optimization perspective this is unfortunate because $\mathbf{w} = 0$ is a saddle point, so taking $\alpha \to 0$ will likely increase the time needed to escape the vicinity of zero.

Thus, there seems to be a tension between generalization and optimization: a smaller α might improve generalization, but it makes optimization trickier. This suggests that one should operate just on the edge of the rich limit, using the largest α that still allows for generalization. This is borne out by our experiments with deep, non-linear neural networks (see Appendix B), where standard initializations correspond to being right on the edge of entering the kernel regime, where we expect models to both generalize well and avoid serious optimization difficulties. Given the extensive efforts put into designing good initialization schemes, this gives further credence to the idea that models will perform best when trained in the intermediate regime between rich and kernel behavior.

This tension can also be seen through a tradeoff between the sample size and the largest α we can use and still generalize. In Figure 1(*c*), for each sample size N, we plot the largest α for which the gradient flow solution $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}$ achieves population risk below some threshold. As N approaches the minimum number of samples for which $\boldsymbol{\beta}^*_{\ell_1}$ generalizes (the vertical dashed line), α must become extremely small. However, generalization is much easier if the number of samples is only slightly larger, and much larger α suffices.

The "Shape" of $\mathbf{w}_0$ and the Implicit Bias So far, we have discussed the implicit bias in the special case $\mathbf{w}_0 = \mathbf{1}$, but we can also characterize it for non-uniform initialization $\mathbf{w}_0$:

Theorem 1 (General case) *For any $0 < \alpha < \infty$ and $\mathbf{w}_0$ with no zero entries, if the gradient flow solution $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0}$ satisfies $X\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} = \mathbf{y}$, then*

$$\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} = \arg\min_{\boldsymbol{\beta}} Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) \;\; s.t. \; X\boldsymbol{\beta} = \mathbf{y}, \tag{5}$$

where $Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \sum_{i=1}^{d} \alpha^2 \mathbf{w}_{0,i}^2 q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2 \mathbf{w}_{0,i}^2}\right)$ and $q(z) = 2 - \sqrt{4+z^2} + z \operatorname{arcsinh}\left(\frac{z}{2}\right)$.

Consider the asymptotic behavior of $Q_{\alpha,\mathbf{w}_0}$. For small z, $q(z) = \frac{z^2}{4} + O(z^4)$ so for $\alpha \to \infty$

$$Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \sum_{i=1}^{d} \alpha^2 \mathbf{w}_{0,i}^2 \, q\Big(\frac{\boldsymbol{\beta}_i}{\alpha^2 \mathbf{w}_{0,i}^2}\Big) = \sum_{i=1}^{d} \frac{\boldsymbol{\beta}_i^2}{4\alpha^2 \mathbf{w}_{0,i}^2} + O\big(\alpha^{-6}\big) \tag{6}$$

In other words, in the $\alpha \to \infty$ limit, $Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta})$ is proportional to a quadratic norm weighted by $\mathrm{diag}(1/\mathbf{w}_0^2)$. On the other hand, for large $|z|$, $q(z) = |z|\log|z| + O(1/|z|)$ so as $\alpha \to 0$

$$\frac{1}{\log(1/\alpha^2)} Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \frac{1}{\log(1/\alpha^2)} \sum_{i=1}^{d} \alpha^2 \mathbf{w}_{0,i}^2 \, q\Big(\frac{\boldsymbol{\beta}_i}{\alpha^2 \mathbf{w}_{0,i}^2}\Big) = \sum_{i=1}^{d} |\boldsymbol{\beta}_i| + O\big(1/\log(1/\alpha^2)\big) \tag{7}$$

So, in the $\alpha \to 0$ limit, $Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta})$ is proportional to $\|\boldsymbol{\beta}\|_1$ regardless of the shape of the initialization $\mathbf{w}_0$! The specifics of the initialization, $\mathbf{w}_0$, therefore affect the implicit bias in the kernel regime (and in the intermediate regime) but *not* in the rich limit.

For wide neural networks with i.i.d. initialized units, the analogue of the "shape" is the distribution used to initialize each unit, including the relative scale of the input weights, output weights, and biases. Indeed, as was explored by Williams et al. (2019) and as we elaborate in Appendix B, changing the unit initialization distribution changes the tangent kernel at initialization and hence the kernel regime behavior. However, we also demonstrate empirically that changing the initialization distribution ("shape") does *not* change the rich regime behavior. These observations match the behavior of $Q_{\alpha,\mathbf{w}_0}$ analyzed above.

Explicit Regularization From the geometry of gradient descent, it is tempting to imagine that its implicit bias would be minimizing the Euclidean norm from initialization:

$$\boldsymbol{\beta}^R_{\alpha,\mathbf{w}_0} := F\Big(\arg\min_{\mathbf{w}} \|\mathbf{w} - \alpha\mathbf{w}_0\|_2^2 \text{ s.t. } L(\mathbf{w}) = 0\Big) = \arg\min_{\boldsymbol{\beta}} R_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) \text{ s.t. } X\boldsymbol{\beta} = y \tag{8}$$

$$\text{where} \quad R_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \min_{\mathbf{w}} \|\mathbf{w} - \alpha\mathbf{w}_0\|_2^2 \text{ s.t. } F(\mathbf{w}) = \boldsymbol{\beta}. \tag{9}$$

It is certainly the case for standard linear regression $f(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$, where from standard analysis, it can be shown that $\boldsymbol{\beta}^\infty_{\alpha,\mathbf{w}_0} = \boldsymbol{\beta}^R_{\alpha,\mathbf{w}_0}$ so the bias is captured by $R_{\alpha,\mathbf{w}_0}$. But does this characterization fully explain the implicit bias for our 2-homogeneous model? Perhaps the behavior in terms of $Q_{\alpha,\mathbf{w}_0}$ can also be explained by $R_{\alpha,\mathbf{w}_0}$? Focusing on the special case $\mathbf{w}_0 = \mathbf{1}$, it is easy to verify that the limiting behavior when $\alpha \to 0$ and $\alpha \to \infty$ of the two approaches match. We can also calculate $R_{\alpha,\mathbf{1}}(\boldsymbol{\beta})$, which decomposes over the coordinates, as: $R_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) = \sum_i r(\boldsymbol{\beta}_i/\alpha^2)$ where $r(z)$ is the unique real root of $p_z(u) = u^4 - 6u^3 + (12 - 2z^2)u^2 - (8 + 10z^2)u + z^2 + z^4$.

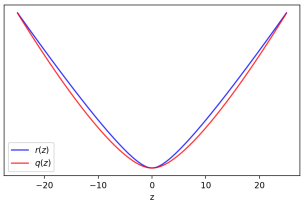
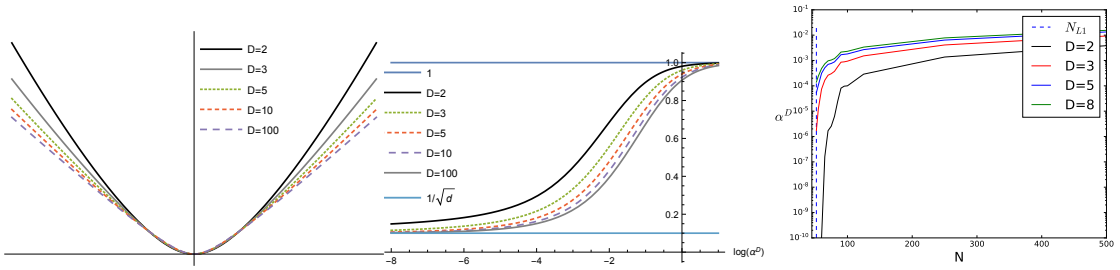


Figure 2: $q(z)$ and $r(z)$.

This function $r(z)$ is shown next to $q(z)$ in Figure 2. They are similar but not the same since $r(z)$ is algebraic (even radical), while $q(z)$ is transcendental. Thus, $Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \neq R_{\alpha,\mathbf{1}}(\boldsymbol{\beta})$ and they are not simple rescalings of each other either. Furthermore, while α needs to be exponentially small in order for $Q_{\alpha,\mathbf{1}}$ to approximate the ℓ_1 norm, the algebraic $R_{\alpha,\mathbf{1}}(\boldsymbol{\beta})$ approaches $\|\boldsymbol{\beta}\|_1$ polynomially in terms of the scale of α. Therefore, the bias of gradient descent and the transition from the kernel regime to the rich limit is more complex and subtle than what is captured simply by distances in parameter space.

5. Higher Order Models

So far, we considered a 2-homogeneous model, corresponding to a simple depth-2 "diagonal" network. Deeper models correspond to higher order homogeneity (*e.g.*, a depth-D



(a) Regularizer (b) Approximation ratio (c) Sparse regression simulation

Figure 3: (a) $q_D(z)$ for several values of D. (b) The ratio $\frac{Q_\alpha^D(e_1)}{Q_\alpha^D(\mathbf{1}_d/\|\mathbf{1}_d\|_2)}$ as a function of α, where $e_1 = [1, 0, 0, \ldots, 0]$ is the first standard basis vector and $\mathbf{1}_d = [1, 1, \ldots, 1]$ is the all ones vector in $\mathbb{R}^d$. This captures the transition between approximating the ℓ_2 norm (where the ratio is 1) and the ℓ_1 norm (where the ratio is $1/\sqrt{d}$). (c) A sparse regression simulation as in Figure 1, using different order models. The y-axis is the largest α^D (the scale of $\boldsymbol{\beta}$ at initialization) that leads to recovery of the planted predictor to accuracy 0.025. The vertical dashed line indicates the number of samples needed in order for $\boldsymbol{\beta}^*_{\ell_1}$ to approximate the plant.

ReLU network is D-homogeneous), motivating us to understand the effect of the order of homogeneity on the transition between the regimes. We therefore generalize our model and consider:

$$F_D(\mathbf{w}) = \boldsymbol{\beta}_{\mathbf{w},D} = \mathbf{w}_+^D - \mathbf{w}_-^D \quad \text{and} \quad f_D(\mathbf{w}, \mathbf{x}) = \left\langle \mathbf{w}_+^D - \mathbf{w}_-^D, \mathbf{x} \right\rangle \tag{10}$$

As before, this is just a linear regression model with an unconventional parametrization, equivalent to a depth-D matrix factorization model with commutative measurement matrices, as studied by Arora et al. (2019a), or a depth-D diagonal linear network. We can again study the effect of the scale of α on the implicit bias. Let $\boldsymbol{\beta}^\infty_{\alpha,D}$ denote the limit of gradient flow on $\mathbf{w}$ when $\mathbf{w}_+(0) = \mathbf{w}_-(0) = \alpha\mathbf{1}$. In Appendix F we prove:

Theorem 3 *For any $0 < \alpha < \infty$ and $D \geq 3$, if $X\boldsymbol{\beta}^\infty_{\alpha,D} = y$, then*

$$\boldsymbol{\beta}^\infty_{\alpha,D} = \arg\min_{\boldsymbol{\beta}} Q_\alpha^D(\boldsymbol{\beta}) \quad s.t. \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

where $Q_\alpha^D(\boldsymbol{\beta}) = \alpha^D \sum_{i=1}^d q_D(\boldsymbol{\beta}_i/\alpha^D)$ and $q_D = \int h_D^{-1}$ is the antiderivative of the unique inverse of $h_D(z) = (1-z)^{-\frac{D}{D-2}} - (1+z)^{-\frac{D}{D-2}}$ on $[-1, 1]$. Furthermore, $\lim_{\alpha\to 0} \boldsymbol{\beta}^\infty_{\alpha,D} = \boldsymbol{\beta}^*_{\ell_1}$ and $\lim_{\alpha\to\infty} \boldsymbol{\beta}^\infty_{\alpha,D} = \boldsymbol{\beta}^*_{\ell_2}$.

In the two extremes, we again get $\boldsymbol{\beta}^*_{\ell_2}$ in the kernel regime, and more interestingly, for any depth $D \geq 2$, we get the $\boldsymbol{\beta}^*_{\ell_1}$ in the rich limit, as has also been observed by Arora et al. (2019a). That the rich limit solution does not change with D is surprising, and disagrees with what would be obtained with explicit regularization (regularizing $\|\mathbf{w}\|_2$ is equivalent to $\|\boldsymbol{\beta}\|_{2/D}$ regularization), nor implicitly on with the logistic loss (which again corresponds to $\|\boldsymbol{\beta}\|_{2/D}$, see, *e.g.*, (Gunasekar et al., 2017; Lyu and Li, 2019)).

Although the two extremes do not change as we go beyond $D = 2$, what does change is the intermediate regime, particularly the sharpness of the transition into the extreme regimes, as illustrated in Figures 3(*a*)-3(*c*). The most striking difference is that, even at

order $D = 3$, the scale of α needed to approximate ℓ_1 is polynomial rather then exponential, yielding a much quicker transition to the rich limit versus the $D = 2$ case above. This allows near-optimal sparse regression with reasonable initialization scales as soon as $D > 2$, and increasing D hastens the transition to the rich limit. This may explain the empirical observations regarding the benefit of depth in deep matrix factorization (Arora et al., 2019a).

6. The Effect of Width

The kernel regime was first discussed in the context of the high (or infinite) *width* of a network, but our treatment so far, following Chizat et al. (2019), identified the *scale* of the initialization as the crucial parameter for entering the kernel regime. So is the width indeed a red herring? Actually, the width indeed plays an important role and allows entering the kernel regime more naturally.

The fixed-width models so far only reach the kernel regime when the initial scale of parameters goes to infinity. To keep this from exploding both the outputs of the model and $F(\mathbf{w}(0))$ itself, we used Chizat and Bach's "unbiasing" trick. However, using unbiased models with $F(\alpha \mathbf{w}_0) = 0$ conceals the unnatural nature of this regime: although the final output may not explode, outputs of internal units do explode in the scaling leading to the kernel regime. Realistic models are not trained like this. We will now use a "wide" generalization of our simple linear model to illustrate how increasing the width can induce kernel regime behavior in a more natural setting where both the initial output and the outputs of all internal units, do not explode and can even vanish.

Consider an (asymmetric) matrix factorization model, *i.e.,* a linear model over matrix-valued observations[4] $\mathbf{X} \in \mathbb{R}^{d \times d}$ described by $f((\mathbf{U}, \mathbf{V}), \mathbf{X}) = \left\langle \mathbf{U}\mathbf{V}^\top, \mathbf{X} \right\rangle$ where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times k}$, and we refer to $k \geq d$ as the "width." We are interested in understanding the behaviour as $k \to \infty$ and the scaling of initialization α of each individual parameter changes with k. Let $\mathbf{M}_{\mathbf{U},\mathbf{V}} = F(\mathbf{U}, \mathbf{V}) = \mathbf{U}\mathbf{V}^\top$ denote the underlying linear predictor. We consider minimizing the squared loss $L(\mathbf{U}, \mathbf{V}) = \tilde{L}(\mathbf{M}_{\mathbf{U},\mathbf{V}}) = \sum_{n=1}^{N} (\langle \mathbf{X}_n, \mathbf{M}_{\mathbf{U},\mathbf{V}} \rangle - y_n)^2$ on N samples using gradient flow on the parameters $\mathbf{U}$ and $\mathbf{V}$. This formulation includes a number of special cases such as matrix completion, matrix sensing, and two layer linear neural networks.

We want to understand how the *scale and width* jointly affect the implicit bias. Since the number of parameters grows with k, it now makes less sense to capture the scale via the magnitude of individual parameters. Instead, we will capture scale via $\sigma = \frac{1}{d}\|\mathbf{M}_{\mathbf{U},\mathbf{V}}\|_F$, *i.e.,* the scale of the model itself at initialization. The initial predictions are also of order σ, *e.g.,* when $\mathbf{X}$ is Gaussian and has unit Frobenius norm. We will now show that the model remains in the kernel regime depending on the relative scaling of k and σ. Unlike the D-homogeneous models of Sections 4 and 5, $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ can be in the kernel regime when σ remains bounded, or even when it goes to zero.

"Lifted" symmetric factorization Does the scale of $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ indeed capture the relevant notion of parameter scale? In case of a *symmetric* matrix factorization model $\mathbf{M}_{\mathbf{W}} = \mathbf{W}\mathbf{W}^\top$, $\mathbf{M}_{\mathbf{W}}$ captures the entire behaviour of the model since the dynamics on $\mathbf{M}_{\mathbf{W}(t)}$ induced by gradient flow on $\mathbf{W}(t)$ given by $\dot{\mathbf{M}}_{\mathbf{W}(t)} = \nabla\tilde{L}(\mathbf{M}_{\mathbf{W}(t)})\mathbf{M}_{\mathbf{W}(t)} + \mathbf{M}_{\mathbf{W}(t)}\nabla\tilde{L}(\mathbf{M}_{\mathbf{W}(t)})$ depends only on $\mathbf{M}_{\mathbf{W}(t)}$ and not on $\mathbf{W}(t)$ itself (Gunasekar et al., 2017).

4. $\mathbf{X}$ need not be square; the results and empirical observations extend for non-square matrices.

For the asymmetric model $\mathbf{M}_{\mathbf{U},\mathbf{V}}$, this is no longer the case, and the dynamics of $\mathbf{M}_{\mathbf{U}(t),\mathbf{V}(t)}$ do depend on the specific factorization $\mathbf{U}(t), \mathbf{V}(t)$ and not only on the product $\mathbf{M}_{\mathbf{U},\mathbf{V}}$. Instead, we can consider an equivalent "lifted" symmetric problem defined by $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}} = \left[\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right]\left[\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right]^\top = \left[\begin{smallmatrix}\mathbf{U}\mathbf{U}^\top & \mathbf{M}_{\mathbf{U},\mathbf{V}}\\ \mathbf{M}_{\mathbf{U},\mathbf{V}}^\top & \mathbf{V}\mathbf{V}^\top\end{smallmatrix}\right]$ and $\bar{\mathbf{X}}_n = \frac{1}{2}\left[\begin{smallmatrix}0 & \mathbf{X}_n\\ \mathbf{X}_n^\top & 0\end{smallmatrix}\right]$ with $\bar{f}((\mathbf{U},\mathbf{V}),\bar{\mathbf{X}}) = \left\langle \bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}, \bar{\mathbf{X}}\right\rangle$. The dynamics over $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$—which on the off diagonal blocks are equivalent to those of $\mathbf{M}_{\mathbf{U},\mathbf{V}}$—are now fully determined by $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$ itself; that is, by the combination of the "observed" part $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ as well as the "unobserved" diagonal blocks $\mathbf{U}\mathbf{U}^\top$ and $\mathbf{V}\mathbf{V}^\top$. To see how this plays out in terms of the width, consider initializing $\mathbf{U}(0)$ and $\mathbf{V}(0)$ with i.i.d. $\mathcal{N}(0,\alpha^2)$ entries. The off-diagonal entries of $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$, and thus σ, will scale with $\alpha^2\sqrt{k}$ while the diagonal entries of $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$ will scale with $\alpha^2 k = \sigma\sqrt{k}$.

By analogy to the models studied in Sections 4 and 5, we can infer that the relevant scale for the problem is that of the entire lifted matrix $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$, which determines the dynamics, and which is a factor of $\sqrt{k}$ *larger* than the scale of the actual predictor $\mathbf{M}_{\mathbf{U},\mathbf{V}}$. We now show that in the special case where the measurements $\mathbf{X}_1,\ldots,\mathbf{X}_N$ commute with each other, the implicit bias is indeed precisely captured by $\sigma\sqrt{k}$—when this quantity goes to zero, we enter the rich limit; when this quantity goes to infinity, we enter the kernel regime; and in the transition we have behavior similar to the 2-homogeneous model from Section 4.

Matrix Sensing with Diagonal/Commutative Measurements Consider the special case where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are all diagonal, or more generally commutative, matrices. The diagonal elements of $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ (the only relevant part when $\mathbf{X}$ is diagonal) are $[\mathbf{M}_{\mathbf{U},\mathbf{V}}]_{ii} = \sum_{j=1}^k \mathbf{U}_{ij}\mathbf{V}_{ij}$, and so the diagonal case can be thought of as an (asymmetric) "wide" analogue to the 2-homogeneous model we considered in Section 4, *i.e.,* a "wide parallel linear network" where each input unit $\mathbf{X}_{ii}$ has its own set of k hidden $(\mathbf{U}_{i1},\mathbf{V}_{i1}),\ldots,(\mathbf{U}_{ik},\mathbf{V}_{ik})$ units. This is depicted in Figure 4. We consider initializing $\mathbf{U}(0)$ and $\mathbf{V}(0)$ with i.i.d. $\mathcal{N}(0,\alpha^2)$ entries, so $\mathbf{M}_{\mathbf{U}(0),\mathbf{V}(0)}$ will be of magnitude $\sigma = \alpha^2\sqrt{k}$, and take $k\to\infty$, scaling α as a function of k.

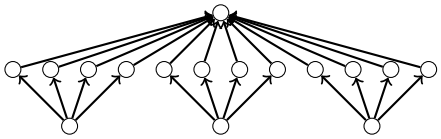


Figure 4: A wide parallel network

Theorem 4, proven in Appendix G, completely characterizes the implicit bias of the model, which corresponds to minimizing Q_μ applied to its spectrum (the "Schatten-Q_μ-norm"). This corresponds to an implicit bias which approximates the trace norm for small μ and the Frobenius norm for large μ. In the diagonal case, this is just the minimum Q_μ solution, but unlike the "width-1" model of Section 4, this is obtained without an "unbiasing" trick.

Theorem 4 *Let $k\to\infty$, $\sigma(k)\to 0$, and $\mu^2 := \frac{1}{2}\lim_{k\to\infty}\sigma(k)\sqrt{k}$, and suppose $\mathbf{X}_1,\ldots,\mathbf{X}_N$ commute. If $\boldsymbol{M}_{\mathbf{U},\mathbf{V}}(t)$ converges to a zero error solution $\boldsymbol{M}^*_{\mathbf{U},\mathbf{V}}$, then*

$$\boldsymbol{M}^*_{\mathbf{U},\mathbf{V}} = \arg\min_{\mathbf{M}} Q_\mu(\mathit{spectrum}(\mathbf{M})) \quad \mathit{s.t.}\ L(\mathbf{M}) = 0$$

Non-Commutative Measurements We might expect that in the general case, there is also a transition around $\sigma \asymp 1/\sqrt{k}$: (a) if $\sigma = \omega(1/\sqrt{k})$, then $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}} \to \infty\cdot I$ and the model should remain in the kernel regime, even in cases where $\sigma = \|\mathbf{M}_{\mathbf{U},\mathbf{V}}\|_F \to 0$; (b) on the other hand, if $\sigma = o(1/\sqrt{k})$ then $\|\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}\|_F \to 0$ and the model should approach some rich limit; (c) at the transition, when $\sigma = \Theta(1/\sqrt{k})$, $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$ will remain bounded and we should

be in an intermediate regime. In light of Theorem 4, if $0 < \mu^2 := \frac{1}{2} \lim \sigma\sqrt{k} < \infty$ exists, we expect an implicit bias resembling Q_μ. Geiger et al. (2019) also study such a transition using different arguments, but they focus on the extremes $\sigma = o(1/\sqrt{k})$ and $\sigma = \omega(1/\sqrt{k})$ and not on the transition. Here, we understand the scaling directly in terms of how the width affects the magnitude of the symmetrized model $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$.

For the symmetric matrix factorization model with non-commutative measurements, we *can* analyze the case $\omega(1/\sqrt{k}) = \sigma = o(1)$ and prove it, unsurprisingly, leads to the kernel regime (see Theorem 9 and Corollary 10 in Appendix H, which closely follow the approach of Chizat et al. (2019)). It would be more interesting to characterize the implicit bias across the full range of the intermediate regime, however, even just the rich limit in this setting has defied generic analysis so far (*q.v.,* the still unresolved conjecture of Gunasekar et al. (2017)), and analyzing the intermediate regime is even harder (in particular, the limit of the intermediate regime describes the rich limit). Nevertheless, we now describe empirical evidence that the behavior of Theorem 4 may also hold for non-commutative measurements.

Low-Rank Matrix Completion Matrix completion is a natural and commonly-studied instance of the general matrix factorization model where the measurements $\mathbf{X}_n = e_{i_n} e_{j_n}^\top$ are indicators of single entries of the matrix (note: these measurements do *not* commute), and so y_n corresponds to observed entries of an unknown matrix $\mathbf{Y}^*$. When $N < d^2$, there are many minimizers of the squared loss which correspond to matching $\mathbf{Y}^*$ on all of the observed entries, and imputing arbitrary values for the unobserved entries. Generally, there is no hope of "generalizing" to unseen entries of $\mathbf{Y}^*$, which need not have any relation to the observed entries. However, when $\mathbf{Y}^*$ is rank-r for $r \ll d$, the minimum nuclear norm solution will recover $\mathbf{Y}^*$ when $N = \tilde{\Omega}(d^{1.2}r)$ (Candès and Recht, 2009). While Theorem 4 *does not* apply for these non-commutative measurements, our experiments described in Appendix A (Figure 5) indicate the same behavior appears to hold: when $\sigma = o(1/\sqrt{k})$, the nuclear norm is nearly minimized and $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ converges to $\mathbf{Y}^*$. On the other hand, the kernel regime corresponds to implicit Frobenius norm regularization, which does not recover $\mathbf{Y}^*$. Therefore, in order to recover $\mathbf{Y}^*$, it is necessary to choose an initialization with $\sigma\sqrt{k} \ll 1$.

Conclusion In this section, we provide evidence that both the *scale*, σ, and *width*, k, of asymmetric matrix factorization models have a role to play in the implicit bias. In particular, we show that the scale of the equivalent "lifted" or "symmetrized" model $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$ is the relevant parameter. Under many natural initialization schemes for $\mathbf{U}$ and $\mathbf{V}$, *e.g.,* with i.i.d. Gaussian entries, the scale of $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$ is $\sqrt{k}$ times larger than the scale of $\mathbf{M}_{\mathbf{U},\mathbf{V}}$. Consequently, wide factorizations can reach the kernel regime even while $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ remains bounded, even without resorting to "unbiasing." On the other hand, reaching the rich limit requires an even smaller initialization for large k.

Acknowledgments

This work was supported by NSF Grant 1764032. BW is supported by a Google PhD Research Fellowship. DS was supported by the Israel Science Foundation (grant No. 31/1031). This work was partially done while the authors were visiting the Simons Institute for the Theory of Computing.

References

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

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6155–6166, 2019.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7411–7422, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019b.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

Amit Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

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*, 2018.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy learning in deep neural networks: an empirical study. *arXiv preprint arXiv:1906.08034*, 2019.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. *arXiv preprint arXiv:1902.01903*, 2019.

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

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*, pages 9461–9471, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

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

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690, 2019.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. In *Advances in Neural Information Processing Systems*, pages 8376–8385, 2019.

Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 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.

Appendix A. Matrix Completion Experiments

We display the results of our matrix completion experiments in Figure 5. The experiments indicate that when $\sigma = o(1/\sqrt{k})$, the implicit regularization indeed appears to correspond to nuclear norm regularization, and $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ converges to $\mathbf{Y}^*$. On the other hand, the kernel regime corresponds to implicit Frobenius norm regularization, which does not recover $\mathbf{Y}^*$ until $N = \Omega(d^2)$.

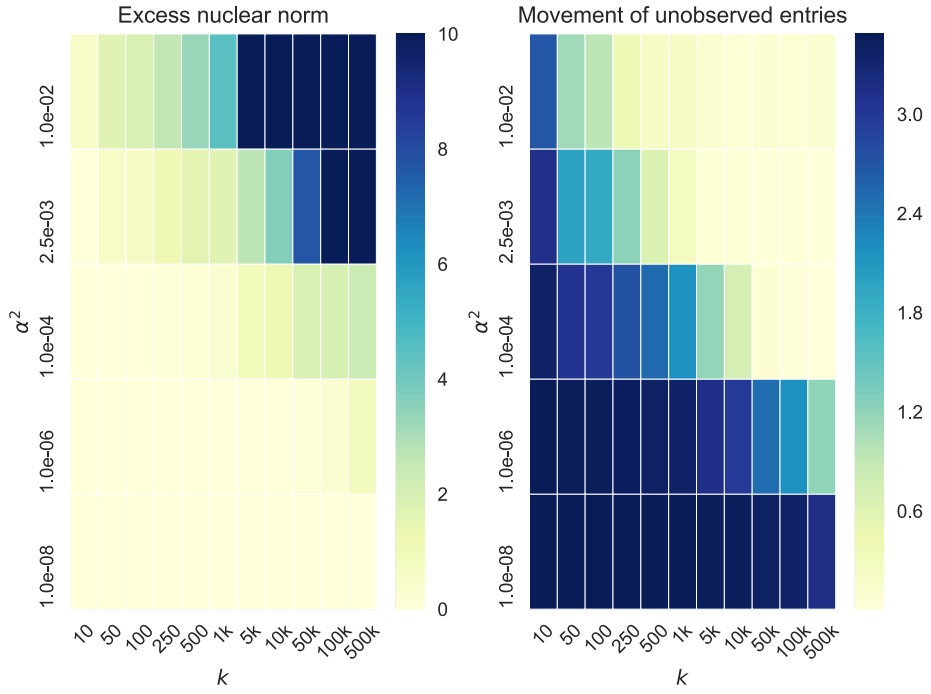


Figure 5: **Matrix Completion** We generate rank-1 ground truth $\mathbf{Y}^* = u^*(v^*)^\top$ where $u^*, v^* \sim \mathcal{N}(0, I_{10\times 10})$ and observe $N = 60$ random entries. We minimize the squared loss on the observed entries of the model $F(U, V) = UV^\top$ with $U, V \in \mathbb{R}^{d\times k}$ using gradient descent with small stepsize 10^{-5}. We initialize $\mathbf{U}(0)_{ij}, \mathbf{V}(0)_{ij} \sim \mathcal{N}(0, \alpha^2)$. For the solution, $\mathbf{M}_{\alpha,k}$, reached by gradient descent, the left heatmap depicts the excess nuclear norm $\|\mathbf{M}_{\alpha,k}\|_* - \|\mathbf{Y}^*\|_*$ (this is conjectured to be zero in the rich limit); and the right heatmap depicts the root mean squared difference between the entries $\mathbf{M}_{\alpha,k}$ and $\mathbf{U}(0)\mathbf{V}(0)^\top$ corresponding to unobserved entries of $\mathbf{Y}^*$ (in the kernel regime, the unobserved entries do not move). Both exhibit a phase transition around $\alpha^2 k = \sigma\sqrt{k} \asymp 1$. For $\sigma\sqrt{k} \ll 1$ the excess nuclear norm is approximately zero, corresponding to the rich limit. For $\sigma\sqrt{k} \gg 1$, the unobserved entries do not change, which corresponds to the kernel regime. This phase transition appears to sharpen somewhat as k increases.

Appendix B. Neural Network Experiments

In Sections 4 and 5, we intentionally focused on the simplest possible models in which a kernel-to-rich transition can be observed, in order to isolate this phenomena and understand it in detail. In those simple models, we were able to obtain a complete analytic description of the transition. Obtaining such a precise description in more complex models is too optimistic

at this point, but we demonstrate the same phenomena empirically for realistic non-linear neural networks.

Figures 6(a) and 6(b) indicate that non-linear ReLU networks remain in the kernel regime when the initialization is large, they exit from the kernel regime as the initialization becomes smaller, and exiting from the kernel regime allows for smaller test error on the synthetic data. On MNIST data, Figure 6(c) shows that previously published successes with training very wide depth-2 ReLU networks without explicit regularization (*e.g.,* Neyshabur et al., 2014) relies on the initialization being small, *i.e.,* being outside of the kernel regime. In fact, the 2.4% test error reached for large initialization is no better than what can be achieved with a linear model over a random feature map. Turning to a more realistic network, 6(d) shows similar behavior when training a VGG11-like network on CIFAR10.

Interestingly, in all experiments, when $\alpha \approx 1$, the models both achieve good test error and are just about to enter the kernel regime, which may be desirable due to the learning vs. optimization tradeoffs discussed in Section 4. Not coincidentally, $\alpha = 1$ corresponds to using the standard out-of-the-box Uniform He initialization. Given the extensive efforts put into designing good initialization schemes, this gives further credence to the idea that model will perform best when trained just outside of the kernel regime.

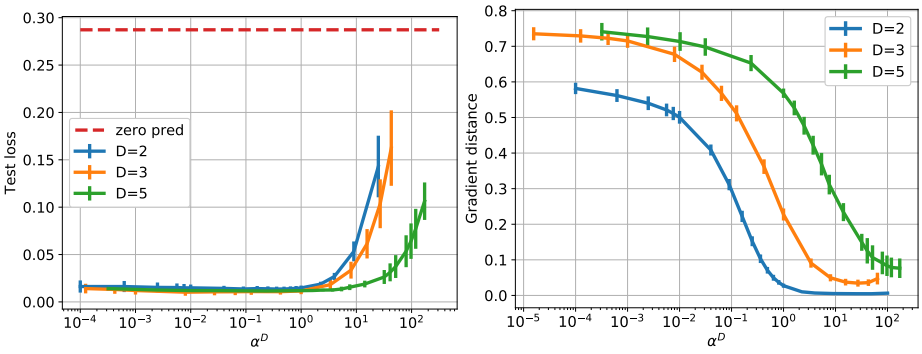
B.1. Univariate 2-layer ReLU Networks

Consider a two layer width-k ReLU network with univariate input $x \in \mathbb{R}$ given by $f((\mathbf{w}, \boldsymbol{b}), x) = \mathbf{w}_2\sigma(\mathbf{w}_1 x + \boldsymbol{b}_1) + \boldsymbol{b}_2$ where $\mathbf{w}_1 \in \mathbb{R}^{k\times 1}, \mathbf{w}_2 \in \mathbb{R}^{1\times k}$ and $\boldsymbol{b}_1 \in \mathbb{R}^{k\times 1}, \boldsymbol{b}_2 \in \mathbb{R}$ are the weights and bias parameters, respectively, for the two layers. This setting is the simplest non-linear model which has been explored in detail both theoretically and empirically (Savarese et al., 2019; Williams et al., 2019). Savarese et al. (2019) show that for an infinite width, univariate ReLU network, the minimal ℓ_2 parameter norm solution for a 1D regression problem, *i.e.,* $\arg\min_{\mathbf{w}} \|\mathbf{w}\|_2^2$ s.t. $\forall n$, $f((\mathbf{w}, \boldsymbol{b}), x_n) = y_n$ is given by a linear spline interpolation. We hypothesize that this bias to corresponds to the rich limit in training univariate 2-layer networks. In contrast, Theorem 5 and Corollary 6 of (Williams et al., 2019), show that the kernel limit corresponds to different cubic spline interpolations, where the exact form of interpolation depends on the relative scaling of weights across the layers. We explored the transition between the two regimes as the scale of initialization changes. We again consider a unbiased model as suggested by (Chizat et al., 2019) to avoid large outputs for large α.

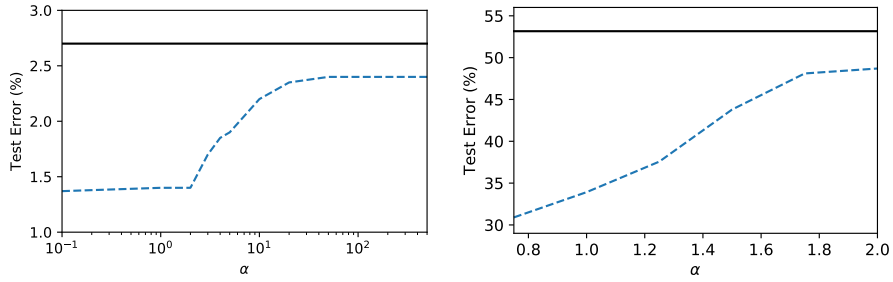
In Figure 7, we fix the width of the network to $k = 10000$ and empirically plot the functions learned with different initialization $\mathbf{w}(0) = \alpha\mathbf{w}_0$ for fixed $\mathbf{w}_0$. Additionally, we also demonstrate the effect of changing $\mathbf{w}_0$, by relatively scaling of layers *without changing the output* as shown in Figure 7-(b,c). First, as we suspected, we see that the rich limit of $\alpha \to 0$ indeed corresponds to linear spline interpolation and is indeed independent of the specific choice $\mathbf{w}_0$ as long as the outputs are unchanged. In contrast, as was also observed by (Williams et al., 2019), the kernel limit (large α), does indeed change as the relative scaling of the two layers changes, leading to what resembles different cubic splines.

B.2. Neural Network Experiment Details

Here, we provide further details about the neural network experiments.



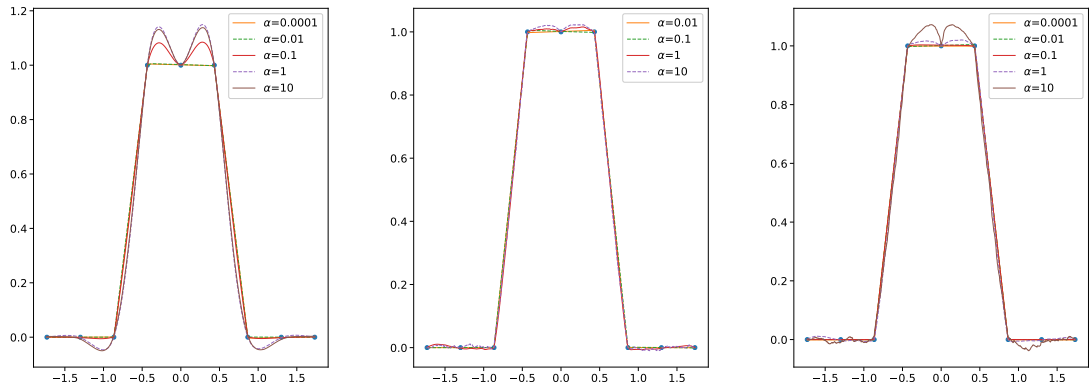
(a) Test RMSE vs scale (b) Grad distance vs scale



(c) MNIST test error vs scale (d) CIFAR10 test error vs scale

Figure 6: **Synthetic Data**: We generated a small regression training set in $\mathbb{R}^2$ by sampling 10 points uniformly from the unit circle, and labelling them with a 1 hidden layer teacher network with 3 hidden units. We trained depth-D, ReLU networks with 30 units per layer with squared loss using full GD and a small stepsize 0.01. The weights of the network are set using the Uniform He initialization, and then multiplied by α. The model is trained until ≈ 0 training loss. Shown in (a) and (b) are the test error and the "grad distance" vs. the depth-adjusted scale of the initialization, α^D. The grad distance is the cosine distance between the tangent kernel feature map at initialization versus at convergence. **MNIST**: We trained a depth-2, 5000 hidden unit ReLU network with cross-entropy loss using SGD until it reached 100% training accuracy. The stepsizes were optimally tuned w.r.t. validation error for each α individually. In (c), the dashed line shows the test error of the resulting network vs. α and the solid line shows the test error of the explicitly trained kernel predictor. **CIFAR10**: We trained a VGG11-like deep convolutional network with cross-entropy loss using SGD and a small stepsize 10^{-4} for 2000 epochs; all models reached 100% training accuracy. In (d), the dashed line shows the final test error vs. α. The solid line shows the test error of the explicitly trained kernel predictor. See Appendix B.2 for further details about all of the experiments.

Synthetic Experiments We construct a synthetic training set with $N = 10$ points drawn uniformly from the unit circle in $\mathbb{R}^2$ and labelled by a teacher model with 1 hidden layer of 3 units. We train fully connected ReLU networks with depths 2, 3, and 5 with 30 units per layer to minimize the square loss using full gradient descent with constant stepsize 0.01 until the training loss is below 10^{-9}. We use Uniform He initialization for the weights and then multiply them by α.



(a) $\mathbf{w}_0 = (\mathbf{w}_1^0, \boldsymbol{b}_1^0, \mathbf{w}_2^0)$ (b) $\mathbf{w}_0 = (k^{-0.5}\mathbf{w}_1^0, k^{-0.5}\boldsymbol{b}_1^0, k^{0.5}\mathbf{w}_2^0)$ (c) $\mathbf{w}_0 = (k^{-0.25}\mathbf{w}_1^0, k^{-0.25}\boldsymbol{b}_1^0, k^{0.25}\mathbf{w}_2^0)$

Figure 7: Each subplot has functions learned by univariate ReLU network of width $k = 10000$ with initialization $\mathbf{w}(0) = \alpha\mathbf{w}_0$, for some fixed $\mathbf{w}_0$. In Figure (a), $\mathbf{w}_0$ are fixed by a standard initialization scheme as $\mathbf{w}_1^0, \boldsymbol{b}_1^0 \sim \mathcal{N}(0, 1)$ and $\mathbf{w}_2^0 \sim \mathcal{N}(0, \sqrt{2/k})$ for second layer. In (b) and (c), the relative scaling of the layers in $\mathbf{w}_0$ is changed without changing the scale of the output.

Here, we describe the details of the neural network implementations for the MNIST and CIFAR10 experiments.

MNIST Since our theoretical results hold for the squared loss and gradient flow dynamics, here we empirically assess whether different regimes can be observed when training neural networks following standard practices.

We train a fully-connected neural network with a single hidden layer composed of 5000 units on the MNIST dataset, where weights are initialized as $\alpha\mathbf{w}_0$, $\mathbf{w}_0 \sim \mathcal{N}\left(0, \sqrt{\frac{2}{n_{in}}}\right)$, n_{in} denoting the number of units in the previous layer, as suggested by He et al. (2015). SGD with a batch size of 256 is used to minimize the cross-entropy loss over the 60000 training points, and error over the 10000 test samples are used as measure of generalization. For each value of α, we search over learning rates $(0.5, 0.01, 0.05, \dots)$ and use the one which resulted in best generalization.

There is a visible phase transition in Figure 6(c) in terms of generalization ($\approx 1.4\%$ error for $\alpha \leq 2$, and $\approx 2.4\%$ error for $\alpha \geq 50$), even though every network reached 100% training accuracy and less than 10^{-5} cross-entropy loss. The black line indicates the test error (2.7%) when training only the output layer of the network, as a proxy for the performance of a linear predictor with features given by a fixed, randomly-initialized hidden layer.

CIFAR10 We trained a VGG11-like architecture, which is as follows: 64-M-128-M-256-256-M-512-512-M-512-512-M-FC (numbers represent the number of channels in a convolution layers with no bias, M is a maxpooling layer, and FC is a fully connected layer). Weights were initialized using Uniform He initialization multiplied by α. No data augmentation was used, and training done using SGD with batch size of 128 and learning rate of 0.0001. All experiments ran for 2000 epochs, and reached 100% train accuracy except when training

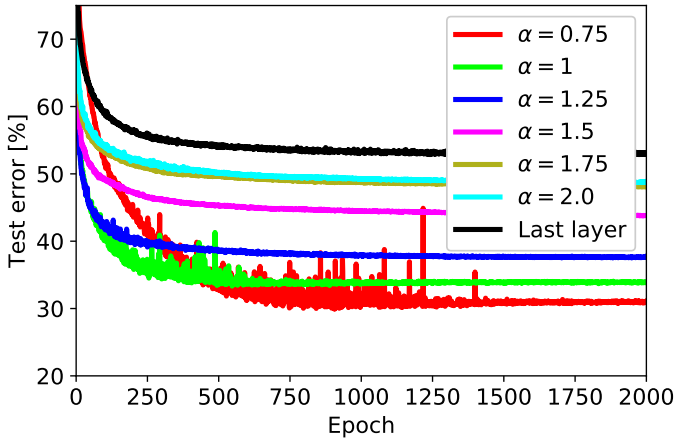
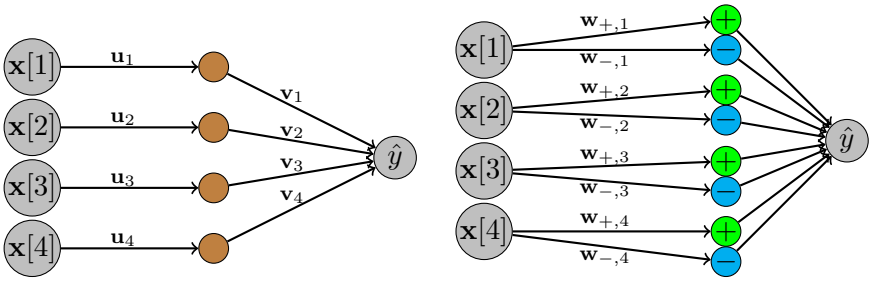


Figure 8: Training curves for the CIFAR10 experiments



(a) A biased diagonal network (b) An unbiased diagonal network

Figure 9: Diagonal linear networks.

only the last layer, which reached 50.38% train accuracy with LR = 0.001 (chosen after hyperparameter tuning).

In addition, to approximate the test error in the kernel regime, we experimented with freezing the bottom layers and only training the output layer for both datasets (the solid lines in Figures 6(c) and 6(d)).

Figure 8 illustrates some of the optimization difficulties that arise from using smaller α as discussed in Section 4.

Appendix C. Diagonal Linear Neural Networks

Consider the model $f((\boldsymbol{u}, \boldsymbol{v}), \mathbf{x}) = \sum_i \boldsymbol{u}_i \boldsymbol{v}_i \mathbf{x}_i$ as described in Section 4, and suppose that $|\mathbf{u}_i(0)| = |\mathbf{v}_i(0)|$, *i.e.,* the input and output weights for each hidden unit are initialized to have the same magnitude. Now, consider the gradient flow dynamics on the weights when minimizing the squared loss:

$$\frac{d}{dt}|\mathbf{u}(t)| = -\text{sign}(\mathbf{u}(t))\dot{\mathbf{u}}(t) \tag{11}$$

$$= -2\sum_{n=1}^{N}\left(\sum_{i=1}^{d} \boldsymbol{u}_i(t)\boldsymbol{v}_i(t)\mathbf{x}_i^{(n)} - \mathbf{y}^{(n)}\right)^2 \text{sign}(\mathbf{u}(t)) \circ \mathbf{v}(t) \circ \mathbf{x}^{(n)} \tag{12}$$

where $a \circ b$ denotes the element-wise multiplication of vectors a and b, and $\text{sign}(a)$ is the vector whose ith entry is $\text{sign}(a_i)$. Similarly,

$$\frac{d}{dt}|\mathbf{v}(t)| = -\text{sign}(\mathbf{v}(t))\dot{\mathbf{v}}(t) \tag{13}$$

$$= -2\sum_{n=1}^{N}\left(\sum_{i=1}^{d}\boldsymbol{u}_i(t)\boldsymbol{v}_i(t)\mathbf{x}_i^{(n)} - \mathbf{y}^{(n)}\right)^2 \text{sign}(\mathbf{v}(t)) \circ \mathbf{u}(t) \circ \mathbf{x}^{(n)} \tag{14}$$

Therefore, if $|\mathbf{u}_i(0)| = |\mathbf{v}_i(0)|$, then $\text{sign}(\mathbf{u}_i(0))\mathbf{v}_i(0) = \text{sign}(\mathbf{v}_i(0))\mathbf{u}_i(0)$, so the dynamics on $|\mathbf{u}_i|$ and $|\mathbf{v}_i|$ are the same, and their magnitudes will remain equal throughout training. Furthermore, the signs of the weights cannot change, since $|\mathbf{u}_i(t)| = |\mathbf{v}_i(t)| = 0$ implies $\dot{\mathbf{u}}_i(t) = \dot{\mathbf{v}}_i(t) = 0$.

Appendix D. Proof of Theorem 1

We prove Theorem 1 using the general approach outlined in Section 4.

Theorem 1 (General case) *For any $0 < \alpha < \infty$ and $\mathbf{w}_0$ with no zero entries, if the gradient flow solution $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0}$ satisfies $X\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} = \mathbf{y}$, then*

$$\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} = \arg\min_{\boldsymbol{\beta}} Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) \;\; s.t.\; X\boldsymbol{\beta} = \mathbf{y}, \tag{5}$$

where $Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \sum_{i=1}^{d}\alpha^2\mathbf{w}_{0,i}^2 q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2\mathbf{w}_{0,i}^2}\right)$ and $q(z) = 2 - \sqrt{4+z^2} + z\,\text{arcsinh}\left(\frac{z}{2}\right)$.

Proof We begin by calculating the gradient flow dynamics on $\mathbf{w}$, since the linear predictor $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0}$ is given by F applied to the limit of the gradient flow dynamics on $\mathbf{w}$. Recalling that $\tilde{X} = \begin{bmatrix} X & -X \end{bmatrix}$,

$$\dot{\mathbf{w}}_\alpha(t) = -\nabla L(\mathbf{w}_\alpha(t)) = -\nabla\left(\|\tilde{X}\mathbf{w}_\alpha(t)^2 - y\|_2^2\right) = -2\tilde{X}^\top r_\alpha(t) \circ \mathbf{w}_\alpha(t) \tag{15}$$

where the residual $r_\alpha(t) \triangleq \tilde{X}\mathbf{w}_\alpha(t)^2 - y$, and $a \circ b$ denotes the element-wise product of a and b. It is easily confirmed that these dynamics have a solution:

$$\mathbf{w}_\alpha(t) = \mathbf{w}_\alpha(0) \circ \exp\left(-2\tilde{X}^\top \int_0^t r_\alpha(s)ds\right) \tag{16}$$

Since $\mathbf{w}_{\alpha,+}(0) = \mathbf{w}_{\alpha,-}(0) = \alpha\mathbf{w}_0$ we can then express $\boldsymbol{\beta}_{\alpha,\mathbf{w}_0}(t)$ as

$$\begin{aligned}\boldsymbol{\beta}_{\alpha,\mathbf{w}_0}(t) &= \mathbf{w}_{\alpha,+}(t)^2 - \mathbf{w}_{\alpha,-}(t)^2 \\ &= \alpha^2\mathbf{w}_0^2 \circ \left(\exp\left(-4X^\top \int_0^t r_\alpha(s)ds\right) - \exp\left(4X^\top \int_0^t r_\alpha(s)ds\right)\right) \\ &= 2\alpha^2\mathbf{w}_0^2 \circ \sinh\left(-4X^\top \int_0^t r_\alpha(s)ds\right)\end{aligned} \tag{17}$$

Supposing also that $\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0}$ is a global minimum with zero error, *i.e.*, $X\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} = \mathbf{y}$. Thus,

$$\begin{aligned}X\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} &= \mathbf{y} \\ \boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{w}_0} &= b_\alpha(X^\top\nu)\end{aligned} \tag{18}$$

for $b_\alpha(z) = 2\alpha^2\mathbf{w}_0^2 \circ \sinh(z)$ and $\nu = -4\int_0^\infty r_\alpha(s)ds$. Following our general approach detailed in Section 4, we conclude

$$\nabla Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = b_\alpha^{-1}(\boldsymbol{\beta}) = \operatorname{arcsinh}\left(\frac{1}{2\alpha^2\mathbf{w}_0^2} \circ \boldsymbol{\beta}\right) \tag{19}$$

where we write $1/\mathbf{w}_0$ to denote the vector whose ith element is $1/\mathbf{w}_{0,i}$. Integrating this expression, we have that

$$Q_{\alpha,\mathbf{w}_0}(\boldsymbol{\beta}) = \sum_{i=1}^{d} \alpha^2\mathbf{w}_{0,i}^2 q\left(\frac{\boldsymbol{\beta}_i}{\alpha^2\mathbf{w}_{0,i}^2}\right) \tag{20}$$

where

$$q(z) = \int_0^z \operatorname{arcsinh}\left(\frac{t}{2}\right)dt = 2 - \sqrt{4+z^2} + z\operatorname{arcsinh}\left(\frac{z}{2}\right) \tag{21}$$

■

Appendix E. Proof of Theorem 2

Lemma 5 *For any $\boldsymbol{\beta} \in \mathbb{R}^d$,*

$$\alpha \le \alpha_1\left(\epsilon, \|\boldsymbol{\beta}\|_1, d\right) := \min\left\{1, \sqrt{\|\boldsymbol{\beta}\|_1}, \left(2\|\boldsymbol{\beta}\|_1\right)^{-\frac{1}{2\epsilon}}, \exp\left(-\frac{d}{2\epsilon\|\boldsymbol{\beta}\|_1}\right)\right\}$$

guarantees that

$$(1-\epsilon)\|\boldsymbol{\beta}\|_1 \le \frac{1}{\ln(1/\alpha^2)} Q_\alpha(\boldsymbol{\beta}) \le (1+\epsilon)\|\boldsymbol{\beta}\|_1$$

Proof We consider only the special case $\mathbf{w}_0 = \mathbf{1}$ and will drop the subscript for brevity. First, we show that $Q_\alpha(\boldsymbol{\beta}) = Q_\alpha(|\boldsymbol{\beta}|)$. Observe that $g(x) = x\arcsin(x/2)$ is even because x and $\arcsin(x/2)$ are odd. Therefore,

$$\begin{aligned}
Q_\alpha(\boldsymbol{\beta}) &= \alpha^2 \sum_{i=1}^{d} 2 - \sqrt{4 + \frac{\boldsymbol{\beta}_i^2}{\alpha^4}} + \frac{\boldsymbol{\beta}_i}{\alpha^2}\operatorname{arcsinh}\left(\frac{\boldsymbol{\beta}_i}{2\alpha^2}\right) \\
&= \alpha^2 \sum_{i=1}^{d} 2 - \sqrt{4 + \frac{\boldsymbol{\beta}_i^2}{\alpha^4}} + g\left(\frac{\boldsymbol{\beta}_i}{\alpha^2}\right) \\
&= \alpha^2 \sum_{i=1}^{d} 2 - \sqrt{4 + \frac{|\boldsymbol{\beta}_i|^2}{\alpha^4}} + g\left(\left|\frac{\boldsymbol{\beta}_i}{\alpha^2}\right|\right) \\
&= Q_\alpha(|\boldsymbol{\beta}|)
\end{aligned} \tag{22}$$

Therefore, we can rewrite

$$\begin{aligned}\frac{1}{\ln(1/\alpha^2)}Q_\alpha(\boldsymbol{\beta}) &= \frac{1}{\ln(1/\alpha^2)}Q_\alpha(|\boldsymbol{\beta}|)\\ &= \sum_{i=1}^{d}\frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{\sqrt{4\alpha^4+\boldsymbol{\beta}_i^2}}{\ln(1/\alpha^2)} + \frac{|\boldsymbol{\beta}_i|}{\ln(1/\alpha^2)}\operatorname{arcsinh}\left(\frac{|\boldsymbol{\beta}_i|}{2\alpha^2}\right)\\ &= \sum_{i=1}^{d}\frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{\sqrt{4\alpha^4+\boldsymbol{\beta}_i^2}}{\ln(1/\alpha^2)} + \frac{|\boldsymbol{\beta}_i|}{\ln(1/\alpha^2)}\ln\left(\frac{|\boldsymbol{\beta}_i|}{2\alpha^2}+\sqrt{1+\frac{\boldsymbol{\beta}_i^2}{4\alpha^4}}\right)\\ &= \sum_{i=1}^{d}\frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{\sqrt{4\alpha^4+\boldsymbol{\beta}_i^2}}{\ln(1/\alpha^2)} + |\boldsymbol{\beta}_i|\left(1+\frac{\ln\left(\frac{|\boldsymbol{\beta}_i|}{2}+\sqrt{\alpha^4+\frac{\boldsymbol{\beta}_i^2}{4}}\right)}{\ln(1/\alpha^2)}\right)\end{aligned} \tag{23}$$

Using the fact that

$$|a| \le \sqrt{a^2+b^2} \le |a|+|b| \tag{24}$$

we can bound for $\alpha < 1$

$$\begin{aligned}\frac{1}{\ln(1/\alpha^2)}Q_\alpha(\boldsymbol{\beta}) &\le \sum_{i=1}^{d}\frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{2\alpha^2}{\ln(1/\alpha^2)} + |\boldsymbol{\beta}_i|\left(1+\frac{\ln\left(\frac{|\boldsymbol{\beta}_i|}{2}+\alpha^2+\frac{|\boldsymbol{\beta}_i|}{2}\right)}{\ln(1/\alpha^2)}\right)\\ &= \sum_{i=1}^{d}|\boldsymbol{\beta}_i|\left(1+\frac{\ln\left(|\boldsymbol{\beta}_i|+\alpha^2\right)}{\ln(1/\alpha^2)}\right)\\ &\le \|\boldsymbol{\beta}\|_1\left(1+\max_{i\in[d]}\frac{\ln\left(|\boldsymbol{\beta}_i|+\alpha^2\right)}{\ln(1/\alpha^2)}\right)\\ &\le \|\boldsymbol{\beta}\|_1\left(1+\frac{\ln\left(\|\boldsymbol{\beta}\|_1+\alpha^2\right)}{\ln(1/\alpha^2)}\right)\end{aligned} \tag{25}$$

So, for any $\alpha \le \min\left\{1, \sqrt{\|\boldsymbol{\beta}\|_1}, (2\|\boldsymbol{\beta}\|_1)^{-\frac{1}{2\epsilon}}\right\}$, then

$$\begin{aligned}\frac{1}{\ln(1/\alpha^2)}Q_\alpha(\boldsymbol{\beta}) &\le \|\boldsymbol{\beta}\|_1\left(1+\frac{\ln\left(\|\boldsymbol{\beta}\|_1+\alpha^2\right)}{\ln(1/\alpha^2)}\right)\\ &\le \|\boldsymbol{\beta}\|_1\left(1+\frac{\ln\left(2\|\boldsymbol{\beta}\|_1\right)}{\ln(1/\alpha^2)}\right)\\ &\le \|\boldsymbol{\beta}\|_1\left(1+\epsilon\right)\end{aligned} \tag{26}$$

On the other hand, using (23) and (24) again,

$$\begin{aligned}\frac{1}{\ln(1/\alpha^2)}Q_\alpha(\boldsymbol{\beta}) &\ge \sum_{i=1}^{d}\frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{|\boldsymbol{\beta}_i|+2\alpha^2}{\ln(1/\alpha^2)} + |\boldsymbol{\beta}_i|\left(1+\frac{\ln\left(|\boldsymbol{\beta}_i|\right)}{\ln(1/\alpha^2)}\right)\\ &= \sum_{i=1}^{d}|\boldsymbol{\beta}_i|\left(1+\frac{\ln\left(|\boldsymbol{\beta}_i|\right)-1}{\ln(1/\alpha^2)}\right)\end{aligned} \tag{27}$$

Using the inequality $\ln(x) \geq 1 - \frac{1}{x}$, this can be further lower bounded by

$$\begin{aligned} \frac{1}{\ln(1/\alpha^2)} Q_\alpha(\boldsymbol{\beta}) &\geq \sum_{i=1}^{d} |\boldsymbol{\beta}_i| - \frac{1}{\ln(1/\alpha^2)} \\ &= \|\boldsymbol{\beta}\|_1 - \frac{d}{\ln(1/\alpha^2)} \end{aligned} \tag{28}$$

Therefore, for any $\alpha \leq \exp\left(-\frac{d}{2\epsilon\|\boldsymbol{\beta}\|_1}\right)$ then

$$\frac{1}{\ln(1/\alpha^2)} Q_\alpha(\boldsymbol{\beta}) \geq \|\boldsymbol{\beta}\|_1 (1-\epsilon) \tag{29}$$

We conclude that for $\alpha \leq \min\left\{1, \sqrt{\|\boldsymbol{\beta}\|_1}, (2\|\boldsymbol{\beta}\|_1)^{-\frac{1}{2\epsilon}}, \exp\left(-\frac{d}{2\epsilon\|\boldsymbol{\beta}\|_1}\right)\right\}$ that

$$(1-\epsilon)\|\boldsymbol{\beta}\|_1 \leq \frac{1}{\ln(1/\alpha^2)} Q_\alpha(\boldsymbol{\beta}) \leq (1+\epsilon)\|\boldsymbol{\beta}\|_1 \tag{30}$$

■

Lemma 6 *Fix any $\epsilon > 0$ and $d \geq \max\{e, 12^{4\epsilon}\}$. Then for any $\alpha \geq d^{-\frac{1}{4}-\frac{1}{8\epsilon}}$, $Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \not\propto \|\boldsymbol{\beta}\|_1$ in the sense that there exist vectors v, w such that*

$$\frac{Q_{\alpha,\mathbf{1}}(v)}{\|v\|_1} \geq (1+\epsilon)\frac{Q_{\alpha,\mathbf{1}}(w)}{\|w\|_1}$$

Proof First, recall that

$$\begin{aligned} q\left(\frac{1}{c\alpha^2}\right) &= 2 - \sqrt{4 + \frac{1}{c^2\alpha^4}} + \frac{1}{c\alpha^2}\operatorname{arcsinh}\left(\frac{1}{2c\alpha^2}\right) \\ &= \frac{1}{c\alpha^2}\left(2c\alpha^2 - \sqrt{4c^2\alpha^4+1} + \ln\left(\frac{1}{2c\alpha^2} + \sqrt{1 + \frac{1}{4c^2\alpha^4}}\right)\right) \\ &= \frac{1}{c\alpha^2}\left(2c\alpha^2 - \sqrt{4c^2\alpha^4+1} + \ln\left(\frac{1}{c\alpha^2}\right) + \ln\left(\frac{1}{2} + \sqrt{c^2\alpha^4 + \frac{1}{2}}\right)\right) \end{aligned} \tag{31}$$

Thus,

$$-1 + \ln\left(\frac{1}{c\alpha^2}\right) \leq c\alpha^2 q\left(\frac{1}{c\alpha^2}\right) \leq 3c\alpha^2 - 1 + \ln\left(\frac{1}{c\alpha^2}\right) \tag{32}$$

Now, consider the ratio

$$\frac{Q_{\alpha,\mathbf{1}}(e_1)}{Q_{\alpha,\mathbf{1}}\left(\frac{\mathbf{1}_d}{\|\mathbf{1}_d\|_2}\right)} = \frac{\alpha^2 q\left(\frac{1}{\alpha^2}\right)}{\alpha^2 d q\left(\frac{1}{\alpha^2\sqrt{d}}\right)} = \frac{1}{\sqrt{d}}\frac{\alpha^2 q\left(\frac{1}{\alpha^2}\right)}{\alpha^2\sqrt{d} q\left(\frac{1}{\alpha^2\sqrt{d}}\right)} \tag{33}$$

Using (32), we conclude

$$\begin{aligned}\sqrt{d}\frac{Q_{\alpha,\mathbf{1}}\left(e_1\right)}{Q_{\alpha,\mathbf{1}}\left(\frac{\mathbf{1}_d}{\|\mathbf{1}_d\|_2}\right)} &\geq \frac{-1+\ln\left(\frac{1}{\alpha^2}\right)}{3\sqrt{d}\alpha^2-1+\ln\left(\frac{1}{\alpha^2\sqrt{d}}\right)} \\ &= \frac{-1+\ln\left(\frac{1}{\alpha^2}\right)}{3\sqrt{d}\alpha^2-1+\ln\left(\frac{1}{\alpha^2}\right)-\frac{1}{2}\ln(d)} \\ &= 1+\frac{\ln(d)-6\sqrt{d}\alpha^2}{6\sqrt{d}\alpha^2-2+2\ln\left(\frac{1}{\alpha^2\sqrt{d}}\right)}\end{aligned} \tag{34}$$

Fix any $\epsilon>0$ and $d\geq\max\left\{e,12^{4\epsilon}\right\}$, and set $\alpha=d^{-\frac{1}{4}-\frac{1}{8\epsilon}}$. Then,

$$\begin{aligned}& 2d^{\frac{1}{4\epsilon}}\geq 6 \quad \text{and} \quad \frac{d^{\frac{1}{4\epsilon}}}{2}\ln d\geq 6 \\ \implies & 2d^{\frac{1}{4\epsilon}}-6+\frac{d^{\frac{1}{4\epsilon}}}{2\epsilon}\ln d-\frac{6}{\epsilon}\geq 0 \\ \implies & \left(\frac{1}{\epsilon}-\frac{1}{2\epsilon}\right)\ln d-\frac{6}{\epsilon}d^{-\frac{1}{4\epsilon}}\geq 6d^{-\frac{1}{4\epsilon}}-2 \\ \implies & \frac{1}{\epsilon}\ln d-\frac{6}{\epsilon}d^{-\frac{1}{4\epsilon}}\geq 6d^{-\frac{1}{4\epsilon}}-2+\frac{1}{2\epsilon}\ln d \\ \implies & \ln\left(d\right)-6\alpha^2\sqrt{d}\geq\epsilon\left(6\alpha^2\sqrt{d}-2+2\ln\left(\frac{1}{\alpha^2\sqrt{d}}\right)\right)\end{aligned} \tag{35}$$

This implies that the second term of (34) is at least ϵ. We conclude that for any $\epsilon>0$ and $d\geq\max\left\{e,12^{4\epsilon}\right\}$, $\alpha=d^{-\frac{1}{4}-\frac{1}{8\epsilon}}$ implies that

$$\frac{Q_{\alpha,\mathbf{1}}\left(e_1\right)}{Q_{\alpha,\mathbf{1}}\left(\frac{\mathbf{1}_d}{\|\mathbf{1}_d\|_2}\right)}\geq(1+\epsilon)\frac{\|e_1\|_1}{\|\frac{\mathbf{1}_d}{\|\mathbf{1}_d\|_2}\|_1} \tag{36}$$

Consequently, for at least one of these two vectors, Q is not proportional to the ℓ_1 norm up to accuracy $O(\epsilon)$ for this value of α. Since

$$\frac{d}{d\alpha}\frac{Q_{\alpha,\mathbf{1}}\left(e_1\right)}{Q_{\alpha,\mathbf{1}}\left(\frac{\mathbf{1}_d}{\|\mathbf{1}_d\|_2}\right)}\geq 0 \tag{37}$$

this conclusion applies also for larger α. ■

Lemma 7 *For any $\boldsymbol{\beta}\in\mathbb{R}^d$,*

$$\alpha\geq\alpha_2(\epsilon,\|\boldsymbol{\beta}\|_2,d):=\sqrt{\|\boldsymbol{\beta}\|_2}\left(1+\epsilon^{-\frac{1}{4}}\right)$$

guarantees that

$$(1-\epsilon)\|\boldsymbol{\beta}\|_2^2\leq 4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta})\leq(1+\epsilon)\|\boldsymbol{\beta}\|_2^2$$

Proof The regularizer $Q_{\alpha,\mathbf{1}}$ can be written

$$Q_{\alpha,\mathbf{1}}(\beta) = \alpha^2 \sum_{i=1}^{d} \int_0^{\beta_i/\alpha^2} \operatorname{arcsinh}\left(\frac{t}{2}\right) dt \tag{38}$$

Let $\phi(z) = \int_0^{z/\alpha^2} \operatorname{arcsinh}\left(\frac{t}{2}\right) dt$, then

$$\begin{aligned}
\phi(0) &= 0 \\
\phi'(0) &= \frac{1}{\alpha^2} \operatorname{arcsinh}\left(\frac{z}{2\alpha^2}\right)\Big|_{z=0} = 0 \\
\phi''(0) &= \frac{1}{\alpha^4 \sqrt{4 + \frac{z^2}{\alpha^4}}}\Bigg|_{z=0} = \frac{1}{2\alpha^4} \\
\phi'''(0) &= \frac{-z}{\alpha^8 \left(4 + \frac{z^2}{\alpha^4}\right)^{3/2}}\Bigg|_{z=0} = 0 \\
\phi''''(z) &= \frac{3z^2}{\alpha^{12}\left(4 + \frac{z^2}{\alpha^4}\right)^{5/2}} - \frac{1}{\alpha^8 \left(4 + \frac{z^2}{\alpha^4}\right)^{3/2}}
\end{aligned} \tag{39}$$

Also, note that

$$\begin{aligned}
|\phi''''(z)| &= \frac{|2z^2 - 4\alpha^4|}{\alpha^{12}\left(4 + \frac{z^2}{\alpha^4}\right)^{5/2}} \\
&\leq \frac{z^2 + 2\alpha^4}{16\alpha^{12}}
\end{aligned} \tag{40}$$

Therefore, by Taylor's theorem, for some ξ with $|\xi| \leq |z|$

$$\begin{aligned}
&\left|\phi(z) - \frac{z^2}{4\alpha^4}\right| = \frac{\phi''''(\xi)}{4!} z^4 \\
\implies &\left|\phi(z) - \frac{z^2}{4\alpha^4}\right| \leq \sup_{|\xi| \leq |z|} \frac{\phi''''(\xi)}{4!} z^4 \leq \frac{z^6 + 2\alpha^4 z^4}{384\alpha^{12}} = \frac{z^2}{4\alpha^4} \frac{z^4 + 2\alpha^4 z^2}{96\alpha^8}
\end{aligned} \tag{41}$$

Therefore, for any $\boldsymbol{\beta} \in \mathbb{R}^d$,

$$\begin{aligned}
\left|4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) - \|\boldsymbol{\beta}\|_2^2\right| &= 4\alpha^4 \left|\sum_{i=1}^{d} \phi(\boldsymbol{\beta}_i) - \frac{\boldsymbol{\beta}_i^2}{4\alpha^4}\right| \\
&\leq 4\alpha^4 \sum_{i=1}^{d} \left|\phi(\boldsymbol{\beta}_i) - \frac{\boldsymbol{\beta}_i^2}{4\alpha^4}\right| \\
&\leq \sum_{i=1}^{d} \boldsymbol{\beta}_i^2 \cdot \frac{\boldsymbol{\beta}_i^4 + 2\alpha^4 \boldsymbol{\beta}_i^2}{96\alpha^8} \\
&\leq \|\boldsymbol{\beta}\|_2^2 \max_i \frac{\boldsymbol{\beta}_i^4 + 2\alpha^4 \boldsymbol{\beta}_i^2}{96\alpha^8}
\end{aligned} \tag{42}$$

Therefore, $\alpha \geq \sqrt{\|\boldsymbol{\beta}\|_2}\left(1+\epsilon^{-\frac{1}{4}}\right)$ ensures

$$(1-\epsilon)\|\boldsymbol{\beta}\|_2^2 \leq 4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \leq (1+\epsilon)\|\boldsymbol{\beta}\|_2^2 \tag{43}$$

■

Theorem 2 *For any $0<\epsilon<d$, under the setting of Theorem 1 with $\mathbf{w}_0=\mathbf{1}$,*

$$\alpha \leq \min\left\{\left(2(1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1\right)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-d/(\epsilon\|\boldsymbol{\beta}^*_{\ell_1}\|_1)\right)\right\} \implies \|\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}\|_1 \leq (1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$$
$$\alpha \geq \sqrt{2(1+\epsilon)(1+2/\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2} \implies \|\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}\|_2^2 \leq (1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2^2$$

Proof We prove the ℓ_1 and ℓ_2 statements separately.

ℓ_1 approximation First, we will prove that $\|\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}\|_1 < (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$. By Lemma 5, since $\alpha \leq \alpha_1\left(\frac{\epsilon}{2+\epsilon}, (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1, d\right)$, for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_1 \leq (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$ we have

$$\left(1-\frac{\epsilon}{2+\epsilon}\right)\|\boldsymbol{\beta}\|_1 \leq \frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \leq \left(1+\frac{\epsilon}{2+\epsilon}\right)\|\boldsymbol{\beta}\|_1 \tag{44}$$

Let $\boldsymbol{\beta}$ be such that $X\boldsymbol{\beta}=y$ and $\|\boldsymbol{\beta}\|_1=(1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$. Then

$$\begin{aligned}
\frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) &\geq \left(1-\frac{\epsilon}{2+\epsilon}\right)\|\boldsymbol{\beta}\|_1 \\
&= \left(1-\frac{\epsilon}{2+\epsilon}\right)(1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1 \\
&\geq \frac{\left(1-\frac{\epsilon}{2+\epsilon}\right)}{\left(1+\frac{\epsilon}{2+\epsilon}\right)}(1+2\epsilon)\frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_1}) \\
&= \frac{1+2\epsilon}{1+\epsilon}\frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_1}) \\
&> \frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_1}) \\
&\geq \frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}})
\end{aligned} \tag{45}$$

Therefore, $\boldsymbol{\beta} \neq \boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}}$. Furthermore, let $\boldsymbol{\beta}$ be any solution $X\boldsymbol{\beta}=y$ with $\|\boldsymbol{\beta}\|_1 > (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$. It is easily confirmed that there exists $c\in(0,1)$ such that the point $\boldsymbol{\beta}'=(1-c)\boldsymbol{\beta}+c\boldsymbol{\beta}^*_{\ell_1}$ is satisfies both $X\boldsymbol{\beta}'=y$ and $\|\boldsymbol{\beta}'\|_1=(1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$. By the convexity of Q, this implies $Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \geq Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}') > Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^\infty_{\alpha,\mathbf{1}})$. Thus a $\boldsymbol{\beta}$ with a large ℓ_1 norm cannot be a solution, even if $\frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \not\approx \|\boldsymbol{\beta}\|_1$.

Since $\|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_1 < (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1$, we conclude

$$\begin{aligned}\|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_1 &\leq \frac{1}{1-\frac{\epsilon}{2+\epsilon}}\frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}})\\ &\leq \frac{1}{1-\frac{\epsilon}{2+\epsilon}}\frac{1}{\ln(1/\alpha^2)}Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_1})\\ &\leq \frac{1+\frac{\epsilon}{2+\epsilon}}{1-\frac{\epsilon}{2+\epsilon}}\|\boldsymbol{\beta}^*_{\ell_1}\|_1\\ &= (1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_1}\|_1\end{aligned} \tag{46}$$

Next, we prove $\|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_2 < (1+2\epsilon)\,\|\boldsymbol{\beta}^*_{\ell_2}\|_2$. By Lemma 7, since $\alpha \geq \alpha_2\left(\frac{\epsilon}{2+\epsilon}, (1+2\epsilon)\,\|\boldsymbol{\beta}^*_{\ell_2}\|_2\right)$, for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_2 \leq (1+2\epsilon)\,\|\boldsymbol{\beta}^*_{\ell_2}\|_2$ we have

$$\|\boldsymbol{\beta}\|_2^2\left(1-\frac{\epsilon}{2+\epsilon}\right) \leq 4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \leq \|\boldsymbol{\beta}\|_2^2\left(1+\frac{\epsilon}{2+\epsilon}\right) \tag{47}$$

Let $\boldsymbol{\beta}$ be such that $X\boldsymbol{\beta} = y$ and $\|\boldsymbol{\beta}\|_2 = (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2$. Then,

$$\begin{aligned}4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) &\geq \left(1-\frac{\epsilon}{2+\epsilon}\right)\|\boldsymbol{\beta}\|_2^2\\ &= \left(1-\frac{\epsilon}{2+\epsilon}\right)(1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2^2\\ &\geq \frac{\left(1-\frac{\epsilon}{2+\epsilon}\right)}{\left(1+\frac{\epsilon}{2+\epsilon}\right)}(1+2\epsilon)4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_2})\\ &= \frac{1+2\epsilon}{1+\epsilon}4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_2})\\ &> 4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_2})\\ &\geq 4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}})\end{aligned} \tag{48}$$

Therefore, $\boldsymbol{\beta} \neq \boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}$. Furthermore, let $\boldsymbol{\beta}$ be any solution $X\boldsymbol{\beta} = y$ with $\|\boldsymbol{\beta}\|_2 > (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2$. It is easily confirmed that there exists $c \in (0,1)$ such that the point $\boldsymbol{\beta}' = (1-c)\boldsymbol{\beta} + c\boldsymbol{\beta}^*_{\ell_2}$ satisfies $X\boldsymbol{\beta}' = y$ and $\|\boldsymbol{\beta}'\|_2 = (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2$. By the convexity of $Q_{\alpha,\mathbf{1}}$, this implies $Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \geq Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}') > Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_2})$. Thus a $\boldsymbol{\beta}$ with a large ℓ_2 norm cannot be a solution, even if $4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}) \not\approx \|\boldsymbol{\beta}\|_2^2$.

Since $\|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_2 < (1+2\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2$, we conclude

$$\begin{aligned}\|\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}}\|_2^2 &\leq \frac{1}{1-\frac{\epsilon}{2+\epsilon}}4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^{\infty}_{\alpha,\mathbf{1}})\\ &\leq \frac{1}{1-\frac{\epsilon}{2+\epsilon}}4\alpha^2 Q_{\alpha,\mathbf{1}}(\boldsymbol{\beta}^*_{\ell_2})\\ &\leq \frac{1+\frac{\epsilon}{2+\epsilon}}{1-\frac{\epsilon}{2+\epsilon}}\|\boldsymbol{\beta}^*_{\ell_2}\|_2^2\\ &= (1+\epsilon)\|\boldsymbol{\beta}^*_{\ell_2}\|_2^2\end{aligned} \tag{49}$$

■

Appendix F. Proof of Theorem 3

Lemma 8 *For $D > 2$ and the D-homogeneous model (10),*

$$\forall_t \left\| X^\top \int_0^t r(\tau) d\tau \right\|_\infty \leq \frac{\alpha^{2-D}}{D(D-2)}$$

Proof For the order-D unbiased model $\boldsymbol{\beta}(t) = \mathbf{w}_+^D - \mathbf{w}_-^D$, the gradient flow dynamics are

$$\dot{\mathbf{w}}_+(t) = -\frac{dL}{d\mathbf{w}_+} = -DX^\top r(t) \circ \mathbf{w}_+^{D-1}(t), \quad \mathbf{w}_+(0) = \alpha 1 \tag{50}$$

$$\implies \mathbf{w}_+(t) = \left(\alpha^{2-D} 1 + D(D-2) X^\top \int_0^t r(\tau) d\tau \right)^{-\frac{1}{D-2}} \tag{51}$$

Where $\circ$ denotes elementwise multiplication, $r(t) = X\boldsymbol{\beta}(t) - y$, and where all exponentiation is elementwise. Similarly,

$$\dot{\mathbf{w}}_-(t) = -\frac{dL}{d\mathbf{w}_-} = DX^\top r(t) \circ \mathbf{w}_-^{D-1}(t), \quad \mathbf{w}_-(0) = \alpha 1 \tag{52}$$

$$\implies \mathbf{w}_-(t) = \left(\alpha^{2-D} 1 - D(D-2) X^\top \int_0^t r(\tau) d\tau \right)^{-\frac{1}{D-2}} \tag{53}$$

First, we observe that $\forall_t \forall_i \ \mathbf{w}_+(t)_i \geq 0$ and $\forall_t \forall_i \ \mathbf{w}_-(t)_i \geq 0$. This is because at time 0, $\mathbf{w}_+(0)_i = \mathbf{w}_-(0)_i = \alpha > 0$; the gradient flow dynamics are continuous; and $\mathbf{w}_+(t)_i = 0 \implies \dot{\mathbf{w}}_+(t)_i = 0$ and $\mathbf{w}_-(t)_i = 0 \implies \dot{\mathbf{w}}_-(t)_i = 0$.

Consequently,

$$\begin{aligned} 0 \leq \mathbf{w}_+(t)_i^{2-D} &= \alpha^{2-D} + D(D-2) \left[X^\top \int_0^t r(\tau) d\tau \right]_i \\ 0 \leq \mathbf{w}_-(t)_i^{2-D} &= \alpha^{2-D} - D(D-2) \left[X^\top \int_0^t r(\tau) d\tau \right]_i \\ \implies -\alpha^{2-D} &\leq D(D-2) \left[X^\top \int_0^t r(\tau) d\tau \right]_i \leq \alpha^{2-D} \end{aligned} \tag{54}$$

which concludes the proof. ■

Theorem 3 *For any $0 < \alpha < \infty$ and $D \geq 3$, if $X\boldsymbol{\beta}_{\alpha,D}^\infty = y$, then*

$$\boldsymbol{\beta}_{\alpha,D}^\infty = \arg\min_{\boldsymbol{\beta}} Q_\alpha^D(\boldsymbol{\beta}) \quad s.t. \quad \mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

where $Q_\alpha^D(\boldsymbol{\beta}) = \alpha^D \sum_{i=1}^d q_D(\boldsymbol{\beta}_i/\alpha^D)$ and $q_D = \int h_D^{-1}$ is the antiderivative of the unique inverse of $h_D(z) = (1-z)^{-\frac{D}{D-2}} - (1+z)^{-\frac{D}{D-2}}$ on $[-1, 1]$. Furthermore, $\lim_{\alpha\to 0} \boldsymbol{\beta}_{\alpha,D}^\infty = \boldsymbol{\beta}_{\ell_1}^*$ and $\lim_{\alpha\to\infty} \boldsymbol{\beta}_{\alpha,D}^\infty = \boldsymbol{\beta}_{\ell_2}^*$.

Proof For the order-D unbiased model $\boldsymbol{\beta}(t) = \mathbf{w}_+^D - \mathbf{w}_-^D$, the gradient flow dynamics are

$$\dot{\mathbf{w}}(t) = \frac{dL}{d\mathbf{w}} = -D\tilde{X}^\top r(t) \circ \mathbf{w}^{D-1}, \quad \mathbf{w}(0) = \alpha 1 \tag{55}$$

$$\implies \mathbf{w}(t) = \left(\alpha^{2-D} + D(D-2)\tilde{X}^\top \int_0^t r(\tau)d\tau\right)^{-\frac{1}{D-2}} \tag{56}$$

$$\implies \boldsymbol{\beta}(t) = \alpha^D \left(1 + \alpha^{D-2}D(D-2)X^\top \int_0^t r(\tau)d\tau\right)^{-\frac{D}{D-2}} - \alpha^D \left(1 - \alpha^{D-2}D(D-2)X^\top \int_0^t r(\tau)d\tau\right)^{-\frac{D}{D-2}} \tag{57}$$

where $\tilde{X} = [X \quad -X]$ and $r(t) = X\boldsymbol{\beta}(t) - y$. Supposing $\boldsymbol{\beta}(t)$ converges to a zero-error solution,

$$X\boldsymbol{\beta}(\infty) = \mathbf{y} \qquad \text{and} \qquad \boldsymbol{\beta}(\infty) = \alpha^D h_D(X^\top \nu(\infty)) \tag{58}$$

where $\nu(\infty) = -\alpha^{D-2}D(D-2)\int_0^\infty r(\tau)d\tau$ and the function h_D is applied elementwise and is defined

$$h_D(z) = (1-z)^{-\frac{D}{D-2}} - (1+z)^{-\frac{D}{D-2}} \tag{59}$$

By Lemma 8, $\|X^\top \nu\|_\infty \leq 1$, so the domain of h_D is the interval $[-1, 1]$, upon which it is monotonically increasing from $h_D(-1) = -\infty$ to $h_D(1) = \infty$. Therefore, there exists an inverse mapping $h_D^{-1}(t)$ with domain $[-\infty, \infty]$ and range $[-1, 1]$.

This inverse mapping unfortunately does not have a simple closed form. Nevertheless, it is the root of a rational equation. Following the general approach outlined in Section 4, we conclude:

$$Q_\alpha^D(\boldsymbol{\beta}) = \alpha^D \sum_i \int_0^{\boldsymbol{\beta}_i/\alpha^D} h_D^{-1}(t)dt \tag{60}$$

Rich Limit Next, we show that if gradient flow reaches a solution $X\boldsymbol{\beta}_{\alpha,D}^\infty = y$, then $\lim_{\alpha\to 0} \boldsymbol{\beta}_{\alpha,D}^\infty = \boldsymbol{\beta}_{\ell_1}^*$ for any D. This is implied by the work of Arora et al. (2019a), but we include it here for an alternative, simpler proof for our special case, and for completeness's sake.

The KKT conditions for $\boldsymbol{\beta} = \boldsymbol{\beta}_{\ell_1}^*$ are $X\boldsymbol{\beta} = y$ and $\exists \nu$ $\text{sign}(\boldsymbol{\beta}) = X^\top \nu$ (where $\text{sign}(0) = [-1, 1]$). The first condition is satisfied by assumption. Define ν as above. We will demonstrate that the second condition holds too in the limit as $\alpha \to 0$.

First, by Lemma 8, $\|X^\top \nu\|_\infty \leq 1$ for all α and D. Thus, for any coordinates i such that $\lim_{\alpha\to 0}[\boldsymbol{\beta}_{\alpha,D}^\infty]_i = 0$, the second KKT condition holds. Consider now i for which $\lim_{\alpha\to 0}[\boldsymbol{\beta}_{\alpha,D}^\infty]_i > 0$. As shown above,

$$\lim_{\alpha\to 0}[\boldsymbol{\beta}_{\alpha,D}^\infty]_i = \lim_{\alpha\to 0} \alpha^D \left(1 - [X^\top \nu]_i\right)^{-\frac{D}{D-2}} - \alpha^D \left(1 + [X^\top \nu]_i\right)^{-\frac{D}{D-2}} > 0 \tag{61}$$

$$\implies \lim_{\alpha\to 0} \alpha^D \left(1 - [X^\top \nu]_i\right)^{-\frac{D}{D-2}} > 0 \tag{62}$$

This and $[X^\top \nu]_i \leq 1$ implies $\lim_{\alpha\to 0}[X^\top \nu]_i = 1$, and thus the positive coordinates satisfy the second KKT condition. An identical argument can be made for the negative coordinates.

Kernel Regime Finally, we show that if gradient flow reaches a solution $X\boldsymbol{\beta}^{\infty}_{\alpha,D} = y$, then $\lim_{\alpha\to\infty} \boldsymbol{\beta}^{\infty}_{\alpha,D} = \boldsymbol{\beta}^{*}_{\ell_2}$ for any D.

First, since X and y are finite, there exists a solution $\boldsymbol{\beta}^*$ whose entries are all finite, and thus all the entries of $\boldsymbol{\beta}^{\infty}_{\alpha,D}$, which is the Q^D_α-minimizing solution, will be finite.

The KKT conditions for $\boldsymbol{\beta} = \boldsymbol{\beta}^{*}_{\ell_2}$ are $X\boldsymbol{\beta} = y$ and $\exists \mu\ \boldsymbol{\beta} = X^\top \mu$. The first condition is satisfied by assumption. Defining ν as above, we have

$$\lim_{\alpha\to\infty} [\boldsymbol{\beta}^{\infty}_{\alpha,D}]_i = \lim_{\alpha\to\infty} \alpha^D \left(1 - [X^\top \nu]_i\right)^{-\frac{D}{D-2}} - \alpha^D \left(1 + [X^\top \nu]_i\right)^{-\frac{D}{D-2}} < \infty \tag{63}$$

$$\implies \lim_{\alpha\to\infty} [X^\top \nu]_i = 0 \tag{64}$$

Consequently, defining $\mu = \frac{2D\alpha^D}{D-2}\nu$, and observing that for small z,

$$(1-z)^{-\frac{D}{D-2}} - (1+z)^{-\frac{D}{D-2}} = \frac{2D}{D-2} z + O(z^3) \tag{65}$$

we conclude

$$\begin{aligned}
\lim_{\alpha\to\infty} \frac{[\boldsymbol{\beta}^{\infty}_{\alpha,D}]_i}{[X^\top \mu]_i} &= \lim_{\alpha\to\infty} \frac{\alpha^D \left(1 - [X^\top \nu]_i\right)^{-\frac{D}{D-2}} - \alpha^D \left(1 + [X^\top \nu]_i\right)^{-\frac{D}{D-2}}}{[X^\top \mu]_i} \\
&= \lim_{\alpha\to\infty} \frac{\alpha^D \left(\frac{2D}{D-2}[X^\top \nu]_i + O([X^\top \nu]_i^3)\right)}{\frac{2D\alpha^D}{D-2}[X^\top \nu]_i} \\
&= 1 + \lim_{\alpha\to\infty} O([X^\top \nu]_i^2) \\
&= 1
\end{aligned} \tag{66}$$

Thus, the KKT conditions are satisfied for $\lim_{\alpha\to\infty} \boldsymbol{\beta}^{\infty}_{\alpha,D} = \boldsymbol{\beta}^{*}_{\ell_2}$. ■

Appendix G. Proof of Theorem 4

Here, we prove Theorem 4:

Theorem 4 *Let $k \to \infty$, $\sigma(k) \to 0$, and $\mu^2 := \frac{1}{2}\lim_{k\to\infty} \sigma(k)\sqrt{k}$, and suppose $\mathbf{X}_1, \ldots, \mathbf{X}_N$ commute. If $\boldsymbol{M}_{\mathbf{U},\mathbf{V}}(t)$ converges to a zero error solution $\boldsymbol{M}^*_{\mathbf{U},\mathbf{V}}$, then*

$$\boldsymbol{M}^*_{\mathbf{U},\mathbf{V}} = \arg\min_{\mathbf{M}} Q_\mu(\textit{spectrum}(\mathbf{M}))\quad \textit{s.t. } L(\mathbf{M}) = 0$$

Proof As $k \to \infty$, $\bar{\mathbf{M}}_{\mathbf{U}(0),\mathbf{V}(0)} \to 2\mu^2 I$, so the four $d \times d$ submatrices of the lifted matrix $\bar{\mathbf{M}}_{\mathbf{U}(0),\mathbf{V}(0)}$ have diagonal structure. The dynamics on $\bar{\mathbf{M}}_{\mathbf{U}(t),\mathbf{V}(t)}$ are linear combination of terms of the form $\bar{\mathbf{M}}_{\mathbf{U}(t),\mathbf{V}(t)}\mathbf{X}_n + \mathbf{X}_n\bar{\mathbf{M}}_{\mathbf{U}(t),\mathbf{V}(t)}$, and each of these terms will share this same block-diagonal structure, which is therefore maintained throughout the course of optimization. We thus restrict our attention to just the main diagonal of $\bar{\mathbf{M}}_{\mathbf{U}(t),\mathbf{V}(t)}$ and the diagonal of $\mathbf{M}_{\mathbf{U}(t),\mathbf{V}(t)}$, all other entries will remain zero. In fact, we only need to track $\Delta(t) := \frac{1}{2}\operatorname{diag}\big(\mathbf{U}(t)\mathbf{U}(t)^\top + \mathbf{V}(t)\mathbf{V}(t)^\top\big) \in \mathbb{R}^d$ and $\delta(t) = \operatorname{diag}\big(\mathbf{U}(t)\mathbf{V}(t)^\top\big) \in \mathbb{R}^d$, with the goal of understanding $\lim_{t\to\infty} \delta(t)$.

Since the dynamics of $\bar{\mathbf{M}}_{\mathbf{U}(t),\mathbf{V}(t)}$ depend only on the observations and $\bar{\mathbf{M}}_{\mathbf{U}(t),\mathbf{V}(t)}$ itself, and *not* on the underlying parameters, we can understand the implicit bias via analyzing *any* initialization $\mathbf{U}(0), \mathbf{V}(0)$ that gives $\bar{\mathbf{M}}_{\mathbf{U}(0),\mathbf{V}(0)} = 2\mu^2 I$. A convenient choice is $\mathbf{U}(0) = [\sqrt{2}\mu I, 0]$ and $\mathbf{V}(0) = [0, \sqrt{2}\mu I]$ so that $\delta(0) = 0$ and $\Delta(0) = 2\mu^2 \mathbf{1}$. Let $\mathcal{X} \in \mathbb{R}^{N\times d}$ denote the matrix whose nth row is $\mathrm{diag}(\mathbf{X}_n)$, and let $r(t)$ be the vector of residuals with $r_n(t) = \langle \bar{\mathbf{M}}_{\mathbf{U}(t),\mathbf{V}(t)}, \bar{\mathbf{X}}_n \rangle - y_n$. A simple calculation then shows that the dynamics are given by $\dot{\delta}(t) = -4\mathcal{X}^\top r(t) \circ \Delta(t)$ and $\dot{\Delta}(t) = -4\mathcal{X}^\top r(t) \circ \delta(t)$ which have as a solution

$$\delta(t) = 2\mu^2 \sinh\Big(-4\mathcal{X}^\top \int_0^t r(s)ds\Big) \qquad \text{and} \qquad \Delta(t) = 2\mu^2 \cosh\Big(-4\mathcal{X}^\top \int_0^t r(s)ds\Big) \tag{67}$$

This solution for $\delta(t)$ is identical to the one derived in the proof of Theorem 1, so if indeed δ reaches a zero-error solution, then using the same argument as for Theorem 1 we conclude that $\mathrm{diag}(\mathbf{M}^\infty_{\mathbf{U},\mathbf{V}}) = \lim_{t\to\infty} \delta(t) = \arg\min_\delta Q_\mu(\delta)$ s.t. $\mathcal{X}\delta = \mathbf{y}$. ■

Appendix H. Kernel Regime in Matrix Factorization

Here, we provide additional kernel regime results in the context of matrix factorization model in Section 6. Recall the notation for $f((\mathbf{U},\mathbf{V}),\mathbf{X})$, $\mathbf{M}_{\mathbf{U},\mathbf{V}}$ and their "lifted" space representations $\bar{f}((\mathbf{U},\mathbf{V}),\mathbf{X})$, $\bar{\mathbf{M}}_{\mathbf{U},\mathbf{V}}$, respectively, from Section 6. Let $\mathbf{W} = \left[\begin{smallmatrix}\mathbf{U}\\ \mathbf{V}\end{smallmatrix}\right]$ be the concatenation of $\mathbf{U}$ and $\mathbf{V}$, let $\mathcal{X} \in \mathbb{R}^{N\times d^2}$ be the matrix whose nth row is $\mathrm{vec}(\mathbf{X}_n)$, let $y^* \in \mathbb{R}^N$ be the vector of targets $y_1, \ldots, y_N$, and let $y(t) = \mathcal{X}\mathrm{vec}\big(\mathbf{M}_{\mathbf{U}(t),\mathbf{V}(t)}\big)$ be the vector of predictions at time t, where $\mathbf{U}(t), \mathbf{V}(t)$ follow the gradient flow dynamics.

Consider the tangent kernel model for the factorized problem in the "lifted" space $\bar{f}((\mathbf{U},\mathbf{V}),\mathbf{X}) = \bar{f}(\mathbf{W},\mathbf{X})$

$$f_{\mathrm{TK}}(\mathbf{W}_{\mathrm{TK}},\mathbf{X}) = \bar{f}(\mathbf{W}(0),\mathbf{X}) + \big\langle \nabla \bar{f}(\mathbf{W}(0),\mathbf{X}), \mathbf{W}_{\mathrm{TK}} - \mathbf{W}(0)\big\rangle \tag{68}$$

Let $y_{\mathrm{TK}} = [f_{\mathrm{TK}}(\mathbf{W}_{\mathrm{TK}},\mathbf{X}_n)]_{n=1}^N \in \mathbb{R}^N$ denote the tangent kernel model's vector of predictions and let $\mathbf{W}_{\mathrm{TK}}(t) = \left[\begin{smallmatrix}\mathbf{U}_{\mathrm{TK}}(t)\\ \mathbf{V}_{\mathrm{TK}}(t)\end{smallmatrix}\right]$ denote gradient flow path wrt the linearized model in (68). The following theorem establishes the conditions under which $\mathbf{W}(t) \approx \mathbf{W}_{\mathrm{TK}}(t)$.

Theorem 9 *Let $k \geq d$ and let $\lambda I \preceq \mathcal{X}\mathcal{X}^\top \preceq \Lambda I$. Fix $\gamma \geq 0$ and $\mu > \frac{4\Lambda\gamma}{\lambda}$, and suppose that $\big\|\mathbf{W}(0)\mathbf{W}(0)^\top - \mu I\big\|_{op} \leq \gamma$ and $\|y(0) - y^*\| \leq \frac{\mu\lambda}{\sqrt{\Lambda}}\Big(1 - \sqrt{(1+\frac{\gamma}{\mu})/(1+\frac{\lambda}{4\Lambda})}\Big)$. Then*

$$\sup_{T\in\mathbb{R}_+} \|\mathbf{W}(T) - \mathbf{W}(0)\|_F \leq \frac{\sqrt{\Lambda + \frac{\lambda}{4}}\|y(0) - y^*\|}{\lambda\sqrt{\mu}} \quad \textit{and}$$

$$\sup_{T\in\mathbb{R}_+} \|\mathbf{W}(T) - \mathbf{W}_{TK}(T)\|_F \leq \frac{\Lambda\sqrt{1+\frac{\lambda}{4\Lambda}}\|y(0) - y^*\|^2}{\lambda^2\mu^{3/2}} + \frac{2\sqrt{\Lambda}\sqrt{1+\frac{\gamma}{\mu}}\|y(0) - y^*\|}{\lambda\sqrt{\mu}}$$

The proof of Theorem 9 follows a similar approach as the proof of (Chizat et al., 2019, Theorem 2.4), except we do not make the assumption that $\bar{F}(\mathbf{W}(0)) = 0$ (see Section H.1).

Additionally, using Theorem 9, we can show the following corollary on the kernel regime for matrix factorization based on the scale of initialization α and the width of the factorization k (proof in Section H.2).

Corollary 10 *Let $\lambda I \preceq \mathcal{X}\mathcal{X}^\top \preceq \Lambda I$, and $\|y^*\| \leq Y$. If $\mathbf{U}(0), \mathbf{V}(0)$ have i.i.d. $\mathcal{N}\big(0, \alpha^2\big)$ entries for $\alpha^2 \geq \Omega(k^{-1})$, then with probability at least $1 - 2\exp(-d)$ over the randomness in the initialization.*

$$\sup_{T\in\mathbb{R}_+} \left\| \begin{bmatrix} \mathbf{U}(T) \\ \mathbf{V}(T) \end{bmatrix} - \begin{bmatrix} \mathbf{U}(0) \\ \mathbf{V}(0) \end{bmatrix} \right\|_F \leq O\left(\frac{1}{\alpha\sqrt{k}} + \alpha \right) \tag{69}$$

$$\sup_{T\in\mathbb{R}_+} \left\| \begin{bmatrix} \mathbf{U}(T) \\ \mathbf{V}(T) \end{bmatrix} - \begin{bmatrix} \mathbf{U}_{TK}(T) \\ \mathbf{V}_{TK}(T) \end{bmatrix} \right\|_F \leq O\left(\frac{1}{\alpha^3 k^{3/2}} + \frac{1}{\alpha\sqrt{k}} + \alpha \right) \tag{70}$$

From Corollary 10, we can infer that the gradient flow over matrix factorization model remains in the kernel regime whenever the scale of the initialization of the prediction matrix $\mathbf{M}_{\mathbf{U}(0),\mathbf{V}(0)}$ given by $\sigma = \alpha^2\sqrt{k}$ satisfies $\sigma = \omega(1/\sqrt{k})$. In particular, unlike width 1 diagonal network model in Section 4 (where the kernel regime is reached only as scale of initialization $\alpha \to \infty$), with a width k model, we see that kernel regime can happen even when $\sigma \to 0$ as long as σ to zero slower than $1/\sqrt{k}$ (or α goes to zero slower than $1/k$).

H.1. Proof of Theorem 9

In order to prove Theorem 9, we require the following lemmas. We use $y_{\mathrm{TK}}(t) \in \mathbb{R}^N$ denote the tangent kernel model's vector of predictions at corresponding to $\mathbf{W}_{\mathrm{TK}}(t)$.

Lemma 11 *Suppose that the weights are initialized such that $\left\|\mathbf{W}(0)\mathbf{W}(0)^\top - \mu I\right\|_{op} \leq \gamma$ and the measurements satisfy $0 \prec \lambda I \preceq \mathcal{X}\mathcal{X}^\top \preceq \Lambda I$. If $\sup_{0\leq t\leq T}\|\mathbf{W}(t) - \mathbf{W}(0)\|_F \leq R$, then for all $t \leq T$*

$$\begin{aligned} \|y(t) - y^*\| &\leq \|y(0) - y^*\| \exp\big(-2\mu\lambda t + 4\Lambda\big(\gamma + R^2 + 2R\sqrt{\mu+\gamma}\big)t\big), \\ \|y_{TK}(t) - y^*\| &\leq \|y_{TK}(0) - y^*\| \exp(-2\mu\lambda t + 4\Lambda\gamma t). \end{aligned}$$

Proof First, consider the dynamics of $y(t)$:

$$\begin{aligned} y'(t) &= \frac{d}{dt}\Big[\Big\langle \mathbf{W}(t)\mathbf{W}(t)^\top, \mathbf{X}_n \Big\rangle\Big]_{n=1}^N \\ &= \Big[\Big\langle 2\dot{\mathbf{W}}(t)\mathbf{W}(t)^\top, \mathbf{X}_n \Big\rangle\Big]_{n=1}^N \\ &= -4\left[\sum_{m=1}^N \Big(\Big\langle \mathbf{W}(t)\mathbf{W}(t)^\top, \mathbf{X}_m \Big\rangle - y_m\Big) \Big\langle \mathbf{W}(t)\mathbf{W}(t)^\top, \mathbf{X}_n\mathbf{X}_m \Big\rangle\right]_{n=1}^N \\ &= -\Sigma(t)(y(t) - y^*) \end{aligned} \tag{71}$$

where the symmetric matrix $\Sigma(t) \in \mathbb{R}^{N\times N}$ has entries

$$\Sigma(t)_{m,n} \coloneqq 4\Big[\Big\langle \mathbf{W}(t)\mathbf{W}(t)^\top, \mathbf{X}_n\mathbf{X}_m \Big\rangle\Big]. \tag{72}$$

This matrix can also be written:

$$\Sigma(t) = 4\mathcal{X}(I_{d\times d} \otimes \mathbf{W}(t)\mathbf{W}(t)^\top)\mathcal{X}^\top \tag{73}$$

where $\otimes$ denotes the Kronecker product. Therefore, for $t \leq T$

$$
\begin{aligned}
&\left\|\Sigma(t) - 4\mu\mathcal{X}\mathcal{X}^\top\right\|_{\text{op}} \\
&= 4\left\|\mathcal{X}\left(I_{d\times d}\otimes \mathbf{W}(t)\mathbf{W}(t)^\top - \mu I_{d\times d}\otimes I_{d\times d}\right)\mathcal{X}^\top\right\|_{\text{op}} \\
&\le 4\left\|I_{d\times d}\otimes \mathbf{W}(t)\mathbf{W}(t)^\top - \mu I_{d\times d}\otimes I_{d\times d}\right\|_{\text{op}}\|\mathcal{X}\|_{\text{op}}^2 \\
&\le 4\Lambda\left\|\mathbf{W}(t)\mathbf{W}(t)^\top - \mu I_{d\times d}\right\|_{\text{op}} \\
&\le 4\Lambda\left(\left\|\mathbf{W}(0)\mathbf{W}(0)^\top - \mu I_{d\times d}\right\|_{\text{op}} + \left\|\mathbf{W}(t)\mathbf{W}(t)^\top - \mathbf{W}(0)\mathbf{W}(0)^\top\right\|_{\text{op}}\right) \\
&\le 4\Lambda\left(\gamma + \left\|(\mathbf{W}(t)-\mathbf{W}(0))(\mathbf{W}(t)-\mathbf{W}(0))^\top\right\|_{\text{op}} + 2\left\|(\mathbf{W}(t)-\mathbf{W}(0))\mathbf{W}(0)^\top\right\|_{\text{op}}\right) \\
&\le 4\Lambda\big(\gamma + R^2 + 2R\|\mathbf{W}(0)\|_{\text{op}}\big) \\
&\le 4\Lambda\big(\gamma + R^2 + 2R\sqrt{\mu+\gamma}\big)
\end{aligned}
\tag{74}
$$

Therefore, for all $t \le T$, $y'(t) = -\Sigma(t)(y(t)-y^*)$ for

$$
\Sigma(t) \succeq 2\mu\lambda - 4\Lambda\big(\gamma + R^2 + 2R\sqrt{\mu+\gamma}\big). \tag{75}
$$

If $\mu\lambda > 2\Lambda\big(\gamma + R^2 + 2R\sqrt{\mu+\gamma}\big)$, then applying (Chizat et al., 2019, Lemma B.1) completes the first half of the proof. Otherwise, noting that $\|y(t)-y^*\|^2$ is non-increasing in t implies $\|y(t)-y^*\| \le \|y(0)-y^*\|$.

Similarly, the dynamics of y_{TK} are

$$
\begin{aligned}
y'_{\text{TK}}(t) &= \frac{d}{dt}\left[\left\langle \mathbf{W}(0)\mathbf{W}(0)^\top, \mathbf{X}_n\right\rangle + 2\left\langle \mathbf{W}_{\text{TK}}(t) - \mathbf{W}(0), \mathbf{X}_n\mathbf{W}(0)\right\rangle\right]_{n=1}^N \\
&= \left[2\left\langle \dot{\mathbf{W}}_{\text{TK}}(t), \mathbf{X}_n\mathbf{W}(0)\right\rangle\right]_{n=1}^N \\
&= -4\left[\sum_{m=1}^N (y_{\text{TK}}(t)_m - y_m)\left\langle \mathbf{W}(0)\mathbf{W}(0)^\top, \mathbf{X}_n\mathbf{X}_m\right\rangle\right]_{n=1}^N \\
&= -\Sigma(0)(y_{\text{TK}}(t) - y^*)
\end{aligned}
\tag{76}
$$

From here, we can follow the same argument to show that

$$
\Sigma(0) \succeq 2\mu\lambda - 4\Lambda\gamma. \tag{77}
$$

Applying (Chizat et al., 2019, Lemma B.1) again concludes the proof. ■

Lemma 12 *Suppose that the weights are initialized such that* $\left\|\mathbf{W}(0)\mathbf{W}(0)^\top - \mu I\right\|_{op} \le \gamma$ *and that the measurements satisfy* $\lambda I \preceq \mathcal{X}\mathcal{X}^\top \preceq \Lambda I$. *Suppose in addition that*

$$
\mu > \frac{4\Lambda\gamma}{\lambda} \quad \textit{and} \quad \|y(0)-y^*\| \le \frac{\mu\lambda}{\sqrt{\Lambda}}\left(1 - \sqrt{\frac{1+\frac{\gamma}{\mu}}{1+\frac{\lambda}{4\Lambda}}}\right).
$$

Then,

$$\sup_{t\geq 0}\|\mathbf{W}(t)-\mathbf{W}(0)\|_F \leq \frac{\sqrt{\Lambda+\frac{\lambda}{4}}\|y(0)-y^*\|}{\lambda\sqrt{\mu}}$$

Proof To begin, define

$$R \coloneqq \sqrt{\mu+\frac{\mu\lambda}{4\Lambda}}-\sqrt{\mu+\gamma}. \tag{78}$$

Since $\mu > \frac{4\Lambda\gamma}{\lambda}$, $R>0$. Note that with this choice

$$2\mu\lambda - 4\Lambda\big(\gamma + R^2 + 2R\big(\sqrt{\mu+\gamma}\big)\big) = \mu\lambda \tag{79}$$

Let $T = \inf\{t \,|\, \|\mathbf{W}(t)-\mathbf{W}(0)\|_F > R\}$, and suppose towards contradiction that $T<\infty$. Then

$$\begin{aligned}
R &\leq \|\mathbf{W}(T)-\mathbf{W}(0)\|_F \\
&= \left\|\int_0^T \dot{\mathbf{W}}(t)dt\right\|_F \\
&\leq \int_0^T \left\|\sum_{n=1}^N (y(t)_n - y_n)\mathbf{X}_n\mathbf{W}(t)\right\|_F dt \\
&= \int_0^T \left\|(I_{d\times d}\otimes \mathbf{W}(t))\mathcal{X}^\top(y(t)-y)\right\|_2 dt \\
&\leq \int_0^T \|\mathbf{W}(t)\|_{\text{op}}\|\mathcal{X}\|_{\text{op}}\|y(t)-y\|_2 dt \\
&\leq \sqrt{\Lambda}\int_0^T \Big(\|\mathbf{W}(0)\|_{\text{op}}+R\Big)\|y(t)-y\|dt \\
&\leq \sqrt{\Lambda}\big(\sqrt{\mu+\gamma}+R\big)\int_0^T \|y(t)-y\|dt \\
&= \sqrt{\Lambda}\sqrt{\mu+\frac{\mu\lambda}{4\Lambda}}\int_0^T \|y(t)-y\|dt
\end{aligned} \tag{80}$$

From here, we apply Lemma 11 and (79) to conclude that

$$\begin{aligned}
R &\leq \sqrt{\mu\Lambda+\frac{\mu\lambda}{4}}\|y(0)-y^*\|\int_0^T \exp(-\mu\lambda t)dt \\
&< \sqrt{\mu\Lambda+\frac{\mu\lambda}{4}}\|y(0)-y^*\|\int_0^\infty \exp(-\mu\lambda t)dt \\
&= \frac{\sqrt{\Lambda+\frac{\lambda}{4}}\|y(0)-y^*\|}{\lambda\sqrt{\mu}} \\
&\leq \frac{\sqrt{\Lambda+\frac{\lambda}{4}}}{\lambda\sqrt{\mu}}\frac{\mu\lambda}{\sqrt{\Lambda}}\left(1-\sqrt{\frac{1+\frac{\gamma}{\mu}}{1+\frac{\lambda}{4\Lambda}}}\right) \\
&= \sqrt{\mu+\frac{\mu\lambda}{4\Lambda}}-\sqrt{\mu+\gamma} \\
&= R
\end{aligned} \tag{81}$$

This is a contradiction, so we conclude $T = \infty$. We conclude the proof by pointing out that the same line of reasoning from the righthand side of (80) through to (81) applies even when $T = \infty$. ■

Theorem 9 *Let $k \geq d$ and let $\lambda I \preceq \mathcal{X}\mathcal{X}^\top \preceq \Lambda I$. Fix $\gamma \geq 0$ and $\mu > \frac{4\Lambda\gamma}{\lambda}$, and suppose that $\left\|\mathbf{W}(0)\mathbf{W}(0)^\top - \mu I\right\|_{op} \leq \gamma$ and $\|y(0) - y^*\| \leq \frac{\mu\lambda}{\sqrt{\Lambda}}\left(1 - \sqrt{(1+\frac{\gamma}{\mu})/(1+\frac{\lambda}{4\Lambda})}\right)$. Then*

$$\sup_{T\in\mathbb{R}_+} \|\mathbf{W}(T) - \mathbf{W}(0)\|_F \leq \frac{\sqrt{\Lambda + \frac{\lambda}{4}}\|y(0) - y^*\|}{\lambda\sqrt{\mu}} \quad \textit{and}$$

$$\sup_{T\in\mathbb{R}_+} \|\mathbf{W}(T) - \mathbf{W}_{TK}(T)\|_F \leq \frac{\Lambda\sqrt{1+\frac{\lambda}{4\Lambda}}\|y(0) - y^*\|^2}{\lambda^2\mu^{3/2}} + \frac{2\sqrt{\Lambda}\sqrt{1+\frac{\gamma}{\mu}}\|y(0) - y^*\|}{\lambda\sqrt{\mu}}$$

Proof Our proof follows the approach of Chizat et al. (2019) closely, but it is specialized to our particular setting and formulation. We also do not require that $F(\mathbf{W}(0)) = 0$.

Consider for some T

$$\begin{aligned}
&\|\mathbf{W}(T) - \mathbf{W}_{\mathrm{TK}}(T)\|_F \\
&= \left\|\int_0^T \dot{\mathbf{W}}(t) - \dot{\mathbf{W}}_{\mathrm{TK}}(t)dt\right\|_F \\
&\leq \int_0^T \left\|\sum_{n=1}^N (y(t)_n - y_n)\mathbf{X}_n\mathbf{W}(t) - (y_{\mathrm{TK}}(t)_n - y_n)\mathbf{X}_n\mathbf{W}(0)\right\|_F dt \\
&= \int_0^T \left\|(I_{d\times d}\otimes \mathbf{W}(t))\mathcal{X}^\top(y(t) - y^*) - (I_{d\times d}\otimes \mathbf{W}(0))\mathcal{X}^\top(y_{\mathrm{TK}}(t) - y^*)\right\|_F dt \\
&= \int_0^T \left\|(I_{d\times d}\otimes (\mathbf{W}(t) - \mathbf{W}(0)))\mathcal{X}^\top(y(t) - y^*) - (I_{d\times d}\otimes \mathbf{W}(0))\mathcal{X}^\top(y_{\mathrm{TK}}(t) - y(t))\right\|_F dt \\
&\leq \sqrt{\Lambda}\int_0^T \|\mathbf{W}(t) - \mathbf{W}(0)\|_{\mathrm{op}}\|y(t) - y^*\|_2 + \|\mathbf{W}(0)\|_{\mathrm{op}}\|y_{\mathrm{TK}}(t) - y(t)\|_2 dt \\
&\leq \sqrt{\Lambda}\int_0^\infty \|\mathbf{W}(t) - \mathbf{W}(0)\|_{\mathrm{op}}\|y(t) - y^*\|_2 + \|\mathbf{W}(0)\|_{\mathrm{op}}\|y_{\mathrm{TK}}(t) - y(t)\|_2 dt
\end{aligned} \tag{82}$$

By Lemma 12,

$$\sup_t \|\mathbf{W}(t) - \mathbf{W}(0)\|_{\mathrm{op}} \leq \sup_t \|\mathbf{W}(t) - \mathbf{W}(0)\|_F \leq \frac{\sqrt{\Lambda + \frac{\lambda}{4}}\|y(0) - y^*\|}{\lambda\sqrt{\mu}} \tag{83}$$

By Lemma 11, for $R = \frac{\sqrt{\Lambda+\frac{\lambda}{4}}\|y(0)-y^*\|}{\lambda\sqrt{\mu}}$, we have

$$\begin{aligned}
\|y(t) - y^*\| &\leq \|y(0) - y^*\|\exp\left(-2\mu\lambda t + 4\Lambda\left(\gamma + R^2 + 2R\sqrt{\mu+\gamma}\right)t\right) \\
\|y_{\mathrm{TK}}(t) - y^*\| &\leq \|y(0) - y^*\|\exp(-2\mu\lambda t + 4\Lambda\gamma t)
\end{aligned} \tag{84}$$

Since $\mu > \frac{4\Lambda\gamma}{\lambda}$ and $\|y(0) - y^*\| \leq \frac{\mu\lambda}{\sqrt{\Lambda}}\left(1 - \sqrt{\frac{1+\frac{\gamma}{\mu}}{1+\frac{\lambda}{4\Lambda}}}\right)$, this further implies

$$\begin{aligned} \|y(t) - y^*\| &\leq \|y(0) - y^*\| \exp(-\mu\lambda t) \\ \|y_{\mathrm{TK}}(t) - y^*\| &\leq \|y(0) - y^*\| \exp(-\mu\lambda t) \end{aligned} \tag{85}$$

Finally,

$$\|y_{\mathrm{TK}}(t) - y(t)\| \leq \|y(t) - y^*\| + \|y_{\mathrm{TK}}(t) - y^*\| \leq 2\|y(0) - y^*\| \exp(-\mu\lambda t) \tag{86}$$

Combining the above inequalities, we have

$$\begin{aligned} &\left\|\mathbf{W}(T) - \bar{\mathbf{W}}(T)\right\|_F \\ &\leq \sqrt{\Lambda}\int_0^\infty \|\mathbf{W}(t) - \mathbf{W}(0)\|_{\mathrm{op}}\|y(t) - y^*\|_2 + \|\mathbf{W}(0)\|_{\mathrm{op}}\|\bar{y}(t) - y(t)\|_2 dt \\ &\leq \sqrt{\Lambda}\int_0^T \left(\frac{\sqrt{\Lambda + \frac{\lambda}{4}}\|y(0) - y^*\|^2}{\lambda\sqrt{\mu}} + 2\sqrt{\mu+\gamma}\|y(0) - y^*\|\right)\exp(-\mu\lambda t)dt \\ &\leq \frac{\sqrt{\Lambda}\left(\frac{\sqrt{\Lambda+\frac{\lambda}{4}}\|y(0)-y^*\|^2}{\lambda\sqrt{\mu}} + 2\sqrt{\mu+\gamma}\|y(0) - y^*\|\right)}{\mu\lambda} \\ &= \frac{\Lambda\sqrt{1+\frac{\lambda}{4\Lambda}}\|y(0) - y^*\|^2}{\lambda^2\mu^{3/2}} + \frac{2\sqrt{\Lambda}\sqrt{1+\frac{\gamma}{\mu}}\|y(0) - y^*\|}{\lambda\sqrt{\mu}} \end{aligned} \tag{87}$$

■

H.2. Proof of Corollary 10

Finally, we prove Corollary 10 using the following:

Lemma 13 (cf. Theorem 6.1 Wainwright (2019)) *Let $W \in \mathbb{R}^{d\times k}$ with $d \leq k$ and with $W_{i,j} \sim \mathcal{N}(0, \sigma^2)$, then*

$$\mathbb{P}\left[\|WW^\top - \sigma^2 kI\|_{op} \geq 8\sigma^2\sqrt{kd}\right] \leq 2\exp\left(-\frac{d}{2}\right)$$

Corollary 10 *Let $\lambda I \preceq \mathcal{X}\mathcal{X}^\top \preceq \Lambda I$, and $\|y^*\| \leq Y$. If $\mathbf{U}(0), \mathbf{V}(0)$ have i.i.d. $\mathcal{N}(0, \alpha^2)$ entries for $\alpha^2 \geq \Omega(k^{-1})$, then with probability at least $1 - 2\exp(-d)$ over the randomness in the initialization.*

$$\sup_{T\in\mathbb{R}_+}\left\|\begin{bmatrix}\mathbf{U}(T)\\ \mathbf{V}(T)\end{bmatrix} - \begin{bmatrix}\mathbf{U}(0)\\ \mathbf{V}(0)\end{bmatrix}\right\|_F \leq O\left(\frac{1}{\alpha\sqrt{k}} + \alpha\right) \tag{69}$$

$$\sup_{T\in\mathbb{R}_+}\left\|\begin{bmatrix}\mathbf{U}(T)\\ \mathbf{V}(T)\end{bmatrix} - \begin{bmatrix}\mathbf{U}_{TK}(T)\\ \mathbf{V}_{TK}(T)\end{bmatrix}\right\|_F \leq O\left(\frac{1}{\alpha^3 k^{3/2}} + \frac{1}{\alpha\sqrt{k}} + \alpha\right) \tag{70}$$

Proof All that is needed is to show the relationship between k and the quantities involved in the statement of Theorem 9. Let $\mathbf{W} \coloneqq \begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix} \in \mathbb{R}^{2d\times k}$. By Lemma 13,

$$\mathbb{P}\left[\left\|\mathbf{W}\mathbf{W}^\top - \alpha^2 k I\right\|_{\text{op}} < 8\alpha^2\sqrt{2kd}\right] \geq 1 - 2\exp(-d) \tag{88}$$

For the remainder of the proof, we condition on the event $\left\|\mathbf{W}\mathbf{W}^\top - \alpha^2 k I\right\|_{\text{op}} < 8\alpha^2\sqrt{2kd}$. Next, we bound $\|y(0) - y^*\|^2$:

$$\begin{aligned}
\|y(0) - y^*\|^2 &\leq 2Y^2 + 2\|y(0)\|^2 \\
&= 2Y^2 + 2\sum_{n=1}^{N}\left\langle \mathbf{W}(0)\mathbf{W}(0)^\top, \bar{\mathbf{X}}_n\right\rangle^2 \\
&\overset{(a)}{=} 2Y^2 + 2\sum_{n=1}^{N}\left\langle \mathbf{W}(0)\mathbf{W}(0)^\top - \alpha^2 k I, \bar{\mathbf{X}}_n\right\rangle^2 \\
&\leq 2Y^2 + 2\sum_{n=1}^{N}\left\|\mathbf{W}(0)\mathbf{W}(0)^\top - \alpha^2 k I\right\|_F^2 \|\bar{\mathbf{X}}_n\|_F^2 \\
&= 2Y^2 + 4d\left\|\mathbf{W}(0)\mathbf{W}(0)^\top - \alpha^2 k I\right\|_{\text{op}}^2 \sum_{n=1}^{N}\frac{1}{2}\|\mathbf{X}_n\|_F^2 \\
&\leq 2Y^2 + 2d\left(8\alpha^2\sqrt{2kd}\right)^2\|\mathcal{X}\|_F^2 \\
&\leq 2Y^2 + 256kd^3\alpha^4\Lambda^2,
\end{aligned} \tag{89}$$

where for (a), we used that $\bar{\mathbf{X}}_n$ is zero on the diagonal. In order to apply Theorem 9 using

$$\gamma = 8\alpha^2\sqrt{2kd} \qquad \text{and} \qquad \mu = \alpha^2 k, \tag{90}$$

we require that

$$\alpha^2 k = \mu > \frac{4\Lambda\gamma}{\lambda} = \frac{32\alpha^2\Lambda\sqrt{2kd}}{\lambda} \iff k > \frac{2048\Lambda^2 d}{\lambda^2} \tag{91}$$

and

$$\|y(0) - y^*\| \leq \frac{\mu\lambda}{\sqrt{\Lambda}}\left(1 - \sqrt{\frac{1+\frac{\gamma}{\mu}}{1+\frac{\lambda}{4\Lambda}}}\right) = \frac{\alpha^2\lambda k}{\sqrt{\Lambda}}\left(1 - \sqrt{\frac{1+\frac{8\sqrt{2kd}}{k}}{1+\frac{\lambda}{4\Lambda}}}\right) \tag{92}$$

By (89), this is implied by

$$\sqrt{2Y^2 + 256kd^3\alpha^4\Lambda^2} \leq \frac{\alpha^2\lambda k}{\sqrt{\Lambda}}\left(1 - \sqrt{\frac{1+\frac{8\sqrt{2d}}{\sqrt{k}}}{1+\frac{\lambda}{4\Lambda}}}\right) \tag{93}$$

$$\impliedby k \geq \max\left\{\frac{8192\Lambda^2 d}{\lambda^2}, \frac{512d^3\Lambda^3\left(4+\frac{16\Lambda}{\lambda}\right)^2}{\lambda^2}, \frac{Y}{2\alpha^2\sqrt{\Lambda}(\lambda+4\Lambda)}\right\} \tag{94}$$

This is because $k \geq \frac{8192\Lambda^2 d}{\lambda^2}$ ensures

$$\sqrt{\frac{1+\frac{8\sqrt{2d}}{\sqrt{k}}}{1+\frac{\lambda}{4\Lambda}}} \leq \sqrt{\frac{1+\frac{\lambda}{8\Lambda}}{1+\frac{\lambda}{4\Lambda}}} = \sqrt{1-\frac{1}{2+\frac{8\Lambda}{\lambda}}} \leq 1 - \frac{1}{4+\frac{16\Lambda}{\lambda}} \tag{95}$$

Consider two cases: either $2Y^2 \leq 256kd^3\alpha^4\Lambda^2$ or it is not. In the first case,

$$\begin{aligned}
\frac{\alpha^2\lambda k}{\sqrt{\Lambda}}\left(1-\sqrt{\frac{1+\frac{8\sqrt{2d}}{\sqrt{k}}}{1+\frac{\lambda}{4\Lambda}}}\right) &\geq \frac{\alpha^2\lambda k}{\sqrt{\Lambda}\left(4+\frac{16\Lambda}{\lambda}\right)} \\
&\geq \frac{\alpha^2\lambda\sqrt{k}}{\sqrt{\Lambda}\left(4+\frac{16\Lambda}{\lambda}\right)} \cdot \frac{\sqrt{512d^3\Lambda^3}\left(4+\frac{16\Lambda}{\lambda}\right)}{\lambda} \\
&= \sqrt{512kd^3\alpha^4\Lambda^2} \\
&\geq \sqrt{2Y^2+256kd^3\alpha^4\Lambda^2}
\end{aligned} \tag{96}$$

For the first inequality, we used (95), for the second inequality we used $k \geq \frac{512d^3\Lambda^3\left(4+\frac{16\Lambda}{\lambda}\right)^2}{\lambda^2}$. Otherwise, $2Y^2 > 256kd^3\alpha^4\Lambda^2$ and

$$\begin{aligned}
\frac{\alpha^2\lambda k}{\sqrt{\Lambda}}\left(1-\sqrt{\frac{1+\frac{8\sqrt{2d}}{\sqrt{k}}}{1+\frac{\lambda}{4\Lambda}}}\right) &\geq \frac{\alpha^2\lambda k}{\sqrt{\Lambda}\left(4+\frac{16\Lambda}{\lambda}\right)} \\
&\geq 2Y \\
&> \sqrt{2Y^2+256kd^3\alpha^4\Lambda^2}
\end{aligned} \tag{97}$$

For the second inequality, we used that $k \geq \frac{Y}{2\alpha^2\sqrt{\Lambda}(\lambda+4\Lambda)}$.

Therefore, for k sufficiently large (94), by Theorem 9

$$\begin{aligned}
\sup_{T\in\mathbb{R}_+} \|\mathbf{W}(T)-\mathbf{W}(0)\|_F &\leq \frac{\sqrt{\Lambda+\frac{\lambda}{4}}\|y(0)-y^*\|}{\lambda\sqrt{\mu}} \\
&\leq \frac{\sqrt{\Lambda+\frac{\lambda}{4}}\sqrt{2Y^2+256kd^3\alpha^4\Lambda^2}}{\lambda\sqrt{\alpha^2 k}} \\
&\leq \frac{2\sqrt{\Lambda}\left(2Y+16\sqrt{kd^3\alpha^4\Lambda^2}\right)}{\lambda\alpha\sqrt{k}} \\
&\leq \frac{4Y\sqrt{\Lambda}}{\lambda\alpha\sqrt{k}} + \frac{32d^{3/2}\Lambda^{3/2}\alpha}{\lambda}
\end{aligned} \tag{98}$$

and

$$\begin{aligned}
&\sup_{T\in\mathbb{R}_+} \left\|\mathbf{W}(T) - \bar{\mathbf{W}}(T)\right\|_F \\
&\leq \frac{\Lambda\sqrt{1+\frac{\lambda}{2\Lambda}}\|y(0)-y^*\|^2}{\lambda^2\mu^{3/2}} + \frac{2\sqrt{\Lambda}\sqrt{1+\frac{\gamma}{\mu}}\|y(0)-y^*\|}{\lambda\sqrt{\mu}} \\
&\leq \frac{2\Lambda\left(2Y^2+512kd^3\alpha^4\Lambda^2\right)}{\lambda^2(\alpha^2k)^{3/2}} + \frac{2\sqrt{\Lambda}\sqrt{1+\frac{8\sqrt{2d}}{\sqrt{k}}}\left(2Y+\sqrt{512kd^3\alpha^4\Lambda^2}\right)}{\lambda\sqrt{\alpha^2k}} \\
&\leq \frac{4\Lambda Y^2}{\lambda^2\alpha^3k^{3/2}} + \frac{1024d^3\Lambda^3\alpha}{\lambda^2\sqrt{k}} + \frac{8Y\sqrt{\Lambda}}{\lambda\alpha\sqrt{k}} + \frac{64d^{3/2}\Lambda^{3/2}\alpha}{\lambda}
\end{aligned} \tag{99}$$

It is clear from (98) and (99) that there is some scalar c which depends only on Λ, λ, d, and Y such that

$$\begin{aligned}
&\sup_{T\in\mathbb{R}_+} \|\mathbf{W}(T) - \mathbf{W}(0)\|_F \leq c\left(\frac{1}{\alpha\sqrt{k}} + \alpha\right), \text{ and} \\
&\sup_{T\in\mathbb{R}_+} \left\|\mathbf{W}(T) - \bar{\mathbf{W}}(T)\right\|_F \leq c\left(\frac{1}{\alpha^3k^{3/2}} + \frac{1}{\alpha\sqrt{k}} + \alpha\right)
\end{aligned} \tag{100}$$

■