arXiv:1906.05827v1 [cs.LG] 13 Jun 2019

Kernel and Deep Regimes in Overparametrized Models

Blake Woodworth
Toyota Technological Institute at Chicago

Suriya Gunasekar
Toyota Technological Institute at Chicago

Jason Lee
University of Southern California

Daniel Soudry
Technion

Nathan Srebro
Toyota Technological Institute at Chicago

Abstract

A recent line of work studies overparametrized neural networks in the "kernel regime," i.e. when the network behaves during training as a kernelized linear predictor, and thus training with gradient descent has the effect of finding the minimum RKHS norm solution. This stands in contrast to other studies which demonstrate how gradient descent on overparametrized multilayer networks can induce rich implicit biases that are not RKHS norms. Building on an observation by Chizat and Bach [4], we show how the ***scale of the initialization*** controls the transition between the "kernel" (aka lazy) and "deep" (aka active) regimes and affects generalization properties in multilayer homogeneous models. We provide a complete and detailed analysis for a simple two-layer model that already exhibits an interesting and meaningful transition between the kernel and deep regimes, and we demonstrate the transition for more complex matrix factorization models.

1 Introduction

A string of recent papers study neural networks trained with gradient descent in the "kernel regime." The main observation is that, in a certain regime, networks trained with gradient descent behave as kernel methods, and so can be studied as such [14, 6, 5]. This allow one to prove convergence to zero error solutions in overparametrized settings [7, 8, 1], but it also implies gradient descent will converge to the minimum norm solution (in the corresponding RKHS) [4, 2, 19] and more generally that models will inherit the inductive bias and generalization behaviour of the RKHS. This would suggest that deep models can be effectively 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 some 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 [13]; weight decay on an infinite width single input ReLU implicitly biases towards minimizing the second order total variations $\int |f''(x)|\, dx$ of the learned function [21]; 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 [11] and can ensure low rank matrix recovery [17]. All these natural inductive biases (L_p bridge penalty for $p < 1$, total variation norm, nuclear norm) are not Hilbert norms, and therefore *cannot* be captured by a kernel. This suggests that training deep models with gradient descent can behave very differently from kernel methods, and have much richer inductive biases.

One might ask whether the kernel approximation indeed captures the behavior of deep learning in a relevant and interesting regime, or does the success of deep learning come when learning escapes this regime? In order to understand this, we must first carefully understand when each of these regimes hold, and how the transition between the "kernel" regime and the "deep" regime happens.

Some investigations of the kernel regime emphasized the number of parameters ("width") going to infinity as leading to this regime. However Chizat and Bach [4] identified the scale of the model as a quantity controlling entry into the kernel regime. Their results suggest that for any number of parameters (any width), a model can be approximated by kernelized linear model when its scale at initialization goes to infinity (see details in Section 3). Considering models with increasing (or infinite) width, the relevant regime (kernel or deep) 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 effects the model behaviour for D-homogeneous models.

In Section 4 we provide a complete and detailed study for a simple 2-homogeneous model that can be viewed as linear regression with squared parametrization, or as a "diagonal" linear neural network. For this model we can exactly characterize the implicit bias of training with gradient descent, as a function of the scale α of initialization, and see how this implicit bias becomes the L2 norm in the $\alpha \to \infty$ kernel regime, but the L1 norm in the $\alpha \to 0$ deep regime. We can therefore understand how, e.g. for a high dimensional problem with underlying sparse structure, we can get good generalization when the initialization scale α is small, but not when α is large. In Section 5 we demonstrate a similar transition in matrix factorization.

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 $h_{\mathbf{w}} = F(\mathbf{w})$ such that $h_{\mathbf{w}}(\mathbf{x}) = f(\mathbf{w}, \mathbf{x})$. Much of our focus will on models, such as linear networks, which are linear in $\mathbf{x}$ (but not on the parameters $\mathbf{w}$!), in which case $F(\mathbf{w}) \in \mathcal{X}^*$ is a linear predictor 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 this change of parametrization is crucial.

We consider models that are D-positive homogeneous in the parameters $\mathbf{w}$, for some integer $D \geq 1$, meaning that for any $c \in \mathbb{R}_+$, $F(c \cdot \mathbf{w}) = c^D F(\mathbf{w})$ and so $f(c \cdot \mathbf{w}, \mathbf{x}) = c^D f(\mathbf{w}, \mathbf{x})$. We will refer to such models simply as D-homogeneous. Such homogeneity is satisfied by many interesting model classes including multi-layer ReLU networks with fully connected and convolutional layers, layered linear neural networks, and matrix factorization where D corresponds to the depth of the network.

Consider a training set $\left\{(\mathbf{x}^{(n)}, y^{(n)})\right\}_{n=1}^N$ consisting of N examples of input label pairs. For a given loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the loss of the model parametrized by $\mathbf{w}$ is $L(\mathbf{w}) = L(F(\mathbf{w})) = \sum_{n=1}^N \ell(f(\mathbf{w}, \mathbf{x}^{(n)}), y^{(n)})$. We will focus on the squared loss $\ell_{\text{sq}}(\hat{y}, y) = (\hat{y} - y)^2$. We slightly abuse notation and use $f(\mathbf{w}, X) \in \mathbb{R}^N$ to denote the vector of predictions $[f(\mathbf{w}, \mathbf{x}^{(1)}), \ldots, f(\mathbf{w}, \mathbf{x}^{(N)})]$ and so we can write $L(\mathbf{w}) = \|f(\mathbf{w}, X) - \mathbf{y}\|_2^2$, where $\mathbf{y} \in \mathbb{R}^N$ is the vector of target labels.

Minimizing the loss $L(\mathbf{w})$ using gradient descent amounts to iteratively updating the parameters

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta \nabla L(\mathbf{w}(k)) \tag{1}$$

We will consider gradient descent with infinitesimally small stepsize η, which is captured by the gradient flow dynamics

$$\dot{\mathbf{w}}(t) = -\nabla L(\mathbf{w}(t)). \tag{2}$$

We will be particularly interested in the scale of initialization and will capture this through a scalar parameter $\alpha \in \mathbb{R}_+$. For each scale α, we will denote by $\mathbf{w}_\alpha(t)$ the dynamics obtained by the gradient flow dynamics (2) with the initial condition $\mathbf{w}_\alpha(0) = \alpha \mathbf{w}_0$ for some fixed $\mathbf{w}_0$. We will also denote $h_\alpha(t) = F(\mathbf{w}_\alpha(t))$, or in the case of linear predictors $\boldsymbol{\beta}_\alpha(t) = \boldsymbol{\beta}_{\mathbf{w}_\alpha(t)}$, the dynamics on the predictor $F(\mathbf{w})$ induced by the gradient flow dynamics on $\mathbf{w}$.

In many cases we expect the gradient flow dynamics to converge to a minimizer of $L(\mathbf{w})$, though establishing that this happens will not be our focus. Rather, we are interested in the underdetermined case, where $N \ll p$, and in general there are multiple minimizers of $L(\mathbf{w})$, all with $f(\mathbf{w}, X) = \mathbf{y}$

and so $L(\mathbf{w}) = 0$. The question we will mostly be concerned with is which of these many minimizers does gradient flow converge to. That is, we would like to characterize $\mathbf{w}_\alpha(\infty) = \lim_{t\to\infty} \mathbf{w}_\alpha(t)$, or more importantly the predictor $h_\alpha(\infty) = F(\mathbf{w}_\alpha(\infty))$ or $\boldsymbol{\beta}_\alpha(\infty) = \boldsymbol{\beta}_{\mathbf{w}_\alpha(\infty)}$ we converge to, and how these depend on the scale α. In underdetermined problems, where there are many zero error solutions, simply fitting the data using the model does not provide enough inductive bias to ensure generalization. But in many cases, the specific solution reached by gradient flow (or some other optimization procedure) has special structure, or minimizes some implicit regularizer, and this structure or regularizer provides the needed inductive bias [13, 12, 22, 15].

3 The Kernel Regime

Gradient descent and gradient flow only consider the first order approximation of a model w.r.t. $\mathbf{w}$ about the current iterate:

$$f(\mathbf{w}, \mathbf{x}) = f(\mathbf{w}(t), x) + \langle \mathbf{w} - \mathbf{w}(t),\, \nabla_{\mathbf{w}} f(\mathbf{w}(t), \mathbf{x})\rangle + O(\|\mathbf{w} - \mathbf{w}(t)\|^2). \tag{3}$$

That is, locally around any $\mathbf{w}(t)$, 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* $K_{\mathbf{w}(t)}(x, x') = \langle \nabla_{\mathbf{w}} f(\mathbf{w}(t), \mathbf{x}),\, \nabla_{\mathbf{w}} f(\mathbf{w}(t), \mathbf{x}')\rangle$ [14, 23, 16]. Of particular interest is the tangent kernel at initialization, $K_{\mathbf{w}_\alpha(0)} = \alpha^{2(D-1)} K_0$ where we denote $K_0 = K_{\mathbf{w}_0}$.

The "kernel regime" refers to a limit in which the tangent kernel $K_{\mathbf{w}(t)}$ does not change over the course of optimization, and less formally to the regime in which it does not change significantly, i.e. where $\forall_t K_{\mathbf{w}(t)} \approx K_{\mathbf{w}(0)}$. In this case, training the model is completely equivalent to training the affine model $\tilde{f}(\mathbf{w}, \mathbf{x}) = \alpha^D f(\mathbf{w}_0, \mathbf{x}) + \left\langle \mathbf{w},\, \alpha^{(D-1)} \phi_0(\mathbf{x})\right\rangle$, or in other words to kernelized gradient descent (or gradient flow) with the kernel K_0 and a "bias term" of $\alpha^D f(\mathbf{w}_0, \mathbf{x})$. In order to not have to worry about this bias term, and in particular its scaling, Chizat and Bach [4] suggest considering "unbiased" initializations such that $F(\mathbf{w}_0) = 0$, and so this bias term vanishes. This can be achieved in many cases by replicating units or components with opposite signs at initialization, and is the approach we take here (see Sections 4 and 5 for examples and details).

For underdetermined problem with multiple solutions $f(\mathbf{w}, X) = \mathbf{y}$, unbiased[1] kernel gradient flow (or gradient descent) converges to the minimum norm solution $\hat{h}_K = \arg\min_{h(X)=\mathbf{y}} \|h\|_K$, where $\|h\|_K$ is the RKHS norm corresponding to the kernel. And so, in the kernel regime, we will have that $h(\infty) = \hat{h}_K$, and the implicit bias of training is precisely given by the kernel.

When does the "kernel regime" happen? Chizat and Bach [4] showed that for any homogeneous[2] model satisfying some technical conditions,[3] the kernel regime is reached as $\alpha \to \infty$. That is, as we increase the scale of initialization, the dynamics converge to the kernel gradient flow dynamics with the kernel K_0, and we have $\lim_{\alpha\to\infty} h_\alpha(\infty) = \hat{h}_K$. In Section 4 we prove this limit directly for our specific model, and we also demonstrate it empirically for matrix factorization in Section 5.

In contrast, and as we shall see in Sections 4 and 5, the $\alpha \to 0$ small initialization limit often leads to a very different and rich inductive bias, e.g. inducing sparsity or low-rank structure [11, 17, 13], that allows for generalization in many settings where kernel methods would not. We refer to this limit reached as $\alpha \to 0$ as the "deep regime." This regime is also referred to as the "active" or "adaptive" regime [4] 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 regime is the regime that truly allows us to exploit the power of depth, and thus is the relevant regime for understanding the success of deep learning.

[1] With a bias term, convergence is to $\hat{h}_K = \arg\min_{h(X)=\mathbf{y}} \|h - h(0)\|_K$, where $h(0) = F(\mathbf{w}(0))$ is the predictor at initialization.

[2] Chizat and Bach 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.

[3] A technical problem with the main result of Chizat and Bach [4], their Theorem 3.2, is that for models obtained by the symmetric initialization of duplicating units and negating their signs, the Jacobian of the model is degenerate at initialization, or in their notation $\sigma_{\min} = 0$, invalidating the assumption in the Theorem. On the other hand, without such symmetric initialization, and for finite width model (i.e. when p is finite), the scale of the prediction at initialization explodes as $\alpha \to \infty$ and violates their assumptions. For this reason, we cannot rely on their result, and instead establish the kernel regime specifically for the model we study in Section 4

4 Detailed Study of a Simple Depth-2 Model

We study in detail a simple 2-homogeneous model. Consider the class of linear functions over $\mathcal{X} = \mathbb{R}^d$, with squared parametrization as follows:

$$f(\mathbf{w}, \mathbf{x}) = \sum_{i=1}^{d} (\mathbf{w}_{+,i}^2 - \mathbf{w}_{-,i}^2)\mathbf{x}_i = \langle \boldsymbol{\beta}_{\mathbf{w}}, \mathbf{x} \rangle \,, \text{ where } \mathbf{w} = \begin{bmatrix} \mathbf{w}_+ \\ \mathbf{w}_- \end{bmatrix} \in \mathbb{R}^{2d} \text{ and } \boldsymbol{\beta}_{\mathbf{w}} = \mathbf{w}_+^2 - \mathbf{w}_-^2 \quad (4)$$

where we use the notation $\mathbf{z}^2$ for $\mathbf{z} \in \mathbb{R}^d$ to denote element-wise squaring. We will consider initializing all weights equally, i.e. using scalings of $\mathbf{w}_0 = \mathbf{1}$.

This is nothing but a linear regression model, except with unconventional over-parametrization. The models can also be thought of as a "diagonal" linear neural network (i.e. where the weight matrix has diagonal structure) with $2d$ units. A standard diagonal linear network would have d units, with each unit connected to just a single input unit with weights u_i and the output with weight v_i, thus implementing the model $f((\boldsymbol{u}, \boldsymbol{v}), \mathbf{x}) = \sum_i \boldsymbol{u}_i \boldsymbol{v}_i \mathbf{x}_i$. But if at initialization $|\boldsymbol{u}_i| = |\boldsymbol{v}_i|$, their magnitude will remain equal and their signs will not flip throughout training, and so we can equivalently replace both with a single weight $\mathbf{w}_i$, yielding the model $f(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}^2, \mathbf{x} \rangle$.

The reason for using both $\mathbf{w}_+$ and $\mathbf{w}_-$ (or $2d$ units) 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 $F(\alpha \mathbf{w}_0) = 0$ without this being a saddle point from which gradient flow will never escape.[4]

The model (4) is perhaps the simplest non-trivial D-homogeneous model for $D > 1$, but, as we shall see, it already exhibits distinct and meaningful kernel and deep regimes. Furthermore, we can completely understand the implicit regularization driving this model analytically, and precisely characterize the transition between the kernel and rich regimes.

Let us consider the behavior of the limit of gradient flow (eq. (2)) as a function of the initialization, in the under-determined $N \ll d$ case where there are many solutions $X\boldsymbol{\beta} = \mathbf{y}$. It is straightforward to compute the tangent kernel at initialization and confirm that $K_0(\mathbf{x}, \mathbf{x}') = 4 \langle \mathbf{x}, \mathbf{x}' \rangle$, i.e. the standard inner inner product kernel (with some scaling), and so $\|\boldsymbol{\beta}\|_{K_0} \propto \|\boldsymbol{\beta}\|_2$. Therefore, in the kernel regime, gradient flow would take us to the minimum L2 norm solution, $\boldsymbol{\beta}^*_{L2} \doteq \arg\min_{X\boldsymbol{\beta}=y} \|\boldsymbol{\beta}\|_2$. Following Chizat and Bach [4] and the discussion in Section 3, we would therefore expect that $\lim_{\alpha\to\infty} \boldsymbol{\beta}_\alpha(\infty) = \boldsymbol{\beta}^*_{L2}$.

In contrast, Gunasekar et al. [11, Corollary 2] shows that when $\alpha \to 0$, gradient flow will lead instead to the minimum L1 norm solution $\lim_{\alpha\to 0} \boldsymbol{\beta}_\alpha(\infty) = \boldsymbol{\beta}^*_{L1} = \arg\min_{X\boldsymbol{\beta}=y} \|\boldsymbol{\beta}\|_1$. This is the "deep regime" in this case. We already see two very distinct behaviors and, in high dimensions, two very different inductive biases, with the deep regime inducing a bias that is *not* an RKHS norm for any choice of kernel. Can we characterize and understand the transition between the two regimes as α transitions from very small to very large? The following theorem does just that.

Theorem 1. *For any $0 < \alpha < \infty$,*

$$\beta_\alpha(\infty) = \hat{\boldsymbol{\beta}}_\alpha \doteq \arg\min_{\boldsymbol{\beta}} Q_\alpha(\boldsymbol{\beta}) \text{ s.t. } X\boldsymbol{\beta} = \mathbf{y}, \quad (5)$$

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

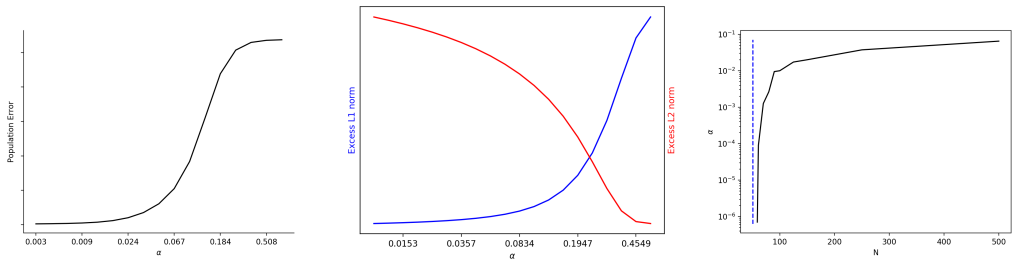
Proof sketch The proof in Appendix A proceeds by showing the gradient flow dynamics on $\mathbf{w}$ lead to a solution of the form

$$\boldsymbol{\beta}_\alpha(\infty) = \alpha^2 \left(\exp\left(-4X^\top \int_0^\infty \boldsymbol{r}_\alpha(t) dt \right) - \exp\left(4X^\top \int_0^\infty \boldsymbol{r}_\alpha(t) dt \right) \right) \quad (6)$$

where $\boldsymbol{r}_\alpha(t) = X\boldsymbol{\beta}_\alpha(t) - \mathbf{y}$. While evaluating the integral would be very difficult, the fact that

$$\boldsymbol{\beta}_\alpha(\infty) \in \left\{ \alpha^2 \left(\exp\left(-X^\top \bar{\boldsymbol{r}}\right) - \exp\left(X^\top \bar{\boldsymbol{r}}\right) : \bar{\boldsymbol{r}} \in \mathbb{R}^N \right) \right\} \quad (7)$$

[4] Our results can be generalized to non-uniform initialization, "biased initiliation" (i.e. where $\mathbf{w}_- \neq \mathbf{w}_+$ at initialization), or the asymmetric parametrization $f((u, v), \mathbf{x}) = \sum_i u_i v_i x_i$, however this complicates the presentation without adding much insight.



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

Figure 1: In Figure 1a, the population error of the gradient flow solution is shown as a function of the initialization. The data are generated by a 5-sparse predictor $\boldsymbol{\beta}^*$ according to $y^{(n)} \sim N(\langle \boldsymbol{\beta}^*, \mathbf{x}^{(n)} \rangle, 0.01)$ with $d = 1000$ and $N = 100$. In Figure 1b, the excess L1 norm of the gradient flow solution $\|\boldsymbol{\beta}_\alpha(\infty)\|_1 - \|\boldsymbol{\beta}^*_{L1}\|_1$ is shown as a function of α in blue. In red is the same for the excess L2 norm $\|\boldsymbol{\beta}_\alpha(\infty)\|_2 - \|\boldsymbol{\beta}^*_{L2}\|_2$. In Figure 1c, for the same sparse regression problem described above, the largest α such that $\boldsymbol{\beta}_\alpha(\infty)$ achieves population error at most 0.025 is shown in black. The blue dashed line indicates the minimum number of samples required for $\boldsymbol{\beta}^*_{L1}$ to achieve this error.

already provides a dual certificate for the KKT conditions for $\min_{\boldsymbol{\beta}} Q_\alpha(\boldsymbol{\beta})$ s.t. $X\boldsymbol{\beta} = \mathbf{y}$. □

The function Q_α, also known as the "hypentropy" regularizer [10], can thus be understood as an implicit regularizer which biases the gradient flow solution towards a particular zero-error solution out of the many possibilities. As α ranges from 0 to ∞, the Q_α regularizer interpolates between the L1 and L2 norms, as illustrated in Figure 2, which shows a single 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 controlled by the behaviour of $q(z)$ around $z = 0$. In this regime $q(z) = \frac{z^2}{4} + O(z^4)$ is quadratic, and so $Q_\alpha(\boldsymbol{\beta}) \approx\propto \sum_i \boldsymbol{\beta}_i^2 = \|\boldsymbol{\beta}\|_2^2$. On the other hand when $\alpha \to 0$, we have that $\boldsymbol{\beta}_i/\alpha^2 \to \infty$ and the behaviour is governed by the asymptotic behaviour $q(z) = \Theta(z \log z)$ as $z \to \infty$. In this regime $Q_\alpha(\boldsymbol{\beta}) \approx \sum_i \frac{\boldsymbol{\beta}_i}{\alpha^2} \log \frac{\boldsymbol{\beta}_i}{\alpha^2} = \sum_i \left(2\frac{\log 1/\alpha}{\alpha^2}\right) \boldsymbol{\beta}_i + O(\frac{1}{\alpha^2}) \propto \|\boldsymbol{\beta}\|_1 + o(1)$. But for any initialization scale α, Q_α describes exactly how training will interpolate between the kernel and deep regimes.

The following Theorems, proven in Appendix B, provide a quantitative statement of how the ℓ_1 and ℓ_2 norms are approached as $\alpha \to 0$ and $\alpha \to \infty$ respectively:

Theorem 2. *For any* $0 < \epsilon < d$,

$$\alpha \leq \min\left\{ (2(1+\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-\frac{d}{\epsilon\|\boldsymbol{\beta}^*_{L1}\|_1}\right)\right\} \implies \left\|\hat{\boldsymbol{\beta}}_\alpha\right\|_1 \leq (1+\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$$

Theorem 3. *For any* $\epsilon > 0$

$$\alpha \geq \sqrt{2(1+\epsilon)\left(1+\frac{2}{\epsilon}\right)\|\boldsymbol{\beta}^*_{L2}\|_2} \implies \left\|\hat{\boldsymbol{\beta}}_\alpha\right\|_2^2 \leq (1+\epsilon)\|\boldsymbol{\beta}^*_{L2}\|_2^2$$

Theorems 2 and 3 and Figure 1b indicate a certain asymmetry between reaching the deep and kernel regimes: a relatively small value of α (polynomial in the accuracy) suffices to approximate the minimum L2 norm solution to a very high degree of accuracy. On the other hand, α needs to be exponentially small in order for the minimum Q_α solution to approximate the minimum L1 norm solution. From an optimization perspective this is unfortunate because $\mathbf{w} = 0$ is a saddle point, so taking $\alpha \to 0$ will quickly create numerical difficulties since the time needed to escape the vicinity of the saddle point will grow drastically.

Generalization In order to understand the effects of initialization on generalization, and how we might need to be in the deep regime in order to generalize well, consider a simple sparse regression problem, where $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \sim N(0, I)$ and $y^{(n)} = \langle \boldsymbol{\beta}^*, \mathbf{x}^{(n)} \rangle + N(0, 0.01)$ where $\boldsymbol{\beta}^*$ is r^*-sparse and its non-zero entries are $1/\sqrt{r^*}$. When $N \leq d$, gradient flow will reach a zero training error solution, however, not all of these solutions will generalize the same. With $N = \Theta(r^* \log d)$ samples, the deep regime, i.e. the minimum ℓ_1 norm solution will generalize well, but even though

we can fit the training data perfectly well, we should not expect any generalization in the kernel regime with this sample size ($N = \Omega(d)$ samples will be required in that regime). This is demonstrated in Figure 1c.

We see that in order to generalize well, we might need to use small initialization, and generalization improves as we decrease the scale of initialization α. There is a tension here between generalization and optimization: a smaller α might improve generalization, but as discussed above makes optimization trickier as we are starting closer to a saddle point. This suggests that in practice we would want to compromise, and operate just at the edge of the deep regime, using the largest α that still allows for generalization. The tension between optimization and generalization can also be seen through a tradeoff between the sample size and the largest α we can use and still generalize. This is illustrated in Figure 1c, where for each sample size N, we plot the largest α for which the gradient flow solution $\hat{\beta}_\alpha$ achieves population risk below some threshold. As N approaches the number of samples needed for the minimum L1 solution to generalize (the vertical dashed line), the initialization α indeed must become extremely small. However, generalization is much easier when the number of samples is only slightly larger, and we can use much more moderate initialization.

The situation we describe here is similar to a situation studied by Mei et al. [19], who considered one-pass stochastic gradient descent (i.e. SGD on the population objective) and analyzed the number of steps, and so also number of samples, required for generalization. Mei et al. showed that even with large initialization one can achieve generalization by optimizing with more one-pass SGD steps. Our analysis suggests that the issue here is not that of optimizing longer or more accurately, but rather of requiring a larger sample size—in studying one-pass SGD this distinction is blurred, but our analysis separates between the two.

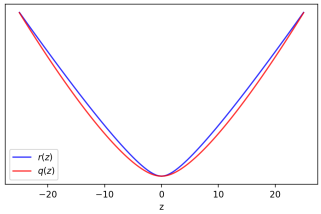


Figure 2: $q(z)$ and $r(z)$ (scaled)

Explicit Regularization It is tempting to imagine that the effect of implicit regularization through gradient descent corresponds to selecting the solution closest to initialization in Euclidean norm:

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

where

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

It is certainly the case for standard linear regression $f(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$, that $\boldsymbol{\beta}_\alpha(\infty) = \boldsymbol{\beta}_\alpha^R$ and the implicit bias is fully captured by this view. Is the implicit bias of $\boldsymbol{\beta}_\alpha(\infty)$ indeed captured by this minimum Euclidean distance solution also for our 2-homogeneous (depth 2) model, and perhaps more generally? Can the behavior discussed above can also be explained by R_α?

Indeed, it is easy to verify that for our square parametrization, the limiting behavior when $\alpha \to 0$ and $\alpha \to \infty$ of the two approaches match, i.e. $\lim_{\alpha\to 0} \boldsymbol{\beta}_\alpha^R = \boldsymbol{\beta}_{L1}^*$ and $\lim_{\alpha\to\infty} \boldsymbol{\beta}_\alpha^R = \boldsymbol{\beta}_{L2}^*$. To check whether the complete behaviour and transition are also captured by (8), we can calculate $R_\alpha(\boldsymbol{\beta})$, which decomposes over the coordinates, as[5]:

$$R_\alpha(\boldsymbol{\beta}) = R(\boldsymbol{\beta}/\alpha^2) = \sum_i r(\boldsymbol{\beta}_i/\alpha^2) \tag{10}$$

Where $r(z)$ is the unique real root of $u^4 - 6u^3 + (12 - 2z^2)u^2 - (8 + 10z^2)u + z^2 + z^4$ w.r.t. u.

As depicted in Figure 2, $r(z)$ is quadratic around $z = 0$ and asymptotically linear as $z \to \infty$, yielding L_1 regularization when $\alpha \to 0$ and L_2 regularization as $\alpha \to \infty$, similarly to $q(z)$. However, $q(z)$ and $r(z)$ are very different: $r(z)$ is quadratic (even radical), while $q(z)$ is transcendental. This implies $Q_\alpha(\boldsymbol{\beta})$ and $R_\alpha(\boldsymbol{\beta})$ are substantially different, are not simple rescaling of each other, and hence will lead to different sets or "paths" of solutions, $\left\{\boldsymbol{\beta}_\alpha^R\right\}_\alpha$ and $\left\{\boldsymbol{\beta}(\infty)_\alpha\right\}_\alpha$. In particular, while α needed to be exponentially small in order for Q_α to approximate the $L1$ norm, and so for the limit

[5]Substituting $w_- = \sqrt{-\beta - w_+^2}$ and equating the gradient w.r.t. w_+ to zero leads to a quadratic equation, the solution of which can be substituted back to evaluate $R_\alpha(\beta)$

of the gradient flow path to approximate the minimum $L1$ norm solution, being algebraic $R_\alpha(\boldsymbol{\beta})/\alpha^2$ converges to $\|\boldsymbol{\beta}\|_1$ polynomially (that is, α only needs to scale polynomially with the accuracy). We see, that implicit regularization effect of gradient descent (or gradient flow), and the transition from the kernel to deep regime, is more complex and subtle than what is captured simply by distances in parameter space.

5 Demonstration in Matrix Completion

We now turn to a more complex depth two model, namely a matrix factorization model, and demonstrate similar transitions empirically. Specifically, we consider the model over matrix inputs $X \in \mathbb{R}^{d\times d}$ defined by $f((U,V),X) = \left\langle UV^\top,\, X\right\rangle$, where $U,V \in \mathbb{R}^{d\times k}$, $k \geq d$. This corresponds to linear predictors over matrix arguments specified by $M_{U,V} = F(U,V) = UV^\top$. For generic inputs $X \in \mathbb{R}^{d\times d}$ this can be thought of as a *matrix sensing* problem, where $X^{(n)}$ are measurement matrices. We consider here a matrix completion problem where $X^{(n)} = e_{i^{(n)}}e_{j^{(n)}}^\top$ represents an observation of entry $(i^{(n)}, j^{(n)})$: $\left\langle M,\, X^{(n)}\right\rangle = M_{i^{(n)},j^{(n)}}$, and we observe some subset of the entries of the matrix and would like to complete the unobserved entries.

In the overparametrized regime $k \geq d$, the model itself does not impose any constraints on the linear predictor $M_{U,V}$, and so for learning with $N < d^2$ samples (as would always be the case for matrix completion), we need to rely on the implicit bias of gradient descent. In particular, consider matrix completion with $N = \Theta(dr\log d)$ observations of a planted rank r matrix, with $r \ll d$. For such underdetermined problems, there are many trivial global minimizers of the loss, most of which are not low rank and hence will not guarantee recovery, and we must rely on some other inductive bias. Indeed, previous work [11, 17] demonstrated rich implicit bias when $\alpha \to 0$, showing (theoretically and/or analytically) that in this regime we would converge to the minimum nuclear norm solution and would be able to generalize (or reconstruct) a low rank model. Crucially, these analysis depend on initialization with scale $\alpha \to 0$. Here we consider what happens with larger scale unbiased initialization (i.e. when $F(U(0),V(0)) = 0$ even though $U(0), V(0) \not\to 0$).

Similar to Section 4, in order to get unbiased initialization, we consider $k = 2k' \geq 2d$ and initialization of the form $U(0) = \alpha\,[U_0, -U_0]$ and $V(0) = \alpha\,[V_0, V_0]$, where $U_0, V_0 \in \mathbb{R}^{d\times k/2}$. We will study implicit bias of gradient flow over the factorized parametrization with above initialization.

We will focus on matrix completion problems where inputs are of the form $X = e_{i_X}e_{j_X}^\top$. The tangent kernel at initialization is given by $K_0(X,X') \propto \langle U_0[i_X,:],\, U_0[i_{X'},:]\rangle\, \mathbf{1}(j_X = j_{X'}) + \langle V_0[j_X,:],\, V_0[j_{X'},:]\rangle\, \mathbf{1}(i_X = i_{X'})$. This defaults to the trivial delta kernel $K(X,X') = \mathbf{1}(i_X = i_{X'})\cdot\mathbf{1}(j_X = j_{X'})$ for the two special cases (a) U_0, V_0 have orthogonal columns (e.g. $U_0 = V_0 = I$), or (b) U_0, V_0 have independent Gaussian entries and $k \to \infty$. In these cases, minimizing the RKHS norm of the tangent kernel corresponds to returning a zero imputed matrix (minimum Frobenius norm solution). Said differently, in the "kernel" regime training is truly lazy: the unobserved entries do not change at all during training, and instead we just adjust the observed entries to fit the observations. We cannot expect any generalization in this regime, no matter what we assume about the observed matrix. In contrast, in the "deep" regime, as was previously observed by Gunasekar et al., training leads to the minimum nuclear norm solution, a rich inductive bias that allows for generalization [3, 20]. Figure 3 demonstrates the transition between the two regimes, and how recovery deteriorates as we move away from the "deep" regime and into the "kernel" regime, changing the unobserved entries less and less.

6 Discussion

The main point of this paper is to emphasize the distinction between the "kernel" regime in training overparametrized multi-layered networks, and the "deep" (rich, active, adaptive) regime, show how the scaling of the initialization can transition between them, and understand this transition in detail. We argue that rich inductive bias that enables generalization may arise in the deep regime, but that focusing on the kernel regime restricts us to only what can be done with an RKHS. By studying the transition we also see a tension between generalization and optimization, which suggests we would tend to operate just on the edge of the deep regime, and so understanding this transition, rather then just the extremes, is important. Furthermore, we see that at this operating regime, at the edge of the deep regime, the implicit bias of gradient descent differs substantively from that of explicit

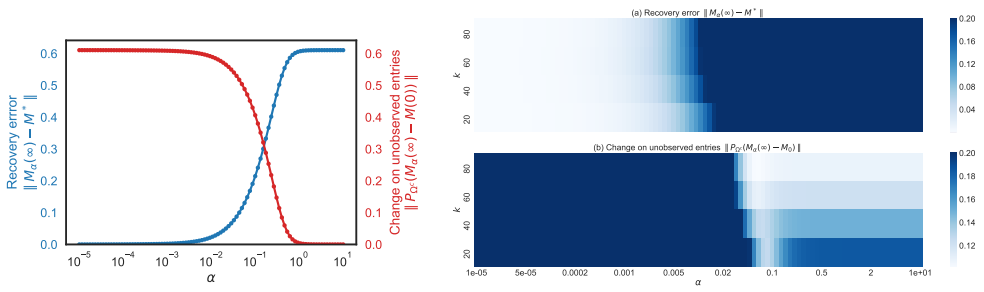


Figure 3: **Regimes in Matrix Completion** We generated a 10×10 rank-one matrix completion problem with ground truth $M^* = u^*(v^*)^\top$ by generating $u^*, v^* \in \mathbb{R}^{10}$ with i.i.d. $\mathcal{N}(0,1)$ entries and observing $N = 60$ random entries Ω. We fit the observed entries by minimizing the squared loss on a matrix factorization model $F(U,V) = UV^\top$ with $U, V \in \mathbb{R}^{d \times 2k}$. For different scalings α, we examine the matrix $M(\infty)$ reached by gradient flow on U, V (solved using python ODE solvers) and plot (i) the reconstruction error on unobserved entries $\sum_{ij \notin \Omega}(M^*_{ij} - M(\infty)_{ij})^2$, and (ii) the amount by which the unobserved entries changed during optimization $\sum_{ij \notin \Omega}(M(\infty)_{ij} - M(0)_{ij})^2$. In (a) we used $k = 2d$ and initialized to $U(0) = \alpha[I, -I]$ and $V(0) = \alpha[I, I]$. In Appendix C we also plot the nuclear norm and Frobenius norms of $M(\infty)$, and observe an almost identical figure. In (b) for varying k, we initialized to $U(0) = \alpha[U_0, -U_0]$ and $V(0) = \alpha[V_0, V_0]$ with $U_0, V_0 \in \mathbb{R}^{d \times k}$ with i.i.d. $\mathcal{N}(0,1)$ entries. For large k, the tangent kernel converges to the kernel corresponding to the Frobenius norm, and so as $\alpha \to \infty$ we again see the unobserved entries do not change. The scaling required to transition between the deep regime (reconstruction) and the kernel regime changes with k.

regularization. Although in our detailed study we focused on a simple model, so that we can carry out a complete and exact analysis analytically, we expect this to be representative of the behaviour also in other homogeneous models, and serve as a basis of a more general understanding.

Effect of Width Our treatment focused on the effect of scale on the transition between the regimes, and we saw that, as pointed out by Chizat and Bach, we can observe a very meaningful transition between a kernel and deep regime even for finite width parametric models. The transition becomes even more interesting if the width of the model (the number of units per layer, and so also the number of parameters) increases towards infinity. In this case, we must be careful as to how the initialization of each individual unit scales when the total number of units increase, and which regime we fall in to is controlled by the relative scaling of the width and the scale of individual units at initialization. This is demonstrated, for example, in Figure 4a-4b, which shows the regime change in matrix factorization problems, from minimum Frobenius norm recovery (the kernel regime) to minimum nuclear norm recovery (the deep regime), as a function of both the number of factors k and the scale of initialization of each factor α . As is expected, the scale α at which we see the transition decreases as the model becomes wider, but further study is necessary in order to obtain a complete understanding of this scaling.

A particularly interesting aspect of infinite width networks is that , unlike for fixed-width networks, it may be possible to scale α relative to the width k such that at the infinite-width limit we would have an (asymptotically) unbiased predictor at initialization $\lim_{k\to\infty} F_k(\mathbf{w}(0)) = 0$, or at least a non-exploding initialization $\limsup_{k\to\infty} \|F_k(\mathbf{w}(0))\| = O(1)$, even with random initialization (without a doubling trick leading to artificially unbiased initialization), while still being in the kernel regime. For two-layer networks with ReLU activation, this is has been established in [2] which showed that with width $k \geq \text{poly}(\frac{1}{\max_{\|x\|_2 \leq 1} \|f_0(x)\|}, N)$ the gradient dynamics stay in the kernel regime forever.

Another interesting question is whether as the width increases, the transition between the kernel and deep regimes becomes sharp, or perhaps for infinite width models there is a wide intermediate regime with distinct and interesting behaviour.

Acknowledgements BW is supported the NSF Graduate Research Fellowship under award 1754881. JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0303. This is part of the collaboration between US DOD, UK MOD and UK Engineering and Physical Research Council (EPSRC) under the Multidisciplinary University Research Initiative. The work of DS was supported by the Israel Science Foundation (grant No. 31/1031), and by the Taub Foundation. NS

is supported by NSF Medium (grant No. NSF-102:1764032) and by NSF BIGDATA (grant No. NSF-104:1546500).

References

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

[2] 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*, 2019.

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

[4] Lenaic Chizat and Francis Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

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

[6] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[7] 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.

[8] 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.

[9] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

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

[11] 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.

[12] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

[13] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018.

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[15] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[16] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

[17] 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.

[18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. In *Proceedings of the National Academy of Sciences*, volume 115, pages E7665–E7671, 2018.

[19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

[20] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[21] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? *arXiv preprint arXiv:1902.05040*, 2019.

[22] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70), 2018.

[23] 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.

[24] Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression. 03 2019.

A Proof of Theorem 1

Theorem 1. *For any* $0 < \alpha < \infty$*,*

$$\beta_\alpha(\infty) = \hat{\boldsymbol{\beta}}_\alpha \doteq \arg\min_{\boldsymbol{\beta}} Q_\alpha(\boldsymbol{\beta}) \text{ s.t. } X\boldsymbol{\beta} = \mathbf{y}, \tag{5}$$

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

Proof. The proof involves relating the set of points reachable by gradient flow on $\mathbf{w}$ to the KKT conditions of the Q_α minimization problem. While it may not be obvious from the expression, $q : \mathbb{R} \to \mathbb{R}$ is the integral of an increasing function and is thus convex, and Q_α is the sum of q applied to individual coordinates of β, and is therefore also convex.

The linear predictor $\boldsymbol{\beta}_\alpha(\infty)$ is given by F applied to the limit of the gradient flow dynamics on $\mathbf{w}$. Recalling that $\tilde{X} = [X \quad -X]$,

$$\dot{\mathbf{w}}_\alpha(t) = -\nabla L(\mathbf{w}_\alpha(t)) = -\nabla\left(\left\|\tilde{X}\mathbf{w}_\alpha(t)^2 - y\right\|_2^2\right) = -2\tilde{X}^\top r_\alpha(t) \circ \mathbf{w}_\alpha(t) \tag{11}$$

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:

$$\mathbf{w}_\alpha(t) = \mathbf{w}_\alpha(0) \circ \exp\left(-2\tilde{X}^\top \int_0^t r_\alpha(s)ds\right) \tag{12}$$

This immediately gives an expression for $\boldsymbol{\beta}_\alpha(t)$:

$$\boldsymbol{\beta}_\alpha(t) = \mathbf{w}_{\alpha,+}(t)^2 - \mathbf{w}_{\alpha,-}(t)^2 \tag{13}$$

$$= \alpha^2\left(\exp\left(-4X^\top \int_0^t r_\alpha(s)ds\right) - \exp\left(4X^\top \int_0^t r_\alpha(s)ds\right)\right) \tag{14}$$

Understanding the limit $\boldsymbol{\beta}_\alpha(\infty)$ exactly requires calculating $\int_0^\infty r_\alpha(s)ds$, which would be a difficult task. However, for our purposes, it is sufficient to know that there is some $\bar{r}_\alpha \in \mathbb{R}^n$ such that $\int_0^\infty r_\alpha(s)ds = \bar{r}_\alpha$. In other words, the vector $\boldsymbol{\beta}_\alpha(\infty)$ is contained in the non-linear manifold

$$\hat{\beta}_\alpha \in \mathcal{M}_\alpha = \left\{\alpha^2\left(\exp\left(-4X^\top\bar{r}\right) - \exp\left(4X^\top\bar{r}\right)\right) : \bar{r} \in \mathbb{R}^n\right\} \tag{15}$$

Setting this aside for a moment, consider the KKT conditions of the convex program

$$\min_\beta Q_\alpha(\boldsymbol{\beta}) \text{ s.t. } X\boldsymbol{\beta} = y \tag{16}$$

which are

$$\exists\nu \quad \nabla Q_\alpha(\boldsymbol{\beta}) = X^\top\nu \tag{17}$$

$$X\boldsymbol{\beta} = y \tag{18}$$

Expanding ∇Q_α in (17), there must exist ν such that

$$\operatorname{arcsinh}\left(\frac{\boldsymbol{\beta}}{2\alpha^2}\right) = X^\top\nu \tag{19}$$

$$\iff \boldsymbol{\beta} = 2\alpha^2\sinh\left(X^\top\nu\right) \tag{20}$$

$$\iff \boldsymbol{\beta} = \alpha^2\left(\exp\left(X^\top\nu\right) - \exp\left(-X^\top\nu\right)\right) \tag{21}$$

Since we already know that the gradient flow solution $\boldsymbol{\beta}_\alpha\infty \in \mathcal{M}_\alpha$, there is some $\bar{r}_\alpha$ such that $\nu = -4\bar{r}_\alpha$ is a certificate of (17). Furthermore, this problem satisfies the strict saddle property [9] [24, Lemma 2.1], therefore gradient flow will converge to a zero-error solution, i.e. $X\boldsymbol{\beta}_\alpha(\infty) = y$. Thus, we conclude that $\boldsymbol{\beta}_\alpha(\infty)$ is a solution to (16). □

B Proofs of Theorems 2 and 3

Lemma 1. *For any $\beta \in \mathbb{R}^d$,*

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

guarantees that

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

Proof. First, we show that $Q(\beta/\alpha^2) = Q_\alpha(|\beta|/\alpha^2)$. Observe that $g(x) = x\arcsin(x/2)$ is even because x and $\arcsin(x/2)$ are odd. Therefore,

$$\begin{aligned}
Q(\beta/\alpha^2) &= \sum_{i=1}^d 2 - \sqrt{4+\frac{\beta_i^2}{\alpha^4}} + \frac{\beta_i}{\alpha^2}\operatorname{arcsinh}\left(\frac{\beta_i}{2\alpha^2}\right) && (22)\\
&= \sum_{i=1}^d 2 - \sqrt{4+\frac{\beta_i^2}{\alpha^4}} + g\left(\frac{\beta_i}{\alpha^2}\right) && (23)\\
&= \sum_{i=1}^d 2 - \sqrt{4+\frac{|\beta_i|^2}{\alpha^4}} + g\left(\left|\frac{\beta_i}{\alpha^2}\right|\right) && (24)\\
&= Q_\alpha(|\beta|/\alpha^2) && (25)
\end{aligned}$$

Therefore, we can rewrite

$$\begin{aligned}
\frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) &= \frac{\alpha^2}{\ln(1/\alpha^2)} Q(|\beta|/\alpha^2) && (26)\\
&= \sum_{i=1}^d \frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{\sqrt{4\alpha^4+\beta_i^2}}{\ln(1/\alpha^2)} + \frac{|\beta_i|}{\ln(1/\alpha^2)}\operatorname{arcsinh}\left(\frac{|\beta_i|}{2\alpha^2}\right) && (27)\\
&= \sum_{i=1}^d \frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{\sqrt{4\alpha^4+\beta_i^2}}{\ln(1/\alpha^2)} + \frac{|\beta_i|}{\ln(1/\alpha^2)}\ln\left(\frac{|\beta_i|}{2\alpha^2} + \sqrt{1+\frac{\beta_i^2}{4\alpha^4}}\right) && (28)\\
&= \sum_{i=1}^d \frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{\sqrt{4\alpha^4+\beta_i^2}}{\ln(1/\alpha^2)} + |\beta_i|\left(1 + \frac{\ln\left(\frac{|\beta_i|}{2} + \sqrt{\alpha^4+\frac{\beta_i^2}{4}}\right)}{\ln(1/\alpha^2)}\right) && (29)
\end{aligned}$$

Using the fact that

$$|a| \leq \sqrt{a^2+b^2} \leq |a| + |b| \tag{30}$$

we can bound for $\alpha < 1$

$$\begin{aligned}
\frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) &\leq \sum_{i=1}^d \frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{2\alpha^2}{\ln(1/\alpha^2)} + |\beta_i|\left(1 + \frac{\ln\left(\frac{|\beta_i|}{2} + \alpha^2 + \frac{|\beta_i|}{2}\right)}{\ln(1/\alpha^2)}\right) && (31)\\
&= \sum_{i=1}^d |\beta_i|\left(1 + \frac{\ln\left(|\beta_i| + \alpha^2\right)}{\ln(1/\alpha^2)}\right) && (32)\\
&\leq \|\beta\|_1\left(1 + \max_{i\in[d]} \frac{\ln\left(|\beta_i| + \alpha^2\right)}{\ln(1/\alpha^2)}\right) && (33)\\
&\leq \|\beta\|_1\left(1 + \frac{\ln\left(\|\beta\|_1 + \alpha^2\right)}{\ln(1/\alpha^2)}\right) && (34)\\
& && (35)
\end{aligned}$$

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

$$\frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) \leq \|\beta\|_1 \left(1 + \frac{\ln\left(\|\beta\|_1 + \alpha^2\right)}{\ln(1/\alpha^2)}\right) \tag{36}$$

$$\leq \|\beta\|_1 \left(1 + \frac{\ln\left(2\|\beta\|_1\right)}{\ln(1/\alpha^2)}\right) \tag{37}$$

$$\leq \|\beta\|_1 (1+\epsilon) \tag{38}$$

On the other hand, using (29) and (30) again,

$$\frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) \geq \sum_{i=1}^{d} \frac{2\alpha^2}{\ln(1/\alpha^2)} - \frac{|\beta_i| + 2\alpha^2}{\ln(1/\alpha^2)} + |\beta_i| \left(1 + \frac{\ln\left(|\beta_i|\right)}{\ln(1/\alpha^2)}\right) \tag{39}$$

$$= \sum_{i=1}^{d} |\beta_i| \left(1 + \frac{\ln\left(|\beta_i|\right) - 1}{\ln(1/\alpha^2)}\right) \tag{40}$$

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

$$\frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) \geq \sum_{i=1}^{d} |\beta_i| - \frac{1}{\ln(1/\alpha^2)} \tag{41}$$

$$= \|\beta\|_1 - \frac{d}{\ln(1/\alpha^2)} \tag{42}$$

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

$$\frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) \geq \|\beta\|_1 (1-\epsilon) \tag{43}$$

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

$$\|\beta\|_1 (1-\epsilon) \leq \frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) \leq \|\beta\|_1 (1+\epsilon) \tag{44}$$

□

Theorem 2. *For any* $0 < \epsilon < d$*,*

$$\alpha \leq \min\left\{(2(1+\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-\frac{d}{\epsilon\|\boldsymbol{\beta}^*_{L1}\|_1}\right)\right\} \implies \left\|\hat{\boldsymbol{\beta}}_\alpha\right\|_1 \leq (1+\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$$

Proof. First, we will prove that $\left\|\hat{\beta}_\alpha\right\|_1 < (1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$. By Lemma 1, since $\alpha \leq \alpha_1\left(\frac{\epsilon}{2+\epsilon}, (1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1, d\right)$, for all β with $\|\beta\|_1 \leq (1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$ we have

$$\|\beta\|_1 \left(1 - \frac{\epsilon}{2+\epsilon}\right) \leq \frac{\alpha^2}{\ln(1/\alpha^2)} Q(\beta/\alpha^2) \leq \|\beta\|_1 \left(1 + \frac{\epsilon}{2+\epsilon}\right) \tag{45}$$

Let β be such that $X\beta = y$ and $\|\beta\|_1 = (1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$. Then

$$\frac{\alpha^2}{\ln(1/\alpha^2)}Q(\beta/\alpha^2) \geq \left(1 - \frac{\epsilon}{2+\epsilon}\right)\|\beta\|_1 \tag{46}$$

$$= \left(1 - \frac{\epsilon}{2+\epsilon}\right)(1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1 \tag{47}$$

$$\geq \frac{\left(1 - \frac{\epsilon}{2+\epsilon}\right)}{\left(1 + \frac{\epsilon}{2+\epsilon}\right)}(1+2\epsilon)\frac{\alpha^2}{\ln(1/\alpha^2)}Q(\boldsymbol{\beta}^*_{L1}/\alpha^2) \tag{48}$$

$$= \frac{1+2\epsilon}{1+\epsilon}\frac{\alpha^2}{\ln(1/\alpha^2)}Q(\boldsymbol{\beta}^*_{L1}/\alpha^2) \tag{49}$$

$$> \frac{\alpha^2}{\ln(1/\alpha^2)}Q(\boldsymbol{\beta}^*_{L1}/\alpha^2) \tag{50}$$

$$\geq \frac{\alpha^2}{\ln(1/\alpha^2)}Q(\hat{\beta}_\alpha/\alpha^2) \tag{51}$$

Therefore, $\beta \neq \hat{\beta}_\alpha$. Furthermore, let β be any solution $X\beta = y$ with $\|\beta\|_1 > (1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$. It is easily confirmed that there exists $c \in (0,1)$ such that the point $\beta' = (1-c)\beta + c\boldsymbol{\beta}^*_{L1}$ is satisfies both $X\beta' = y$ and $\|\beta'\|_1 = (1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$. By the convexity of Q, this implies $Q(\beta/\alpha^2) \geq Q(\beta'/\alpha^2) > Q_\alpha(\hat{\beta}_\alpha/\alpha^2)$. Thus a β with a large L1 norm cannot be a solution, even if $\frac{\alpha^2}{\ln(1/\alpha^2)}Q(\beta/\alpha^2) \not\approx \|\beta\|_1$.

Since $\left\|\hat{\beta}_\alpha\right\|_1 < (1+2\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1$, we conclude

$$\left\|\hat{\beta}_\alpha\right\|_1 \leq \frac{1}{1 - \frac{\epsilon}{2+\epsilon}}\frac{\alpha^2}{\ln(1/\alpha^2)}Q(\hat{\beta}_\alpha/\alpha^2) \tag{52}$$

$$\leq \frac{1}{1 - \frac{\epsilon}{2+\epsilon}}\frac{\alpha^2}{\ln(1/\alpha^2)}Q(\boldsymbol{\beta}^*_{L1}/\alpha^2) \tag{53}$$

$$\leq \frac{1 + \frac{\epsilon}{2+\epsilon}}{1 - \frac{\epsilon}{2+\epsilon}}\|\boldsymbol{\beta}^*_{L1}\|_1 \tag{54}$$

$$= (1+\epsilon)\|\boldsymbol{\beta}^*_{L1}\|_1 \tag{55}$$

□

Lemma 2. *For any $\beta \in \mathbb{R}^d$,*

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

guarantees that

$$(1-\epsilon)\|\beta\|_2^2 \leq 4\alpha^4 Q(\beta/\alpha^2) \leq (1+\epsilon)\|\beta\|_2^2$$

Proof. The regularizer Q can be written

$$Q(\beta/\alpha^2) = \sum_{i=1}^{d}\int_0^{\beta_i/\alpha^2} \operatorname{arcsinh}\left(\frac{t}{2}\right)dt \tag{56}$$

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

$$\phi(0) = 0 \tag{57}$$

$$\phi'(0) = \frac{1}{\alpha^2} \operatorname{arcsinh}\left(\frac{z}{2\alpha^2}\right)\Big|_{z=0} = 0 \tag{58}$$

$$\phi''(0) = \frac{1}{\alpha^4\sqrt{4+\frac{z^2}{\alpha^4}}}\Bigg|_{z=0} = \frac{1}{2\alpha^4} \tag{59}$$

$$\phi'''(0) = \frac{-z}{\alpha^8\left(4+\frac{z^2}{\alpha^4}\right)^{3/2}}\Bigg|_{z=0} = 0 \tag{60}$$

$$\phi''''(z) = \frac{3z^2}{\alpha^{12}\left(4+\frac{z^2}{\alpha^4}\right)^{5/2}} - \frac{1}{\alpha^8\left(4+\frac{z^2}{\alpha^4}\right)^{3/2}} \tag{61}$$

Also, note that

$$|\phi''''(z)| = \frac{\left|2z^2 - 4\alpha^4\right|}{\alpha^{12}\left(4+\frac{z^2}{\alpha^4}\right)^{5/2}} \tag{62}$$

$$\leq \frac{z^2+2\alpha^4}{16\alpha^{12}} \tag{63}$$

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

$$\left|\phi(z) - \frac{z^2}{4\alpha^4}\right| = \frac{\phi''''(\xi)}{4!}z^4 \tag{64}$$

$$\implies \left|\phi(z) - \frac{z^2}{4\alpha^4}\right| \leq \sup_{|\xi|\leq|z|}\frac{\phi''''(\xi)}{4!}z^4 \leq \frac{z^6+2\alpha^4z^4}{384\alpha^{12}} = \frac{z^2}{4\alpha^4}\frac{z^4+2\alpha^4z^2}{96\alpha^8} \tag{65}$$

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

$$\left|4\alpha^4 Q_\alpha(\beta) - \|\beta\|_2^2\right| = 4\alpha^4\left|\sum_{i=1}^d \phi(\beta_i) - \frac{\beta_i^2}{4\alpha^4}\right| \tag{66}$$

$$\leq 4\alpha^4\sum_{i=1}^d\left|\phi(\beta_i) - \frac{\beta_i^2}{4\alpha^4}\right| \tag{67}$$

$$\leq \sum_{i=1}^d \beta_i^2 \cdot \frac{\beta_i^4+2\alpha^4\beta_i^2}{96\alpha^8} \tag{68}$$

$$\leq \|\beta\|_2^2 \max_i \frac{\beta_i^4+2\alpha^4\beta_i^2}{96\alpha^8} \tag{69}$$

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

$$(1-\epsilon)\|\beta\|_2^2 \leq 4\alpha^4 Q(\beta/\alpha^2) \leq (1+\epsilon)\|\beta\|_2^2 \tag{70}$$

□

Theorem 3. *For any* $\epsilon > 0$

$$\alpha \geq \sqrt{2(1+\epsilon)\left(1+\frac{2}{\epsilon}\right)\|\boldsymbol{\beta}_{L2}^*\|_2} \implies \left\|\hat{\boldsymbol{\beta}}_\alpha\right\|_2^2 \leq (1+\epsilon)\|\boldsymbol{\beta}_{L2}^*\|_2^2$$

Proof. First, we will prove that $\left\|\hat{\beta}_\alpha\right\|_2 < (1+2\epsilon)\|\boldsymbol{\beta}_{L2}^*\|_2$. By Lemma 2, since $\alpha \geq \alpha_2\left(\frac{\epsilon}{2+\epsilon}, (1+2\epsilon)\|\boldsymbol{\beta}_{L2}^*\|_2\right)$, for all β with $\|\beta\|_2 \leq (1+2\epsilon)\|\boldsymbol{\beta}_{L2}^*\|_2$ we have

$$\|\beta\|_2^2\left(1-\frac{\epsilon}{2+\epsilon}\right) \leq 4\alpha^4 Q(\beta/\alpha^2) \leq \|\beta\|_2^2\left(1+\frac{\epsilon}{2+\epsilon}\right) \tag{71}$$

Let β be such that $X\beta = y$ and $\|\beta\|_2 = (1+2\epsilon)\|\boldsymbol{\beta}^*_{L2}\|_2$. Then

$$4\alpha^4 Q(\beta/\alpha^2) \geq \left(1 - \frac{\epsilon}{2+\epsilon}\right)\|\beta\|_2^2 \tag{72}$$

$$= \left(1 - \frac{\epsilon}{2+\epsilon}\right)(1+2\epsilon)\|\boldsymbol{\beta}^*_{L2}\|_2^2 \tag{73}$$

$$\geq \frac{\left(1 - \frac{\epsilon}{2+\epsilon}\right)}{\left(1 + \frac{\epsilon}{2+\epsilon}\right)}(1+2\epsilon)4\alpha^4 Q(\boldsymbol{\beta}^*_{L2}/\alpha^2) \tag{74}$$

$$= \frac{1+2\epsilon}{1+\epsilon}4\alpha^4 Q(\boldsymbol{\beta}^*_{L2}/\alpha^2) \tag{75}$$

$$> 4\alpha^4 Q(\boldsymbol{\beta}^*_{L2}/\alpha^2) \tag{76}$$

$$\geq 4\alpha^4 Q(\hat{\beta}_\alpha/\alpha^2) \tag{77}$$

Therefore, $\beta \neq \hat{\beta}_\alpha$. Furthermore, let β be any solution $X\beta = y$ with $\|\beta\|_2 > (1+2\epsilon)\|\boldsymbol{\beta}^*_{L2}\|_2$. It is easily confirmed that there exists $c \in (0,1)$ such that the point $\beta' = (1-c)\beta + c\boldsymbol{\beta}^*_{L2}$ satisfies $X\beta' = y$ and $\|\beta'\|_2 = (1+2\epsilon)\|\boldsymbol{\beta}^*_{L2}\|_2$. By the convexity of Q, this implies $Q(\beta/\alpha^2) \geq Q(\beta'/\alpha^2) > Q(\boldsymbol{\beta}^*_{L2}/\alpha^2)$. Thus a β with a large L2 norm cannot be a solution, even if $4\alpha^4 Q(\beta/\alpha^2) \not\approx \|\beta\|_2^2$.

Since $\left\|\hat{\beta}_\alpha\right\|_2 < (1+2\epsilon)\|\beta_2^*\|_2$, we conclude

$$\left\|\hat{\beta}_\alpha\right\|_2^2 \leq \frac{1}{1 - \frac{\epsilon}{2+\epsilon}}4\alpha^4 Q(\hat{\beta}_\alpha/\alpha^2) \tag{78}$$

$$\leq \frac{1}{1 - \frac{\epsilon}{2+\epsilon}}4\alpha^4 Q(\boldsymbol{\beta}^*_{L2}/\alpha^2) \tag{79}$$

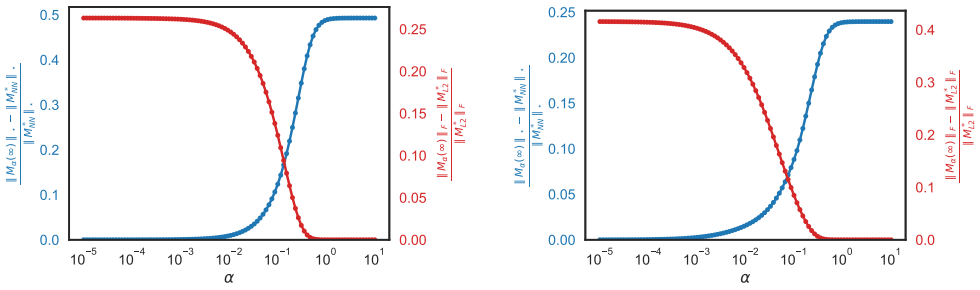
$$\leq \frac{1 + \frac{\epsilon}{2+\epsilon}}{1 - \frac{\epsilon}{2+\epsilon}}\|\boldsymbol{\beta}^*_{L2}\|_2^2 \tag{80}$$

$$= (1+\epsilon)\|\boldsymbol{\beta}^*_{L2}\|_2^2 \tag{81}$$

□

C Matrix experiments

Here, we provide additional results similar to those in Section 5. First, in Figure 4a, we plot the implicit regularization behavior of gradient flow limits with identity initialization as in Figure 3(a): in "deep" regime (small α) we recover the minimum nuclear norm solution $M^*_{NN} = \arg\min_{P_\Omega(M)=y}\|M\|_\star$, while "kernel" regime recovers the minimum Frobenius norm solution $M^*_{L2} = \arg\min_{P_\Omega(M)=y}\|M\|_2$. Figure 4b, is the same plot for an instance matrix sensing problem where the inputs X_n are random i.i.d. Gaussian measurements rather than matrix completion measurements.



(a) Matrix completion.

(b) i.i.d. Gaussian measurements.

Figure 4: The plots demonstrate the regimes where gradient flow with $U_0 = V_0 = I$, implicitly minimizes nuclear norm (blue plots) and L2 norm (red plots), respectively.