Implicit Bias of the Step Size in Linear Diagonal Neural Networks

Mor Shpigel Nacson [1] **Kavya Ravichandran** [2] **Nati Srebro** [2] **Daniel Soudry** [1]

Abstract

Focusing on diagonal linear networks as a model for understanding the implicit bias in underdetermined models, we show how the gradient descent step size can have a large qualitative effect on the implicit bias, and thus on generalization ability. In particular, we show how using large step size for non-centered data can change the implicit bias from a ”kernel” type behavior to a ”rich” (sparsity-inducing) regime — even when gradient flow, studied in previous works, would not escape the ”kernel” regime. We do so by using dynamic stability, proving that convergence to dynamically stable global minima entails a bound on some weighted ℓ_1-norm of the linear predictor, i.e. a ”rich” regime. We prove this leads to good generalization in a sparse regression setting.

1. Introduction

It is becoming evident that implicit regularization guided by the optimization procedure plays a crucial role in learning using underdetermined (overparameterized) models, including deep networks (Neyshabur et al., 2015), and that this optimization-induced bias can ensure learning and generalization (e.g., Moroshko et al. (2020); Li et al. (2019a); Arora et al. (2019); Woodworth et al. (2020)). Different optimization choices, such as using different optimization methods (Gunasekar et al., 2018b), or different optimization parameters can significantly change the algorithmic bias, and thus completely change the effective inductive bias of learning and the ability to generalize in specific scenarios. In order to understand learning with underdetermined models, and be able to make informed principled choices about the optimization procedure to use, it is important to understand how such optimization choices affect the algorithmic bias. In this paper we focus on the effect of the gradient descent step size on the learned predictor in an underdetermined model. We do so in the squared parameterization linear regression (diagonal linear network) model.

The squared parameterization linear regression model (Woodworth et al., 2020) is perhaps the simplest non-linear (in the parameters) model that displays rich, non-trivial algorithmic bias. It can also be thought of as the simplest possible “deep” (that is, deeper than a single unit) model, namely a diagonal linear network (a depth two network with linear activation and diagonal weight matrices). As such, it has been used to study the implicit regularization phenomenon, and in particular the effect of different optimization choices on the algorithmic bias. Even this very simple model displays rich inductive bias that depend on optimization choices: although no explicit ℓ_1 regularization is imposed, under certain optimization choices (but not others!), optimization biases us to a low ℓ_1-norm predictor, which is sufficient for ensuring generalization, e.g. in a sparse regression setting. But under other optimization choices, the model behaves as a kernel machine, leading to implicit ℓ_2 regularization, which is useless in a sparse regression setting and does not lead to generalization.

What are these choices that switch between a kernel regime and a regime where sparse regression is possible? Previous work studied how initialization scale (Woodworth et al., 2020), stochasticity (Damian et al., 2021; Pesme et al., 2021; Blanc et al., 2020), and relative scale between layers (Azulay et al., 2021) affect the algorithmic bias in the squared parametrization model, can take optimization outside the kernel regime and allow generalization. In particular, (Woodworth et al., 2020) showed how with infinitesimal step size, generalization in a sparse regression model is possible only with small initialization scale, while a larger initialization scale forces us into the kernel regime and does not allow generalization. This is problematic since very small initialization scales, particularly those required for leaving the kernel regime in wide models, correspond to initializing very close to a saddle point which might be difficult to escape. Unlike this previous analysis which focused on infinitesimal step size (like much of the implicit regularization analysis), here we consider gradient descent (GD) with different (positive, finite) step sizes, to study the effect of step size on the implicit regularization. In particular, our results demonstrate and elucidate how a large step size can also allow us to escape the kernel regime, biasing opti-

[1] Technion, Israel [2] TTI Chicago, USA. Correspondence to: Mor Shpigel Nacson < mor.shpigel@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

mization toward low ℓ_1 norm predictors and thus to sparse learning and generalization. This remains true even at initialization scales where small (or infinitesimal) step sizes would correspond to the kernel regime, thereby leading to ℓ_2 implicit bias and not allowing for sparse learning.

Figure 1 demonstrates the effect of the step size on the algorithmic bias and generalization ability in a sparse regression setting. Here, we trained an underdetermined squared parameterization linear model on data generated from an unknown sparse model. With very small step size (e.g. 10^{-4}, top solid curve), generalization is only possible when the initialization scale α is small, while larger initialization scales behaves similar to the minimum ℓ_2 norm predictor (dashed blue line, on top), which does not generalize well. However, increasing the stepsize (after appropriate warmup—-see details in Section 7), leads to lower ℓ_1 norm, and better generalizing predictors, regardless of the initialization scale. Using the largest possible step size (on the edge of stability), allows for generalizing essentially as well as using explicit ℓ_1 regularization, even at large initialization scales.

We explain the effect of step size on the algorithmic bias as follows: we first discuss how using a larger step size only allows us to converge to "dynamically stable" solutions, i.e. predictors whose "stability" in parameter space (maximal eigenvalue of the Hessian) is small (Section 2). After giving background on the kernel and rich regime (Section 3), we show how this "stability" implies a bound on a weighted ℓ_1-norm of the predictor, where coordinates of the predictor are weighted by the empirical mean of data coordinates (Section 4). If the data is not centered (i.e., it has non-zero mean), then this empirically weighted ℓ_1 norm can bound the unweighted ℓ_1-norm. We use this to explain how, for sparse regression problems with non-centered data (as was used in Figure 1), large step sizes lead a low ℓ_1-norm solution, and how this entails generalization (Section 5).

The data being non-centered turns out to be crucial, not only for our analysis, but also to reap the benefits of a large step size: consistent with our theory, we see how for zero-mean data, increasing the step size cannot make up for using a large initialization scale, and does not substantially help in generalization. We thus provide a novel connection between data being non-centered[1] and the beneficial implicit bias effects of using large step sizes.

Lastly, in Section 6 we extend our results to "deeper" linear diagonal models (higher order parametrizations). We observe that for large step sizes, deeper models (i.e., with depth greater then two) are biased towards less sparse solutions. Interestingly, this is very different behavior from what was observed in previous works.

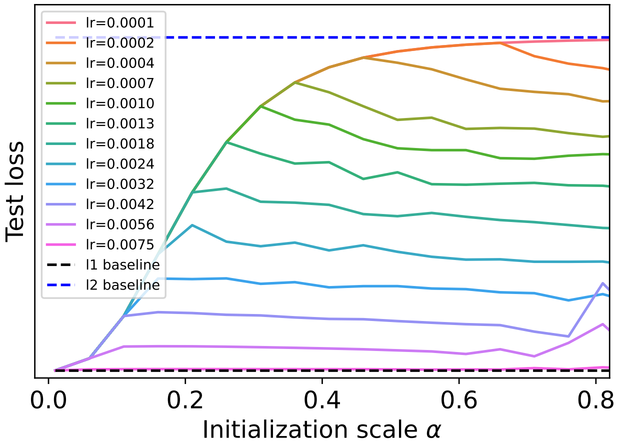


Figure 1. The test loss of GD solution vs. the initialization scale α in the sparse regression problem described in Section 7. We observe that for small step size, the test loss transitions from the ℓ_1 baseline to the ℓ_2 baseline as the initialization scale α increases, as expected from (Woodworth et al., 2020). However, we see that using larger step sizes reduces the error significantly. In fact, **for large step size the test loss is close to the ℓ_1 baseline regardless of the initialization**.

2. Preliminaries: Setup and Minima stability

Notations. For vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, we denote by $\boldsymbol{u}^k$ the element-wise kth power, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ as their dot product, $\boldsymbol{u} \circ \boldsymbol{v}$ as the element-wise multiplication, $\|\boldsymbol{u}\|$ as the L_2 norm of $\boldsymbol{u}$, $\mathbb{B}_r(\boldsymbol{u}) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{u}\| < r\}$ as the r-ball around some point $\boldsymbol{u}$, and $\mathbb{S}^d = \left\{\boldsymbol{x} \in \mathbb{R}^{d+1} : \|\boldsymbol{x}\| = 1\right\}$ as the d dimensional unit sphere. Finally, given N vectors $\boldsymbol{z}_1, ..., \boldsymbol{z}_N$, we denote the empirical mean as $\hat{\mathbb{E}}\boldsymbol{z} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{z}_n$.

Given a dataset of N samples $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N) \in \mathbb{R}^{d \times N}$ with corresponding labels $\boldsymbol{Y} = (y_1, ..., y_N)^\top \in \mathbb{R}^N$ and a prediction function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, we consider the problem of minimizing the empirical squared loss

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\hat{\mathbb{E}}\left((y - f_{\boldsymbol{\theta}}(\boldsymbol{x}))^2\right) = \frac{1}{2N}\sum_{n=1}^{N}\left(y_n - f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)\right)^2 , \tag{1}$$

using GD with step-size η:

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \nabla \mathcal{L}(\boldsymbol{\theta}(t)) \ . \tag{2}$$

We assume that the problem is over-determined, as is often the case when training neural networks, meaning there are multiple global minima that minimize the empirical loss. However, not all minima are accessible by GD with a particular step size (Nar & Sastry, 2018).

We will use Lyapunov stablity (Vidyasagar, 2002; Sastry, 1999) to determine which minima are stable, and thus can

[1] Which is more relevant for ReLU nets, see Remark 4.3 below.

be obtained by GD algorithm with step size η, and which are not. From this point, we refer to global minima of the loss function as solutions.

Definition 2.1. We say that $\boldsymbol{\theta}^*$ is a solution if it satisfies $\mathcal{L}(\boldsymbol{\theta}) = 0$. This implies $\forall n \in [N] : \ f_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = y_n$ and $\nabla\mathcal{L}(\boldsymbol{\theta}^*) = 0$.

We recall the definition of stability in the sense of Lyapunov (Vidyasagar, 2002; Sastry, 1999).

Definition 2.2 (Stability in the sense of Lyapunov)**.** Consider GD iterate $\boldsymbol{\theta}(t)$. A solution $\boldsymbol{\theta}^*$ is said to be *Lyapunov stable* if $\forall\epsilon > 0, \exists\delta > 0$ so that for any $\boldsymbol{\theta}(0) \in \mathbb{B}_\delta(\boldsymbol{\theta}^*)$:

$$\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\| < \epsilon .$$

In words, a solution is Lyapunov stable if once the iterate is sufficiently close to the solution (at time 0, without loss of generality), it always stays close to the solution.

In the following derivations, we will use a well known connection between stability and the eigenvalues of the Hessian (Vidyasagar, 2002; Bof et al., 2018):

Lemma 2.3. *If $\mathcal{L}$ is $\mathcal{C}^1$ and $\boldsymbol{\theta}^*$ is a Lyapunov stable solution, then*

$$\lambda_{\max}\left(\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)\right) \le \frac{2}{\eta} . \tag{3}$$

Note that the assumption that the loss is $\mathcal{C}^1$, i.e., continuously differentiable, is satisfied for all the models we examine in the paper. Throughout the paper we will use $\lambda_{\max} \triangleq \lambda_{\max}\left(\nabla^2\mathcal{L}(\boldsymbol{\theta}^*)\right)$ and $\min\lambda_{\max}$ to denote the minimal $\lambda_{\max}$ achievable by a solution. This Lemma states that using the maximal step size which allows convergence, i.e., $\eta = 2/\min\lambda_{\max}$, we effectively minimize the Hessian maximal eigenvalue. Interestingly, this lemma is also a necessary condition for stability for SGD (not only GD), and one can use this to prove all our results below also for SGD. But, for simplicity, we will focus on GD.

3. Background: Kernel and Rich Regimes in Linear Diagonal Models

In this work we mainly focus on a depth two diagonal linear network,

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left\langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2, \boldsymbol{x}\right\rangle = \langle\boldsymbol{\beta}, \boldsymbol{x}\rangle . \tag{4}$$

This 2-positive homogeneous model was analyzed in several previous works (Woodworth et al., 2020; Gissin et al., 2020; Moroshko et al., 2020; Pesme et al., 2021), and is referred to as squared regression model[2]. Despite its simplicity, previous works demonstrated that the squared regression model exhibits non-trivial kernel and rich behaviours.

Kernel regime. Any prediction function $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \triangleq f(\boldsymbol{\theta}, \boldsymbol{x})$ can be locally approximated around the weights at the initialization, i.e., $\boldsymbol{\theta}(0)$, using

$$f(\boldsymbol{\theta}, \boldsymbol{x}) \approx f(\boldsymbol{\theta}(0), \boldsymbol{x}) + \langle\nabla_\theta f(\boldsymbol{\theta}(0), \boldsymbol{x}), \boldsymbol{\theta} - \boldsymbol{\theta}(0)\rangle . \tag{5}$$

Thus, if the gradients $\nabla_\theta f(\boldsymbol{\theta}, \boldsymbol{x})$ do not change too much during training, then the model $f(\boldsymbol{\theta}, \boldsymbol{x})$ behaves like a kernelized linear predictor

$$\tilde{f}(\boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{\theta}(0), \boldsymbol{x}) + \langle\phi(\boldsymbol{x}), \boldsymbol{\theta} - \boldsymbol{\theta}(0)\rangle , \tag{6}$$

with feature map $\phi(\boldsymbol{x}) = \nabla_\theta f(\boldsymbol{\theta}(0), \boldsymbol{x})$ that corresponds to the Tangent Kernel at initialization $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle\nabla_\theta f(\boldsymbol{\theta}(0), \boldsymbol{x}_1), \nabla_\theta f(\boldsymbol{\theta}(0), \boldsymbol{x}_2)\rangle$ (Jacot et al., 2018; Du et al., 2019). This case was referred to in previous works as the ”kernel regime”. Training this linear model using Gradient Flow (GF) leads to implicit regularization that corresponds to the RKHS norm associated with the kernel. Chizat et al. (2018) demonstrated empirically that we need to escape the kernel regime to obtain good generalization. Specifically, for the squared regression model, the kernel regime corresponds to ℓ_2 regularization (Chizat et al., 2018; Woodworth et al., 2020).

Rich regime. In contrast, previous works showed different implicit biases which cannot be captured by a kernel (Gunasekar et al., 2018c). This regime seems to be more relevant for understanding neural networks practical success. Specifically, for the squared regression model, the rich regime corresponds to sparsity-inducing ℓ_1 regularization (Woodworth et al., 2020). Such ℓ_1 bias can be thought of as doing feature selection, which is a crucial component of representation learning. In representation learning we would like to select from a continuum of possible features. To see how these are related, consider a matrix factorization transfer learning model (Amit et al.; Ando & Zhang; Argyriou et al.), where spectral sparsity is used to learn new features that are linear combinations of input features. Such spectral sparsity can also be induced as the implicit bias in a non-diagonal extension of the squared parameterization (diagonal linear net) model (Gunasekar et al., 2018a; Li et al., 2019a).

For the squared regression model, previous works (Chizat et al., 2018; Gunasekar et al., 2018c; Woodworth et al., 2020), identified the initialization scale α as a dominant factor in the transition between the kernel and rich regimes:

- For small initialization scale, i.e. $\alpha \to 0$, gradient flow

[2] It is ”squared” since the weights are shared (i.e., matched) between the layers. Other works (Azulay et al., 2021; Zhao et al., 2022; Vaskevicius et al., 2019) also extended the analysis to cases where the different layer have different (i.e., non-shared) weights. We show in appendix B how to extend our results to this model with non-shared weights.

induces ℓ_1 regularization on the predictor $\boldsymbol{\beta}$.

- For large initialization scale, i.e. $\alpha \to \infty$, gradient flow induces weighted ℓ_2 regularization on the predictor $\boldsymbol{\beta}$, where the weighting depend on the initialization shape.
- For intermediate initialization scale α, gradient flow induces some sort of interpolation (a hypentropy regularization) between the ℓ_1 and ℓ_2 regularization.

4. Stable Minima Correspond to Predictors with a Bounded Weighted ℓ_1 norm

To the best of our knowledge, all previous works that studied the implicit bias of this model only focused on the case of small or infinitesimal step size, i.e., GF. However, as seen in Figure 1, the step size seems to be an important factor that can help the model escape the kernel regime. Our next lemma connects $\lambda_{\max}$ to a weighted ℓ_1 norm of the model squared weights. Combining this key lemma with Lemma 2.3 leads to Theorem 4.2, which shows how the step size affects which solutions are accessible by GD. In addition, we use Lemma 4.1 to obtain Lemma 5.1, a crucial step in obtaining the generalization result given in Theorem 5.2.

Lemma 4.1. *Let $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{2d}$ be a solution of the squared loss (Eq. (1)). Then,*

$$\left\|\left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2\right\|_1 \leq \frac{\lambda_{\max}}{4} \leq \left\|\left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ \hat{\mathbb{E}}\boldsymbol{x}^2\right\|_1 . \tag{7}$$

Combining the lower bound above with Lemma 2.3 entails

Theorem 4.2. *Let $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{2d}$ be a Lyapunov stable solution of GD with step size η. Then,*

$$\left\|\boldsymbol{\beta} \circ 4\left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2\right\|_1 \leq \left\|\left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ 4\left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2\right\|_1 \leq \frac{2}{\eta} . \tag{8}$$

From this theorem, we see that for non-centered data ($\hat{\mathbb{E}}\boldsymbol{x} \neq 0$), larger step sizes correspond to solutions with lower weighted ℓ_1 norm. That is, larger step sizes result in smaller weights, when the empirical mean is non-zero.

Remark 4.3 (non-zero mean). The analysis in our paper relies on a simplified model for a neural network. We are interested in the transition between the kernel and rich regimes. Therefore, the relevant “real” kernel is the neural network’s kernel, which is a sum of the kernels from all the layers. While the input layer might be centralized, the hidden layers are typically not centralized due to their activation functions (e.g., ReLU, which is not centralized). Thus, the relevant features are typically non-centralized.

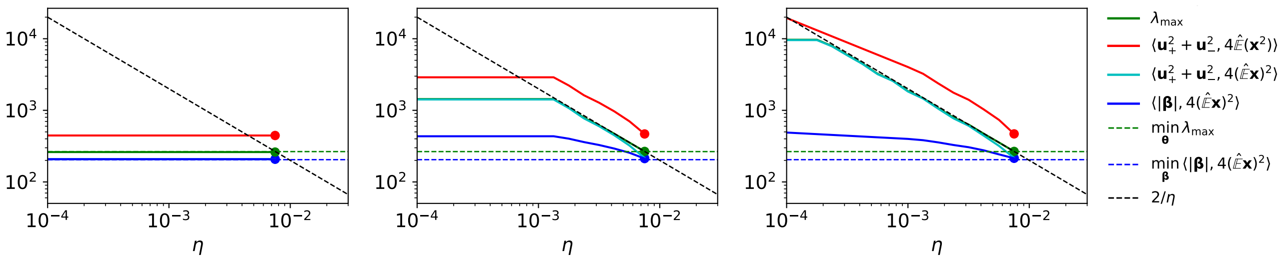
Stability bound visualization. For the same sparse regression problem as in Figure 1 (See section 7 for implementation details), Figure 2 illustrates how the stability measure $\left\|\left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ 4(\hat{\mathbb{E}}\boldsymbol{x})^2\right\|_1$ defined in Eq. (8) changes with step size and initialization. Specifically, for a given initialization scale, we observe that the stability measure does not change with the step size until reaching to the edge of the stability region that corresponds to the line $\frac{2}{\eta}$ (dashed black line). Then, when reaching the border of the stable region, the stability measure starts to decrease in order to satisfy the bound in Theorem 4.2. The figure demonstrates the tightness of our bound. Specifically, note that the stability measure (the light blue line) almost coincides with the maximal eigenvalue of the loss Hessian matrix for any step size. In addition, note that once reaching the stable region border, the maximal eigenvalue of the Hessian, and with it, our stability measure, always remains on the border, meaning that GD converges at the edge of stability. This is a known phenomenon discussed in (Cohen et al., 2021). Additionally, we observe that our stability measure increases with the initialization scale α. Thus, for a given step size we expect that solutions with large weighted ℓ_1 norm will lose stability as the initialization scale increase. Finally, we see that when increasing the step size, GD effectively minimizes the maximal eigenvalue of the loss Hessian, i.e., $\lambda_{\max}\left(\nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)\right)$ and with it the weighted ℓ_1 norm of the obtained solution and corresponding predictor (light blue and blue lines, respectively).

5. When do Large Step Sizes Exit the Kernel Regime and Improve Generalization?

So far we saw the relationship between stable solutions with step size η and their *empirically weighted* ℓ_1-norm $\left\|\boldsymbol{\beta} \circ 4(\hat{\mathbb{E}}\boldsymbol{x})^2\right\|_1$. For the minimal $\lambda_{\max}$, both bounds described in Lemma 4.1 can be translated to function space, i.e., to the weighted ℓ_1 norm of the predictor $\boldsymbol{\beta}$ (Lemma D.2 in the appendix). This is an essential step for our study of generalization, as it is what allows us to connect minimizing $\lambda_{\max}$ with the minimal ℓ_1 norm interpolator. This connection relies on the concentration of $\left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2$ and $\hat{\mathbb{E}}\boldsymbol{x}^2$.

Particularly, when sufficiently many samples are drawn and the coordinates of the data are identically distributed, we may rewrite $\left\|\boldsymbol{\beta} \circ \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2\right\|_1$ and $\left\|\boldsymbol{\beta} \circ \hat{\mathbb{E}}\boldsymbol{x}^2\right\|_1$ as $\gamma_1 \left\|\boldsymbol{\beta}\right\|_1$ and $\gamma_2 \left\|\boldsymbol{\beta}\right\|_1$, respectivelym where γ_1 and γ_2 are distribution-dependent factors. Thus, we can show that the predictor borne of minimizing $\lambda_{\max}$ yields a constant factor multiplicative approximation of the ℓ_1 norm of the predictor with optimal ℓ_1 norm:

Lemma 5.1. *Suppose the data $\boldsymbol{X} \in \mathbb{R}^{d \times N}$ is drawn independently and identically (i.i.d.) with mean μ and such that $X - \mu$ is sub-Gaussian (per the Definition in Appendix D)*



(a) $\alpha = 0.01$ (b) $\alpha = 0.26$ (c) $\alpha = 0.71$

Figure 2. This figure illustrates the stability bounds given in Lemma 4.1 and Theorem 4.2. The figure includes all convergent step sizes. Each plot is for a fixed initialization scale α and shows how various quantities vary with step size: the empirically-calculated maximum eigenvalue of the Hessian (green solid line, $\lambda_{\max}$), empirical value of the weighted norm of the parameters $\langle \boldsymbol{u}_+^2 + \boldsymbol{u}_-^2, 4\hat{\mathbb{E}}(\boldsymbol{x}^2)\rangle$ ($\lambda_{\max}$ upper bound, red solid line), empirical value of the weighted norm of the parameters $\langle \boldsymbol{u}_+^2 + \boldsymbol{u}_-^2, 4(\hat{\mathbb{E}}\boldsymbol{x})^2\rangle$ ($\lambda_{\max}$ lower bound termed stability measure, light blue solid line), which is an upper bound on $\left\|\boldsymbol{\beta} \circ 4(\hat{\mathbb{E}}\boldsymbol{x})^2\right\|_1$ (blue solid line), the minimum $\lambda_{\max}$ achievable by a solution of this problem (green dotted line), the minimum stability measure achievable by a solution (blue dotted line), and twice the reciprocal of the step size (black dotted line). Note the largest step size that converged (green dot) is exactly in the intersection of the dotted green and black lines, i.e., the largest stable step size $2/\min\lambda_{\max}$. There are several things of interest in this figure: (1) the empirical quantities $\lambda_{\max}$ and $\langle \boldsymbol{u}_+^2 + \boldsymbol{u}_-^2, 4(\hat{\mathbb{E}}\boldsymbol{x})^2\rangle$ track each other very closely, indicating that our proposed stability measure is indeed a good proxy for measuring the stability via the maximum eigenvalue of the Hessian. (2) For sufficiently large step size η there are regions where the $\langle \boldsymbol{u}_+^2 + \boldsymbol{u}_-^2, 4(\hat{\mathbb{E}}\boldsymbol{x})^2\rangle$ lower bound appears tight with the $2/\eta$ line (these regions are wide especially for large α). (3) For the maximal step size, the bounds $\langle \boldsymbol{u}_+^2 + \boldsymbol{u}_-^2, 4(\hat{\mathbb{E}}\boldsymbol{x})^2\rangle$ and $\left\|\boldsymbol{\beta} \circ 4(\hat{\mathbb{E}}\boldsymbol{x})^2\right\|_1$ become tight and reach the minimal weighted ℓ_1 norm predictor. Every pairwise relationship between the largest eigenvalue of the Hessian, the step size, and the weighted norm stability measure is interesting. While the relationship between the first two was known previously, these plots indicate that there is also a relationship between the step size and the stability measure, implying a connection between the largest eigenvalue of the Hessian and the stability measure.

with parameter σ . Define $\tilde{\boldsymbol{u}}_+$, $\tilde{\boldsymbol{u}}_-$ such that:

$$\tilde{\boldsymbol{u}}_+ , \tilde{\boldsymbol{u}}_- := \underset{\langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y}}{\arg\min} \lambda_{\max}(\boldsymbol{u}_+, \boldsymbol{u}_-) , \tag{9}$$

and let $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{u}}_+^2 - \tilde{\boldsymbol{u}}_-^2$. Then, we have that, provided $N = \Omega\left(\max\left\{\frac{\sigma^2}{\sigma^2+\mu^2}\log(d/\delta), \frac{\sigma\sqrt{\log(d/\delta)}}{\mu}\right\}^2\right)$, with probability at least $1-\delta$,

$$\left\|\tilde{\boldsymbol{\beta}}\right\|_1 \le \left(1 + \frac{\sigma^2}{\mu^2}\right) \cdot 1.1 \inf_{\langle \boldsymbol{\beta}, \boldsymbol{X}\rangle = \boldsymbol{y}} \|\boldsymbol{\beta}\|_1 . \tag{10}$$

Note that the tightness of the approximation depends on the ratio $\frac{\sigma}{\mu}$. In particular, if $\mu = 0$ then the predictor that corresponds to minimizing $\lambda_{\max}$ does not approximately minimize the ℓ_1 norm. This, again, emphasizes the significance of non-zero mean data.

Establishing this tight relationship between the relevant weighted ℓ_1 norms of the predictor $\boldsymbol{\beta}$ and the ℓ_1 norm of the ℓ_1 norm minimizing predictor allows us to equate minimizing $\lambda_{\max}$ to minimizing, up to a constant factor, the ℓ_1 norm of the predictor. From there, we extend existing Radaemacher complexity analyses of the generalization of bounded ℓ_1 norm predictors (from (Srebro et al.)) to the Gaussian data setting (Lemma D.6 in the appendix). Putting these pieces together, we get the following generalization result, with proof details in Appendix D.

Theorem 5.2. *Suppose the entries of $\boldsymbol{X}$ are independently and identically drawn from a Gaussian with variance σ^2 and mean $\mu = O(\sigma)$. Let $\boldsymbol{y}$ be the labels produced by a k-sparse linear predictor, i.e., $\boldsymbol{y} = \langle \boldsymbol{\beta}^\star, \boldsymbol{X}\rangle$, for some $\|\boldsymbol{\beta}^\star\|_0 \le k$. For $\tilde{\boldsymbol{\beta}}$ as defined in in Lemma 5.1 and $\zeta = 1 + \frac{\sigma^2}{\mu^2}$, the population loss $\mathcal{L}_{\mathcal{D}}$ obeys, with probability at least $1-\delta$:*

$$\frac{\mathcal{L}_{\mathcal{D}}(\tilde{\boldsymbol{\beta}})}{\mathcal{L}_{\mathcal{D}}(\mathbf{0})} \le \mathcal{O}\left(\frac{\zeta^2 k \operatorname{polylog}(dN/\delta)}{N}\right) . \tag{11}$$

Theorem 5.2 ensures generalization using $N = \tilde{\mathcal{O}}(\zeta^2\, k \operatorname{polylog}(d/\delta))$ samples.[3] The sample complexity is increased by a factor of $\zeta^2 = \left(1 + \sigma^2/\mu^2\right)^2$, which controls how well the weighted ℓ_1 norm approximates the true ℓ_1 norm. When the data mean μ is too small, the guarantee becomes meaningless and the sample test loss explodes, matching the observed behavior (Figure 3). However, for non-centered data, when $\mu = \Theta(\sigma)$, the sample complexity remains $N = \tilde{\mathcal{O}}(k \operatorname{polylog}(d/\delta))$. Learning to within an error that's a fraction of the null error $\mathcal{L}_{\mathcal{D}}(\mathbf{0}) = \mathbb{E}\left(y^2\right)$, as in Theorem 5.2, is the best possible using concentration guarantees, or without exploiting incoherence assumptions (see Foygel & Srebro, 2011; Zhang et al., 2014). This seems disappointing here, since under the i.i.d. and noiseless assumptions we make, exact recovery is possible. However,

[3] We use $f = \tilde{\mathcal{O}}(g)$ to mean that there exist constants a, b, c such that $f \le a + bg\log^c g$.

exact recovery would require exactly minimizing $\|\beta\|_1$, and here we are only relying on a constant factor approximation to the ℓ_1 norm. With such a constant factor approximation, an error scaling with $\mathbb{E}(y^2)$ is unavoidable.

Now that we have shown that minimizing the maximum eigenvalue of the Hessian leads to generalization, let us consider the link between this and running gradient descent with an initialization $\boldsymbol{\theta}_0$ sampled from some distribution $\mathcal{P}_0$.

Assumption 5.3. Given a distribution $\mathcal{P}_0$, from which sample the initialization $\boldsymbol{\theta}_0$, then for any step size η, GD will converge to a non-stable solution with probability zero.

In other words, GD generically does not converge to unstable minima. This common assumption holds empirically and is made in previous works, either implicitly or explicitly (e.g., Nar & Sastry (2018); Mulayoff et al. (2021)). However, existing proofs of this assumption require conditions which do not hold for our setting (e.g., Ahn et al. (2022) requires that η^{-1} is not a eigenvalue of the Hessian, at *all* stationary points), so we leave its proof for future work.

Next, we assume that GD converges to zero training loss.

Assumption 5.4. For any step size η , if there exists an η-stable solution (as in Lemma 2.3 , that is, a Lyapunov stable solution with maximum Hessian eigenvalue less than $2/\eta$), then GD with initialization $\boldsymbol{\theta}_0 \sim \mathcal{P}_0$ and step size η converges with probability one.

Empirically, we observed that this assumption always hold for any reasonable initialization (i.e., when the initialization scale is not extremely large) when using long enough warmup. Again, we leave the proof of this interesting observation for future work.

With both assumptions above, GD converges as long as there exist a stable solution ($\eta \leq \frac{2}{\min \lambda_{\max}}$). Then, applying the previous result, we obtain the following corollary:

Corollary 5.5. *In the same setting as in Lemma 5.1, under Assumptions 5.3 and 5.4, and with initialization $\boldsymbol{\theta}_0 \sim \mathcal{P}_0$, consider running GD with the largest step size η such that gradient descent converges. Recall $\zeta = 1 + \frac{\sigma^2}{\mu^2}$. Then, GD converges to a predictor $\boldsymbol{\beta} = \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2$, such that the population loss $\mathcal{L}_\mathcal{D}$ satisfies, with probability at least $1-\delta$:*

$$\frac{\mathcal{L}_\mathcal{D}(\boldsymbol{\beta})}{\mathcal{L}_\mathcal{D}(\mathbf{0})} \leq \mathcal{O}\left(\frac{\zeta^2 k \operatorname{polylog}(dN/\delta)}{N}\right).$$

Connection to Figure 1. Woodworth et al. (2020) characterized the implicit regularization that leads from the kernel ℓ_2 regularization to the rich ℓ_1 regularization as the initialization scale vanishes, under the assumption of infinitesimal step size (i.e., GF). To demonstrate their theoretical findings, they used a sparse regression problem in which the number of samples N is sufficient to obtain good recovery with ℓ_1 regularization and yet insufficient with ℓ_2 regularization. In this setting, the authors showed empirically that for small initialization scale gradient flow obtains small generalization error and as the initialization increases the generalization degrades. This behaviour is captured in Figure 1 when we use sufficiently small step sizes.

However, Figure 1 also reveals that there is more to the story when going beyond the infinitesimal step size case. Specifically, we observe that increasing the step size improves generalization in the sparse regression problem. In fact, for large step size we obtain small test loss, i.e., rich behaviour, regardless of the initialization scale. From this figure, we identify the step size as a much more prominent factor that can help the iterator escape the kernel regime. Theorem 4.2 along with Corollary 5.5 offers an explanation to this empirical observations. Since the step size inversely bounds the weighted ℓ_1 norm of GD accessible solutions, for any non-centered data we expect that increasing the step size will lead to solutions with smaller weighted ℓ_1 norm, which, per the analysis in the proof of Theorem 5.2 (particularly, Lemma 5.1), corresponds to smaller ℓ_1 norm and thus improves generalization in the sparse regression problem. We expect this behaviour to become more significant when the empirical mean is large.

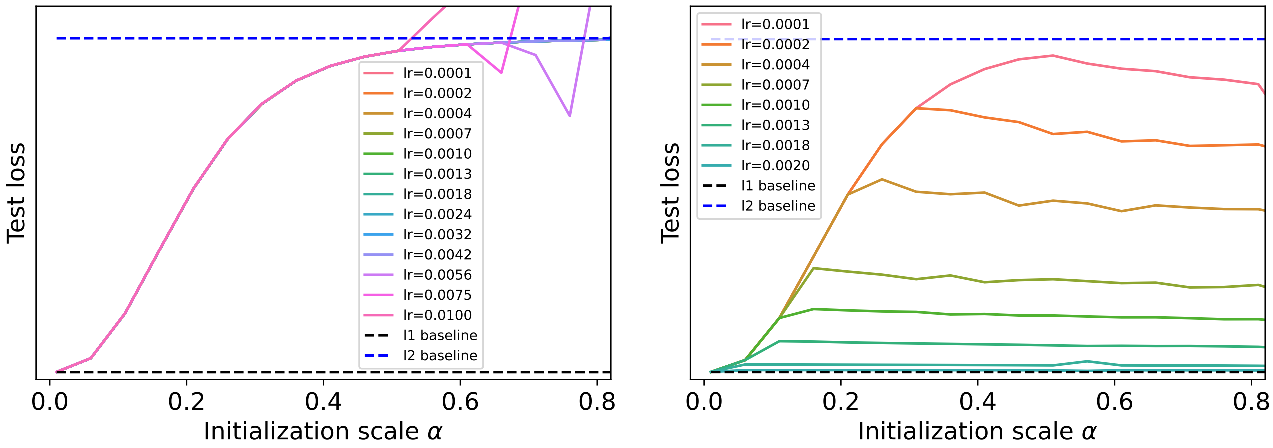
Empirical mean significance. The LHS of our theoretical bound in Eq. (8) increases with the empirical mean of the training set. Consequently, Corollary 5.5 only applies when the distribution expectation μ is non-zero and the generalization bound improves when $|\mu|$ increases. This implies that the phenomenon discussed so far will diminish when the training data is sampled from a centered data distribution, i.e., $\mu = 0$. In this case, we expect the step size will only have a mild effect on the test loss. Figure 3 demonstrates this result. We observe that for $\mu = 0$, most step sizes resulted in the same test loss behaviour. Only the two largest step sizes that converged, resulted in the expected test loss drop, and even then, the drop is mild and does not bring us all the way to the rich ℓ_1 regime. Note that even for $\mu = 0$ we expect the empirical mean to be small yet non-zero. Thus, Theorem 4.2 still gives us valuable insight on this case. For large data μ, we observe that the test loss decrease per step size magnitude is more significant than what we observed in Figure 2.

6. The Effect of Depth on Implicit Bias

So far, we focused on the squared regression model, corresponding to a depth two diagonal model. In this section, we generalize our results to deeper (higher order) linear diagonal models. This will allow us to study how the model depth affect the stability criterion.

Formally, we consider a depth D diagonal model:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left\langle \boldsymbol{u}_+^D - \boldsymbol{u}_-^D, \boldsymbol{x} \right\rangle = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle \ . \tag{12}$$



(a) $\mu = 0$

(b) $\mu = 10$

Figure 3. In this figure, we observe the effect of the expectation of the data (μ) on the test loss profile against the initialization scale α for increasing step sizes. In particular, we observe that when $\mu = 0$ (left), there is no generalization benefit to increasing the step size, since the "rich" regime is only achieved for small initialization scale. On the other hand, for $\mu = 10$ (right), increasing step sizes allows gradient descent to pick solutions that generalize well, indeed approaching the baseline error of the ℓ_1 minimizing solution. These empirical results correspond well to the theoretical results that show that larger means result in better generalization (Corollary 5.5).

Note that D corresponds to the model homogeneity degree. The next theorem generalizes the $\lambda_{\max}$ bounds given in Theorem 4.1 for the $D = 2$ case to any depth diagonal networks.

Lemma 6.1. *Let $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{2d}$ be a solution of the squared loss (Eq. (1)). Then,*

$$\left\| \boldsymbol{\theta}^{*D} \circ \boldsymbol{w}_1 \right\|_p^p \leq \lambda_{\max} \leq \left\| \boldsymbol{\theta}^{*D} \circ \boldsymbol{w}_2 \right\|_p^p , \tag{13}$$

where $p = \frac{2(D-1)}{D}$ and

$$\boldsymbol{w}_1 = D^2 \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^{\frac{D}{D-1}} , \ \boldsymbol{w}_2 = D^2 \begin{pmatrix} \hat{\mathbb{E}}\left(\boldsymbol{x}^2\right) \\ \hat{\mathbb{E}}\left(\boldsymbol{x}^2\right) \end{pmatrix}^{\frac{D}{2(D-1)}} , \tag{14}$$

Similarly to lemma 4.1 for the $D = 2$ case, this lemma draws a connection between minimizing $\lambda_{\max}$ and minimizing a weighted ℓ_p norm of the weights, with $p = \frac{2(D-1)}{D}$. In addition, this lemma enables the generalization of Theorem 4.2 to obtain a stability bound for deeper models.

Theorem 6.2. *Let $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{2d}$ be a Lyapunov stable solution of GD with step size η. Then,*

$$\left\| \begin{pmatrix} \boldsymbol{u}_+^\top \\ \boldsymbol{u}_-^\top \end{pmatrix}^D \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^{\frac{D}{D-1}} \right\|_p^p \leq \frac{2}{D^2 \eta} \tag{15}$$

where $p = \frac{2(D-1)}{D}$. If D is even, or $\forall i : \theta_i^* \geq 0$, then this implies $\left\| \boldsymbol{\beta} \circ \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^{\frac{D}{D-1}} \right\|_p^p \leq \frac{2}{D^2 \eta}$.

From this theorem, we see that large step sizes effectively induce regularization on the weighed ℓ_p norm of the predictor, where $p = \frac{2(D-1)}{D}$. For $D = 2$, this exactly gives us ℓ_1 regularization as discussed above. Note that as D increases we are getting closer to the ℓ_2 norm. Specifically, when $D \to \infty$ we get ℓ_2 regularization.

Connection to previous results. In the classification setting with the exponential loss, (Gunasekar et al., 2018a) characterized the implicit bias for linear convolutional networks—which are equivalent to linear diagonal networks in the frequency domain. Specifically, they showed that GD induces $\ell_{2/D}$ regularization in the frequency domain on the network equivalent linear predictor. This result does not depend on the step size, as long as it is sufficiently small to reach zero training loss. This $\ell_{2/D}$ regularization on the predictor can also be induced explicitly (for any loss) using (vanishing) ℓ_2 regularization on the parameters. Since both Gunasekar et al. (2018a)'s results and our Theorem 6.2 apply to linear diagonal models, it is interesting to compare the two. Gunasekar et al. (2018a) showed that the implicit regularization with exponential loss, or small explicit ℓ_2 regularization, cause the predictor bias to change from ℓ_1 when $D = 2$ to ℓ_0 as $D \to \infty$, implying that deeper models achieve higher sparsity. In contrast, our result states that the inductive bias changes from ℓ_1 for $D = 2$ to ℓ_2 for $D \to \infty$, implying only the shallow model $D = 2$ model has a sparse penalty, while deeper models are not biased to sparse solutions.

Empirical results. From our theoretical results, we expect large step sizes to bias the solution towards small weighted

ℓ_p norm with $p=\frac{2(D-1)}{D}$. This regularization should somewhat improve the generalisation in comparison to the ℓ_2 predictor, yet not as effectively as for the case $D=2$ where we effectively obtained ℓ_1 regularization. Figure 4 shows the test loss vs. initialization scale for different depths and step sizes. As expected from (Woodworth et al., 2020), for small step size we get rich ℓ_1 behaviour when the initialization scale is small and kernel ℓ_2 behaviour for large initialization scale. In addition, we observe that the test loss decreased as the step size increased, i.e. the implicit regularization induced by stability still improves the generalization. However, for deep networks, the step size does not take us all the way to the ℓ_1 rich regime. This is due to ℓ_p being less efficient regularization for this type of sparse problems and also because $\lambda_{\max}$ grows with depth and so the maximal step size that allows convergence is more limited. Interestingly, we observe the generalization vs. optimization trade-off discussed in (Woodworth et al., 2020). For GF, good generalization requires small initialization. However, this is problematic from an optimization perspective since $\boldsymbol{\beta}=0$ is a saddle point and thus using small initialization scale is likely to increase the time it takes to escape the vicinity of zero. This corresponds to the results observed in Figure 4, where we see that for small initialization scale the networks did not converge. This emphasizes the crucial role of the step size as a more realistic hyper-parameter that can take us from the kernel regime to the rich regime.

7. Numerical Simulations Details

Sparse regression problem. To understand the step size influence, we consider a simple sparse regression problem, similar to problem used in (Woodworth et al., 2020). Specifically, we define $\boldsymbol{x}_1,...,\boldsymbol{x}_N \sim \mathcal{N}\left(\boldsymbol{\mu},\sigma^2\boldsymbol{I}\right)$ and $y_n\sim\mathcal{N}\left(\langle\boldsymbol{\beta}^*,\boldsymbol{x}_n\rangle,0.01\right)$ where $\boldsymbol{\beta}^*$ is r^* sparse with non-zero entries equal to $1/\sqrt{r^*}$. Unless explicitly mentioned, we used $\mu=5\cdot\mathbf{1}$ and $\sigma^2=5$, $r^*=5$, $d=100$. The number of data samples was chosen so there will be a sufficient number of training samples to generalize well in the rich ℓ_1 regime ($N=\Omega\left(r^*\log d\right)$), but not in the kernel ℓ_2 regime (good generalization in that regime requires $N=\Omega\left(d\right)$). Thus, we chose $N=50$ which satisfies these requirements.

Learning Procedure We study the diagonal deep model, i.e., parameterization $\boldsymbol{\beta}=\boldsymbol{u}_+^D-\boldsymbol{u}_-^D$ for different depths D. We initialize $\boldsymbol{u}_+=\boldsymbol{u}_-=\alpha\cdot\mathbf{1}$ for chosen scale α. We used this initialization shape for simplicity and since it was used in (Woodworth et al., 2020). However, we observe empirically the same behaviour regardless of the initialization shape (see Appendix E). We use GD to minimize the loss until convergence, i.e. achieving (extremely close to) 0 training error, so that we find an interpolating solution. We use warmup procedure to avoid exploding gradients at the beginning of training when using large step sizes with large initialization scales. In other words, we start the training with low learning rate and linearly increase the step size until reaching the desired step size η. Note that even without warmup, we see good generalization at initialization moderate scales, but warmup seems to be required to be able to converge at larger scales. It is an interesting question for future work to understand why warmup is needed, as warmup is also a popular standard practice in many deep learning models.

8. Minima Stability Results Proof Outline

In this section, we explain the proof idea for obtaining the lower bound in Lemma 4.1. The general scheme closely resembles the ideas in (Mulayoff et al., 2021). From Lemma 2.3, we have that any stable solution $\boldsymbol{\theta}^*$ must satisfy

$$\lambda_{\max}\left(\nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)\right)\leq\frac{2}{\eta}\,. \tag{16}$$

Thus, our goal is to lower bound $\lambda_{\max}\left(\nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)\right)$. We first calculate the loss gradient defined in Eq. (1):

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}\left(\boldsymbol{\theta}\right)=\frac{1}{N}\sum_{n=1}^{N}\left(f\left(\boldsymbol{x}_n\right)-y_n\right)\nabla_{\boldsymbol{\theta}}f\left(\boldsymbol{x}_n\right)\,, \tag{17}$$

and the Hessian matrix

$$\begin{aligned}\nabla^2_{\boldsymbol{\theta}}\mathcal{L}\left(\boldsymbol{\theta}\right)=&\frac{1}{N}\sum_{n}\nabla_{\boldsymbol{\theta}}f\left(\boldsymbol{x}_n\right)\nabla_{\boldsymbol{\theta}}f\left(\boldsymbol{x}_n\right)^\top\\&+\frac{1}{N}\sum_{n}\left(f\left(\boldsymbol{x}_n\right)-y_n\right)\nabla^2_{\boldsymbol{\theta}}f\left(\boldsymbol{x}_n\right)\,.\end{aligned} \tag{18}$$

Since $\boldsymbol{\theta}^*$ is a solution, $\forall n\in[N]:\ f_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}_n\right)=y_n$ and thus we can write the Hessian matrix as

$$\nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)=\frac{1}{N}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\,, \tag{19}$$

where we define

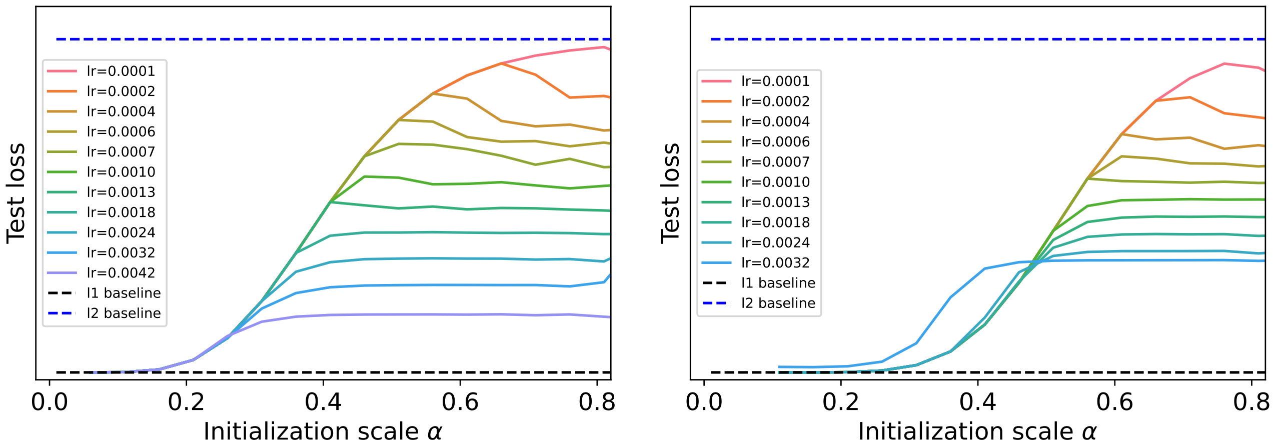
$$\boldsymbol{\Phi}=\begin{bmatrix}\nabla_{\boldsymbol{\theta}^*}f\left(\boldsymbol{x}_1\right) & \ldots & \nabla_{\boldsymbol{\theta}^*}f\left(\boldsymbol{x}_N\right)\end{bmatrix}\in\mathbb{R}^{d\times N}\,. \tag{20}$$

Thus, $\lambda_{\max}\left(\nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)\right)$ can be expressed as

$$\begin{aligned}\lambda_{\max}\left(\nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)\right)&=\max_{\boldsymbol{b}\in\mathbb{S}^{d-1}}\boldsymbol{b}^\top\nabla^2\mathcal{L}\left(\boldsymbol{\theta}^*\right)\boldsymbol{b}\\&=\max_{\boldsymbol{b}\in\mathbb{S}^{d-1}}\frac{1}{N}\left\|\boldsymbol{\Phi}^\top\boldsymbol{b}\right\|^2=\max_{\boldsymbol{p}\in\mathbb{S}^{N-1}}\frac{1}{N}\left\|\boldsymbol{\Phi}\boldsymbol{p}\right\|^2\,,\end{aligned} \tag{21}$$

and we can lower bound the RHS of the last equation taking any[4] $\boldsymbol{p}\in\mathbb{S}^{N-1}$. In Appendix A and C, we analyze the RHS of Eq. (21) for each one of the models discussed in this paper, and prove a lower bound that corresponds to the lower bound in the theorem associated with that model.

[4] Note that some choices of $\boldsymbol{p}$ might result in a loose bound. However, our empirical observations demonstrate that our bound is tight.



(a) $D = 3$

(b) $D = 4$

Figure 4. In this figure, we explore the effect of depth on the test loss of models trained with different step sizes starting at different initialization scales. The plot includes all convergent step sizes. We note that the maximal step size that converged satisfies $\eta \leq \frac{2}{\min \lambda_{\max}}$ with near equality, as expected (For $D = 3$: $\frac{2}{\min \lambda_{\max}} = 0.0061$, and for $D = 4$: $\frac{2}{\min \lambda_{\max}} = 0.0034$). Like in the $D = 2$ case, we see improvements in generalization with larger step sizes, even for large initialization scales. However, we do not approach the loss of ℓ_1 minimizer as closely as we did in the $D = 2$ case. Of course, this is not unexpected, because an inspection of Theorem 6.2 suggests that even for $D = 3$, the norm that is approximately minimized is $p = 4/3$, which is already quite far away from sparsity, and for large D , this norm approaches $p = 2$, so gradient descent would remain squarely in the kernel regime.

9. Conclusion

Deep models are often trained with large step sizes, since this was observed to be beneficial for generalization (e.g., Hoffer et al. (2017); Li et al. (2019b)). Therefore, we believe large step sizes are the more relevant mechanism for escaping the kernel regime and reaching the rich regime. This is in contrast to other mechanisms previously considered for this purpose, such as small initialization size (Chizat et al., 2018; Woodworth et al., 2020), which are not typically used in practice since they can hurt the optimization speed.

With the aim of understanding this behavior, in this paper we studied the effect of non-zero step sizes on the implicit bias of gradient descent in the squared regression model. Our work identifies and analyzes a setting in which gradient descent with large stable step size achieves a “rich” implicit bias regime. Specifically, through the lens of dynamical stability, we show that stable solutions satisfy a stability condition on the weighted ℓ_1 norm in function space. Moreover, under additional assumptions, we show that by taking the maximal stable step size we guarantee generalization in a sparse regression setting. That is, to the best of our knowledge, we provide the first theoretical example that shows that large step size can lead to good generalization. The step size influences on the implicit bias was previously studied in other context, e.g., (Nakkiran, 2020; You et al., 2020; Barrett & Dherin, 2021; Smith et al., 2021), yet not directly linked (theoretically) to generalization. The only exception is (Ma & Ying, 2021). Specifically, Ma & Ying (2021) provided sufficient and necessary conditions for dynamical stability of SGD based on the expectation of high-order moments of the gradient noise. They use these condition to prove a generalization result. However, their generalization bound depends on the norm of the first layer of the network, which in general can be arbitrarily large.

Interestingly, our results show that the benefits of using large step sizes only apply for non-centered inputs, which are quite common in deep learning (e.g., due to the non-negativity of ReLU activations). Lastly, we extend these results to deep diagonal linear networks, where we see that large step sizes the predictor becomes less sparse with depth— which is very different from how depth affected the implicit bias in previous works.

Acknowledgements

The research of DS was supported by the Israel Science Foundation (grant No. 1308/18), and by the Israel Innovation Authority (the Avatar Consortium). KR and NS were partially supported by the NSF-Simons Collaboration on the Mathematics of Deep Learning and by an NSF IIS award.

References

Ahn, K., Zhang, J., and Sra, S. Understanding the unstable convergence of gradient descent. *arXiv preprint arXiv:2204.01050*, 2022.

Amit, Y., Fink, M., Srebro, N., and Ullman, S. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pp. 17–24. Association for Computing Machinery. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273499.

Ando, R. K. and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. 6(61):1817–1853. ISSN 1533-7928.

Argyriou, A., Evgeniou, T., and Pontil, M. Multi-task feature learning.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization, 2019.

Azulay, S., Moroshko, E., Nacson, M. S., Woodworth, B., Srebro, N., Globerson, A., and Soudry, D. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent, 2021.

Barrett, D. and Dherin, B. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process, 2020.

Bof, N., Carli, R., and Schenato, L. Lyapunov theory for discrete time systems. *arXiv preprint arXiv:1809.05289*, 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. *ICLR*, 2021.

Damian, A., Ma, T., and Lee, J. D. Label noise sgd provably prefers flat global minimizers, 2021.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Foygel, R. and Srebro, N. Fast-rate and optimistic-rate error bounds for l1-regularized regression. 2011.

Gissin, D., Shalev-Shwartz, S., and Daniely, A. The implicit bias of depth: How incremental learning drives generalization. *International Conference on Learning Representations (ICLR)*, 2020.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018a.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018b. ISSN: 2640-3498.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018c.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *NeuriPS*, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 2018.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations, 2019a.

Li, Y., Wei, C., and Ma, T. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *NeurIPS*, 2019b.

Ma, C. and Ying, L. On linear stability of SGD and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Moroshko, E., Gunasekar, S., Woodworth, B., Lee, J. D., Srebro, N., and Soudry, D. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Mulayoff, R., Michaeli, T., and Soudry, D. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34, 2021.

Nakkiran, P. Learning rate annealing can provably help generalization, even for convex problems. *arXiv preprint arXiv:2005.07360*, 2020.

Nar, K. and Sastry, S. S. Step size matters in deep learning. *arXiv preprint arXiv:1805.08890*, 2018.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning, 2015.

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *arXiv preprint arXiv:2106.09524*, 2021.

Sastry, S. *Nonlinear systems: analysis, stability, and control*, volume 10. Springer Science & Business Media, 1999.

Smith, S. L., Dherin, B., Barrett, D., and De, S. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low-noise and fast rates. pp. 21.

Vaskevicius, T., Kanade, V., and Rebeschini, P. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Vidyasagar, M. *Nonlinear systems analysis*. SIAM, 2002.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

You, C., Zhu, Z., Qu, Q., and Ma, Y. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. *Advances in Neural Information Processing Systems*, 33:17733–17744, 2020.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. pp. 28, 2014.

Zhao, P., Yang, Y., and He, Q.-C. High-dimensional linear regression via implicit regularization. *Biometrika*, 2022.

A. Proof of Theorems 4.2 and 6.2 and Lemmas 4.1 and 6.1

A.1. Proof of Theorems 4.2 and 6.2

Note that Theorem 4.2 is just a special case of Theorem 6.2, and thus it's sufficient to prove Theorem 6.2. To prove Theorem 6.2 we need to show that any linearly stable solution $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{2d}$ of GD with step size η, satisfies

$$\left\| \begin{pmatrix} \boldsymbol{u}_+^\top \\ \boldsymbol{u}_-^\top \end{pmatrix}^D \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^{\frac{D}{D-1}} \right\|_p^p \leq \frac{2}{D^2\eta} \tag{22}$$

where $\hat{\mathbb{E}}\boldsymbol{x} = \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n$ is the empirical mean and $p = \frac{2(D-1)}{D}$. In addition, we need to show that if D is even, or $\forall i : \theta_i^* \geq 0$, then $\left\| \boldsymbol{\beta} \circ \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^{\frac{D}{D-1}} \right\|_p^p \leq \frac{2}{D^2\eta}$.

Proof. Let $\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{2d}$ be a linearly stable solution of GD with step size η. Eq. 22 is a direct result of combining Lemma 4.1 with Lemma 2.3. Specifically, from Lemma 2.3 and Lemma 4.1 we have that

$$D^2 \left\| \begin{pmatrix} \boldsymbol{u}_+^\top \\ \boldsymbol{u}_-^\top \end{pmatrix}^D \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^{\frac{D}{D-1}} \right\|_p^p \leq \lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\boldsymbol{\theta}\right)\right) \leq \frac{2}{\eta}\,. \tag{23}$$

If, in addition, D is even, or $\forall i : \theta_i^* \geq 0$, then

$$\left\| \begin{pmatrix} \boldsymbol{u}_+^\top \\ \boldsymbol{u}_-^\top \end{pmatrix}^D \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^{\frac{D}{D-1}} \right\|_p^p = \left(\left(\boldsymbol{u}_+^D\right)^{\frac{2(D-1)}{D}} + \left(\boldsymbol{u}_-^D\right)^{\frac{2(D-1)}{D}} \right)^\top \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \geq \left(|\boldsymbol{\beta}|^{\frac{2(D-1)}{D}} \right)^\top \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 = \left\| \boldsymbol{\beta} \circ \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^{\frac{D}{D-1}} \right\|_p^p \tag{24}$$

where in the inequality we used the fact that $\left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \geq 0$ and $\left(\left(\boldsymbol{u}_+^D\right)^{\frac{2(D-1)}{D}} + \left(\boldsymbol{u}_-^D\right)^{\frac{2(D-1)}{D}} \right) \geq |\boldsymbol{\beta}|^{\frac{2(D-1)}{D}}$ (element-wise). To prove the last inequality, $\forall i = 1, ..., d$ we define $a_i \triangleq \boldsymbol{u}_{+,i}^D \geq 0$ and $b_i \triangleq \boldsymbol{u}_{-,i}^D \geq 0$ and apply the following technical lemma.

Lemma A.1. *Let $a, b \in \mathbb{R}_+$ and $p \geq 1$. Then, $a^p + b^p \geq |a-b|^p$.*

Lemma A.1 is proved in appendix section A.3. Combining Eqs. (23) and (24) we obtain

$$\left\| \boldsymbol{\beta} \circ \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^{\frac{D}{D-1}} \right\|_p^p \leq \frac{2}{D^2\eta}\,. \tag{25}$$

This completes our proof. □

A.2. Proof of Lemmas 4.1 and 6.1

Note that Lemma 4.1 is just a special case of Lemma 6.1, and thus it's sufficent to prove Lemma 6.1.

Proof. Let $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{2d}$ be a solution of the squared loss (Eq. (1)) with the deep diagonal model $f_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \left\langle \boldsymbol{u}_+^D - \boldsymbol{u}_-^D, \boldsymbol{x} \right\rangle = \left\langle \boldsymbol{\theta}^D, \hat{\boldsymbol{x}} \right\rangle$, where we defined $\hat{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x}^\top & -\boldsymbol{x}^\top \end{pmatrix}^\top$. From Eq. (21) we have that

$$\begin{aligned} \lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\boldsymbol{\theta}\right)\right) &= \max_{\boldsymbol{b}\in\mathbb{S}^{d-1}} \boldsymbol{b}^\top \nabla^2 \mathcal{L}\left(\boldsymbol{\theta}\right) \boldsymbol{b} \\ &= \max_{\boldsymbol{b}\in\mathbb{S}^{d-1}} \frac{1}{N} \left\| \boldsymbol{\Phi}^\top \boldsymbol{b} \right\|^2 \\ &= \max_{\boldsymbol{p}\in\mathbb{S}^{N-1}} \frac{1}{N} \left\| \boldsymbol{\Phi}\boldsymbol{p} \right\|^2 \,, \end{aligned} \tag{26}$$

where

$$\boldsymbol{\Phi} = \left(\nabla_{\boldsymbol{\theta}^*} f\left(\boldsymbol{x}_1\right) \quad \ldots \quad \nabla_{\boldsymbol{\theta}^*} f\left(\boldsymbol{x}_N\right)\right) \in \mathbb{R}^{d \times N}\,. \tag{27}$$

For the deep diagonal model

$$\nabla_{\boldsymbol{\theta}^*} f\left(\boldsymbol{x}\right) = D\hat{\boldsymbol{x}} \circ \boldsymbol{\theta}^{D-1} \tag{28}$$

and thus we obtain

$$\boldsymbol{\Phi} = D \cdot \left(\begin{array}{ccc} \hat{\boldsymbol{x}}_1 \circ \boldsymbol{\theta}^{D-1} & \cdots & \hat{\boldsymbol{x}}_N \circ \boldsymbol{\theta}^{D-1} \end{array}\right) \in \mathbb{R}^{2d \times N}\,. \tag{29}$$

Next, we use this result to lower and upper bound $\lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\boldsymbol{\theta}\right)\right)$.

Lower bound:

Using Eqs. (26) and (29), and taking $\boldsymbol{p} = \frac{1}{\sqrt{N}} \mathbf{1}_{N\times 1}$, we obtain:

$$\begin{aligned}
\lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\boldsymbol{\theta}\right)\right) &= \frac{1}{N} \max_{\boldsymbol{p}\in\mathbb{S}^{N-1}} \left\|\boldsymbol{\Phi}\boldsymbol{p}\right\|^2 \\
&\geq \frac{D^2}{N^2} \sum_{i=1}^{d} \left[\sum_n \boldsymbol{u}_{+,i}^{D-1} \boldsymbol{x}_{n,i}\right]^2 + \left[\sum_n \boldsymbol{u}_{-,i}^{D-1} \boldsymbol{x}_{n,i}\right]^2 \\
&= \frac{D^2}{N^2} \sum_{i=1}^{d} \left(\boldsymbol{u}_{+,i}^{2(D-1)} + \boldsymbol{u}_{-,i}^{2(D-1)}\right) \left[\sum_n \boldsymbol{x}_{n,i}\right]^2 \\
&= D^2 \sum_{i=1}^{d} \left(\boldsymbol{u}_{+,i}^{2(D-1)} + \boldsymbol{u}_{-,i}^{2(D-1)}\right) \left(\hat{\mathbb{E}}\boldsymbol{x}_i\right)^2 \\
&= D^2 \left(\boldsymbol{u}_{+}^{2(D-1)} + \boldsymbol{u}_{-}^{2(D-1)}\right)^\top \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2\,,
\end{aligned} \tag{30}$$

where we defined $\hat{\mathbb{E}}\boldsymbol{x} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$. Lastly, note that

$$\left(\boldsymbol{u}_{+}^{2(D-1)} + \boldsymbol{u}_{-}^{2(D-1)}\right)^\top \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 = \sum_{i=1}^{d} \left(\boldsymbol{u}_{+,i}^{2(D-1)} + \boldsymbol{u}_{-,i}^{2(D-1)}\right) \left(\left(\hat{\mathbb{E}}\boldsymbol{x}\right)^{\frac{D}{D-1}}\right)^{\frac{2(D-1)}{D}} = \left\|\boldsymbol{\theta}^D \circ \left(\begin{array}{c} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{array}\right)^{\frac{D}{D-1}}\right\|_p^p\,, \tag{31}$$

where $p = \frac{2(D-1)}{D}$. Thus, we obtain

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\boldsymbol{\theta}\right)\right) \geq D^2 \left\|\boldsymbol{\theta}^D \circ \left(\begin{array}{c} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{array}\right)^{\frac{D}{D-1}}\right\|_p^p\,. \tag{32}$$

Upper bound:

Let $\boldsymbol{p} = \arg \max\limits_{\boldsymbol{p} \in \mathbb{S}^{N-1}} \|\Phi \boldsymbol{p}\|^2$. Then, we can write

$$\begin{aligned}
\lambda_{\max}\left(\nabla^2 \mathcal{L}(\boldsymbol{\theta})\right) &= \frac{1}{N} \max_{\boldsymbol{p} \in \mathbb{S}^{N-1}} \|\Phi \boldsymbol{p}\|^2 \\
&= \frac{D^2}{N} \sum_{i=1}^{d} \left[\sum_n \boldsymbol{u}_{+,i}^{D-1} \boldsymbol{x}_{n,i} p_n\right]^2 + \left[\sum_n \boldsymbol{u}_{-,i}^{D-1} \boldsymbol{x}_{n,i} p_n\right]^2 \\
&= \frac{D^2}{N} \sum_{i=1}^{d} \left[\boldsymbol{u}_{+,i}^{2D-2} \left[\sum_n \boldsymbol{x}_{n,i} \boldsymbol{p}_n\right]^2 + \boldsymbol{u}_{-,i}^{2D-2} \left[\sum_n \boldsymbol{x}_{n,i} \boldsymbol{p}_n\right]^2\right] \\
&\leq \frac{D^2}{N} \sum_{i=1}^{d} \left[\boldsymbol{u}_{+,i}^{2D-2} \left[\sum_n \boldsymbol{x}_{n,i} \frac{\boldsymbol{x}_{n,i}}{\sqrt{\sum_n \boldsymbol{x}_{n,i}^2}}\right]^2 + \boldsymbol{u}_{-,i}^{2D-2} \left[\sum_n \boldsymbol{x}_{n,i} \frac{\boldsymbol{x}_{n,i}}{\sqrt{\sum_n \boldsymbol{x}_{n,i}^2}}\right]^2\right] \\
&= \frac{D^2}{N} \sum_{i=1}^{d} \left[\boldsymbol{u}_{+,i}^{2D-2} + \boldsymbol{u}_{-,i}^{2D-2}\right] \sum_n \boldsymbol{x}_{n,i}^2 \,, \qquad (33)
\end{aligned}$$

where in the inequality we used the fact that $\max\limits_{\boldsymbol{p} \in \mathbb{S}^{N-1}} \sum_n \boldsymbol{x}_{n,i} \boldsymbol{p}_n \leq \sum_n \boldsymbol{x}_{n,i} \frac{\boldsymbol{x}_{n,i}}{\sqrt{\sum_n \boldsymbol{x}_{n,i}^2}}$.

Combining the lower and upper bounds of $\lambda_{\max}\left(\nabla^2 \mathcal{L}(\boldsymbol{\theta})\right)$ completes our proof. □

A.3. Proof of Lemma A.1

Proof. Let $a, b \in \mathbb{R}_+$ and $p \geq 1$. We need to show that $a^p + b^p - |a-b|^p \geq 0$. First, note that if $a = b$ then the inequality holds. We assume without loss of generality that $a > b$ and define $m \triangleq a - b > 0$. Using Bernoulli inequality we obtain

$$a^p + b^p - |a-b|^p = (m+b)^p + b^p - m^p = m^p \left(1 + \frac{b}{m}\right)^p + b^p - m^p \geq m^p \left(1 + p\frac{b}{m}\right) + b^p - m^p = pbm^{p-1} + b^p \geq 0 \,,$$

where the first inequality relies on Bernoulli inequality and $\frac{b}{m} \geq 0$. □

B. Extension: UV model

In this section, we show that our results can also be extended to the two different layers diagonal linear network, known as the UV model (Azulay et al., 2021), case.

Formally, we consider a depth 2 diagonal model:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{u}_+ \circ \boldsymbol{v}_+ - \boldsymbol{u}_- \circ \boldsymbol{v}_-, \boldsymbol{x} \rangle = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle \;. \qquad (34)$$

This model has also been studied by (Azulay et al., 2021), who conclude that initialization shape and size affect implicit bias under gradient flow. We consider the finite-step-size case and again characterize the properties of stable solutions.

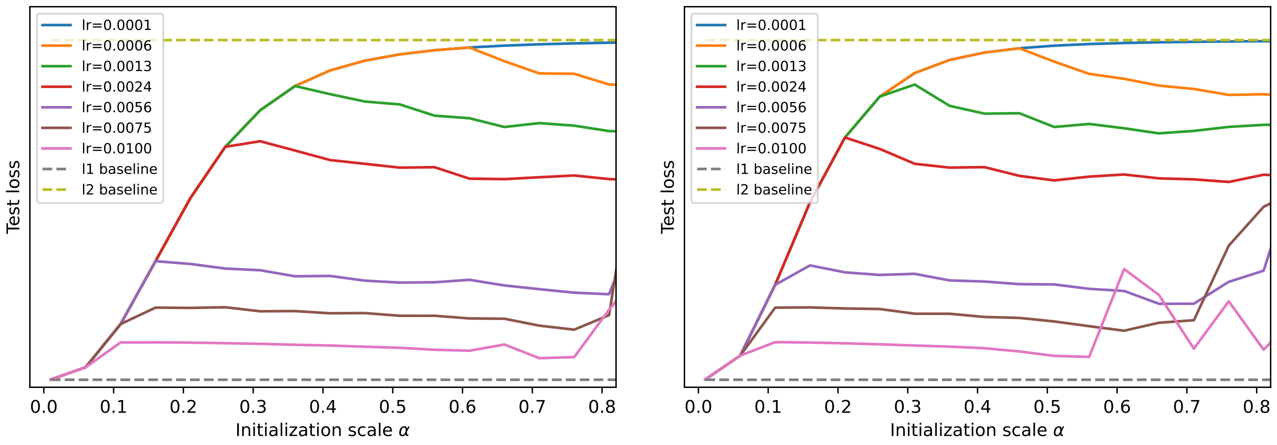
Theorem B.1. *Let $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{v}_+^\top & \boldsymbol{u}_-^\top & \boldsymbol{v}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{4d}$ be a Lyapunov stable solution of GD with step size η. Then,*

$$\left\| \boldsymbol{\theta}^2 \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^2 \right\|_1 \leq \frac{2}{\eta} \,, \qquad (35)$$

where $\hat{\mathbb{E}}\boldsymbol{x} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$ is the empirical mean.

As before, the stability properties induce a bound on the weighted norm of the parameters of the model. We prove this theorem in Appendix C.

Figure 5 demonstrates empirically that the same trends regarding the step size influence on the test loss discussed in the main text for the squared regression model also seem to apply for the UV model.



(a) $s = 0$

(b) s=0.5

Figure 5. The test loss of vs. the initialization scale α in the sparse regression problem described in Section 7, for the UV model. We observe that for small step size, the test loss transitions from the ℓ_1 baseline to the ℓ_2 baseline as the initialization scale α increases, as expected from (Azulay et al., 2021). However, we see that using larger step sizes reduces the error significantly.

C. Proof of Theorem B.1

The proof is identical to Theorem 6.2 proof given in Section A. The only difference is that instead of using Lemma 6.1, we rely on the following key Lemma:

Lemma C.1. *Let* $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{v}_+^\top & \boldsymbol{u}_-^\top & \boldsymbol{v}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{4d}$ *be a solution of the squared loss (Eq. (1)) with* $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{u}_+ \circ \boldsymbol{v}_+ - \boldsymbol{u}_- \circ \boldsymbol{v}_-, \boldsymbol{x} \rangle = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle$. *Then,*

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}(\boldsymbol{\theta})\right) \geq \left\| \boldsymbol{\theta}^2 \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^2 \right\|_1 , \tag{36}$$

where $\hat{\mathbb{E}}\boldsymbol{x} = \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n$ *is the empirical mean.*

We prove this Lemma in Appendix C.1.

C.1. Proof of Lemma C.1

Let $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{v}_+^\top & \boldsymbol{u}_-^\top & \boldsymbol{v}_-^\top \end{pmatrix}^\top \in \mathbb{R}^{4d}$ be a solution of the squared loss (Eq. (1)) with the two layers diagonal model $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{u}_+ \circ \boldsymbol{v}_+ - \boldsymbol{u}_- \circ \boldsymbol{v}_-, \boldsymbol{x} \rangle = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle = \langle \boldsymbol{u} \circ \boldsymbol{v}, \hat{\boldsymbol{x}} \rangle$, where we defined $\hat{\boldsymbol{x}} = \begin{pmatrix} \boldsymbol{x}^\top & -\boldsymbol{x}^\top \end{pmatrix}^\top$, $\boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_+^\top & \boldsymbol{u}_-^\top \end{pmatrix}^\top$, and $\boldsymbol{v} = \begin{pmatrix} \boldsymbol{v}_+^\top & \boldsymbol{v}_-^\top \end{pmatrix}^\top$. From Eq. (21) we have that

$$\begin{aligned} \lambda_{\max}\left(\nabla^2 \mathcal{L}(\boldsymbol{\theta})\right) &= \max_{\boldsymbol{b} \in \mathbb{S}^{d-1}} \boldsymbol{b}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta})\, \boldsymbol{b} \\ &= \max_{\boldsymbol{b} \in \mathbb{S}^{d-1}} \frac{1}{N} \left\| \boldsymbol{\Phi}^\top \boldsymbol{b} \right\|^2 \\ &= \max_{\boldsymbol{p} \in \mathbb{S}^{N-1}} \frac{1}{N} \left\| \boldsymbol{\Phi} \boldsymbol{p} \right\|^2 , \end{aligned} \tag{37}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \nabla_{\boldsymbol{\theta}^*} f(\boldsymbol{x}_1) & \dots & \nabla_{\boldsymbol{\theta}^*} f(\boldsymbol{x}_N) \end{pmatrix} \in \mathbb{R}^{d \times N} . \tag{38}$$

For the two layers diagonal model

$$\nabla_{\boldsymbol{\theta}^*} f(\boldsymbol{x}) = \begin{pmatrix} \hat{\boldsymbol{x}}_n \circ \boldsymbol{v}(t) \\ \hat{\boldsymbol{x}}_n \circ \boldsymbol{u}(t) \end{pmatrix} \tag{39}$$

and thus we obtain

$$\boldsymbol{\Phi} = \begin{pmatrix} \hat{\boldsymbol{x}}_1 \circ \boldsymbol{v} & \cdots & \hat{\boldsymbol{x}}_n \circ \boldsymbol{v} \\ \hat{\boldsymbol{x}}_1 \circ \boldsymbol{u} & \cdots & \hat{\boldsymbol{x}}_n \circ \boldsymbol{u} \end{pmatrix} \in \mathbb{R}^{4d\times n}. \tag{40}$$

We substitute this result into Eq. (21), and take $\boldsymbol{p} = \frac{1}{\sqrt{N}}\mathbf{1}_{N\times 1}$:

$$\begin{aligned}
\lambda_{\max}\left(\nabla_u^2 \mathcal{L}(u)\right) &= \frac{1}{N} \max_{\boldsymbol{p}\in\mathbb{S}^{N-1}} \|\boldsymbol{\Phi}\boldsymbol{p}\|^2 \\
&\geq \frac{1}{N^2}\sum_{i=1}^{d}\left(\left[\sum_n \boldsymbol{v}_{+,i}\boldsymbol{x}_{n,i}\right]^2 + \left[\sum_n \boldsymbol{v}_{-,i}\boldsymbol{x}_{n,i}\right]^2 + \left[\sum_n \boldsymbol{u}_{+,i}\boldsymbol{x}_{n,i}\right]^2 + \left[\sum_n \boldsymbol{u}_{-,i}\boldsymbol{x}_{n,i}\right]^2\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{d}\left(\boldsymbol{u}_{+,i}^2 + \boldsymbol{u}_{-,i}^2 + \boldsymbol{v}_{+,i}^2 + \boldsymbol{v}_{-,i}^2\right)\left[\sum_n \boldsymbol{x}_{n,i}\right]^2 \\
&= \sum_{i=1}^{d}\left(\boldsymbol{u}_{+,i}^2 + \boldsymbol{u}_{-,i}^2 + \boldsymbol{v}_{+,i}^2 + \boldsymbol{v}_{-,i}^2\right)\left(\hat{\mathbb{E}}\boldsymbol{x}_i\right)^2 \\
&= \left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2 + \boldsymbol{v}_+^2 + \boldsymbol{v}_-^2\right)^\top \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2,
\end{aligned} \tag{41}$$

where we defined $\hat{\mathbb{E}}\boldsymbol{x} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$. Lastly, note that

$$\left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2 + \boldsymbol{v}_+^2 + \boldsymbol{v}_-^2\right)^\top \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 = \left\|\boldsymbol{\theta}^2 \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^2\right\|_1 \tag{42}$$

Thus, we obtain

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}(\boldsymbol{\theta})\right) \geq \left\|\boldsymbol{\theta}^2 \circ \begin{pmatrix} \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \\ \hat{\mathbb{E}}\boldsymbol{x} \end{pmatrix}^2\right\|_1 . \tag{43}$$

D. Generalization Result

In this section, we show that under some assumptions, our procedure returns a predictor that generalizes well. Our main result in this section shows that that a parameter setting that minimizes the maximum eigenvalue of the Hessian yields a predictor that generalizes well. We show this by arguing that the maximum eigenvalue of the Hessian is upper and lower bounded by weighted ℓ_1 norms of the predictor, weighted by different weights. We then argue that with sufficiently many samples, these weighted norms come within a constant factor of the ℓ_1 norm of the ℓ_1 norm minimizing interpolator. To connect this result to the procedure of running gradient descent with large step sizes, we appeal to the stability analysis and an assumption regarding the convergence of gradient descent.

First, for convenience we reproduce the definition of a sub-Gaussian random variable from (Vershynin) and comment on how we use this.

(Proposition 2.5.2 in (Vershynin)). Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.

(i) The tails of X satisfy:

$$\mathbb{P}\left[|X| \geq t\right] \leq 2\exp\left(-t^2/K_1^2\right) \quad \text{for all } t \geq 0\,.$$

(ii) The moments of X satisfy

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p} \text{ for all } p \geq 1\,.$$

(iii) the MGF of X^2 satisfies

$$\mathbb{E}\exp(\sigma^2 X^2) \leq \exp(K_3^2\,\sigma^2) \text{ for all } \sigma \text{ such that } |\sigma| \leq \frac{1}{K_3}\,.$$

(iv) The MGF of X^2 is bounded at some point, namely

$$\mathbb{E}\exp(X^2/K_4^2) \leq 2\,.$$

Moreover, if $\mathbb{E}X = 0$, the properties i-iv are also equivalent to the following one.

(v) The MGF of X satisfies

$$\mathbb{E}\exp(\sigma X) \leq \exp(K_5^2\sigma^2) \text{ for all } \sigma \in \mathbb{R}\,.$$

We also use the following definition.
(Definition 2.5.6 in (Vershynin)). A random variable X that satisfies one of the equivalent properties i-iv in Proposition D is called a sub-Gaussian random variable. The *sub-Gaussian norm* of X, denoted $\|X\|_{\Psi_2}$, is defined to be the smallest K_4 in property iv. In other words, define:

$$\|X\|_{\Psi_2} = \inf\left\{t > 0\ \ \mathbb{E}\exp(X^2/t^2) \leq 2\right\}\,.$$

For a sub-Gaussian random variable V, we define σ as $\sigma := \|V\|_{\Psi_2}$. From there it is easy to verify the following proposition.

Proposition D.1. *Suppose for a sub-Gaussian random variable V we have $\sigma := \|V\|_{\Psi_2}$, where $\|\cdot\|_{\Psi_2}$ represents the sub-Gaussian norm. Then we have that $\mathbb{P}\left[|V| \geq t\right] \leq 2\exp\left(\frac{-t^2}{\sigma^2}\right)$. We shall call σ the sub-Gaussian parameter.*

Proof. Starting from the definition of sub-Gaussian norm from (Vershynin), we have that:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{V^2}{\sigma^2}\right)\right] &\leq 2 \\
e^{-\frac{t^2}{\sigma^2}}\mathbb{E}\left[\exp\left(\frac{V^2}{\sigma^2}\right)\right] &\leq 2\,e^{-\frac{t^2}{\sigma^2}} && \text{multiply both sides} \\
\mathbb{P}\left[e^{\frac{V^2}{\sigma^2}} \geq e^{\frac{t^2}{\sigma^2}}\right] \leq e^{-\frac{t^2}{\sigma^2}}\mathbb{E}\left[\exp\left(\frac{V^2}{\sigma^2}\right)\right] &\leq 2\,e^{-\frac{t^2}{\sigma^2}} && \text{Markov's Inequality} \\
\mathbb{P}\left[e^{\frac{V^2}{\sigma^2}} \geq e^{\frac{t^2}{\sigma^2}}\right] = \mathbb{P}\left[|V| \geq t\right]& \\
\mathbb{P}\left[|V| \geq t\right] &\leq 2\,e^{-\frac{t^2}{\sigma^2}}\,.
\end{aligned}$$

□

Now we state our main generalization result:

Theorem 5.2. *Suppose the entries of $\boldsymbol{X}$ are independently and identically drawn from a Gaussian with variance σ^2 and mean $\mu = O(\sigma)$. Let $\boldsymbol{y}$ be the labels produced by a k-sparse linear predictor, i.e., $\boldsymbol{y} = \langle\boldsymbol{\beta}^\star, \boldsymbol{X}\rangle$, for some $\|\boldsymbol{\beta}^\star\|_0 \leq k$. For $\tilde{\boldsymbol{\beta}}$ as defined in in Lemma 5.1 and $\zeta = 1 + \frac{\sigma^2}{\mu^2}$, the population loss $\mathcal{L}_\mathcal{D}$ obeys, with probability at least $1-\delta$:*

$$\frac{\mathcal{L}_\mathcal{D}(\tilde{\boldsymbol{\beta}})}{\mathcal{L}_\mathcal{D}(\mathbf{0})} \leq \mathcal{O}\left(\frac{\zeta^2 k\,\mathrm{polylog}(dN/\delta)}{N}\right). \tag{11}$$

Proof. In order to prove this theorem, we require three main ingredients:

1. First, we bound $\lambda_{\max}$ in terms of a weighed norm of the predictor. To this end, we use the upper and lower bounds on $\lambda_{\max}$ from Theorem 4.1 and the lower bound extension to the predictor $\boldsymbol{\beta}$ space from Theorem 4.2. In addition, we extend the upper bound to also be in terms of the predictor, rather than the parameters $\boldsymbol{\theta}$. (Lemma D.2)
2. Second, we show that if sufficiently many samples are used, the respective weighted ℓ_1 norms of a predictor β are bounded by multiplicative factors of $||\boldsymbol{\beta}||_1$. (Lemma 5.1).
3. Finally, we apply a generalization result from (Srebro et al.) that applies for predictors within a constant multiplicative factor of $||\boldsymbol{\beta}^\star||_1$, where $\boldsymbol{\beta}^\star := \arg\min_{\langle \boldsymbol{\beta}, \boldsymbol{X}\rangle = \boldsymbol{y}} ||\boldsymbol{\beta}||_1$. (Lemma D.6)

D.1. Bounding $\lambda_{\max}$ in terms of a weighted norm of the predictor

Lemma D.2. *For $\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_- := \arg\min_{\boldsymbol{u}_+, \boldsymbol{u}_- : \langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y}} \lambda_{\max}(\boldsymbol{u}_+, \boldsymbol{u}_-)$:*

$$\left\| \boldsymbol{\beta}(\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_-) \circ 4 \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \right\|_1 \leq \lambda_{\max}(\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_-) \leq \left\| \left(\boldsymbol{\beta}(\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_-) \circ 4\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)\right)\right\|_1 .$$

Proof. From Lemma 4.1, we have that:

$$\left\| \left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ 4 \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \right\|_1 \leq \lambda_{\max}(\boldsymbol{u}_+, \boldsymbol{u}_-) \leq \left\| \left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ 4\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)\right\|_1 .$$

First, on the lower bound side, we have that for any set of parameters $\boldsymbol{u}_+, \boldsymbol{u}_-$:

$$\left\| \beta(\boldsymbol{u}_+, \boldsymbol{u}_-) \circ 4 \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \right\|_1 = \left\| \left(\boldsymbol{u}_+^2 - \boldsymbol{u}_-^2\right) \circ 4 \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \right\|_1 \leq \left\| \left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ 4 \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \right\|_1 \leq \lambda_{\max}(\boldsymbol{u}_+, \boldsymbol{u}_-)$$

On the upper bound side, we have that:

$$\begin{aligned}
\lambda^\star &:= \inf_{\boldsymbol{u}_+, \boldsymbol{u}_- : \langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y}} \lambda_{\max}(\boldsymbol{u}_+, \boldsymbol{u}_-) \\
&\leq \inf_{\boldsymbol{u}_+, \boldsymbol{u}_- : \langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y}} \left\| \left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ 4\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)\right\|_1 \\
&\leq \inf_{\boldsymbol{u}_+, \boldsymbol{u}_- : \langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y} \, , \, \boldsymbol{u}_+ \circ \boldsymbol{u}_- = 0} \left\| \left(\boldsymbol{u}_+^2 + \boldsymbol{u}_-^2\right) \circ 4\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)\right\|_1 \\
&= \inf_{\boldsymbol{u}_+, \boldsymbol{u}_- : \langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y}} \left\| \left(\beta(\boldsymbol{u}_+, \boldsymbol{u}_-) \circ 4\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)\right)\right\|_1
\end{aligned}$$

Thus, we have that for $\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_- := \arg\min_{\boldsymbol{u}_+, \boldsymbol{u}_- : \langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y}} \lambda_{\max}(\boldsymbol{u}_+, \boldsymbol{u}_-)$:

$$\left\| \boldsymbol{\beta}(\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_-) \circ 4 \left(\hat{\mathbb{E}}\boldsymbol{x}\right)^2 \right\|_1 \leq \lambda_{\max}(\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_-) \leq \left\| \left(\boldsymbol{\beta}(\tilde{\boldsymbol{u}}_+, \tilde{\boldsymbol{u}}_-) \circ 4\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)\right)\right\|_1 .$$

□

D.2. Concentration of Weighting Implies Bounds by Multiplicative Factor of Optimal ℓ_1 Norm

Having established bounds on the maximum eigenvalue of the Hessian at the relevant parameters in terms of the predictor borne of those parameters, let us now consider the concentration of the bounds.

Lemma D.3. *Suppose the data $\boldsymbol{X} \in \mathbb{R}^{d \times N}$ is drawn independently and identically (i.i.d.) with mean μ and such that $X - \mu$ is sub-Gaussian (per the Definition in Appendix D) with parameter σ . Define $\tilde{\mathbf{u}}_+$, $\tilde{\mathbf{u}}_-$ such that:*

$$\tilde{\boldsymbol{u}}_+ \, , \tilde{\boldsymbol{u}}_- := \operatorname*{arg\,min}_{\langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2 , \boldsymbol{X}\rangle = \boldsymbol{y}} \lambda_{\max}(\boldsymbol{u}_+, \boldsymbol{u}_-) \, , \tag{9}$$

and let $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{u}}_+^2 - \tilde{\boldsymbol{u}}_-^2$ *. Then, we have that, provided* $N = \Omega\left(\max\left\{\frac{\sigma^2 \log(d/\delta)}{\sigma^2+\mu^2}, \frac{\sigma\sqrt{\log(d/\delta)}}{\mu}\right\}^2\right)$ *, with probability at least* $1-\delta$ *,*

$$\left\|\tilde{\boldsymbol{\beta}}\right\|_1 \leq \left(1+\frac{\sigma^2}{\mu^2}\right)\cdot 1.1 \inf_{\langle \boldsymbol{\beta}, \boldsymbol{X}\rangle = \boldsymbol{y}} \|\boldsymbol{\beta}\|_1 . \tag{10}$$

Proof. In order to show this, we will separately show concentration of the lower bound (Lemma D.4) and the upper bound (Lemma D.5) on λ_{max} . We do so by showing that with sufficiently many samples, the weightings concentrate to the mean. Following this, we can combine the results to bound the ℓ_1 norm of the predictor gotten by minimizing λ_{max} .

Lemma D.4. *Suppose* $\hat{\mathbb{E}}(\boldsymbol{x}) \coloneqq \frac{1}{N}\sum_{n=1}^N \boldsymbol{x}_n$ *, where* x_n *are the columns of* $\boldsymbol{X} \in \mathbb{R}^{d\times N}$ *, which is a matrix with each entry an independent, identically distributed sub-Gaussian around* μ *with parameter* σ *. Then, for any vector* $\boldsymbol{w}$ *, with probability at least* $1-\delta$*:*

$$\left(\mu - \sigma\sqrt{\frac{\log(d/\delta)}{c\,n}}\right)^2 ||\boldsymbol{w}||_1 \leq ||\boldsymbol{w}\circ\hat{\mathbb{E}}(\boldsymbol{x})^2||_1 . \tag{44}$$

Proof. We start by expanding the weighted ℓ_1 norm for any vector w:

$$||\boldsymbol{w}\circ\hat{\mathbb{E}}(\boldsymbol{x})^2||_1 = \sum_{i=1}^d \left| w_i \left(\hat{\mathbb{E}}(\boldsymbol{x})\right)_i^2 \right| .$$

For ease of notation, let $\bar{x}_i = \left(\hat{\mathbb{E}}(\boldsymbol{x})\right)_i$ For any term in this sum where $\bar{x}_i \geq 1$, we can lower bound that term by $|w_i|$. However, this might not be the case for all the μ_i . Thus, we normalize:

$$\begin{aligned}
||\boldsymbol{w}\circ\hat{\mathbb{E}}(\boldsymbol{x})^2||_1 &= \sum_{i=1}^d \left|w_i\,\bar{x}_i^2\right| \\
&= \sum_{i=1}^d \left|w_i\,\bar{x}_i^2 \cdot \frac{\min_i \bar{x}_i^2}{\min_i \bar{x}_i^2}\right| = \left(\min_i \bar{x}_i^2\right)\sum_{i=1}^d \left|w_i \cdot \frac{\bar{x}_i^2}{\min_i \bar{x}_i^2}\right| \\
&\geq \left(\min_i \bar{x}_i^2\right)\sum_{i=1}^d |w_i| = \left(\min_i \bar{x}_i^2\right) ||\boldsymbol{w}||_1 .
\end{aligned}$$

Next, we wish to bound the value of $\min_i \bar{x}_i^2$. To do so, we want to find t such that:

$$\begin{aligned}
&\mathbb{P}\left[\forall i\, \bar{x}_i > \mu - t\right] \geq 1-\delta \\
\iff &\mathbb{P}\left[\exists\, i\, : \bar{x}_i - \mu < -t\right] \leq \delta
\end{aligned}$$

Now, we apply Proposition D.1 and Hoeffding's Inequality as stated in Theorem 2.6.2 in (Vershynin), which we reproduce here for clarity (where c is some absolute constant).

(Theorem 2.6.2 in (Vershynin)). Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables. Then for every $t \geq 0$, we have:

$$\mathbb{P}\left[\left|\sum_{i=1}^N X_i\right| \geq t\right] \leq 2\exp\left(-\frac{c\,t^2}{\sum_{i=1}^N \|X_i\|_{\psi_2}^2}\right) .$$

We can use Hoeffing's Inequality and the union bound to get:

$$\mathbb{P}\left[\exists\, i\, : \bar{x}_i - \mu < -t\right] \leq d\exp\left(\frac{-Nct^2}{\sigma^2}\right)$$

$$\text{so} \quad t = \sqrt{\frac{\sigma^2\log(d/\delta)}{c\,N}} .$$

We get that :

$$\left(\mu - \sigma\sqrt{\frac{\log(d/\delta)}{c\,N}}\right)^2 ||\boldsymbol{w}||_1 \leq ||\boldsymbol{w} \circ \hat{\mathbb{E}}(\boldsymbol{x})^2||_1 \quad \text{ with probability } \geq 1 - \delta \tag{45}$$

□

Next, we consider the upper bound.

Lemma D.5. *Suppose* $\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right) \coloneqq \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n^2$ *, where* $\boldsymbol{x}_n$ *are the columns of* $X \in \mathbb{R}^{d\times N}$ *, which is a matrix with each entry an independent, identically distributed sub-Gaussian around* μ *with parameter* σ *. Then, for any vector* $\mathbf{w}$ *, with probability at least* $1-\delta$ *:*

$$||\boldsymbol{w} \circ \hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)||_1 \leq ||\boldsymbol{w}||_1 \cdot \left(\mu^2 + \sigma^2 + \frac{8\sigma^2 e}{\sqrt{n}} \log \frac{2d}{\delta}\right) \tag{46}$$

Proof. Let us bound the value of $\hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)$. Observe that $y_{n,i} \coloneqq x_{n,i}^2$ is a subexponential random variable due to it being the square of a subgaussian random variable. For clarity, we reproduce Bernstein's inequality (Corollary 2.8.3 in (Vershynin)) before applying it.

(Corollary 2.8.3 in (Vershynin)). Let $X_1, \ldots, X_N$ be independent, mean 0, sub-exponential random variables. Then, for every $t \geq 0$, we have:

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right| \geq t\right] \leq 2\exp\left(-c\,N\,\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right),$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Applying Bernstein's inequality, we get that:

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{n=1}^{N} y_{n,i} - \mu^2 - \sigma^2\right| \geq t\right] \leq 2\exp\left(-c\,N \min\left\{\frac{t^2}{\|y_{n,i}\|_{\psi_1}^2}, \frac{t}{\|y_{n,i}\|_{\psi_1}}\right\}\right).$$

By Lemma 2.7.6 in (Vershynin), we have that $\|y_{n,i}\|_{\psi_1} = \left\|x_{n,i}^2\right\|_{\psi_1} = \|x_{n,i}\|_{\psi_2}^2 = \sigma^2$.

This implies that with probability at least $1-\delta$, if t is small enough, i.e., $t \leq \sigma^2$, then:

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{n=1}^{N} y_{n,i} - \mu^2 - \sigma^2\right| \geq t\right] \leq 2\exp\left(-c\,N\frac{t^2}{\sigma^4}\right).$$

Then, applying union bound:

$$\mathbb{P}\left[\exists\, i\,:\, \left|\frac{1}{N}\sum_{n=1}^{N} y_{n,i} - \mu^2 - \sigma^2\right| \geq t\right] \leq 2d\exp\left(-c\,N\frac{t^2}{\sigma^4}\right).$$

Setting the probability of failure to δ , we get:

$$\mathbb{P}\left[\exists\, i\,:\, \left|\frac{1}{N}\sum_{n=1}^{N} y_{n,i} - \mu^2 - \sigma^2\right| \geq \sigma^2\sqrt{\frac{\log(2d/\delta)}{cN}}\right] \leq \delta.$$

To match the earlier condition on t , we require that: $t \leq \sigma^2 \Leftrightarrow N \geq \frac{\log(2d/\delta)}{c}$. Thus, this gives us that with probability at least $1-\delta$,

$$\forall\, i \in [d] \quad \mu^2 + \sigma^2 - \frac{\sigma^2}{\sqrt{N}}\sqrt{\log\frac{2d}{\delta}} \leq \frac{1}{N}\sum_{n=1}^{N} y_{n,i} \leq \mu^2 + \sigma^2 + \frac{\sigma^2}{\sqrt{N}}\sqrt{\log\frac{2d}{\delta}}, \qquad \text{if } N \geq \left(\frac{\log(2d/\delta)}{c}\right).$$

This then gives us that with probability $\geq 1-\delta$, provided $N \geq \frac{\log(2d/\delta)}{c}$:

$$||\boldsymbol{\beta} \circ \hat{\mathbb{E}}\left(\boldsymbol{x}^2\right)||_1 \leq ||\boldsymbol{\beta}||_1 \cdot \left(\mu^2 + \sigma^2 + \frac{\sigma^2}{\sqrt{N}} \log \frac{2d}{\delta}\right) \tag{47}$$

□

With these bounds on the upper and lower bounds of λ_{max} , we can now consider the ℓ_1 norm of the predictor we output in terms of the ℓ_1 norm of the ℓ_1 norm-minimizing interpolator.

First, define

$$\hat{\boldsymbol{u}}_+, \hat{\boldsymbol{u}}_- \coloneqq \arg \min_{\langle \boldsymbol{\beta}(\boldsymbol{u}_+, \boldsymbol{u}_-), \boldsymbol{X}\rangle = \boldsymbol{y}} \left\|\boldsymbol{\beta}(\boldsymbol{u}_+, \boldsymbol{u}_-) \circ \hat{\mathbb{E}}(\boldsymbol{x})^2\right\|_1$$

and

$$\lambda^\star = \min_{\langle \boldsymbol{\beta}(\boldsymbol{u}_+, \boldsymbol{u}_-), \boldsymbol{X}\rangle = \boldsymbol{y}} \lambda_{\max}(\boldsymbol{u}_+\,, \boldsymbol{u}_-)\,.$$

Further, define $\boldsymbol{\beta}^\star \coloneqq \arg\min_{\langle \boldsymbol{\beta}(\hat{\boldsymbol{u}}_+, \hat{\boldsymbol{u}}_-), \boldsymbol{X}\rangle = \boldsymbol{y}} \|\boldsymbol{\beta}\|_1$.

By these definitions, we have:

$$\left\|\boldsymbol{\beta}(\tilde{\boldsymbol{u}}_+\,, \tilde{\boldsymbol{u}}_-) \circ \hat{\mathbb{E}}(x)^2\right\|_1 \leq \lambda^\star \leq \lambda(\hat{\boldsymbol{u}}_+, \hat{\boldsymbol{u}}_-) \leq \left\|\boldsymbol{\beta}(\hat{\boldsymbol{u}}_+\,, \hat{\boldsymbol{u}}_-) \circ \hat{\mathbb{E}}(\boldsymbol{x}^2)\right\|_1 \leq \left\|\boldsymbol{\beta}^\star \circ \hat{\mathbb{E}}(\boldsymbol{x}^2)\right\|_1$$

$$\text{with probability } \geq 1-\delta\,, \qquad \|\boldsymbol{\beta}(\tilde{\boldsymbol{u}}_+\,, \tilde{\boldsymbol{u}}_-)\|_1 \leq \frac{\left(\mu^2+\sigma^2+\frac{8\,e\,\sigma^2}{\sqrt{N}}\ln(\frac{4d}{\delta})\right)}{\left(\mu - \sigma\sqrt{\frac{\log(2d/\delta)}{c\,n}}\right)^2} \|\boldsymbol{\beta}^\star\|_1\ ,$$

which follows as a result of applying Lemmas D.4 and D.5, each with probability $\delta/2$. Next, we bound the multiplicative factor on the left.

$$
\begin{aligned}
&\frac{\left(\mu^2+\sigma^2+\frac{8\,e\,\sigma^2}{\sqrt{N}}\ln(\frac{4d}{\delta})\right)}{\left(\mu-\sigma\sqrt{\frac{\log(2d/\delta)}{c\,n}}\right)^2} \\
&= \frac{\mu^2+\sigma^2}{\mu^2}\cdot\frac{\left(1+\frac{8\,e\,\sigma^2}{(\mu^2+\sigma^2)}\frac{\log(4d/\delta)}{\sqrt{N}}\right)}{\left(1-\frac{\sigma}{\mu}\sqrt{\frac{\log(2d/\delta)}{c\,N}}\right)^2} \\
&\leq \left(1+\frac{\sigma^2}{\mu^2}\right)\cdot\frac{\left(1+\frac{8\,e\,\sigma^2}{(\mu^2+\sigma^2)}\frac{\log(4d/\delta)}{\sqrt{N}}\right)}{1-2\frac{\sigma}{\mu}\sqrt{\frac{\log(2d/\delta)}{c\,N}}} \qquad (1-x)^2 \geq 1-2x \\
&= \left(1+\frac{\sigma^2}{\mu^2}\right)\cdot\left(1+\frac{\left(\frac{8\,e\,\sigma^2}{(\mu^2+\sigma^2)}\frac{\log(4d/\delta)}{\sqrt{N}}\right)+2\frac{\sigma}{\mu}\sqrt{\frac{\log(2d/\delta)}{c\,N}}}{1-2\frac{\sigma}{\mu}\sqrt{\frac{\log(d/\delta)}{c\,N}}}\right)\cdot\,||\boldsymbol{\beta}^\star||_1 \\
&\leq \left(1+\frac{\sigma^2}{\mu^2}\right)\cdot\left(1+\frac{2\cdot b}{1-b}\right)\cdot\,||\boldsymbol{\beta}^\star||_1 \quad \text{where } b=\max\left\{\frac{8\,e\,\sigma^2}{(\mu^2+\sigma^2)}\frac{\log(2d/\delta)}{\sqrt{N}}\,,\,\frac{2\,\sigma}{\mu}\sqrt{\frac{\log(d/\delta)}{c\,N}}\right\} \\
&\leq \left(1+\frac{\sigma^2}{\mu^2}\right)\cdot(1+4\cdot b)\,\cdot\,||\boldsymbol{\beta}^\star||_1 \qquad \text{provided } b\leq\frac{1}{2}
\end{aligned}
$$

$$\text{because } \frac{2\epsilon}{1-\epsilon}\leq 4\epsilon \text{ if } \epsilon<1/2\,.$$

With probability at least $1-\delta$, we have that, provided $N=\Omega\left(\max\left\{\frac{\sigma^2\,\log(d/\delta)}{\sigma^2+\mu^2},\,\frac{\sigma\,\sqrt{\log(d/\delta)}}{\mu}\right\}^2\right)$

$$
\begin{aligned}
\|\boldsymbol{\beta}(\tilde{\boldsymbol{u}}_+\,,\tilde{\boldsymbol{u}}_-)\| &\leq \left(1+\frac{\sigma^2}{\mu^2}\right)\left(1+\mathcal{O}\left(\max\left\{\frac{\sigma^2}{(\mu^2+\sigma^2)}\frac{\log(d/\delta)}{\sqrt{N}}\,,\,\frac{2\,\sigma}{\mu}\sqrt{\frac{\log(d/\delta)}{N}}\right\}\right)\right)\inf_{\langle\boldsymbol{\beta},\boldsymbol{X}\rangle=\boldsymbol{y}}\|\boldsymbol{\beta}\|_1 \\
&\leq \left(1+\frac{\sigma^2}{\mu^2}\right)\cdot 1.1\inf_{\langle\boldsymbol{\beta},\boldsymbol{X}\rangle=\boldsymbol{y}}\|\boldsymbol{\beta}\|_1 \qquad (48)
\end{aligned}
$$

□

D.3. Applying Generalization Result

Several works have considered the question of generalization in the sparse regression problem. We apply the following result of (Srebro et al.), which applies for any hypothesis class $\mathcal{H}$ in terms of the Rademacher complexity, $\mathcal{R}_N(\mathcal{H})$. The predictor $\hat{h}$ in the second line is $\hat{h} \coloneqq \arg\min_{h\in\mathcal{H}}\mathcal{L}(h)$.

(Theorem 1 in (Srebro et al.), restated). For an H-smooth non-negative loss ϕ s.t. $\forall_{x,y,h}\,|\phi(h(x),y)|\leq b$, for any $\delta>0$ we have that with probability at least $1-\delta$ over a random sample of size n , for any $h\in\mathcal{H}$,

$$\mathcal{L}_{\mathcal{D}}(h)\leq\mathcal{L}(h)+K\left(\sqrt{\mathcal{L}(h)}\left(\sqrt{H}\,\log^{1.5}n\,\mathcal{R}_n(\mathcal{H})\right)+H\log^3 n\mathcal{R}_n^2(\mathcal{H})+\frac{b\,\log(1/\delta)}{n}\right) \qquad (49)$$

and so:

$$\mathcal{L}_{\mathcal{D}}(\hat{h})\leq\inf_{h\in H}\mathcal{L}(h)+K\left(\sqrt{\inf_{h\in H}\mathcal{L}(h)}\left(\sqrt{H}\,\log^{1.5}N\,\mathcal{R}_N(\mathcal{H})\right)+H\log^3 N\mathcal{R}_N^2(\mathcal{H})+\frac{b\,\log(1/\delta)}{N}\right)$$

where $K<10^5$ is a numeric constant.

In this theorem, they show that the population loss of any predictor in a class when considered under smooth, bounded loss is bounded by a term involving the training loss, the Rademacher complexity, and several other standard parameters. Now, we extend this result to show that a slight variant of the predictor we find with bounded ℓ_1 norm predictor generalizes well.

Lemma D.6. *Suppose in the setting stated in Theorem 5.2, $\hat{\boldsymbol{\beta}}$ interpolates the training data such that* $||\hat{\boldsymbol{\beta}}||_1 \leq 1.1\left(1+\frac{\sigma^2}{\mu^2}\right)||\boldsymbol{\beta}^\star||_1$ *, where $\beta^\star$ is the ground truth predictor. Let $\mathcal{H}$ be the set of linear predictors with ℓ_1 norm bounded by* $1.1\ \left(1+\frac{\sigma^2}{\mu^2}\right)\ ||\boldsymbol{\beta}^\star||_1$ *. Define:*

$$\zeta := 1+\frac{\sigma^2}{\mu^2}$$

Then, we have the following with probability at least $1-\delta$:

$$L(\hat{\boldsymbol{\beta}}) \leq \mathbb{E}(y^2)\mathcal{O}\left(\frac{\zeta^2 k \operatorname{polylog}(dN/\delta)}{N}\right)$$

Proof. In order to show this, we use that the set of predictors that might be output by minimizing $\lambda_{\max}$ is a set of predictors with ℓ_1 norm bounded by $B := \mathcal{O}\left(\sqrt{\zeta}\cdot\|\boldsymbol{\beta}^\star\|_1\right)$.

In order to apply their results here, we first consider a variant of the hypothesis class that we are actually studying. Let the main hypothesis class in which $\hat{\beta}$ lies be defined as follows:

$$\mathcal{H} := \left\{h_{\boldsymbol{w}} \,:\, \boldsymbol{x}\mapsto\langle\boldsymbol{w}\,,\boldsymbol{x}\rangle\middle|\, \|\boldsymbol{w}\|_1 \leq B\right\}.$$

Then, we define:

$$\tilde{\mathcal{H}} := \left\{\tilde{h}_{\boldsymbol{w}} \,:\, \boldsymbol{x}\mapsto\begin{cases}\langle\boldsymbol{w}\,,\boldsymbol{x}\rangle & \text{when } \|\boldsymbol{x}-\boldsymbol{\mu}\|_\infty \leq R\,,\\ 0 & \text{otherwise}\end{cases}\right\}.$$

With the definition above, we can bound the loss as follows:

$$\begin{aligned}\max\left\{\max_{\boldsymbol{x},y,h}(y-h(\boldsymbol{x}))^2\,,\operatorname{var}(y)\right\} &= \max\left\{\max_{x,w}(\langle\boldsymbol{w}^\star-\boldsymbol{w},\boldsymbol{x}\rangle)^2\,,\mathbb{E}(y^2)\right\}\\ &\leq \max\left\{\|\boldsymbol{w}^\star-\boldsymbol{w}\|_1^2\,\|\boldsymbol{x}\|_\infty^2\,,k\,R^2\right\} \leq \max\left\{(2\,B)^2(R+\mu)^2\,,B^2(\mu^2+\sigma^2)\right\}\\ &\leq 4B^2\,(R+\mu)^2\end{aligned}$$

Let us also compute the scaling of $\|\boldsymbol{\beta}^\star\|_1$:

$$\mathbb{E}(y^2) = \mathbb{E}(\langle\boldsymbol{\beta}^\star,\boldsymbol{x}\rangle^2) \leq \mathbb{E}\left(\left(\sum_i \boldsymbol{\beta}_i^\star\boldsymbol{x}_i\right)^2\right) = E\left(\left(\sum_i\boldsymbol{\beta}_i^\star\mu+\sum_i\boldsymbol{\beta}_i^\star(\boldsymbol{x}_i-\mu)\right)^2\right)$$

$$\begin{aligned}\Rightarrow \frac{\|\boldsymbol{\beta}^\star\|_1^2}{k}\sigma^2 \leq \|\boldsymbol{\beta}^\star\|_2^2\,\sigma^2 \leq \mathbb{E}(y^2) \leq \|\boldsymbol{\beta}^\star\|_1^2\,\mu^2+\|\boldsymbol{\beta}^\star\|_2^2\,\sigma^2\\ \Rightarrow \|\boldsymbol{\beta}^\star\|_1^2 \leq \frac{k\,\mathbb{E}(y^2)}{\sigma^2}\end{aligned}$$

Let us compare the population loss of a function in $\mathcal{H}$ to the population loss of the corresponding function in $\tilde{\mathcal{H}}$. Let us consider this by conditioning on whether the datapoint over which we are taking the expectation lies inside or outside the ball of radius R centered at $\boldsymbol{\mu}$. For ease of notation, we define $\bar{\boldsymbol{x}} := \boldsymbol{x}-\boldsymbol{\mu}$. We have, for any $h_w \in \mathcal{H}$ and the corresponding

$\tilde{h}_{\boldsymbol{w}}$:

$$\begin{aligned}
\mathcal{L}(h_w) &= \mathbb{E}\left((y - h_w(\boldsymbol{x}))^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \leq R\right) \mathbb{P}\left[\|\bar{\boldsymbol{x}}\|_\infty \leq R\right] + \mathbb{E}\left((y - h_w(\boldsymbol{x}))^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) \mathbb{P}\left[\|\bar{\boldsymbol{x}}\|_\infty \geq R\right] \\
&= \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) - \mathbb{E}\left((y - \mathbb{E}[y])^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) \mathbb{P}\left[\|\bar{\boldsymbol{x}}\|_\infty \geq R\right] + \mathbb{E}\left((y - h_w(\boldsymbol{x}))^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) \mathbb{P}\left[\|\bar{\boldsymbol{x}}\|_\infty \geq R\right] \\
&= \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) - \mathbb{E}\left((y - \mathbb{E}[y])^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) \mathbb{P}\left[\|\bar{\boldsymbol{x}}\|_\infty \geq R\right] + \mathbb{E}\left((h_{w^\star}(\boldsymbol{x}) - h_w(\bar{\boldsymbol{x}}))^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) \mathbb{P}\left[\|\bar{\boldsymbol{x}}\|_\infty \geq R\right] \\
&= \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) - \text{var}(Y)\, \delta_1 + \mathbb{E}\left((\langle \boldsymbol{w}^\star - \boldsymbol{w}, \boldsymbol{x}\rangle)^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) \delta_1 \qquad\qquad \text{define } \boldsymbol{g} \coloneqq \boldsymbol{w}^\star - \boldsymbol{w} \\
&= \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) - \text{var}(Y)\, \delta_1 + \mathbb{E}\left((\langle \boldsymbol{g}, \bar{\boldsymbol{x}} + \boldsymbol{\mu}\rangle)^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) \delta_1 \\
&= \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) - \text{var}(Y)\, \delta_1 + \mathbb{E}\left((\langle \boldsymbol{g}, \bar{\boldsymbol{x}}\rangle)^2 + 2\langle \boldsymbol{g}, \bar{\boldsymbol{x}}\rangle\langle \boldsymbol{g}, \boldsymbol{\mu}\rangle + \langle \boldsymbol{g}, \boldsymbol{\mu}\rangle^2 \middle| \, \|\bar{\boldsymbol{x}}\|_2 \geq R\right) \delta_1 \\
&\leq \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) + \delta_1\, \mathbb{E}\left((\langle \boldsymbol{g}, \bar{\boldsymbol{x}}\rangle)^2 \middle| \, \|\bar{\boldsymbol{x}}\|_2 \geq R\right) + \delta_1 \;\cdot 2\langle \boldsymbol{g}, \boldsymbol{\mu}\rangle \mathbb{E}\left(\langle \boldsymbol{g}, \bar{\boldsymbol{x}}\rangle \middle| \, \|\bar{\boldsymbol{x}}\|_2 \geq R\right) + \delta_1\, (\langle \boldsymbol{g}, \boldsymbol{\mu}\rangle)^2 \\
&\leq \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) + \delta_1\, \mathbb{E}\left(\|\boldsymbol{g}\|_1^2\, \|\bar{\boldsymbol{x}}\|_\infty^2 \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) + \delta_1 \;\cdot 2\langle \boldsymbol{g}, \boldsymbol{\mu}\rangle \mathbb{E}\left(\langle \|g\|_1\, \|\bar{\boldsymbol{x}}\|_\infty\rangle \middle| \, \|\bar{\boldsymbol{x}}\|_\infty \geq R\right) + \delta_1\, (\langle \boldsymbol{g}, \boldsymbol{\mu}\rangle)^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{by Hölder} \\
&\leq \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) + \delta_1 \left(4B^2\, R + 4B\mu\, B\, R + (B\, \mu)^2\right) = \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) + \delta_1\, 4B^2(R + \mu R + \mu^2) \\
&\leq \mathcal{L}(\tilde{h}_{\boldsymbol{w}}) + d\, e^{-R^2/\sigma^2}\, 4\, B^2(R + \mu)^2 \qquad\qquad\qquad\qquad \text{sub Gaussian}
\end{aligned}$$

We bound $\mathcal{L}(\tilde{h}_{\boldsymbol{w}})$ using Theorem D.3 and the fact that the predictor we output is an interpolator. Namely:

$$\mathcal{L}(\tilde{h}_{\boldsymbol{w}}) \leq K \left(H\, \log^3 N\, \mathcal{R}_N^2(\tilde{\mathcal{H}}) + \frac{b \log(1/\delta)}{N}\right)$$

We bound the Rademacher complexity of the hypothesis class and bounding the value of the loss. The Rademacher complexity of $\tilde{\mathcal{H}}$ is upper bounded by that of the original class, since the clipping procedure only limits the expressivity of the class:

$$\mathcal{R}_N(\tilde{\mathcal{H}}) \leq \|\boldsymbol{x}\|_\infty\, B\sqrt{\frac{2\log d}{N}} \leq (R + \mu)B\sqrt{\frac{2\log d}{N}}\,,$$

which gives us:

$$\mathcal{L}(\tilde{h}_{\boldsymbol{w}}) \leq K \left(H\, \log^3 N\, \left((R + \mu)B\sqrt{\frac{\log d}{N}}\right)^2 + \frac{b \log(1/\delta)}{N}\right)$$

Putting these pieces together, and using that square loss is 1-smooth, we have:

$$\begin{aligned}
\mathcal{L}(h_{\boldsymbol{w}}) &\leq \mathcal{O}\left(\frac{(R+\mu)^2\, B^2\, \log d}{N/\log^3 N} + \frac{B^2(R+\mu)^2\log(1/\delta)}{N}\right) + d\, e^{-R^2/\sigma^2} 4B^2\, (R+\mu)^2 && (50)\\
&\leq \mathcal{O}\left(B^2(R+\mu)^2 \cdot \left(\frac{\log(d/\delta)}{N/\log^3 N} + d\, e^{-R^2/\sigma^2}\right)\right) \qquad \zeta \coloneqq 1 + \frac{\sigma^2}{\mu^2} && (51)\\
&= \mathcal{O}\left(\zeta^2 \frac{(R+\mu)^2}{\sigma^2}\left(\frac{k\,\mathbb{E}\big(y^2\big)\, \log(d/\delta)}{N/\log^3 N} + d\, e^{-R^2/\sigma^2}\right)\right) && (52)\\
&= \mathcal{O}\left(\zeta^2 \left(\frac{\mu + \sigma\sqrt{\log(d/\delta_1)}}{\sigma}\right)^2 \left(\frac{k\,\mathbb{E}\big(y^2\big)\, \log(d/\delta)}{N/\log^3 N} + \delta_1\right)\right) && (53)\\
&= \mathcal{O}\left(\zeta^2 \left(\frac{\mu + \sigma\sqrt{\log \frac{d\,N}{k\log(d/\delta)k\,\mathbb{E}(y^2)\, \log(d/\delta)\log^3 N}}}{\sigma}\right)^2 \left(\frac{k\,\mathbb{E}\big(y^2\big)\, \log(d/\delta)}{N/\log^3 N}\right)\right) && (54)\\
&\leq \mathcal{O}\left(\zeta^2 \left(\frac{\mu + \sigma\sqrt{\log d\,N}}{\sigma}\right)^2 \left(\frac{k\,\mathbb{E}\big(y^2\big)\, \log(d/\delta)}{N/\log^3 N}\right)\right) && (55)\\
&= \mathbb{E}\big(y^2\big)\,\mathcal{O}\left(\frac{\zeta^2 k\,\mathrm{polylog}(dN/\delta)}{N}\right) && (56)
\end{aligned}$$

Equation 54 follows from setting δ_1 to be the same size as the first term, $k\,\mathbb{E}\big(y^2\big)\, \log(d/\delta)/(N/\log^3 N)$. Note that when $\mu = O(\sigma)$, the additive part of the second term is a constant. When we expand R to include all but δ_1 fraction of the points, we incur the extra multiplicative $\log(d/\delta_1)$ factor in the generalization error.

□

□

E. Experiment with Random initialization

In this section, we show that using random initialization instead of uniform ones initilization does not affect the observed qualitative behaviour. Specifically, we see that the test loss still improves with the step size.

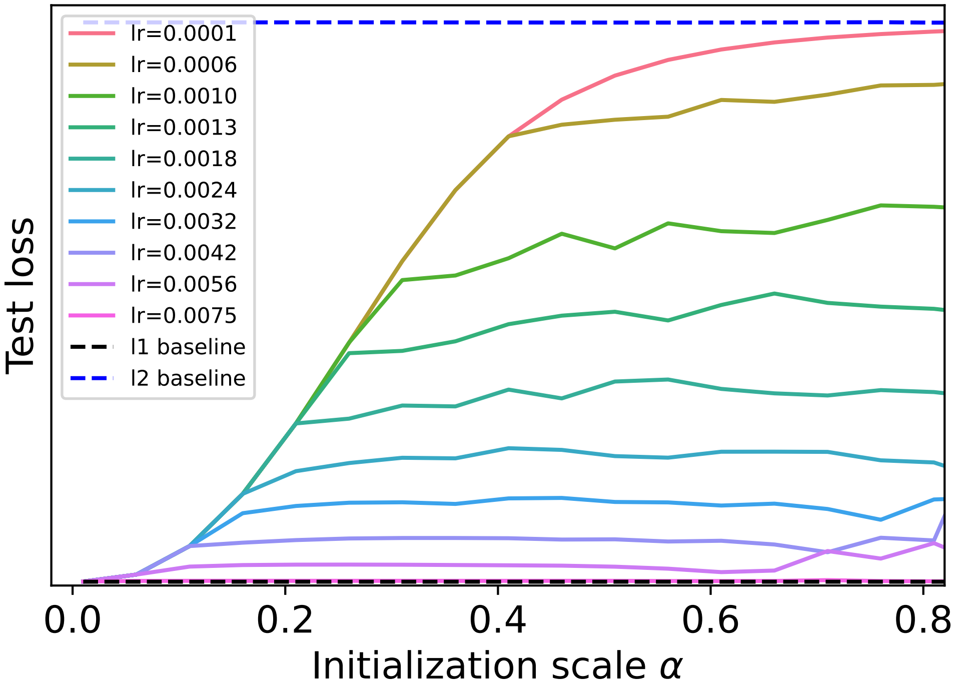


Figure 6. The test loss of GD solution vs. the initialization scale α in the sparse regression problem described in Section 7. This is the same setting as in Fig. 1, only here we used random normal initialization. We observe the same qualitative behaviour as observed in Fig. 1. Specifically, we observe that for small step size, the test loss transitions from the ℓ_1 baseline to the ℓ_2 baseline as the initialization scale α increases, as expected from (Woodworth et al., 2020). However, we see that using larger step sizes reduces the error significantly. In fact, **for large step size the test loss is close to the ℓ_1 baseline regardless of the initialization**.