DOES SGD SEEK FLATNESS OR SHARPNESS? AN EXACTLY SOLVABLE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

A large body of theory and empirical work hypothesizes a connection between the flatness of a neural network's loss landscape during training and its performance. However, there have been conceptually opposite pieces of evidence regarding when SGD prefers flatter or sharper solutions. In this work, we partially but causally clarify the flatness-seeking behavior of SGD by identifying and exactly solving an analytically solvable model that exhibits both flattening and sharpening behavior. In this model, the SGD training has no *a priori* preference for flatness, but only a preference for minimal gradient fluctuations. This leads to the insight that, at least within this model, it is data distribution that uniquely determines the sharpness at convergence, and that a flat minimum is preferred if and only if the noise in the labels is isotropic across all output dimensions. When the noise in the labels is anisotropic, the model instead prefers sharpness and can converge to an arbitrarily sharp solution, depending on the imbalance in the noise in the labels spectrum. We reproduce this key insight in controlled settings with different model architectures such as MLP, RNN, and transformers.

1 INTRODUCTION

The flatness and sharpness of the loss landscape have long been an essential part of our understanding of optimization and generalization in neural networks (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016). A common view is that stochastic gradient descent (SGD) training implicitly favors flatter minima: even when many parameter configurations achieve essentially the same training loss, SGD training is more likely to go to a flatter region. Meanwhile, flatter solutions are often argued to be more robust to perturbations and thus generalize better. This intuition has motivated a substantial body of work connecting the stochasticity and finite step size of SGD to an implicit preference for flat regions of the loss landscape, typically quantified through Hessian-based notions of sharpness (Keskar et al., 2016; Wu et al., 2018; Xie et al., 2020; Wu & Su, 2023). It has also inspired optimization methods that explicitly seek flatter solutions, such as sharpness-aware minimization (Foret et al., 2020).

Yet, the statement that "SGD seeks flatness" is stronger than what is empirically observed. In an overparameterized model, the set of global minima can form a large degenerate manifold containing solutions with different Hessian spectra and sharpness levels. The simplest question is therefore: *when many equally good minima coexist, does SGD converge to the flattest one?* This distinction is important because the evolution of sharpness during training can depend strongly on initialization and on the geometry of the optimization trajectory. For example, sharpness is known to increase during training in a variety of settings, including the progressive-sharpening and edge-of-stability regimes (Cohen et al., 2021; Andreyev & Beneventano, 2024; Cai et al., 2026). These phenomena concern the training dynamics and dynamical stability of optimization. In contrast, the question we study here concerns the *selection at convergence*.

Our starting point is the empirical observation that, in the small-but-nonvanishing learning rate regime, SGD can be described by an effective objective containing a correction that penalizes gradient fluctuations (Roberts, 2021; Smith et al., 2021; Beneventano, 2023). To leading order, this induces a preference among otherwise equivalent minima for solutions with smaller gradient fluctuation Ziyin et al. (2025). The key observation is that *gradient fluctuation and sharpness are different quantities*. Consequently, minimizing fluctuation is equivalent to minimizing sharpness only under special conditions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We make this mechanism exact in a deep linear network. The model admits a continuous manifold of global minima containing solutions with arbitrarily large sharpness. Nevertheless, in the minimum-fluctuation limit, the selected sharpness can be solved in closed form. In particular, with a full-row-rank teacher, isotropic label noise makes fluctuation minimization coincide with sharpness minimization. When the label noise is anisotropic, the two objectives become misaligned: the minimum-fluctuation solution is generally sharper than the minimum-sharpness solution, and its sharpness grows as the square root of the condition number of the label-noise covariance.

Our main contributions are:

1. We identify a general distinction between gradient fluctuation and Hessian sharpness and show why minimizing fluctuation need not imply minimizing sharpness (Section 3).
2. We prove the exact minimum-fluctuation solutions of a deep linear network with arbitrary widths and depths, and derive their sharpness in closed form. Although there are globally optimal solutions with arbitrarily sharpness, all minimum-fluctuation solutions have the same Hessian spectrum (Section 4).
3. We characterize when SGD's fluctuation bias coincides with a flatness bias. Under a full-row-rank teacher, the minimum-fluctuation solution attains the minimum possible sharpness if and only if the label-noise covariance is isotropic. For anisotropic label noise, the selected sharpness grows as the square root of the condition number of the label-noise covariance (Section 5).
4. We test these predictions beyond the solvable setting. Across nonlinear architectures including MLPs, RNNs, and Transformers, and across regression and classification tasks, increasing the anisotropy of the label noise consistently leads to sharper solutions (Section 5).

2 SHARP OR FLAT?

The statement that "SGD seeks flatness" is fundamentally a statement about *implicit selection*: when many parameter configurations achieve the same optimal loss, does SGD preferentially select those with smaller sharpness? This question should be distinguished from whether the sharpness increases or decreases along a particular optimization trajectory. The latter can be strongly affected by initialization, deterministic landscape geometry, and finite-step-size dynamics, whereas the former concerns the implicit bias of SGD. Our goal in this work is to isolate this selection problem.

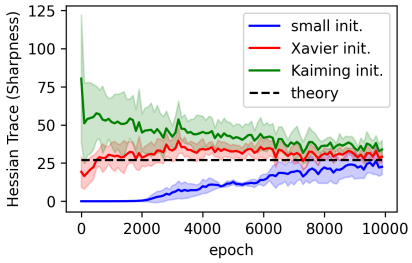


Figure 1: Deep linear networks trained with SGD converges to the same sharpness, even if the initial sharpness depends strongly on the initialization scale. The theoretical line is the prediction of Theorem 4.1.

Arbitrariness of the training trajectory. A first difficulty in inferring an implicit preference from the evolution of sharpness is the arbitrariness of initialization. Suppose, for example, that the learning dynamics eventually select solutions with a characteristic nonzero sharpness. Starting from a flatter initialization will then necessarily produce an apparent sharpening along the trajectory, whereas starting from a sharper initialization will produce flattening toward the same final sharpness. The sign of the change in sharpness therefore does not by itself reveal whether the learning algorithm intrinsically prefers flat or sharp solutions. A second difficulty is that different initializations or sampling realizations may converge to different minima with different sharpness, making comparisons ambiguous. Figure 1 illustrates the first issue in a deep matrix-factorization problem: changing the initialization scale strongly alters the initial sharpness, while SGD nevertheless converges to essentially the same final value.

Intrinsic and selected sharpness. For a fixed task and architecture, let $T_{\min} := \min_{\theta \in \mathcal{M}} T(\theta)$, where $\mathcal{M}$ denotes the global-minimum manifold and T is a chosen measure of sharpness. For a solution selected by the learning algorithm, we can write

$$T = T_{\min} + \Delta_{\text{select}}, \tag{1}$$

where $T_{\min}$ is the smallest curvature compatible with solving the learning problem within the given parameterization, while $\Delta_{\text{select}} \geq 0$ measures the excess sharpness induced by the algorithm's selection among equivalent minima. The question studied here is what determines Δ_{select}, and in particular whether SGD necessarily drives it to zero.

Small-learning-rate selection The regime considered in this work is the small-learning-rate $\eta \to 0$ limit. More precisely, we study the leading correction to gradient flow induced by finite-step-size

stochastic gradient descent. In this limit, the correction to the objective is small in magnitude, but it can still resolve a degeneracy among global minima: once the leading population loss is identical along $\mathcal{M}$, an $O(\eta)$ term can determine which points on $\mathcal{M}$ are selected. This is the sense in which the limit $\eta \to 0^+$ remains nontrivial for implicit bias.

This regime is conceptually distinct from progressive-sharpening and edge-of-stability phenomena associated with discrete optimization at finite or large learning rates (Xing et al., 2018; Jastrzębski et al., 2019; 2020; Cohen et al., 2021; 2022; Andreyev & Beneventano, 2024; Cai et al., 2026). Here we instead consider learning rates sufficiently small that such a stability boundary is not the mechanism determining the selected curvature. The object of interest is the stochastic selection among already degenerate minimizers in the small-η regime.

Gradient fluctuation as the selection principle. Recent work suggests that SGD admits an effective objective whose leading stochastic correction penalizes gradient fluctuations (Roberts, 2021; Smith et al., 2021; Beneventano, 2023; Ziyin et al., 2025). As we formalize in Section 3.2, in the limit $\eta \to 0^+$, this leads to a selection rule: the leading loss first restricts the solution to the global-minimum manifold $\mathcal{M}$, after which the SGD-induced correction selects minima with smaller gradient fluctuation.

The crucial observation is that gradient fluctuation and Hessian sharpness are not the same object. When these two objectives are misaligned, SGD may prefer a solution with smaller gradient fluctuation even if that solution is sharper than other global minima. This distinction provides the central mechanism studied in the remainder of the paper. In particular, it allows the data distribution to influence the sharpness selected by SGD even when it leaves the minimizer itself unchanged.

3 General Theory

Notation: We use θ to denote the set of all trainable parameters. When used without a subscript, ∇ denotes the gradient with respect to θ. $\lambda_{\max}(\Sigma)$ denotes the maximal eigenvalue of Σ. $\nabla^2 z$ denotes the Hessian matrix of a scalar function z. Throughout this work, we use $\mathbb{E}$ to denote the average over the training set $\mathcal{D}_{\text{train}}$, which might be finite or infinite. I_d is the identity operator for $\mathbb{R}^d$, and d is often omitted when the context is clear. Other notations are introduced in the context.

3.1 Sharpness

For most of the manuscript, we use the Hessian trace as a measure of sharpness:

Definition 3.1. The sharpness $T(\theta) := \text{Tr}\mathbb{E}[\nabla^2 \ell(x, y, \theta)]$.

T has been used in many prior works as a metric of sharpness and is often found to better represent the stability of the learning dynamics (Li et al., 2021; Wu & Su, 2023; Ziyin et al., 2024), which is one of the primary reasons for studying the sharpness of the landscape. It also offers an upper and lower bound to the largest eigenvalue $\lambda_{\max}$ of a positive semi-definite Hessian: $T/p \leq \lambda_{\max} \leq T$, where p is the number of parameters. Thus, $T(\theta)$ serves as a reliable proxy for the stability of the solution. when $T \to \infty$ (e.g. in Lemma 3.2), $\lambda_{\max}$ must also diverge.

The necessity of studying SGD's implicit bias is highlighted by the fact that the loss landscape contains paths to arbitrarily sharp minima. The following lemma shows that for the class of symmetries called exponential symmetries (Ziyin et al., 2024), we can always find a global minimum with infinite sharpness:

Lemma 3.2. *Assume that A is a symmetric matrix and that for any x, y, θ, $\ell(x, y, e^{\lambda A}\theta) = \ell(x, y, \theta)$. Moreover, assume that $A\mathbb{E}_{x,y}\nabla^2 \ell(x, y, \theta) \neq 0$ and that $\mathbb{E}_{x,y}\nabla^2 \ell(x, y, \theta) \succeq 0$. Then,* $\limsup_{|\lambda|\to+\infty} |T(e^{\lambda A}\theta)| = +\infty$.

The matrix rescaling symmetry is a type of exponential symmetry and obeys Lemma 3.2. That symmetry implies the existence of sharp solutions is first identified and proved in Dinh et al. (2017) for the case of simple rescaling symmetries, and lemma 3.2 can be seen as the most general form of this result.

As an example of its meaning, consider its application to the deep linear network with $D = 2$. For $\ell(x, y, \theta) = \frac{1}{2}||y - Z_2 Z_1 x||^2$, we have

$$T(Z_2, Z_1) = ||Z_2||_F^2 \text{Tr}\Sigma_x + d_y \text{Tr}(Z_1^\top Z_1 \Sigma_x), \tag{2}$$

In this case, for any Z_1, Z_2 we have $\ell(x, y, Z_2, Z_1) = \ell(x, y, \lambda Z_2, \lambda^{-1} Z_1)$ but

$$\lim_{|\lambda| \to +\infty} T(\lambda Z_2, \lambda^{-1} Z_1) = +\infty. \tag{3}$$

As the main theorem 4.1 below will prove, at the global minima, the deep linear network will (1) have a unique minimal sharpness value $T_{\min}$ that only depends on the problem and will (2) converge to a unique sharpness value $T \geq T_{\min}$, this means that the actually learned level of sharpness can be decomposed into two contributions:

$$T = T_{\min} + \Delta_{\rm SGD}, \tag{4}$$

where $\Delta_{\rm SGD} \geq 0$, $T_{\min}$ is the minimal level of sharpness that is required in order to fit the data distribution (given the architecture) and needs to be obeyed by any learning algorithm that is capable of reaching the global minimum, and $\Delta_{\rm SGD}$ is the level of sharpness due to the implicit bias of SGD (namely, due to the fluctuation minimization effect that we will introduce in the next section).

The quantity $\lambda_{\max}$ is also of interest. In general, while our theory allows one to compute $\lambda_{\max}$ directly, its functional form is so complicated that we do not present it explicitly. That being said, we will discuss in Section 5 a few cases where $\lambda_{\max}$ can written in an interpretable form.

3.2 Minimal fluctuation ansatz

Training with SGD at a small learning rate and minibatch sampling have been shown to be equivalent to solving the following perturbed optimization problem, which we refer to as the "entropic" loss (Roberts, 2021; Smith et al., 2021; Beneventano, 2023; Ziyin et al., 2025):

$$F_\eta(\theta) = \underbrace{\mathbb{E}_{x,y} \ell(x, y, \theta)}_{\text{learning, symmetry}} + \underbrace{\frac{\eta}{4} \mathbb{E}_{\mathcal{B}} || \mathbb{E}_{(x,y)\in\mathcal{B}} \nabla \ell(x, y, \theta)||^2}_{\text{due to discretization error, noise, } := \mathcal{V}(\theta)}, \tag{5}$$

where $\mathcal{B}$ denotes a mini-batch and η is the learning rate. Using this modified entropic loss[1] allows us to characterize the implicit bias of SGD via the following lemma.

Lemma 3.3 (Minimal Fluctuation Lemma, informal)**.** *In the limit $\eta \to 0^+$, the global minima of $F_\eta(\theta)$ are*

$$\arg\min_{\theta \in \mathcal{M}} \mathcal{V}(\theta), \tag{6}$$

where $M = \{\theta \mid \mathbb{E}_{x,y} \ell(x, y, \theta) = \inf_u \mathbb{E}_{x,y} \ell(x, y, u)\}$ is the the global minima manifold.

Since modern neural networks are typically over-parameterized, the loss landscape $\mathbb{E}_{x,y} \ell(x, y, \theta)$ contains a manifold of global minima. The lemma suggests that SGD does not treat these minima equally but implicitly prefers those with minimal fluctuation S. Notably, this bias vanishes for full-batch gradient descent (GD), as the fluctuation term disappears at any global optimum.

A key mechanism of the proof is that different directions along the matrix rescaling symmetry actually have different levels of fluctuation (often also corresponding to different levels of sharpness, as shown by Lemma 3.2) – thus, the preference of the learning algorithm for the minimal-fluctuation solution causes the learning algorithm to also moves towards a certain level of sharpness.

Because

$$\mathbb{E}_{\mathcal{B}} \|\mathbb{E}_{(x,y)\in\mathcal{B}} \nabla \ell(x, y, \theta)\|^2 = \frac{1}{|\mathcal{B}|} \mathbb{E}_{x,y} \|\nabla \ell(x, y, \theta)\|^2 + \frac{|\mathcal{B}| - 1}{|\mathcal{B}|} \left\|\mathbb{E}_{x,y} \nabla \ell(x, y, \theta)\right\|^2 \tag{7}$$

is proportional to $\mathbb{E}_{x,y} \|\nabla \ell(x, y, \theta)\|^2$ at the minimum where $\mathbb{E}_{x,y} \nabla \ell(x, y, \theta) = 0$, we will choose $|\mathcal{B}| = 1$ in the following without loss of generality.

3.3 Fluctuation and sharpness

The minimal-fluctuation principle does not, in general, imply a preference for flat solutions. The key reason is that sharpness and gradient fluctuation measure two different properties of the model.

[1] Do not confuse with the "entropy" in physics.

In this section, we make this distinction precise for the mean-squared-error (MSE) loss and show that label-noise anisotropy is what permits the two objectives to become misaligned.

Consider $\ell(x, y, \theta) = \frac{1}{2}\|f_\theta(x) - y\|_2^2$, where $f_\theta(x) \in \mathbb{R}^{d_y}$. Let $J_\theta(x) := \frac{\partial f_\theta(x)}{\partial \theta} \in \mathbb{R}^{d_y \times p}$, where p is the total number of trainable parameters, and define the residual $e(x, y, \theta) := y - f_\theta(x)$. The loss gradient is $\nabla_\theta \ell(x, y, \theta) = -J_\theta(x)^\top e(x, y, \theta)$, and its Hessian is

$$\nabla_\theta^2 \ell(x, y, \theta) = J_\theta(x)^\top J_\theta(x) - \sum_{j=1}^{d_y} e_j(x, y, \theta) \nabla_\theta^2 [f_\theta(x)]_j. \tag{8}$$

This is the standard Gauss–Newton decomposition (Martens et al., 2010).

Assuming that at the minimizer, $\mathbb{E}[e \mid x] = 0$ and $\mathbb{E}[ee^\top \mid x] = \Sigma_\epsilon$. Taking the expectation of Eq. (8), the second term vanishes, and the Hessian becomes (Wu & Su, 2023; MacDonald et al., 2023)

$$H(\theta) = \mathbb{E}_x \left[J_\theta(x)^\top J_\theta(x) \right]. \tag{9}$$

It is useful to introduce $\Gamma(\theta) := \mathbb{E}_x \left[J_\theta(x) J_\theta(x)^\top \right] \succeq 0$. The sharpness is then simply

$$T(\theta) := \mathrm{Tr} H(\theta) = \mathrm{Tr}\Gamma(\theta) = \mathbb{E}_x \|J_\theta(x)\|_F^2. \tag{10}$$

In contrast, consider the fluctuation (second moment of the gradient),

$$\mathcal{V}(\theta) := \mathbb{E}_{x,y} \left[\|\nabla_\theta \ell(x, y, \theta)\|_2^2 \right]. \tag{11}$$

Using $\nabla_\theta \ell = -J_\theta^\top e$, we obtain

$$\mathcal{V}(\theta) = \mathbb{E}_{x,y} \left[e^\top J_\theta(x) J_\theta(x)^\top e \right] = \mathrm{Tr} \left[\Sigma_\epsilon \Gamma(\theta) \right]. \tag{12}$$

Thus, sharpness and fluctuation are respectively

$$T(\theta) = \mathrm{Tr}\Gamma(\theta), \qquad \mathcal{V}(\theta) = \mathrm{Tr}[\Sigma_\epsilon \Gamma(\theta)]. \tag{13}$$

Equation (13) reveals the essential difference between sharpness and fluctuation. Gradient fluctuation weights each direction according to the covariance Σ_ϵ. This immediately gives the following sharpness–fluctuation relation:

$$\frac{\mathcal{V}(\theta)}{\lambda_{\max}(\Sigma_\epsilon)} \leq T(\theta) \leq \frac{\mathcal{V}(\theta)}{\lambda_{\min}(\Sigma_\epsilon)}. \tag{14}$$

Thus when the label noise is isotropic, sharpness and fluctuation induce the same ordering over solutions. Once the label noise is anisotropic, such an ordering need not exist. To see this, let $\lambda_{\min} < \lambda_{\max}$ be the smallest and largest eigenvalues of Σ_ϵ, with corresponding unit eigenvectors $u_{\min}$ and $u_{\max}$. Choose any $r \in \left(\frac{\lambda_{\min}}{\lambda_{\max}}, 1 \right)$ and consider two positive-semidefinite matrices $\Gamma_1 = u_{\min} u_{\min}^\top$ and $\Gamma_2 = r\, u_{\max} u_{\max}^\top$. Their sharpnesses satisfy

$$\mathrm{Tr}\Gamma_1 = 1 > r = \mathrm{Tr}\Gamma_2, \tag{15}$$

whereas their fluctuations obey

$$\mathrm{Tr}[\Sigma_\epsilon \Gamma_1] = \lambda_{\min} < r\lambda_{\max} = \mathrm{Tr}[\Sigma_\epsilon \Gamma_2]. \tag{16}$$

Thus, a solution can simultaneously have *smaller gradient fluctuation and larger sharpness*. In other words, anisotropic label noise makes it possible for fluctuation minimization to favor a sharper solution.

This observation has a simple geometric interpretation. The matrix $\Gamma(\theta)$ measures how strongly the model output responds to parameter perturbations in each output direction. Sharpness penalizes all such directions equally, whereas SGD-induced fluctuation assigns a larger penalty to directions with larger residual variance. Consequently, minimizing fluctuation encourages the model to suppress parameter sensitivity in high-noise output directions, while allowing substantially larger sensitivity in low-noise directions. The latter can increase the sharpness even while the fluctuation decreases.

The argument above only establishes that anisotropy makes such a reversal possible. The deep linear model studied in the next section makes this mechanism exact. There, the global-minimum manifold contains solutions with arbitrarily different sharpnesses, and the minimum-fluctuation solution can be solved in closed form. We will show that isotropic noise selects the minimum-sharpness solution, whereas anisotropic noise generically selects a sharper one controlled by the label-noise spectrum.

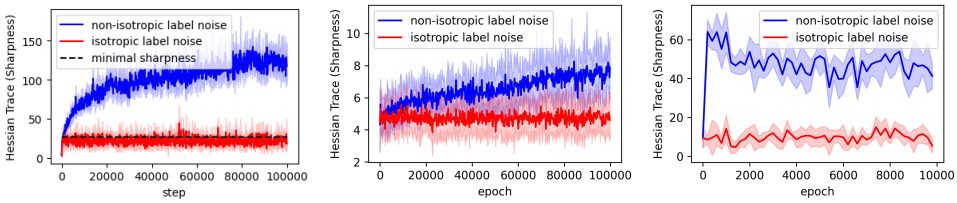


Figure 2: Anisotropic noise in the labels leads to progressive sharpening. **Left**: linear networks, with the minimal sharpness predicted by (23). **Middle**: Two layer ReLU networks trained under a teacher-student setting. **Right**: Two layer ReLU networks trained on MNIST. Each line is averaged over five trials shown with the standard error. See Appendix C for experiment details and more experiments.

4 Deep Residual Linear Network: An Exactly Solvable Model

In this section, we first motivate and formally define the problem setting, and we then present the main theorem, whose proof is given in the appendix. We will discuss the various implications of the theory in the next main section.

4.1 Model and Loss Function

We study a D-layer deep residual linear network with possible residual connections. For simplicity, we assume $D \geq 2$. Specifically, the model is defined as

$$f_\theta(x) = (W_D + R_D) \cdots (W_1 + R_1)x. \tag{17}$$

where $\theta := \{W_k + R_k\}_{k=1}^D$ denotes the collection of all the weight matrices, and x is the input data. $\{W_k\}_{k=1}^D$ are trainable and $\{R_k\}_{k=1}^D$ are the fixed but arbitrary residual connections. This covers the standard residue networks ($R_k = I$) and networks without residues ($R_k = 0$) but is more general. For the following we denote $Z_k := W_k + R_k$. The model is trained with the MSE loss on a regression task, where the data-wise stochastic loss is

$$\ell(x, y, \theta) = \frac{1}{2}||y - Z_D \cdots Z_1 x||^2, \tag{18}$$

where y is the label associated with x. In practice, we train this model with SGD, but the training dynamics of SGD on this model are difficult to solve analytically, and we thus adopt an alternative approach, which can be seen as answering the question of "if the training proceeds with SGD, then what kind of objective function is the algorithm actually minimizing?" This alternative problem, in contrast, can be exactly solved to its full generality. This alternative objective function will be discussed in Section 3.2.

The labels are generated by a linear teacher V with noise ϵ

$$y = Vx + \epsilon \in \mathbb{R}^{d_y}, \tag{19}$$

where ϵ and x are independent and we define the covariances as $\Sigma_x := \mathbb{E}_x[xx^\top]$ and $\Sigma_\epsilon := \mathbb{E}_\epsilon[\epsilon\epsilon^\top]$. Note that all input, label, V, and noise are allowed to have arbitrary dimensions as long as their dimensions are mutually consistent. We consider the population (infinite training set) limit such that the global minima of the population loss satisfy (assuming $\Sigma_x \succ 0$):

$$Z_D \cdots Z_1 x = Vx. \tag{20}$$

As a key part of the analysis, note that the model has many symmetries (namely, loss-invariant directions), which we call the matrix rescaling symmetry. For an arbitrary invertible matrix A, and an arbitrary layer index i, we have

$$f(Z_i, Z_{i-1}) := Z_D \cdots Z_i Z_{i-1} \cdots Z_1 x = f(Z_i A, A^{-1} Z_{i-1}), \tag{21}$$

which implies that the deep linear network has infinitely many global minima. Then Lemma 3.2 implies the existence of a global minimum with arbitrarily high sharpness.

4.2 Main Result

Let $V' := \sqrt{\Sigma_\epsilon} V \sqrt{\Sigma_x}$. Let its SVD decomposition be given by $V' = LSR^\top$. We assume the network width $d \geq \text{rank}(V')$[2]. Under the SGD limit $\eta \to 0^+$, we obtain the following closed-form result for the sharpness:

[2] d is defined as the the smallest dimension of the output and input dimensions of W_i for all $i \in [D]$. When $d < \text{rank}(V')$, it learns the largest d singular values, and the result is similar.

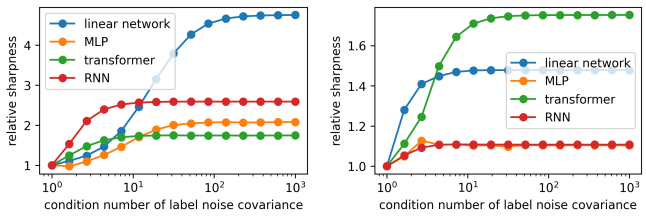


Figure 3: Across all four models, and for both regression and classification tasks, we observe that the learning process yields a sharper solution as the noise in the labels becomes increasingly anisotropic. We define relative sharpness as the ratio of the sharpness under anisotropic noise to that under isotropic noise. **Left**: MSE loss. **Right**: Cross entropy loss.

Theorem 4.1. *Assume that $\Sigma_x, \Sigma_\epsilon \succ 0$. For all global minima of* (5)*, the spectrum of the Hessian is identical (determined by $\Sigma_x, \Sigma_\epsilon$ and V), and the sharpness is given by*

$$T(\theta) = (D-1)\,(\mathrm{Tr}S)^{(D-2)/D}\,(\mathrm{Tr}\Sigma_\epsilon \mathrm{Tr}\Sigma_x)^{1/D}\mathrm{Tr}[\Sigma_\epsilon^{-1} LSL^\top] + d_y\,(\mathrm{Tr}S)^{2(D-1)/D}\,\frac{(\mathrm{Tr}\Sigma_x)^{1/D}}{(\mathrm{Tr}\Sigma_\epsilon)^{(D-1)/D}}. \tag{22}$$

In addition, the minimal sharpness of the global minimum of $\mathbb{E}_x \ell(x, \theta)$ is given by

$$T_{\min} := \min_{\theta \in \mathcal{M}} T(\theta) = D(\mathrm{Tr}\hat{S})^{2(D-1)/D}(d_y \mathrm{Tr}\Sigma_x)^{1/D}, \tag{23}$$

where $\hat{S}$ is the singular values of $V\sqrt{\Sigma_x}$.

Theorem 4.1 reveals a striking property: although there exist infinitely many global minima with varying (and even infinite) sharpness, SGD consistently selects a solution with a **unique** sharpness level. This suggests that the final sharpness is independent of initialization, provided the model converges to a global optimum. The result is surprisingly also independent of the residual connectivities R_i, which means that the sharpness reached is quite insensitive to architecture details, at least in this model. Moreover, the first term of (22) suggests that $D-1$ layers contribute to the sharpness in the same way, while the last layer (the second term of (22)) contributes differently. At the minimal sharpness solution (23), however, all layers contribute in the same way. We will analyze the implications of this theorem in detail in Section 5.

5 Insights

A great merit of exactly solvable models is that every effect and phenomenon within the model can be precisely and causally understood. This section is thus devoted to naming and understanding the key observations and insights of Theorem 4.1. We then present experiments on a diverse set of model architectures to demonstrate the qualitative relevance of the prediction. The proofs of the theoretical results are presented in Appendix B.

Noise Imbalance as a Source of Sharpness.

Corollary 5.1. *Suppose that $\Sigma_\epsilon \succ 0$ and that $A := V\sqrt{\Sigma_x}$ has full row rank. Then there exist constants $c_1, c_2 > 0$, depending only on A, Σ_x, and d_y, such that the sharpness at* (77) *satisfies*

$$c_1 D\sqrt{\kappa(\Sigma_\epsilon)} \le T(\theta) \le c_2 D\sqrt{\kappa(\Sigma_\epsilon)}, \tag{24}$$

where $\kappa(\Sigma_\epsilon) := \frac{\lambda_{\max}(\Sigma_\epsilon)}{\lambda_{\min}(\Sigma_\epsilon)}$.

Corollary 5.1 suggests that the sharpness mainly depends on the imbalance of noise in the labels, and does not depend much on the depth, in the sense that $\frac{T(\theta)}{T_{\min}} = \Theta(\sqrt{\kappa(\Sigma_\epsilon)})$.

Recovery of Minimal Sharpness under Isotropic Noise.

Corollary 5.2. *Suppose that $\Sigma_\epsilon \succ 0$ and that $V\sqrt{\Sigma_x}$ has full row rank. Then the minimum-fluctuation solutions characterized by Eq.* (77) *attain the minimum possible sharpness if and only if $\Sigma_\epsilon = cI$ for some $c > 0$.*

The reason behind Corollary 5.2 is that for deep linear networks with $\Sigma_\epsilon = I$, the sharpness is exactly proportional to the fluctuation (the second order derivative in (8) disappears for linear networks), and thus minimizing fluctuation is equivalent to minimizing sharpness.

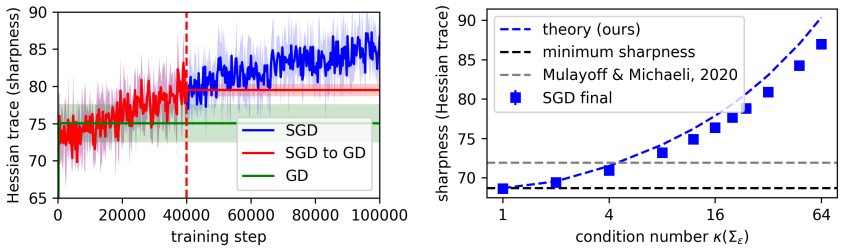


Figure 4: **Left**: Progressive sharpening with anisotropic noise in the labels is unique for SGD. The red dotted line denotes the time when we switch to GD. **Right**: Across various condition numbers of the label noise, our theory consistently provides a good prediction for the final sharpness that SGD reaches after 300000 steps. The blue dotted line denotes the prediction (22) of our theory. The black dotted line denotes the minimal sharpness (23). The gray dotted line refers to the solution that minimizes the maximal eigenvalue of the Hessian (Mulayoff & Michaeli, 2020).

Largest Eigenvalue Has a Similar Property

Corollary 5.3 (Spectral sharpness under anisotropic label noise)**.** *Suppose that $VV^\top$ commutes with Σ_ϵ and $V^\top V$ commutes with Σ_x. In addition, assume that $\Sigma_x \succ 0$ and that V is full rank. Then there exist constants $c_1, c_2 > 0$, depending only on V, Σ_x, and the output dimension, such that*

$$c_1 D\sqrt{\kappa(\Sigma_\epsilon)} \le \lambda_{\max}(H) \le c_2 D\sqrt{\kappa(\Sigma_\epsilon)}, \tag{25}$$

where $\kappa(\Sigma_\epsilon) := \frac{\lambda_{\max}(\Sigma_\epsilon)}{\lambda_{\min}(\Sigma_\epsilon)}$.

Corollary 5.3 suggests that in a special case where all the matrices are aligned, both the maximal eigenvalue and the trace of the Hessian grow at the same rate $\sqrt{\kappa(\Sigma_\epsilon)}$. We introduce the commutation assumption because under this assumption Theorem B.5 can precisely characterize the Hessian spectrum. We expect that Corollary 5.3 should be qualitatively true in general.

Feature noise can promote flatness or sharpness Suppose that the data are generated according to $y = Vx+\epsilon$, but the model is trained on a noisy observation of the input, $x' = x+\epsilon_x$, where x, ϵ_x, ϵ are zero-mean, mutually independent Gaussian random variables with covariances $\Sigma_x, \Sigma_{\epsilon_x}, \Sigma_\epsilon$, respectively. In this case, the feature noise can either increase or decrease the sharpness of the selected solution. Isotropic feature noise tends to decrease the sharpness while anisotropic feature noise tends to increase the sharpness. See Appendix B.9 for some discussions.

Sharpening or flattening. Theorem 4.1 suggests that SGD converges to a sharpness level that is independent of the initialization. Consequently, progressive sharpening (at least on average and throughout the trajectory) occurs if the initialization is sufficiently small, while progressive flattening occurs if it is large. In Proposition B.4, we characterize the initial sharpness and observe that it remains independent of the labels for standard initialization schemes. Comparing Proposition B.4 with Theorem 4.1 shows many mechanisms of progressive sharpening. Noting that the initial sharpness is independent of the label or noise distribution, while the final sharpness is label and noise-dependent, one finds a simple and straightforward mechanism for introducing flattening and sharpening: for a fixed initialization, rescaling the labels y to be smaller causes a flattening, while rescaling y to be larger can cause sharpening. Again, this showcases the key message: SGD intrinsically may not have a preference for sharpness, and it is the data distribution that changes, if not determines, the sharpness.

A *new* mechanism of progressive sharpening. As mentioned in Section 2, most previous known mechanisms were about gradient flow and full batch GD, but the progressive sharpening discovered in this paper is about the noise covariance, and cannot be seen in a deterministic training pipeline. For GD, the second term of (5) is always zero at any local minimum and does not lead to the progressive sharpening, as demonstrated in Figure 4 (Left). Moreover, Figure 4 (Right) also suggests that the performance of SGD can be explained by our theory (blue dashed line), but not the solution that minimizes the maximal eigenvalue of the Hessian (Mulayoff & Michaeli, 2020) (gray dashed line). The gap between our theory and the experiments for large κ is because we fix the number of training steps, but it takes much longer for SGD to converge with larger κ.

Experiments. To empirically validate our theory, we perform experiments on both deep linear networks and several nonlinear architectures (MLP, RNN, and Transformer) across regression and classification tasks. These results, shown in Figures 2 and 3, demonstrate that the observed phenomena are not unique to linear models; specifically, imbalanced label noise induces sharper solutions as predicted by Corollary 5.1. Furthermore, the left side of Figure 2 verifies Corollary 5.2, where linear networks converge to the minimal sharpness solution under isotropic label noise.

In Figure 3, we observe that sharpness increases with the non-isotropy of label noise. The plateau at high condition numbers is a consequence of the finite-horizon effect, where highly imbalanced noise necessitates a very long convergence time to reach the sharpest minima. In Appendix C, we further verify that this is also true for a real vision transformer trained on the CIFAR-10 dataset. To isolate the role of stochasticity, Figure 4 (Left) presents a direct intervention: when switching to GD in the linear network setting, progressive sharpening disappears. This identifies the mechanism as unique to SGD, aligning with (5) where the second term vanishes for GD at local minima, exerting no implicit bias. Comprehensive details and additional experiments on eigenvalues, condition numbers, and Transformers can be found in Appendix C.

6 Related Works

A closely related work is Gatmiry et al. (2023), which identifies the solution for minimal sharpness (trace of the Hessian) for deep linear networks, and it is not surprising that our solution agrees with theirs for isotropic noise. Similarly, Mulayoff & Michaeli (2020) analyzes the solution that minimizes the maximal eigenvalue of the Hessian for deep linear networks. Therefore, there is no sharpness-seeking behavior in these models. Another closely-related work is Ziyin et al. (2025), which solves the sharpness problem for a two-layer linear network and does not investigate whether the theory is relevant to nonlinear models. In contrast, we solve a model with arbitrary residual connections and arbitrary depths, and demonstrated the relevance to a broad class of nonlinear models.

Other theoretical modesl of the progressive sharpening also exist. Agarwala et al. (2022) analytically demonstrated progressive sharpening in a quadratic model as a direct consequence of the edge of stability (EoS). Damian et al. (2022) attributed these effects to the influence of cubic terms in the Taylor expansion of the loss function. Furthermore, Ahn et al. (2023) proved the existence of EoS in two-layer ReLU networks and highlighted its benefits for generalization, while Beneventano & Woodworth (2025) argued that large learning rates in linear models implicitly bias GD toward flatter solutions. Our work departs from these studies by demonstrating that such phenomena can arise from the intrinsic stochasticity of SGD, rather than being solely a function of the step size (Nar & Sastry, 2018; Mulayoff & Michaeli, 2020). We also discuss more related works on deep linear networks and roles of stochastic noise in training in Appendix A.

7 Conclusion

In this work, we studied how and when SGD training can have a flatness or sharpness-seeking behavior. We provide an exact analytical solution for the deep linear networks under the minimal gradient fluctuation constraint. In this model, we demonstrate that the converged sharpness is unique and independent of initialization. A key insight is that noise in the labels imbalance is a primary driver of the sharpening phenomenon – a conclusion further validated through experiments on realistic datasets across model types. This is a novel form of progressive sharpening, of a different nature from the ones previously discussed in the literature.

There are various limitations, each of which is worthwhile to be addressed by future work. First of all, our theory is effective in nature and does not directly reveal the dynamical aspects of training. A valuable open problem is thus to study the dynamics of reaching sharper or flatter solutions. Secondly, our theory only studies the leading order effect in η during training, which tends to suppress fluctuation. This essentially corresponds to the small learning rate regime of training (but not as small as what is needed for gradient flow), and it is unclear what role higher-order terms in the entropic loss could play. Thus, understanding the flattening and sharpening effects with larger learning rates remains open. For example, it could well be the case that at a larger learning rate, the sharpness-seeking behavior is suppressed again. For a similar reason, within our framework, only progressive sharpening can be explained, but not EoS. This is because to derive the entropic loss (5), one needs to assume that the learning is stable (Smith et al., 2021). Solving these problems will certainly advance our understanding of deep learning optimization.

REFERENCES

Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. *arXiv preprint arXiv:2210.04860*, 2022.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36:19540–19569, 2023.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd. *arXiv preprint arXiv:2412.20553*, 2024.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

Pierfrancesco Beneventano. On the trajectories of sgd without replacement. *arXiv preprint arXiv:2312.16143*, 2023.

Pierfrancesco Beneventano and Blake Woodworth. Gradient descent converges linearly to flatter minima than gradient flow in shallow linear networks. *arXiv preprint arXiv:2501.09137*, 2025.

Pierfrancesco Beneventano, Andrea Pinto, and Tomaso Poggio. How Neural Networks Learn the Support is an Implicit Regularization Effect of SGD. *arXiv:2406.11110 [cs, math, stat]*, June 2024. doi: 10.48550/arXiv.2406.11110. URL http://arxiv.org/abs/2406.11110. arXiv:2406.11110 [cs, math, stat].

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020.

Yuhang Cai, Haofeng Huang, Haodong Wen, Deyi Liu, Yiyuan Ma, and Kaifeng Lyu. Does llm pre-training typically occur at the edge of stability? In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026.

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. *arXiv preprint arXiv:2103.00065*, 2021.

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.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp Minima Can Generalize For Deep Nets. *ArXiv e-prints*, March 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Khashayar Gatmiry, Zhiyuan Li, Tengyu Ma, Sashank Reddi, Stefanie Jegelka, and Ching-Yao Chuang. What is the inductive bias of flatness regularization? a study of deep matrix factorization models. *Advances in Neural Information Processing Systems*, 36:28040–28052, 2023.

Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pp. 2315–2357. PMLR, 2021.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Wei Huang, Andi Han, Yujin Song, Yilan Chen, Denny Wu, Difan Zou, and Taiji Suzuki. How does label noise gradient descent improve generalization in the low snr regime? *arXiv preprint arXiv:2510.17526*, 2025.

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the Relation Between the Sharpest Directions of DNN Loss and the SGD Step Length, December 2019. URL http://arxiv.org/abs/1807.05031. arXiv:1807.05031 [stat].

Stanisław Jastrzębski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The Break-Even Point on Optimization Trajectories of Deep Neural Networks. *arXiv:2002.09572 [cs, stat]*, February 2020. URL http://arxiv.org/abs/2002.09572. arXiv:2002.09572.

Anil Kamber and Rahul Parhi. Sharpness of minima in deep matrix factorization: Exact expressions. *arXiv preprint arXiv:2509.25783*, 2025.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

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. *arXiv preprint arXiv:1609.04836*, 2016.

Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pp. 2902–2907. PMLR, 2018.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2021.

Lachlan Ewen MacDonald, Jack Valmadre, and Simon Lucey. On progressive sharpening, flat minima and generalisation. *arXiv preprint arXiv:2305.14683*, 2023.

Pierre Marion and Lénaïc Chizat. Deep linear networks for regression are implicitly regularized towards flat minima. *Advances in Neural Information Processing Systems*, 37:76848–76900, 2024.

James Martens et al. Deep learning via hessian-free optimization. In *Icml*, volume 27, pp. 735–742, 2010.

Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *International conference on machine learning*, pp. 7108–7118. PMLR, 2020.

Kamil Nar and Shankar Sastry. Step size matters in deep learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. *Advances in Continuous and Discrete Models*, 2024(1):23, 2024.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

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

Daniel A. Roberts. SGD Implicitly Regularizes Generalization Error. *arXiv:2104.04874 [cs, stat]*, April 2021. URL http://arxiv.org/abs/2104.04874. arXiv:2104.04874.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

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.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A Walk with SGD, May 2018. URL `http://arxiv.org/abs/1802.08770`. arXiv:1802.08770 [cs, stat].

Ziqing Xu, Hancheng Min, Salma Tarmoun, Enrique Mallada, and René Vidal. Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2262–2284. PMLR, 2023.

Liu Ziyin, Botao Li, and Xiangming Meng. Exact solutions of a deep linear network. *Advances in Neural Information Processing Systems*, 35:24446–24458, 2022.

Liu Ziyin, Mingze Wang, Hongchao Li, and Lei Wu. Parameter symmetry and noise equilibrium of stochastic gradient descent. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Liu Ziyin, Yizhou Xu, and Isaac Chuang. Neural thermodynamics: Entropic forces in deep and universal representation learning. *Advances in Neural Information Processing Systems*, 2025.

A More Related Works

Effect of noise in the labels. There are a lot of studies on the implicit bias of noise in the labels. It is shown that the noise in the labels has an implicit regularization effect that biases the model to minima of weaker noise Blanc et al. (2020); Li et al. (2021); HaoChen et al. (2021) or flatter minimum Damian et al. (2021). They also show that such a regularization effect is beneficial for generalization Pillaud-Vivien et al. (2022); Huang et al. (2025). Importantly, the cited theoretical analyses predominantly focus on scalar-output settings, where output-noise anisotropy is absent by construction. Our work, in contrast, shows that the noisy labeling can also bias the model to sharper solutions, dependent on the condition number.

Deep linear networks. The key object of our study is the deep linear networks (also known as deep matrix factorization). Deep linear networks serve as essential proxies for understanding the training dynamics of complex architectures. However, most existing literature focuses on their behavior under gradient descent or gradient flow (Saxe et al., 2013; Arora et al., 2018; Xu et al., 2023; Nguegnang et al., 2024), where some of them suggest that GD possesses an implicit bias toward flatter minima in these models (Marion & Chizat, 2024; Beneventano & Woodworth, 2025). While some studies have also characterized their loss landscapes—showing that all local minima are global (Kawaguchi, 2016; Laurent & Brecht, 2018)—the role of noise remains less explored. Although Ziyin et al. (2022) analyzed the global minima of deep linear networks with weight decay, our analysis specifically addresses the implicit bias induced by the stochasticity of SGD. Other closely related works are Pesme et al. (2021); Beneventano et al. (2024), which analyze diagonal neural networks (multiple dynamics with 1-dimensional label covariances), and show that the stochasticity is beneficial.

B THEORY

B.1 PROOF OF LEMMA 3.2

Proof. By the definition of A-symmetry, the loss function satisfies the following invariance property for any λ:

$$\ell(x, y, e^{\lambda A}\theta) = \ell(x, y, \theta). \tag{26}$$

To analyze the sharpness, we take the second-order derivative with respect to θ on both sides. Applying the chain rule, we obtain:

$$e^{\lambda A}\nabla^2_{e^{\lambda A}\theta}\ell(x, y, e^{\lambda A}\theta)e^{\lambda A} = \nabla^2\ell(x, y, \theta), \tag{27}$$

where we use the fact that A is symmetric.

Using the linearity and cyclic property of the trace, we have:

$$T(e^{\lambda A}\theta) := \mathrm{Tr}[\mathbb{E}_{x,y}\nabla^2_{e^{\lambda A}\theta}\ell(x, y, e^{\lambda A}\theta)] = \mathrm{Tr}[e^{-2\lambda A}\mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta)]. \tag{28}$$

Let $A := \sum_i \mu_i n_i n_i^\top$. The sharpness can then be expressed as a weighted sum:

$$T(e^{\lambda A}\theta) = \sum_i e^{-2\lambda\mu_i}(n_i^\top \mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta)n_i). \tag{29}$$

By the assumption $A\mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta) \neq 0$ and $\mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta) \succeq 0$, there must exist at least one index i such that $\mu_i \neq 0$ and $n_i^\top \mathbb{E}_{x,y}\nabla^2\ell(x, y, \theta)n_i \neq 0$. Consequently, the term $e^{-2\lambda\mu_i}$ will dominate as λ tends to infinity with the appropriate sign. Specifically, we have $\lim_{\lambda\to+\infty} |T(e^{\lambda A}\theta)| = +\infty$ if there exists $\mu_i < 0$, and $\lim_{\lambda\to-\infty} |T(e^{\lambda A}\theta)| = +\infty$ if there exists $\mu_i > 0$. □

B.2 PROOF OF LEMMA 3.3

We first provide a formal statement of Lemma 3.3.

Lemma B.1. *Let $f, \mathcal{V}$ be continuous functions, and let f be coercive (i.e., $\lim_{|\theta|\to\infty} f(\theta) = +\infty$). Define the perturbed objective function*

$$F_\eta(\theta) = f(\theta) + \eta\mathcal{V}(\theta). \tag{30}$$

Let $\{\eta_k\}_{k=1}^\infty$ be a sequence of positive numbers such that $\eta_k \to 0$ as $k \to \infty$. Suppose

$$\theta_k \in \arg\min_\theta f(\theta) + \eta_k\mathcal{V}(\theta). \tag{31}$$

and $\theta_k \to \bar{\theta}$[3] as $k \to \infty$. Then $\bar{\theta}$ minimizes $\mathcal{V}$ over the set of minimizers of f, i.e.,

$$\bar{\theta} \in \arg\min_{\theta\in M} \mathcal{V}(\theta), \quad \textit{where } M := \arg\min_\theta f(\theta). \tag{32}$$

Proof. Since f is coercive and continuous, the set of minimizers M is non-empty and compact. Let $y \in M$ be any global minimizer of f. By the optimality of θ_k for F_{η_k}, we have:

$$f(\theta_k) + \eta_k\mathcal{V}(\theta_k) \leq f(y) + \eta_k\mathcal{V}(y). \tag{33}$$

Rearranging the terms, we obtain:

$$f(\theta_k) - f(y) \leq \eta_k(\mathcal{V}(y) - \mathcal{V}(\theta_k)) \tag{34}$$

Since $y \in M$, we have $f(\theta_k) \geq f(y)$, which implies $f(\theta_k) - f(y) \geq 0$. Consequently, from (34) and the fact that $\eta_k > 0$, it follows that:

$$\mathcal{V}(\theta_k) \leq \mathcal{V}(y). \tag{35}$$

Taking the limit as $k \to \infty$ in (34), and noting that $\mathcal{V}$ is continuous (thus $\mathcal{V}(\theta_k)$ remains bounded as $\theta_k \to \bar{\theta}$), the right-hand side vanishes:

$$0 \leq \lim_{k\to\infty}(f(\theta_k) - f(y)) \leq \lim_{k\to\infty} \eta_k(\mathcal{V}(y) - \mathcal{V}(\theta_k)) = 0. \tag{36}$$

[3]By the optimality of θ_k, for any $y \in M$ we have $f(\theta_k) \leq f(y) + \eta_k\mathcal{V}(y)$. Hence, for sufficiently large k, the sequence θ_k lies in a bounded sublevel set of f. Since f is coercive, these sublevel sets are compact, and therefore θ_k admits a convergent subsequence. For simplicity we assume that θ_k converges.

By the continuity of f, we have $f(\bar{\theta}) = f(y)$, which implies $\bar{\theta} \in M$.

Finally, we consider the inequality (35). Taking the limit $k \to \infty$ on both sides and utilizing the continuity of $\mathcal{V}$, we obtain:

$$\mathcal{V}(\bar{\theta}) \leq \mathcal{V}(y) \quad \text{for all } y \in M. \tag{37}$$

This inequality demonstrates that $\bar{\theta}$ minimizes $\mathcal{V}$ over M, i.e., $\bar{\theta} \in \arg\min_{\theta \in M} \mathcal{V}(\theta)$, which completes the proof. □

B.3 Proof of Theorem 4.1

Proof. Step 1: Computing the Hessian Trace.

We begin by evaluating the trace of the Hessian with respect to the weight matrices $\{W_k\}_{k=1}^{D}$. For each layer k, let $A_k := Z_D \cdots Z_{k+1}$ and $B_k := Z_{k-1} \cdots Z_1$, with the conventions $A_D = I_{d_y}$ and $B_1 = I_{d_x}$. The loss function for a single sample can be written as $\ell(x, y, \theta) = \frac{1}{2}\|y - A_k(W_k + R_k)B_k x\|^2$. The second-order partial derivative with respect to W_k yields the following identity:

$$\mathrm{Tr}\frac{\partial^2 \ell}{\partial W_k^2} = \mathrm{Tr}A_k^\top A_k \mathrm{Tr}B_k x x^\top B_k^\top. \tag{38}$$

By taking the expectation over the data distribution and summing over all layers $k = 1, \ldots, D$, the total sharpness $T(\theta)$ is given by:

$$T(\theta) = \sum_{k=1}^{D} \mathbb{E}_{x,y}\mathrm{Tr}\frac{\partial^2 \ell}{\partial W_k^2} = \sum_{k=1}^{D} \mathrm{Tr}[A_k^\top A_k]\mathrm{Tr}[B_k \Sigma_x B_k^\top]. \tag{39}$$

This establishes the structural form of the sharpness. The minimization of $T(\theta)$ as per (23) then follows directly from applying Lemma B.2 to the summation in (39).

Step 2: Analysis of Sharpness at the Minimum of (5).

For the solutions from Lemma B.3, the product matrices are decomposed as

$$A_k = \sqrt{\Sigma_\epsilon^{-1}} L \Sigma_D \cdots \Sigma_{k+1} U_k^\top,\ B_k = U_{k-1}\Sigma_{k-1} \cdots \Sigma_1 R^\top \Sigma_x^{-1/2}. \tag{40}$$

For the intermediate layers $k \in \{2, \ldots, D-1\}$, the trace product simplifies to:

$$\begin{aligned} \mathrm{Tr}[A_k^\top A_k]\mathrm{Tr}[B_k \Sigma_x B_k^\top] &= \mathrm{Tr}[\Sigma_\epsilon^{-1} L \Sigma_D^2 \cdots \Sigma_{k+1}^2 L^\top]\mathrm{Tr}[\Sigma_1^2 \cdots \Sigma_{k-1}^2] \\ &= (\mathrm{Tr}S)^{-2/D} \left(\mathrm{Tr}[\Sigma_\epsilon]\mathrm{Tr}[\Sigma_x]\right)^{1/D} \mathrm{Tr}[\Sigma_\epsilon^{-1} L S L^\top]\mathrm{Tr}[S] \end{aligned} \tag{41}$$

The boundary layers require separate treatment. For $k = 1$ and $k = D$, we have:

$$\begin{aligned} \mathrm{Tr}[A_1^\top A_1]\mathrm{Tr}[B_1 \Sigma_x B_1^\top] &= \mathrm{Tr}[\Sigma_\epsilon^{-1} L \Sigma_D^2 \cdots \Sigma_2^2 L^\top]\mathrm{Tr}[\Sigma_x] \\ &= (\mathrm{Tr}S)^{(D-2)/D} \frac{\mathrm{Tr}[\Sigma_\epsilon]^{1/D}}{\mathrm{Tr}[\Sigma_x]^{(D-1)/D}} \mathrm{Tr}[\Sigma_\epsilon^{-1} L S L^\top]\mathrm{Tr}[\Sigma_x] \end{aligned} \tag{42}$$

and

$$\mathrm{Tr}[A_D^\top A_D]\mathrm{Tr}[B_D \Sigma_x B_D^\top] = \mathrm{Tr}[I_{d_y}]\mathrm{Tr}[\Sigma_1^2 \cdots \Sigma_{D-1}^2] = d_y \left(\mathrm{Tr}S\right)^{(D-2)/D} \frac{\mathrm{Tr}[\Sigma_x]^{1/D}}{\mathrm{Tr}[\Sigma_\epsilon]^{(D-1)/D}} \mathrm{Tr}[S]. \tag{43}$$

Summing these contributions over $k = 1, \ldots, D$ yields the final expression for $T(\theta)$:

$$\begin{aligned} T(\theta) &= \sum_{k=1}^{D} \mathrm{Tr}[A_k^\top A_k]\mathrm{Tr}[B_k \Sigma_x B_k^\top] \\ &= (D-1)\left(\mathrm{Tr}S\right)^{(D-2)/D} \left(\mathrm{Tr}[\Sigma_\epsilon]\mathrm{Tr}[\Sigma_x]\right)^{1/D} \mathrm{Tr}[\Sigma_\epsilon^{-1} L S L^\top] \\ &\quad + d_y \left(\mathrm{Tr}S\right)^{2(D-1)/D} \frac{\mathrm{Tr}[\Sigma_x]^{1/D}}{\mathrm{Tr}[\Sigma_\epsilon]^{(D-1)/D}}, \end{aligned} \tag{44}$$

which proves (22).

Step 3: Invariance of the Hessian Spectrum.

Finally, we verify that the spectrum of the Hessian $H(\theta)$ is constant across the manifold of solutions defined in Lemma B.3. Consider two sets of weight matrices $\theta = \{Z_k\}$ and $\theta' = \{Z'_k\}$. According to the lemma, any two such solutions are related by a sequence of orthogonal transformations $R_k \in O(d_k)$, such that $Z'_k = R_k Z_k R_{k-1}^\top$. This transformation can be represented as $\theta' = \mathcal{M}\theta$, where $\mathcal{M}$ is a block-diagonal orthogonal matrix. Since the loss function is invariant under these internal rotations, the Hessian transforms as $H(\theta') = \mathcal{M} H(\theta) \mathcal{M}^\top$. Because $\mathcal{M}$ is orthogonal, this is a similarity transformation, which implies that the eigenvalues (and thus the spectrum) of the Hessian remain invariant. □

The following are technical lemmas for Theorem 4.1.

Lemma B.2. *Consider*

$$T(Z_1, \dots, Z_D) = \sum_{k=1}^{D} \|Z_D \cdots Z_{k+1}\|_F^2 \left\| Z_{k-1} \cdots Z_1 \Sigma^{1/2} \right\|_F^2, \tag{45}$$

where empty products are identity matrices. Then

$$\min_{Z_D \cdots Z_1 = V} T = D\, \tau^{2(D-1)/D} (d_y X)^{1/D}. \tag{46}$$

Moreover, all minimizers are characterized as follows:

$$\begin{aligned} Z_D &= \tau^{-\frac{D-2}{2D}} d_y^{\frac{D-1}{2D}} X^{-\frac{1}{2D}} \hat{L} \hat{S}^{1/2} U_{D-1}^\top, \\ Z_k &= \tau^{1/D} (d_y X)^{-1/(2D)} U_k U_{k-1}^\top + (I - U_k U_k^\top) Q_k (I - U_{k-1} U_{k-1}^\top), \qquad 2 \le k \le D-1, \\ Z_1 \Sigma^{1/2} &= \tau^{-\frac{D-2}{2D}} d_y^{-\frac{1}{2D}} X^{\frac{D-1}{2D}} U_1 \hat{S}^{1/2} \hat{R}^\top, \end{aligned} \tag{47}$$

where $U_k \in \mathbb{R}^{d_{k+1} \times r}$ are arbitrary matrices satisfying $U_k^\top U_k = I_r$ with $r := \operatorname{rank}(V\Sigma^{1/2})$ and $Q_k \in \mathbb{R}^{d_{k+1} \times d_k}$ are arbitrary.

Proof. Define, for $k = 1, \dots, D-1$,

$$A_k := Z_D \cdots Z_{k+1}, \qquad P_k := Z_k \cdots Z_1 \Sigma^{1/2}, \tag{48}$$

and use the conventions $A_D := I_{d_y}$ and $P_0 := \Sigma^{1/2}$. Then

$$T = \sum_{k=1}^{D} t_k, \qquad t_k := \|A_k\|_F^2 \|P_{k-1}\|_F^2. \tag{49}$$

By the AM–GM inequality,

$$T \ge D \left(\prod_{k=1}^{D} t_k \right)^{1/D}. \tag{50}$$

Using $\|A_D\|_F^2 = d_y$ and $\|P_0\|_F^2 = \operatorname{Tr}\Sigma = X$, we can rearrange the product as

$$\prod_{k=1}^{D} t_k = d_y X \prod_{k=1}^{D-1} \|A_k\|_F^2 \|P_k\|_F^2. \tag{51}$$

For every $k = 1, \dots, D-1$, the constraint $Z_D \cdots Z_1 = V$ implies

$$A_k P_k = V \Sigma^{1/2} = \hat{L} \hat{S} \hat{R}^\top. \tag{52}$$

Therefore,

$$\|A_k\|_F \|P_k\|_F \ge \|A_k P_k\|_* = \|V \Sigma^{1/2}\|_* = \tau, \tag{53}$$

where $\|\cdot\|_*$ denotes the nuclear norm. Consequently,

$$\prod_{k=1}^{D} t_k \ge d_y X\, \tau^{2(D-1)}. \tag{54}$$

Substituting this into (50) gives

$$T \geq D\,\tau^{2(D-1)/D}(d_y X)^{1/D}. \tag{55}$$

We next characterize when equality holds in (53). For any factorization $AB = \hat{L}\hat{S}\hat{R}^\top$, we have

$$\tau = \mathrm{Tr}(\hat{L}^\top AB\hat{R}) = \langle A^\top\hat{L}, B\hat{R}\rangle_F \leq \|A^\top\hat{L}\|_F\|B\hat{R}\|_F \leq \|A\|_F\|B\|_F. \tag{56}$$

Equality holds if and only if there exist a scalar $a > 0$ and a matrix U satisfying $U^\top U = I_r$ such that

$$A = a\hat{L}\hat{S}^{1/2}U^\top, \qquad B = a^{-1}U\hat{S}^{1/2}\hat{R}^\top. \tag{57}$$

This is because equality in the last inequality of (56) requires $A = \hat{L}\hat{L}^\top A$, $B = B\hat{R}\hat{R}^\top$, while equality in the second-to-last inequality of (56) requires $B\hat{R} = \lambda A^\top\hat{L}$ for some $\lambda > 0$. Writing $C := \hat{L}^\top A$, these conditions imply $A = \hat{L}C$, $B = \lambda C^\top\hat{R}^\top$, and the constraint $AB = \hat{L}\hat{S}\hat{R}^\top$ gives $\lambda CC^\top = \hat{S}$. Hence $C = \lambda^{-1/2}\hat{S}^{1/2}U^\top$ for some $U^\top U = I_r$, which gives (57) with $a = \lambda^{-1/2}$.

Thus, if T reaches the lower bound (55), then for every $k = 1, \ldots, D-1$,

$$A_k = a_k\hat{L}\hat{S}^{1/2}U_k^\top, \qquad P_k = a_k^{-1}U_k\hat{S}^{1/2}\hat{R}^\top \tag{58}$$

for some $a_k > 0$ and $U_k^\top U_k = I_r$.

Equality in (50) requires $t_1 = t_2 = \cdots = t_D =: t_\star$. Since their product equals $d_y X\tau^{2(D-1)}$,

$$t_\star = \tau^{2(D-1)/D}(d_y X)^{1/D}. \tag{59}$$

From (58),

$$\|A_k\|_F^2 = a_k^2\tau, \qquad \|P_k\|_F^2 = a_k^{-2}\tau. \tag{60}$$

Therefore,

$$t_1 = a_1^2\tau X, \tag{61}$$

$$t_k = \frac{a_k^2}{a_{k-1}^2}\tau^2, \qquad 2 \leq k \leq D-1, \tag{62}$$

and

$$t_D = d_y\,a_{D-1}^{-2}\tau. \tag{63}$$

Equating these terms with $t_\star$ yields

$$a_k = \tau^{\frac{D-2k}{2D}}\,d_y^{\frac{k}{2D}}\,X^{-\frac{D-k}{2D}}, \qquad k = 1, \ldots, D-1. \tag{64}$$

In particular,

$$\frac{a_{k-1}}{a_k} = \tau^{1/D}(d_y X)^{-1/(2D)} =: c, \tag{65}$$

which is independent of k.

Since $A_{D-1} = Z_D$, Eqs. (58) and (64) immediately give

$$Z_D = \tau^{-\frac{D-2}{2D}}\,d_y^{\frac{D-1}{2D}}\,X^{-\frac{1}{2D}}\,\hat{L}\hat{S}^{1/2}U_{D-1}^\top. \tag{66}$$

Similarly, since $P_1 = Z_1\Sigma^{1/2}$,

$$Z_1\Sigma^{1/2} = \tau^{-\frac{D-2}{2D}}\,d_y^{-\frac{1}{2D}}\,X^{\frac{D-1}{2D}}\,U_1\hat{S}^{1/2}\hat{R}^\top. \tag{67}$$

For an intermediate layer $2 \leq k \leq D-1$, the relations

$$P_k = Z_k P_{k-1}, \qquad A_{k-1} = A_k Z_k$$

together with (58) imply

$$Z_k U_{k-1} = cU_k, \qquad U_k^\top Z_k = cU_{k-1}^\top. \tag{68}$$

Let $\Pi_k := U_k U_k^\top$. Decomposing Z_k gives

$$Z_k = \Pi_k Z_k \Pi_{k-1} + \Pi_k Z_k(I - \Pi_{k-1}) + (I - \Pi_k)Z_k\Pi_{k-1} + (I - \Pi_k)Z_k(I - \Pi_{k-1}). \tag{69}$$

The two relations in (68) imply

$$\Pi_k Z_k \Pi_{k-1} = c U_k U_{k-1}^\top, \tag{70}$$

and

$$\Pi_k Z_k (I - \Pi_{k-1}) = 0, \qquad (I - \Pi_k) Z_k \Pi_{k-1} = 0. \tag{71}$$

Hence the most general possible intermediate layer is

$$Z_k = c U_k U_{k-1}^\top + (I - U_k U_k^\top) Q_k (I - U_{k-1} U_{k-1}^\top), \tag{72}$$

where Q_k is arbitrary. Using (65) gives exactly (47).

Finally, take arbitrary $U_k^\top U_k = I_r$ and arbitrary Q_k, and define $Z_1, \dots, Z_D$ by (47). Because

$$(I - U_k U_k^\top) U_k = 0, \qquad U_k^\top (I - U_k U_k^\top) = 0, \tag{73}$$

all Q_k-dependent terms vanish. Therefore,

$$Z_k U_{k-1} = c U_k, \qquad U_k^\top Z_k = c U_{k-1}^\top. \tag{74}$$

It follows by induction that the partial products satisfy (58) with the scales (64). In particular,

$$Z_D \cdots Z_1 \Sigma^{1/2} = \hat{L} \hat{S} \hat{R}^\top = V \Sigma^{1/2}. \tag{75}$$

Since $\Sigma \succ 0$, $Z_D \cdots Z_1 = V$. Moreover, every inequality used in (50)–(53) is saturated. Hence $T = D\,\tau^{2(D-1)/D} (d_y X)^{1/D}$. Thus every factorization in (47) is a minimizer, and the necessity argument above shows that every minimizer has this form. □

The following lemma is a restatement of Theorem 12 of Ziyin et al. (2025).

Lemma B.3. *Consider a $D-$layer linear network which has more than d hidden units in every layer. Then at the global minimum of* (5)*, we have*

$$\sqrt{\Sigma_\epsilon} Z_D = L \Sigma_D U_{D-1}^\top,\ Z_i = U_i \Sigma_i U_{i-1}^\top,\ Z_1 \sqrt{\Sigma_x} = U_1 \Sigma_1 R^\top, \tag{76}$$

for $i = 2, \cdots, D-1$, where $\{U_i\}_{i=1}^{D-1}$ are arbitrary matrices satisfying $U_i^\top U_i = I_{d \times d}$, and $\Sigma_x = \mathbb{E}[x x^\top]$, $\Sigma_\epsilon = \mathbb{E}[\epsilon \epsilon^\top]$. Moreover,

$$\begin{aligned}
\Sigma_1 &= (\mathrm{Tr} S)^{-\frac{D-2}{2D}} \frac{\mathrm{Tr}[\Sigma_x]^{\frac{D-1}{2D}}}{\mathrm{Tr}[\Sigma_\epsilon]^{\frac{1}{2D}}} \sqrt{S}, \\
\Sigma_D &= (\mathrm{Tr} S)^{-\frac{D-2}{2D}} \frac{\mathrm{Tr}[\Sigma_\epsilon]^{\frac{D-1}{2D}}}{\mathrm{Tr}[\Sigma_x]^{\frac{1}{2D}}} \sqrt{S}, \\
\Sigma_i &= (\mathrm{Tr} S)^{1/D} (\mathrm{Tr}[\Sigma_\epsilon] \mathrm{Tr}[\Sigma_x])^{-\frac{1}{2D}} I_d.
\end{aligned} \tag{77}$$

B.4 Initial sharpness

To understand the optimization landscape at the start of training, we analyze the expected sharpness under standard random initialization schemes.

Proposition B.4. *Consider a D-layer linear network where each weight matrix $W_k \in \mathbb{R}^{d_{k+1} \times d_k}$ is initialized with i.i.d. entries $w_{i,j}^{(k)} \sim \mathcal{N}(0, \sigma_k^2)$, where $d_1 = d_x$ and $d_{D+1} = d_y$. For simplicity we assume $R_k = 0$. At initialization, the expected sharpness is given by:*

$$\mathbb{E}[T(\theta)] = d_y \Pi_{k=2}^D d_k \mathrm{Tr}[\Sigma_x] \sum_{k=1}^D \Pi_{j \neq k} \sigma_j^2. \tag{78}$$

If we further assume $\sigma_k^2 = \sigma^2$ and $d_k = d$ for all k, this simplifies to:

$$\mathbb{E}[T(\theta)] = D \cdot d_y \mathrm{Tr}[\Sigma_x] (d \sigma^2)^{D-1}. \tag{79}$$

Proof. Recall that $T(\theta) = \sum_{k=1}^{D} \mathrm{Tr}[A_k^\top A_k]\mathrm{Tr}[B_k \Sigma_x B_k^\top]$. Since the weight matrices $\{W_k\}$ are initialized independently, the expectation of the trace product decomposes into the product of expectations. For A_k, we have

$$\mathbb{E}[\mathrm{Tr}[A_k^\top A_k]] = \mathbb{E}[\mathrm{Tr}[W_{k+1}^\top \cdots W_D W_D^\top \cdots W_{k+1}]] = d_y \Pi_{j=k+1}^{D} d_j \sigma_j^2. \tag{80}$$

Similarly, for B_k:

$$\mathbb{E}[\mathrm{Tr}[B_k \Sigma_x B_k^\top]] = \mathbb{E}[\mathrm{Tr}[W_{k-1} \cdots W_1 \Sigma_x W_1^\top \cdots W_{k-1}]] = \mathrm{Tr}[\Sigma_x] \Pi_{j=1}^{k-1} d_{j+1} \sigma_j^2. \tag{81}$$

Substituting these into the total sharpness expression (39) yields (78). □

For the Xavier initialization, we have $\sigma_k^2 = \frac{2}{d_k + d_{k+1}}$. Taking $d_k = d$ for all k, we have $\mathbb{E}[T(\theta)] = D d_y \mathrm{Tr}[\Sigma_x]$. For the Kaiming initialization with linear activation, we have $\sigma^2 = \frac{1}{d_k}$. Then the sharpness is the same if $d_k = d$ for any k.

B.5 Hessian spectrum

The following theorem characterizes the Hessian at the optimal solutions.

Theorem B.5. *Assume that $VV^\top$ commutes with Σ_ϵ and $V^\top V$ commutes with Σ_x, i.e., $\Sigma_\epsilon = L\Lambda_\epsilon L^\top$, $\Sigma_x = R\Lambda_x R^\top$ and $\sqrt{\Sigma_\epsilon} V \sqrt{\Sigma_x} = LSR^\top$, where $\Lambda_\epsilon = \mathrm{diag}(e_1, \ldots, e_d)$, $\Lambda_x = \mathrm{diag}(x_1, \ldots, x_d)$ and $S = \mathrm{diag}(s_1, \ldots, s_d)$. Define*

$$\begin{aligned} a &= (\mathrm{Tr}S)^{-2/D} (\mathrm{Tr}\Sigma_\epsilon \mathrm{Tr}\Sigma_x)^{1/D}, \\ b &= (\mathrm{Tr}S)^{(D-2)/D} (\mathrm{Tr}\Sigma_\epsilon)^{1/D} (\mathrm{Tr}\Sigma_x)^{-(D-1)/D}, \\ c &= (\mathrm{Tr}S)^{(D-2)/D} (\mathrm{Tr}\Sigma_\epsilon)^{-(D-1)/D} (\mathrm{Tr}\Sigma_x)^{1/D}. \end{aligned} \tag{82}$$

At any solution characterized by (77)*, the nonzero eigenvalues of the Hessian are*

$$h_{ij} = (D-2) a \frac{s_i}{e_i} s_j + b \frac{s_i}{e_i} x_j + c s_j. \tag{83}$$

Consequently, we have

$$\lambda_{\max}(H) = \max_j \left\{ [(D-2) a q_\star + c] s_j + b q_\star x_j \right\}, \tag{84}$$

where $q_\star = \max_i s_i / e_i$.

Proof. Let $A_k := Z_D \cdots Z_{k+1}$ and $B_k := Z_{k-1} \cdots Z_1$ with the conventions $A_D = I$ and $B_1 = I$, and define $C_k := B_k \sqrt{\Sigma_x}$.

For a perturbation $\Delta = (\Delta_1, \ldots, \Delta_D)$, its quadratic form is

$$\Delta^\top H \Delta = \mathbb{E}_x \left\| \sum_{k=1}^{D} A_k \Delta_k B_k x \right\|_2^2 = \left\| \sum_{k=1}^{D} A_k \Delta_k C_k \right\|_F^2. \tag{85}$$

Define the linear operator $\mathcal{J}(\Delta) := \sum_{k=1}^{D} A_k \Delta_k C_k$. Then $H = \mathcal{J}^* \mathcal{J}$. Consequently, the nonzero eigenvalues of H coincide with those of $K := \mathcal{J}\mathcal{J}^*$. Using

$$\mathrm{vec}(A_k \Delta_k C_k) = (C_k^\top \otimes A_k) \, \mathrm{vec}(\Delta_k), \tag{86}$$

we obtain

$$K = \sum_{k=1}^{D} C_k^\top C_k \otimes A_k A_k^\top. \tag{87}$$

We now evaluate the individual terms in (87) at the solutions characterized by (77). Let $\tau := \mathrm{Tr}S$, $E := \mathrm{Tr}\Sigma_\epsilon$ and $X := \mathrm{Tr}\Sigma_x$, and introduce

$$\begin{aligned} \alpha &= \tau^{-\frac{D-2}{2D}} X^{\frac{D-1}{2D}} E^{-\frac{1}{2D}}, \\ \beta &= \tau^{-\frac{D-2}{2D}} E^{\frac{D-1}{2D}} X^{-\frac{1}{2D}}, \\ \gamma &= \tau^{1/D} (EX)^{-1/(2D)}. \end{aligned} \tag{88}$$

Then (77) can be written as

$$\Sigma_1 = \alpha S^{1/2}, \qquad \Sigma_D = \beta S^{1/2}, \qquad \Sigma_k = \gamma I, \quad 2 \le k \le D-1. \tag{89}$$

Under the simultaneous diagonalization assumption,

$$Z_D = L\Lambda_\epsilon^{-1/2}\Sigma_D U_{D-1}^\top, \qquad Z_1\sqrt{\Sigma_x} = U_1\Sigma_1 R^\top. \tag{90}$$

For an intermediate layer $2 \le k \le D-1$, we have $A_k = L\Lambda_\epsilon^{-1/2}\beta S^{1/2}\gamma^{D-k-1}U_k^\top$ and $C_k = U_{k-1}\gamma^{k-2}\alpha S^{1/2}R^\top$. Therefore, $A_kA_k^\top = \beta^2\gamma^{2(D-k-1)}L\Lambda_\epsilon^{-1}SL^\top$, and $C_k^\top C_k = \alpha^2\gamma^{2(k-2)}RSR^\top$ with $\alpha^2\beta^2\gamma^{2(D-3)} = \tau^{-2/D}(EX)^{1/D} = a$. Thus every intermediate layer contributes

$$a\,RSR^\top \otimes L\Lambda_\epsilon^{-1}SL^\top.$$

For the first layer, $C_1 = \sqrt{\Sigma_x}$, so $C_1^\top C_1 = \Sigma_x = R\Lambda_x R^\top$. Moreover, $A_1A_1^\top = \beta^2\gamma^{2(D-2)}L\Lambda_\epsilon^{-1}SL^\top$, where $\beta^2\gamma^{2(D-2)} = \tau^{(D-2)/D}E^{1/D}X^{-(D-1)/D} = b$. Hence the first-layer contribution is

$$b\,R\Lambda_x R^\top \otimes L\Lambda_\epsilon^{-1}SL^\top.$$

For the final layer, $A_D = I$, while $C_D = U_{D-1}\gamma^{D-2}\alpha S^{1/2}R^\top$. Thus $C_D^\top C_D = \alpha^2\gamma^{2(D-2)}RSR^\top$, and $\alpha^2\gamma^{2(D-2)} = \tau^{(D-2)/D}E^{-(D-1)/D}X^{1/D} = c$. The final-layer contribution is therefore

$$c\,RSR^\top \otimes I.$$

Combining the contributions of all layers gives

$$K = (D-2)a\,RSR^\top \otimes L\Lambda_\epsilon^{-1}SL^\top + b\,R\Lambda_x R^\top \otimes L\Lambda_\epsilon^{-1}SL^\top + c\,RSR^\top \otimes I. \tag{91}$$

All three terms in (91) are simultaneously diagonalized by the orthogonal basis $R \otimes L$:

$$(R\otimes L)^\top K(R\otimes L) = (D-2)a\,S\otimes\Lambda_\epsilon^{-1}S + b\,\Lambda_x\otimes\Lambda_\epsilon^{-1}S + c\,S\otimes I. \tag{92}$$

Since $S = \mathrm{diag}(s_1,\dots,s_d)$, $\Lambda_\epsilon^{-1}S = \mathrm{diag}\left(\frac{s_1}{e_1},\dots,\frac{s_d}{e_d}\right)$, and $\Lambda_x = \mathrm{diag}(x_1,\dots,x_d)$, we obtain all the non-zero eigenvalues of the Hessian:

$$h_{ij} = (D-2)a\frac{s_i}{e_i}s_j + b\frac{s_i}{e_i}x_j + cs_j. \tag{93}$$

Finally, for any fixed j,

$$h_{ij} = \frac{s_i}{e_i}\left[(D-2)as_j + bx_j\right] + cs_j. \tag{94}$$

The coefficient of s_i/e_i is nonnegative, so the maximizing index i satisfies $\frac{s_i}{e_i} = q_\star := \max_i \frac{s_i}{e_i}$. Therefore,

$$\lambda_{\max}(H) = \max_j\left\{[(D-2)aq_\star + c]s_j + bq_\star x_j\right\}, \tag{95}$$

which proves the result. □

Remark B.6. (85) also gives a general expression for the spectral sharpness. Let $C_k := B_k\sqrt{\Sigma_x}$, and define $\mathcal{J} := \begin{bmatrix}C_1^\top\otimes A_1 & \cdots & C_D^\top\otimes A_D\end{bmatrix}$. For $\delta = \begin{bmatrix}\mathrm{vec}(\Delta_1)\\ \vdots \\ \mathrm{vec}(\Delta_D)\end{bmatrix}$, we have

$$\Delta^\top H(\theta)\Delta = \|\mathcal{J}\delta\|_2^2. \tag{96}$$

Hence $H = \mathcal{J}^\top\mathcal{J}$, and therefore

$$\lambda_{\max}(H) = \sigma_{\max}(\mathcal{J})^2 = \lambda_{\max}\left(\sum_{k=1}^{D} C_k^\top C_k \otimes A_kA_k^\top\right). \tag{97}$$

This recovers the general spectral characterization of Kamber & Parhi (2025).

In general, however, the positive-semidefinite matrices appearing in the different Kronecker terms of (97) need not be simultaneously diagonalizable. Consequently, even after substituting the minimum-fluctuation solution (77), the largest eigenvalue does not admit a simple expression. The commutation assumptions in Theorem B.5 ensure that the relevant matrices share common eigenbases L and R, respectively. As a result, all terms in (97) are simultaneously diagonalized by $R \otimes L$, yielding the explicit Hessian spectrum stated in the theorem.

Remark B.7. As a sanity check, summing the eigenvalues in Theorem B.5 recovers the sharpness in Theorem 4.1. Indeed,

$$\mathrm{Tr}H = \sum_{i,j} h_{ij} = (D-2)a\left(\sum_i \frac{s_i}{e_i}\right)\left(\sum_j s_j\right) + b\left(\sum_i \frac{s_i}{e_i}\right)\left(\sum_j x_j\right) + c\sum_{i,j} s_j. \tag{98}$$

Using $\sum_j s_j = \mathrm{Tr}S$, $\sum_j x_j = \mathrm{Tr}\Sigma_x$, $\sum_i \frac{s_i}{e_i} = \mathrm{Tr}\big[\Sigma_\epsilon^{-1} LSL^\top\big]$, and $\sum_i 1 = d_y$, the first two terms combine because

$$a\,\mathrm{Tr}S = b\,\mathrm{Tr}\Sigma_x = (\mathrm{Tr}S)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon \mathrm{Tr}\Sigma_x)^{1/D}. \tag{99}$$

Therefore,

$$\begin{aligned}\mathrm{Tr}H = (D-1)(\mathrm{Tr}S)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon\mathrm{Tr}\Sigma_x)^{1/D}\mathrm{Tr}\big[\Sigma_\epsilon^{-1}LSL^\top\big] \\ + d_y(\mathrm{Tr}S)^{2(D-1)/D}\frac{(\mathrm{Tr}\Sigma_x)^{1/D}}{(\mathrm{Tr}\Sigma_\epsilon)^{(D-1)/D}},\end{aligned} \tag{100}$$

which is precisely (22).

B.6 Proof of Corollary 5.1

Proof. Let $A := V\sqrt{\Sigma_x}$ and denote its smallest and largest singular values by $m := \sigma_{\min}(A)$ and $M := \sigma_{\max}(A)$. The assumption that A has full row rank implies $m > 0$.

Recall that $\sqrt{\Sigma_\epsilon}A = LSR^\top$. Therefore,

$$P := LSL^\top = \left(\sqrt{\Sigma_\epsilon}AA^\top\sqrt{\Sigma_\epsilon}\right)^{1/2}. \tag{101}$$

Since $m^2 I \preceq AA^\top \preceq M^2 I$, multiplying on both sides by $\sqrt{\Sigma_\epsilon}$ gives

$$m^2\Sigma_\epsilon \preceq \sqrt{\Sigma_\epsilon}AA^\top\sqrt{\Sigma_\epsilon} \preceq M^2\Sigma_\epsilon. \tag{102}$$

The matrix square-root function is operator monotone on the positive-semidefinite cone, and hence

$$m\sqrt{\Sigma_\epsilon} \preceq P \preceq M\sqrt{\Sigma_\epsilon}. \tag{103}$$

Taking traces in (103), and using $\mathrm{Tr}P = \mathrm{Tr}S$, yields

$$m\mathrm{Tr}\sqrt{\Sigma_\epsilon} \le \mathrm{Tr}S \le M\mathrm{Tr}\sqrt{\Sigma_\epsilon}. \tag{104}$$

Similarly, multiplying (103) by the positive-definite matrix Σ_ϵ^{-1} inside the trace gives

$$m\mathrm{Tr}\Sigma_\epsilon^{-1/2} \le \mathrm{Tr}[\Sigma_\epsilon^{-1}LSL^\top] \le M\mathrm{Tr}\Sigma_\epsilon^{-1/2}. \tag{105}$$

We next use the fact that the sharpness in Theorem 4.1 is invariant under an overall rescaling of the label-noise covariance. We may therefore normalize, without loss of generality, $\lambda_{\max}(\Sigma_\epsilon) = 1$ and $\lambda_{\min}(\Sigma_\epsilon) = \kappa(\Sigma_\epsilon)^{-1}$. Writing $\kappa := \kappa(\Sigma_\epsilon)$ for brevity, we consequently have

$$1 \le \mathrm{Tr}\sqrt{\Sigma_\epsilon} \le d_y, \qquad 1 \le \mathrm{Tr}\Sigma_\epsilon \le d_y, \tag{106}$$

and

$$\sqrt{\kappa} \le \mathrm{Tr}\Sigma_\epsilon^{-1/2} \le d_y\sqrt{\kappa}. \tag{107}$$

Combining (104)–(107), there exist positive constants c, C, depending only on A and d_y, such that

$$c \le \mathrm{Tr}S \le C \tag{108}$$

and

$$c\sqrt{\kappa} \le \mathrm{Tr}[\Sigma_\epsilon^{-1}LSL^\top] \le C\sqrt{\kappa}. \tag{109}$$

We now substitute these bounds into the sharpness formula

$$\begin{aligned}T(\theta) = (D-1)(\mathrm{Tr}S)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon\mathrm{Tr}\Sigma_x)^{1/D}\mathrm{Tr}[\Sigma_\epsilon^{-1}LSL^\top] \\ + d_y(\mathrm{Tr}S)^{2(D-1)/D}\frac{(\mathrm{Tr}\Sigma_x)^{1/D}}{(\mathrm{Tr}\Sigma_\epsilon)^{(D-1)/D}}.\end{aligned} \tag{110}$$

Since $\mathrm{Tr}S$ and $\mathrm{Tr}\Sigma_\epsilon$ are bounded above and below by positive constants independent of κ, and Σ_x is fixed, the first term satisfies

$$c'(D-1)\sqrt{\kappa} \le T_1 \le C'(D-1)\sqrt{\kappa} \tag{111}$$

for constants $c', C' > 0$ independent of D and Σ_ϵ. Here we have used the fact that all exponents appearing in (110) lie in bounded intervals, so the corresponding factors remain uniformly bounded above and below.

The second term is bounded independently of κ:

$$0 < T_2 \le C'', \tag{112}$$

for a constant C'' depending only on A, Σ_x, and d_y. Since $\kappa \ge 1$ and $D \ge 2$, $T_2 \le C'' \le \frac{C''}{2} D\sqrt{\kappa}$. Together with $D - 1 \ge D/2$, (111) therefore gives

$$\frac{c'}{2} D\sqrt{\kappa} \le T(\theta) \le \left(C' + \frac{C''}{2}\right) D\sqrt{\kappa}. \tag{113}$$

Absorbing the fixed constants into c_1 and c_2 finishes the proof. □

B.7 Proof of Corollary 5.2

Proof. Let $A := V\sqrt{\Sigma_x}$ and $A = \hat{L}\hat{S}\hat{R}^\top$ be the singular value decomposition of A. By assumption, A has full row rank, and hence $C := AA^\top \succ 0$.

We first prove the "if" direction. Suppose that $\Sigma_\epsilon = cI$ for some $c > 0$. Then $\sqrt{\Sigma_\epsilon} A = \sqrt{c}\, A = \hat{L}(\sqrt{c}\,\hat{S})\hat{R}^\top$. Hence, in the notation of Lemma B.3, $L = \hat{L}$, $R = \hat{R}$ and $S = \sqrt{c}\,\hat{S}$. Substituting these relations, together with $\mathrm{Tr}\Sigma_\epsilon = cd_y$, into Eq. (77), we obtain

$$\begin{aligned} Z_D &= (\mathrm{Tr}\hat{S})^{-\frac{D-2}{2D}} d_y^{\frac{D-1}{2D}} (\mathrm{Tr}\Sigma_x)^{-\frac{1}{2D}} \hat{L}\hat{S}^{1/2} U_{D-1}^\top, \\ Z_k &= (\mathrm{Tr}\hat{S})^{1/D} (d_y \mathrm{Tr}\Sigma_x)^{-1/(2D)} U_k U_{k-1}^\top, \qquad 2 \le k \le D-1, \\ Z_1\sqrt{\Sigma_x} &= (\mathrm{Tr}\hat{S})^{-\frac{D-2}{2D}} d_y^{-\frac{1}{2D}} (\mathrm{Tr}\Sigma_x)^{\frac{D-1}{2D}} U_1 \hat{S}^{1/2}\hat{R}^\top. \end{aligned} \tag{114}$$

This is a special case of the minimum-sharpness solution in Lemma B.2, corresponding to $Q_k = 0$ for all k. Therefore every minimum-fluctuation solution characterized by Eq. (77) attains the minimum possible sharpness.

We next prove the converse. Suppose that a minimum-fluctuation solution characterized by Eq. (77) also attains the minimum sharpness.

By the necessity part of Lemma B.2, every minimum-sharpness solution must satisfy

$$Z_D = \alpha\, \hat{L}\hat{S}^{1/2} U_{D-1}^\top \tag{115}$$

for some scalar $\alpha > 0$. Consequently,

$$Z_D Z_D^\top = \alpha^2 \hat{L}\hat{S}\hat{L}^\top = \alpha^2 (AA^\top)^{1/2} = \alpha^2 C^{1/2}. \tag{116}$$

On the other hand, let $\sqrt{\Sigma_\epsilon} A = LSR^\top$. Equation (77) implies

$$\sqrt{\Sigma_\epsilon} Z_D = \beta L S^{1/2} U_{D-1}^\top \tag{117}$$

for some scalar $\beta > 0$. Therefore,

$$Z_D Z_D^\top = \beta^2 \Sigma_\epsilon^{-1/2} L S L^\top \Sigma_\epsilon^{-1/2}. \tag{118}$$

Since

$$LS^2 L^\top = \Sigma_\epsilon^{1/2} A A^\top \Sigma_\epsilon^{1/2} = \Sigma_\epsilon^{1/2} C \Sigma_\epsilon^{1/2}, \tag{119}$$

we have

$$LSL^\top = \left(\Sigma_\epsilon^{1/2} C \Sigma_\epsilon^{1/2}\right)^{1/2}. \tag{120}$$

Hence

$$Z_D Z_D^\top = \beta^2 \Sigma_\epsilon^{-1/2} \left(\Sigma_\epsilon^{1/2} C \Sigma_\epsilon^{1/2}\right)^{1/2} \Sigma_\epsilon^{-1/2}. \tag{121}$$

Because the same solution also minimizes the sharpness, Eqs. (116) and (121) must agree. Thus there exists a scalar $\rho > 0$ such that

$$\Sigma_\epsilon^{-1/2} \left(\Sigma_\epsilon^{1/2} C \Sigma_\epsilon^{1/2}\right)^{1/2} \Sigma_\epsilon^{-1/2} = \rho C^{1/2}. \tag{122}$$

Multiplying from the left and right by $\Sigma_\epsilon^{1/2}$ gives $\left(\Sigma_\epsilon^{1/2} C \Sigma_\epsilon^{1/2}\right)^{1/2} = \rho \Sigma_\epsilon^{1/2} C^{1/2} \Sigma_\epsilon^{1/2}$. Squaring both sides yields $\Sigma_\epsilon^{1/2} C \Sigma_\epsilon^{1/2} = \rho^2 \Sigma_\epsilon^{1/2} C^{1/2} \Sigma_\epsilon C^{1/2} \Sigma_\epsilon^{1/2}$. Since $\Sigma_\epsilon \succ 0$, multiplying from the left and right by $\Sigma_\epsilon^{-1/2}$ gives

$$C = \rho^2 C^{1/2} \Sigma_\epsilon C^{1/2}. \tag{123}$$

Because A has full row rank, $C \succ 0$, and therefore $C^{-1/2}$ exists. Multiplying from the left and right by $C^{-1/2}$ gives

$$\Sigma_\epsilon = \rho^{-2} I, \tag{124}$$

so the label-noise covariance is isotropic.

Therefore, the minimum-fluctuation solutions characterized by Eq. (77) attain the minimum possible sharpness if and only if $\Sigma_\epsilon = cI$ for some $c > 0$. □

B.8 Proof of Corollary 5.3

Proof. Under the simultaneous diagonalization assumption, write $\Sigma_\epsilon = L \Lambda_\epsilon L^\top$ and $\Sigma_x = R \Lambda_x R^\top$, where $\Lambda_\epsilon = \mathrm{diag}(e_1, \ldots, e_d)$ and $\Lambda_x = \mathrm{diag}(x_1, \ldots, x_d)$. Since V has full rank and is simultaneously diagonalizable with these matrices, there exist nonzero scalars v_i such that $V = L \Lambda_V R^\top$ with $\Lambda_V = \mathrm{diag}(v_1, \ldots, v_d)$. Consequently, $\sqrt{\Sigma_\epsilon} V \sqrt{\Sigma_x} = L S R^\top$, with

$$s_i = |v_i| \sqrt{e_i x_i}. \tag{125}$$

It follows that

$$\frac{s_i}{e_i} = |v_i| \sqrt{\frac{x_i}{e_i}}. \tag{126}$$

Recall from Theorem B.5 that

$$\lambda_{\max}(H) = \max_j \left\{ [(D-2) a q_\star + c] s_j + b q_\star x_j \right\}, \tag{127}$$

where $q_\star := \max_i \frac{s_i}{e_i}$.

We first note that the expression in (127) is invariant under an overall rescaling $\Sigma_\epsilon \mapsto \alpha \Sigma_\epsilon$.. Therefore, without loss of generality, we normalize

$$\lambda_{\max}(\Sigma_\epsilon) = 1. \tag{128}$$

Writing $\kappa := \kappa(\Sigma_\epsilon)$, we then have $\min_i e_i = \kappa^{-1}$ and $\max_i e_i = 1$.

Since V is full rank and $\Sigma_x \succ 0$, define positive constants

$$v_{\min} := \min_i |v_i| > 0, \qquad v_{\max} := \max_i |v_i|, \tag{129}$$

and

$$x_{\min} := \min_i x_i > 0, \qquad x_{\max} := \max_i x_i. \tag{130}$$

(126) then implies

$$v_{\min} \sqrt{x_{\min}} \sqrt{\kappa} \le q_\star \le v_{\max} \sqrt{x_{\max}} \sqrt{\kappa}. \tag{131}$$

Hence $q_\star = \Theta(\sqrt{\kappa})$.

We next bound the remaining quantities independently of κ. Since at least one e_i equals 1, Eq. (125) implies

$$v_{\min} \sqrt{x_{\min}} \le \max_i s_i \le v_{\max} \sqrt{x_{\max}}. \tag{132}$$

Moreover,

$$1 \le \mathrm{Tr} \Sigma_\epsilon \le d, \; v_{\min} \sqrt{x_{\min}} \le \mathrm{Tr} S \le d \, v_{\max} \sqrt{x_{\max}}. \tag{133}$$

Since $\mathrm{Tr}\Sigma_x$ is fixed, (133) imply that the coefficients $a = (\mathrm{Tr}S)^{-2/D}(\mathrm{Tr}\Sigma_\epsilon \mathrm{Tr}\Sigma_x)^{1/D}$, $b = (\mathrm{Tr}S)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon)^{1/D}(\mathrm{Tr}\Sigma_x)^{-(D-1)/D}$ and $c = (\mathrm{Tr}S)^{(D-2)/D}(\mathrm{Tr}\Sigma_\epsilon)^{-(D-1)/D}(\mathrm{Tr}\Sigma_x)^{1/D}$ are uniformly bounded above and below by positive constants independent of κ and $D \geq 2$. Thus, for suitable constants

$$0 < a_- \leq a \leq a_+, \qquad 0 < b_- \leq b \leq b_+, \qquad 0 < c_- \leq c \leq c_+. \tag{134}$$

We first establish the upper bound. From (127), (131), and the bounds on s_j and x_j,

$$\begin{aligned} \lambda_{\max}(H) &\leq [(D-2)a_+ q_\star + c_+] \max_j s_j + b_+ q_\star x_{\max} \\ &\leq C_1(D-2)\sqrt{\kappa} + C_2 + C_3\sqrt{\kappa} \end{aligned} \tag{135}$$

for constants $C_1, C_2, C_3 > 0$ independent of D and Σ_ϵ. Because $D \geq 2$ and $\kappa \geq 1$, this gives

$$\lambda_{\max}(H) \leq c_2 D\sqrt{\kappa}. \tag{136}$$

For the lower bound, let j_0 be an index such that $e_{j_0} = 1$, which exists by the normalization (128). Then

$$s_{j_0} = |v_{j_0}|\sqrt{x_{j_0}} \geq v_{\min}\sqrt{x_{\min}}, \tag{137}$$

and $x_{j_0} \geq x_{\min}$. Evaluating (127) at $j = j_0$ yields

$$\begin{aligned} \lambda_{\max}(H) &\geq (D-2)aq_\star s_{j_0} + bq_\star x_{j_0} \\ &\geq q_\star \left[(D-2)a_- v_{\min}\sqrt{x_{\min}} + b_- x_{\min}\right]. \end{aligned} \tag{138}$$

Using the lower bound in (131), we obtain

$$\lambda_{\max}(H) \geq C\big[(D-2)+1\big]\sqrt{\kappa} \tag{139}$$

for some $C > 0$ depending only on V and Σ_x. Since $D - 1 \geq D/2 \qquad (D \geq 2)$, there exists $c_1 > 0$ such that

$$\lambda_{\max}(H) \geq c_1 D\sqrt{\kappa}. \tag{140}$$

Combining (136) and (140) proves

$$c_1 D\sqrt{\kappa(\Sigma_\epsilon)} \leq \lambda_{\max}(H) \leq c_2 D\sqrt{\kappa(\Sigma_\epsilon)}. \tag{141}$$

□

B.9 The influence of feature noise

Corollary B.8. *Training on (x', y) is equivalent to a linear teacher problem*

$$y = V_{\text{eff}} x' + \epsilon_{\text{eff}}, \tag{142}$$

where ϵ_{eff} is independent of x', and

$$\Sigma_{x'} = \Sigma_x + \Sigma_{\epsilon_x}, \tag{143}$$

$$V_{\text{eff}} = V\Sigma_x(\Sigma_x + \Sigma_{\epsilon_x})^{-1}, \tag{144}$$

and

$$\Sigma_{\epsilon,\text{eff}} = \Sigma_\epsilon + V\left[\Sigma_x - \Sigma_x(\Sigma_x + \Sigma_{\epsilon_x})^{-1}\Sigma_x\right]V^\top. \tag{145}$$

Consequently, the sharpness of the minimum-fluctuation solution (77) *is obtained from Theorem 4.1 by the substitutions $(\Sigma_x, V, \Sigma_\epsilon) \longrightarrow (\Sigma_{x'}, V_{\text{eff}}, \Sigma_{\epsilon,\text{eff}})$.*

Proof. Since $x' = x + \epsilon_x$, we have

$$\Sigma_{x'} = \mathbb{E}[x'x'^\top] = \Sigma_x + \Sigma_{\epsilon_x}. \tag{146}$$

Moreover, $\mathbb{E}[yx'^\top] = V\mathbb{E}[xx^\top] = V\Sigma_x$, because x, ϵ_x, ϵ are mutually independent and zero-mean. Since x and x' are jointly Gaussian, $\mathbb{E}[x \mid x'] = \Sigma_x(\Sigma_x + \Sigma_{\epsilon_x})^{-1}x'$. Therefore,

$$\mathbb{E}[y \mid x'] = V\Sigma_x(\Sigma_x + \Sigma_{\epsilon_x})^{-1}x' = V_{\text{eff}}x', \tag{147}$$

where $V_{\text{eff}} = V\Sigma_x(\Sigma_x+\Sigma_{\epsilon_x})^{-1}$. Define the effective residual $\epsilon_{\text{eff}} := y - V_{\text{eff}}x'$. Its cross-covariance with x' is

$$\mathbb{E}[\epsilon_{\text{eff}}x'^{\top}] = \mathbb{E}[yx'^{\top}] - V_{\text{eff}}\mathbb{E}[x'x'^{\top}] \tag{148}$$

$$= V\Sigma_x - V\Sigma_x(\Sigma_x + \Sigma_{\epsilon_x})^{-1}(\Sigma_x + \Sigma_{\epsilon_x}) \tag{149}$$

$$= 0. \tag{150}$$

Because all variables are jointly Gaussian, zero cross-covariance implies that ϵ_{eff} and x' are independent. The covariance of the effective residual is

$$\Sigma_{\epsilon,\text{eff}} = \text{Cov}(y) - \text{Cov}(y,x')\Sigma_{x'}^{-1}\,\text{Cov}(x',y) \tag{151}$$

$$= \Sigma_\epsilon + V\Sigma_x V^{\top} - V\Sigma_x(\Sigma_x + \Sigma_{\epsilon_x})^{-1}\Sigma_x V^{\top} \tag{152}$$

$$= \Sigma_\epsilon + V\left[\Sigma_x - \Sigma_x(\Sigma_x + \Sigma_{\epsilon_x})^{-1}\Sigma_x\right]V^{\top}, \tag{153}$$

which proves Eqs. (143)–(145). Thus, after replacing the original input covariance, teacher matrix, and label-noise covariance by $\Sigma_{x'}, V_{\text{eff}}, \Sigma_{\epsilon,\text{eff}}$, the noisy-input problem satisfies exactly the same assumptions as the original model. Theorem 4.1 therefore applies directly. □

To see explicitly why the effect of feature noise need not have a fixed sign, consider the simultaneously diagonalizable case

$$\begin{aligned}\Sigma_x &= R\ \text{diag}(x_1,\dots,x_d)R^{\top},\\ \Sigma_{\epsilon_x} &= R\ \text{diag}(r_1,\dots,r_d)R^{\top},\\ \Sigma_\epsilon &= L\ \text{diag}(e_1,\dots,e_d)L^{\top},\\ V &= L\ \text{diag}(v_1,\dots,v_d)R^{\top}.\end{aligned} \tag{154}$$

Then each mode transforms independently as

$$x_i' = x_i + r_i, \tag{155}$$

$$v_i^{\text{eff}} = v_i\frac{x_i}{x_i + r_i}, \tag{156}$$

and

$$e_i^{\text{eff}} = e_i + v_i^2\frac{x_i r_i}{x_i + r_i}. \tag{157}$$

Therefore, feature noise simultaneously attenuates the effective teacher (flattening) and adds an additional, generally anisotropic contribution to the effective label noise (sharpening).

For example, take $\Sigma_x = I, V = I, \Sigma_\epsilon = cI$. If the feature noise is isotropic: $\Sigma_{\epsilon_x} = rI$, then $\Sigma_{x'} = (1+r)I$, $V_{\text{eff}} = (1+r)^{-1}I$, and $\Sigma_{\epsilon,\text{eff}} = \left(c + \frac{r}{1+r}\right)I$. The effective label noise remains isotropic, so by Corollary 5.2 the selected solution is the minimum-sharpness solution. From Eq. (23),

$$T_{\text{noisy}} = Dd^2(1+r)^{-(D-2)/D}. \tag{158}$$

Thus, for $D > 2$, isotropic feature noise strictly decreases the sharpness.

Conversely, starting from the same isotropic problem, consider anisotropic feature noise applied only to one input direction: $\Sigma_{\epsilon_x} = \text{diag}(r, 0, \dots, 0)$. Then $x_1' = 1+r$, $v_1^{\text{eff}} = \frac{1}{1+r}$, $e_1^{\text{eff}} = c + \frac{r}{1+r}$, while all remaining modes are unchanged. As $r \to \infty$, the input trace satisfies $\text{Tr}\Sigma_{x'} = \Theta(r)$, whereas $\text{Tr}S$, $\text{Tr}\Sigma_{\epsilon,\text{eff}}$, and $\text{Tr}[\Sigma_{\epsilon,\text{eff}}^{-1}LSL^{\top}]$ remain bounded above and below by positive constants. Equation (22) therefore gives

$$T_{\text{noisy}} = \Theta(r^{1/D}), \qquad r \to \infty, \tag{159}$$

and hence the sharpness can become arbitrarily larger than that of the original problem.

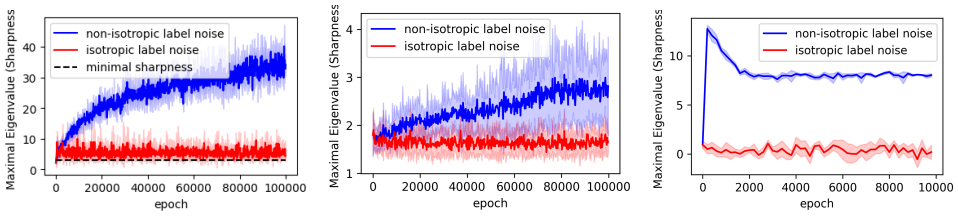


Figure 5: The maximal eigenvalue of the Hessian evolves similarly to the trace. **Left**: linear networks, with the minimum predicted by Theorem B.5. **Middle**: Two layer ReLU networks trained under a teacher-student setting. **Right**: Two layer ReLU networks trained on MNIST. Each line is averaged over five trials shown with the standard error.

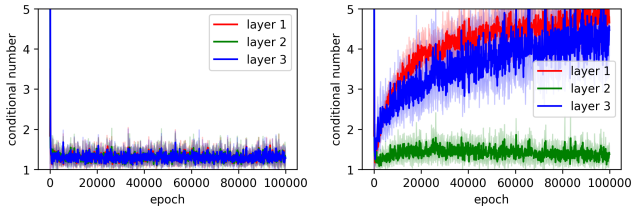


Figure 6: The evolution of the condition numbers for three-layer linear networks. **Left**: Balanced noise in the labels. **Right**: Imbalanced noise in the labels.

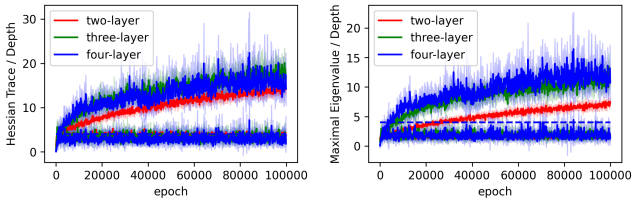


Figure 7: The evolution of sharpness for linear networks of various depths. **Left**: Traces of Hessian. **Right**: Maximal eigenvalues of Hessian.

C EXPERIMENTS

Experiments on linear networks For the linear network settings, we set $D = 3$, $d_x = d_y = 2$, and $V = I$, with data drawn from a standard Gaussian distribution ($\Sigma_x = I$). The residues are always set to zero. We configure the noise scales as $\sqrt{\Sigma_\epsilon} = \text{diag}(10, 0.1)$ for the anisotropic case and $\sqrt{\Sigma_\epsilon} = \text{diag}(10, 10)$ for the isotropic case. Training is performed using online SGD for 10^5 epochs (learning rate 0.01, batch size 10). The Hessian is estimated via 200 random samples, where the trace is calculated exactly and the maximal eigenvalue is computed using the power method. For experiments in Figure 1, we maintain these hyperparameters but adjust the initialization strength and learning rate (to 0.001), or transition to a full-batch gradient descent (GD) setting with 1000 samples.

In Figure 4, we use a three-layer width-2 linear residual network $f(x) = Z_3 Z_2 Z_1 x$ with $Z_k = I + W_k$ trained on $x \sim \mathcal{N}(0, I)$ with targets $y = Vx + \epsilon$, where $V = \text{diag}(0.4, 4)$. All weights are initialized i.i.d. from $\mathcal{N}(0, 0.3^2)$. We use learning rate 0.02, batch size 64, and no weight decay. The label-noise covariance is diagonal with fixed total variance $\text{Tr}(\Sigma_\epsilon) = 2$; its condition number is varied by setting $v_2/v_1 = \kappa(\Sigma_\epsilon)$. For Figure 4 (Left) we use $\kappa = 64$ and compare SGD, GD, and SGD followed by GD after 4×10^4 steps. For Figure 4 (Right), the reported converged value is averaged over the final 20% of training, with error bars denoting the standard error across seeds. All runs are trained for 3×10^5 steps over five random seeds.

To complement the observations in Figure 2, we evaluate the maximal eigenvalue of the Hessian matrix, as shown in Figure 5. The resulting curves exhibit consistent trends, further corroborating Corollary 5.3.

Figure 6 illustrates the condition numbers across layers for a three-layer linear network. Aligning with our analytical solution (77), we observe that balanced noise in the labels leads to small condition numbers across all layers. Conversely, imbalanced noise primarily inflates the condition numbers of the first and last layers, while intermediate layers remain stable. Furthermore, Figure 7 demonstrates that depth acts only as a multiplicative factor for sharpness, as suggested by Corollary 5.1.

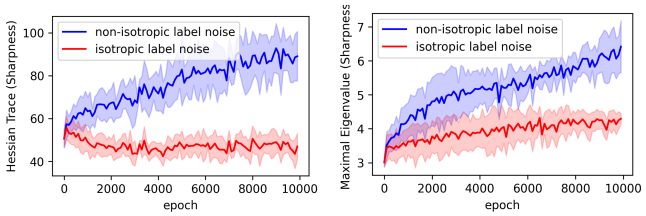


Figure 8: anisotropic noise leads to progressive sharpening also for transformers. **Left**: The evolution of the sharpness. **Right**: The evolution of the maximal eigenvalue.

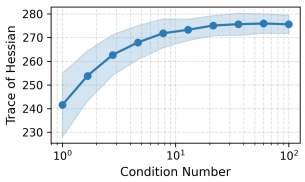


Figure 9: The trace of the Hessian of a vision transformer increases when the condition numbers of the label noise increase. The vision transformer (2 layers with 4 attention heads each) is trained on the CIFAR-10 dataset until convergence. The shaded area represents the standard error estimated over 5 batches.

Experiments on non-linear settings In the ReLU teacher-student setting (in Figures 2 and 5), we employ a single-hidden-layer teacher-student framework with $d_x = d_y = 2$. The teacher and student networks consist of 100 and 4 hidden units, respectively, both initialized using the Kaiming method. We use MSE loss and online SGD (learning rate 0.1, batch size 100) for 10^5 epochs. Noise scales are set to $\mathrm{diag}(3, 0.03)$ and $\mathrm{diag}(3, 3)$. Hessian statistics are computed using 200 samples, with the trace approximated by the Hutchinson estimator.

For the MNIST dataset (in Figures 2 and 5), a two-layer ReLU network (128 hidden units) is trained using cross-entropy loss. noise in the labels is introduced by adding Gaussian noise to one-hot labels, followed by a softmax operation. The anisotropic noise covariance $\sqrt{\Sigma_\epsilon}$ is set to $\mathrm{diag}(10, \ldots, 10, 1, \ldots, 1)$ (five entries each), while the isotropic case uses $\sqrt{\Sigma_\epsilon} = 10I$. We run SGD for 10^4 epochs (learning rate 0.01, batch size 128) and evaluate the Hessian on a test batch of size 1024.

To verify the universality of our findings beyond linear networks and MLPs, we extend our experiments to transformers and RNNs in a teacher-student framework. In Figures 3 and 8, we employ a simplified Transformer comprising a two-head self-attention layer followed by an MLP with residual connections. The model takes inputs of dimension 8 and sequence length 4, with the output dimension matching the input. The hidden dimension of the MLP is set to 100 for the teacher and 4 for the student, both initialized via the Kaiming method.

Training is conducted using online SGD (learning rate 0.03, batch size 300) for 10^4 epochs on standard Gaussian data. The Hessian is evaluated on a test set of size 1000. We configure the noise scales as $\sqrt{\Sigma_\epsilon} = \mathrm{diag}(4, 0.04, \ldots, 0.04)$ for the anisotropic case and $\sqrt{\Sigma_\epsilon} = \mathrm{diag}(4, \ldots, 4)$ for the isotropic case. The results in Figure 8 confirm that anisotropic noise in the labels induces the same progressive sharpening phenomenon in transformers as observed in simpler models.

Furthermore, we investigate the specific impact of the noise covariance condition number on sharpness in Figure 3. Maintaining the same experimental setup, we systematically vary the smaller eigenvalues of Σ_ϵ while keeping other hyperparameters fixed. This analysis is also extended to RNNs, where the teacher and student hidden units are set to 100 and 4, respectively. For experiments involving cross-entropy loss, an additional softmax layer is applied to the network output. Across all these diverse settings—including different architectures and loss functions—the observed trends remain highly consistent, further validating the robustness of our theoretical predictions.

To narrow the gap between our analytical model and real-world architectures, we have also conducted an addition experiment on a small Vision Transformer Dosovitskiy et al. (2020) (2 layers with 4 attention heads each) trained on the CIFAR-10 dataset in Figure 9. These results demonstrate

that even in this more complex setting, the sharpness at convergence increases monotonically with the condition number of the label noise, corroborating our theoretical prediction in Corollary 5.1.

D DECLARATION OF LLM USAGE

We used LLMs in the following stages of the preparation of this work: editing (e.g., grammar, spelling, word choice), drafting sections of the paper, facilitating or running experiments, assisting in the writing of proofs, visualizing results for submission. All LLM-generated material used in this manuscript was manually verified by the authors to ensure that it performs as intended, in particular with regard to the proof and the numerical experiments.