Conflicting Biases at the Edge of Stability: Norm versus Sharpness Regularization

Anonymous authors
Paper under double-blind review

Abstract

A widely believed explanation for the remarkable generalization capacities of overparameterized neural networks is that the optimization algorithms used for training induce an implicit bias towards benign solutions. To grasp this theoretically, recent works examine gradient descent and its variants in simplified training settings, often assuming vanishing learning rates. These studies reveal various forms of implicit regularization, such as norm minimizing parameters in regression and max-margin solutions in classification. Concurrent findings show that moderate to large learning rates exceeding standard stability thresholds lead to faster, albeit oscillatory, convergence in the so-called Edge-of-Stability regime, and induce an implicit bias towards minima of low sharpness (norm of training loss Hessian).
In this work, we argue that a comprehensive understanding of the generalization performance of gradient descent requires analyzing the interaction between these various forms of implicit regularization. We empirically demonstrate that the learning rate balances between low parameter norm and low sharpness of the trained model. We furthermore prove for diagonal linear networks trained on a simple regression task that neither implicit bias alone minimizes the generalization error. These findings demonstrate that focusing on a single implicit bias is insufficient to explain good generalization, and they motivate a broader view of implicit regularization that captures the dynamic trade-off between norm and sharpness induced by non-negligible learning rates.

1 Introduction

First-order methods such as *gradient descent (GD)* are at the core of optimization in deep learning, used to train models which generalize remarkably well to unseen data while being able to interpolate random noise (Zhang et al., 2021). A widely believed explanation for this impressive generalization ability on meaningful data is that GD and its variants exhibit an implicit bias — a tendency of the optimization algorithm to favor well-structured solutions.

When rigorously characterizing this implicit bias for full-batch GD, recent works often consider small learning rates or even the corresponding *gradient flow (GF)*, which is GD's continuous time limit under infinitely small learning rates. For classification tasks, GF has been shown to favor max-margin solutions (Soudry et al., 2018). In regression tasks using diagonal linear networks initialized near the origin, GF induces an implicit bias toward parameters of minimal norm (Woodworth et al., 2020). In practice, however, optimization relies on finite learning rates that are bounded away from zero, raising the question of whether these explanations remain valid also in such scenarios.

At the same time, it was observed for standard architectures that full-batch GD can minimize the training loss even with learning rates that are larger than what classical optimization theory would require (Jastrzębski et al., 2019; Cohen et al., 2021). To be more precise, when

optimizing a (locally) L-smooth[1] loss function $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ via full-batch GD, i.e.,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_k) \tag{1}$$

with fixed learning rate $\eta > 0$, it is well-known (Bubeck et al., 2015) that

$$\mathcal{L}(\boldsymbol{\theta}_{k+1}) \leq \mathcal{L}(\boldsymbol{\theta}_k) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla \mathcal{L}(\boldsymbol{\theta}_k)\|_2^2, \tag{2}$$

which means that monotonic decrease of GD is only ensured for $\eta < 2/L$. This suggests for general twice differentiable $\mathcal{L}$ that GD with learning rate η becomes unstable if $\|\nabla^2 \mathcal{L}(\boldsymbol{\theta}_k)\| > 2/\eta$. As a result, the training loss $\mathcal{L}$ is not to be expected to decrease in these sharp regions of the loss landscape.
When training neural networks via GD with fixed $\eta > 0$, it was however confirmed in extensive simulations (Cohen et al., 2021) that the *sharpness* $S_{\mathcal{L}}(\boldsymbol{\theta}_k) = \|\nabla^2 \mathcal{L}(\boldsymbol{\theta}_k)\|$ of the training loss $\mathcal{L}$ at iterate $\boldsymbol{\theta}_k$ increases along the GD trajectory until it exceeds the critical value $2/\eta$ at some $\boldsymbol{\theta}_{k_0}$. For $k > k_0$, the sharpness of the iterates starts hovering around and slightly above this value (see Figure 13 for illustration). In this phase, the loss decreases non-monotonically and faster than when using adaptive learning rates that stay in the stable regime $\eta_k < 2/S_{\mathcal{L}}(\boldsymbol{\theta}_k)$. Accordingly, the authors dubbed the phases $k < k_0$ *"Progressive Sharpening"* and the phase $k > k_0$ *"Edge of Stability (EoS)"*. In practice, convergence in the EoS regime is attractive due to the fast average loss decay. It was even suggested that large learning rates and thus EoS might be necessary to learn certain functions (Ahn et al., 2023). More importantly, recent works on EoS showed that large learning rates induce an implicit bias of GD towards minimizers with low sharpness (Ahn et al., 2022). Indeed, for fixed $\eta > 0$ and twice differentiable $\mathcal{L}$, GD can only converge towards stationary points $\boldsymbol{\theta}_\star$ with $S_{\mathcal{L}}(\boldsymbol{\theta}_\star) < 2/\eta$.

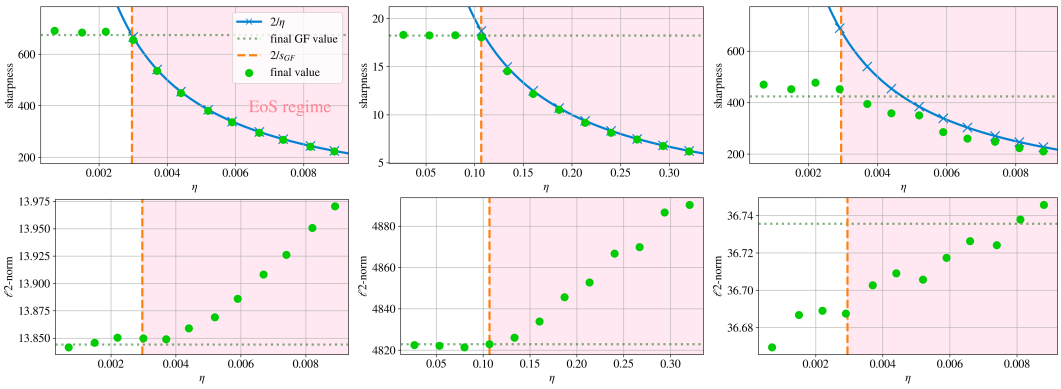
In summary, these different lines of works suggest that GD in (1) exhibits at least two distinct but entangled forms of implicit bias; one stemming from the underlying GF $\boldsymbol{\theta}' = -\nabla \mathcal{L}(\boldsymbol{\theta})$ and one induced by its learning rate η. To fully understand the success of GD-based training via implicit bias, it is therefore insufficient to analyze each bias in isolation. Instead, it is essential to understand the trade-off between various biases and answer the central question: How do different implicit biases interact when GD is used for training neural networks? A better understanding of this interaction may ultimately lead to more principled choices in the design of training algorithms and hyperparameters.

1.1 Contribution

Our work focuses on the two previously mentioned biases: the sharpness regularization induced by large learning rates (Ahn et al., 2022) and the norm-regularization induced by vanishing learning rates due to the compositional structure of *feedforward networks (FFNs)* (Woodworth et al., 2020; Chou et al., 2023). Our contribution consists of three major points:

(i) **Implicit bias trade-off in training:** Across a wide range of settings, we empirically demonstrate that at the end of training there is a trade-off between small norm of the parameters and small sharpness of the training loss. This trade-off is controlled by the learning rate. When comparing the final solutions across a range of learning rates (see Section 2), we observe a sharp phase transition at a data- and model-dependent critical learning rate η_c. Below η_c, both the norm and sharpness remain nearly constant. Above η_c, increasing the learning rate leads to an overall trend of increasing norm and decreasing sharpness. *We emphasize that this phase transition occurs when comparing final GD iterates over the choice of learning rate, and does not correspond to the transition from Progressive Sharpening to EoS observed for fixed learning rate η over the iterates $\boldsymbol{\theta}_k$ of GD (Cohen et al., 2021). To highlight that our observations do not depend on the specific choice of norm, we present different norms in Figures 1 – 3, and compare different norm choices in Appendix H.9.*

[1] A differentiable function $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ is called L-smooth if $\nabla \mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}^p$ is L-Lipschitz. If $\mathcal{L}$ is twice differentiable, this is equivalent to the Hessian having operator norm $\|\nabla^2 \mathcal{L}\|$ bounded by L.



(a) FCN on CIFAR-10-5k (b) FCN on MNIST (c) ViT on MNIST-5k[2]

Figure 1: A critical learning rate $\eta_c = 2/s_{GF}$ marks a sharp phase transition between two regimes, a flow-aligned regime, where solutions match gradient flow in sharpness and norm, and an Edge-of-Stability (EoS) regime, where sharpness decreases while the ℓ_2-norm increases, indicating a trade-off between low sharpness and small norm. Here, three models are trained with full-batch gradient descent with varying learning rates. This behavior is observed consistently across a wide range of experiments, see Section 2.1.

(ii) **Impact on generalization:** Remarkably, low generalization error often does not align with either extreme of the learning rate spectrum and never aligns with minimal norm. In some settings, the test error follows a U-shaped curve, with the best generalization occurring at intermediate learning rates where norm and sharpness biases are balanced, see Section 2.2. The learning rate can be interpreted as a regularization hyperparameter that controls generalization capacity of the resulting model, cf. Andriushchenko et al. (2023a).

(iii) **Theoretical analysis of a simple model:** Restricting ourselves to the strongly simplified setting of training a shallow diagonal linear network with shared weights for regression on a single data point with square loss, in Section 3 we analyze how the norm- and sharpness minimizers on the solution manifold $\mathcal{L} = 0$ are related and how they compare in terms of generalization. In fact, we provide scenarios where the lowest expected generalization error is attained by neither of them and the learning rate controls the generalization performance of the GD solution. Serving as a basic counterexample in which single biases do not generalize optimally, this supports our conjecture that the generalization behavior of neural networks can not be explained by a single implicit bias of GD. We analyze a comparably simple classification setting in Appendix F.

To illustrate the effect of bias entanglement and the influence of the learning rate on the resulting trade-off right away, we present a prototypical experiment in Figure 1.

1.2 Notation and outline

In the remainder of the paper, we denote vectors $\mathbf{x} \in \mathbb{R}^d$ and matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$ by bold lower and upper case letters, and abbreviate $[n] := \{1, \dots, n\}$. For vectors/matrices of ones and zeros we write $\mathbf{1}$ and $\mathbf{0}$, where the respective dimensions are clear from the context. The sharpness of a twice differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ at a point $\boldsymbol{\theta}$ is defined as

$$S_f(\boldsymbol{\theta}) := \|\nabla^2 f(\boldsymbol{\theta})\| = \max_{\lambda \in \sigma(\nabla^2 f(\boldsymbol{\theta}))} |\lambda|,$$

[2]The properties shown in the two left columns correspond to fully-connected FFNs (FCNs) trained with mean squared error (MSE), while the Vision Transformer (ViT) in the right column uses cross-entropy loss. We discuss the resulting qualitative differences between both losses in Appendix H.4.

where $\|\cdot\|$ denotes the operator norm and $\sigma(\mathbf{M})$ the spectrum of a matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$. By $\odot$ we denote the (entry-wise) Hadamard product between two vectors/matrices and write $\mathbf{z}^{\odot k} = \mathbf{z} \odot \cdots \odot \mathbf{z}$ for the k-th Hadamard power. The support of a vector $\mathbf{z} \in \mathbb{R}^d$ is denoted by $\operatorname{supp}(\mathbf{z}) = \{i \in [d] \colon z_i \neq 0\}$ and the diagonal matrix with diagonal $\mathbf{z}$ by $\mathbf{D}_\mathbf{z} \in \mathbb{R}^{d\times d}$. For any index set $I \subset [d]$ and $\mathbf{z} \in \mathbb{R}^d$, we furthermore write $\mathbf{z}|_I \in \mathbb{R}^d$ for the vector that is zero on I^c and $\mathbf{z}$ on I.

Our numerical results are presented in Section 2. To shed some light on the observed phenomena, we analyze a simple regression model in Section 3. Finally, we conclude in Section 4 with a discussion of our results. All proofs and further insights are deferred to the appendix.

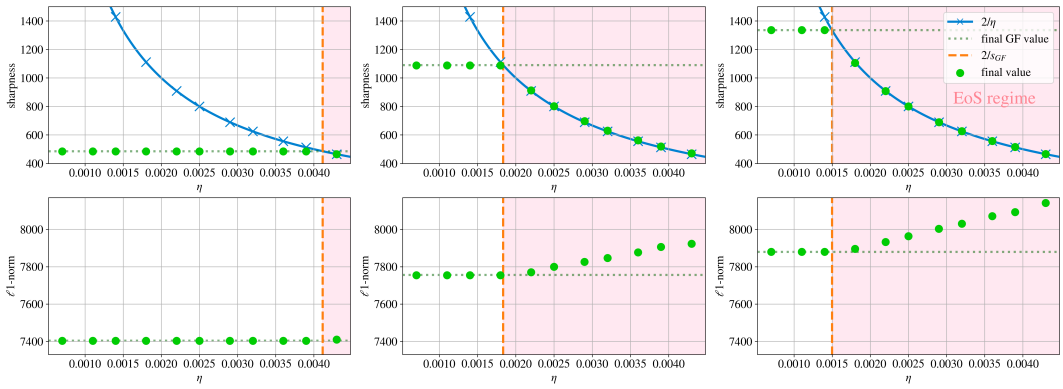
1.3 Related works

Before presenting our results in detail, let us review the current state of the art on analyzing the implicit bias of GF and GD, on EoS, which represent the two forms of regularization we study. Thereafter we discuss the question how generalization relates to each implicit bias. This section serves as a synopsis of Appendix A.

Implicit bias of GF. To understand the remarkable generalization properties of unregularized gradient-based learning procedures for deep neural networks (Zhang et al., 2021; Belkin et al., 2019), a recent line of works has been analyzing the implicit bias of GD towards parsimoniously structured solutions in simplified settings such as linear classification (Soudry et al., 2018), matrix factorization (Gunasekar et al., 2017), training linear networks (Geyer et al., 2020), training two-layer networks for classification (Chizat & Bach, 2020), and training linear diagonal networks for regression (Vaskevicius et al., 2019). All of these results analyze GD with small or vanishing learning rate, i.e., the implicit biases identified therein can be ascribed to the underlying GF dynamics. It is worth noting that there are other mechanisms inducing algorithmic regularization such as label noise (Pesme et al., 2021) or weight normalization (Chou et al., 2024b).

Edge of Stability. Whereas most of the above studies rely on vanishing learning rates, results by Cohen et al. (2021) on EoS suggest that GD under finite, realistic learning rates behaves notably differently from its infinitesimal limit. Recently, a thorough analysis of EoS has been provided for training linear classifiers (Wu et al., 2024) and shallow near-homogeneous networks (Cai et al., 2024) on the logistic loss via GD. In particular, GD with fixed learning rate $\eta > 0$ can only converge to sufficiently flat minima (Ahn et al., 2022), i.e., stationary points $\boldsymbol{\theta}_\star$ of a loss $\mathcal{L}$ with bounded sharpness $S_\mathcal{L}(\boldsymbol{\theta}_\star) < 2/\eta$. Note that EoS was first observed for *stochastic gradient descent (SGD)* (Wu et al., 2018), for which the analogous sharpness bounds also depend on the batch size (Wu et al., 2022). Ghosh et al. (2025) show that large learning rates in deep linear networks induce a so-called beyond–EoS regime in which GD oscillates stabely around the minimal sharpness solution.

Generalization and sharpness. In the past, various notions of sharpness have been studied in connection to generalization. The idea that flat minima benefit generalization dates back to Wolpert (1993). Since then, many authors have conjectured that flatter solutions should generalize better. Nevertheless, the relationship between flatness and generalization remains disputed. Studies have found little correlation between sharpness and generalization performance (Kaur et al., 2023), even when using scaling invariant sharpness measures like *adaptive sharpness* (Kwon et al., 2021). On the contrary, in various cases the correlation is negative, i.e., sharper minima generalize better. Notably, one of these works by Andriushchenko et al. (2023a) observe correlation of generalization with parameters such as the learning rate, which agrees with the herein presented idea of an implicit bias trade-off that is governed by hyperparameters of GD.
We emphasize that with the present work we do not contribute to resolving the question of which notion of sharpness (Tahmasebi et al., 2024) might be most accurate as a measure of generalization. In fact, we restrict ourselves to the so-called worst-case sharpness $S_\mathcal{L}$ defined as the operator norm of the loss Hessian since this version of sharpness is provably regularized by GD with large learning rates (Ahn et al., 2022).



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.01$ (c) $\varepsilon = 0.001$

Figure 2: Sharpness and ℓ_1-norm of final FCN models with tanh activation trained via MSE loss on CIFAR-10-5k for three different loss thresholds ε. Axis scales are equal for all three instances. Each plot illustrates a sharp regime transition as the learning rate crosses the critical threshold $\eta_c \approx 2/s^{\varepsilon}_{\mathrm{GF}}$, shifting from the flow-aligned regime with nearly constant sharpness and norm to the EoS regime where sharpness decreases and the norm increases.

Generalization and ℓ_1-norm. A possible explanation for the occasionally observed correlation between flatness and generalization can be deduced from Ding et al. (2024). The authors show for (overparameterized) matrix regression that sharpness and nuclear norm (ℓ_1-norm on the spectrum) minimizers lie close to each other. In view of the well-established theory of sparse resp. low-rank recovery via ℓ_1- resp. nuclear norm minimization (Foucart & Rauhut, 2013), good generalization of flat minima might just be consequence of flat minima lying close to nuclear norm minimizers, which provably generalize well in low-rank recovery. The observation that a single bias causes generalization might only stem from special situations in which several independent biases agree. This is also the case in scalar factorization Wang et al. (2022a, Appendix F.2). This point of view is supported by Wen et al. (2023) and aligns with our observations.

2 Conflicting biases

Across a wide range of training setups with varying architectures, activations, loss functions, and datasets, we consistently observe a trade-off between sharpness and norm of the final parameters as soon as the learning rate increases above a critical value. In Figure 1 we show examples of this transition, revealing two distinct regimes: The *flow-aligned regime* where both final sharpness and norm remain nearly constant with respect to the learning rate, and the *Edge-of-Stability (EoS) regime* where sharpness decreases hyperbolically and the ℓ_1-norm increases approximately linearly. For GD trained until loss ε the critical learning rate at which this phase transition occurs depends on the gradient flow solution and is approximately given by $\eta_c := 2/s^{\varepsilon}_{\mathrm{GF}}$. Here, $s^{\varepsilon}_{\mathrm{GF}} := \max_{t \le t_\varepsilon} S_{\mathcal{L}}(\boldsymbol{\theta}(t_\varepsilon))$ denotes the maximal sharpness of the GF solution $\boldsymbol{\theta}$ until time $t_\varepsilon := \inf\{t \colon \widetilde{\mathcal{L}}(\boldsymbol{\theta}(t)) \le \varepsilon\}$, see Figure 2. When ε is clear from the context, we just write s_{GF}. We emphasize that this regime transition occurs when comparing final GD iterates initialized identically over the choice of learning rate, and does not correspond to the transition from Progressive Sharpening to EoS at $t_\eta := \inf\{t \colon S_{\mathcal{L}}(\boldsymbol{\theta}_t) \ge 2/\eta\}$ observed for fixed learning rate η over the iterates $\boldsymbol{\theta}_k$ of GD (Cohen et al., 2021).

2.1 Systematic experimental analysis

To systematically investigate the trade-off between sharpness and norm minimization, we conduct experiments on standard vision datasets using both simple and moderately complex architectures. Since computing the sharpness during training involves estimating the largest

eigenvalue of the Hessian, which scales with both model and dataset size, we primarily use compact models to allow for evaluation across a broad range of learning rates.

Following the experimental setup of Cohen et al. (2021), our base configuration consists of a fully connected ReLU network with two dense layers with 200 hidden neurons each, trained on the first 5,000 training examples from both MNIST and CIFAR-10 (LeCun et al., 2010; Krizhevsky et al., 2014). These two datasets provide complementary complexity levels and help ensure that the observed effects are not specific to a single data distribution.

We train using full-batch gradient descent in order to cleanly isolate the fundamental trade-off between norm and sharpness bias driven by the learning rate η. This allows us to study the biases GD and GF induce without further confounding factors such as stochasticity or momentum. To ensure comparable convergence across settings, we train until we reach a fixed (training) loss threshold depending on the model.

Once we fix a setup, we use the same weight initialization across all learning rates to isolate the effect of the step size. The exact choice of the learning rate schedule, along with further experimental details, is available in Appendix G.

We perform a systematic investigation by varying the following core components of the training setup.

(i) **Dataset size.** When training on the full MNIST and CIFAR-10 dataset, the phase transition persists, see Appendix H.1.

(ii) **Architecture.** We vary the architecture of the fully-connected network (FCN), as well as extend the FCN to a convolutional neural network, a ResNet and a Vision Transformer (Lecun et al., 1998; He et al., 2016; Dosovitskiy et al., 2021), see Appendix H.2.

(iii) **Activation function.** We study ReLU and tanh activations. The phase transition occurs in both settings, see Appendix H.3.

(iv) **Loss function.** On most settings, we compare both cross-entropy loss (CE) and mean squared error (MSE). The phase transitions are similar though differences in the time evolution exist, see Appendix H.4.

(v) **Loss threshold.** For every experiment, we vary the loss threshold to which we train, cf. Figure 2 and Appendix H.5. Note that varying the loss threshold can be interpreted as early stopping.

(vi) **Initialization.** When varying the initialization, the properties of the GF solution s_{GF} are changed. Consequently, the transition between both regimes happens at a different learning rate, see Section H.6.

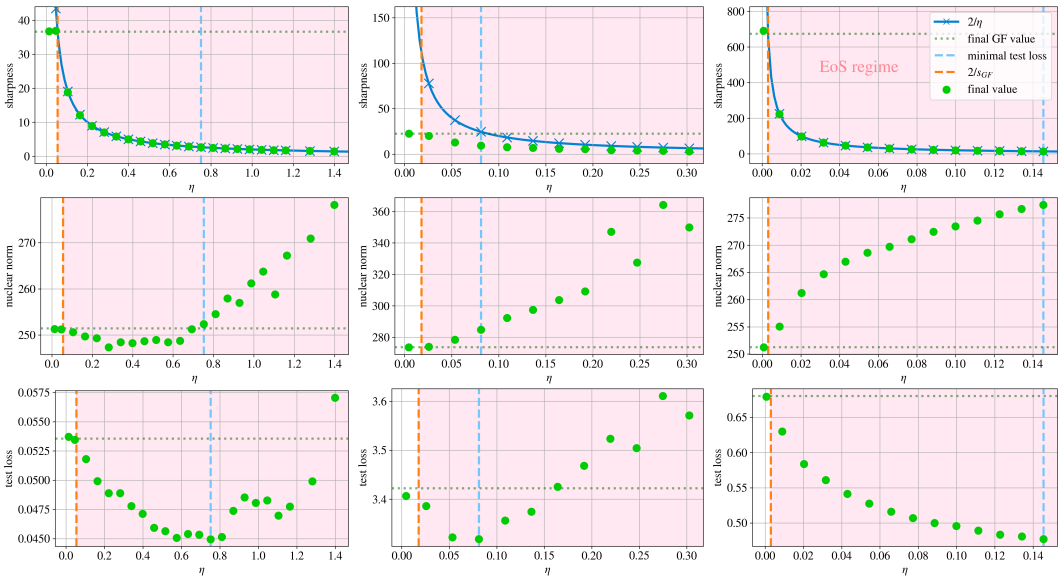
(vii) **Parametrization**. We train FCNs with varying widths in the μP and kernel parameterizations (Yang et al., 2022; Jacot et al., 2018) in Appendix H.7 where for μP we observe a certain width-independence of the spectral properties, cf. Noci et al. (2024).

Across all variations, we consistently observe the same trade-off between sharpness and norm, and the emergence of the flow-aligned and EoS regimes. Most figures showing these variations are deferred to Appendix I due to the page limit, along with further noteworthy observations from our experiments being noted in Appendix H.

2.2 Interpretation of the experiments

We now provide a high-level summary of our findings.

Flow-aligned regime. In the flow-aligned regime ($\eta < \eta_{\mathrm{c}}$), the behavior of GD closely mirrors that of continuous-time gradient flow. This regime is characterized by stable convergence of GD and minimal deviation from the gradient flow dynamics in terms of sharpness and norm. Intuitively, the sharpness of the solution in this regime stays within the stability limits set by the learning rate in (2), i.e., $S_{\mathcal{L}}(\theta_k) \leq 2/\eta$, allowing the discrete updates to track the continuous trajectory. However, we note that contrary to previous findings such



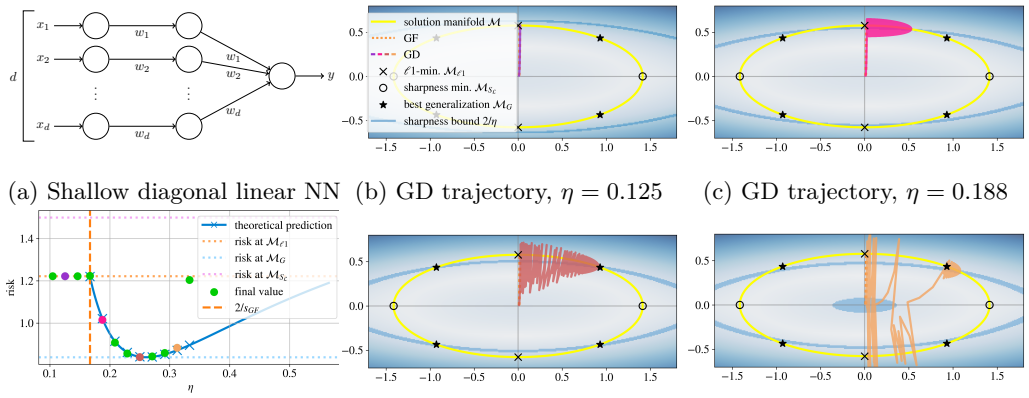
(a) MNIST-5k, MSE (b) CIFAR-10-5k, CE (c) CIFAR-10-5k, MSE

Figure 3: Final sharpness, nuclear norm, and test loss versus learning rate for three FCNs. On MNIST-5k with MSE loss (left), a clear U-shaped test loss indicates a trade-off between low sharpness and low nuclear norm. CIFAR-10-5k with CE loss (middle) shows a similar, though weaker trend. The best generalization typically occurs at intermediate learning rates where norm and sharpness biases are balanced. However, this is not universal — for instance CIFAR-10-5k with MSE loss (right) does not follow this pattern.

as by Arora et al. (2022), the absolute deviation from the GF trajectory is not necessarily negligible, see Appendix H.10. Nonetheless, the limits of GF and GD share nearly equal sharpness and norm values.

Edge-of-Stability regime. As the learning rate exceeds the critical threshold $\eta_c = 2/s_{\mathrm{GF}}$, the dynamics of GD enter the EoS regime. Here, training is governed by EoS (Cohen et al., 2021): while the loss continues to decrease on average over time, the decrease is no longer monotone and the curvature of the loss at the iterates (as measured by $S_{\mathcal{L}}$) fluctuates just above $2/\eta$. As GD is unable to converge to an overly sharp solution (cf. Theorem B.2), the iterates oscillate towards flatter regions. If training ends during or just after this EoS phase, the solution sharpness will therefore be near $2/\eta$.
In this regime, the sharpness $S_{\mathcal{L}}$ of the final network parameters thus decreases hyperbolically with the learning rate, closely tracking the function $\eta \mapsto 2/\eta$. At the same time, the norm of the final parameters increases. In some cases, there is an initial, temporary decrease in norm before the overarching trend of increasing norm and decreasing sharpness takes over at larger learning rates. We highlight that this increase in norm is not specific to the choice of norm: we observe the same qualitative trend for the ℓ_1, ℓ_2-norm and the nuclear norm, suggesting a general increase in model complexity as the learning rate increases, see Appendix H.9.

Generalization. When comparing the test error of the produced solutions, see Figure 3, we note that minimal norm solutions in the flow-aligned regime never lead to optimal generalization, i.e., if the test error decreases towards one extreme, it is always towards higher learning rates and increasing norm. In some of the cases we even observe a U-curve of the test error suggesting that GD generalizes best when norm and sharpness biases are well-balanced, see Figure 3. The learning rate can then be interpreted as a regularization hyperparameter that controls generalization capacity of the resulting model. This aligns with recent independent experiments by Andriushchenko et al. (2023a).



(a) Shallow diagonal linear NN (b) GD trajectory, $\eta = 0.125$ (c) GD trajectory, $\eta = 0.188$

(d) Generalization (e) GD trajectory, $\eta = 0.25$ (f) GD trajectory, $\eta = 0.312$

Figure 4: Two-layer diagonal linear model with weight sharing, shown in (4a). In (4b), (4c), and (4e) (4f), evolutions of weight iterates throughout training can be seen for different learning rates, where (4b) operates in the flow-aligned regime, and the others in the EoS regime. The background color map represents loss sharpness from low (white) to high (blue). The U-shaped generalization error is shown in (4d).

3 An elementary study of how implicit biases interact

To shed some light on the empirical observations of Section 2, we study the implicit biases of GF and GD in the EoS regime in a simple regression task and show that for this setup, the norm and sharpness minimizers of the interpolating manifold are distinct, and neither is sufficient for best generalization. Assuming a *single data point* $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, we train a shallow diagonal linear network with shared weights $\mathbf{w} \in \mathbb{R}^d$ and without bias

$$\phi_\mathbf{w} \colon \mathbb{R}^d \to \mathbb{R}, \qquad \phi_\mathbf{w}(\mathbf{z}) = \mathbf{w}^T \mathbf{D}_\mathbf{w} \mathbf{z}, \tag{3}$$

see Figure 4a, via the square loss $\mathcal{L}(y', y) = \frac{1}{2}(y' - y)^2$. The training objective is then

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\phi_\mathbf{w}(\mathbf{x}), y) = \min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) := \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \left(\langle \mathbf{w}^{\odot 2}, \mathbf{x} \rangle - y \right)^2, \tag{4}$$

where we overload the notation $\mathcal{L}$ for the sake of simplicity. Note that $\odot$ denotes the Hadamard product and $\mathbf{z}^{\odot k} = \mathbf{z} \odot \cdots \odot \mathbf{z}$ the k-th Hadamard power. We define the set of parameters of interpolating solutions $\phi_\mathbf{w}$ as

$$\mathcal{M} = \{\mathbf{w} \in \mathbb{R}^d \colon \mathcal{L}(\mathbf{w}) = 0\} \tag{5}$$

and note in the following lemma that $\mathcal{M}$ is a Riemannian manifold in general. We provide the proof in Appendix C.

Lemma 3.1. *For $\mathcal{L}$ as in (4), define $\mathcal{M}$ as in (5) and assume that $\mathcal{M} \neq \emptyset$. If $\mathbf{x} \in \mathbb{R}^d_{\neq 0}$ and $y \neq 0$, then $\mathcal{M}$ is a Riemannian manifold with tangent space $T_\mathbf{w}\mathcal{M} = (\mathbf{x} \odot \mathbf{w})^\perp$ at $\mathbf{w} \in \mathcal{M}$.*

While this training model is strongly simplistic, it allows us to explicitly compare the implicit biases induced by GF and by EoS, and to compute their generalization errors w.r.t. the realization of $(\mathbf{x}, y)$. Indeed, it is known that in this setting GF initialized at $\mathbf{w}_0 = \alpha \mathbf{1}$, for $\alpha > 0$ small, converges to an end-to-end model $\mathbf{w}_\star^{\odot 2}$ that approximately minimizes the ℓ_1-norm among all interpolating solutions (Chou et al., 2023), see Theorem B.1 in Appendix B.[3] Similarly, under mild technical conditions on $\mathcal{L}$, which are fulfilled in the present study, it is well-known for GD with learning rate $\eta > 0$ that for almost every initialization $\mathbf{w}_0 \in \mathbb{R}^d$ the iterates $\mathbf{w}_k$ can only converge to stationary points $\mathbf{w}_\infty$ with $S_\mathcal{L}(\mathbf{w}_\infty) \leq 2/\eta$ (Ahn et al., 2022), see Theorem B.2 in Appendix B. In consequence, GD is implicitly restricted to limits with low sharpness if η is chosen sufficiently large.

[3] In consequence, the network parameters $\mathbf{w}_\star$ minimize the squared ℓ_2-norm.

The following result now characterizes how the norm- and sharpness-minimizers of (4) relate. In particular, it illustrates that they are clearly distinct in general.

Proposition 3.2. *For $\mathbf{x} \in \mathbb{R}^d_{\neq 0}$ and $\mathcal{L}$ as in (4) with $\mathcal{M} \neq \emptyset$ as in (5), the following hold:*

(i) To have

$$\mathbf{w} \in \mathcal{M}_{\ell_1} := \arg\min_{\mathbf{z} \in \mathcal{M}} \|\mathbf{z}^{\odot 2}\|_1,$$

it is necessary that $\mathbf{x}|_{\mathrm{supp}(\mathbf{w})} = x_{\max} \cdot \mathbf{1}|_{\mathrm{supp}(\mathbf{w})}$, for $x_{\max} = \max_i |x_i|$.
If $\mathbf{x} \in \mathbb{R}^d_{>0}$, this condition is also sufficient. In particular, we have in this case that

$$\mathcal{M}_{\ell_1} = \left\{ \mathbf{w} \in \mathbb{R}^d \colon \|\mathbf{w}\|_2^2 = \frac{y}{x_{\max}} \text{ and } \mathrm{supp}(\mathbf{w}) \subset \arg\max_i x_i \right\}. \tag{6}$$

(ii) To have

$$\mathbf{w} \in \mathcal{M}_{S_\mathcal{L}} := \arg\min_{\mathbf{z} \in \mathcal{M}} S_\mathcal{L}(\mathbf{z}),$$

it is necessary that $\mathbf{x}|_{\mathrm{supp}(\mathbf{w})} = x_0 \cdot \mathbf{1}|_{\mathrm{supp}(\mathbf{w})}$, for some $x_0 \in \mathbb{R}$.
If $\mathbf{x} \in \mathbb{R}^d_{>0}$, it is necessary and sufficient that the previous condition holds with $x_0 = x_{\min} = \min_i x_i$. In particular, we have in this case that

$$\mathcal{M}_{S_\mathcal{L}} = \left\{ \mathbf{w} \in \mathbb{R}^d \colon \|\mathbf{w}\|_2^2 = \frac{y}{x_{\min}} \text{ and } \mathrm{supp}(\mathbf{w}) \subset \arg\min_i x_i \right\}. \tag{7}$$

Proof sketch: To derive the necessary conditions, we calculate Riemannian gradients and Hessians along $\mathcal{M}$ and use the respective first- and second-order necessary conditions. To derive the sufficient conditions and the explicit representations in (6) and (7), we construct simple minimizers based on canonical basis elements. The full proof is in Appendix D. □

Proposition 3.2 shows that, in general, the norm- and sharpness-minimizer on $\mathcal{M}$ do not agree. We mention that the assumption $\mathbf{x} \in \mathbb{R}^d_{\neq 0}$ is not restrictive since any zero coordinate of $\mathbf{x}$ can be removed by reducing the problem dimension. In view of Theorems B.1 and B.2, we see that depending on the learning rate, GD with initialization $\mathbf{w}_0 = \alpha\mathbf{1}$, for $\alpha > 0$ close to zero, is implicitly more biased to two disjoint sets. For $\eta \to 0$, the limit of stable GD will lie close to the set in (6); as η increases, the limit of unstable GD (as far as it exists) will lie close to the set in (7). For $d = 2$, the situation is illustrated in Figure 4. We further note that the restriction of Theorem B.1 to non-negative parameters is not limiting the analysis since (6) always contains such solutions, i.e., in our setting an ℓ_1-minimizer on $\mathcal{M} \cap \mathbb{R}^d_{\geq 0}$ is also a minimizer on $\mathcal{M}$.

Despite its simplicity, our toy model can reproduce the characteristic phase transitions of norm and sharpness (Figure 1) and the U-shaped generalization curve (Figure 3). For this, let us assume that the data follows a simple linear regression model with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $y = \langle \mathbf{1}, \mathbf{x} \rangle + \varepsilon$, for independent $\varepsilon \sim \mathcal{N}(0, 1)$. Then, the risk $\mathcal{R}$ under $\mathcal{L}$ can be computed explicitly and the best achievable generalization error of $\phi_\mathbf{w}$ trained via (4) can be identified, see Lemma E.1.

Assume we are given a generic draw of the single data point $(\mathbf{x}_0, y_0) \sim (\mathbf{x}, y)$ with $\mathbf{x}_0 \in \mathbb{R}^d_{\geq 0}$, i.e., we consider a draw $(\mathbf{x}_0, y_0)$ from the conditional distribution $p((\mathbf{x}, y) | \mathbf{x} \geq \mathbf{0})$.[4] Note that almost surely $\mathbf{x}_0$ will satisfy $|\mathrm{supp}(\mathbf{x}_0)| \geq 2$, and have a unique minimal entry $x_{\min}$ at index $k_{\min}$ and a unique maximal entry $x_{\max}$ at index $k_{\max}$ such that the sets in (6) and (7) consist of two points each which only differ by a sign.

On this model, GD with learning rate η will minimize $\mathcal{L}$ under constraints $S_\mathcal{L} \leq \frac{2}{\eta}$ due to its implicit sharpness regularization. We can now compare the limit of GD with initialization $\mathbf{w}_0 = \alpha\mathbf{1}$, for $\alpha > 0$ small, to three *idealized* training algorithms which, given input $(\mathbf{x}_0, y_0)$, output the weight vector $\mathbf{w} \in \mathbb{R}^d$ of an interpolating solution $\phi_\mathbf{w}$:

[4] In this discussion, $(\mathbf{x}_0, y_0)$ takes the role of the single data point $(\mathbf{x}, y)$ from before and we condition to non-negative data in order to apply Proposition 3.2. We examine removing the latter limitation in Section E.1.

(i) **Minimal norm:** $\mathcal{A}_{\ell_1}\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ with $\mathcal{A}_{\ell_1}(\mathbf{x}_0, y_0) = \sqrt{\frac{y_0}{x_{\max}}}\mathbf{e}_{k_{\max}}$. This corresponds to the solution computed by GD with vanishing learning rate.

(ii) **Minimal sharpness:** $\mathcal{A}_{S_\mathcal{L}}\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ with $\mathcal{A}_{S_\mathcal{L}}(\mathbf{x}_0, y_0) = \sqrt{\frac{y_0}{x_{\min}}}\mathbf{e}_{k_{\min}}$. This corresponds to the solution that would be computed by GD with extremely large learning rate if convergence still happened.

(iii) **Minimal generalization error:** $\mathcal{A}_{\text{opt}}\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ with $\mathcal{A}_{\text{opt}}(\mathbf{x}_0, y_0)$ returning a risk minimizer in $\mathcal{M}_G$ (best generalizing points in $\mathcal{M}$).

Figure 4 shows four snapshots of the training dynamics for growing η. Figure 4b reflects the situation where GD has no sharpness induced restrictions on $\mathcal{M}$ and converges to a minimizer in $\mathcal{M}_{\ell_1}$, i.e. the output of $\mathcal{A}_{\ell_1}$. As long as η is not too large (Figure 4c), the generalization minimizer fall inside the feasible set. Due to EoS, the model finds a solution with sharpness around $2/\eta$ yielding suboptimal generalization error, though risk improves over $\mathcal{M}_{\ell_1}$. For carefully tuned η, Figure 4e shows convergence of GD to a point close to the output of $\mathcal{A}_{\text{opt}}$. For too large η, the sharpness constraints exclude $\mathcal{M}_{\text{G}}$ and GD moves closer to $\mathcal{M}_{S_\mathcal{L}}$. As Figure 4d illustrates, our toy model exhibits the U-shaped generalization curve observed in various training simulations, and explains it by an interpolation between implicit norm- and sharpness biases.

We note that in this example both $\mathcal{M}_{\ell_1}$ and $\mathcal{M}_{S_\mathcal{L}}$ lead to suboptimal generalization with $\mathcal{R}(\mathcal{M}_{\ell_1}) < \mathcal{R}(\mathcal{M}_{S_\mathcal{L}})$. Due to its instability, GD already diverges for many η where the feasible set of the constrained optimization problem is non-empty, i.e., although there exist points on the solution manifold with sharpness $< 2/\eta$. Consequently, all convergent trajectories in the EoS regime achieve better generalization than $\mathcal{R}(\mathcal{M}_{\ell_1})$, although the sharpness minimizer induces a higher risk. This might be an explanation for why the U-shaped generalization curve is not always visible in our experiments.

We provide additional numerical experiments for the diagonal network in Appendix H.13. In particular, note that the GD limit is often close to a KKT point of a sharpness-restricted risk minimization on $\mathcal{M}$ (Figure 16 and Lemma E.1). In Appendix F, we analyze a comparably simplified classification model for which sharpness minimization leads to better generalization performance than norm-minimization.

4 Discussion

Our experiments suggest that a single implicit bias of gradient descent is not sufficient to explain the good generalization performance in deep learning. While solutions obtained with vanishing learning rates may have an implicit bias towards simple structures, the bias changes with increasing learning rate. This insight provides an explanation for the strong empirical influence of the learning rate on model performance. Our theoretical analysis further indicates that the learning rate balances between various implicit biases, and that good generalization performance is only reached by careful fine-tuning of such hyperparameters of GD. These insights from our simplified model open the door to a broader perspective on implicit regularization which accounts for the interaction between multiple biases shaped by the optimization dynamics. Future work extending our insights to additional known biases and more realistic optimizers (e.g., SGD, Adam) will be important to fully translate these insights into practical training settings.

4.1 Limitations

Our theoretical analysis is restricted to simple models due to the difficulty in explicitly characterizing the implicit biases of GD in more general setups. In combination with our empirical studies, it nevertheless provides consistent evidence for the observed phenomena. Our study is further limited by only considering full-batch gradient descent as well as two specific manifestations of implicit bias. Further empirical validation on other popular optimizers, network classes and datasets would be desirable.

Reproducibility Statement

The complete experimental methodology is described in detail in Appendix G, and all experiments are fully reproducible. Source code will be released upon acceptance and is also provided as part of the supplementary material. Proofs of the main statements are included in Appendix C and D, with additional theoretical results and their corresponding proofs in E and F.

Ethics Statement

The presented work on implicit regularization is foundational in nature. The theory part is not tied to an application and also uses a simplified model. The experiments utilize established architectures, algorithms, and datasets. We therefore do not identify any specific ethical issues arising from this work.

LLM Usage Statement

We used large language models (OpenAI's ChatGPT, Google's Gemini, Writefull) for editorial assistance such as grammar, spelling, and word choice. In addition, we used OpenAI's ChatGPT for limited coding support including plotting routines, assistance with bash scripts and resolving error messages. No substantive ideas, research contributions, or results were generated by AI tools.

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International conference on machine learning*, pp. 247–257. PMLR, 2022.

Kwangjun Ahn, Sebastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19540–19569. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/3e592c571de69a43d7a870ea89c7e33a-Paper-Conference.pdf`.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 639–668. PMLR, 2022. URL `https://proceedings.mlr.press/v162/andriushchenko22a.html`.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023a.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with large step sizes learns sparse features. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 903–925. PMLR, 2023b.

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

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024. PMLR, 17–23 Jul 2022.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pp. 468–477. PMLR, 2021.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. *Advances in Neural Information Processing Systems*, 37:71306–71351, 2024.

Dennis Chemnitz and Maximilian Engel. Characterizing dynamical stability of stochastic gradient descent in overparameterized learning. *arXiv preprint arXiv:2407.20209*, 2024.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *International Conference on Machine Learning*, pp. 4330–4391. PMLR, 2023.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pp. 1305–1338. PMLR, 2020.

Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: inducing sparsity via overparameterization. *Information and Inference: A Journal of the IMA*, 12(3):1437–1460, 2023.

Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis*, 68:101595, 2024a.

Hung-Hsu Chou, Holger Rauhut, and Rachel Ward. Robust implicit regularization via weight normalization. *Information and Inference: A Journal of the IMA*, 13(3):iaae022, 2024b.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=jh-rTtvkGeM`.

Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 2024.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=uAyElhYKxg`.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *CoRR*, abs/2010.01412, 2020. URL `https://arxiv.org/abs/2010.01412`.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, New York, NY, 2013.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022.

Khashayar Gatmiry, Zhiyuan Li, Sashank J Reddi, and Stefanie Jegelka. Simplicity bias via global convergence of sharpness minimization. *arXiv preprint arXiv:2410.16401*, 2024.

Kelly Geyer, Anastasios Kyrillidis, and Amir Kalev. Low-rank regularization and solution uniqueness in over-parameterized matrix sensing. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 930–940, 2020.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep linear networks beyond the edge of stability. *arXiv preprint arXiv:2502.20531*, 2025.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/file/01882513d5fa7c329e940dda99b12147-Paper.pdf.

Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1):1–42, 01 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1. URL https://doi.org/10.1162/neco.1997.9.1.1.

Tom Jacobs, Chao Zhou, and Rebekka Burkholz. Mirror, mirror of the flow: How does regularization shape implicit bias? *CoRR*, abs/2504.12883, 2025. doi: 10.48550/ARXIV.2504.12883. URL https://doi.org/10.48550/arXiv.2504.12883.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in neural information processing systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf. Read_Status: In Progress Read_Status_Date: 2023-10-05T07:33:41.285Z.

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amost Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkgEaj05t7.

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

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Simran Kaur, Jeremy Cohen, and Zachary Chase Lipton. On the maximum hessian eigenvalue and generalization. In *Proceedings on*, pp. 51–65. PMLR, 2023.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. https://www.cs.toronto.edu/~kriz/cifar.html, 2014.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–50. Springer, 2002.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd international conference on artificial intelligence and statistics*, pp. 888–896. PMLR, 2019.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35: 34689–34708, 2022.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018. ISBN 0262039400.

Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 102696–102743. Curran Associates, Inc., 2024.

Hristo Papazov, Scott Pesme, and Nicolas Flammarion. Leveraging continuous time to understand momentum when training diagonal linear networks. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3556–3564. PMLR, 2024. URL `https://proceedings.mlr.press/v238/papazov24a.html`.

Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29218–29230, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/f4661398cb1a3abd3ffe58600bf11322-Abstract.html`.

Carl Runge. Ueber die numerische Auflösung von Differentialgleichungen. *Mathematische Annalen*, 46(2):167–178, 1895. doi: 10.1007/BF01446807.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

Behrooz Tahmasebi, Ashkan Soleymani, Dara Bahri, Stefanie Jegelka, and Patrick Jaillet. A universal class of sharpness-aware minimization algorithms. *arXiv preprint arXiv:2406.03682*, 2024.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pp. 2972–2983, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the LASSO for quadratic parametrisation. In *Proceedings of Conference on Learning Theory*, volume 178, pp. 2127–2159. PMLR, 2022.

Shuyang Wang and Diego Klabjan. A mirror descent perspective of smoothed sign descent. In Silvia Chiappa and Sara Magliacane (eds.), *Conference on Uncertainty in Artificial Intelligence, Rio Othon Palace, Rio de Janeiro, Brazil, 21-25 July 2025*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4515–4542. PMLR, 2025. URL https://proceedings.mlr.press/v286/wang25h.html.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022a.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=thgItcQrJ4y.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1024–1035. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0354767c6386386be17cabe4fc59711b-Paper-Conference.pdf.

David H. Wolpert. Bayesian backpropagation over i-o functions rather than weights. In J. Cowan, G. Tesauro, and J. Alspector (eds.), *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993. URL https://proceedings.neurips.cc/paper_files/paper/1993/file/d4c2e4a3297fe25a71d030b67eb83bfc-Paper.pdf.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 5019–5073. PMLR, 2024.

Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, pp. 37656–37684. PMLR, 2023.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Lei Wu, Mingze Wang, and Weijie Su. The alignment property of sgd noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35: 4680–4693, 2022.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer, 2022. URL https://arxiv.org/abs/2203.03466.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115, 2021.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023.

Supplement to the paper "Conflicting Biases at the Edge of Stability: Norm versus Sharpness Regularization"

In this supplement, we provide additional numerical simulations and proofs that were skipped in the main paper.

Contents

A Related works — Extended discussion

We provide a more detailed review of the related literature here.

Implicit bias of GF. To understand the remarkable generalization properties of unregularized gradient-based learning procedures for deep neural networks (Zhang et al., 2021; Belkin et al., 2019), a recent line of works has been analyzing the implicit bias of GD towards parsimoniously structured solutions in simplified settings such as linear classification (Soudry et al., 2018; Ji & Telgarsky, 2019), matrix factorization (Gunasekar et al., 2017; Arora et al., 2019; Chou et al., 2024a), training linear networks (Geyer et al., 2020; Stöger & Soltanolkotabi, 2021), training two-layer networks for classification (Chizat & Bach, 2020; Frei et al., 2022), and training linear diagonal networks for regression (Vaskevicius et al., 2019; Woodworth et al., 2020; Azulay et al., 2021; Chou et al., 2023). All of these results analyze GD with small or vanishing learning rate, i.e., the implicit biases identified therein can be ascribed to the underlying GF dynamics.

Other types of implicit regularization of GD. It is worth noting that there are other mechanisms inducing algorithmic regularization such as label noise (Pesme et al., 2021; Vivien et al., 2022) or weight normalization (Chou et al., 2024b), momentum gradient descent (Papazov et al., 2024), smoothed sign descent (Wang & Klabjan, 2025) and explicit regularization into the mirror flow (Jacobs et al., 2025). In (Andriushchenko et al., 2023b; Even et al., 2023) an intriguing connection regarding implicit regularization induced by large step sizes coupled with SGD noise has been discussed. In particular, for shallow diagonal linear networks it has been shown that SGD with large learning rates implicitly regularizes certain parameter norms (Wu & Su, 2023). For a broader overview on the topic including further references we refer to the survey by Vardi (2023).

Edge of Stability. Whereas most of the above works rely on vanishing learning rates, results by Cohen et al. (2021) on EoS suggest that GD under finite, realistic learning rates behaves notably differently from its infinitesimal limit. In the past few years, subsequent works have started to theoretically analyze the EoS regime. It is noted in Ahn et al. (2022) that GD with fixed learning rate $\eta > 0$ can only converge to stationary points $\boldsymbol{\theta}_\star$ of a loss $\mathcal{L}$ if $S_\mathcal{L}(\boldsymbol{\theta}_\star) < 2/\eta$. In Chemnitz & Engel (2024), this stability criterion of stationary points has been generalized to SGD. Note that EoS was first observed for SGD (Wu et al., 2018), for which the analogous sharpness bounds also depend on the batch size (Wu et al., 2022). Arora et al. (2022) relate normalized GD on a loss $\mathcal{L}$ to GD on the modified loss $\sqrt{\mathcal{L}}$ and show that EoS occurs $\mathcal{O}(\eta)$-close to the manifold of interpolating solutions. Under various restrictive assumptions, progressive sharpening and EoS have been analyzed by Wang et al. (2022b); Chen & Bruna (2023); Zhu et al. (2023); Kreisler et al. (2023). Recently, a thorough analysis of EoS has been provided for training linear classifiers (Wu et al., 2024) and shallow near-homogeneous networks (Cai et al., 2024) on the logistic loss via GD. The authors show that large learning rates allow a loss decay of $\mathcal{O}(1/k^2)$ which exceeds the best known rates for vanilla GD from classical optimization. Cohen et al. (2021) extended their empirical study of EoS to adaptive GD-methods for which the stability criterion becomes more involved (Cohen et al., 2022). Finally, let us mention that applying early stopping to label noise SGD with small learning rate can also induce sharpness minimization and structural simplicity of the learned weights (Gatmiry et al., 2024). As opposed to our definition of sharpness, sometimes called *worst-case sharpness*, in the latter work sharpness is measured by the trace of $\nabla^2\mathcal{L}$ also known as *average-case sharpness*. Additionally, Ghosh et al. (2025) show that when deep linear networks are trained with very large learning rates, gradient descent operates in a so-called beyond–EoS regime characterized by sustained oscillations around the balanced minimum which is of minimum sharpness. In contrast, we only consider converged trajectories, not ones which are in stable oscillations. Finally, we highlight that for models with normalization layers, the sharpness scales inversely with the squared parameter norm (Li et al., 2020; Lyu et al., 2022). Although this corresponds to a different GD dynamics due to the explicit regularization, the resulting trade-off aligns with our main observation.

Sharpness and generalization. In the past, various notions of sharpness have been studied in connection to generalization. The idea that flat minima benefit generalization dates back to Wolpert (1993), who argued this from a minimal description length perspective. Later, Hochreiter & Schmidhuber (1994; 1997) proposed an algorithm designed to locate flat minima, defining them as "large regions of connected acceptable minima," where an acceptable minimum is any point with empirical mean squared error below a certain threshold. Notably, their formulation does not explicitly involve the Hessian. Following these early works, many authors have conjectured that flatter solutions should generalize better (Xing et al., 2018; Zhou et al., 2020; Park & Kim, 2022; Lyu et al., 2022). The prevailing intuition is that solutions lying in flatter regions of the loss landscape are more robust to perturbations (Keskar et al., 2017), which may contribute to improved generalization.

Inspired by this idea, sharpness-aware minimization (SAM) has been proposed by Foret et al. (2020) as an explicit regularization method that penalizes sharpness, successfully applied in improving model generalization on benchmark datasets such as CIFAR-10 and CIFAR-100. In Tahmasebi et al. (2024), SAM was extended to sharpness measures that are general functions of the (spectrum of the) Hessian of the loss. The general sharpness formulation presented therein encompasses various common notions of sharpness such as worst-case and average-case sharpness.

Despite these theoretical and empirical arguments, the relationship between flatness and generalization remains disputed (Andriushchenko & Flammarion, 2022). Studies have found little correlation between sharpness and generalization performance (Jiang et al., 2019; Kaur et al., 2023). Furthermore, a re-parametrization argument by Dinh et al. (2017) shows that sharpness measures such as $S_{\mathcal{L}}$ can be made arbitrarily large without affecting generalization, challenging the notion that flatness is a necessary condition for good performance. Even when using scaling invariant sharpness measures like *adaptive sharpness* (Kwon et al., 2021), the empirical studies performed by Andriushchenko et al. (2023a) show that there is no notable correlation between low sharpness and good generalization. On the contrary, in various cases the correlation is negative, i.e., sharper minima generalize better. What is most interesting about the latter work from our perspective, is that it observes correlation of generalization with parameters such as the learning rate, which agrees with the herein presented idea of an implicit bias trade-off that is governed by hyperparameters of GD.

Generalization and ℓ_1-norm. A possible explanation for the occasionally observed correlation between flatness and generalization can be deduced from Ding et al. (2024). Therein the authors show for (overparameterized) matrix factorization of $\mathbf{X}_\star \in \mathbb{R}^{d_1 \times d_2}$ via

$$\min_{\mathbf{U}\in\mathbb{R}^{d_1\times k},\mathbf{V}\in\mathbb{R}^{d_2\times k}} \|\mathbf{U}\mathbf{V}^T - \mathbf{X}_\star\|_F^2,$$

where $k \geq \operatorname{rank}(\mathbf{X}_\star)$ is arbitrarily large, that sharpness and nuclear norm (ℓ_1-norm on the spectrum) minimizers coincide. For (overparameterized) matrix regression

$$\min_{\mathbf{U}\in\mathbb{R}^{d_1\times k},\mathbf{V}\in\mathbb{R}^{d_2\times k}} \|\mathcal{A}(\mathbf{U}\mathbf{V}^T) - \mathbf{y}\|_2^2, \tag{8}$$

where $\mathbf{y} = \mathcal{A}(\mathbf{X}_\star) + \mathbf{e}$, for $\mathcal{A}\colon \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ and unknown noise $\mathbf{e} \in \mathbb{R}^m$, they relate the distance between sharpness and nuclear norm minimizers to how close the measurement operator $\mathcal{A}$ is to identity. Good generalization of a solution $(\hat{\mathbf{U}}, \hat{\mathbf{V}})$ of (8), i.e., $\hat{\mathbf{U}}\hat{\mathbf{V}}^T \approx \mathbf{X}_\star$, is then proved if $\mathcal{A}$ satisfies an appropriate *restricted isometry property (RIP)* for low-rank matrices. However, it is not really clear which of the two types of regularization explains the generalization. In view of the well-established theory of sparse resp. low-rank recovery via ℓ_1- resp. nuclear norm minimization (Foucart & Rauhut, 2013), one may assume in this specific setting that good generalization of flat minima is just a consequence of the fact that flat minima lie close to nuclear norm minimizers, which provably generalize well in low-rank recovery. The observation that a single bias causes generalization might only stem from special situations in which several independent biases agree. This is also the case in scalar factorization Wang et al. (2022a, Appendix F.2.), where the sharpness of a minimizer is equal to squared norm and the biases thus coincide. This point of view is supported by Wen et al. (2023) and aligns with our observations.

B Implicit norm and sharpness regularization

In this section, we recall two established results on implicit bias of GF and GD. In the setting of Section 3, it is known that GF converges to an end-to-end model $\mathbf{w}_\star^{\odot 2}$ that approximately minimizes a weighted ℓ_1-norm among all interpolating solutions $\phi_{\mathbf{w}}(\mathbf{x}) = y$ if initialized close to the origin (Chou et al., 2023) where the weights of the ℓ_1-norm depend on the chosen initialization. To avoid unnecessary technicalities, we formulate the result only for $\mathbf{w}_0 = \alpha\mathbf{1}$ which induces a bias towards the unweighted ℓ_1-norm.

Theorem B.1 (Implicit ℓ_1-bias of GF (Chou et al., 2023))**.** *Let $\mathcal{L}$ be defined as in (4) with $\mathcal{M}$ as in (5). Assume that $\mathcal{M} \cap \mathbb{R}^d_{\geq 0}$ is non-empty and GF is applied with $\mathbf{w}_0 = \alpha\mathbf{1}$, for $\alpha > 0$. Then, GF converges to $\mathbf{w}_\infty \in \mathbb{R}^d$ with*

$$\|\mathbf{w}_\infty^{\odot 2}\|_1 \leq \Big(\min_{\mathbf{w} \in \mathcal{M} \cap \mathbb{R}^d_{\geq 0}} \|\mathbf{w}^{\odot 2}\|_1 \Big) + \varepsilon(\alpha),$$

where $\varepsilon(\alpha) > 0$ satisfies $\varepsilon(\alpha) \searrow 0$, for $\alpha \to 0$.

The implicit sharpness regularization of GD for large learning rates can be deduced from the following result.

Theorem B.2 (Dynamic stability of GD (Ahn et al., 2022))**.** *Let $\eta > 0$ and $X \subset \mathbb{R}^p$. Let $\mathcal{L}$ be twice continuously differentiable such that the operator $F \colon \mathbb{R}^p \to \mathbb{R}^p$, $F(w) = w - \eta\nabla\mathcal{L}(w)$ satisfies that $F^{-1}(S)$ is a set of Lebesgue-measure zero, for any set $S \subset \mathbb{R}^p$ of measure zero. Assume furthermore that $\frac{1}{\eta}$ is not an eigenvalue of $\nabla^2\mathcal{L}(w_\star)$ for every stationary point $w_\star$ of $\mathcal{L}$. Let w_k be the iterates of GD with learning rate η. If $\|\nabla^2\mathcal{L}(w)\|_2 > 2/\eta$ for every $w \in X$, then there exists a zero Lebesgue measure set A_X such that*

- *either $w_0 \in A_X$*
- *or w_k does not converge to any $w \in X$.*

C Proof of Lemma 3.1

Lemma 3.1 is a special case of the following result for training diagonal linear L-layer networks with shared weights on a single data point. In this case, the loss $\mathcal{L}$ is given by

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}(\langle \mathbf{x}, \mathbf{w}^{\odot L} \rangle - y)^2. \tag{9}$$

Lemma C.1. *For $\mathcal{L}$ as in (9), define $\mathcal{M}$ as in (5). If $\mathbf{x} \in \mathbb{R}^n_{\neq 0}$ and $y \neq 0$, then $\mathcal{M}$ is a Riemannian manifold with tangent space $T_{\mathbf{w}}\mathcal{M} = (\mathbf{x} \odot \mathbf{w}^{\odot(L-1)})^\perp$ at $\mathbf{w} \in \mathcal{M}$.*

Proof. Note that $\mathbf{w} \in \mathcal{M}$ is equivalent to

$$h(\mathbf{w}) := \langle \mathbf{x}, \mathbf{w}^{\odot L} \rangle - y = 0,$$

where $h \colon \mathbb{R}^d \to \mathbb{R}$. Since $Dh(\mathbf{w}) = L(\mathbf{x} \odot \mathbf{w}^{\odot L-1})^T$ and $\mathbf{w} \neq \mathbf{0}$ for any $\mathbf{w} \in \mathcal{M}$ due to $y \neq 0$, we have that $\operatorname{rank}(Dh(\mathbf{w})) = 1$ for all $\mathbf{w} \in \mathcal{M}$. Hence, $\mathcal{M}$ is a $(d-1)$-dimensional submanifold in $\mathbb{R}^d$ with tangent spaces

$$T_{\mathbf{w}}\mathcal{M} = \ker(Dh(\mathbf{w})) = (\mathbf{x} \odot \mathbf{w}^{L-1})^\perp,$$

e.g., see Boumal (2023). Smoothness of the manifold follows by equipping $T_{\mathbf{w}}\mathcal{M}$ with the Euclidean metric of $\mathbb{R}^d$. □

D Proof of Proposition 3.2

Before we prove Proposition 3.2, we note that the ℓ_1-norm of $\mathbf{w}^{\odot 2}$ can be written as

$$\|\mathbf{w}^{\odot 2}\|_1 = \|\mathbf{w}\|_2^2 \tag{10}$$

and that the sharpness $S_{\mathcal{L}}(\mathbf{w})$ of $\mathcal{L}$ at $\mathbf{w}$ satisfies

$$S_{\mathcal{L}}(\mathbf{w}) = 4\|\mathbf{x} \odot \mathbf{w}\|_2^2, \tag{11}$$

for any $\mathbf{w} \in \mathcal{M}$, where we used that

$$\nabla^2 \mathcal{L}(\mathbf{w}) = \mathbf{D}_{2(\langle \mathbf{x}, \mathbf{w}^{\odot 2}\rangle - y)\cdot \mathbf{x}} + 4(\mathbf{x} \odot \mathbf{w})(\mathbf{x} \odot \mathbf{w})^T.$$

The necessary conditions of Proposition 3.2 are proven in the following lemma.

Lemma D.1. *For $\mathbf{x} \in \mathbb{R}^d_{\neq 0}$ and $\mathcal{L}$ as in (4) with $\mathcal{M}$ as in (5), the following hold:*

(i) To have

$$\mathbf{w} \in \arg\min_{\mathbf{z} \in \mathcal{M}} \|\mathbf{z}^{\odot 2}\|_1,$$

it is necessary that $\mathbf{x}|_{\mathrm{supp}(\mathbf{w})} = x_0 \cdot \mathbf{1}|_{\mathrm{supp}(\mathbf{w})}$, for $x_0 = \max_i |x_i|$.

(ii) To have

$$\mathbf{w} \in \arg\min_{\mathbf{z} \in \mathcal{M}} S_{\mathcal{L}}(\mathbf{z}),$$

it is necessary that $\mathbf{x}|_{\mathrm{supp}(\mathbf{w})} = x_0 \cdot \mathbf{1}|_{\mathrm{supp}(\mathbf{w})}$, for some $x_0 \in \mathbb{R}$. Furthermore, if $\mathbf{x} \in \mathbb{R}^d_{>0}$, it is additionally necessary that $x_0 = \min_i x_i$.

Proof. In the proof we compute the Riemannian gradient $\mathrm{grad} f$ and the Riemannian Hessian $\mathrm{Hess} f$ of a function f on $\mathcal{M}$. Note that

$$\mathrm{grad} f(\mathbf{w}) = \mathbb{P}_{T_{\mathbf{w}}\mathcal{M}} \nabla f(\mathbf{w})$$

and

$$[\mathrm{Hess} f(\mathbf{w})](\mathbf{u}) = \mathbb{P}_{T_{\mathbf{w}}\mathcal{M}}([\nabla \mathrm{grad} f(\mathbf{w})](\mathbf{u})),$$

for any $\mathbf{w} \in \mathcal{M}$ and $\mathbf{u} \in T_{\mathbf{w}}\mathcal{M}$, where $\mathbb{P}_U$ denotes the orthogonal projection onto the linear subspace $U \subset \mathbb{R}^d$ (Boumal, 2023).

We begin with (i). Define $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$ and note that $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}^{\odot 2}\|_1$ by (10). Hence,

$$\mathrm{grad} f(\mathbf{w}) = \mathbb{P}_{T_{\mathbf{w}}\mathcal{M}} \nabla f(\mathbf{w}) = \mathbf{w} - \frac{1}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \mathbf{D}_{\mathbf{x}} \mathbf{w}\mathbf{w}^T \mathbf{D}_{\mathbf{x}} \cdot \mathbf{w}.$$

To have $\mathrm{grad} f(\mathbf{w}) = \mathbf{0}$, $\mathbf{w}$ has to be an eigenvector of $\mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}}$ with eigenvalue $\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2$ which is equivalent to $\mathbf{x}|_{\mathrm{supp}(\mathbf{w})} = x_0 \cdot \mathbf{1}|_{\mathrm{supp}(\mathbf{w})}$, for some $x_0 \in \mathbb{R}$. This is the first necessary condition.

Now define $G(\mathbf{w}) = \mathrm{grad} f(\mathbf{w})$. Then,

$$\begin{aligned}
[\nabla G(\mathbf{w})]_{ij} &= \partial_j G(\mathbf{w})_i \\
&= \begin{cases} \frac{2}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^4} \cdot x_j^2 w_j \cdot x_i w_i \langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}\mathbf{w}\rangle - \frac{2}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \cdot x_i x_j w_i w_j & i \neq j, \\ 1 - \frac{1}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \cdot (x_i \langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}\mathbf{w}\rangle + 2x_i^2 w_i^2) + \frac{2}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^4} x_i^2 w_i \cdot x_i w_i \langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}\mathbf{w}\rangle & i = j, \end{cases}
\end{aligned}$$

such that

$$\nabla G(\mathbf{w}) = \mathbf{D}_{\mathbf{1} - \frac{\langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}\mathbf{w}\rangle}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \cdot \mathbf{x}} - \frac{2}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}} + \frac{2\langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}\mathbf{w}\rangle}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^4} \mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}}^2.$$

Consequently, we have that

$$\begin{aligned}
[\mathrm{Hess} f(\mathbf{w})](\mathbf{u}) &= \mathbb{P}_{T_{\mathbf{w}}\mathcal{M}}([\nabla G(\mathbf{w})](\mathbf{u})) \\
&= (\mathbf{I} - \frac{1}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}}) \cdot \left[(\mathbf{1} - \frac{\langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}\mathbf{w}\rangle}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \cdot \mathbf{x}) \odot \mathbf{u} \right].
\end{aligned}$$

For any $\mathbf{w}$ satisfying the first necessary condition, we thus have that

$$\langle \mathbf{u}, [\mathrm{Hess} f(\mathbf{w})](\mathbf{u})\rangle = \mathbf{u}^T \cdot (\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^2}) \cdot (\mathbf{1} - \frac{\mathbf{x}}{x_0}) \odot \mathbf{u} = \|\mathbf{u}\|_2^2 - \langle \mathbf{u}, \frac{\mathbf{x}}{x_0} \odot \mathbf{u}\rangle,$$

where we used in the second equality that $\mathbf{x}|_{\mathrm{supp}(\mathbf{w})} = x_0 \cdot \mathbf{1}|_{\mathrm{supp}(\mathbf{w})}$ by which $(\mathbf{1} - \frac{\mathbf{x}}{x_0})|_{\mathrm{supp}(\mathbf{w})} = \mathbf{0}$. Hence, $\langle \mathbf{u}, [\mathrm{Hess} f(\mathbf{w})](\mathbf{u}) \rangle \geq 0$ can only hold for all $\mathbf{u} \in T_{\mathbf{w}}\mathcal{M}$ if $x_0 = \arg\max_i |x_i|$.

To show (ii), we proceed analogously but consider $f(\mathbf{w}) = \frac{1}{2}\mathbf{D}_{\mathbf{x}}\mathbf{w}^T\mathbf{w}\mathbf{D}_{\mathbf{x}}$, and note that $f(\mathbf{w}) = \frac{1}{8}S_{\mathcal{L}}(\mathbf{w})$ by (11). Then, one can easily check that

$$\mathrm{grad} f(\mathbf{w}) = \mathbf{D}_{\mathbf{x}}^2\mathbf{w} - \frac{1}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2}\mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}}^3 \cdot \mathbf{w},$$

which implies the same first necessary condition. Now assume $\mathbf{x} \in \mathbb{R}^d_{>0}$. Then,

$$\nabla^2 G(\mathbf{w}) = \mathbf{D}_{\mathbf{x}^{\odot 2} - \frac{\langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}^3\mathbf{w}\rangle}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2}\cdot\mathbf{x}} - \frac{2}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2}\mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}}^3 + \frac{2\langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}^3\mathbf{w}\rangle}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^4}\mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}}^2,$$

such that

$$[\mathrm{Hess} f(\mathbf{w})](\mathbf{u}) = \Big(\mathbf{I} - \frac{1}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2}\mathbf{D}_{\mathbf{x}}\mathbf{w}\mathbf{w}^T\mathbf{D}_{\mathbf{x}}\Big) \cdot \Big(\mathbf{x}^{\odot 2} - \frac{\langle \mathbf{w}, \mathbf{D}_{\mathbf{x}}^3\mathbf{w}\rangle}{\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2} \cdot \mathbf{x}\Big) \odot \mathbf{u}.$$

For any $\mathbf{w}$ satisfying the first necessary condition, we thus have that

$$\langle \mathbf{u}, [\mathrm{Hess} f(\mathbf{w})](\mathbf{u}) \rangle = \langle \mathbf{u}, \mathbf{D}_{\mathbf{x}}^2\mathbf{u} \rangle - x_0 \langle \mathbf{u}, \mathbf{D}_{\mathbf{x}}\mathbf{u} \rangle$$

which implies for $\mathbf{x} \in \mathbb{R}^d_{>0}$ that $\langle \mathbf{u}, [\mathrm{Hess} f(\mathbf{w})](\mathbf{u}) \rangle \geq 0$ can only hold for all $\mathbf{u} \in T_{\mathbf{w}}\mathcal{M}$ if $x_0 = \arg\min_i x_i$. □

The sufficient conditions are stated in the following lemma.

Lemma D.2. *For $\mathbf{x} \in \mathbb{R}^d_{>0}$ and $\mathcal{L}$ as in (4) with $\mathcal{M}$ as in (5), we have the following:*

(i) To have

$$\mathbf{w} \in \arg\min_{\mathbf{z} \in \mathcal{M}} \|\mathbf{z}^{\odot 2}\|_1,$$

it is sufficient for $\mathbf{w} \in \mathcal{M}$ that $\mathrm{supp}(\mathbf{w}) \subset \arg\max_k x_k$.

(ii) To have

$$\mathbf{w} \in \arg\min_{\mathbf{z} \in \mathcal{M}} S_{\mathcal{L}}(\mathbf{z}),$$

it is sufficient for $\mathbf{w} \in \mathcal{M}$ that $\mathrm{supp}(\mathbf{w}) \subset \arg\min_k x_k$.

Proof. First recall (10) and (11). We begin with (i). Let $k_* \in \arg\max_k x_k$. Since $\|\mathbf{w}\|_2^2 < y/x_{k_\star}$ implies by our assumption on $\mathbf{x}$ that $\langle \mathbf{x}, \mathbf{w}^{\odot 2} \rangle \leq x_{k_\star}\|\mathbf{w}\|_2^2 < y$, i.e., $\mathbf{w} \notin \mathcal{M}$, and

$$\sqrt{\frac{y}{x_{k_\star}}}\mathbf{e}_{k_*} \in \mathcal{M} \quad \text{satisfies} \quad \Big\|\sqrt{\frac{y}{x_{k_\star}}}\mathbf{e}_{k_*}\Big\|_2^2 = \frac{y}{x_{k_\star}},$$

we know by (10) that

$$\min_{\mathbf{z} \in \mathcal{M}} \|\mathbf{z}^{\odot 2}\|_1 = \frac{y}{x_{k_\star}}.$$

For any $\mathbf{w} \in \mathcal{M}$ with $\mathrm{supp}(\mathbf{w}) \subset \arg\max_k x_k$, we have that

$$y = \langle \mathbf{x}, \mathbf{w}^{\odot 2} \rangle = x_{k_\star}\|\mathbf{w}\|_2^2 = x_{k_\star}\|\mathbf{w}^{\odot 2}\|_1$$

and the claim in (i) follows.

To see (ii) we proceed analogously. Let $k_* \in \arg\min_k x_k$. Since $\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2 < yx_{k_\star}$ implies by our assumption on $\mathbf{x}$ that $\langle \mathbf{x}, \mathbf{w}^{\odot 2} \rangle \leq \frac{1}{x_{k_\star}}\|\mathbf{D}_{\mathbf{x}}\mathbf{w}\|_2^2 < y$, i.e., $\mathbf{w} \notin \mathcal{M}$, and

$$\sqrt{\frac{y}{x_{k_\star}}}\mathbf{e}_{k_*} \in \mathcal{M} \quad \text{satisfies} \quad \Big\|\mathbf{D}_{\mathbf{x}} \cdot \sqrt{\frac{y}{x_{k_\star}}}\mathbf{e}_{k_*}\Big\|_2^2 = yx_{k_\star},$$

we know by (11) that

$$\min_{\mathbf{z}\in\mathcal{M}} S_{\mathcal{L}}(\mathbf{z}) = y x_{k_\star}.$$

For any $\mathbf{w} \in \mathcal{M}$ with $\mathrm{supp}(\mathbf{w}) \subset \arg\min_k x_k$, we have that

$$y = \langle \mathbf{x}, \mathbf{w}^{\odot 2}\rangle = x_{k_\star}\|\mathbf{w}\|_2^2 = \frac{1}{x_{k_\star}} S_{\mathcal{L}}(\mathbf{w})$$

and the claim in (ii) follows. □

The specific shape of the minimizing sets (6) and (7) can easily be derived from the previous two lemmas.

E An elementary study of how implicit biases interact — Generalization

Recalling the setting outlined in Section 3, let us assume that our data follows a simple linear regression model with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $y = \langle \mathbf{1}, \mathbf{x}\rangle + \varepsilon$, for independent $\varepsilon \sim \mathcal{N}(0,1)$. Then, the risk under $\mathcal{L}$ can be computed explicitly and, given a single training data point $(\mathbf{x}_0, y_0)$ with $\mathbf{x}_0 \in \mathbb{R}^d_{\geq 0}$, the best achievable generalization error of $\phi_{\mathbf{w}}$ trained via (4) can be computed as follows.[5]

Lemma E.1. *Let $\mathcal{L}$ be as in (4) and let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$ and $y = \langle \mathbf{1}, \mathbf{x}\rangle + \varepsilon$, for independent $\varepsilon \sim \mathcal{N}(0,1)$. Then,*

$$\mathcal{R}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)}\, \mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_4^4 - \|\mathbf{w}\|_2^2 + \frac{1}{2}(d+1).$$

Let now $\eta > 0$ and $(\mathbf{x}_0, y_0) \in \mathbb{R}^d_{\geq 0} \times \mathbb{R}$, and define the corresponding risk minimization under sharpness constraints $S_{\mathcal{L}}(\mathbf{w}) \leq \frac{2}{\eta}$ as

$$\min_{\mathbf{w}\in\mathbb{R}^d} \mathcal{R}(\mathbf{w}), \quad s.t. \quad \langle \mathbf{x}_0, \mathbf{w}^{\odot 2}\rangle = y_0, \quad S_{\mathcal{L}}(\mathbf{w}) \leq \frac{2}{\eta}. \tag{12}$$

Fix any support $S_w \subset [d]$ with $S_w \cap \mathrm{supp}(\mathbf{x}_0) \neq \emptyset$. Let $\mathbf{w}$ be any vector such that $\mathrm{supp}(\mathbf{w}) = S_w$ and

$$\mathbf{w}|_{S_w}^{\odot 2} = (\mathbf{1} - 2\lambda\eta \mathbf{x}_0^{\odot 2} - \nu \mathbf{x}_0)|_{S_w},$$

for (λ, ν) as defined below:

- *If $S_{\mathcal{L}}(\mathbf{w}) \leq \frac{2}{\eta}$ and*

$$\lambda = 0$$
$$\nu = \frac{\|\mathbf{x}_0|_{S_w}\|_1 - y_0}{\|\mathbf{x}_0|_{S_w}\|_2^2}$$

  with $\nu\|\mathbf{x}_0|_{S_w}\|_\infty < 1$, then $\mathbf{w}$ is a KKT point of (12).

- *If $\mathbf{x}_0 \neq \alpha\mathbf{1}$, for all $\alpha \neq 0$, and*

$$\begin{aligned}\lambda &= \frac{y_0\|\mathbf{x}_0|_{S_w}\|_3^3 + \|\mathbf{x}_0|_{S_w}\|_2^4 - \|\mathbf{x}_0|_{S_w}\|_1\|\mathbf{x}_0|_{S_w}\|_3^3 - \frac{1}{2\eta}\|\mathbf{x}_0|_{S_w}\|_2^2}{2\eta(\|\mathbf{x}_0|_{S_w}\|_2^2\|\mathbf{x}_0|_{S_w}\|_4^4 - \|\mathbf{x}_0|_{S_w}\|_3^6)} \\ \nu &= \frac{y_0\|\mathbf{x}_0|_{S_w}\|_4^4 + \|\mathbf{x}_0|_{S_w}\|_3^3\|\mathbf{x}_0|_{S_w}\|_2^2 - \|\mathbf{x}_0|_{S_w}\|_1\|\mathbf{x}_0|_{S_w}\|_4^4 - \frac{1}{2\eta}\|\mathbf{x}_0|_{S_w}\|_3^3}{\|\mathbf{x}_0|_{S_w}\|_3^6 - \|\mathbf{x}_0|_{S_w}\|_2^2\|\mathbf{x}_0|_{S_w}\|_4^4}\end{aligned} \tag{13}$$

  or $\mathbf{x}_0 = \alpha\mathbf{1}$, for some $\alpha \neq 0$, and (λ, ν) satisfying

$$\|\mathbf{x}_0|_{S_w}\|_1 - 2\eta\lambda\|\mathbf{x}_0|_{S_w}\|_3^3 - \nu\|\mathbf{x}_0|_{S_w}\|_2^2 = y_0, \tag{14}$$

  both with $\lambda \geq 0$ and $2\lambda\eta(x_0)_i + \nu(x_0)_i < 1$, for all $i \in S_w$, then $\mathbf{w}$ is a KKT point of (12).

[5] Note that $(\mathbf{x}_0, y_0)$ takes in this section the role of the single data point $(\mathbf{x}, y)$ from before and that we condition to non-negative data in order to apply Proposition 3.2.

This characterizes all KKT points of (12).

Proof. First note that

$$\begin{aligned}\mathcal{R}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)}\,\mathcal{L}(\mathbf{w}) &= \frac{1}{2}\,\mathbb{E}_{(\mathbf{x},y)}(\langle \mathbf{w}^{\odot 2}, \mathbf{x}\rangle - y)^2 \\ &= \frac{1}{2}\Big((\mathbf{w}^{\odot 2})^T\,\mathbb{E}\,(\mathbf{x}\mathbf{x}^T)\mathbf{w}^{\odot 2} - 2\,\mathbb{E}\,(y\mathbf{x}^T)\mathbf{w}^{\odot 2} + \mathbb{E}\,y^2\Big) \\ &= \frac{1}{2}\|\mathbf{w}^{\odot 2}\|_2^2 - \langle \mathbf{1}, \mathbf{w}^{\odot 2}\rangle + \frac{1}{2}(d+1) \\ &= \frac{1}{2}\|\mathbf{w}\|_4^4 - \|\mathbf{w}\|_2^2 + \frac{1}{2}(d+1),\end{aligned}$$

where we used in the penultimate line that $\mathbb{E}\,(y\mathbf{x}^T) = \mathbf{1}^T$ and $\mathbb{E}\,(y^2) = d+1$, and in the ultimate line that $\langle \mathbf{1}, \mathbf{w}^{\odot 2}\rangle = \|\mathbf{w}\|_2^2$ and $\|\mathbf{w}^{\odot 2}\|_2^2 = \|\mathbf{w}\|_4^4$.

For the KKT analysis of Equation (12), we will drop the additive constant $\frac{1}{2}(d+1)$. We first re-write Equation (12) as

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}), \quad \text{s.t.} \quad h(\mathbf{w}) = 0, \quad g(\mathbf{w}) \le 0.$$

where

$$\begin{aligned}f(\mathbf{w}) &= \frac{1}{2}\|\mathbf{w}\|_4^4 - \|\mathbf{w}\|_2^2 \\ h(\mathbf{w}) &= \langle \mathbf{x}_0, \mathbf{w}^{\odot 2}\rangle - y_0 \\ g(\mathbf{w}) &= 2\eta\|\mathbf{x}_0 \odot \mathbf{w}\|_2^2 - 1.\end{aligned}$$

The point $\mathbf{w}$ satisfies the KKT conditions if there exists $\lambda, \nu \in \mathbb{R}$ such that

$$\begin{aligned}\nabla f(\mathbf{w}) + \nu\nabla h(\mathbf{w}) + \lambda\nabla g(\mathbf{w}) &= \mathbf{0} \\ h(\mathbf{w}) &= 0 \\ g(\mathbf{w}) &\le 0 \\ \lambda g(\mathbf{w}) &= 0 \\ \lambda &\ge 0.\end{aligned}$$

Plugging in, we obtain

$$2\mathbf{w}^{\odot 3} - 2\mathbf{w} + 2\nu\mathbf{x}_0 \odot \mathbf{w} + 4\lambda\eta\mathbf{x}_0^{\odot 2} \odot \mathbf{w} = \mathbf{0} \tag{15}$$

$$\langle \mathbf{x}_0, \mathbf{w}^{\odot 2}\rangle - y_0 = 0 \tag{16}$$

$$2\eta\|\mathbf{x}_0 \odot \mathbf{w}\|_2^2 - 1 \le 0 \tag{17}$$

$$\lambda(2\eta\|\mathbf{x}_0 \odot \mathbf{w}\|_2^2 - 1) = 0 \tag{18}$$

$$\lambda \ge 0. \tag{19}$$

By rewriting (15) as

$$(\mathbf{w}^{\odot 2} - \mathbf{1} + \nu\mathbf{x}_0 + 2\lambda\eta\mathbf{x}_0^{\odot 2}) \odot \mathbf{w} = \mathbf{0},$$

we see that, for any $i \in [d]$, we have

$$w_i = 0 \qquad \text{or} \qquad w_i^2 = 1 - \nu(x_0)_i - 2\lambda\eta(x_0)_i^2. \tag{20}$$

Consider any $\mathbf{w}$ with $\operatorname{supp}(\mathbf{w}) = S_w$ satisfying the KKT conditions.

If $\lambda = 0$, we get that $\mathbf{w}|_{S_w}^{\odot 2} = (\mathbf{1} - \nu\mathbf{x}_0)|_{S_w}$ such that (16) yields that

$$\|\mathbf{x}_0|_{S_w}\|_1 - \nu\|\mathbf{x}_0|_{S_w}\|_2^2 = y_0 \qquad \Leftrightarrow \qquad \nu = \frac{\|\mathbf{x}_0|_{S_w}\|_1 - y_0}{\|\mathbf{x}_0|_{S_w}\|_2^2},$$

which implies that a suitable ν exists iff $S_w \cap \operatorname{supp}(\mathbf{x}_0) \neq \emptyset$ and $\nu < \min_{i\in S_w\cap\operatorname{supp}(\mathbf{x}_0)} \frac{1}{(x_0)_i}$. The latter condition stems from the fact that non-zero entries of $\mathbf{w}^{\odot 2}$ have to be positive. Finally, to be a KKT point, $\mathbf{w}$ has to satisfy (17).

If $\lambda \neq 0$, we get that $\mathbf{w}|_{S_w}^{\odot 2} = (\mathbf{1} - 2\lambda\eta\mathbf{x}_0^{\odot 2} - \nu\mathbf{x}_0)|_{S_w}$ such that (16) and (18) yield that

$$\|\mathbf{x}_0|_{S_w}\|_1 - 2\eta\lambda\|\mathbf{x}_0|_{S_w}\|_3^3 - \nu\|\mathbf{x}_0|_{S_w}\|_2^2 = y_0$$
$$\|\mathbf{x}_0|_{S_w}\|_2^2 - 2\eta\lambda\|\mathbf{x}_0|_{S_w}\|_4^4 - \nu\|\mathbf{x}_0|_{S_w}\|_3^3 = \frac{1}{2\eta},$$

which is a solvable linear system iff $S_w \cap \operatorname{supp}(\mathbf{x}_0) \neq \emptyset$. If $\mathbf{x}_0|_{S_w} \neq \alpha\mathbf{1}|_{S_w}$, for all $\alpha \neq 0$, the unique solution is given by (13). Else, the system is underdetermined and only yields the relation in (14). Finally, if $\lambda \geq 0$ and $(2\lambda\eta\mathbf{x}_0^{\odot 2} + \nu\mathbf{x}_0)|_{S_w} < \mathbf{1}|_{S_w}$ (positivity constraint for non-zero entries of $\mathbf{w}^{\odot 2}$), any resulting $\mathbf{w}$ yields the second type of KKT point. □

While it is cumbersome to analytically extract for general d which of the KKT points of Lemma E.1 corresponds to a global minimizer, we can easily evaluate this numerically in our toy example from Figure 4, see Section 3.

E.1 A more general regression analysis

Since it is more natural to have unconditioned training data, let us now assume that our data follows a general distribution $(\mathbf{x}, y) \sim \mathcal{D}$. Then, the risk for a parameter choice $\mathbf{w}$ under the model in (3)-(4) is given by

$$\mathcal{R}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)} \mathcal{L}(\mathbf{w}) = \frac{1}{2}\Big((\mathbf{w}^{\odot 2})^T\boldsymbol{\Sigma}\mathbf{w}^{\odot 2} - 2\boldsymbol{\mu}^T\mathbf{w}^{\odot 2} + \sigma^2\Big), \tag{21}$$

where we define $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{x}\mathbf{x}^T)$, $\boldsymbol{\mu} = \mathbb{E}(y\mathbf{x})$, and $\sigma^2 = \mathbb{E}\, y^2$. Under mild technical assumptions on $\mathcal{D}$ and considering a single training data point $(\mathbf{x}_0, y_0) \sim (\mathbf{x}, y)$, we can compare the three (idealized) training algorithms $\mathcal{A}_{\ell_1}$, $\mathcal{A}_{S_\mathcal{L}}$, and $\mathcal{A}_{\text{opt}}$ from above which minimize ℓ_1-norm, sharpness, and generalization error on $\mathcal{M}$, respectively.

Proposition E.2. *Assume that $\mathcal{D}$ is a distribution such that $\boldsymbol{\Sigma}, \boldsymbol{\mu}, \sigma^2$ are well-defined and finite, that $\boldsymbol{\Sigma}$ is invertible, that $\mathbf{x} \in \mathbb{R}^d_{\geq 0}$ a.s., and that the entries of $\mathbf{x}$ are a.s. distinct. Then, given a single training data point $(\mathbf{x}_0, y_0) \sim (\mathbf{x}, y)$ we have that*

(i) $\mathcal{A}_{\ell_1}(\mathbf{x}_0, y_0) = \sqrt{\frac{y_0}{x_{\max}}}\mathbf{e}_{k_{\max}}$, *where $k_{\max}$ is the index of the maximal entry of $\mathbf{x}_0$. The expected generalization error is given by*

$$\mathbb{E}_{(\mathbf{x}_0,y_0)} \mathcal{R}(\mathcal{A}_{\ell_1}(\mathbf{x}_0, y_0)) = \frac{1}{2}\left(\sigma^2 + \mathbb{E}\left(\frac{\Sigma_{k_{\max}k_{\max}} y_0^2}{x_{\max}^2}\right) + \mathbb{E}\left(\frac{\mu_{k_{\max}} y_0}{x_{\max}}\right)\right).$$

(ii) $\mathcal{A}_{S_\mathcal{L}}(\mathbf{x}_0, y_0) = \sqrt{\frac{y_0}{x_{\min}}}\mathbf{e}_{k_{\min}}$, *where $k_{\min}$ is the index of the minimal entry of $\mathbf{x}_0$. The expected generalization error is given by*

$$\mathbb{E}_{(\mathbf{x}_0,y_0)} \mathcal{R}(\mathcal{A}_{S_\mathcal{L}}(\mathbf{x}_0, y_0)) = \frac{1}{2}\left(\sigma^2 + \mathbb{E}\left(\frac{\Sigma_{k_{\min}k_{\min}} y_0^2}{x_{\min}^2}\right) + \mathbb{E}\left(\frac{\mu_{k_{\min}} y_0}{x_{\min}}\right)\right).$$

(iii) $\mathcal{A}_{opt}(\mathbf{x}_0, y_0) = \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\mathcal{P}^\perp_{\mathbf{x}_\boldsymbol{\Sigma}}\boldsymbol{\mu}_\boldsymbol{\Sigma} + \frac{y_0}{\|\mathbf{x}_\boldsymbol{\Sigma}\|_2^2}\mathbf{x}_\boldsymbol{\Sigma}\right)\right)^{\odot\frac{1}{2}}$, *where $\mathcal{P}_\mathbf{z}$ denotes the orthogonal projection onto* $\operatorname{span}\{\mathbf{z}\}$, *$\mathbf{x}_\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{x}_0$, and $\boldsymbol{\mu}_\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}$. The expected generalization error is given by*

$$\begin{aligned}&\mathbb{E}_{(\mathbf{x}_0,y_0)} \mathcal{R}(\mathcal{A}_{opt}(\mathbf{x}_0, y_0)) \\ &\quad= \frac{1}{2}\left(\sigma^2 + \mathbb{E}\left(\frac{y_0^2}{\|\mathbf{x}_\boldsymbol{\Sigma}\|_2^2}\right) - 2\boldsymbol{\mu}_\boldsymbol{\Sigma}^T\,\mathbb{E}\left(\frac{y_0}{\|\mathbf{x}_\boldsymbol{\Sigma}\|_2^2}\mathbf{x}_\boldsymbol{\Sigma}\right) - \boldsymbol{\mu}_\boldsymbol{\Sigma}^T\,\mathbb{E}\,\mathcal{P}^\perp_{\mathbf{x}_\boldsymbol{\Sigma}}\boldsymbol{\mu}_\boldsymbol{\Sigma}\right).\end{aligned}$$

Although it is not possible to analytically evaluate the expectations on this level of generality, the expected generalization error of $\mathcal{A}_{S_\mathcal{L}}(\mathbf{x}_0, y_0)$ will presumably be larger than the one of $\mathcal{A}_{\ell_1}(\mathbf{x}_0, y_0)$ since $x_{\min} < x_{\max}$; just like in the specific setting in the beginning of Section E.

Proof of Proposition E.2. By our assumptions on the distribution of $\mathbf{x}_0$, Points (i) and (ii) follow from applying Proposition 3.2, and inserting the resulting minimizer into (21).

To derive (iii), we abbreviate $\tilde{\mathbf{w}} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{w}^{\odot 2}$, $\boldsymbol{\mu}_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}$, and $\mathbf{x}_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{x}_0$, and consider the linearly constrained optimization problem

$$\min_{\mathbf{w}\in\mathcal{M}} \mathcal{R}(\mathbf{w}) = \frac{1}{2}\min_{\tilde{\mathbf{w}}\in\mathbb{R}^d} \|\tilde{\mathbf{w}}\|_2^2 - 2\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\tilde{\mathbf{w}} + \sigma^2, \quad \text{s.t. } \mathbf{x}_{\boldsymbol{\Sigma}}^T\tilde{\mathbf{w}} = y_0. \tag{22}$$

Since the objective is convex and the constraints are linear, the KKT-conditions of (22)

$$\begin{cases} 2\tilde{\mathbf{w}} - 2\boldsymbol{\mu}_{\boldsymbol{\Sigma}} + \lambda\mathbf{x}_{\boldsymbol{\Sigma}} = 0 \\ \mathbf{x}_{\boldsymbol{\Sigma}}^T\tilde{\mathbf{w}} = y_0 \end{cases} \iff \begin{cases} \tilde{\mathbf{w}} = \boldsymbol{\mu}_{\boldsymbol{\Sigma}} - \frac{1}{2}\lambda\mathbf{x}_{\boldsymbol{\Sigma}} \\ \mathbf{x}_{\boldsymbol{\Sigma}}^T\boldsymbol{\mu}_{\boldsymbol{\Sigma}} - \frac{1}{2}\lambda\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2 = y_0 \end{cases} \iff \begin{cases} \tilde{\mathbf{w}} = \boldsymbol{\mu}_{\boldsymbol{\Sigma}} - \frac{1}{2}\lambda\mathbf{x}_{\boldsymbol{\Sigma}} \\ \frac{1}{2}\lambda = \frac{1}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}(\mathbf{x}_{\boldsymbol{\Sigma}}^T\boldsymbol{\mu}_{\boldsymbol{\Sigma}} - y_0) \end{cases}$$

are sufficient and necessary, and yield the unique minimizer

$$\tilde{\mathbf{w}}_\star = \Big(\mathbf{I} - \frac{\mathbf{x}_{\boldsymbol{\Sigma}}\mathbf{x}_{\boldsymbol{\Sigma}}^T}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\Big)\boldsymbol{\mu}_{\boldsymbol{\Sigma}} + \frac{y_0}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\mathbf{x}_{\boldsymbol{\Sigma}}$$

with

$$\begin{aligned}
\mathcal{R}(\mathcal{A}_{\text{opt}}(\mathbf{x}_0, y_0)) &= \frac{1}{2}\big(\|\tilde{\mathbf{w}}_\star\|_2^2 - 2\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\tilde{\mathbf{w}}_\star + \sigma^2\big) \\
&= \frac{1}{2}\left(\Big\|\mathcal{P}_{\mathbf{x}_{\boldsymbol{\Sigma}}}^{\perp}\boldsymbol{\mu}_{\boldsymbol{\Sigma}} + \frac{y_0}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\mathbf{x}_{\boldsymbol{\Sigma}}\Big\|_2^2 - 2\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\Big(\mathcal{P}_{\mathbf{x}_{\boldsymbol{\Sigma}}}^{\perp}\boldsymbol{\mu}_{\boldsymbol{\Sigma}} + \frac{y_0}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\mathbf{x}_{\boldsymbol{\Sigma}}\Big) + \sigma^2\right) \\
&= \frac{1}{2}\left(\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\mathcal{P}_{\mathbf{x}_{\boldsymbol{\Sigma}}}^{\perp}\boldsymbol{\mu}_{\boldsymbol{\Sigma}} + \Big\|\frac{y_0}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\mathbf{x}_{\boldsymbol{\Sigma}}\Big\|_2^2 - 2\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\mathcal{P}_{\mathbf{x}_{\boldsymbol{\Sigma}}}^{\perp}\boldsymbol{\mu}_{\boldsymbol{\Sigma}} - 2\frac{y_0}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\mathbf{x}_{\boldsymbol{\Sigma}} + \sigma^2\right) \\
&= \frac{1}{2}\left(\frac{y_0^2}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2} - 2\frac{y_0}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\mathbf{x}_{\boldsymbol{\Sigma}} - \boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\mathcal{P}_{\mathbf{x}_{\boldsymbol{\Sigma}}}^{\perp}\boldsymbol{\mu}_{\boldsymbol{\Sigma}} + \sigma^2\right).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x}_0,y_0)}\,\mathcal{R}(\mathcal{A}_{\text{opt}}(\mathbf{x}_0, y_0)) \\
&\quad= \frac{1}{2}\left(\sigma^2 + \mathbb{E}\left(\frac{y_0^2}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\right) - 2\boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\,\mathbb{E}\left(\frac{y_0}{\|\mathbf{x}_{\boldsymbol{\Sigma}}\|_2^2}\mathbf{x}_{\boldsymbol{\Sigma}}\right) - \boldsymbol{\mu}_{\boldsymbol{\Sigma}}^T\,\mathbb{E}\,\mathcal{P}_{\mathbf{x}_{\boldsymbol{\Sigma}}}^{\perp}\boldsymbol{\mu}_{\boldsymbol{\Sigma}}\right).
\end{aligned}$$

□

We can now use Proposition E.2 to examine a regression task in which the feature distribution is a folded Gaussian and thus restricted to the positive orthant. Let $\mathbf{x} \sim |\mathcal{N}(0, \mathbf{I}_n)|$ and $y = \langle \mathbf{1}, \mathbf{x}\rangle$. Then $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$, and σ^2 are given by

$$\begin{aligned}
\Sigma_{ij} = \mathbb{E}(\mathbf{x}_i\mathbf{x}_j) &= \begin{cases} 1 & \text{if } i = j \\ \frac{2}{\pi} & \text{if } i \neq j \end{cases} \\
\mu_i = \mathbb{E}(y\mathbf{x}_i) &= \mathbb{E}(\mathbf{x}_i^2) + \sum_{j:j\neq i}\mathbb{E}(\mathbf{x}_i\mathbf{x}_j) = 1 + \frac{2(n-1)}{\pi} \\
\sigma^2 = \mathbb{E}(y^2) &= \sum_i \mathbb{E}(\mathbf{x}_i^2) + \sum_{i,j:i\neq j}\mathbb{E}(\mathbf{x}_i\mathbf{x}_j) = n + \frac{2n(n-1)}{\pi}
\end{aligned}$$

By Proposition E.2, we obtain the following results: For $\mathcal{A}_{\ell_1}(\mathbf{x}_0, y_0)$, the expected generalization error is given by

$$\frac{1}{2}\left(\frac{n(2n-2+\pi)}{\pi} + \mathbb{E}\left(\frac{\langle \mathbf{1}, \mathbf{x}_0\rangle^2}{x_{\max}^2}\right) + \frac{2n-2+\pi}{\pi}\,\mathbb{E}\left(\frac{\langle \mathbf{1}, \mathbf{x}_0\rangle}{x_{\max}}\right)\right).$$

Since $\langle \mathbf{1}, \mathbf{x}_0\rangle \leq nx_{\max}$, the above expectation terms are bounded by

$$\mathbb{E}\,\frac{\langle \mathbf{1}, \mathbf{x}_0\rangle^2}{x_{\max}^2} \leq n^2, \quad \mathbb{E}\,\frac{\langle \mathbf{1}, \mathbf{x}_0\rangle}{x_{\max}} \leq n.$$

For $\mathcal{A}_{S_\mathcal{L}}(\mathbf{x}_0, y_0)$, the expected generalization error is given by

$$\frac{1}{2}\left(\frac{n(2n-2+\pi)}{\pi} + \mathbb{E}\left(\frac{\langle \mathbf{1}, \mathbf{x}_0\rangle^2}{x_{\min}^2}\right) + \frac{2n-2+\pi}{\pi}\,\mathbb{E}\left(\frac{\langle \mathbf{1}, \mathbf{x}_0\rangle}{x_{\min}}\right)\right).$$

However, in this case due to $x_{\min}$ the expectation blows up to infinity as shown below.

$$\begin{aligned}\mathbb{E}\,\frac{\langle \mathbf{1}, \mathbf{x}_0\rangle}{x_{\min}} &\geq \left(\frac{2}{\pi}\right)^{n/2} \int_{[0,1]\times[1,2]^{n-1}} \frac{x_1 + \cdots + x_n}{\min_i x_i} e^{-\frac{1}{2}(x_1^2+\cdots+x_n^2)} dx_1 \cdots dx_n \\ &\geq \left(\frac{2}{\pi}\right)^{n/2} \underbrace{\int_{[0,1]} \frac{n-1}{x_1} e^{-\frac{1}{2}x_1^2} dx_1}_{=\infty} \underbrace{\int_{[1,2]^{n-1}} e^{-\frac{1}{2}(x_2^2+\cdots+x_n^2)} dx_2 \cdots dx_n}_{>0} = \infty.\end{aligned}$$

Consequently, as in the simpler setting above we see that the implicit GF-regularization leads to smaller generalization error than the sharpness regularization.

F An elementary study of how implicit biases interact II — Classification

In this section, we extend our insights from Section 3 to a simple classification set-up. To this end, define for data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^{d+1} \times \{0,1\}$ the logistic loss

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \left(y_i \log(g(\langle \mathbf{w}, \mathbf{x}_i\rangle)) + (1-y_i)\log(1-g(\langle \mathbf{w}, \mathbf{x}_i\rangle))\right),$$

where

$$g\colon \mathbb{R} \to \mathbb{R} \quad \text{with} \quad g(z) = \frac{1}{1+e^{-z}}$$

is the logistic function. Here, we assume that $\mathbf{w} = (\tilde{\mathbf{w}}, b)^T$ and that the data points are of the form $\mathbf{x} = (\tilde{\mathbf{x}}, 1)^T$ such that the linear classifier $h_\mathbf{w}$ corresponding to parameters $\mathbf{w}$ is given by

$$h_\mathbf{w}(\mathbf{x}) = \mathbf{1}_{\{\mathbf{z}=(\tilde{\mathbf{z}},1)\colon \langle \mathbf{w}, \mathbf{z}\rangle > 0\}}(\mathbf{x}) = \mathbf{1}_{\{\tilde{\mathbf{z}}\colon \langle \tilde{\mathbf{w}}, \tilde{\mathbf{z}}\rangle + b > 0\}}(\mathbf{x}).$$

In the simplest possible case, we only have two data points with different labels. W.l.o.g. we assume that one of the two data points is centered at the origin and that their distance is normalized to one. Then we know the following.

Theorem F.1. *Let $D = \{(\mathbf{x}_1, 0), (\mathbf{x}_2, 1)\} \subset \mathbb{R}^{d+1} \times \{0,1\}$ where $\mathbf{x}_i = (\tilde{\mathbf{x}}_i, 1)^T$ with $\tilde{\mathbf{x}}_1 = \mathbf{0}$ and $\|\tilde{\mathbf{x}}_2\|_2 = 1$. Then,*

(i) the max-margin classifier of D is parametrized by any positive scalar multiple of $\mathbf{w} = (\tilde{\mathbf{w}}, b)^T$ with $\tilde{\mathbf{w}} = \tilde{\mathbf{x}}_2$ and $b = -1/2$.

(ii) the parameters minimizing the sharpness of $\mathcal{L}$ over

$$\mathcal{M} = \{\mathbf{w} = (\tilde{\mathbf{w}}, b)\colon h_\mathbf{w}(\mathbf{x}_1) = 0,\ h_\mathbf{w}(\mathbf{x}_2) = 1,\ \text{and } \|\tilde{\mathbf{w}}\|_2 = 1\}$$

are given by a min-margin classifier parametrized by $\mathbf{w} = (\tilde{\mathbf{w}}, b)$ with $\tilde{\mathbf{w}} = \tilde{\mathbf{x}}_2$ and $b = 0$.

Proof. To see (i), just note that the decision boundary of the max-margin classifier in $\mathbb{R}^d$ must be orthogonal to $\tilde{\mathbf{x}}_2 - \tilde{\mathbf{x}}_1$ with $h_\mathbf{w}(\mathbf{x}_2) = 1$, i.e., $\tilde{\mathbf{w}} = \alpha(\tilde{\mathbf{x}}_2 - \tilde{\mathbf{x}}_1) = \alpha\tilde{\mathbf{x}}_2$, for $\alpha > 0$, and that it must contain $\frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2)$ which implies that $0 = \langle \tilde{\mathbf{w}}, \frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2)\rangle + b = \frac{1}{2}\alpha\|\tilde{\mathbf{x}}_2\|_2^2 + b$, i.e., $b = -\frac{1}{2}\alpha$.

For (ii), we compute that

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= \frac{1}{2}\left(\log(1-g(\langle \mathbf{w}, \mathbf{x}_1\rangle)) + \log(g(\langle \mathbf{w}, \mathbf{x}_2\rangle))\right) \\ &= \frac{1}{2}\left(\log(1-g(b)) + \log(g(\langle \mathbf{w}, \mathbf{x}_2\rangle))\right).\end{aligned}$$

By using that $g'(z) = g(z)(1 - g(z))$, we then get that

$$\nabla \mathcal{L}(\mathbf{w}) = \frac{1}{2}\left(-g(\langle \mathbf{w}, \mathbf{x}_1\rangle) \cdot \mathbf{x}_1 + (1 - g(\langle \mathbf{w}, \mathbf{x}_2\rangle)) \cdot \mathbf{x}_2\right)$$

and

$$\nabla^2 \mathcal{L}(\mathbf{w}) = -\frac{1}{2}\left(g'(\langle \mathbf{w}, \mathbf{x}_1\rangle) \cdot \mathbf{x}_1\mathbf{x}_1^T + g'(\langle \mathbf{w}, \mathbf{x}_2\rangle) \cdot \mathbf{x}_2\mathbf{x}_2^T\right).$$

To deduce the sharpness $S(\mathbf{w}) = \left\|\nabla^2 \mathcal{L}(\mathbf{w})\right\|$, we will compute the eigenvalues of the Hessian. First note, that any vector in the image of $\nabla^2 \mathcal{L}(\mathbf{w})$ can be expressed as $\mathbf{x} = \alpha \mathbf{e}_{d+1} + \beta \mathbf{x}_2$. Now assume $\mathbf{x} \neq \mathbf{0}$ is an eigenvector with eigenvalue $\lambda \neq 0$. Then, since $\mathbf{x}_1 = \mathbf{e}_{d+1}$,

$$\begin{aligned}\nabla^2 \mathcal{L}(\mathbf{w})\mathbf{x} &= -\frac{1}{2}\Big(g'(b)\,(\alpha + \beta)\,\mathbf{e}_{d+1} + g'(\langle \mathbf{w}, \mathbf{x}_2\rangle)\,(\alpha + 2\beta)\,\mathbf{x}_2\Big) \\ &= \lambda(\alpha \mathbf{e}_{d+1} + \beta \mathbf{x}_2),\end{aligned}$$

where we used that $\mathbf{x}_2^T \mathbf{e}_{d+1} = \mathbf{e}_{d+1}^T \mathbf{x}_2 = 1$, $\mathbf{x}_2^T \mathbf{x}_2 = 2$, and $\mathbf{e}_{d+1}^T \mathbf{e}_{d+1} = 1$. Matching coefficients, we obtain the system

$$\begin{pmatrix} \frac{1}{2}g'(b) + \lambda & \frac{1}{2}g'(b) \\ \frac{1}{2}g'(\langle \mathbf{w}, \mathbf{x}_2\rangle) & g'(\langle \mathbf{w}, \mathbf{x}_2\rangle) + \lambda \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0.$$

Since $(\alpha, \beta) \neq \mathbf{0}$, this implies that the matrix has determinant zero and leads to the quadratic equation

$$\lambda^2 + \Big(\frac{1}{2}g'(b) + g'(\langle \mathbf{w}, \mathbf{x}_2\rangle)\Big)\lambda + \frac{1}{4}g'(b) \cdot g'(\langle \mathbf{w}, \mathbf{x}_2\rangle) = 0.$$

Since $g'(b), g'(\langle \mathbf{w}, \mathbf{x}_2\rangle) > 0$, the maximal solution of the latter system, i.e., the leading eigenvalue of $\nabla^2 \mathcal{L}(\mathbf{w})$, is

$$\begin{aligned}S(\mathbf{w}) = \left\|\nabla^2 \mathcal{L}(\mathbf{w})\right\| &= \frac{\frac{1}{2}g'(b) + g'(\langle \mathbf{w}, \mathbf{x}_2\rangle) + \sqrt{\frac{1}{4}g'(b)^2 + g'(\langle \mathbf{w}, \mathbf{x}_2\rangle)^2}}{2} \\ &= \frac{1}{4}\left(g'(b) + 2g'(\langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}_2\rangle + b) + \sqrt{g'(b)^2 + 4 \cdot g'(\langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}_2\rangle + b)^2}\right).\end{aligned}$$

The parameter minimizing the sharpness is then

$$\begin{aligned}&\min_{\mathbf{w} \in \mathcal{M}} S_{\mathcal{L}}(\mathbf{w}) \\ &= \frac{1}{4} \min_{\|\tilde{\mathbf{w}}\|_2 = 1} g'(b) + 2g'(\langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}_2\rangle + b) + \\ &\qquad\qquad \sqrt{g'(b)^2 + (2g'(\langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}_2\rangle + b))^2}, \qquad \text{s.t. } \begin{cases} b = \langle \mathbf{w}, \mathbf{x}_1\rangle \leq 0 \\ \langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}}_2\rangle + b > 0 \end{cases} \\ &= \frac{1}{4} \min_{z \in (0,1]} g'(b) + 2g'(z + b) + \sqrt{g'(b)^2 + (2g'(z + b))^2}, \qquad \text{s.t. } -z < b \leq 0 \\ &\approx 0.277\end{aligned}$$

The minimum of the function is attained at $(z, b) = (1, 0)$ which means that $\tilde{\mathbf{w}} = \tilde{\mathbf{x}}_2$. □

Analogously to the regression case, we can now evaluate the max-margin and the sharpness minimizing classifiers in terms of their expected generalization error in a toy set-up that assumes only two samples. To satisfy the requirements of Theorem F.1, we propose the following simple data generation process.

Let the samples be generated as $(\mathbf{x}_1, y_1)$ with $\tilde{\mathbf{x}}_1 = \mathbf{0}$ and $y_1 = 0$, and, for $k \geq 2$, as $(\mathbf{x}_k, y_k) \sim (\mathbf{x}, 1)$ which follows a joint distribution with $\mathbf{x} \sim \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$, where $\mathbf{g} \sim \mathcal{N}(\boldsymbol{\mu}, I)$ for $\boldsymbol{\mu} \neq \mathbf{0}$. The classification task is thus to separate a Gaussian cluster that is projected to the unit sphere from the origin. Given two samples $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$ one can use Theorem F.1 and numerically evaluate that the expected generalization error (Mohri et al., 2018). To get a feeling of it, let us consider the two cases where $\|\boldsymbol{\mu}\| \ll 1$ and $\|\boldsymbol{\mu}\| \gg 1$. Let $\mathbf{g}_0$ and $\mathbf{g}_0'$ be independent and distributed as $\mathcal{N}(0, I)$.

Suppose $\|\boldsymbol{\mu}\| \ll 1$. The expected generalization error for the max-margin classifier $\mathbf{w}_{max} = (\tilde{\mathbf{w}}_{max}, b_{max})^T$ is

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{x}}_2}\mathbb{P}_{\mathbf{x}}[h_{\mathbf{w}_{max}(\mathbf{x})\neq 1}] &= \mathbb{E}_{\tilde{\mathbf{x}}_2}\mathbb{P}_{\mathbf{g}}\Big[\Big\langle \tilde{\mathbf{x}}_2, \frac{\mathbf{g}}{\|\mathbf{g}\|_2}\Big\rangle \leq \frac{1}{2}\Big] \\ &\approx \mathbb{E}_{\mathbf{g}_0'}\mathbb{P}_{\mathbf{g}_0}\Big[\Big\langle \frac{\mathbf{g}_0'}{\|\mathbf{g}_0'\|_2}, \frac{\mathbf{g}_0}{\|\mathbf{g}_0\|_2}\Big\rangle \leq \frac{1}{2}\Big] \\ &\approx \frac{\gamma(\frac{d}{2}+\frac{1}{2})}{\gamma(\frac{d}{2})\gamma(\frac{1}{2})}\int_{-1}^{\frac{1}{2}}(1-x^2)^{\frac{d}{2}-1}dx \\ &\to 1 \text{ (as } d \text{ grows)}\end{aligned}$$

because $(1-x^2)^{\frac{d}{2}-1}$ concentrates well around $x=0$. On the other hand, the expected generalization error for the sharpness minimizing classifier $\mathbf{w}_{min} = (\tilde{\mathbf{w}}_{min}, b_{min})$ is

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{x}}_2}\mathbb{P}_{\mathbf{x}}[h_{\mathbf{w}_{min}(\mathbf{x})\neq 1}] &= \mathbb{E}_{\tilde{\mathbf{x}}_2}\mathbb{P}_{\mathbf{g}}\Big[\Big\langle \tilde{\mathbf{x}}_2, \frac{\mathbf{g}}{\|\mathbf{g}\|_2}\Big\rangle \leq 0\Big] \\ &\approx \mathbb{E}_{\mathbf{g}_0'}\mathbb{P}_{\mathbf{g}_0}\Big[\Big\langle \frac{\mathbf{g}_0'}{\|\mathbf{g}_0'\|_2}, \frac{\mathbf{g}}{\|\mathbf{g}_0\|_2}\Big\rangle \leq 0\Big] \\ &= \frac{1}{2},\end{aligned}$$

where we used symmetry of the distribution in the last step. We see that in contrast to Section E here the sharpness minimizer leads to a significantly smaller expected generalization error than the GF-induced regularization.

Now suppose that $\|\boldsymbol{\mu}\| \gg 1$. The expected generalization error for the max-margin classifier is

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{x}}_2}\mathbb{P}_{\mathbf{x}}[h_{\mathbf{w}_{max}(\mathbf{x})\neq 1}] &= \mathbb{E}_{\mathbf{g}'}\mathbb{P}_{\mathbf{g}}\Big[\Big\langle \frac{\mathbf{g}'}{\|\mathbf{g}'\|_2}, \frac{\mathbf{g}}{\|\mathbf{g}\|_2}\Big\rangle \leq \frac{1}{2}\Big] \\ &\approx \mathbb{E}_{\mathbf{g}'}\mathbb{P}_{\mathbf{g}}\Big[\Big\langle \frac{\mathbf{g}_0'+\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}, \frac{\mathbf{g}_0+\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}\Big\rangle \leq \frac{1}{2}\Big] \\ &\approx \mathbb{E}_{\mathbf{g}'}\mathbb{P}_{\mathbf{g}}\Big[\langle \mathbf{g}_0'+\mathbf{g}_0, \boldsymbol{\mu}\rangle \leq -\frac{1}{2}\|\boldsymbol{\mu}\|_2^2\Big] \\ &= \frac{1}{\sqrt{2}(2\pi)^{\frac{d}{2}}}\int_{-\infty}^{-\frac{1}{2}\|\boldsymbol{\mu}\|_2} e^{-\frac{1}{4}x^2}dx \\ &= \frac{1}{(2\pi)^{\frac{d-1}{2}}}\cdot\Phi\Big(-\frac{1}{2\sqrt{2}}\|\boldsymbol{\mu}\|_2\Big)\end{aligned}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Similarly, the expected generalization error for the sharpness minimizing classifier is

$$\begin{aligned}\mathbb{E}_{\tilde{\mathbf{x}}_2}\mathbb{P}_{\mathbf{x}}[h_{\mathbf{w}_{min}(\mathbf{x})\neq 1}] &= \mathbb{E}_{\tilde{\mathbf{x}}_2}\mathbb{P}_{\mathbf{g}}\Big[\Big\langle \tilde{\mathbf{x}}_2, \frac{\mathbf{g}}{\|\mathbf{g}\|_2}\Big\rangle \leq 0\Big] \\ &\approx \mathbb{E}_{\mathbf{g}_0'}\mathbb{P}_{\mathbf{g}_0}\Big[\langle \mathbf{g}_0'+\mathbf{g}_0, \boldsymbol{\mu}\rangle \leq -\|\boldsymbol{\mu}\|_2^2\Big] \\ &= \frac{1}{\sqrt{2}(2\pi)^{d/2}}\int_{-\infty}^{-\|\boldsymbol{\mu}\|_2} e^{-\frac{1}{4}x^2}dx \\ &= \frac{1}{(2\pi)^{\frac{d-1}{2}}}\cdot\Phi\Big(-\frac{1}{\sqrt{2}}\|\boldsymbol{\mu}\|_2\Big).\end{aligned}$$

Here, both expected generalization errors are small.

G Methodology

To ensure reproducibility, we follow a standard procedure for each experimental configuration, which is defined by a specific combination of dataset, architecture, activation function,

and loss function. To isolate the effect of the learning rate, we fix the initialization across all runs within a configuration. We initialize using the default PyTorch scheme, which is a modified LeCun initialization (LeCun et al., 2002): Fixing a random seed, initial entries of each weight matrix are uniformly sampled from the interval $\left(-1/\sqrt{n_{l-1}}, 1/\sqrt{n_{l-1}}\right)$, where n_{l-1} is the input dimension of the respective matrix.

We begin by computing the gradient flow solution using a fourth-order Runge-Kutta integrator (Runge, 1895). At each iteration step, we record the sharpness of the training loss. We also save model checkpoints whenever the training loss first drops below a power of ten (i.e., 10^{-1}, 10^{-2}, etc.). From this gradient flow trajectory, we extract two key statistics: the sharpness at initialization (s_0) and the maximum sharpness observed during the trajectory (s_{GF}). The values $1/s_0$ and $2/s_{\mathrm{GF}}$ are of particular interest. Taking the learning rate of $1/s_0$ has been suggested as a heuristic for optimal step size selection for non-adaptive GD (Cohen et al., 2021), and for learning rates above $2/s_{\mathrm{GF}}$, the well-known stability condition (2) is violated at some point of the gradient flow trajectory, suggesting that the loss decrease is not guaranteed there.

We construct the learning rate schedule for each configuration using two regular grids: a fine grid focused on the critical transition region, and a coarse grid which allows us to study the trade-off of the regularization in the EoS regime.
The fine grid consists of 12 points uniformly spaced with step size $\frac{1}{2s_{\mathrm{GF}}}$ in the interval $\left[\frac{1}{2s_{\mathrm{GF}}}, \frac{6}{s_{\mathrm{GF}}}\right]$. The coarse grid includes nine uniformly spaced learning rates interpolated in the interval $\left[\frac{6}{s_{\mathrm{GF}}}, \frac{2}{s_0}\right]$, and additionally includes all learning rates sampled at the step size $\frac{1}{8} \cdot \left(\frac{2}{s_0} - \frac{6}{s_{\mathrm{GF}}}\right)$ which are strictly greater than zero, and above until divergence. If we observe divergence already within the $\left[\frac{6}{s_{\mathrm{GF}}}, \frac{2}{s_0}\right]$ interval, we manually refine the schedule by decreasing the step size.

For each learning rate in the schedule, we train the model using full-batch gradient descent until the training loss falls below a fixed threshold (see table 1 for the exact configuration). During training, we record the sharpness and ℓ_1-norm every 10 epochs, and similar to the gradient flow experiments, we save the model checkpoints at every power-of-ten loss threshold. To compute the Hessian, we approximate its leading eigenvalues using the Lanczos algorithm applied to Hessian-vector products, which can be efficiently computed via backpropagation (Pearlmutter, 1994).

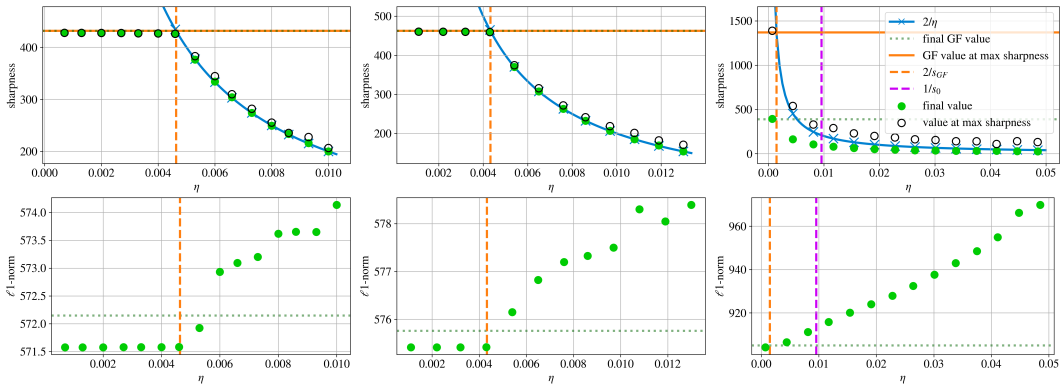
All experiments are fully reproducible, and the code is available in the supplementary material. Our implementation builds upon the original code by Cohen et al. (2021).

We ran the experiments on a heterogeneous computing infrastructure. Our hardware included NVIDIA A100, RTX 2080 Ti, TITAN RTX, RTX 3090 Ti, and RTX A6000 GPUs. Because GPU performance and availability varied across machines, we do not report a precise total runtime. However, the study required substantial computational effort: for each of the more than a dozen model configurations, we evaluated at least 20 learning rates, with individual runs ranging from a few minutes (for small models) to hundreds of hours (for larger models).

H Effect of training configuration on sharpness–norm trade-off

As described in Section 2.1, we systematically investigate variants of our base configuration (fully-connected ReLU feed-forward network (FCN) with three layers, 200 hidden neurons each, trained on the first $5,000$ examples of MNIST or CIFAR with mean squared error) to demonstrate the relationship between sharpness and implicit regularization for varying step size.

We vary the dataset size, architecture, activation functions, loss functions, initialization and parameterization. While quantitative metrics such as the critical learning rate η_c and absolute sharpness values differ, we consistently observe the norm-sharpness regularization trade-off.



(a) MNIST-5k, CNN-ReLU, MSE (b) MNIST, CNN-ReLU, MSE (c) CIFAR-10-5k, CNN-ReLU with Batch Normalization, CE

Figure 5: Different configurations using the CNN architecture. We observe that the ℓ_1-norm increase flattens out more towards larger η in comparison to the FCN.

In the following sections, we describe the findings on each variation and illustrate it with few representative plots. In all cases, we observe the same overall qualitative behavior. Additional supporting plots are included in the systematic overview of all experimental runs across configurations, provided in Appendix I and summarized in Table 1. For each of these configurations, we present both the coarse and fine-grained learning rate schedules to emphasize the transition region around η_c as well as the behavior at larger learning rates.

H.1 Dataset size

Most of our experiments use a subset of $5,000$ training examples of MNIST and CIFAR-10 respectively, chosen to allow tractable estimation of sharpness across a wide range of learning rates. To confirm that our findings are not specific to the small dataset sizes, we run a limited number of configurations on the full MNIST and CIFAR-10 training sets. In Figure 5, we show the comparison of the sharpness and ℓ_1-norm for a CNN with ReLU activation for MSE loss. The GF solution changes slightly, but the overall phenomena persists and the values are relatively similar. We present additional figures on the full MNIST (see Appendix I.1.3, I.3.2, I.4.1) and full CIFAR (I.1.4) in Appendix I.

H.2 Architecture

Our base model is a two-hidden-layer fully connected neural network (FCN), where each hidden layer consists of 200 neurons, with input and output layer sizes depending on the dataset.

To study the influence of the FCN architecture, we vary its widths and depths, namely experiments with $2\times$, $3\times$, and $10\times$ width, while keeping depth fixed, $2\times$ and $3\times$ depth, keeping width fixed, and $2\times$ and $3\times$ both width and depth. In other words, the considered FCN model shapes are: 200×2, 400×2, 600×2, 2000×2, 200×4, 200×6, 400×4, and 600×6 where the first number is the number of hidden neurons per hidden layer, and the second corresponds to the number of hidden layers.

While across most of these experiments the sharpness-norm tradeoff is ever-present and consistent with the behavior of the standard model, increasing width alone on the MNIST-5k dataset leads to a dissolution of the trend of increasing norm. Here in the EoS regime the norm first decreases and then stays near constant (Figures 37, 38, and 39). However, we believe this to be the result of the limited range of learning rates, since for experiments increasing both width and depth we can see a similar decrease in norm at first, but a robust overall increase afterwards (Figures 42 and 43).

We further extend our analysis beyond the fully connected baseline by evaluating several alternative architectures: Convolutional networks (CNNs) with ReLU activations (Figure 5and Appendix I.3), ResNet (Appendix I.5), and a Vision Transformer (Appendix I.4). For CNNs, the ℓ_1-norm flattens out more for increasing η in comparison to the FCN. For the CNN with Batch Normalization, comparably higher learning rates still converge. We do not observe a qualitative change of the phenomena for the ResNet and ViT architectures.

The CNNs (Lecun et al., 1998) consist of two convolutional layers with 32 filters, each using 3×3 kernels, stride 1, and padding 1. Each convolution is followed by an activation function (ReLU or tanh) and a 2×2 maximum pooling operation. A fully connected layer after flattening maps the features to class logits. We further include an alternative architecture that applies batch normalization within the CNN.

The ResNet-20 model (He et al., 2016) consists of three residual layers, with three blocks per layer. Each block contains two 3×3 convolutions followed by batch normalization and ReLU activation. Between stages, spatial down-sampling is performed using average pooling. To match feature dimensions across residual connections, the skip paths are adjusted using batch normalization and zero-padding along the channel dimension.

The Vision Transformer (ViT) (Dosovitskiy et al., 2021) splits the input image into non-overlapping patches (7×7 for MNIST, 4×4 for CIFAR-10), embeds each patch into a latent space (dimension 64 for MNIST, 128 for CIFAR-10), and processes the resulting sequences with transformer encoder layers (4 for MNIST; 6 for CIFAR-10), using 4 attention heads per layer. Each configuration includes a learnable class token and positional embeddings, and ends with a linear classifier applied to the class token output.

H.3 Activation function

We evaluate the effect of activation functions by comparing ReLU and tanh in fully connected networks on MNIST-5k (Appendix I.1.1, I.2.1) and on CIFAR-10-5k (Appendix I.1.2, I.2.2). Across all configurations, the sharpness–norm trade-off and the transition between flow-aligned and EoS regimes are consistently observed.

H.4 Loss function

We compare the behavior of cross-entropy (CE) and mean squared error (MSE) for both the base configuration and additional architectures, see Figure 5 for a comparison of the trade-off comparing both MSE and CE for MNIST-5k for a ReLU CNN and Appendix I for all other setups.

Compared to MSE, the sharpness profile for varying η when training with CE differs. In the flow-aligned phase, the final sharpness values for CE are still similar in magnitude but consistently below the maximum sharpness of its corresponding GF. In contrast, for MSE the final sharpness is at s_{GF}. The transition to the EoS regime still occurs approximately at $\eta = 2/s_{\mathrm{GF}}$. For large η, the sharpness values remain below the $2/\eta$ curve but qualitatively still decrease as η increases for the EoS regime.

We observe for the sharpness of the iterates during training that after an initial increase (progressive sharpening) and an oscillatory phase around $2/\eta$, the sharpness subsequently decreases again significantly. This phenomenon, originally remarked in Cohen et al. (2021), appears more pronounced in our results, as they used a higher loss-threshold beyond which the strong decrease starts occurring. Although the final sharpness values therefore do not follow the $2/\eta$ relationship, the training iterates rise toward this value and oscillate around it before the sharpness drops. In our plots, we visualize the smoothed sharpness around its maximum to highlight this trend. The effect during the training is illustrated in Figure 13 for selected learning rates.

Training with CE often fails to converge at learning rates even below $1/s_0$ (s_0 denoting the sharpness at initialization), while training with MSE often converges at comparatively higher values. This aligns with previous findings on the geometry of the log-loss landscape (Soudry et al., 2018), which indicate that the loss surface becomes flatter as the parameter

norm increases. Because of the exponential in the CE loss equation, the loss decreases with growing parameter norm and, as a result, parameters only converge in direction. However, when the learning rate is too high early in the training, the high curvature of the loss landscape leads to instability or stagnation before this directional convergence effect.

H.5 Loss threshold

In Section 2, we show how the loss threshold ε directly affects the critical learning rate η_c at which (approximately) the sharpness–norm phase transition occurs, given by $2/s^{\varepsilon}_{\mathrm{GF}}$. This effect is illustrated in Figure 2 for an FCN with tanh activation on CIFAR-10-5k, trained with MSE loss. Comparing identical models trained to different loss thresholds, we observe that smaller ε values yield higher $s^{\varepsilon}_{\mathrm{GF}}$, resulting in a lower η_c and thus shifting the transition point between the flow-aligned and EoS regimes. We confirm this trend across multiple architectures in Appendix I.7.1.

This dependence on ε is naturally related to early stopping: A higher loss threshold corresponds to a point before the model begins to overfit on the training set, where the test loss is still decreasing. In contrast, very small loss thresholds reflect the late phase of training, where the characteristic U-shaped test loss curve over time is evident. There, the training loss continues to drop, but the test loss increases slowly. By varying ε, we can thus study the sharpness and norm trade-offs under different degrees of overfitting. However, note that we do not link ε to the validation loss, as it is commonly done when using early stopping as a regularizer during training.

H.6 Initialization

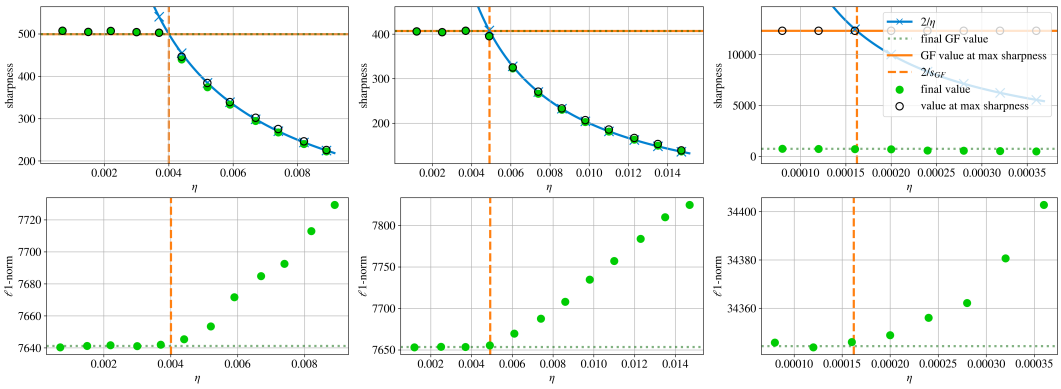
We vary the initialization seed in fully connected networks trained on CIFAR-10-5k to test the sensitivity of the transition to random initialization, see Figure 6. While the critical learning rate η_c shifts with initialization, due to a different initial sharpness s_0 and maximum of the flow trajectory s_{GF}, the qualitative structure remains intact.

We also perform experiments with increased initialization scale, scaling all initial weights $\times 5$ and $\times 10$. As a result, the maximal sharpness along the trajectory occurs already at initialization, which drastically alters the optimization dynamics and sharpness evolution. The sharpness decreases at first, and, if reaching the $2/\eta$ threshold, oscillates around this value. In general, the training is highly unstable with leads to divergence of the training at many small learning rates. Still, the $\times 5$-scaled initializations result in somewhat similar qualitative behaviors in the observed values as our default scale. For $10\times$ scaling, the training diverges already at learning rates smaller than η_c. In addition, the final ℓ_1-norm reaches very high values and decreases with increasing learning rate. These results suggest that the mechanism of implicit regularization differs at such large scales. We note that this aligns with previous works on EoS which often implicitly assume a sufficiently small initialization to permit progressive sharpening.

We provide further figures with varying initialization seeds and scales in Appendix I.7.2 and I.7.3, respectively.

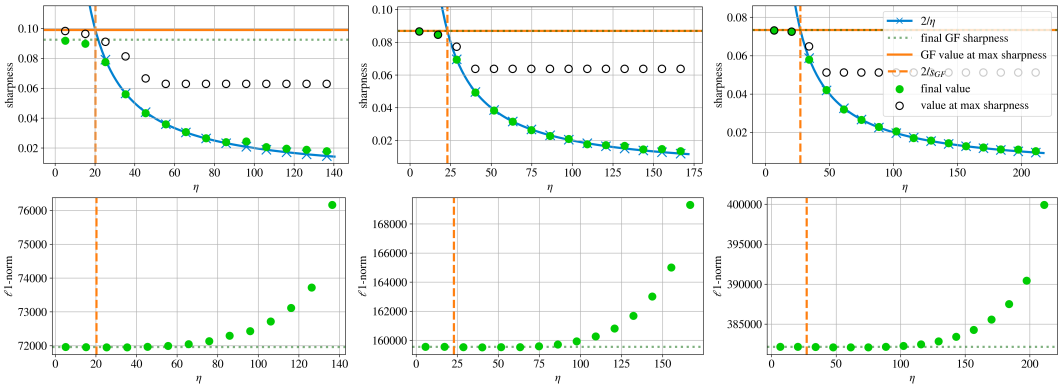
H.7 Parameterization

Different parameterizations of the forward pass are known to place training in qualitatively different regimes with respect to feature learning (Noci et al., 2024), which is why we test the norm-sharpness tradeoff for this setup. We focus on the μP and kernel parameterizations (Yang et al., 2022; Jacot et al., 2018). The kernel parameterization corresponds to NTK-like scaling, where feature learning diminishes with width, while μP remains in the feature-learning regime with width-independent gradient magnitudes and transferable learning-rates for models of varying widths (Yang et al., 2022). Recent work by Noci et al. (2024) further suggests that the Hessian spectrum also transfers for μP.



(a) Init. seed 43 (b) Init. seed 44 (c) Init. seed 43, scaled $\times 5$

Figure 6: Effect of varying initialization seed and scaling at initialization on the sharpness–norm trade-off. All columns show sharpness and ℓ_1-norm curves for the same architecture (FCN-ReLU), dataset (CIFAR-10-5k), and loss function (MSE), all trained until loss 0.01. While the different seed does not affect the overall behavior, scaling disrupts adherence of solution sharpness to the $2/\eta$ curve. Effect on norm is however preserved.



(a) Width 100 (b) Width 200 (c) Width 400

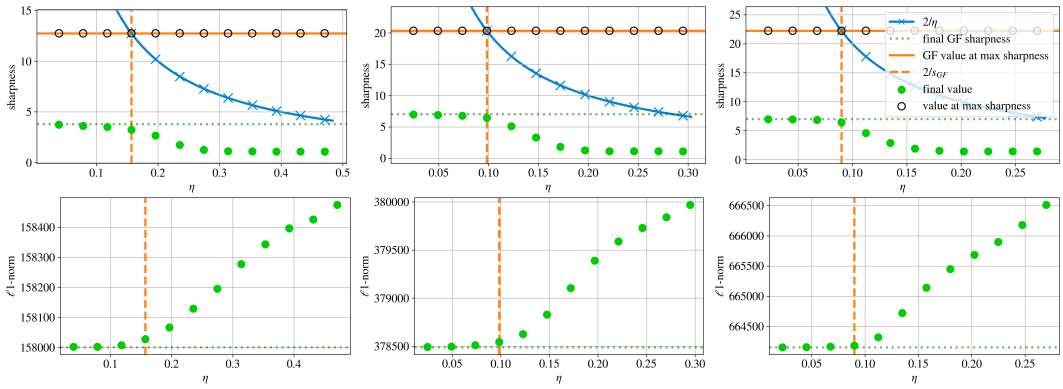
Figure 7: Sharpness (top row) and ℓ_1-norm of final classifiers (bottom row) for μP parametrization with widths 100, 200, and 400 on MNIST-5k with MSE and loss goal 0.1.

Both used parameterizations use fully connected feed-forward networks with ReLU activations. Each hidden layer of width n_l computes

$$h_l = \frac{1}{\sqrt{n_{l-1}}}\sigma\left(W_l h_{l-1}\right)$$

with weights initialized as $(W_l)_{ij} \sim \mathcal{N}(0,1)$. In the kernel parametrization the final layer is obtained as $f(x) = W_L h_L$, while in the μP parametrization the logits are rescaled by the width of the last hidden layer $f(x) = \frac{1}{\sqrt{n_L}} W_L h_L$. This differs from the normal parameterization in all other experiments where the $1/\sqrt{n_{l-1}}$ factor in the forward pass is missing and the weights are initialized uniformly with variance $1/(3n_{l-1})$. The hypothesis spaces are the same in both settings, however the reparameterization changes the dynamics and is hence of interest with respect to implicit regularization.

For the μP parameterization, the sharpness plots (top row of Figure 7) show approximately constant sharpness for small learning rates and a decrease along the $2/\eta$ curve for larger learning rates, with similar values in the flow-aligned regime across widths. The ℓ_1-norm plots (bottom row) reveal the usual pattern across widths of increasing final parameter ℓ_1 for increasing learning rate. The absolute norms differ due to model size, but the growth



(a) Width 200 (b) Width 400 (c) Width 600

Figure 8: Sharpness (top row) and ℓ_1-norm of final classifiers (bottom row) for kernel parametrization with widths 200, 400, and 600 on MNIST-5k with MSE and loss goal 0.1.

of the norm as η increases is approximately consistent (though divergence happens slightly earlier for smaller models). This is expected for the μP parametrization, as the parameter update magnitudes are independent of the model width. After rescaling the learning rate proportionally to width, the results align across the models of different widths which matches the results by Noci et al. (2024).

For the kernel parametrization we observe that the ℓ_1-norm of the parameters (bottom row of Figure 8) remains stable for small learning rates and starts to increase once η crosses the critical threshold, with the transition occurring at learning rates of the same order across widths[6]. The sharpness plots (top row) show that the maximum sharpness coincides with the sharpness at initialization, similar to the large-initialization experiments in Section H.6. Because of the different parameterization, sharpness no longer tracks the $2/\eta$ curve, yet the qualitative pattern is consistent across widths: sharpness stays flat below the threshold and decreases gradually thereafter.

H.8 Number of iterations

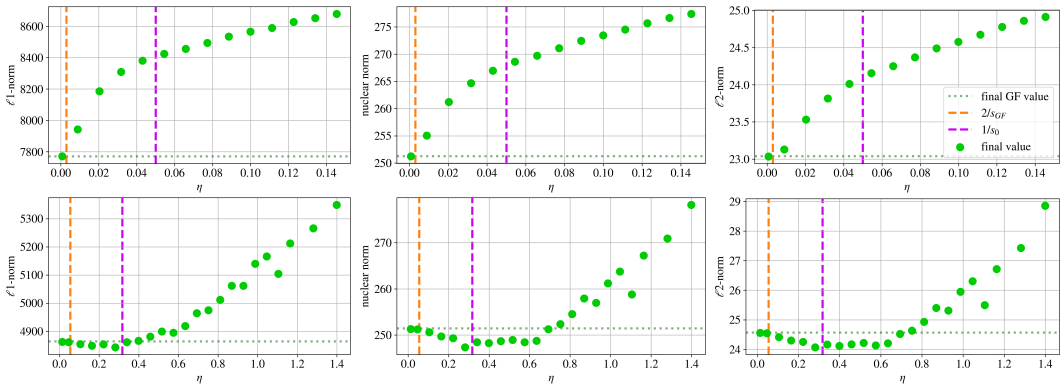
A notable difference between the two regimes lies in the relationship between learning rate and convergence speed. While the small learning rates of the flow-aligned regime lead to slower convergence in absolute terms, increasing the step size within this regime significantly accelerates optimization, with the number of iterations required to reach a fixed training loss decreasing at an approximate rate of $1/\eta$. As further shown in Section I.8.1, this rate of convergence speed acceleration with respect to the learning rate is higher in the flow-aligned regime than in the EoS regime.

H.9 Alternative norms and sharpness measures

In most of the paper, we focus on the ℓ_1-norm of the GD solution. In Figure 9, we compare the ℓ_1-norm to the nuclear and ℓ_2-norms, which look qualitatively similar. We provide more examples in Section I.8.1.

Similarly, as our primary measure of sharpness we use throughout most of the paper the top eigenvalue of the loss Hessian. This notion of sharpness, though commonly used, has been shown to allow for being made arbitrarily large by means of reparametrization without affecting generalization (Dinh et al. (2017)). This can can make it ill-suited for studying connections to generalization performance. Therefore in Figure 10 we compare different notions of sharpness, including re-scaling invariant measures such as adaptive sharpness

[6]Note that the norm of the weight matrices (after adjusting for the different widths) differs slightly due to the randomness. The change in randomness is comparable to the variance indicated by experiments when changing the initialization seed, see Section H.6.



(a) ℓ_1-norm (b) Nuclear norm (c) ℓ_2-norm

Figure 9: Each row shows the ℓ_1-norm, the nuclear norm, and the ℓ_2-norm of the solution for different models - both use FCN-ReLU with MSE loss, in the top row on CIFAR-10-5k, in the bottom row on MNIST-5k. As expected, the behavior of the different norms is approximately equivalent

(Kwon et al. (2021)), showing they share the overall decreasing behavior in the EoS regime similar to the worst-case sharpness.

H.10 Gradient descent solution distance

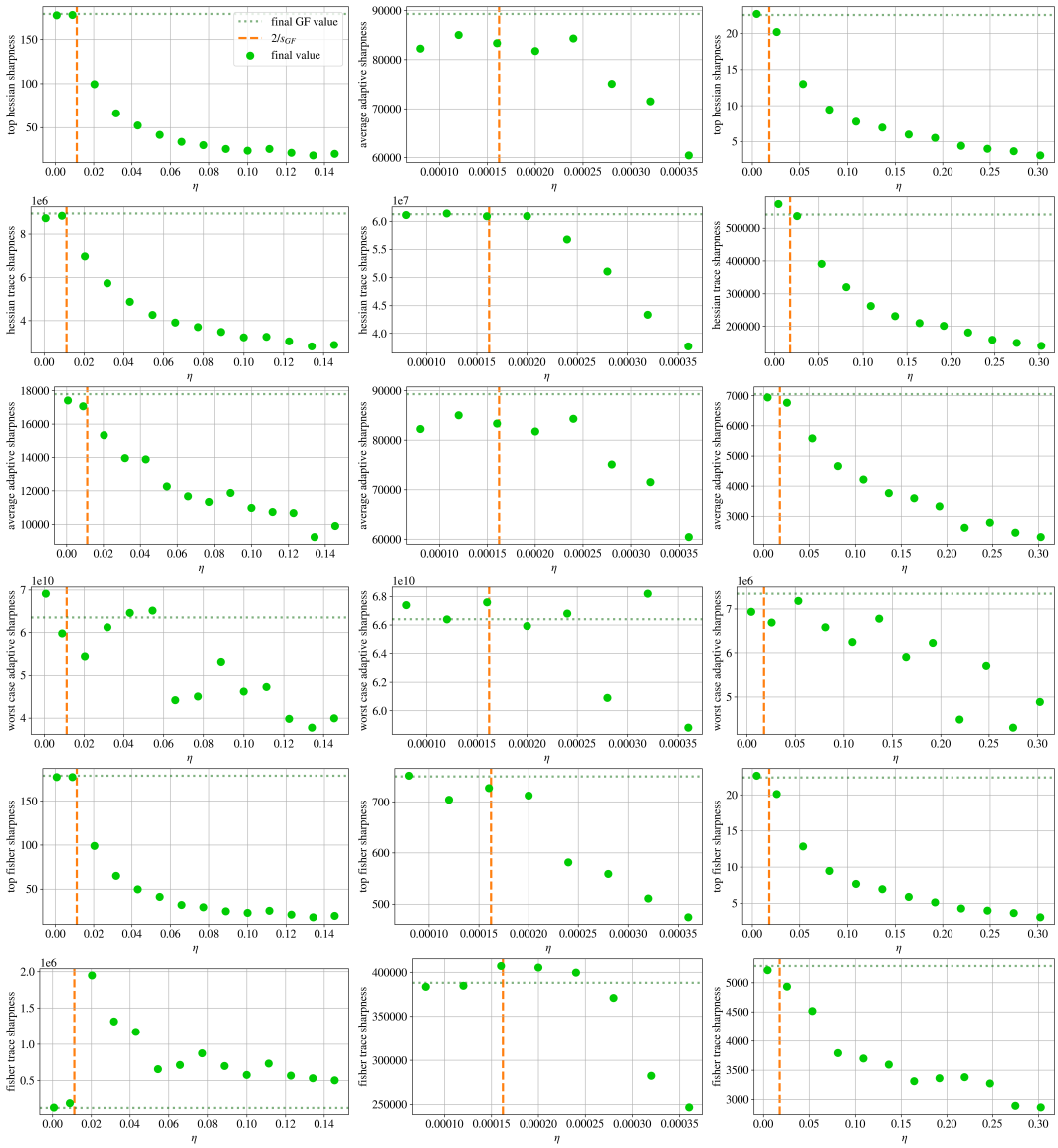
We measure the distance between the final solutions of GF and GD across different learning rates. This analysis provides insight into how closely GD tracks the continuous-time dynamics and how this relationship evolves as we move through the flow-aligned and EoS regimes.

In Figure 11, we show this relationship for two of our standard models. Comparing this figure with Figure 9, we can see that even though the qualitative behavior of the ℓ_1-norm and ℓ_1-distance from the GF solution are nearly equal, the distance of solutions for $\eta < \eta_c$ is already relatively high. This suggests that while in the flow-aligned regime, GD reaches solutions of similar sharpness and norm as GF, in absolute terms these solutions are non-negligibly different. Furthermore, comparing the scales of the two figures shows, that the increase in distance from the GF solution is much larger than the increase in absolute ℓ_1-norm. Therefore, increasing the learning rate within the EoS regime likely results in movement of the solution in a direction more misaligned with the GF solution than the origin. Section I.8.1 shows this for further configurations.

Additionally, in Figure 12 we compare the parameter ℓ_1-norm to the ℓ_1-distance from the untrained model at initialization. When examining this quantity for the final learned models plotted against the learning rate, the distance from initialization shows a similar qualitative trend as the parameter norm. In the flow-aligned regime, the distance to initialization is still approximately constant, before robustly increasing in the EoS regime. This is consistent with what can be expected since the models are initialized small relative to the norm of the final parameters.

H.11 Evolution during training

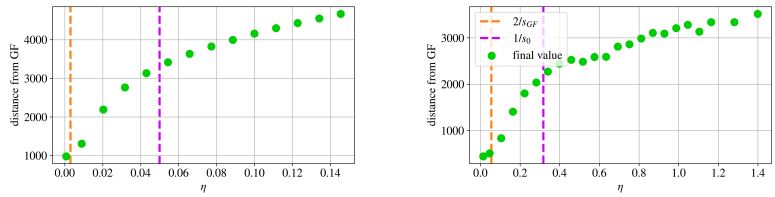
In Figure 13, we illustrate how sharpness, ℓ_1-norm and loss evolve over the course of training in intrinsic time, i.e $\eta \cdot \#$ iterations. The sharpness increases initially (progressive sharpening) until reaching $2/\eta$, and then oscillates around this value. For very small learning rates, the increase stops earlier (aligned with the maximum sharpness of the corresponding GF). The norm rises without oscillation, suggesting that the oscillation occurs along a direction that preserves the parameter norm. The norm grows faster for larger learning rates. The loss decreases monotonically at first, then with oscillation after the sharpness has risen to $2/\eta$.



(a) MSE loss, train loss 0.1 (b) MSE loss, train loss 0.1, initialization scaled $\times 5$ (c) CE loss, train loss 0.01

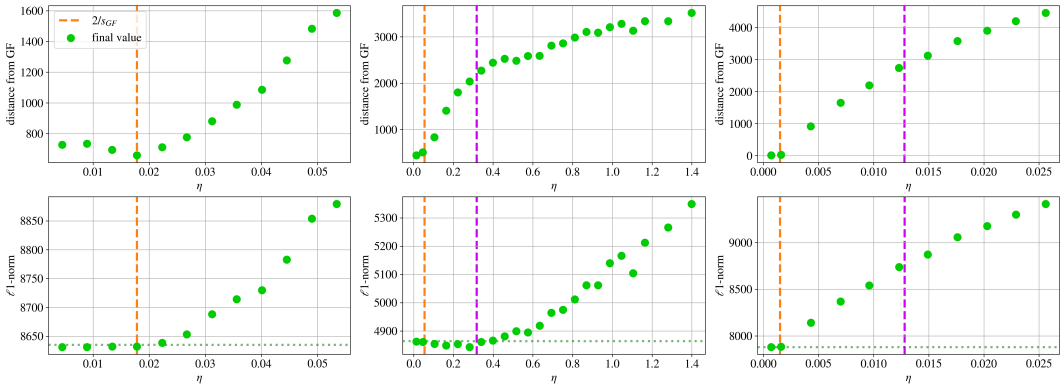
Figure 10: Each column represents a different setting: All display an FCN-ReLU network on CIFAR-10-5k, but in the first we show MSE loss with standard initialization, in the second MSE loss with scaled initialization and in the last CE loss. Each row shows a different measure of sharpness. Top to bottom these are: top eigenvalue of the loss Hessian (used throughout the paper), trace of the loss Hessian, average-case and worst-case adaptive sharpness (Kwon et al. (2021)), and top eigenvalue and trace of the Fisher information matrix (Liang et al. (2019)). Note that all measures display a general decreasing behavior with the exception of the Fisher trace on standard MSE loss (bottom left), where there is a sharp increase around the critical threshold η_c, from which the decreasing behavior starts. The scaled experiments show slightly more irregularity, but still preserve this general decrease.

In contrast to MSE loss, for training with CE loss, the sharpness decreases again after a period of oscillation. These dynamics in sharpness and loss were first systematically studied by Cohen et al. (2021). Our primary focus is on the dependence of final values on the learning rate, which complements these observations.



(a) FCN-ReLU on CIFAR-10-5k (b) FCN-ReLU on MNIST-5k

Figure 11: ℓ_1-distance of the GD solution from the GF solution. Not to be confused with distance from the GF trajectory - here we measure only final values. On both examples we can see an increasing behavior similar to that of solution ℓ_1-norm.



(a) FCN-ReLU on CIFAR-10-5k with CE loss (b) FCN-ReLU on MNIST-5k with MSE loss (c) FCN-tanh on CIFAR-10-5k with MSE loss

Figure 12: The top row shows for each setting the ℓ_1-distance of the final models from their initialization, while the bottom row shows the absolute norm. As expected, the qualitative behavior remains almost identical.

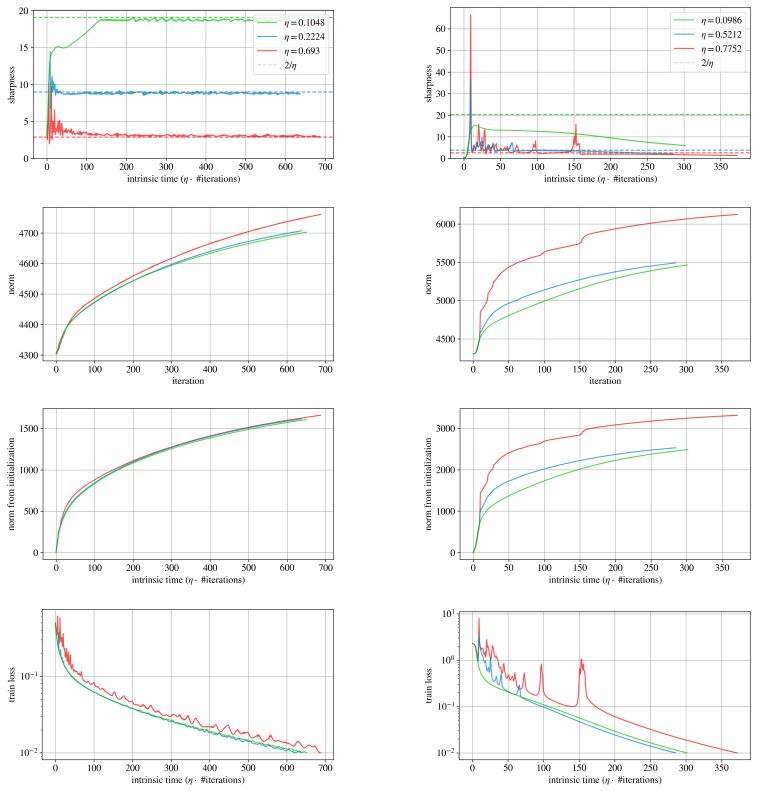
Similar to Figure 12, we compare the evolution of the parameter norm and the distance to the initialization in the second and third row of Figure 13. We observe that the distance follows closely a translated and scaled version of the parameter norm's trajectory. It naturally starts at 0 and then grows significantly before entering the Edge of Stability. In comparison to the parameter norm evolution, here the rate of growth slows down to a larger extent after entering EoS, which supports the intuition that the chaotic EoS updates have a smaller cumulative effect on the solution's magnitude.

H.12 Per-layer norms

In Figure 14 we present the layer norms when training the standard ReLU FCN on MNIST-5k and CIFAR-10-5k. As one can see, all layers show an increasing trend. As one might expect, the increase is relative to the number of parameters of the respective layer.

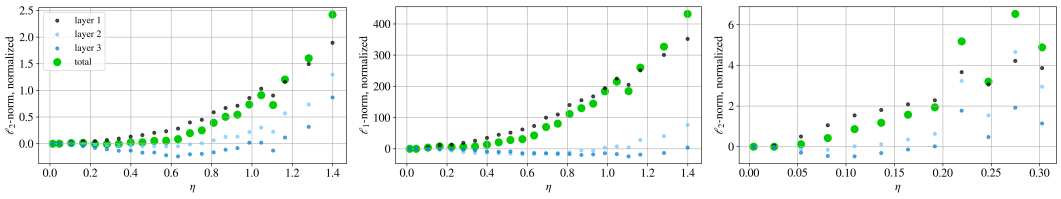
H.13 The diagonal network

For the diagonal network discussed in Section 3, we present the sharpness, norm, and generalization values for different learning rates in Figure 15. We can explicitly compute the ℓ_1-norm on the solution manifold under the sharpness constraint $2/\eta$, yielding the predicted line in Figure 15b. We emphasize that these curves look qualitatively similar to the more realistic models on MNIST and CIFAR-10 described throughout the empirical experiments section. Note that divergence occurs already for learning rates η below the theoretical divergence threshold when the sharpness of all points on the solution manifold is above $2/\eta$.



(a) MSE loss. (b) CE loss.

Figure 13: For three different learning rates, we display the sharpness, ℓ_1-norm, norm from initialization and train loss for both MSE (left) and CE loss (right column), both on MNIST-5k, FCN-ReLU, loss goal 0.01. We clearly observe the progressive sharpening and oscillations once the sharpness reaches $2/\eta$. For CE loss, the sharpness at the iterates drop after a oscillatory phase.

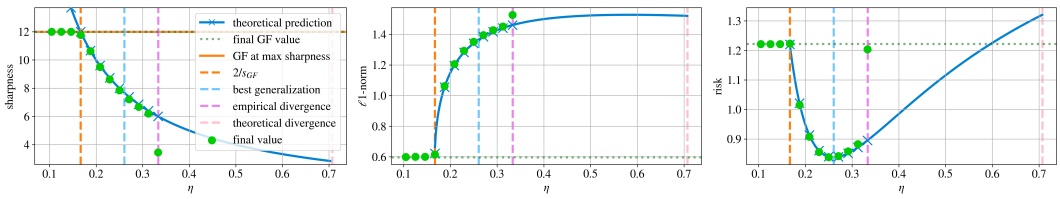


(a) MNIST-5k, MSE, ℓ_2-norm (b) MNIST-5k, MSE, ℓ_1-norm (c) CIFAR-10-5k, CE, ℓ_2-norm

Figure 14: Layer-wise norms of the final solution our ReLU-FCN on MNIST-5k and CIFAR-10-5k for different learning rates. We individually normalize each group by subtracting the value of the norm at the smallest learning rate. All layers show an increasing trend, which is relative to the layer size.

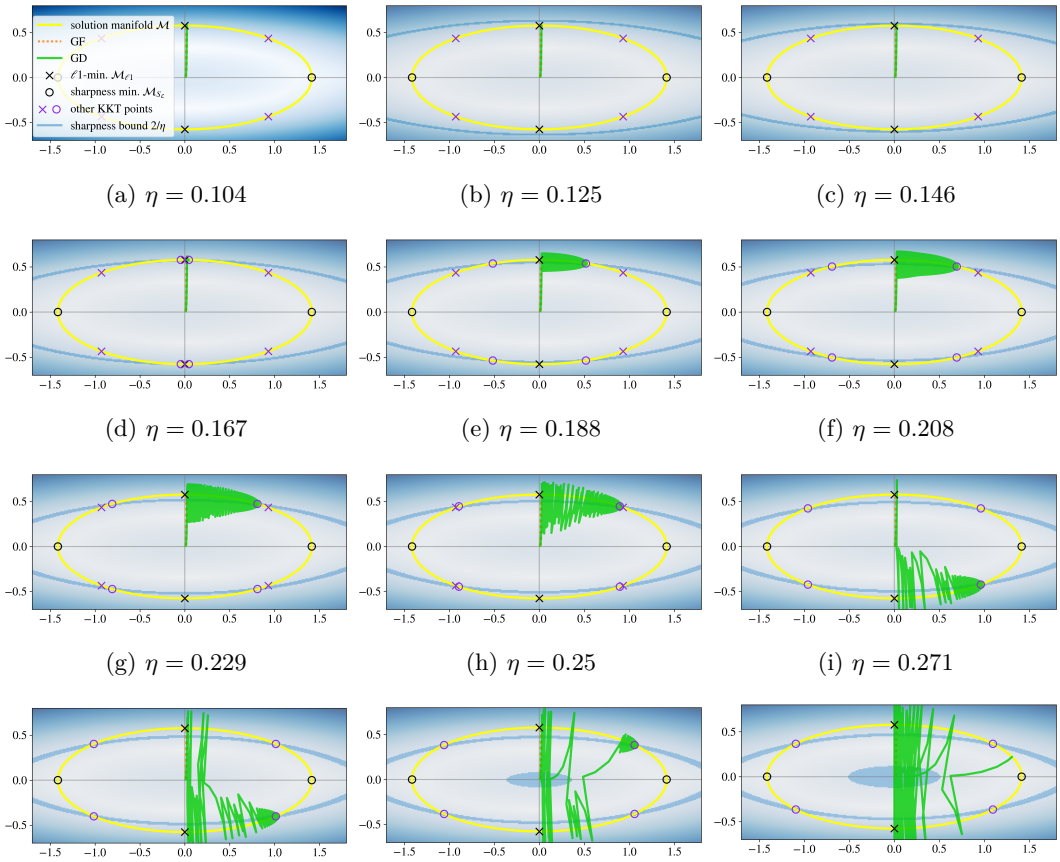
We model generalization using a simple Gaussian data distribution (see Appendix E), which produces an (idealized) U-shaped curve, consistent with the behavior observed for many other realistic setups.

In Figure 16, we provide all trajectories of the iterates (cf. Figure 4).



(a) Sharpness (b) ℓ_1-norm (c) Generalization

Figure 15: Final sharpness, ℓ_1-norm and generalization of a two-dimensional diagonal linear network with weight sharing, described in Section 3. The behavior corresponds to that of more realistic models studied throughout the paper.



(a) $\eta = 0.104$ (b) $\eta = 0.125$ (c) $\eta = 0.146$

(d) $\eta = 0.167$ (e) $\eta = 0.188$ (f) $\eta = 0.208$

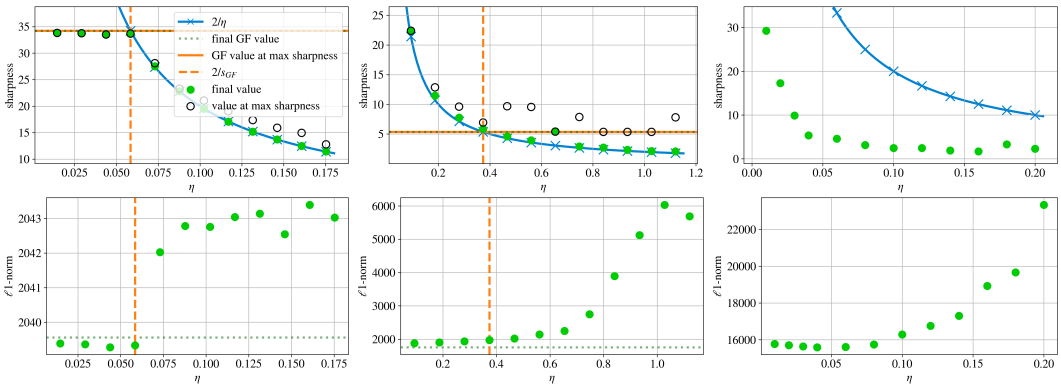
(g) $\eta = 0.229$ (h) $\eta = 0.25$ (i) $\eta = 0.271$

(j) $\eta = 0.291$ (k) $\eta = 0.312$ (l) $\eta = 0.333$

Figure 16: Iterates of weights of the two-dimensional diagonal linear network throughout training, for increasing learning rate. There is a clear distinction between the flow-aligned regime (16a)-(16d), where GD closely tracks the GF trajectory, and the EoS regime (16e)-(16l), where at some point GD begins to oscillate away from GF, until converging to one of the first solutions whose sharpness is less than $2/\eta$ (intersection of the yellow solution manifold $\mathcal{M}$ and blue sharpness bound). This aligns with the intuition stemming from Theorem B.2. In purple, we mark the KKT points from Lemma E.1.

H.14 Other data modalities

While the systematic evaluation presented in this paper focuses on the image domain, we also include examples suggesting that the observed trade-off is not limited to images. We consider



(a) Breastcancer, MSE (b) California Housing, MSE (c) Reverse Sequence, CE[7]

Figure 17: We show the sharpness, ℓ_1-norm and test loss for two tabular and a sequence-to-sequence data set. This indicates that our results extend beyond the image sector.

a synthetic sequence-reversal task and two tabular tasks, one for binary classification and one for regression.

For the sequence domain, we use a synthetic sequence-reversal task with a fixed sequence length 10 and vocabulary size 9. Each input is sequence of 10 tokens sampled uniformly from $\{1, \ldots, 9\}$. The target is its exact reversal. We train using teacher forcing. The model is a standard encoder-decoder transformer (Vaswani et al., 2017) with two encoder and two decoder layers, each using four attention heads, a model dimension of 64, and a feed-forward width of 128. The inputs pass through learned token embeddings and fixed sinusoidal positional encodings, and the decoder uses a causal mask for autoregressive prediction. A linear layer maps decoder outputs to vocabulary logits.

For the tabular tasks, we use the california housing (regression) and breastcaster (classification) dataset by scikit-learn (Pedregosa et al., 2011). The california housing dataset contains aggregated demographic and housing features (e.g., average number of rooms) and to be predicted is the median house value. The beast cancer Wisconsin dataset contains 30 cell nuclei features such as radius or texture, and the goal is to identify whether a tumor sample is malignant or benign. For both datasets, we standardize all input features by subtracting the training-set mean and dividing by the training-set standard deviation for each feature dimension, and we apply the same transformation to the targets. The model is our standard feed-forward network with two-hidden layers and width 200.

For both data modalities, we observe the similar characteristic trade-off of sharpness and norm which we show in Figure 17 In contrast, the sharpness value is not constant but increasing when decreasing the learning rate.

I Systematic overview of experiments

All performed experiments are summarized in Table 1. For most of these configurations, we present both coarse and fine-grained learning rate schedules to emphasize the transition region between flow-aligned and EoS regime around η_c, as well as the behavior at larger learning rates, demonstrating the trade-off between increasing ℓ_1-norm and decreasing sharpness for varying the learning rate. Table 1 specifies for each setting the following attributes:

- **Model.** We state the model architecture (see Section H.2) and activation used. For the FCN models where we vary width and depth, we also indicate the size. When we do not specify a size, we refer to the standard architecture of 200×2.

[7] We do not include the GF lines as we only run GD for this setup.

- **Dataset.** MNIST or CIFAR-10, with the "-5k" suffix indicating that we train only on the first 5000 data points of the train set, while still testing on the full test set.

- **Loss.** Mean square error (MSE) or cross-entropy (CE).

- **Seed.** The random seed used for generating weights at initialization. For experiments using a scaled initialization, the scaling factor is given.

- **Loss Goal.** We stop training gradient flow and gradient descent for each learning rate upon reaching this train loss value.

- **U-Shape.** For each setting we state whether optimal test loss aligns with either learning rate extreme, indicating a generalization advantage of either low-norm or low-sharpness bias. Settings where the optimum is attained for mid-range learning rates are marked by ✓, settings with an alignment towards either extreme by ×, and somewhat inconclusive settings by either mark in brackets. In our experiments, in all cases with a clear optimum extreme alignment, the alignment is always towards high learning rates, that is, towards low sharpness solutions.

- **Figures.** List of figures throughout the paper where the respective setting appears.

In the main part of the systematic review, we present for each setting sharpness, ℓ_1-norm and test loss plots, for both a fine-grained set of learning rate values focused around the critical threshold and a coarse set showing large-scale behaviors. In the plots we show

- the final respective value attained for each learning rate represented by green dots;

- a horizontal dotted green line indicating the final value reached by the gradient flow;

- a vertical dashed orange line showing the critical learning rate threshold of $2/\eta_{GF}$, for the transition from the flow-aligned to the EoS regime;

- for coarse-grained plots, a vertical dashed purple line, indicating the inverse value of sharpness at initialization, which has been proposed as a heuristic for learning rate initialization, if the line is missing this means that the GD did not converge for such learning rate;

- for sharpness plots, the $2/\eta$ curve, for η being the learning rate variable, shown in blue with crosses at each used learning rate value;

- for sharpness plots, the maximum value reached throughout training, indicated by black circles;

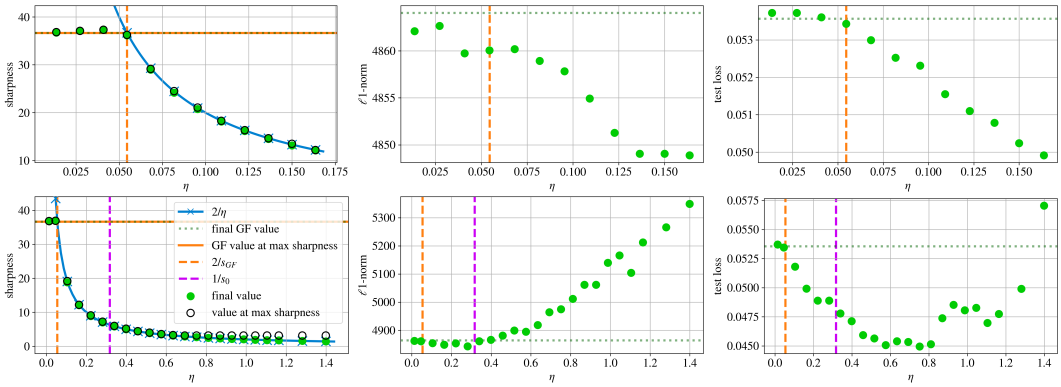
- for sharpness plots, a horizontal orange line showing the maximal GF sharpness.

Table 1: Full list of experimental configurations.

Model	Dataset	Loss	Seed	Loss Goal	U-Shape	Figures
FCN-ReLU	MNIST-5k	MSE	43	0.0001	✓	3a,13,9,11b, 18,64,72
FCN-ReLU	MNIST-5k	MSE	43	0.001	✓	51
FCN-ReLU	MNIST-5k	MSE	43	0.01	✓	52
FCN-ReLU	MNIST-5k	MSE	43	0.1	✓	53
FCN-ReLU	MNIST-5k	CE	43	0.01	✓	19,65,73
FCN-ReLU	MNIST-5k	CE	43	0.1	✓	54
FCN-ReLU	CIFAR-10-5k	MSE	43	0.0001	×	1a,3c,9,11a, 20,68,76
FCN-ReLU	CIFAR-10-5k	MSE	43	0.001	×	55
FCN-ReLU	CIFAR-10-5k	MSE	43	0.01	×	6a,56
FCN-ReLU	CIFAR-10-5k	MSE	43	0.1	(×)	57,10
FCN-ReLU	CIFAR-10-5k	MSE	44	0.01	×	6b,59
FCN-ReLU	CIFAR-10-5k	MSE	45	0.01	×	60
FCN-ReLU	CIFAR-10-5k	MSE	43, ×5	0.1	×	6c,61,10
FCN-ReLU	CIFAR-10-5k	CE	43	0.01	✓	3b,21,69,77,10
FCN-ReLU	CIFAR-10-5k	CE	43	0.1	✓	58
FCN-ReLU	CIFAR-10-5k	CE	43, ×5	0.01	×	62
FCN-ReLU	CIFAR-10-5k	CE	43, ×10	0.01	×	63
FCN-ReLU	MNIST	MSE	43	0.01	✓	1b,22,66,74
FCN-ReLU	MNIST	CE	43	0.01	(✓)	23,67,75
FCN-ReLU	CIFAR-10	CE	43	0.1	×	24
FCN-ReLU 400 × 2	MNIST-5k	MSE	43	0.01	×	37
FCN-ReLU 600 × 2	MNIST-5k	MSE	43	0.01	(×)	38
FCN-ReLU 2000 × 2	MNIST-5k	MSE	43	0.01	×	39
FCN-ReLU 200 × 4	MNIST-5k	MSE	43	0.01	(×)	40
FCN-ReLU 200 × 6	MNIST-5k	MSE	43	0.01	(✓)	41
FCN-ReLU 400 × 4	MNIST-5k	MSE	43	0.01	×	42
FCN-ReLU 600 × 6	MNIST-5k	MSE	43	0.01	(✓)	43
FCN-ReLU 400 × 2	CIFAR-10-5k	MSE	43	0.01	×	44
FCN-ReLU 600 × 2	CIFAR-10-5k	MSE	43	0.01	×	45
FCN-ReLU 2000 × 2	CIFAR-10-5k	MSE	43	0.01	×	46
FCN-ReLU 200 × 4	CIFAR-10-5k	MSE	43	0.01	✓	47
FCN-ReLU 200 × 6	CIFAR-10-5k	MSE	43	0.01	✓	48
FCN-ReLU 400 × 4	CIFAR-10-5k	MSE	43	0.01	✓	49
FCN-ReLU 600 × 6	CIFAR-10-5k	MSE	43	0.01	✓	50
FCN-tanh	MNIST-5k	MSE	43	0.1	×	25
FCN-tanh	MNIST-5k	CE	43	0.01	(✓)	26
FCN-tanh	CIFAR-10-5k	MSE	43	0.001	×	2c,27,70,78
FCN-tanh	CIFAR-10-5k	MSE	43	0.01	×	2b
FCN-tanh	CIFAR-10-5k	MSE	43	0.1	(×)	2a
FCN-tanh	CIFAR-10-5k	CE	43	0.01	✓	28,71,79
CNN-ReLU	MNIST-5k	MSE	43	0.1	✓	5a,29
CNN-ReLU	MNIST-5k	CE	43	0.01	✓	30
CNN-ReLU	MNIST	MSE	43	0.1	(×)	5b,31
CNN-ReLU	MNIST	CE	43	0.01	✓	32
CNN-ReLU BN	CIFAR-10-5k	CE	43	0.01	5c,33	
ViT-ReLU	MNIST-5k	CE	43	0.1	(✓)	1c,34
ViT-ReLU	CIFAR-10-5k	CE	43	1	(✓)	35
ResNet20-ReLU	CIFAR-10-5k	CE	43	0.1	(×)	36

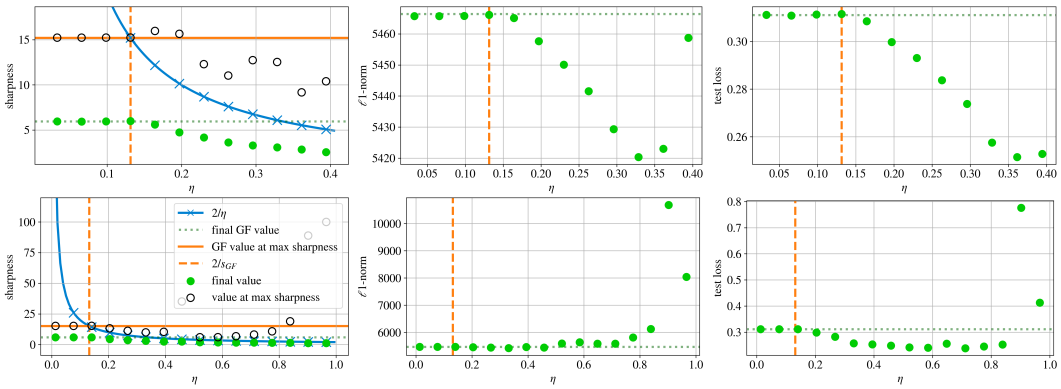
I.1 FCNs with ReLU activation

I.1.1 On MNIST-5k



(a) Sharpness (b) ℓ1-norm (c) Test loss

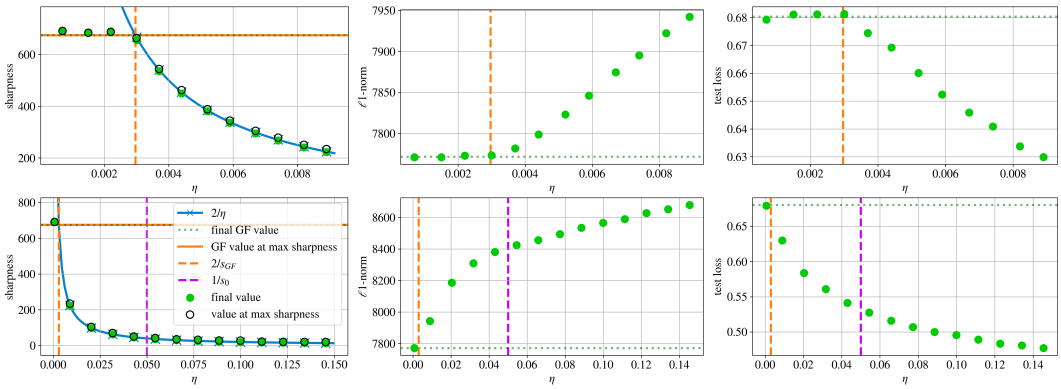
Figure 18: **MSE loss.** FCN-ReLU, MNIST-5k, train loss 0.0001. Both rows show the same setting, but different ranges of learning rate η - the top row includes the fine grid, focused on the transition from the flow-aligned to the EoS regime, while the coarse grid in the bottom row displays more large-scale behavior, going typically up to diverging learning rates.



(a) Sharpness (b) ℓ1-norm (c) Test loss

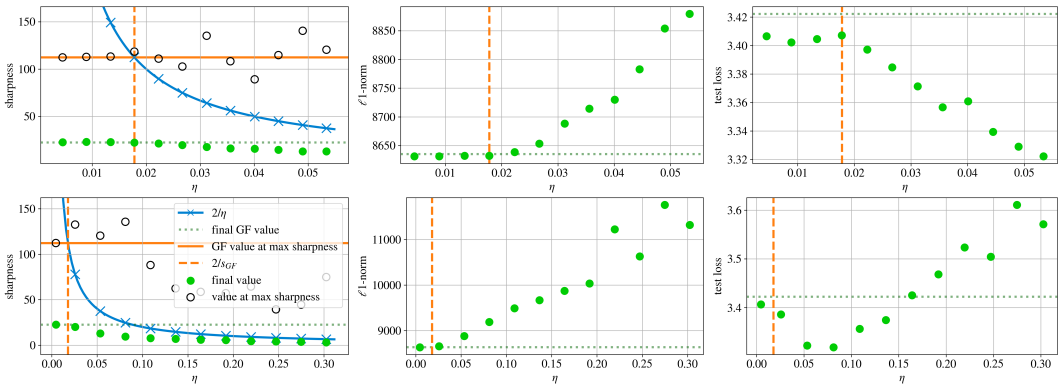
Figure 19: **CE loss.** FCN-ReLU, MNIST-5k, train loss 0.01

I.1.2 On CIFAR-10-5k



(a) Sharpness (b) ℓ1-norm (c) Test loss

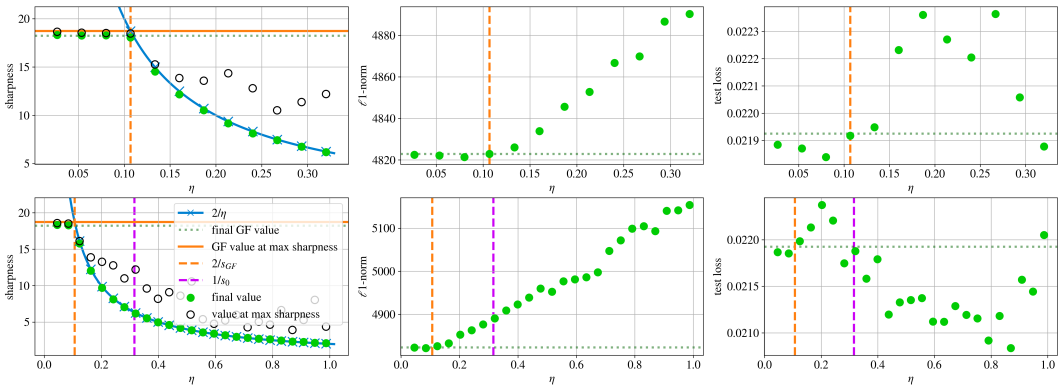
Figure 20: **MSE loss.** FCN-ReLU, CIFAR-10-5k, train loss 0.0001



(a) Sharpness (b) ℓ1-norm (c) Test loss

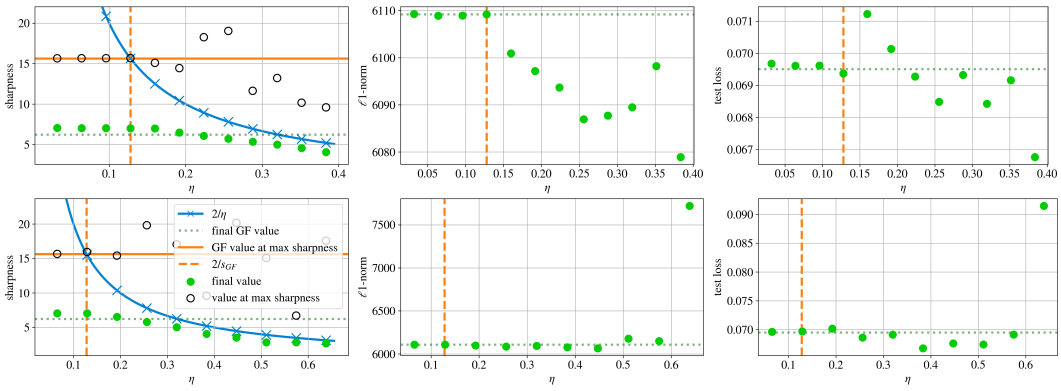
Figure 21: **CE loss.** FCN-ReLU, CIFAR-10-5k, train loss 0.01

I.1.3 On full MNIST



(a) Sharpness (b) ℓ1-norm (c) Test loss

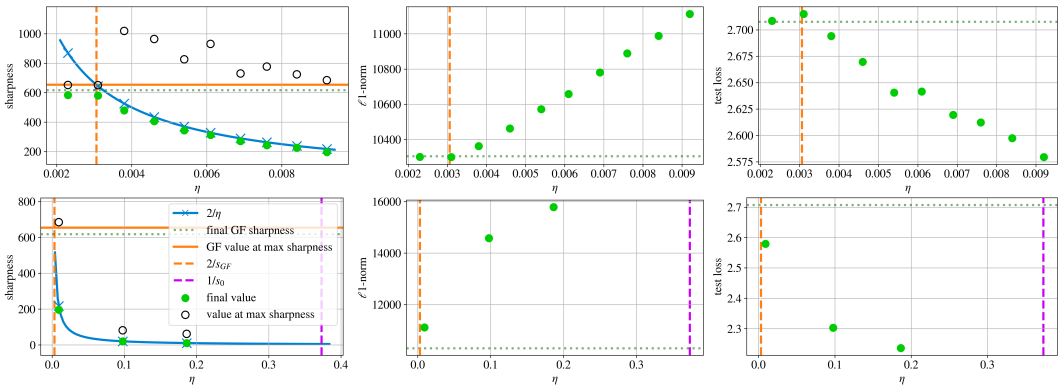
Figure 22: **MSE loss.** FCN-ReLU, MNIST, train loss 0.01



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 23: **CE loss.** FCN-ReLU, MNIST, train loss 0.01

I.1.4 On full CIFAR-10

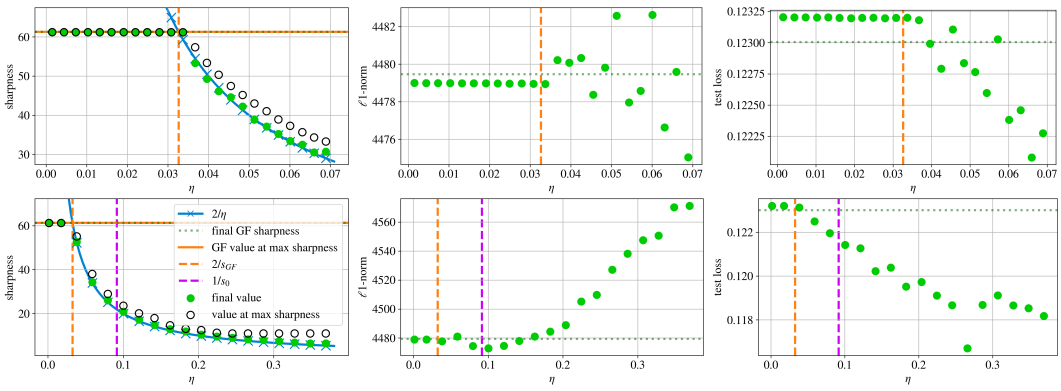


(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 24: **CE loss.** FCN-ReLU, CIFAR-10, train loss 0.1

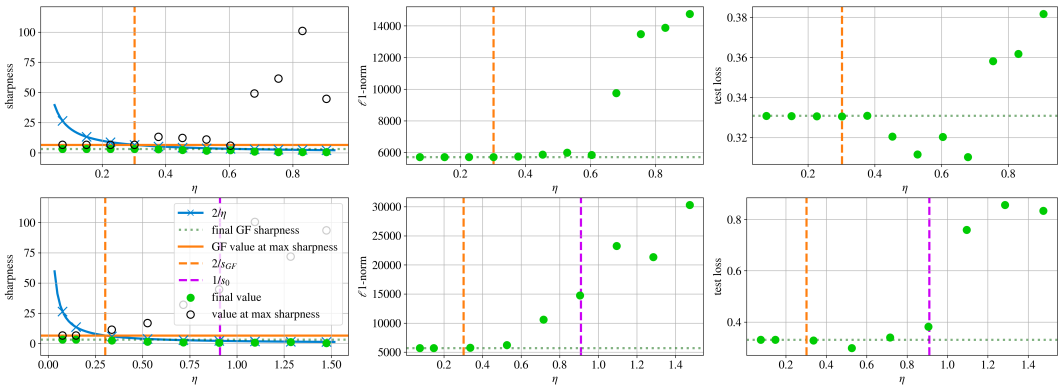
I.2 FCNs WITH TANH ACTIVATION

I.2.1 ON MNIST-5K



(a) Sharpness (b) ℓ1-norm (c) Test loss

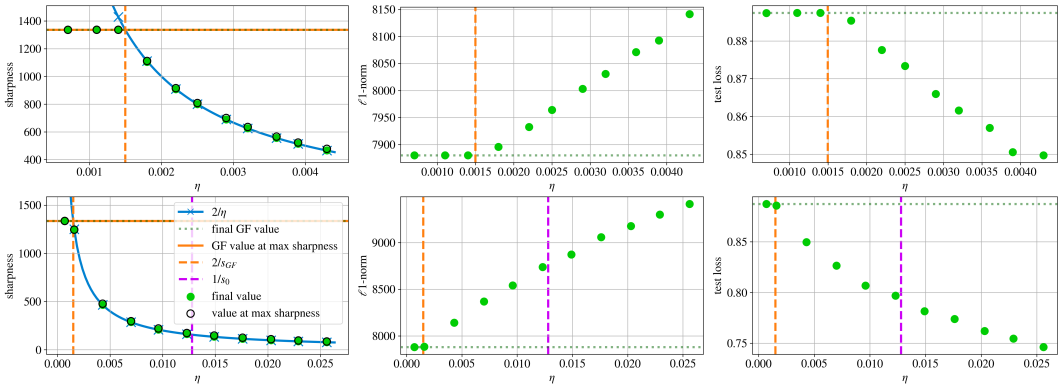
Figure 25: **MSE loss.** FCN-tanh, MNIST-5k, train loss 0.1



(a) Sharpness (b) ℓ1-norm (c) Test loss

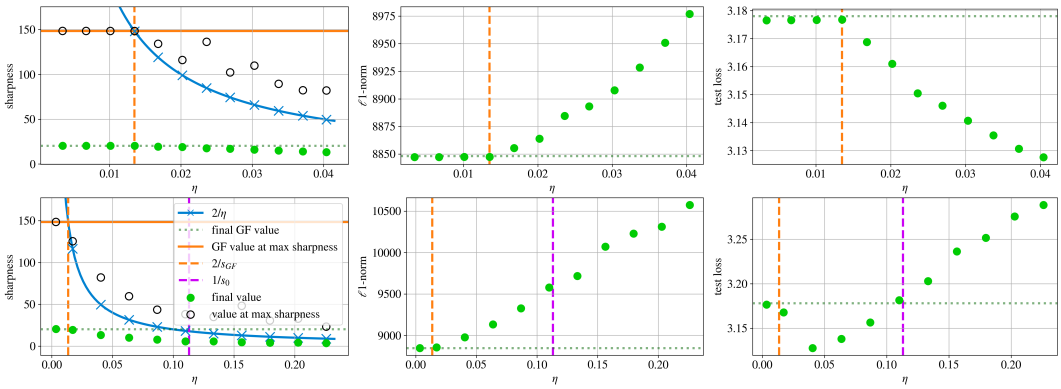
Figure 26: **CE loss.** FCN-tanh, MNIST-5k, train loss 0.01

I.2.2 On CIFAR-10-5k



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 27: **MSE loss.** FCN-tanh, CIFAR-10-5k, train loss 0.001

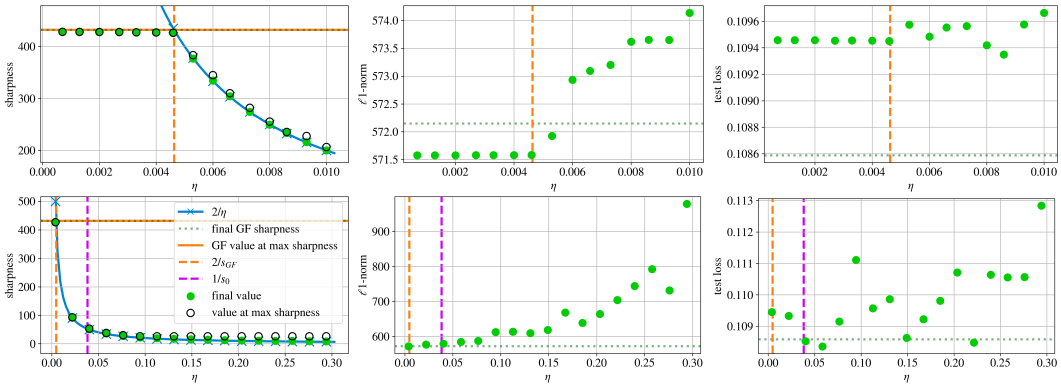


(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 28: **CE loss.** FCN-tanh, CIFAR-10-5k, train loss 0.01

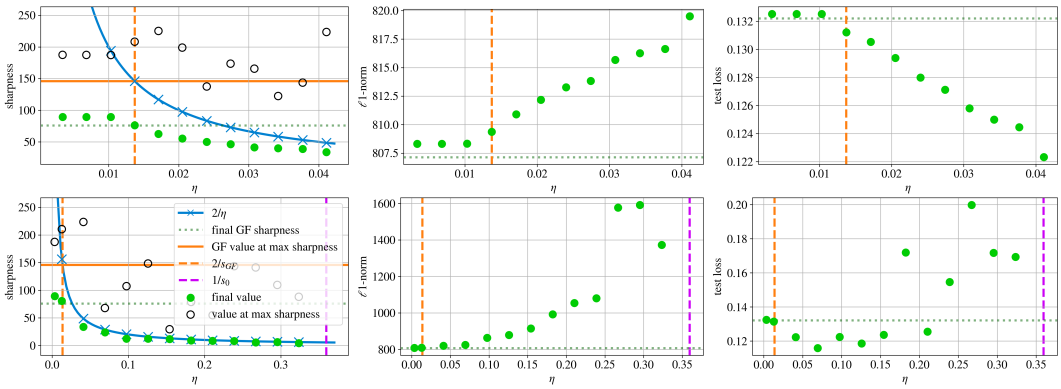
I.3 CNNs with ReLU activation

I.3.1 On MNIST-5k



(a) Sharpness (b) ℓ1-norm (c) Test loss

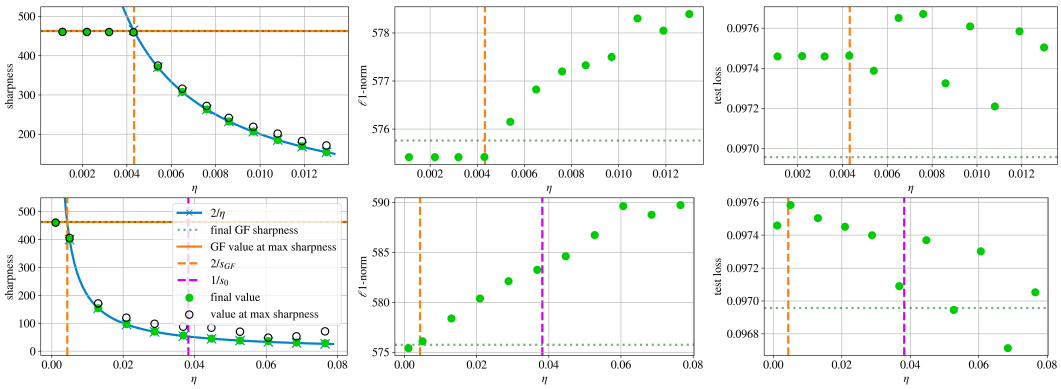
Figure 29: **MSE loss.** CNN-ReLU, MNIST-5k, train loss 0.1



(a) Sharpness (b) ℓ1-norm (c) Test loss

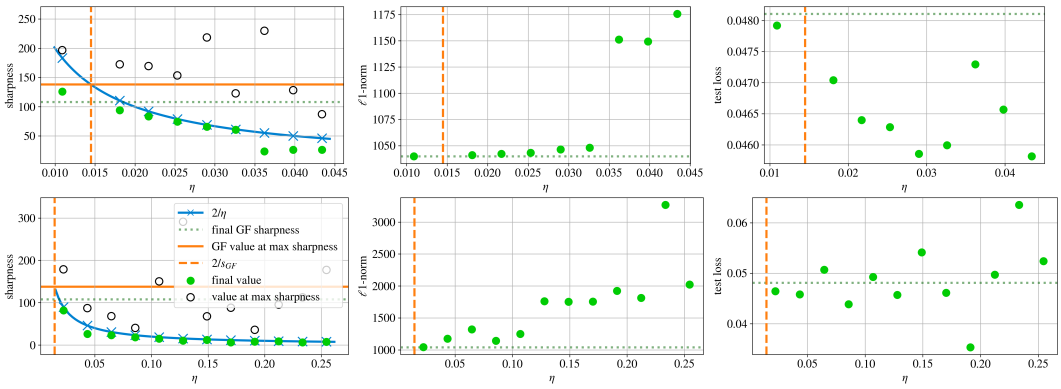
Figure 30: **CE loss.** CNN-ReLU, MNIST-5k, train loss 0.01

I.3.2 On full MNIST



(a) Sharpness (b) ℓ1-norm (c) Test loss

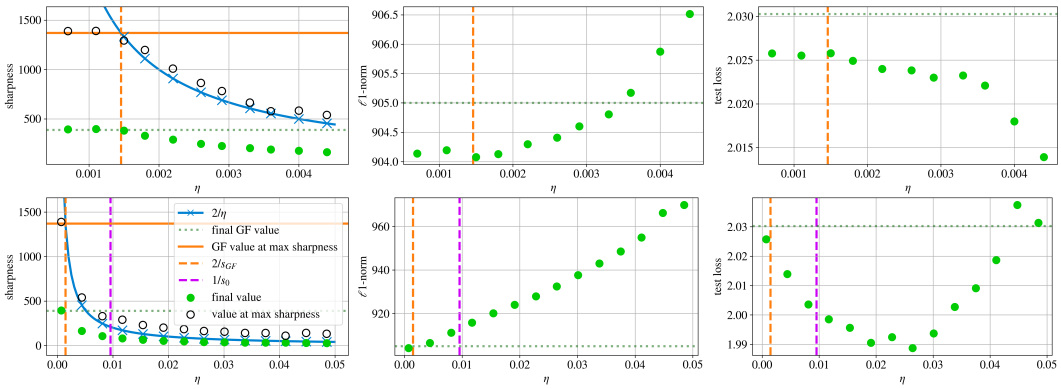
Figure 31: **MSE loss.** CNN-ReLU, MNIST, train loss 0.1



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 32: **CE loss.** CNN-ReLU, MNIST, train loss 0.01

I.3.3 On CIFAR-10-5k

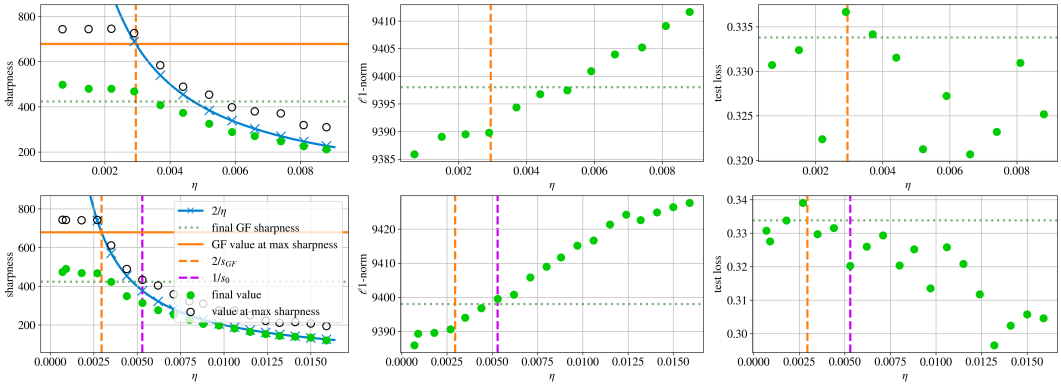


(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 33: **CE loss.** CNN-ReLU with Batch Normalization, CIFAR-10-5k, train loss 0.01

I.4 Vision Transformer

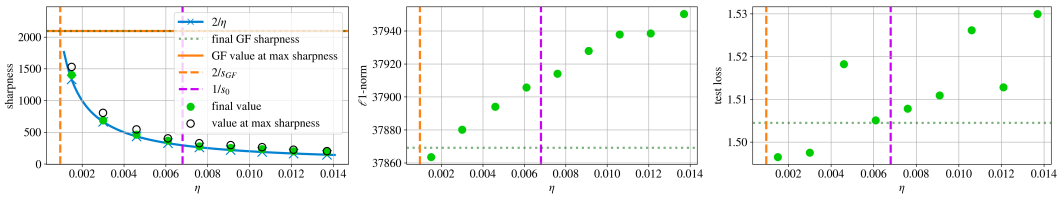
I.4.1 On MNIST-5k



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 34: **CE loss.** ViT, MNIST-5k, train loss 0.1

I.4.2 On CIFAR-10-5k

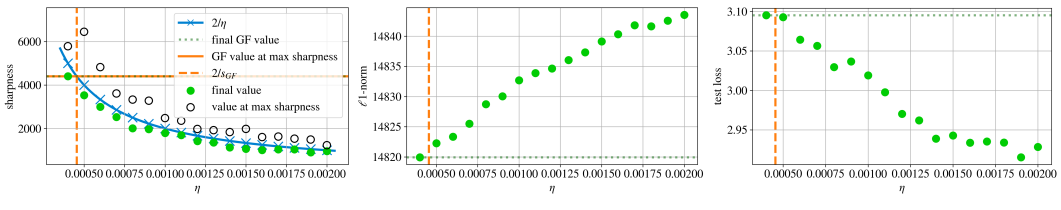


(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 35: **CE loss.** ViT, CIFAR-10-5k, train loss 1

I.5 ResNet20

I.5.1 On CIFAR-10-5k

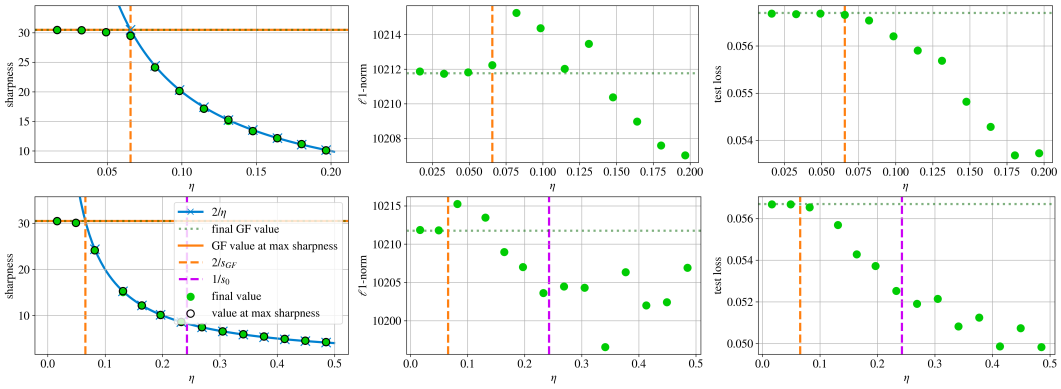


(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 36: **CE loss.** ResNet20, CIFAR-10-5k, train loss 0.1

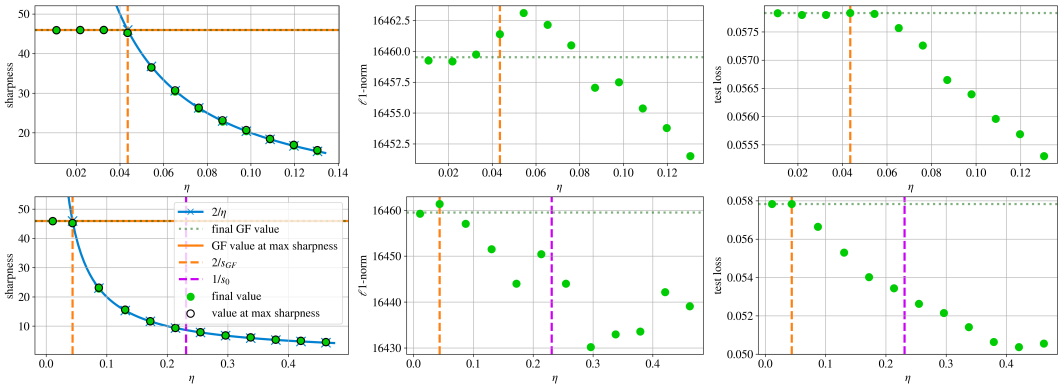
I.6 Varying width and depth

I.6.1 On MNIST-5k



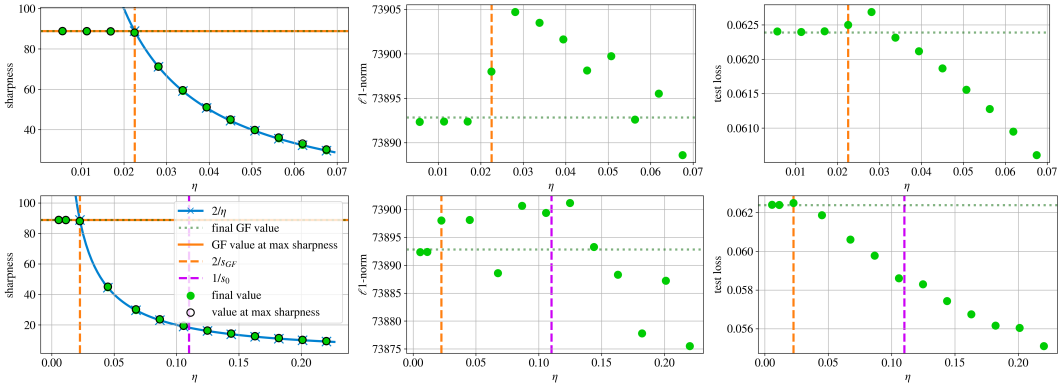
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 37: **FCN-ReLU, 2× width (400×2).** Train loss 0.01, MNIST-5k, MSE loss



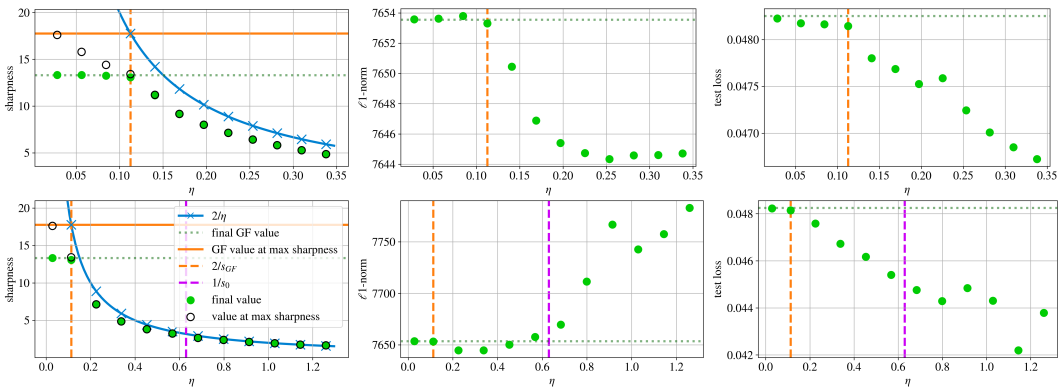
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 38: **FCN-ReLU, 3× width (600×2).** Train loss 0.01, MNIST-5k, MSE loss



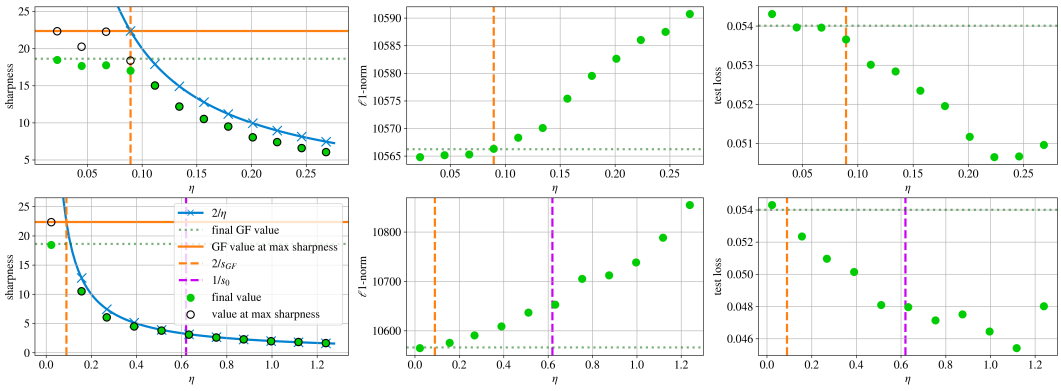
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 39: **FCN-ReLU, 10× width (2000×2).** Train loss 0.01, MNIST-5k, MSE loss



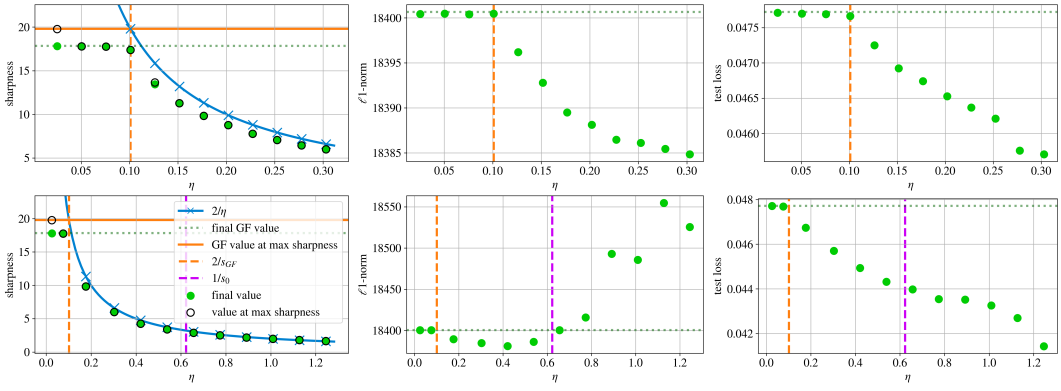
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 40: **FCN-ReLU, 2× depth (200×4).** Train loss 0.01, MNIST-5k, MSE loss



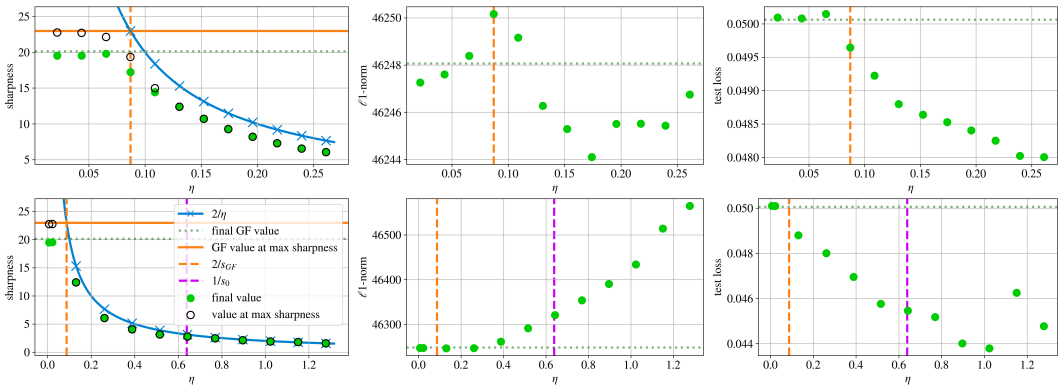
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 41: **FCN-ReLU, 3× depth (200×6).** Train loss 0.01, MNIST-5k, MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

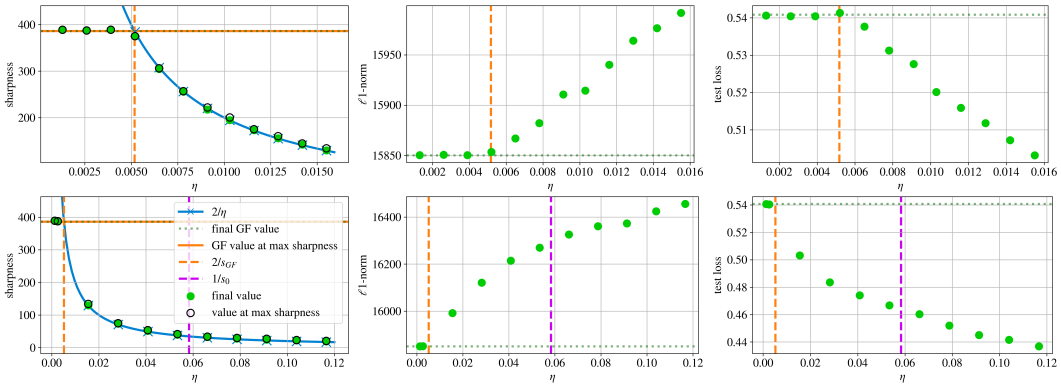
Figure 42: **FCN-ReLU, 2× width and depth (400×4).** Train loss 0.01, MNIST-5k, MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

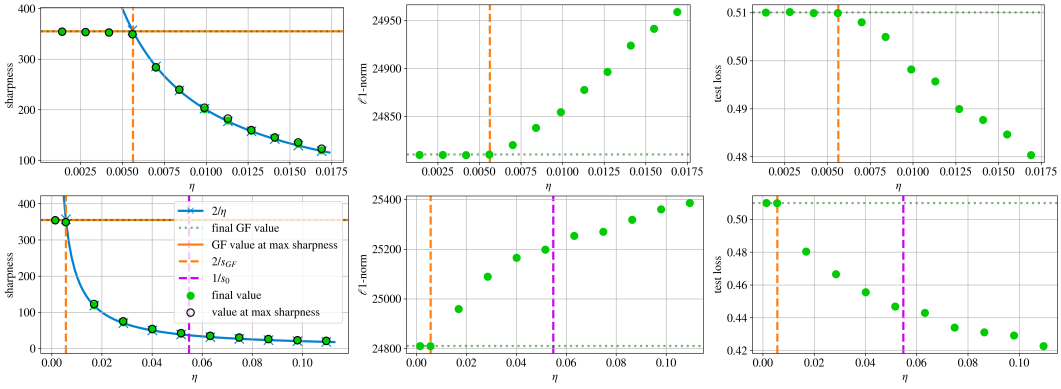
Figure 43: **FCN-ReLU, 3× width and depth (600 × 6).** Train loss 0.01, MNIST-5k, MSE loss

I.6.2 On CIFAR-10-5k



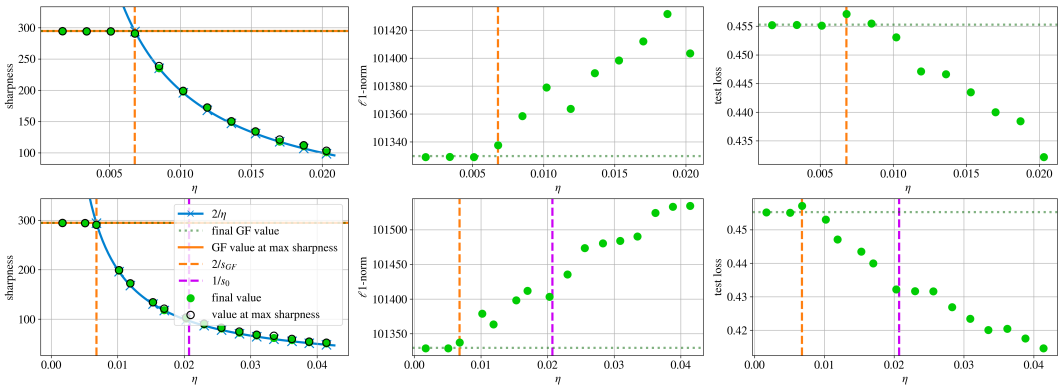
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 44: **FCN-ReLU, 2× width (400 × 2).** Train loss 0.01, CIFAR-10-5k, MSE loss



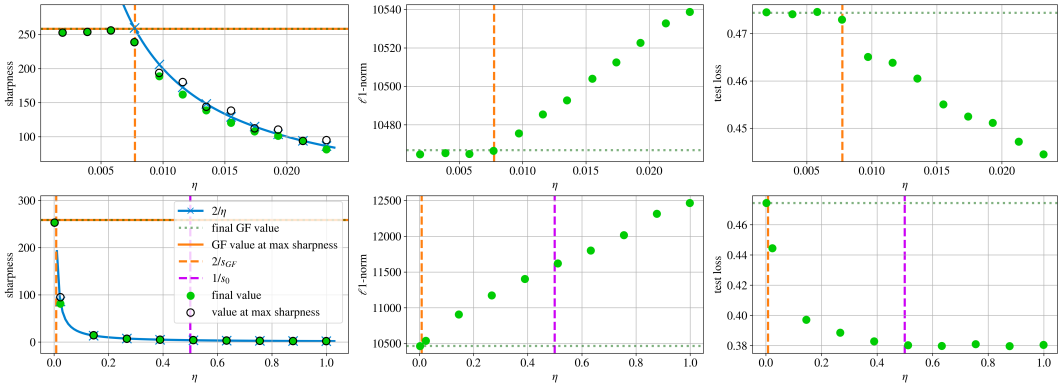
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 45: **FCN-ReLU, 3× width (600 × 2).** Train loss 0.01, CIFAR-10-5k, MSE loss



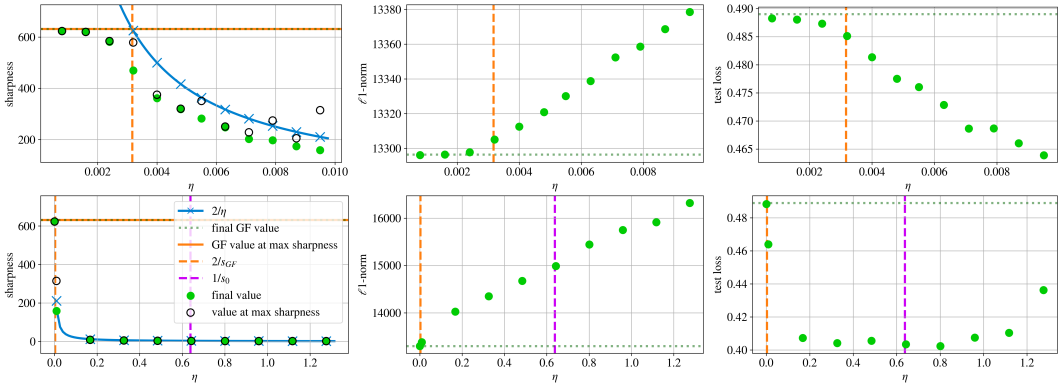
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 46: **FCN-ReLU, 10× width (2000 × 2).** Train loss 0.01, CIFAR-10-5k, MSE loss



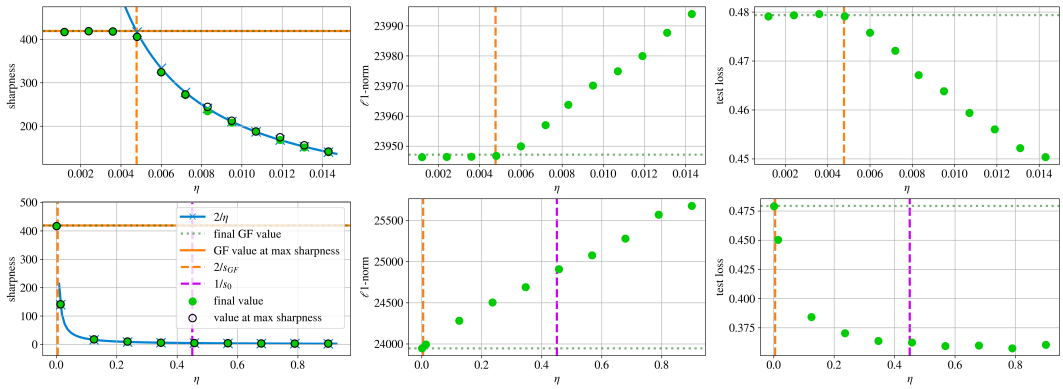
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 47: **FCN-ReLU, 2× depth (200 × 4).** Train loss 0.01, CIFAR-10-5k, MSE loss



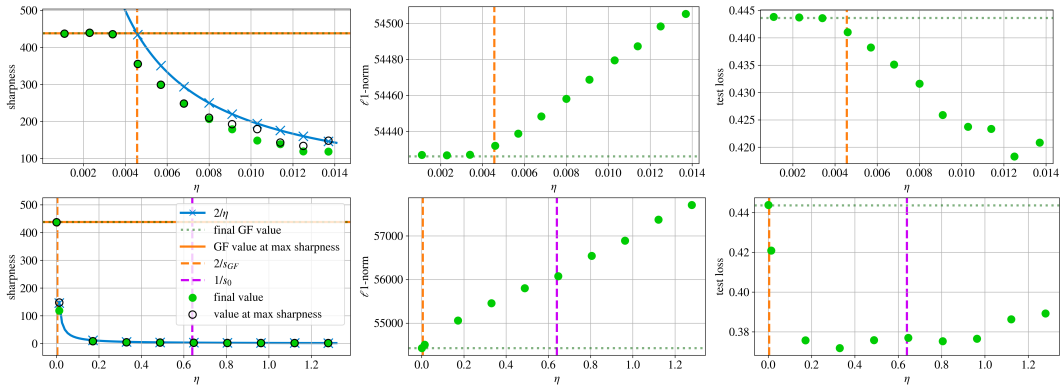
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 48: **FCN-ReLU, 3× depth (200 × 6).** Train loss 0.01, CIFAR-10-5k, MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 49: **FCN-ReLU, 2× width and depth (400×4).** Train loss 0.01, CIFAR-10-5k, MSE loss



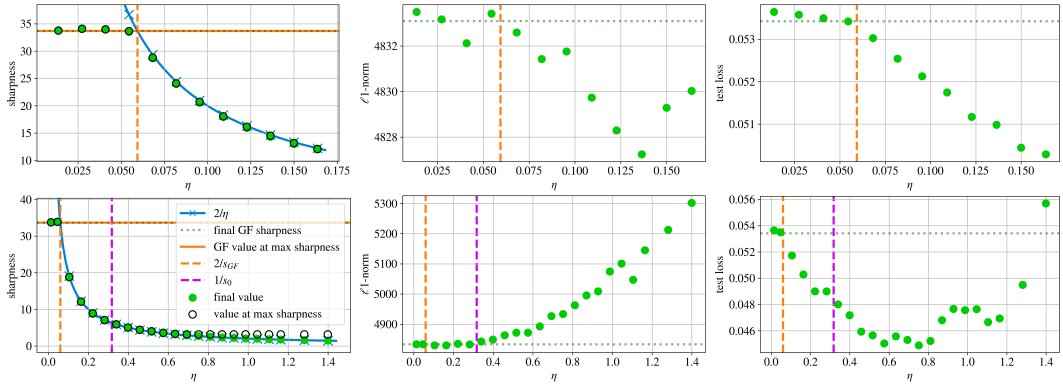
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 50: **FCN-ReLU, 3× width and depth (600×6).** Train loss 0.01, CIFAR-10-5k, MSE loss

I.7 Further configurations

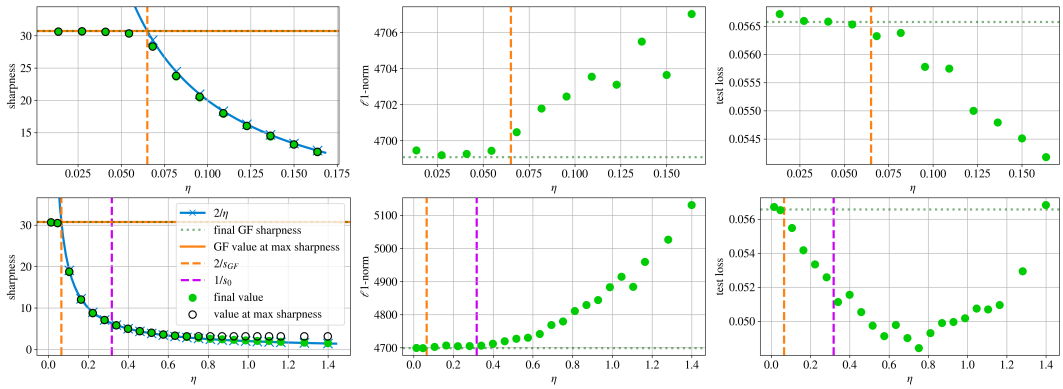
I.7.1 Different loss goals

FCN-ReLU on MNIST-5k with the MSE loss



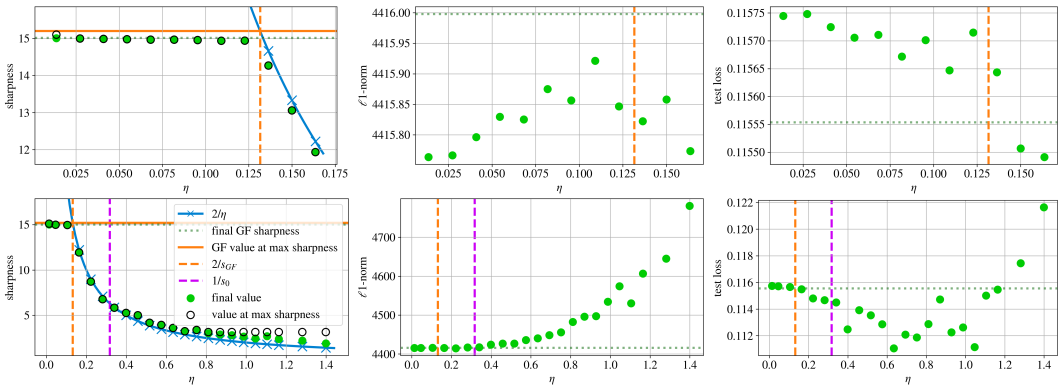
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 51: **Train loss** 0.001**.** FCN-ReLU, MNIST-5k, MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

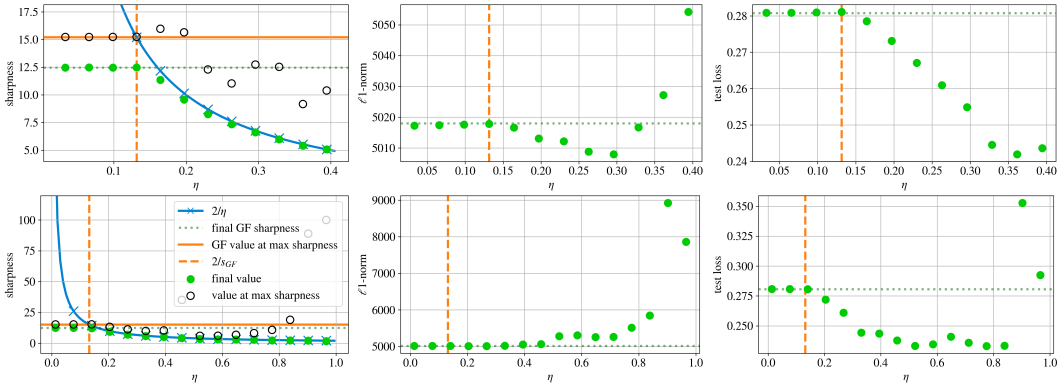
Figure 52: **Train loss** 0.01**.** FCN-ReLU, MNIST-5k, MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 53: **Train loss** 0.1**.** FCN-ReLU, MNIST-5k, MSE loss

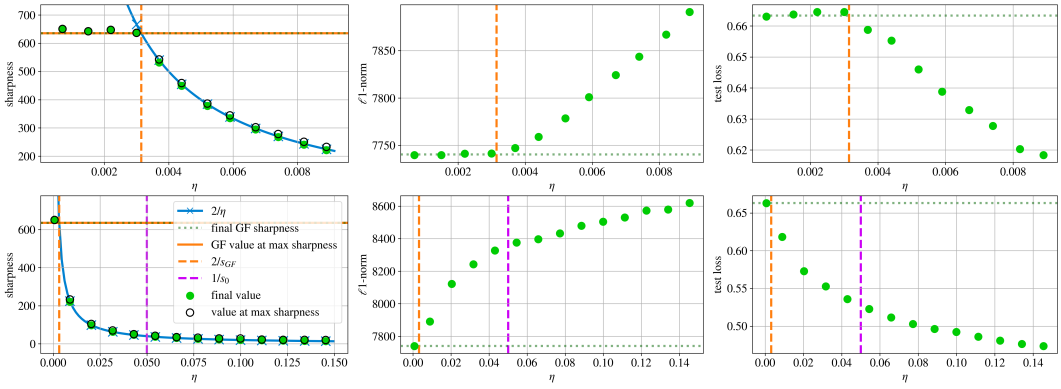
FCN-ReLU on MNIST-5k with the CE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

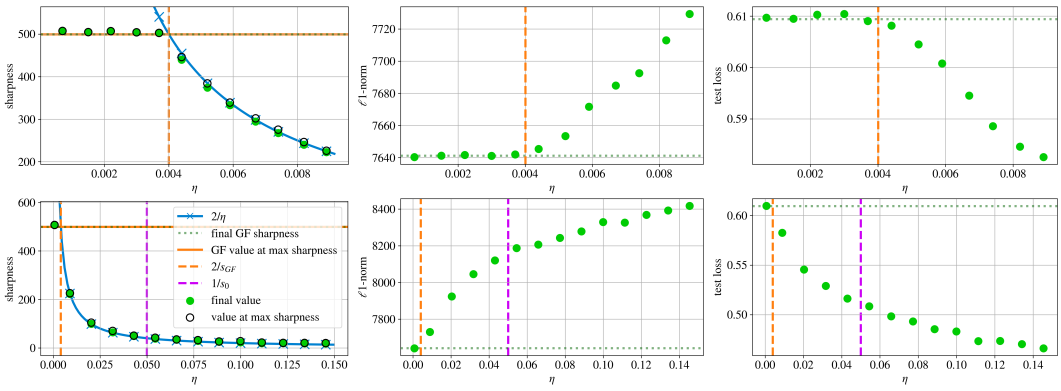
Figure 54: **Train loss** 0.1**.** FCN-ReLU, MNIST-5k, CE loss

FCN-ReLU on CIFAR-10-5k with the MSE loss



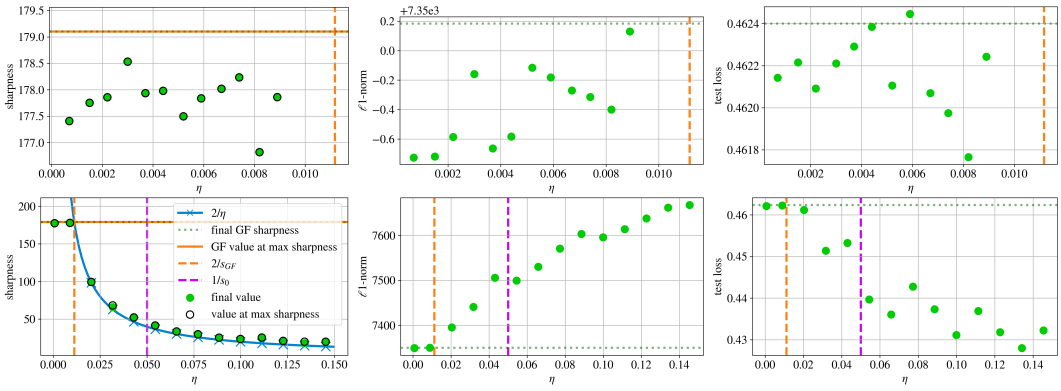
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 55: **Train loss** 0.001**.** FCN-ReLU, CIFAR-10-5k, MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

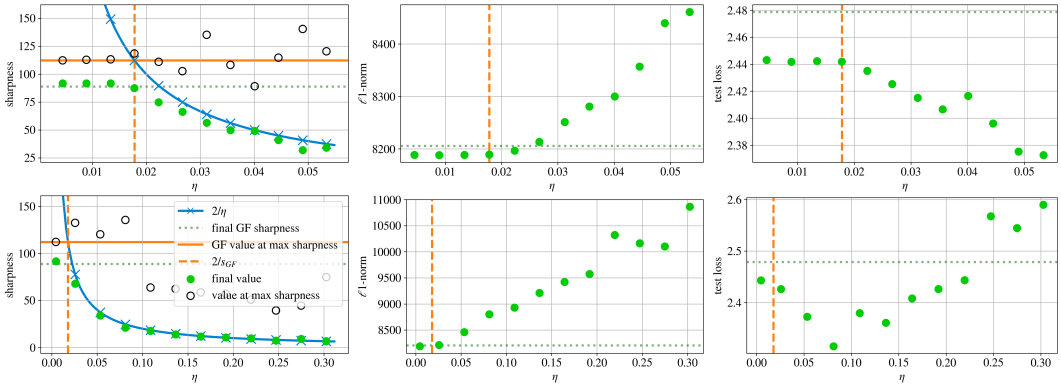
Figure 56: **Train loss** 0.01**.** FCN-ReLU, CIFAR-10-5k, MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 57: **Train loss** 0.1**.** FCN-ReLU, CIFAR-10-5k, MSE loss

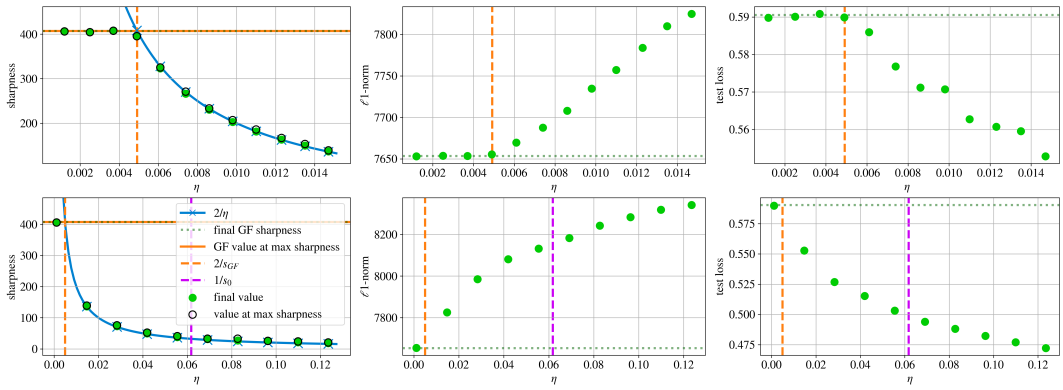
FCN-ReLU on CIFAR-10-5k with the CE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

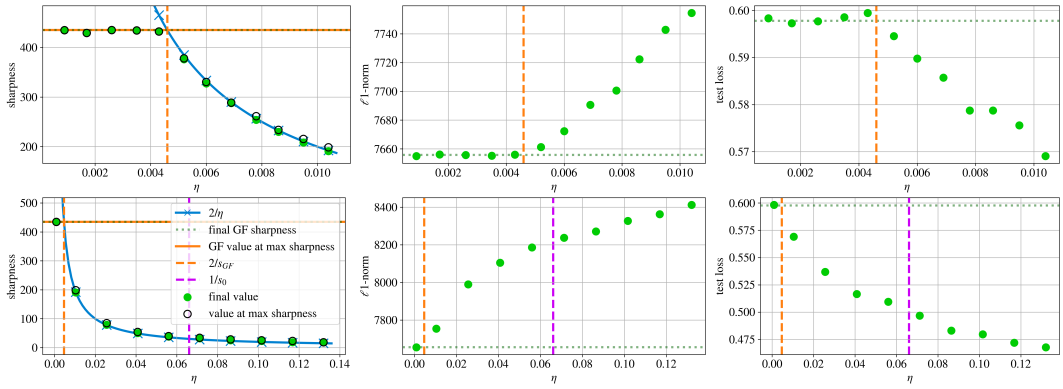
Figure 58: **Train loss** 0.1**.** FCN-ReLU, CIFAR-10-5k, CE loss

I.7.2 Other initialization seeds for FCN-ReLU on CIFAR-10-5k with the MSE loss



(a) Sharpness (b) ℓ1-norm (c) Test loss

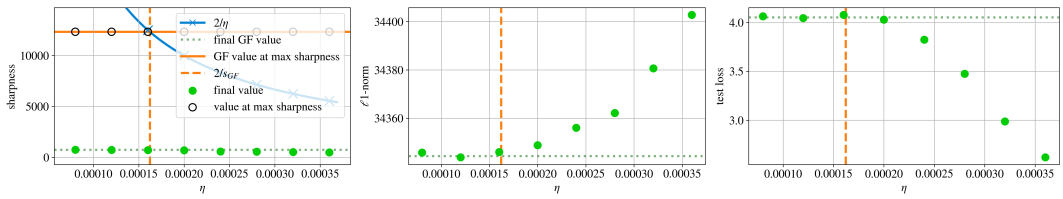
Figure 59: **Seed 44.** FCN-ReLU, CIFAR-10-5k, MSE loss, train loss 0.01



(a) Sharpness (b) ℓ1-norm (c) Test loss

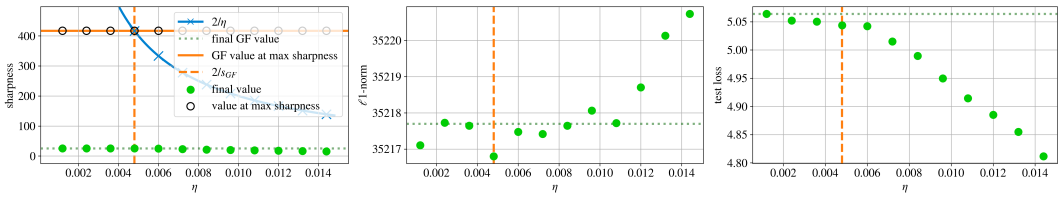
Figure 60: **Seed 45.** FCN-ReLU, CIFAR-10-5k, MSE loss, train loss 0.01

I.7.3 Scaled initialization for FCN-ReLU on CIFAR-10-5k with the MSE loss



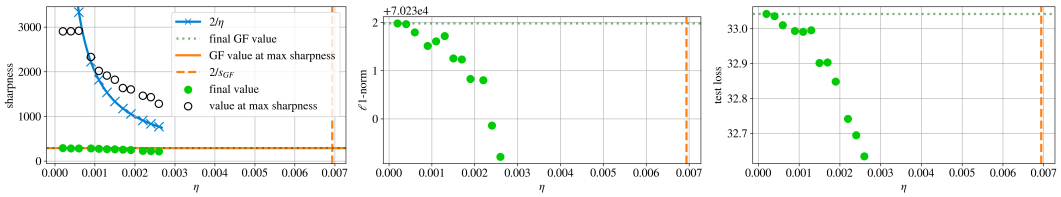
(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 61: **Initialization from seed 43 scaled $\times 5$.** FCN-ReLU, CIFAR-10-5k, MSE loss, train loss 0.1



(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 62: **Initialization from seed 43 scaled $\times 5$.** FCN-ReLU, CIFAR-10-5k, CE loss, train loss 0.01

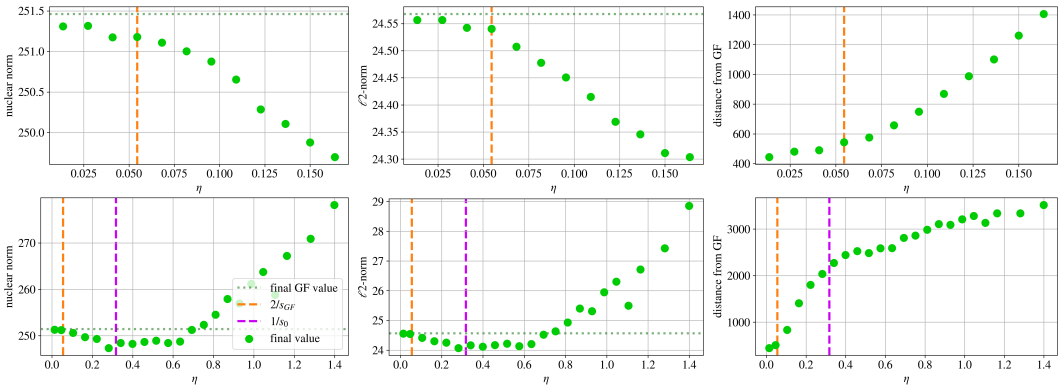


(a) Sharpness (b) ℓ1-norm (c) Test loss

Figure 63: **Initialization from seed 43 scaled $\times 10$.** FCN-ReLU, CIFAR-10-5k, CE loss, train loss 0.01

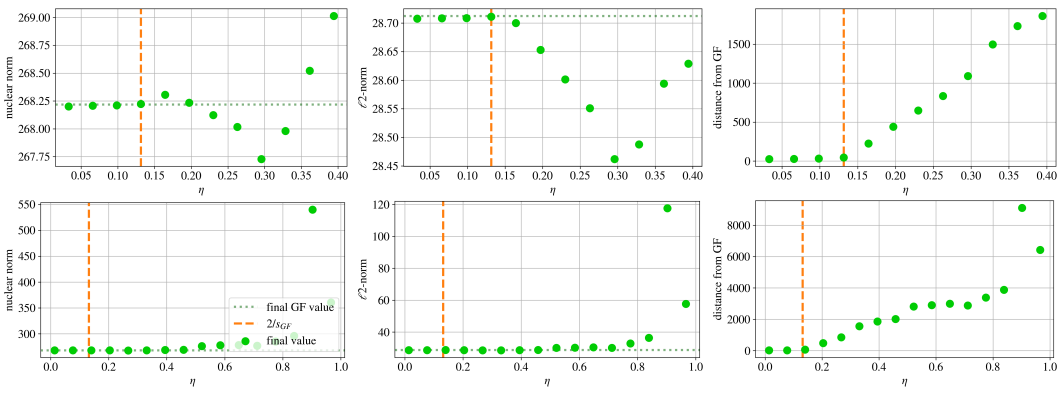
I.8 Further properties

I.8.1 Alternative norms and distance from GF solution



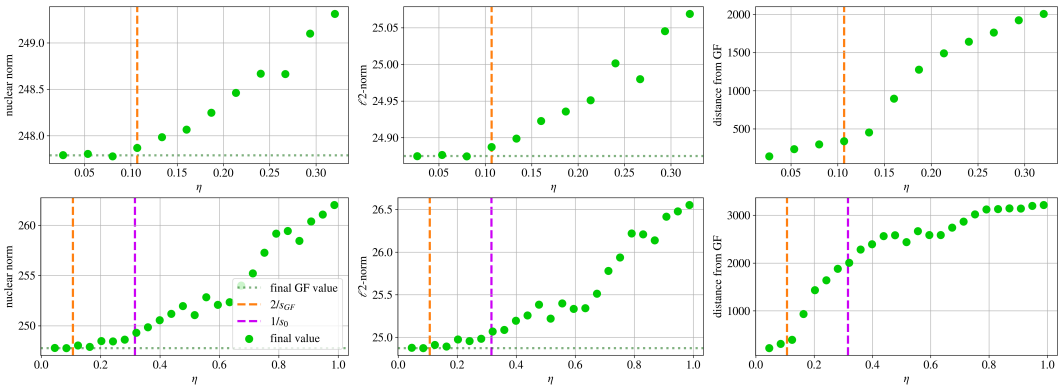
(a) Nuclear norm (b) ℓ2-norm (c) ℓ1-distance from GF solution

Figure 64: **FCN-ReLU on MNIST-5k with the MSE loss.** Train loss 0.0001



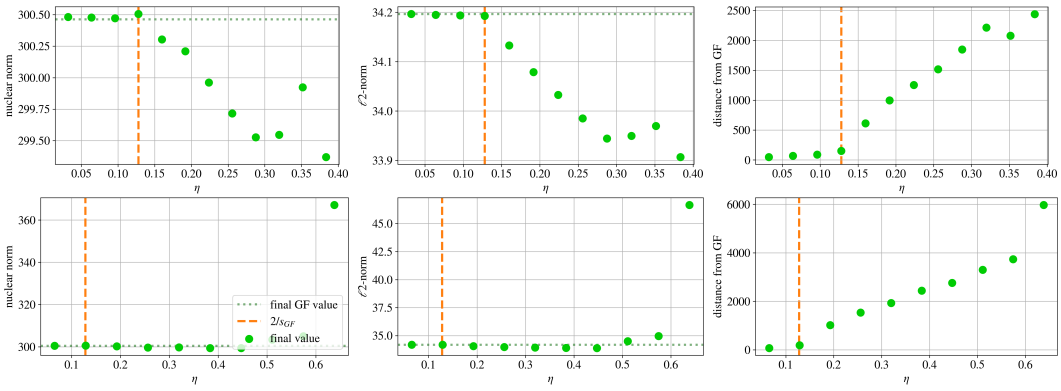
(a) Nuclear norm (b) $\ell 2$-norm (c) $\ell 1$-distance from GF solution

Figure 65: **FCN-ReLU on MNIST-5k with the CE loss.** Train loss 0.01



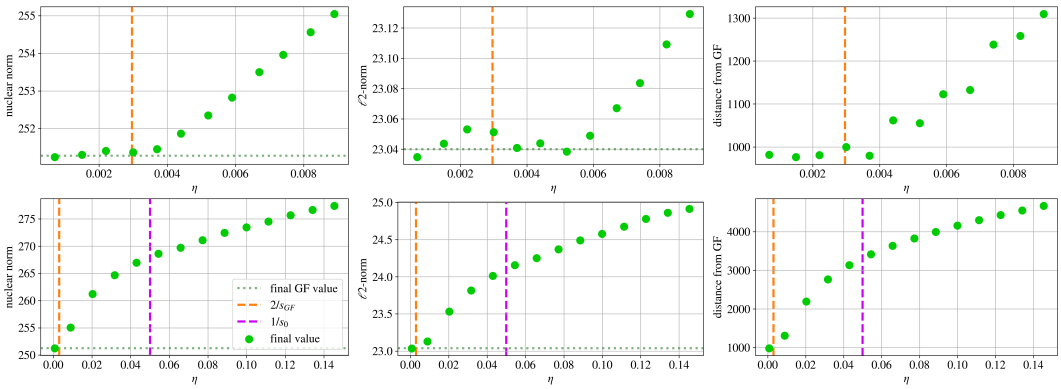
(a) Nuclear norm (b) $\ell 2$-norm (c) $\ell 1$-distance from GF solution

Figure 66: **FCN-ReLU on full MNIST with the MSE loss.** Train loss 0.01



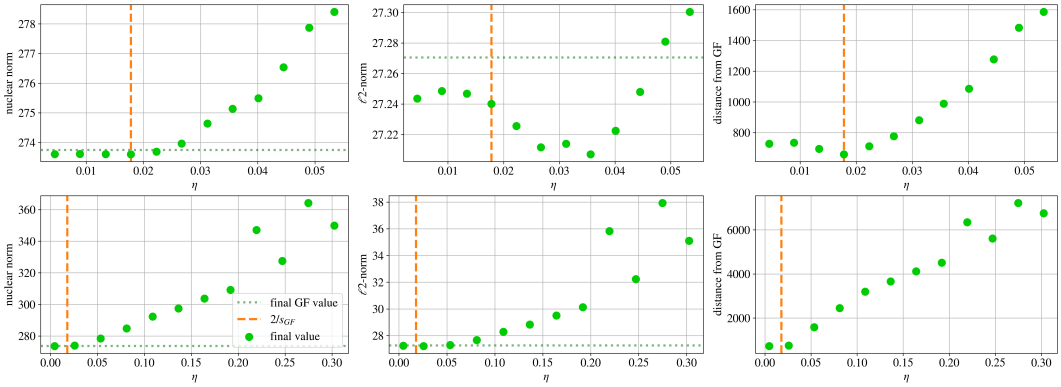
(a) Nuclear norm (b) $\ell 2$-norm (c) $\ell 1$-distance from GF solution

Figure 67: **FCN-ReLU on full MNIST with the CE loss.** Train loss 0.01



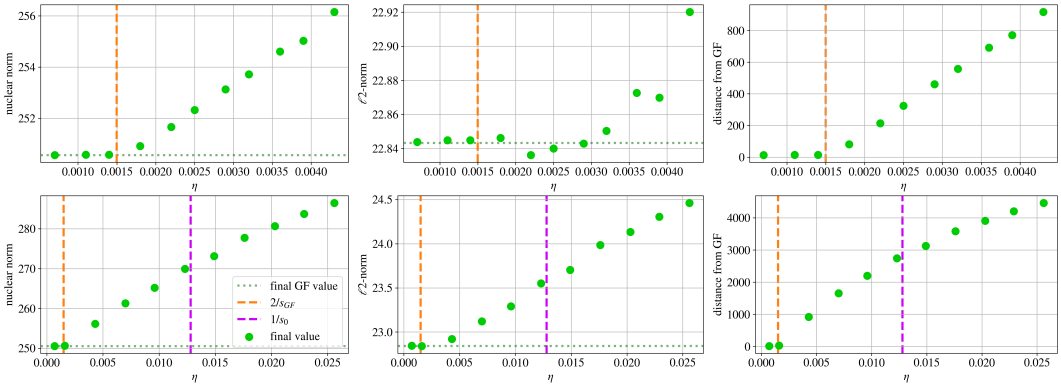
(a) Nuclear norm (b) ℓ2-norm (c) ℓ1-distance from GF solution

Figure 68: **FCN-ReLU on CIFAR-10-5k with the MSE loss.** Train loss 0.0001



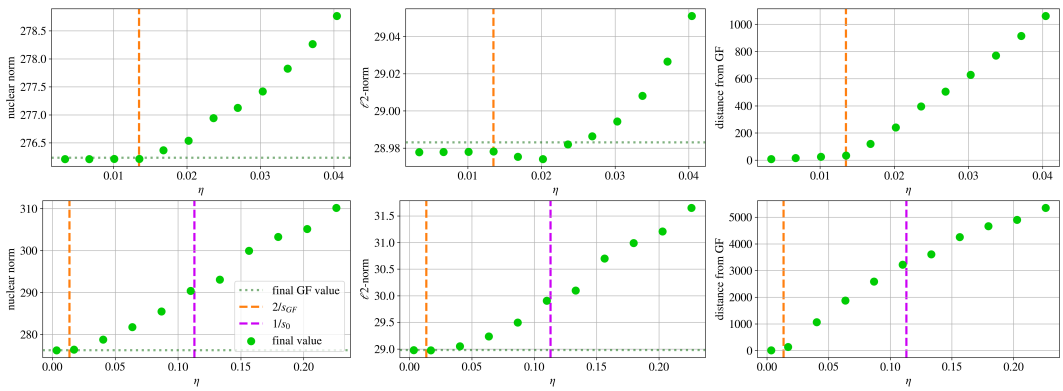
(a) Nuclear norm (b) ℓ2-norm (c) ℓ1-distance from GF solution

Figure 69: **FCN-ReLU on CIFAR-10-5k with the CE loss.** Train loss 0.01



(a) Nuclear norm (b) ℓ2-norm (c) ℓ1-distance from GF solution

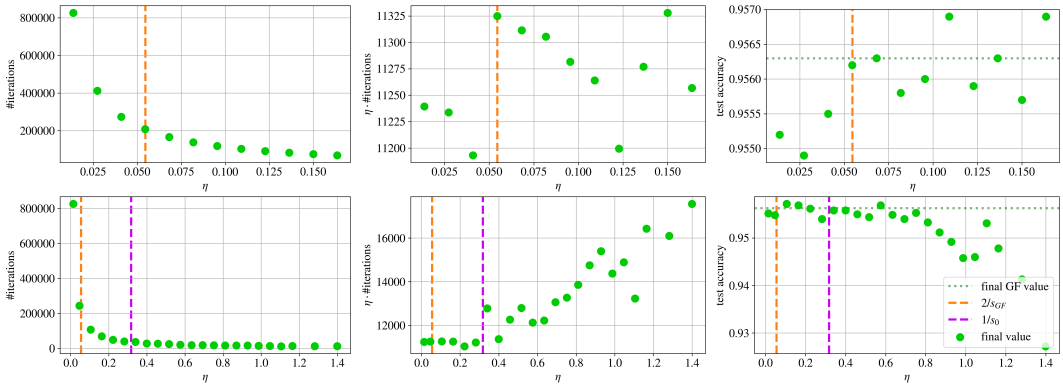
Figure 70: **FCN-tanh on CIFAR-10-5k with the MSE loss.** Train loss 0.001



(a) Nuclear norm (b) $\ell 2$-norm (c) $\ell 1$-distance from GF solution

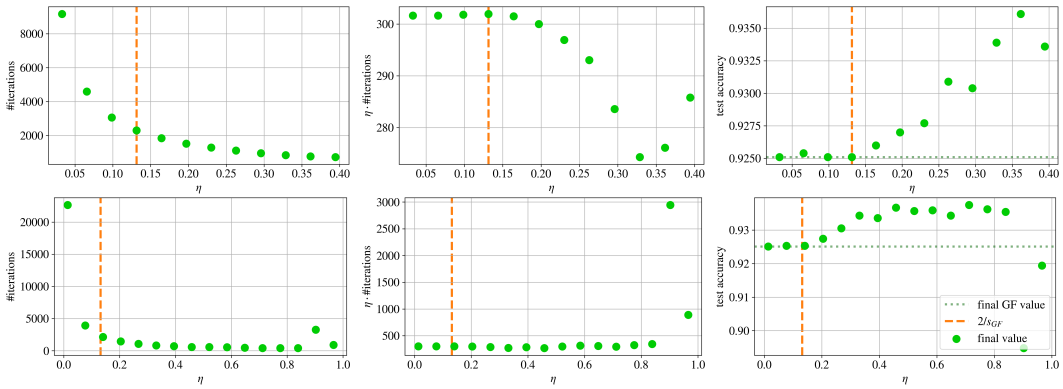
Figure 71: **FCN-tanh on CIFAR-10-5k with the CE loss.** Train loss 0.01

I.8.2 Convergence speed and test accuracy



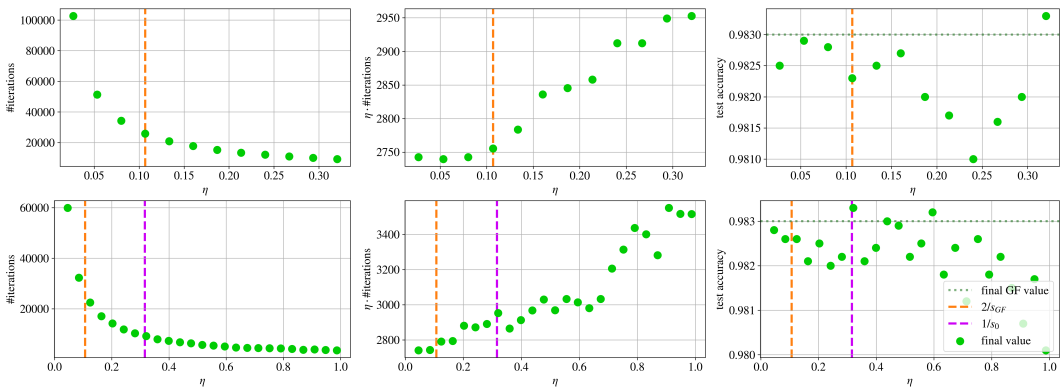
(a) Iterations till loss 0.0001 (b) Curve length (c) Test accuracy

Figure 72: **FCN-ReLU on MNIST-5k with the MSE loss.** Train loss 0.0001



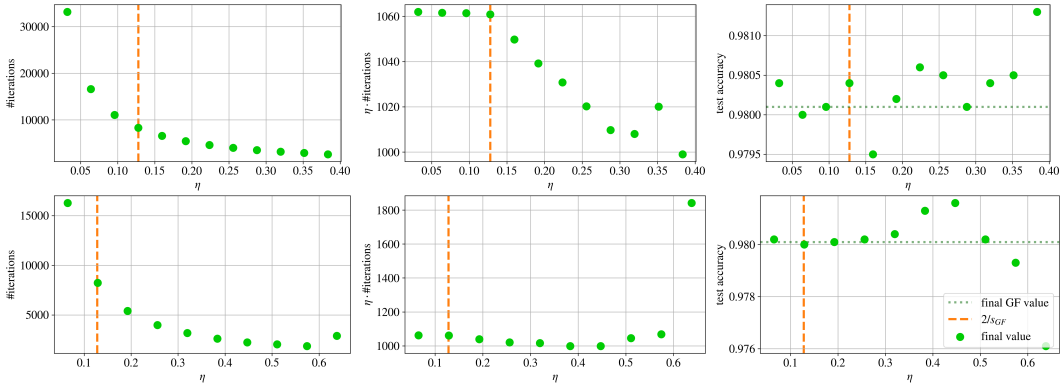
(a) Iterations till loss 0.01 (b) Curve length (c) Test accuracy

Figure 73: **FCN-ReLU on MNIST-5k with the CE loss.** Train loss 0.01



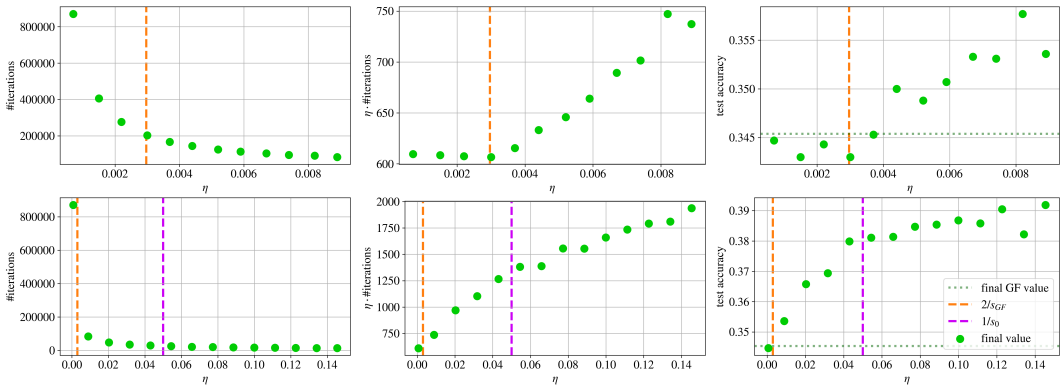
(a) Iterations till loss 0.01 (b) Curve length (c) Test accuracy

Figure 74: **FCN-ReLU on full MNIST with the MSE loss.** Train loss 0.01



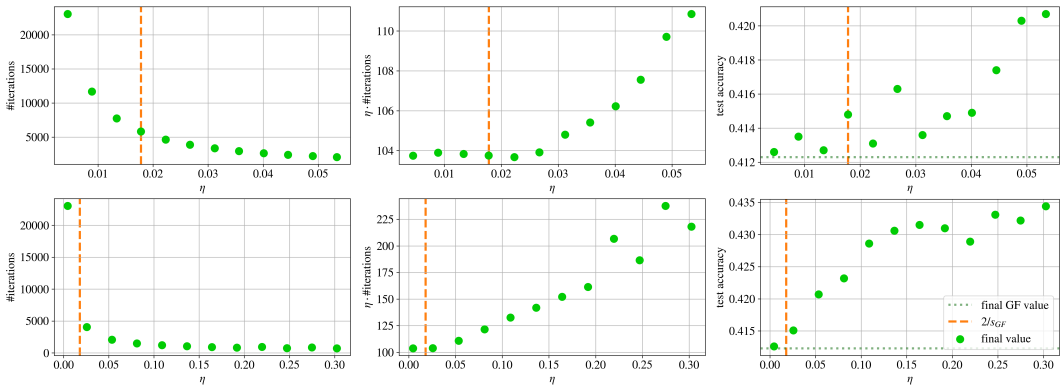
(a) Iterations till loss 0.01 (b) Curve length (c) Test accuracy

Figure 75: **FCN-ReLU on full MNIST with the CE loss.** Train loss 0.01



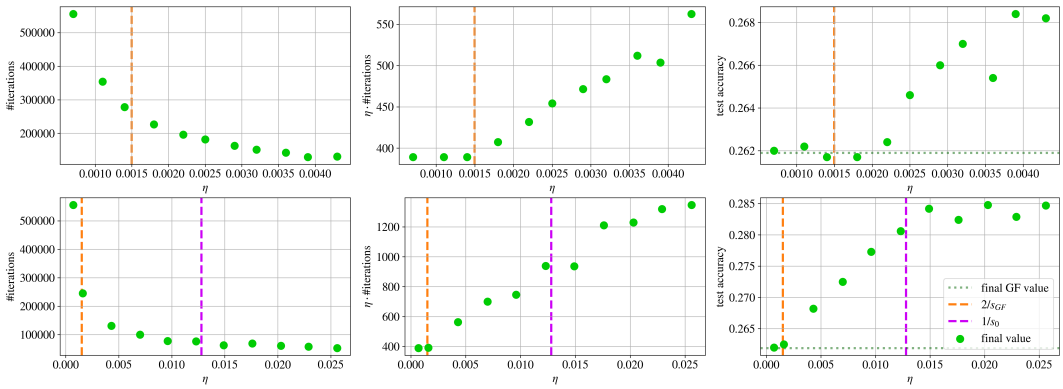
(a) Iterations till loss 0.0001 (b) Curve length (c) Test accuracy

Figure 76: **FCN-ReLU on CIFAR-10-5k with the MSE loss.** Train loss 0.0001



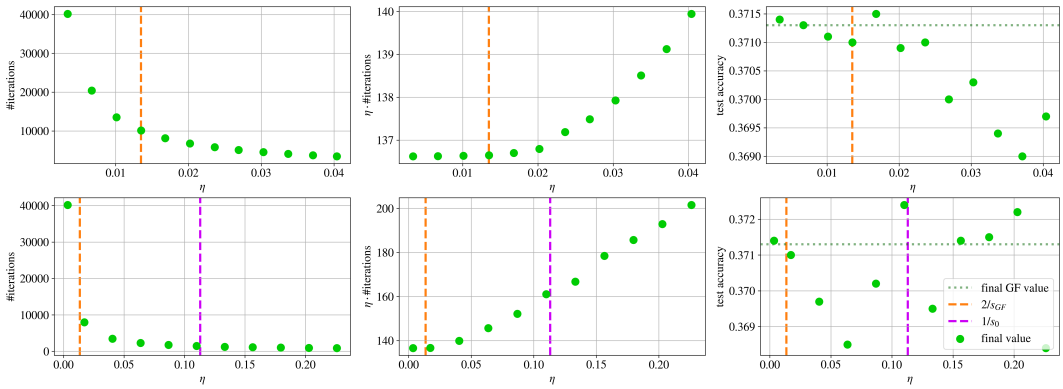
(a) Iterations till loss 0.01 (b) Curve length (c) Test accuracy

Figure 77: **FCN-ReLU on CIFAR-10-5k with the CE loss.** Train loss 0.01



(a) Iterations till loss 0.001 (b) Curve length (c) Test accuracy

Figure 78: **FCN-tanh on CIFAR-10-5k with the MSE loss.** Train loss 0.001



(a) Iterations till loss 0.01 (b) Curve length (c) Test accuracy

Figure 79: **FCN-tanh on CIFAR-10-5k with the CE loss.** Train loss 0.01