SGD at the Edge of Stability: The Stochastic Sharpness Gap

Fangshuo Liao*
Department of Computer Science
Rice University
Houston, TX 77005
Fangshuo.Liao@rice.edu

Afroditi Kolomvaki
Department of Computer Science
Rice University
Houston, TX 77005
ak203@rice.edu

Anastasios Kyrillidis
Department of Computer Science
Rice University
Houston, TX 77005
anastasios@rice.edu

Abstract

Edge of Stability (EoS) refers to the phenomenon where full-batch gradient descent (GD) training of neural networks with step size η pushes the largest eigenvalue of the Hessian, i.e. the sharpness, to $2/\eta$ and hovers there. Damian et al. [2023] explains the hovering behavior of the sharpness by *self-stabilization*, a mechanism driven by third-order structure of the loss, and shows that GD implicitly follows projected gradient descent (PGD) on the set where the sharpness is below $2/\eta$. For mini-batch stochastic gradient descent (SGD), the sharpness stabilizes *below* $2/\eta$, with the gap widening as the batch size decreases. However, no theoretical explanation exists for this suppression. In this paper, we introduce *stochastic self-stabilization* that extends the self-stabilization framework to SGD. Our key insight is that gradient noise injects variance into the oscillatory dynamics along the top Hessian eigenvector, strengthening the sharpness-reducing force and shifting the equilibrium below $2/\eta$. Following the approach of Damian et al. [2023], we define *stochastic predicted dynamics* that tracks the deviation of SGD from the PGD trajectory, and prove a stochastic coupling theorem that relates the SGD sharpness and loss to those of PGD. Based on our predicted dynamic, we derive a closed-form equilibrium sharpness gap that scales with the variance of the gradient noise projected onto the top eigenvector of the Hessian. This formula predicts that smaller batch sizes yield flatter solutions, and recovers GD under full-batch.

1 Introduction

The choice of optimizer and its hyperparameters such as step size profoundly affects both the training dynamics and the generalization performance of neural networks. Classical optimization theory relies on the *descent lemma* which guarantees monotonic loss decrease only when the step size satisfies $\eta < 2/S(\boldsymbol{\theta})$, where $S(\boldsymbol{\theta}) = \lambda_{\max}(\nabla^2 L(\boldsymbol{\theta}))$ is the largest eigenvalue of the Hessian of the loss function $L(\cdot)$. In recent years, a line of empirical work [Cohen et al., 2021, Jastrzebski et al., 2019, 2020, Xing et al., 2018] established that gradient-based optimizers routinely operate and decrease the loss in regimes that classical optimization theory deems unstable. For full-batch gradient descent (GD), Cohen et al. [2021] documented two striking phenomena: *progressive sharpening*, where the

*Work done while at Rice University. Currently at NYU. Alternative email address: Fangshuo.Liao@nyu.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

sharpness steadily increases throughout early training until it reaches the instability threshold $2/\eta$, and *Edge of Stability* (EoS), where once the sharpness reaches $2/\eta$, it ceases to grow and oscillates in a narrow band around this value, while the loss continues to decrease non-monotonically.

Among the two striking phenomena, *progressive sharpening* is believed to be specific to neural network training, while EoS is shown to be a more universal behavior in optimization. In particular, Damian et al. [2023] provided a theoretical explanation for EoS by identifying a *self-stabilization* mechanism. Through a cubic Taylor expansion of the loss, they showed that as the iterates diverge along the top Hessian eigenvector due to instability, the third-order term acts as a restoring force that drives the sharpness back below $2/\eta$. On a technical level, Damian et al. [2023] defines a *constrained trajectory* of neural network parameters $\{\boldsymbol{\theta}_t^\dagger\}$ via projected gradient descent (PGD) onto the set where the sharpness is below $2/\eta$, and show that at EoS, the GD trajectory oscillates around this constrained trajectory, which justifies the "hovering" behavior of the sharpness.

For mini-batch SGD, the situation is both empirically richer and theoretically more opaque. Cohen et al. [2021] observed that during SGD training, the sharpness stabilizes *below* the $2/\eta$ threshold, with the suppressed magnitude correlating strongly with the batch size. More recently, Andreyev and Beneventano [2025] identified the notion of *Batch Sharpness*, defined as the expected directional curvature of the mini-batch Hessian along the mini-batch gradient. They showed empirically the *Edge of Stochastic Stability* (EoSS), where the *Batch Sharpness* saturates at $2/\eta$ for SGD. Meanwhile, Lee and Jang [2023] introduced *Interaction-Aware Sharpness* and showed that it characterizes the onset of oscillatory behavior in SGD. Despite the progress above that aims to fill in the gap of EoS in mini-batch SGD, a fundamental question remains unanswered:

By what mechanism does SGD *self-stabilize, and why does the full-batch sharpness settle at a level below $2/\eta$ that depends on the batch size?*

Main Contributions. In this paper, we provide an answer of this question by extending the self-stabilization framework of Damian et al. [2023] from GD to SGD. Our contributions are:

1. **Stochastic Self-Stabilization Mechanism.** We show that gradient noise injects additional variance into the oscillatory dynamics along the top Hessian eigenvector. Through the same cubic nonlinearity that drives GD self-stabilization, this variance acts as an additional sharpness-suppressing force, shifting the equilibrium sharpness below $2/\eta$.
2. **Stochastic Predicted Dynamics and Coupling.** Following the methodology of Damian et al. [2023], we define *enhanced stochastic predicted dynamics* that extend their deterministic predicted dynamics to include gradient noise, and we prove a *stochastic coupling theorem* showing that the true SGD trajectory is well-approximated by the constrained trajectory $\boldsymbol{\theta}_t^\dagger$ plus the stochastic predicted dynamics.
3. **Closed-Form Sharpness Gap.** From the stationarity conditions of the stochastic predicted dynamics in a simplified setting, we derive the full-batch sharpness at equilibrium under SGD as $S_{\mathrm{eq}} \approx \frac{2}{\eta} - \frac{\eta\beta\sigma_{\boldsymbol{u}}^2}{4\alpha}$, where α is the progressive sharpening coefficient, β is the self-stabilization strength, and $\sigma_{\boldsymbol{u}}^2$ is the gradient noise variance projected onto the top eigenvector of the Hessian.
4. **Experimental Verification.** We verify all predictions on MLPs and CNNs trained on CIFAR-10 with varying batch sizes and step sizes, demonstrating quantitative agreement between the predicted and measured sharpness gaps.

1.1 Related Work

Edge of Stability. The *Edge of Stability* (EoS) phenomenon was systematically documented by Cohen et al. [2021], who observed that full-batch gradient descent on neural networks operates in a regime where the sharpness $S(\boldsymbol{\theta}) = \lambda_{\max}(\nabla^2 L(\boldsymbol{\theta}))$ rises steadily until it reaches $2/\eta$ and then hovers near this instability threshold, while the training loss continues to decrease non-monotonically over long timescales. Earlier empirical work by Jastrzebski et al. [2019, 2020] and Xing et al. [2018] had already suggested that the sharpness of the loss landscape along the SGD trajectory is strongly influenced by the hyperparameters, such as the learning rate and batch size.

Subsequent work has analyzed the EoS from multiple angles. Arora et al. [2022] studied EoS for normalized gradient descent and established conditions under which training dynamics evolve along

a sharpness-minimizing flow on the manifold of minimizers. Damian et al. [2023] provided the first rigorous mechanism for GD self-stabilization via a cubic Taylor expansion of the loss. Zhu et al. [2023] analyzed EoS dynamics through a minimalist example, deriving the sharpness behavior near the stability threshold. Wang et al. [2022] analyzed sharpness dynamics in two-layer linear networks, providing further theoretical grounding for the progressive sharpening and edge-of-stability behavior. Chen and Bruna [2023] studied gradient descent behavior beyond the edge of stability via two-step gradient updates, characterizing a local implicit bias toward symmetric solutions. Kreisler et al. [2023] showed that the sharpness of the gradient flow solution decreases monotonically during EoS training in scalar networks. Agarwala et al. [2023] demonstrated that second-order regression models exhibit progressive sharpening and EoS, suggesting these phenomena are not unique to neural networks. On the SGD side, Andreyev and Beneventano [2025] identified *Batch Sharpness* as the quantity that saturates at $2/\eta$ during SGD training. Lee and Jang [2023] similarly introduced *Interaction-Aware Sharpness* and characterized the onset of oscillatory behavior in SGD. Recently, Emmanouilidis et al. [2026] studied the stabilization and convergence property of SGD with a large step size at the EoS. In addition, Cai et al. [2026] questions the validity of EoS in large language models trained with SGD, and proposed Edge-of-Convexity as a generalized notion.

Implicit Bias of SGD Related to Sharpness. A foundational empirical observation is that SGD with smaller batch sizes or larger learning rates tends to find flatter minima that generalize better [Jastrzebski et al., 2019]. Keskar et al. [2017] showed that large-batch SGD converges to sharper minimizers with worse generalization, while small-batch SGD consistently finds flatter minima due to the inherent noise in gradient estimation. The relationship between batch size and learning rate in controlling sharpness was further studied by Jastrzebski et al. [2018], who showed that the equilibrium properties of SGD depend primarily on the ratio η/b, and by Wu et al. [2018], who analyzed the dynamical stability of global minima under SGD from a linear stability perspective. Smith et al. [2021] characterized the implicit regularizer of SGD using backward error analysis, showing that the scale of implicit regularization is proportional to η/b. Wu et al. [2022] analyzes the alignment between SGD noise and the sharp directions of the loss landscape, and proved that linearly stable minima under SGD must satisfy a Frobenius-norm bound on the Hessian that depends on the ratio b/η. HaoChen et al. [2021] showed that the *shape* of the noise covariance determines which minimizers SGD converges to, demonstrating that parameter-dependent noise finds flatter minima than spherical Gaussian noise. Damian et al. [2021] proved that label noise SGD provably prefers flat global minimizers by penalizing sharpness through its parameter-dependent noise structure. These works establish the empirical and theoretical backdrop for the batch-size-dependent sharpness we explain. In a separate direction, Mulayoff and Michaeli [2024] focus on the linear stability of the whole quadratic system under SGD, which is different from the focus of our paper.

2 Preliminaries

We consider the empirical loss $L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i; \boldsymbol{\theta})$ over a training set $\{\boldsymbol{x}_i\}_{i=1}^{n}$, where $\boldsymbol{\theta} \in \mathbb{R}^d$. Stochastic gradient descent with constant step size $\eta > 0$ and mini-batch size b performs:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{g}_{B_t}(\boldsymbol{\theta}_t), \qquad \boldsymbol{g}_{B_t}(\boldsymbol{\theta}) = \frac{1}{b}\sum_{i \in B_t} \nabla \ell(\boldsymbol{x}_i; \boldsymbol{\theta}), \tag{1}$$

where $B_t \subseteq \{1, \ldots, n\}$ with $|B_t| = b$ is drawn uniformly at random. We write $\boldsymbol{g}_{B_t}(\boldsymbol{\theta}_t) = \nabla L(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_t$ where $\boldsymbol{\xi}_t$ is the gradient noise satisfying $\mathbf{E}[\boldsymbol{\xi}_t | \boldsymbol{\theta}_t] = \mathbf{0}$. Full-batch GD is the special case $b = n$.

Sharpness and Its Gradient. When the top eigenvalue of $\mathbf{H}(\boldsymbol{\theta}) = \nabla^2 L(\boldsymbol{\theta})$ is simple[2], the sharpness $S(\boldsymbol{\theta}) = \lambda_{\max}(\mathbf{H}(\boldsymbol{\theta}))$ is differentiable, and its gradient is given by relating the third derivative of the loss to the sharpness gradient. Damian et al. [2023] shows that the gradient of the sharpness is given by the fundamental identity $\nabla S(\boldsymbol{\theta}) = \nabla^3 L(\boldsymbol{\theta})\big(\boldsymbol{u}(\boldsymbol{\theta}), \boldsymbol{u}(\boldsymbol{\theta})\big)$, i.e., tensor-vector product between $\nabla^3 L(\boldsymbol{\theta})$ and two $\boldsymbol{u}(\boldsymbol{\theta})$, where $\boldsymbol{u}(\boldsymbol{\theta})$ denotes the top eigenvector of $\mathbf{H}(\boldsymbol{\theta})$.

Progressive Sharpening. Following Damian et al. [2023], we define the *progressive sharpening coefficient* as:

$$\alpha(\boldsymbol{\theta}) := -\langle \nabla L(\boldsymbol{\theta}), \nabla S(\boldsymbol{\theta}) \rangle. \tag{2}$$

When $\alpha(\boldsymbol{\theta}) > 0$, the negative gradient direction $-\nabla L(\boldsymbol{\theta})$ has a positive component along $\nabla S(\boldsymbol{\theta})$, leading to a negative first-order term in the Taylor expansion of $S(\boldsymbol{\theta})$. This implies a force that increases the sharpness when traveling along $-\nabla L(\boldsymbol{\theta})$.

[2]An eigenvalue λ of a matrix $\mathbf{A}$ is called *simple* if it has *algebraic multiplicity equal to 1*.

The Constrained Trajectory and Self-Stabilization under GD. We present the central construction of Damian et al. [2023] that serves as the foundation for our analysis. Define the *stable set*:

$$\mathcal{M} := \{\boldsymbol{\theta} : S(\boldsymbol{\theta}) \le 2/\eta \text{ and } \langle \nabla L(\boldsymbol{\theta}), \boldsymbol{u}(\boldsymbol{\theta})\rangle = 0\}. \tag{3}$$

The *constrained trajectory* $\{\boldsymbol{\theta}_t^\dagger\}$ is defined by projected gradient descent on $\mathcal{M}$:

$$\boldsymbol{\theta}_0^\dagger := \mathrm{proj}_{\mathcal{M}}(\boldsymbol{\theta}_0), \qquad \boldsymbol{\theta}_{t+1}^\dagger := \mathrm{proj}_{\mathcal{M}}\big(\boldsymbol{\theta}_t^\dagger - \eta \nabla L(\boldsymbol{\theta}_t^\dagger)\big), \tag{4}$$

where $\mathrm{proj}_{\mathcal{M}}(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}' \in \mathcal{M}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ is the orthogonal projection. Damian et al. [2023, Lemma I.3] showed that $S(\boldsymbol{\theta}_t^\dagger) = 2/\eta$ for all t, and the PGD update takes the approximate form $\boldsymbol{\theta}_{t+1}^\dagger = \boldsymbol{\theta}_t^\dagger - \eta P^\perp_{\boldsymbol{u}_t, \nabla S_t} \nabla L_t + O(\epsilon_t^2 \eta \|\nabla L_t\|)$, where $P^\perp_{\boldsymbol{u}_t, \nabla S_t}$ projects out the $\boldsymbol{u}_t$ and ∇S_t directions. That is, the constrained trajectory performs gradient descent with the sharpness-increasing components projected out.

Lemma 2.1 (Properties of $\boldsymbol{\theta}_t^\dagger$, Damian et al. [2023, Lemma 5, 6])**.** *Under Assumptions 3–4 below, for all $t \le \mathscr{T}$:*

$$S(\boldsymbol{\theta}_t^\dagger) = 2/\eta, \qquad \boldsymbol{\theta}_{t+1}^\dagger = \boldsymbol{\theta}_t^\dagger - \eta P^\perp_{\boldsymbol{u}_t, \nabla S_t} \nabla L_t + O(\epsilon_t^2 \cdot \eta \|\nabla L_t\|). \tag{5}$$

Moreover, $L(\boldsymbol{\theta}_{t+1}^\dagger) \le L(\boldsymbol{\theta}_t^\dagger) - \Omega(\eta \left\|P^\perp_{\boldsymbol{u}_t, \nabla S_t} \nabla L_t\right\|^2)$ (descent lemma for $\boldsymbol{\theta}_t^\dagger$).

The simplified update rule of $\boldsymbol{\theta}_t^\dagger$ in (5) serves as a base for our analysis of the SGD dynamic. Moreover, since the loss decreases along the PGD trajectory, as shown in Lemma 2.1, as long as the SGD trajectory oscillates around $\boldsymbol{\theta}_t^\dagger$, the loss will also decrease in this oscillation.

Self-stabilization under GD. We briefly summarize the self-stabilization mechanism of Damian et al. [2023] for full-batch GD, as our extension to SGD follows the same conceptual structure. Consider GD with step size η: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t)$. Suppose the iterate reaches a point $\boldsymbol{\theta}^\star \in \mathcal{M}$. Define the displacements:

$$x_t := \langle \boldsymbol{u}, \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star \rangle, \qquad y_t := \langle \nabla S, \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star \rangle,$$

where $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{\theta}^\star)$ and $\nabla S = \nabla S(\boldsymbol{\theta}^\star)$. Note that y_t approximates the sharpness deviation by $S(\boldsymbol{\theta}_t) \approx 2/\eta + y_t$. Conceptually, an increased y_t caused by the progressive sharpening coefficient $\alpha > 0$ will lead to unstable dynamic along the $\boldsymbol{u}$, thus causing $|x_t|$ to increase. A third-order Taylor expansion shows that $y_{t+1} - y_t$ becomes negative when $|x_t|$ is large. Therefore, when the unstable quantity x_t reaches a certain threshold, y_t starts to decrease, thus returning to stability.

3 The Four Stages of EoS under SGD

In this section, we give a heuristic derivation of the four-stage dynamic of the sharpness under SGD. The presented dynamic generalizes the setting in Damian et al. [2023], and reduces to their setting by letting the noise to be zero. For the simplicity of the demonstration, here we use a fixed reference $\boldsymbol{\theta}^\star$ and ignore the time dependence of α, β, δ. The rigorous analysis in Section 4 addresses both issues.

Suppose the SGD trajectory first reaches a point where the sharpness is approximately $2/\eta$. Let $\boldsymbol{\theta}^\star$ denote the projection of this point onto the manifold $\mathcal{M} := \{\boldsymbol{\theta} : S(\boldsymbol{\theta}) \le 2/\eta, \langle \nabla L(\boldsymbol{\theta}), \boldsymbol{u}(\boldsymbol{\theta})\rangle = 0\}$. We fix the reference point $\boldsymbol{\theta}^\star$ and define the displacement vector:

$$\boldsymbol{v}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star, \tag{6}$$

with projections $x_t := \langle \boldsymbol{u}, \boldsymbol{v}_t \rangle$ and $y_t := \langle \nabla S, \boldsymbol{v}_t \rangle$, where $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{\theta}^\star)$ and $\nabla S = \nabla S(\boldsymbol{\theta}^\star)$.

Stage 1: Progressive Sharpening under SGD. When $\boldsymbol{v}_t$ is small, $\nabla L(\boldsymbol{\theta}_t) \approx \nabla L(\boldsymbol{\theta}^\star) = \nabla L$. The SGD update is $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\nabla L + \boldsymbol{\xi}_t)$, where $\boldsymbol{\xi}_t = \boldsymbol{g}_{B_t}(\boldsymbol{\theta}_t) - \nabla L(\boldsymbol{\theta}_t)$. Therefore,

$$\mathbf{E}[y_{t+1} - y_t] = \langle \nabla S, \mathbf{E}[\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t] \rangle = -\eta \langle \nabla L, \nabla S \rangle = \eta\alpha. \tag{7}$$

This shows that the *expected* rate of progressive sharpening is identical to GD.

Stage 2: Blowup under SGD. The displacement x_t evolves as:

$$x_{t+1} = \langle \boldsymbol{u}, \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^\star \rangle = x_t - \eta \langle \boldsymbol{u}, \boldsymbol{g}_{B_t}(\boldsymbol{\theta}_t) \rangle = x_t - \eta \langle \boldsymbol{u}, \nabla L(\boldsymbol{\theta}_t) \rangle - \eta \zeta_t. \tag{8}$$

where $\zeta_t = \langle \boldsymbol{u}, \boldsymbol{\xi}_t \rangle$ represents the noise projected onto the top eigenvector of the Hessian. To evaluate $\langle \boldsymbol{u}, \nabla L(\boldsymbol{\theta}_t) \rangle$, we Taylor-expand ∇L around $\boldsymbol{\theta}^\star$. To leading order (see Appendix A for the full expansion),

$$\langle \boldsymbol{u}, \nabla L(\boldsymbol{\theta}_t) \rangle \approx S(\boldsymbol{\theta}^\star) x_t + y_t x_t = \left(\frac{2}{\eta} + y_t\right) x_t, \tag{9}$$

where we used $S(\boldsymbol{\theta}^\star) = 2/\eta$. Substituting (9) into (8):

$$x_{t+1} = -(1 + \eta y_t)x_t - \eta\zeta_t. \tag{10}$$

To recover the heuristic dynamic of GD in Damian et al. [2023], we only need to set the additive noise term $-\eta\zeta_t$ to zero. When $y_t > 0$, the deterministic part $-(1 + \eta y_t)x_t$ causes $|x_t|$ to grow, but the noise ζ_t continuously perturbs the trajectory.

Stage 3: Self-Stabilization under SGD. Similar to Damian et al. [2023], expanding the cubic term in the Taylor expansion gives

$$\frac{1}{2}\nabla^3 L(\boldsymbol{\theta}_t - \boldsymbol{\theta}^\star, \boldsymbol{\theta}_t - \boldsymbol{\theta}^\star) \approx \frac{x_t^2}{2}\nabla^3 L(\boldsymbol{u}, \boldsymbol{u}) = \frac{x_t^2}{2}\nabla S$$

When x_t^2 is large, the additional cubic term causes self-stabilization effect

$$\mathbf{E}\big[y_{t+1} - y_t\big] = \eta\alpha - \eta\left\langle \nabla S, \frac{x_t^2}{2}\nabla S\right\rangle = \eta\left(\alpha - \frac{\beta}{2}\mathbf{E}\big[x_t^2\big]\right)$$

where in this case $\beta = \|\nabla S\|^2$. However, x_t is random with noise-inflated second moment. Thus, this term should have a magnitude that grows with the noise scale.

Stage 4: Role of Noise Variance. In this stage, x_t grows to a large enough magnitude that drives the sharpness back to stability condition. In particular, the dynamic of x_t is governed by

$$\mathbf{E}\big[x_{t+1}^2\big] = \mathbf{E}\big[(1 + \eta y_t)^2 x_t^2\big] + \eta^2\sigma_{\boldsymbol{u}}^2 \tag{11}$$

This stage marks a particular difference from Damian et al. [2023]. In the deterministic scenario, the expectation can be dropped and we have that $\sigma_{\boldsymbol{u}} = 0$. However, our framework handles the SGD noise explicitly. Indeed, apply a mean-field approximation that decouples the randomness in x_t and y_t and solve for the sharpness at equilibrium gives

$$\bar{y} = -\frac{\eta\beta\sigma_{\boldsymbol{u}}^2}{4\alpha}, \qquad \bar{S} \approx \frac{2}{\eta} - \frac{\eta\beta\sigma_{\boldsymbol{u}}^2}{4\alpha} = \frac{2}{\eta} - \Delta S. \tag{12}$$

The gradient noise projection ζ_t introduces an additional growth to the sequence $\mathbf{E}\big[x_t^2\big]$. In order to stabilize, the $(1 + \eta y_t)$ has to be smaller than 1 to contract the magnitude of the first term in (11). Thus, to maintain equilibrium, the system shifts to $\bar{y} < 0$, resulting in a sharpness gap $\Delta S \propto \sigma_{\boldsymbol{u}}^2$.

4 Main Theoretical Results

In this section, we derive a rigorous approximation of the heuristic dynamic in Section 3. Similar to Damian et al. [2023], our analysis also relies on a reference trajectory induced by the projected gradient descent defined in (3). We make the following definitions along the reference trajectory.

Definition 4.1 (Taylor Expansion Quantities at $\boldsymbol{\theta}_t^\dagger$)**.**

$$\nabla L_t := \nabla L(\boldsymbol{\theta}_t^\dagger), \quad \mathbf{H}_t := \nabla^2 L(\boldsymbol{\theta}_t^\dagger), \quad \boldsymbol{u}_t := \boldsymbol{u}(\boldsymbol{\theta}_t^\dagger),$$
$$\nabla S_t := \nabla S(\boldsymbol{\theta}_t^\dagger), \quad \nabla S_t^\perp := P_{\boldsymbol{u}_t}^\perp \nabla S(\boldsymbol{\theta}_t^\dagger), \quad \kappa_t := \langle \nabla S_t, \boldsymbol{u}_t\rangle.$$

Definition 4.2 (Landscape Parameters)**.** *Define*

$$\alpha_t := -\langle \nabla L_t, \nabla S_t\rangle, \qquad \beta_t := \left\|\nabla S_t^\perp\right\|^2, \qquad \delta_t := \sqrt{2\alpha_t/\beta_t},$$
$$\epsilon_t := \eta\sqrt{\alpha_t}, \qquad \delta := \sup_t \delta_t, \qquad \epsilon := \sup_t \epsilon_t.$$

The coefficient α_t is the *progressive sharpening force*: when $\alpha_t > 0$, unconstrained gradient descent at $\boldsymbol{\theta}_t^\dagger$ would increase the sharpness. The coefficient β_t is the *strength of the self-stabilization force*: larger β_t means stronger cubic restoring. The scale δ_t sets the amplitude of the oscillations and serves as the fixed-point value of $|x_t|$ in the simplified dynamics.

Definition 4.3 (Sharpness Propagation Factor)**.** *Define* $\mathbf{A}_k := (I - \eta\mathbf{H}_k)P_{\boldsymbol{u}_k}^\perp$ *and*

$$\beta_{s\to t} := (\nabla S_{t+1}^\perp)^\top \left[\prod_{k=t}^{s+1} \mathbf{A}_k\right] \nabla S_s^\perp. \tag{13}$$

The factor $\beta_{s\to t}$ measures the change in sharpness between times s and $t+1$ caused by a displacement of $\nabla S_s^\perp$ at time s. At time s, a displacement in the $\nabla S_s^\perp$ direction increases the sharpness. This displacement then evolves under the linearized dynamics, being multiplied by $\mathbf{A}_k$ at each step k. After $t-s$ steps, the resulting displacement is $[\prod_{k=t}^{s+1}\mathbf{A}_k]\nabla S_s^\perp$, and its effect on the sharpness at time $t+1$ is measured by the inner product with $\nabla S_{t+1}^\perp$. When the landscape quantities are constant (as in the heuristic Section 3), $\beta_{s\to t}\equiv\beta$. Our theoretical result depend on the following assumptions.

Assumption 1 (Eigengap)**.** *For some constant $c_{\mathrm{gap}}<2$, $\lambda_2(\mathbf{H}_t)<c_{\mathrm{gap}}/\eta$ for all $t\le\mathscr{T}$.*

Assumption 2 (Higher-Order Derivative Regularity)**.** *Let ρ_3,ρ_4 be the minimum constant such that there exists a neighborhood $\mathcal{N}$ of $\{\boldsymbol{\theta}_t^\dagger\}_{t\le\mathscr{T}}$ where every sample loss ℓ_i is C^4 on $\mathcal{N}$, and $\left\|\nabla^3 L(\boldsymbol{\theta})\right\|_{\mathrm{op}}\le\rho_3$, $\left\|\nabla^3 L(\boldsymbol{\theta})-\nabla^3 L(\boldsymbol{\theta}')\right\|_{\mathrm{op}}\le\rho_4\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|$ for all $\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathcal{N}$.*

Assumption 3 (Progressive Sharpening)**.** *There exists some constant $c_\alpha,c_\beta\ge 0$ such that for all $t\le\mathscr{T}$, $\alpha_t\ge c_\alpha\left\|\nabla L_t\right\|\left\|\nabla S_t^\perp\right\|$, and $\left\|\nabla S_t^\perp\right\|\ge c_\beta\rho_3$.*

Assumption 4 (Non-Worst-Case Bounds)**.** *For all $t\le\mathscr{T}$ and all $\boldsymbol{v},\boldsymbol{w}\perp\boldsymbol{u}_t$:*

$$\left\|\nabla^3 L_t(\boldsymbol{v},\boldsymbol{w})\right\|\le O(\epsilon\rho_3)\left\|\boldsymbol{v}\right\|\left\|\boldsymbol{w}\right\|,\quad |\boldsymbol{v}^\top\mathbf{H}_t\boldsymbol{v}|\le O(\epsilon)\left\|\mathbf{H}_t\right\|\left\|\boldsymbol{v}\right\|^2,\quad |\lambda_{\min}(\mathbf{H}_t)|\le O(\epsilon)\left\|\mathbf{H}_t\right\|.$$

Assumption 5 (Noise Projection Non-Degeneracy)**.** *$\sigma_{\boldsymbol{u},t}^2>0$ for all t.*

Assumption 6 (Scaling conditions)**.** *We assume that $\mathscr{T}\le c_{\mathscr{T}}/\epsilon$ for some constant $c_{\mathscr{T}}$, $\mathbf{E}\left[\left\|\boldsymbol{\xi}_t\right\|^2\right]\le O\left(\epsilon\delta^2/\eta^2\right)$, and $\rho_4\le O(\eta\rho_3^2)$.*

Here, Assumption 1-4 are inherited from Damian et al. [2023]. Assumption 5 assumes that the gradient noise vector has a non-vanishing component when projected onto the top eigenvector of the Hessian. In later analysis we will see that this quantity determines the sharpness gap, and Assumption 5 is made to make sure that the sharpness gap is non-trivial. Assumption 6 contains requirements on $\mathscr{T},\rho_4$, which are inherited from Damian et al. [2023]. Furthermore, it also requires an upper bound on the noise variance $\mathbf{E}\left[\left\|\boldsymbol{\xi}_t\right\|^2\right]\le O\left(\frac{\epsilon\delta^2}{\eta^2}\right)$. This is to make sure that the noise scale does not drive the dynamic of the sharpness into the chaotic regime where the training could potentially blow-up. With these assumptions, we are ready to state our theoretical results.

4.1 Stochastic Coupling Theorem

Defining the displacement from the constrained trajectory as $\boldsymbol{v}_t:=\boldsymbol{\theta}_t-\boldsymbol{\theta}_t^\dagger$, and its projections:

$$x_t:=\langle\boldsymbol{u}_t,\boldsymbol{v}_t\rangle,\qquad y_t:=\langle\nabla S_t^\perp,\boldsymbol{v}_t\rangle. \tag{14}$$

The dynamic of $\boldsymbol{v}_t$, and thus x_t and y_t, during training are chaotic and hard to analyze. To cope with this difficulty, we define a predicted dynamic $\hat{v}_t$ and correspondingly $\hat{x}_t$ and $\hat{y}_t$ that approximate $\boldsymbol{v}_t$ and x_t,y_t reasonably well, but at the same time enjoys easier interpretability.

Definition 4.4 (Enhanced Stochastic Predicted Dynamics)**.** *Define $\hat{\boldsymbol{v}}_0=\boldsymbol{v}_0$, and for $t\ge 0$:*

$$\hat{\boldsymbol{v}}_{t+1}=P_{\boldsymbol{u}_{t+1}}^\perp\left[(I-\eta\mathbf{H}_t)P_{\boldsymbol{u}_t}^\perp\hat{\boldsymbol{v}}_t+\eta\nabla S_t^\perp\tfrac{\delta_t^2-\hat{x}_t^2}{2}\right]-\left((1+\eta\hat{y}_t)\hat{x}_t+\tfrac{\eta}{2}\kappa_t\hat{x}_t^2\right)\boldsymbol{u}_{t+1}-\eta\boldsymbol{\xi}_t, \tag{15}$$

where $\hat{x}_t:=\langle\boldsymbol{u}_t,\hat{\boldsymbol{v}}_t\rangle$ and $\hat{y}_t:=\langle\nabla S_t^\perp,\hat{\boldsymbol{v}}_t\rangle$.

Definition 4.5 (Deterministic and Stochastic Step Map)**.** *The deterministic part of the step map, namely $\widetilde{\mathtt{step}}_t(\boldsymbol{v})$, is*

$$P_{\boldsymbol{u}_{t+1}}^\perp\widetilde{\mathtt{step}}_t(\boldsymbol{v})=P_{\boldsymbol{u}_{t+1}}^\perp\left[\mathbf{A}_t\boldsymbol{v}+\eta\nabla S_t^\perp\tfrac{\delta_t^2-x^2}{2}\right],$$
$$\langle\boldsymbol{u}_{t+1},\widetilde{\mathtt{step}}_t(\boldsymbol{v})\rangle=-(1+\eta y)x-\tfrac{\eta}{2}\kappa_t x^2,$$

where $x=\langle\boldsymbol{u}_t,\boldsymbol{v}\rangle$, $y=\langle\nabla S_t^\perp,\boldsymbol{v}\rangle$. The stochastic step map is $\widetilde{\mathtt{step}}_t^S(\boldsymbol{v}):=\widetilde{\mathtt{step}}_t(\boldsymbol{v})-\eta\boldsymbol{\xi}_t$.

Lemma 4.6 (Projected Dynamics for $\hat{x},\hat{y}$)**.** *Define $\mathcal{I}_{t+1}:=(\nabla S_{t+1}^\perp)^\top[\prod_{k=t}^0\mathbf{A}_k]\hat{\boldsymbol{v}}_0^\perp$. Then:*

$$\hat{x}_{t+1}=-(1+\eta\hat{y}_t)\hat{x}_t-\tfrac{\eta}{2}\kappa_t\hat{x}_t^2-\eta\zeta_t, \tag{16}$$

$$\hat{y}_{t+1}=\mathcal{I}_{t+1}+\eta\sum_{s=0}^t\beta_{s\to t}\frac{\delta_s^2-\hat{x}_s^2}{2}-\eta\sum_{s=0}^t\langle\boldsymbol{\gamma}_{s\to t},\boldsymbol{\xi}_s\rangle, \tag{17}$$

where $\boldsymbol{\gamma}_{s\to t}:=(\nabla S_{t+1}^\perp)^\top[\prod_{k=t}^{s+1}\mathbf{A}_k]P_{\boldsymbol{u}_{s+1}}^\perp$.

See Appendix C for the full proof. We highlight here that our predicted dynamic differs in form from Damian et al. [2023]. Except for introducing the additional disturbance caused by the noise ζ_t and ξ_s, our dynamic involves and additional term $\kappa_t \hat{x}_t^2$ and an explicit resolution of $\mathcal{I}_{t+1}$. We make these choices in the construction to resolve the technical difficulty of handling the stochasticity.

Our main theoretical result is a coupling theorem that shows a strong dependence of the loss and sharpness dynamic on the predicted dynamic we defined above. Establishing this result will allow us to derive the explicit sharpness gap in later sections.

Theorem 1 (Stochastic Coupling in Mean Square)**.** *Let Assumption 1-6 hold with a small enough constant $c_{\mathscr{T}}$. If $\|\boldsymbol{v}_0\|^2 \le c_0\delta^2$ for some small enough constant c_0, then for any small constant ω such that $\omega^{-1} \le O(1)$, we have that with probability at least $1-\omega$, the following holds for all $t \le \mathscr{T}$:*

$$\max\{\|\boldsymbol{v}_t\|, \|\hat{\boldsymbol{v}}_t\|\} \le O(\delta); \quad \|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\| \le O(\epsilon\delta), \qquad \text{(Norm and Deviation Bound)}$$

$$L(\boldsymbol{\theta}_t) = L(\boldsymbol{\theta}_t^\dagger) + \frac{1}{\eta}\hat{x}_t^2 + O\left(\epsilon\delta^2/\eta\right), \qquad \text{(Loss)}$$

$$S(\boldsymbol{\theta}_t) = \frac{2}{\eta} + \hat{y}_t + \kappa_t\hat{x}_t + O\left(\epsilon^2/\eta\right). \qquad \text{(Sharpness)}$$

Full details are in Appendix E. In particular, the theorem states three results. First, both the true deviation from constrained trajectory, namely $\boldsymbol{v}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^\dagger$, and the predicted dynamic $\hat{\boldsymbol{v}}_t$ has a norm upper bounded by $O(\delta)$. Second, the difference between $\boldsymbol{v}_t$ and $\hat{\boldsymbol{v}}_t$ is upper bounded by $O(\epsilon\delta)$, showing a relatively small difference compared to the magnitude of the two vectors $O(\delta)$. This shows that the predicted dynamic roughly describes the true deviation from the constrained trajectory. Moreover, the theorem shows that the loss of the SGD training deviates from the loss of the constrained loss by a magnitude of $\hat{x}_t^2/\eta$ with an additional error of $O\left(\epsilon\delta^2/\eta\right)$. As $\hat{x}_t$ roughly scales with $O(\delta)$, the error term $O\left(\epsilon\delta^2/\eta\right)$ has a small influence in the loss dynamic. Combined with Lemma 2.1, this shows a decrease of the SGD loss with an oscillation cased by the instability term $\hat{x}_t$. Lastly, we can approximate the SGD trajectory's sharpness by a combination of $\hat{y}_t$ and $\hat{x}_t$. In the next section, we will demonstrate how this approximation of the sharpness leads to a sharpness gap that depends on the gradient noise variance.

4.2 Derivation of the Sharpness Gap

To derive a meaningful closed-form sharpness gap, we need to make certain simplifications to avoid excessive dependencies. In particular, we work in a frozen-coefficient regime with an explicit closure assumption, and assume a mean-field approximation of the coupling between $\hat{y}_t$ and $\hat{x}_t$.

Assumption 7 (Stationarity and Decorrelation)**.** *The frozen-coefficient stochastic predicted dynamics admit a stationary distribution for $(\hat{x}_t, \hat{y}_t)$. At stationarity:*

$$\left|\mathbf{E}\left[(1+\eta\hat{y}_t)^2\hat{x}_t^2\right] - (1+\eta\bar{y})^2\mathbf{E}\left[\hat{x}_t^2\right]\right| \le O(\epsilon^2\delta^2), \qquad \bar{y} := \mathbf{E}[\hat{y}_t].$$

Assumption 8 (Frozen Coefficients)**.** *Over the equilibrium window $[T_0, T_0+\mathscr{T}]$, the following quantities evaluated along the reference trajectory $\boldsymbol{\theta}_t^\dagger$ are approximately constant. Specifically:*

1. Landscape geometry: *The Hessian $\mathbf{H}_t \equiv \mathbf{H}$, the top eigenvector $\boldsymbol{u}_t \equiv \boldsymbol{u}$, and the sharpness gradient $\nabla S_t^\perp \equiv \nabla S^\perp$ are constant. As a consequence, the derived scalar quantities are also constant: $\alpha_t \equiv \alpha > 0$, $\beta_t \equiv \beta > 0$, $\delta_t \equiv \delta = \sqrt{2\alpha/\beta}$, $\kappa_t \equiv \kappa = \langle \nabla S, \boldsymbol{u}\rangle \le O(\epsilon\rho_3)$.*
2. Sharpness propagation: $\nabla S^\perp \in \ker(\mathbf{H}|_{\boldsymbol{u}^\perp})$.

Assumption 7 decouples the expectation of $(1+\eta\hat{y}_t)^2\hat{x}_t^2$ into expectation of $\hat{y}_t$ and $\hat{x}_t^2$. Here, $\hat{y}_t$ and $\hat{x}_t$ are the projection of $\hat{v}_t$ onto orthogonal directions, so the correlation between the two are small. We make Assumption 8 so that the dynamic of $\hat{x}_t$ and $\hat{y}_t$ can be well understood. Assuming fixed values of $\mathbf{H}, \alpha, \beta, \delta$, and κ allows us to express the sharpness gap in a simple form without the hard-to-interpret dependencies. The assumption of small κ and $\nabla S^\perp \in \ker(\mathbf{H}|_{\boldsymbol{u}^\perp})$ are simplification of the real-world scenario where high dimensionality implies near-orthogonality between vectors, and the neural network Hessian are effectively low-rank, especially after entering the EoS regime. The condition in these two assumptions are verified experimentally in Appendix H.7 and H.9.

Under the assumptions above, we study the sharpness gap induced by SGD. The predicted dynamics are at equilibrium when the joint distribution of $(\hat{x}_t, \hat{y}_t)$ is at stationary, and thus so are the statistics of the joint distribution. In this section, we will focus on the following statistics:

$\mathbf{E}[\hat{y}_{t+1}] = \mathbf{E}[\hat{y}_t]$ (mean stationarity of $\hat{y}$); $\mathbf{E}\left[\hat{x}_{t+1}^2\right] = \mathbf{E}\left[\hat{x}_t^2\right]$ (second-moment stationarity of $\hat{x}$)

These two conditions yield a system of two equations that determine $\mathbf{E}\left[\hat{x}_t^2\right]$ and $\mathbf{E}[\hat{y}_t]$, which gives us the sharpness gap theorem below.

Theorem 2 (Equilibrium Sharpness under SGD)**.** *Under Assumptions 1–6,7, and 8, with probability at least 1 - ω, at stationary distribution of the predicted dynamic $\hat{y}_t$ and $\hat{x}_t$, we have that:*

$$\mathbf{E}\left[\hat{x}_t^2\right] = \delta^2; \quad \mathbf{E}[\hat{y}_t] = -\frac{\eta\beta\sigma_{\boldsymbol{u}}^2}{4\alpha} + O\left(\frac{\epsilon^2}{\eta}\right)$$

and the sharpness at stability is given by

$$\mathbf{E}[S(\boldsymbol{\theta}_t)] = \frac{2}{\eta} - \frac{\eta\beta\sigma_{\boldsymbol{u}}^2}{4\alpha} + O\left(\frac{\epsilon^2}{\eta}\right). \tag{18}$$

The proof is provided in Appendix F. Theorem 2 states that, at the stationary distribution of $\hat{x}_t$ and $\hat{y}_t$, the unstable force $\hat{x}_t$ has a second moment that matches the stabilizing force δ^2, recovering the condition in Damian et al. [2023]. In the meantime, the sharpness deviation from $2/\eta$, namely $\hat{y}_t$, has a mean at $-\eta\beta\sigma_{\boldsymbol{u}}^2/(4\alpha) + O\left(\epsilon^2/\eta\right)$. Ignoring the small error term $O\left(\epsilon^2/\eta\right)$, we derive a sharpness gap of $-\eta\beta\sigma_{\boldsymbol{u}}^2/(4\alpha)$, showcasing that the suppressed sharpness under the injection of gradient noise. This sharpness gap depends linearly on the variance of the noise when projected onto the top Hessian eigenvector $\sigma_{\boldsymbol{u}}$, which matches the prior work that a smaller batch size result in a less sharp solution Cohen et al. [2021], Andreyev and Beneventano [2025]. The sharpness gap in Theorem 2 can shown to scale as $1/b$ with a simple characterization of $\sigma_{\boldsymbol{u}}^2$, which we present in Appendix G.

5 Experimental Validation

We validate the predictions of our theoretical analysis through comprehensive experiments on neural networks trained with SGD at the edge of stability. Our primary setup follows Cohen et al. [2021]: a two-hidden-layer FC-Tanh network trained with MSE loss on CIFAR-10, with additional experiments on a CNN and FC-ReLU network (with both MSE and cross-entropy losses) in the appendix to verify generality. We train on 5,000 CIFAR-10 images using vanilla SGD with default learning rate $\eta = 0.01$ (stability threshold $2/\eta = 200$), varying $\eta \in \{0.005, 0.008, 0.01, 0.015, 0.02\}$ for scaling experiments, and repeat each run over 3–5 random seeds. Throughout training we periodically compute the top Hessian eigenvalue, projected gradient noise variance, batch sharpness, and the sharpness gradient quantities α and β[3]; equilibrium sharpness is taken as the average of $\lambda_{\max}$ over the final 20 measurements.

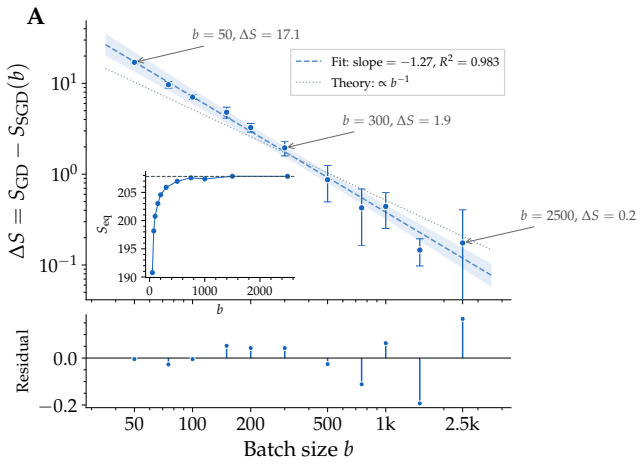

Figure 1: Log-log plot of the sharpness gap ΔS versus batch size b. The dotted line shows the theoretical slope. The shaded band indicates the 95% confidence interval. Error bars indicate ± 1 standard deviation.

5.1 Batch-Size and Noise Variance Scaling of Sharpness Gap

In this section we study the scaling behavior of the sharpness gap with different batch sizes and the noise variances. We choose to study the scaling behavior rather than directly validating the prediction $\Delta S = \eta\beta\sigma_u^2/(4\alpha)$ because of the hardness of measuring α and β in real training. In Appendix H.9.1, we explicitly test the dependency on $\sigma_{\boldsymbol{u}}$ by injecting artificial noise in the gradient.

[3]Notice that in our theory these quantities are measured along the constrained trajectory. In experiments we use the approximation of the measurements along the SGD trajectory.

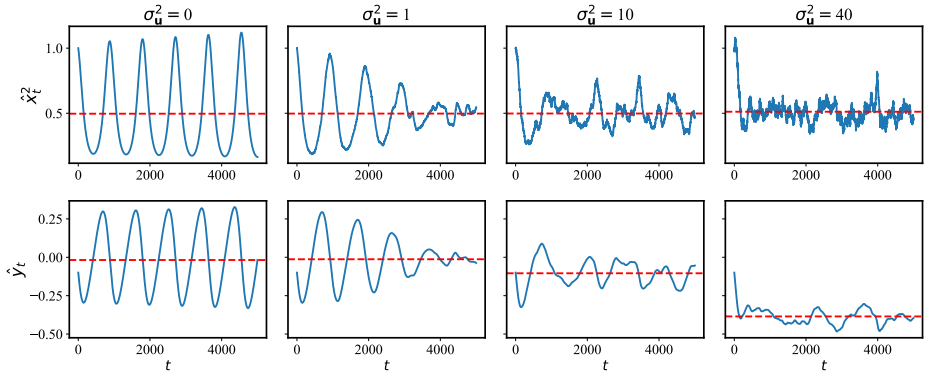

Figure 3: Simulation of the $\hat{x}_t^2, \hat{y}_t$ dynamic given in Lemma 4.6. We choose $\beta_{s\to t} = \beta = 1$, and $\delta_t^2 = \delta^2 = 0.5$. Also, we choose $\eta = 0.01, \kappa_t = 0, \boldsymbol{\gamma}_{s\to t} = \mathbf{0}$, and $\zeta_t \sim \mathcal{N}(0, \sigma_{\boldsymbol{u}}^2)$ for difference choice of $\sigma_{\boldsymbol{u}}$. We choose initial value $\hat{x}_t = 1$ and $\hat{y}_t = 0$. Trajectories are averaged over 100 runs.

Scaling with Batch Size. Figure 1 plots the sharpness gap $\Delta S = S_{\mathrm{GD}} - S_{\mathrm{SGD}}(b)$ against batch size on a log-log scale. The theory predicts $\Delta S \propto 1/b$, corresponding to a slope of -1 in log-log coordinates. We observe a fitted slope of -1.27 ($R^2 = 0.98, p < 10^{-5}$), in good agreement with the theoretical prediction slope of -1. The slightly steeper slope may reflect higher-order corrections or the fact that β and α have mild batch-size dependence not captured by the leading-order formula. The residual panel confirms that the power-law fit captures the data well, with no systematic deviations across the range of batch sizes tested.

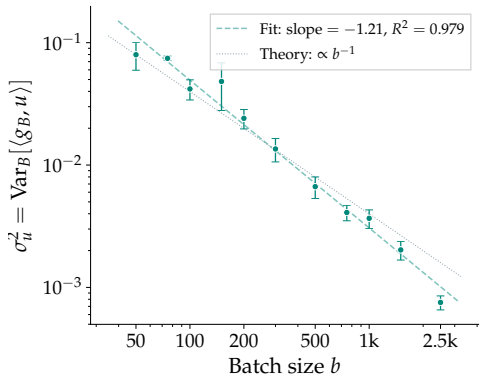

Figure 2: Gradient noise variance σ_u^2 versus batch size on a log-log scale.

Scaling with Noise Variance. The formula $\Delta S = \eta\beta\sigma_u^2/(4\alpha)$ predicts that the sharpness gap is mediated by the projected gradient noise $\sigma_{\boldsymbol{u}}^2 = \boldsymbol{u}^\top \Sigma_b \boldsymbol{u}$, where Σ_b is the covariance of the mini-batch gradient. Since mini-batch gradients are averages over b independent per-sample gradients, $\Sigma_b = \Sigma_1/b$ and hence $\sigma_{\boldsymbol{u}}^2 \propto 1/b$. Figure 2 confirms this prediction: the measured σ_u^2 scales with batch size with a fitted log-log slope of -1.21 ($R^2 = 0.979$), close to the theoretical -1. Combined with the batch-size scaling of ΔS, this establishes the causal chain: smaller batches lead to larger projected noise σ_u^2, which lead to stronger stochastic stabilization and lower equilibrium sharpness.

5.2 Simulation of Predicted Dynamic

We simulate the predicted dynamic $\hat{x}_t$ and $\hat{y}_t$ given in Lemma 4.6. To isolate the interesting factors, we simplify the dynamic to $\beta_{s\to t} = \beta, \delta_t = \delta$, and $\kappa_t = 0, \boldsymbol{\gamma}_{s\to t} = \mathbf{0}$. The simulation result is shown in Figure 3, where the red dashed line shows the approximation of the stationary condition by taking a time-average of trajectory. Each column in Figure 3 shows the dynamic of a difference $\sigma_{\mathbf{u}}^2$ value, ranging from 0 (deterministic case) to 40. The $\hat{x}_t^2, \hat{y}_t$ dynamic in the first column reproduces the behavior in Damian et al. [2023]. Observing the first row that demonstrate the $\hat{x}_t^2$ dynamic, one could see that increasing the noise suppresses the oscillation scale of $\hat{x}_t^2$, but keeping the mean $\mathbf{E}\left[\hat{x}_t^2\right]$ at $\delta^2 = 0.5$ and breaking the periodic structure. In the second row, one could observe that the mean at stationary (dashed red line) gradually shifts down as $\sigma_{\boldsymbol{u}}^2$ grows. The scale of the oscillation seems to also be suppressed. Together, Figure 3 validates the theoretical result in Theorem 2.

5.3 Slope Decomposition: Closing the Gap Between $\Delta S \propto b^{-1.27}$ and Theory

Here we account for the difference between Theorem 2's prediction $\Delta S \propto b^{-1}$ via $\sigma_{\boldsymbol{u}}^2$ and Section 5.1's measurement $\Delta S \propto b^{-1.27}$ by measuring the per-batch dependence of all three landscape quantities and combining them through the equilibrium formula.

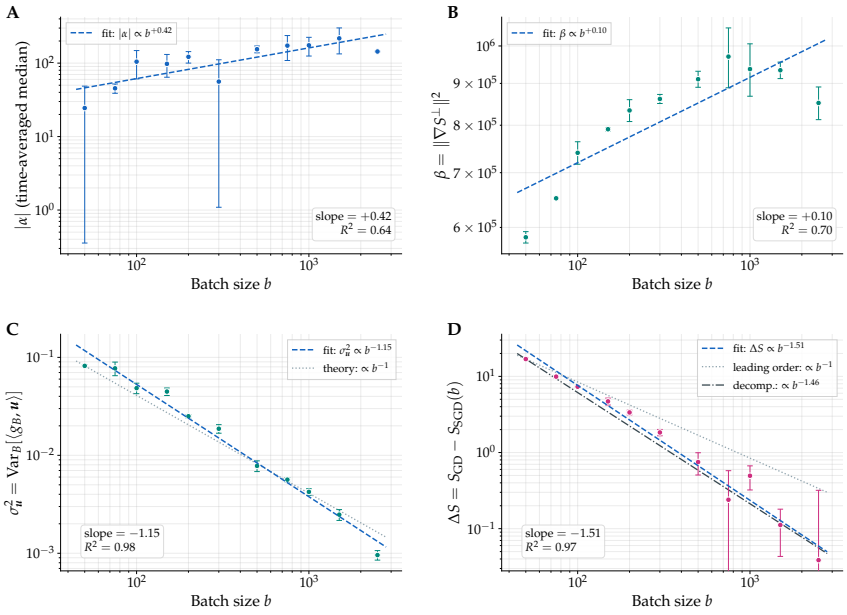

Figure 4: Slope decomposition. **(A)** $|\alpha|$; **(B)** $\beta = \|\nabla S^{\perp}\|^2$; **(C)** $\sigma_{\boldsymbol{u}}^2$; **(D)** ΔS.

From $\Delta S = \eta\beta\sigma_{\boldsymbol{u}}^2/(4\alpha)$, if each quantity has a power-law dependence on b over the swept range, namely $\alpha(b) \propto b^{a_\alpha}$, $\beta(b) \propto b^{a_\beta}$, and $\sigma_{\boldsymbol{u}}^2(b) \propto b^{a_\sigma}$, then $\Delta S(b)$ inherits a log-log slope

$$\mathrm{slope}_{\mathrm{pred}}(\Delta S) = a_\beta + a_\sigma - a_\alpha. \quad (19)$$

A perfect b^{-1} scaling requires $a_\beta = 0$, $a_\sigma = -1$, and $a_\alpha = 0$. Empirically we find none of these hold exactly.

Table 1: Log-log fits of quantities versus batch size.

Quantity	**Slope (log-log)**	$\boldsymbol{R^2}$	$\boldsymbol{n}$
ΔS (measured)	-1.510	0.970	11
$\sigma_{\boldsymbol{u}}^2$	-1.146	0.982	11
$\beta = \|\nabla S^{\perp}\|^2$	$+0.104$	0.696	11
$\lvert\alpha_{\mathrm{med}}\rvert$	$+0.418$	0.640	11
ΔS (predicted, α_{med})	-1.460	—	—
ΔS (predicted, α_{last})	-1.660	—	—

Per-batch fits. We refit each landscape quantity by taking the median of $\alpha_t, \beta_t, \sigma_{\boldsymbol{u},t}^2$ over the last 5 measurements. We report the time-averaged median fit ($a_{\alpha,\mathrm{med}} = +0.42$, $R^2 = 0.64$). The per-batch values for b ranging from 50 to 2500 and the three log-log fits appear in Table 1 and Figure 4. Substituting the time-averaged exponents into Equation (19),

$$\mathrm{slope}_{\mathrm{pred}}(\Delta S) = a_\beta + a_\sigma - a_{\alpha,\mathrm{med}} = (+0.104) + (-1.146) - (+0.418) = -1.460,$$

to be compared with the directly measured slope of -1.510. The point residual is -0.05 in slope units. This decomposition accounts for gap between theory and experiment (-1 versus -1.27) with the explanation that the main theorem captures not just the leading-order scaling but also the sub-leading corrections, once taking into consideration the dependency on the landscape quantities.

6 Conclusion and Limitations

We investigate of why SGD stabilizes at sharpness below the $2/\eta$ threshold observed in GD. By extending the framework of Damian et al. [2023] to the stochastic setting, we identified gradient noise as a sharpness-suppressing force. From the stationarity conditions of the predicted dynamics, we derived the closed-form sharpness gap $\Delta S = \eta\beta\sigma_{\boldsymbol{u}}^2/(4\alpha)$. Our experimental validation confirms the predicted power-law scaling of the sharpness gap with batch size and the noise variance.

Limitations. Our analysis assumes simplified settings such as frozen landscape coefficients, noise bounded by a small magnitude, and a mean-field decorrelation for the simplicity of the analysis and the presentation of the result. Extending the theory to handle more general, time-varying landscapes in a larger noise regime would be a natural next step. Our analysis of the EoS behavior focus on the sharpness dynamic, leaving the stability of other eigenvalues of the Hessian open to study. Our experiments reveal that EoS does not always emerge, depending on the loss function and architecture, and characterizing precisely when and why these dynamics arise is an intriguing open question.

Acknowledgments and Disclosure of Funding

This research was funded in part by Rice University, the U.S. National Science Foundation (CAREER award no. 2145629), and a Microsoft Research Award. AK also thanks the Ken Kennedy Institute at Rice University for its support through the research cluster QuanTAS. We thank the anonymous reviewers and the area chair for comments that improved this paper.

References

Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 169–195. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/agarwala23b.html.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd, 2025. URL https://arxiv.org/abs/2412.20553.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 948–1024. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

Yuhang Cai, Haodong Wen, Haofeng Huang, Deyi Liu, Yiyuan Ma, and Kaifeng Lyu. Does llm pre-training typically occur at the edge of stability?, 2026. URL https://arxiv.org/abs/2610.36778.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4330–4391. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23b.html.

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

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27449–27461. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e6af401c28c1790eaef7d55c92ab6ab6-Paper.pdf.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *ICLR*, 2023. URL https://openreview.net/forum?id=nhKHA59gXz.

Konstantinos Emmanouilidis, Lachlan MacDonald, Salma Tarmoun, and Rene Vidal. Convergence of large-stepsize sgd for the cross-entropy loss via stochastic lyapunov stability, 2026. URL https://arxiv.org/abs/2606.30930.

Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2315–2357. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/haochen21a.html.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd, 2018. URL https://arxiv.org/abs/1711.04623.

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

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1g87C4KwB.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17684–17744. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kreisler23a.html.

Sungyoon Lee and Cheongjae Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bH-kCY6LdKg.

Rotem Mulayoff and Tomer Michaeli. Exact mean square linear stability analysis for sgd, 2024. URL https://arxiv.org/abs/2306.07850.

Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rq_Qr0c1Hyo.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9983–9994. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/40bb79c081828bebdc39d65a82367246-Paper-Conference.pdf.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6651526b6fb8f29a00507de6a49ce30f-Paper.pdf.

Lei Wu, Mingze Wang, and Weijie Su. The alignment property of sgd noise and how it helps select flat minima: A stability analysis, 2022. URL https://arxiv.org/abs/2207.02628.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd, 2018. URL https://arxiv.org/abs/1802.08770.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p7EagBsMAEO.

A Cubic Taylor Expansion Details

Lemma A.1 (Cubic expansion of the mini-batch gradient)**.** *Let $\boldsymbol{\theta}^\star \in \mathcal{M}$ Let $\boldsymbol{\theta}_t = \boldsymbol{\theta}^\star + \boldsymbol{v}_t$ with $\|\boldsymbol{v}_t\| \leq R$ for some $R > 0$. Under Assumptions 2, the mini-batch gradient satisfies:*

$$\nabla L(\boldsymbol{\theta}_t) = \nabla L(\boldsymbol{\theta}^\star) + \nabla^2 L(\boldsymbol{\theta}^\star)\boldsymbol{v}_t + \frac{1}{2}\nabla^3 L(\boldsymbol{\theta}^\star)(\boldsymbol{v}_t, \boldsymbol{v}_t) + \boldsymbol{r}_t, \tag{20}$$

where $\|\boldsymbol{r}_t\| \leq O(\rho_4)\|\boldsymbol{v}_t\|^3$.

Proof. By the integral form of Taylor's theorem applied to $\boldsymbol{g}_{B_t}(\cdot) = \nabla L_{B_t}(\cdot)$ around $\boldsymbol{\theta}^\star$:

$$\nabla L(\boldsymbol{\theta}^\star + \boldsymbol{v}) = \nabla L(\boldsymbol{\theta}^\star) + \nabla^2 L(\boldsymbol{\theta}^\star)\boldsymbol{v} + \frac{1}{2}\nabla^3 L(\boldsymbol{\theta}^\star)(\boldsymbol{v},\boldsymbol{v}) + \underbrace{\int_0^1 \frac{(1-s)^2}{2}\nabla^4 L(\boldsymbol{\theta}^\star + s\boldsymbol{v})(\boldsymbol{v},\boldsymbol{v},\boldsymbol{v})\mathrm{d}s}_{\boldsymbol{r}_t}.$$

By Assumption 2, we have that $\left\|\nabla^4 L(\boldsymbol{\theta}^\star + s\boldsymbol{v})(\boldsymbol{v},\boldsymbol{v},\boldsymbol{v})\right\|_{\mathrm{op}} \le \rho_4$. Therefore,

$$\|\boldsymbol{r}_t\| \le \frac{1}{6}\sup_{s\in[0,1]}\left\|\nabla^4 L(\boldsymbol{\theta}^\star + s\boldsymbol{v}_t)\right\|_{\mathrm{op}}\|\boldsymbol{v}_t\|^3 \le \frac{\rho_4}{6}\|\boldsymbol{v}_t\|^3.$$

□

B Lipschitz Properties of Landscape Quantities

Lemma B.1 (Lipschitz Properties, Damian et al. [2023])**.** *Define $\mathcal{S}_t := B(\boldsymbol{\theta}_t^\dagger, (2-c_{\mathrm{gap}})/(4\eta\rho_3))$. Under Assumption 1,2, it holds that*

(1) $\boldsymbol{\theta} \mapsto \nabla L(\boldsymbol{\theta})$ is $O(\eta^{-1})$-Lipschitz in $\mathcal{S}_t$.

(2) $\boldsymbol{\theta} \mapsto \mathbf{H}(\boldsymbol{\theta})$ is ρ_3-Lipschitz.

(3) $\boldsymbol{\theta} \mapsto \lambda_i(\mathbf{H}(\boldsymbol{\theta}))$ is ρ_3-Lipschitz.

(4) $\boldsymbol{\theta} \mapsto \boldsymbol{u}(\boldsymbol{\theta})$ is $O(\eta\rho_3)$-Lipschitz in $\mathcal{S}_t$.

(5) $\boldsymbol{\theta} \mapsto \nabla S(\boldsymbol{\theta})$ is L_S-Lipschitz in $\mathcal{S}_t$ with $L_S := O(\eta\rho_3^2)$.

Proof. The proof is identical to Damian et al. [2023], Lemma 12. □

C Proof of Lemma 4.6

Proof. We first derive the $\hat{x}$ update, then unroll the recursion for $\hat{\boldsymbol{v}}^\perp$ to obtain the $\hat{y}$ formula.

Part A: Derivation of the $\hat{x}$ update. From Definition 4.4, project $\hat{\boldsymbol{v}}_{t+1}$ onto $\boldsymbol{u}_{t+1}$:

$$\hat{x}_{t+1} = \langle \boldsymbol{u}_{t+1}, \hat{\boldsymbol{v}}_{t+1}\rangle = \underbrace{\langle \boldsymbol{u}_{t+1}, P^\perp_{\boldsymbol{u}_{t+1}}[\cdots]\rangle}_{=0 \text{ by definition}} - (1+\eta\hat{y}_t)\hat{x}_t - \tfrac{\eta}{2}\kappa_t\hat{x}_t^2 - \eta\underbrace{\langle \boldsymbol{u}_{t+1}, \tilde{\boldsymbol{\xi}}_t\rangle}_{\tilde{\zeta}_t}.$$

The first term vanishes because $P^\perp_{\boldsymbol{u}_{t+1}}$ projects orthogonally to $\boldsymbol{u}_{t+1}$. This yields (16).

Part B: Unrolling the recursion for $\hat{\boldsymbol{v}}^\perp_{t+1}$. Define $\hat{\boldsymbol{v}}^\perp_t := P^\perp_{\boldsymbol{u}_t}\hat{\boldsymbol{v}}_t$. Then we have that

$$\hat{y}_{t+1} = \langle \nabla S^\perp_{t+1}, \hat{\boldsymbol{v}}_{t+1}\rangle = \nabla S_{t+1}^\top P^\perp_{\boldsymbol{u}_t}\hat{\boldsymbol{v}}_t = \langle \nabla S_{t+1}, \hat{\boldsymbol{v}}^\perp_{t+1}\rangle$$

By definition of $\hat{\boldsymbol{v}}_t$, we have

$$\begin{aligned}\hat{\boldsymbol{v}}^\perp_{t+1} &= P^\perp_{\boldsymbol{u}_{t+1}}\left[(\mathbf{I}-\eta\mathbf{H}_t)\hat{\boldsymbol{v}}^\perp_t + \frac{\eta}{2}\nabla S^\perp_t(\delta_t^2 - \hat{x}_t^2)\right] - P^\perp_{\boldsymbol{u}_{t+1}}\tilde{\boldsymbol{\xi}}_t \\ &= P^\perp_{\boldsymbol{u}_{t+1}}(\mathbf{I}-\eta\mathbf{H}_t)\hat{\boldsymbol{v}}^\perp_t + \frac{\eta}{2}P^\perp_{\boldsymbol{u}_{t+1}}\nabla S^\perp_t(\delta_t^2 - \hat{x}_t^2) - P^\perp_{\boldsymbol{u}_{t+1}}\tilde{\boldsymbol{\xi}}_t\end{aligned}$$

We shall prove by induction that

$$\hat{\boldsymbol{v}}^\perp_{t+1} = P^\perp_{\boldsymbol{u}_{t+1}}\prod_{k=t}^{0}\mathbf{A}_k\hat{\boldsymbol{v}}^\perp_0 + \eta P^\perp_{\boldsymbol{u}_{t+1}}\sum_{s=0}^{t}\prod_{k=t}^{s+1}\mathbf{A}_k\left(\nabla S^\perp_s\frac{\delta_s^2 - \hat{x}_s^2}{2} - P^\perp_{\boldsymbol{u}_{s+1}}\tilde{\boldsymbol{\xi}}_s\right)$$

where $\mathbf{A}_k := (I - \eta \mathbf{H}_k) P^{\perp}_{\boldsymbol{u}_k}$. For base case, it is straight forward to hold as $\hat{\boldsymbol{v}}_0^{\perp} = \hat{\boldsymbol{v}}_0^{\perp}$. Assume that $\hat{\boldsymbol{v}}_t^{\perp}$ takes the desired form. Then we have that

$$\begin{aligned}
\hat{\boldsymbol{v}}_{t+1}^{\perp} &= P^{\perp}_{\boldsymbol{u}_{t+1}} (\mathbf{I} - \eta \mathbf{H}_t) \left(P^{\perp}_{\boldsymbol{u}_t} \prod_{k=t-1}^{0} \mathbf{A}_k \hat{\boldsymbol{v}}_0^{\perp} + \eta P^{\perp}_{\boldsymbol{u}_t} \sum_{s=0}^{t-1} \prod_{k=t-1}^{s+1} \mathbf{A}_k \left(\nabla S_s^{\perp} \frac{\delta_s^2 - \hat{x}_s^2}{2} - P^{\perp}_{\boldsymbol{u}_{s+1}} \tilde{\boldsymbol{\xi}}_s \right) \right) \\
&\quad + \eta P^{\perp}_{\boldsymbol{u}_{t+1}} \sum_{s=0}^{t} \prod_{k=t}^{s+1} \mathbf{A}_k \left(\nabla S_s^{\perp} \frac{\delta_s^2 - \hat{x}_s^2}{2} - P^{\perp}_{\boldsymbol{u}_{s+1}} \tilde{\boldsymbol{\xi}}_s \right) \\
&= P_{\boldsymbol{u}_{t+1}} \prod_{k=t}^{0} \mathbf{A}_t \hat{\boldsymbol{v}}_0^{\perp} + \eta P^{\perp}_{\boldsymbol{u}_{t+1}} \sum_{s=0}^{t} \prod_{k=t}^{s+1} \mathbf{A}_k \left(\nabla S_s^{\perp} \frac{\delta_s^2 - \hat{x}_s^2}{2} - P^{\perp}_{\boldsymbol{u}_{s+1}} \tilde{\boldsymbol{\xi}}_s \right)
\end{aligned}$$

finishing the induction proof. This gives that

$$\begin{aligned}
\hat{y}_{t+1} &= \langle \nabla S_{t+1}, \hat{\boldsymbol{v}}_{t+1}^{\perp} \rangle \\
&= \nabla S_{t+1}^{\top} P^{\perp}_{\boldsymbol{u}_{t+1}} \prod_{k=t}^{0} \mathbf{A}_k \hat{\boldsymbol{v}}_0^{\perp} + \eta \nabla S_{t+1}^{\top} P^{\perp}_{\boldsymbol{u}_{t+1}} \sum_{s=0}^{t} \prod_{k=t}^{s+1} \mathbf{A}_k \nabla S_s^{\perp} \frac{\delta_s^2 - \hat{x}_s^2}{2} \\
&\quad - \eta \nabla S_{t+1}^{\top} P^{\perp}_{\boldsymbol{u}_{t+1}} \sum_{s=0}^{t} \prod_{k=t}^{s+1} \mathbf{A}_k P^{\perp}_{\boldsymbol{u}_{s+1}} \tilde{\boldsymbol{\xi}}_s \\
&= \mathcal{I}_{t+1} + \eta \sum_{s=0}^{t} \beta_{s \to t} \frac{\delta_s^2 - \hat{x}_s^2}{2} - \eta \sum_{s=0}^{t} \langle \boldsymbol{\gamma}_{s \to t}, \tilde{\boldsymbol{\xi}}_s \rangle
\end{aligned}$$

□

D Heuristic Derivation of the Predicted Dynamics

The enhanced stochastic predicted dynamics (Definition 4.4) arise from a systematic Taylor expansion of the SGD update rule around the constrained trajectory $\boldsymbol{\theta}_t^{\dagger}$, keeping all leading-order terms and dropping subleading corrections. We carry out this derivation in full.

Lemma D.1. *Under Assumptions 1-6, we have that*

$$\begin{aligned}
x_{t+1} &= -(1 + \eta y_t) x_t - \frac{1}{2} \eta \kappa_t x_t^2 - \eta \langle \boldsymbol{\xi}_t, \boldsymbol{u}_{t+1} \rangle + \langle \boldsymbol{u}_{t+1} - \boldsymbol{u}_t, \boldsymbol{v}_{t+1} \rangle + O(E_t) \\
P^{\perp}_{\boldsymbol{u}_{t+1}} \boldsymbol{v}_{t+1} &= P^{\perp}_{\boldsymbol{u}_{t+1}} \left[(\mathbf{I} - \eta \mathbf{H}) P^{\perp}_{\boldsymbol{u}_t} \boldsymbol{v}_t + \eta \nabla S_t^{\perp} \frac{\delta_t^2 - x_t^2}{2} \right] - \eta P^{\perp}_{\boldsymbol{u}_{t+1}} \boldsymbol{\xi}_t + O(E_t)
\end{aligned}$$

where $E_t = \frac{\epsilon^2}{\delta} \|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2} \|\boldsymbol{v}_t\|^3 + \epsilon^3 \delta$.

Proof. Recall that the SGD update is $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{g}_{B_t}(\boldsymbol{\theta}_t)$. Writing $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^{\dagger} + \boldsymbol{v}_t$ and decomposing $\boldsymbol{g}_{B_t}(\boldsymbol{\theta}_t) = \nabla L(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_t$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla L(\boldsymbol{\theta}_t) - \eta \boldsymbol{\xi}_t. \tag{21}$$

The constrained trajectory satisfies (Lemma 2.1):

$$\boldsymbol{\theta}_{t+1}^{\dagger} = \boldsymbol{\theta}_t^{\dagger} - \eta P^{\perp}_{\boldsymbol{u}_t, \nabla S_t} \nabla L_t + O(\epsilon^2 \eta \|\nabla L_t\|).$$

Subtracting the two gives:

$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - \eta \left[\nabla L(\boldsymbol{\theta}_t^{\dagger} + \boldsymbol{v}_t) - P^{\perp}_{\boldsymbol{u}_t, \nabla S_t} \nabla L_t \right] - \eta \boldsymbol{\xi}_t + O(\epsilon^2 \eta \|\nabla L_t\|). \tag{22}$$

To expand the full-batch gradient, we invoke Lemma A.1:

$$\nabla L(\boldsymbol{\theta}_t^{\dagger} + \boldsymbol{v}_t) = \nabla L_t + \mathbf{H}_t \boldsymbol{v}_t + \tfrac{1}{2} \nabla^3 L_t(\boldsymbol{v}_t, \boldsymbol{v}_t) + O(\rho_4 \|\boldsymbol{v}_t\|^3).$$

Plugging into (22) gives

$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - \eta \left[\nabla L_t - P^{\perp}_{\boldsymbol{u}_t, \nabla S_t} \nabla L_t \right] - \mathbf{H}_t \boldsymbol{v}_t - \tfrac{1}{2} \nabla^3 L_t(\boldsymbol{v}_t, \boldsymbol{v}_t) - \eta \boldsymbol{\xi}_t + O(\eta \rho_4 \|\boldsymbol{v}_t\|^3 + \epsilon^2 \eta \|\nabla L_t\|). \tag{23}$$

The projected gradient satisfies $\nabla L_t - P^{\perp}_{\boldsymbol{u}_t,\nabla S_t}\nabla L_t = P_{\boldsymbol{u}_t,\nabla S_t}\nabla L_t$. Since $\langle \boldsymbol{u}_t, \nabla L_t\rangle = 0$ on $\mathcal{M}$, this projection lies entirely in the $\nabla S_t^{\perp}$ direction. Therefore,

$$\nabla L_t - P^{\perp}_{\boldsymbol{u}_t,\nabla S_t}\nabla L_t = P_{\boldsymbol{u}_t,\nabla S_t}\nabla L_t = \frac{\langle \nabla L_t, \nabla S_t^{\perp}\rangle}{\left\|\nabla S_t^{\perp}\right\|^2}\nabla S_t^{\perp} = -\frac{\alpha_t}{\beta_t}\nabla S_t^{\perp} = -\frac{\delta_t^2}{2}\nabla S_t^{\perp}, \tag{24}$$

where we used $\alpha_t = -\langle \nabla L_t, \nabla S_t^{\perp}\rangle$ (since $\langle \boldsymbol{u}_t, \nabla L_t\rangle = 0$), $\beta_t = \left\|\nabla S_t^{\perp}\right\|^2$, and $\delta_t^2 = 2\alpha_t/\beta_t$. To handle the cubic term, we decompose $\boldsymbol{v}_t = x_t\boldsymbol{u}_t + \boldsymbol{v}_t^{\perp}$ where $\boldsymbol{v}_t^{\perp} = P^{\perp}_{\boldsymbol{u}_t}\boldsymbol{v}_t$. The cubic term expands bilinearly (using symmetry of $\nabla^3 L_t$):

$$\tfrac{1}{2}\nabla^3 L_t(\boldsymbol{v}_t, \boldsymbol{v}_t) = \tfrac{1}{2}x_t^2 \cdot \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{u}_t) + x_t \cdot \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^{\perp}) + \tfrac{1}{2}\nabla^3 L_t(\boldsymbol{v}_t^{\perp}, \boldsymbol{v}_t^{\perp}). \tag{25}$$

By Assumption 2,4, and noticing that $|x_t|, \left\|\boldsymbol{v}_t^{\perp}\right\| \le \|\boldsymbol{v}_t\|$ we have that $\nabla^3 L_t(\boldsymbol{v}_t^{\perp}, \boldsymbol{v}_t^{\perp}) \le O(\epsilon\rho_3 \|\boldsymbol{v}_t\|^2)$. Moreover, since $\nabla S_t = \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{u}_t)$, we have

$$\frac{1}{2}x_t^2 \cdot \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{u}_t) = \frac{x_t^2}{2}\nabla S_t = \frac{x_t^2}{2}\nabla S_t^{\perp} + \frac{x_t^2}{2}\langle \nabla S_t, \boldsymbol{u}_t\rangle \boldsymbol{u}_t = \frac{x_t^2}{2}\nabla S_t^{\perp} + \frac{\kappa_t}{2}x_t^2 \cdot \boldsymbol{u}_t$$

Combining gives

$$\begin{aligned}\boldsymbol{v}_{t+1} = (\mathbf{I} - \eta\mathbf{H}_t)\boldsymbol{v}_t + \eta \cdot \frac{\delta_t^2 - x_t^2}{2}\nabla S_t^{\perp} - \frac{1}{2}\eta\kappa_t x_t^2 \cdot \boldsymbol{u}_t - \eta x_t \cdot \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^{\perp}) - \eta\boldsymbol{\xi}_t \\ + O(\eta\epsilon\rho_3 \|\boldsymbol{v}_t\|^2 + \rho_4 \|\boldsymbol{v}_t\|^3 + \epsilon^2\eta \|\nabla L_t\|)\end{aligned}$$

Notice that, by Assumption 3 and the definition that $\beta_t = \|\nabla S_t\|^2 = \Theta(\rho_3^2)$, we have

$$\alpha_t = \Theta\left(\|\nabla L_t\| \left\|\nabla S_t^{\perp}\right\|\right) = \Theta(\rho_3 \|\nabla L_t\|) \Rightarrow \eta \|\nabla L_t\| = \Theta\left(\frac{\eta\alpha_t}{\rho_3}\right) = \Theta\left(\eta\sqrt{\alpha_t} \cdot \frac{\sqrt{\alpha_t}}{\sqrt{\beta_t}}\right) = \Theta(\epsilon\delta)$$

Also, we have that

$$\eta\rho_3\delta = \Theta\left(\eta\sqrt{\beta_t} \cdot \frac{\sqrt{\alpha_t}}{\sqrt{\beta_t}}\right) = \Theta(\eta\sqrt{\alpha_t}) = \Theta(\epsilon)$$

Therefore, combining with Assumption 6, we have that

$$\begin{aligned}\boldsymbol{v}_{t+1} = (\mathbf{I} - \eta\mathbf{H}_t)\boldsymbol{v}_t + \eta \cdot \frac{\delta_t^2 - x_t^2}{2}\nabla S_t^{\perp} - \frac{1}{2}\eta\kappa_t x_t^2 \cdot \boldsymbol{u}_t - \eta x_t \cdot \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^{\perp}) - \eta\boldsymbol{\xi}_t \\ + O\left(\frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta\right)\end{aligned}$$

We will track the projection of $\boldsymbol{v}_{t+1}$ onto $\boldsymbol{u}_{t+1}$ and $P^{\perp}_{\boldsymbol{u}_{t+1}}$. To start, we have that

$$\begin{aligned}x_{t+1} &= \langle \boldsymbol{v}_{t+1}, \boldsymbol{u}_{t+1}\rangle \\ &= \langle \boldsymbol{v}_{t+1}, \boldsymbol{u}_t\rangle + \langle \boldsymbol{u}_{t+1} - \boldsymbol{u}_t, \boldsymbol{v}_{t+1}\rangle \\ &= \boldsymbol{u}_t^{\top}(\mathbf{I} - \eta H_t)\boldsymbol{v}_t - \frac{1}{2}\eta\kappa_t x_t^2 - \eta x_t\langle \nabla S_t, \boldsymbol{v}_t^{\perp}\rangle - \eta\langle \boldsymbol{\xi}_t, \boldsymbol{u}_{t+1}\rangle \\ &\qquad + \langle \boldsymbol{u}_{t+1} - \boldsymbol{u}_t, \boldsymbol{v}_{t+1}\rangle + O\left(\frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta\right) \\ &= -(1 + \eta y_t)x_t - \frac{1}{2}\eta\kappa_t x_t^2 - \eta\langle \boldsymbol{\xi}_t, \boldsymbol{u}_{t+1}\rangle + \langle \boldsymbol{u}_{t+1} - \boldsymbol{u}_t, \boldsymbol{v}_{t+1}\rangle \\ &\qquad + O\left(\frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta\right)\end{aligned}$$

Projecting onto $P^{\perp}_{\boldsymbol{u}_{t+1}}$ gives

$$\begin{aligned}P^{\perp}_{\boldsymbol{u}_{t+1}}\boldsymbol{v}_{t+1} = P^{\perp}_{\boldsymbol{u}_{t+1}}\left[(\mathbf{I} - \eta\mathbf{H})\boldsymbol{v}_t + \eta\nabla S_t^{\perp}\frac{\delta_t^2 - x_t^2}{2}\right] - \eta x_t P^{\perp}_{\boldsymbol{u}_{t+1}}\nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^{\perp}) - \eta P^{\perp}_{\boldsymbol{u}_{t+1}}\boldsymbol{\xi}_t \\ + O\left(\frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta\right)\end{aligned}$$

We have that
$\eta x_t P^\perp_{\boldsymbol{u}_{t+1}} \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^\perp) = O\left(\eta|x_t| \left\|P^\perp_{\boldsymbol{u}_t} \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^\perp)\right\|\right) + O\left(\eta|x_t| \left\|P_{\boldsymbol{u}_{t+1}} - P_{\boldsymbol{u}_t}\right\| \left\|\nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^\perp)\right\|\right)$
where, by Assumption 4,

$$\left\|P^\perp_{\boldsymbol{u}_t} \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^\perp)\right\| \le O(\epsilon\rho_3 \|\boldsymbol{v}_t\|)$$

and by Lemma B.1

$$\left\|P_{\boldsymbol{u}_{t+1}} - P_{\boldsymbol{u}_t}\right\| \le O(\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|) \le O(\eta^2\rho_3 \|\nabla L_t\|) \le O(\epsilon^2)$$

Thus, we have that

$$\left\|\eta x_t P^\perp_{\boldsymbol{u}_{t+1}} \nabla^3 L_t(\boldsymbol{u}_t, \boldsymbol{v}_t^\perp)\right\| \le O(\eta\epsilon\rho_3 \|\boldsymbol{v}_t\|^2 + \eta\epsilon^2\rho_3 \|\boldsymbol{v}_t\|^2) \le O\left(\epsilon^2 \frac{\|\boldsymbol{v}_t\|^2}{\delta}\right)$$

Therefore, the projection of $\boldsymbol{v}_{t+1}$ onto $P_{\boldsymbol{u}_{t+1}}$ is given by

$$\begin{aligned} P^\perp_{\boldsymbol{u}_{t+1}} \boldsymbol{v}_{t+1} &= P^\perp_{\boldsymbol{u}_{t+1}} \left[(\mathbf{I} - \eta\mathbf{H})\boldsymbol{v}_t + \eta\nabla S_t^\perp \frac{\delta_t^2 - x_t^2}{2}\right] - \eta P^\perp_{\boldsymbol{u}_{t+1}} \boldsymbol{\xi}_t + O\left(\frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta\right) \\ &= P^\perp_{\boldsymbol{u}_{t+1}} \left[(\mathbf{I} - \eta\mathbf{H})P^\perp_{\boldsymbol{u}_t}\boldsymbol{v}_t + \eta\nabla S_t^\perp \frac{\delta_t^2 - x_t^2}{2}\right] - \eta P^\perp_{\boldsymbol{u}_{t+1}} \boldsymbol{\xi}_t + O\left(\frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta\right) \end{aligned}$$

where the last equality follows from Assumption 1. □

E Complete Proof of the Stochastic Coupling Theorem

The proof follows a four-lemma structure: (i) local growth of the enhanced step map; (ii) one-step local truncation error; (iii) mean-square boundedness; (iv) mean-square coupling.

E.1 Local Growth of the Enhanced Step Map

Lemma E.1 (Local Growth Bound)**.** *Let Assumption 1-6 hold. If $\|\boldsymbol{v}\| \le R\delta$ for some R satisfying $1 \le R \le O(\epsilon^{-1})$, then*

$$\left\|\widetilde{\texttt{step}}_t(\boldsymbol{v})\right\|^2 \le (1 + O(\epsilon R))\|\boldsymbol{v}\|^2 + O(\epsilon R^3\delta^2). \tag{26}$$

Proof. Let $x = \langle \boldsymbol{u}_t, \boldsymbol{v}\rangle$, $y = \langle \nabla S_t^\perp, \boldsymbol{v}\rangle$. To start, we bound the $\boldsymbol{u}_{t+1}$-component. By definition of the enhanced step map:

$$\langle \boldsymbol{u}_{t+1}, \widetilde{\texttt{step}}_t(\boldsymbol{v})\rangle = -(1 + \eta y)x - \tfrac{\eta}{2}\kappa_t x^2.$$

We bound each factor. Since $\eta|y| \le \eta\left\|\nabla S_t^\perp\right\|\|\boldsymbol{v}\| \le \eta\rho_3 R\delta$ and $\eta\rho_3\delta = O(\epsilon)$, we have $\eta|y| = O(\epsilon R)$. Similarly, $\frac{\eta}{2}|\kappa_t||x| \le \frac{\eta}{2}\rho_3 R\delta = O(\epsilon R)$. Therefore,

$$|\langle \boldsymbol{u}_{t+1}, \widetilde{\texttt{step}}_t(\boldsymbol{v})\rangle| \le (1 + O(\epsilon R))|x| \tag{27}$$

Squaring: $\langle \boldsymbol{u}_{t+1}, \widetilde{\texttt{step}}_t(\boldsymbol{v})\rangle^2 \le (1 + O(\epsilon R))^2|x|^2 \le (1 + O(\epsilon R))|x|^2$ since $R \le O(\epsilon^{-1})$. Next, we bound the $P^\perp_{\boldsymbol{u}_{t+1}}$-component.

$$P^\perp_{\boldsymbol{u}_{t+1}} \widetilde{\texttt{step}}_t(\boldsymbol{v}) = = P^\perp_{\boldsymbol{u}_{t+1}} \left[(I - \eta\mathbf{H}_t)P^\perp_{\boldsymbol{u}_t}\boldsymbol{v} + \eta\nabla S_t^\perp \tfrac{\delta_t^2 - x^2}{2}\right].$$

Notice that for all $\boldsymbol{v}^\perp \perp \boldsymbol{u}_t$, we have that

$$\left\|(I - \eta\mathbf{H}_t)\boldsymbol{v}^\perp\right\| \le \max_{i\ge 2}|1 - \lambda_i(\mathbf{H}_t)|\left\|\boldsymbol{v}^\perp\right\|$$

By Assumption 1 and Assumption 4, we obtain that

$$\max_{i\ge 2}|1 - \lambda_i(\mathbf{H}_t)| \le 1 + O(\epsilon) \Rightarrow \|(I - \eta\mathbf{H}_t)P_{\boldsymbol{u}_t}\boldsymbol{v}\| \le (1 + O(\epsilon))\|\boldsymbol{v}\|$$

To bound the term $\eta\nabla S_t^\perp \frac{\delta_t^2 - x^2}{2}$, we notice that $\left\|\nabla S_t^\perp\right\| \le \rho_3$ and $|\delta_t^2 - x_t^2| \le (1 + R^2)\delta^2$. Therefore,

$$\left\|\eta\nabla S_t^\perp \tfrac{\delta_t^2 - x^2}{2}\right\| \le O((1 + R^2)\eta\rho_3\delta^2) \le O((1 + R^2)\epsilon\delta)$$

Therefore,

$$\left\|P^\perp_{\boldsymbol{u}_{t+1}} \widetilde{\texttt{step}}_t(\boldsymbol{v})\right\|^2 \le (1 + O(\epsilon R))\|\boldsymbol{v}\|^2 + O(\epsilon R^3\delta^2)$$

which implies that

$$\left\|\widetilde{\texttt{step}}_t(\boldsymbol{v})\right\|^2 \le (1 + O(\epsilon R))\|\boldsymbol{v}\|^2 + O(\epsilon R^3\delta^2)$$

□

E.2 One-Step Local Truncation Error

Lemma E.2 (One-Step Local Truncation Error)**.** *Let Assumption 1-6 hold. Define* $\boldsymbol{\varepsilon}_t = \boldsymbol{v}_{t+1} - \widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t)$*. If* $\|\boldsymbol{v}_t\| \le R\delta$ *for some* R *satisfying* $1 \le R \le O(\epsilon^{-1})$*, then we have that*

$$\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big] \le O(\epsilon^4\eta^2)\mathbf{E}\Big[\|\boldsymbol{\xi}_t\|^2 \mid \mathcal{F}_t\Big] + O\left(\frac{\epsilon^4}{\delta^2}\|\boldsymbol{v}_t\|^4 + \frac{\epsilon^4}{\delta^6}\|\boldsymbol{v}_t\|^6 + \epsilon^4\|\boldsymbol{v}_t\|^2 + \epsilon^5 R^3\delta^2\right)$$

Proof. Recall that, by definition, $\widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t) = \widetilde{\texttt{step}}_t(\boldsymbol{v}_t) - \eta\boldsymbol{\xi}_t s$ with

$$P^\perp_{\boldsymbol{u}_{t+1}}\widetilde{\texttt{step}}_t(\boldsymbol{v}_t) = P^\perp_{\boldsymbol{u}_{t+1}}\left[\mathbf{A}_t\boldsymbol{v}_t + \eta\nabla S_t^\perp\frac{\delta_t^2 - x_t^2}{2}\right]$$

$$\langle \boldsymbol{u}_{t+1}, \widetilde{\texttt{step}}_t(\boldsymbol{v}_t)\rangle = -(1+\eta y_t)x_t - \frac{\eta}{2}\kappa_t x_t^2$$

Thus, let $\zeta_t = \langle \boldsymbol{u}_{t+1}, \boldsymbol{\xi}_t\rangle$, we have

$$P^\perp_{\boldsymbol{u}_{t+1}}\widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t) = P^\perp_{\boldsymbol{u}_{t+1}}\left[\mathbf{A}_t\boldsymbol{v}_t + \eta\nabla S_t^\perp\frac{\delta_t^2 - x_t^2}{2}\right] - \eta P^\perp_{\boldsymbol{u}_{t+1}}\boldsymbol{\xi}$$

$$\langle \boldsymbol{u}_{t+1}, \widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t)\rangle = -(1+\eta y_t)x_t - \frac{\eta}{2}\kappa_t x_t^2 - \eta\zeta_t$$

By Lemma D.1, we have that

$$P^\perp_{\boldsymbol{u}_{t+1}}\boldsymbol{v}_{t+1} = P^\perp_{\boldsymbol{u}_{t+1}}\left[\mathbf{A}_t\boldsymbol{v}_t + \eta\nabla S_t^\perp\frac{\delta_t^2 - x_t^2}{2}\right] - \eta P^\perp_{\boldsymbol{u}_{t+1}}\boldsymbol{\xi}_t + O(E_t)$$

$$\langle \boldsymbol{u}_{t+1}, \boldsymbol{v}_{t+1}\rangle = -(1+\eta y_t)x_t - \frac{1}{2}\eta\kappa_t x_t^2 - \eta\zeta_t + \langle \boldsymbol{u}_{t+1} - \boldsymbol{u}_t, \boldsymbol{v}_{t+1}\rangle + O(E_t)$$

where $E_t = \frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta$. Therefore $\left\|P^\perp_{\boldsymbol{u}_{t+1}}\boldsymbol{v}_{t+1} - P^\perp_{\boldsymbol{u}_{t+1}}\widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t)\right\| = O(E_t)$ and

$$\langle \boldsymbol{u}_{t+1}, \boldsymbol{v}_{t+1} - \widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t)\rangle = \langle \boldsymbol{u}_{t+1} - \boldsymbol{u}_t, \boldsymbol{v}_{t+1}\rangle + O(E_t)$$

This gives that

$$\boldsymbol{\varepsilon}_t = \langle \boldsymbol{u}_{t+1} - \boldsymbol{u}_t, \boldsymbol{v}_{t+1}\rangle\boldsymbol{u}_{t+1} + O\left(\frac{\epsilon^2}{\delta}\|\boldsymbol{v}_t\|^2 + \frac{\epsilon^2}{\delta^2}\|\boldsymbol{v}_t\|^3 + \epsilon^3\delta\right)$$

which implies that

$$\begin{aligned}\|\boldsymbol{\varepsilon}_t\|^2 &\le 2\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|^2\|\boldsymbol{v}_{t+1}\|^2 + O\left(\frac{\epsilon^4}{\delta^2}\|\boldsymbol{v}_t\|^4 + \frac{\epsilon^4}{\delta^4}\|\boldsymbol{v}_t\|^6 + \epsilon^6\delta^2\right)\\ &\le O(\epsilon^4)\|\boldsymbol{v}_{t+1}\|^2 + O\left(\frac{\epsilon^4}{\delta^2}\|\boldsymbol{v}_t\|^4 + \frac{\epsilon^4}{\delta^4}\|\boldsymbol{v}_t\|^6 + \epsilon^6\delta^2\right)\end{aligned}$$

Next we bound $\|\boldsymbol{v}_{t+1}\|$. We notice that

$$\boldsymbol{v}_{t+1} = \widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t) + \boldsymbol{\varepsilon}_t = \widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t) + \boldsymbol{\varepsilon}_t = \widetilde{\texttt{step}}_t(\boldsymbol{v}_t) - \eta\boldsymbol{\xi}_t + \boldsymbol{\varepsilon}_t$$

Therefore,

$$\begin{aligned}\mathbf{E}\Big[\|\boldsymbol{v}_{t+1}\|^2 \mid \mathcal{F}_t\Big] &= 2\mathbf{E}\Big[\|\widetilde{\texttt{step}}_t(\boldsymbol{v}_t)\|^2 \mid \mathcal{F}_t\Big] + 2\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big] + \eta^2\mathbf{E}\Big[\|\boldsymbol{\xi}_t\|^2 \mid \mathcal{F}_t\Big]\\ &\le O(1)\|\boldsymbol{v}_t\|^2 + O(\epsilon R^3\delta^2) + 2\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big] + \eta^2\mathbf{E}\Big[\|\boldsymbol{\xi}_t\|^2 \mid \mathcal{F}_t\Big]\end{aligned}$$

Plugging back into the bound on $\boldsymbol{\varepsilon}_t$ gives

$$\begin{aligned}\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big] \le{}& O(\epsilon^4)\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big] + O(\epsilon^4\eta^2)\mathbf{E}\Big[\|\boldsymbol{\xi}_t\|^2 \mid \mathcal{F}_t\Big]\\ &O\left(\frac{\epsilon^4}{\delta^2}\|\boldsymbol{v}_t\|^4 + \frac{\epsilon^4}{\delta^6}\|\boldsymbol{v}_t\|^6 + \epsilon^4\|\boldsymbol{v}_t\|^2 + \epsilon^5 R^3\delta^2\right)\end{aligned}$$

With the assumption that $\epsilon \le o(1)$ gives

$$\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big] \le O(\epsilon^4\eta^2)\mathbf{E}\Big[\|\boldsymbol{\xi}_t\|^2 \mid \mathcal{F}_t\Big] + O\left(\frac{\epsilon^4}{\delta^2}\|\boldsymbol{v}_t\|^4 + \frac{\epsilon^4}{\delta^6}\|\boldsymbol{v}_t\|^6 + \epsilon^4\|\boldsymbol{v}_t\|^2 + \epsilon^5 R^3\delta^2\right)$$

□

E.3 Mean-Square Boundedness

Lemma E.3 (Mean-Square Boundedness of the Enhanced Predicted Dynamics)**.** *Let Assumption 1-6 hold. If* $\|\boldsymbol{v}_t\| \le R\delta$ *for some* R *satisfying* $1 \le R \le O(\epsilon^{-1})$*, then*

$$\mathbf{E}\Big[\,\|\hat{\boldsymbol{v}}_{t+1}\|^2 \mid \mathcal{F}_t\Big] \le (1 + O(\epsilon R))\,\|\hat{\boldsymbol{v}}_t\|^2 + O(\epsilon R^3\delta^2). \tag{28}$$

Proof. By definition, $\hat{\boldsymbol{v}}_{t+1} = \widetilde{\texttt{step}}_t^S(\hat{\boldsymbol{v}}_t) = \widetilde{\texttt{step}}_t(\hat{\boldsymbol{v}}_t) - \eta\boldsymbol{\xi}_t$. Since $\mathbf{E}[\boldsymbol{\xi}_t \mid \mathcal{F}_t] = \mathbf{0}$, the cross-term vanishes when squaring:

$$\mathbf{E}\Big[\,\|\hat{\boldsymbol{v}}_{t+1}\|^2 \mid \mathcal{F}_t\Big] = \mathbf{E}\Big[\,\big\|\widetilde{\texttt{step}}_t(\hat{\boldsymbol{v}}_t)\big\|^2\Big] + \eta^2\mathbf{E}\Big[\,\|\boldsymbol{\xi}_t\|^2\Big]. \tag{29}$$

Taking full expectations and applying Lemma E.1:

$$\mathbf{E}\Big[\,\|\hat{\boldsymbol{v}}_{t+1}\|^2 \mid \mathcal{F}_t\Big] \le (1 + O(\epsilon R))\,\|\hat{\boldsymbol{v}}_t\|^2 + O(\epsilon R^3\delta^2 + \eta^2\sigma^2).$$

where we also applied that $\mathbf{E}\Big[\,\|\boldsymbol{\xi}_t\|^2\Big] \le \sigma^2$. By the scaling condition in Assumption 6, $\eta^2\sigma^2 \le O(\epsilon\delta^2)$. Therefore:

$$\mathbf{E}\Big[\,\|\hat{\boldsymbol{v}}_{t+1}\|^2 \mid \mathcal{F}_t\Big] \le (1 + O(\epsilon R))\,\|\hat{\boldsymbol{v}}_t\|^2 + O(\epsilon R^3\delta^2). \tag{30}$$

□

E.4 Lipschitz Bound of Enhanced Step Map

Lemma E.4 (Lipschitz Bound of Enhanced Step Map)**.** *Let Assumption 1-6 hold. For* $\boldsymbol{v}, \boldsymbol{w}$ *satisfying* $\|\boldsymbol{v}\|, \|\boldsymbol{w}\| \le R\delta$ *for some* $1 \le R \le O(\epsilon^{-1})$*, we have that*

$$\Big\|\widetilde{\texttt{step}}_t^S(\boldsymbol{v}) - \widetilde{\texttt{step}}_t^S(\boldsymbol{w})\Big\| \le (1 + O(\epsilon R))\,\|\boldsymbol{v} - \boldsymbol{w}\|$$

Proof. By definition, we have that

$$\begin{aligned}
\widetilde{\texttt{step}}_t^S(\boldsymbol{v}) - \widetilde{\texttt{step}}_t^S(\boldsymbol{w}) &= \widetilde{\texttt{step}}_t(\boldsymbol{v}) - \widetilde{\texttt{step}}_t(\boldsymbol{w}) \\
&= P^{\perp}_{\boldsymbol{u}_{t+1}}\mathbf{A}_t(\boldsymbol{v} - \boldsymbol{w}) + \frac{\eta}{2}P^{\perp}_{\boldsymbol{u}_{t+1}}\nabla S_t^{\perp}(x_v^2 - x_w^2) \\
&\qquad - ((1 + \eta y_v)x_v - (1 + \eta y_w)x_w)\,\boldsymbol{u}_{t+1} - \frac{\eta}{2}\kappa_t(x_v^2 - x_w^2)\boldsymbol{u}_{t+1} \\
&= \mathbf{A}_t(\boldsymbol{v} - \boldsymbol{w}) + \frac{\eta}{2}\nabla S_t^{\perp}(x_v^2 - x_w^2) - \eta\,(y_v x_v - y_w x_w)\,\boldsymbol{u}_{t+1}
\end{aligned}$$

where $x_v = \langle \boldsymbol{u}_t, \boldsymbol{v}\rangle, x_w = \langle \boldsymbol{u}_t, \boldsymbol{w}\rangle$ and $y_v = \langle \nabla S_t^{\perp}, \boldsymbol{v}\rangle, y_w = \langle \nabla S_t^{\perp}, \boldsymbol{w}\rangle$. Therefore

$$\Big\|\widetilde{\texttt{step}}_t^S(\boldsymbol{v}) - \widetilde{\texttt{step}}_t^S(\boldsymbol{w})\Big\| \le \underbrace{\|\mathbf{A}_t\|\,\|\boldsymbol{v} - \boldsymbol{w}\|}_{T_1} + \underbrace{\frac{\eta}{2}\,\|\nabla S_t\|\,\big|x_v^2 - x_w^2\big|}_{T_2} + \underbrace{|y_v x_v - y_w x_w|}_{T_3}.$$

Next, we will analyze each term individually. To start, for T_1 we have that

$$\|\mathbf{A}_t\| \le \|\mathbf{I} - \eta\mathbf{H}_t\| \le 1 + \eta\lambda_{\min}(\mathbf{H}_t) \le 1 + O(\eta\epsilon)\,\|\mathbf{H}_t\| \le 1 + O(\epsilon)$$

Therefore, $T_1 \le (1 + O(\epsilon))\,\|\boldsymbol{v} - \boldsymbol{w}\|$. For T_2, we have that $\|\nabla S_t\| \le O(\rho_3)$ and

$$\big|x_v^2 - x_w^2\big| \le |x_v + x_w| \cdot |x_v - x_w| \le (\|\boldsymbol{v}\| + \|\boldsymbol{w}\|)\,\|\boldsymbol{v} - \boldsymbol{w}\| \le O(R\delta)\,\|\boldsymbol{v} - \boldsymbol{w}\|$$

Therefore, $T_2 \le O(\eta\rho_3\delta R)\,\|\boldsymbol{v} - \boldsymbol{w}\| \le O(\epsilon R)\,\|\boldsymbol{v} - \boldsymbol{w}\|$. Lastly, for T_3 we have that

$$\begin{aligned}
T_3 &\le \eta\,|y_v|\,|x_v - x_w| + \eta|x_w||y_v - y_w| \\
&\le O(\eta\rho_3\delta R)\,\|\boldsymbol{v} - \boldsymbol{w}\| + \eta\delta \cdot O(\rho_3)\,\|\boldsymbol{v} - \boldsymbol{w}\| \\
&\le O(\eta\rho_3\delta R)\,\|\boldsymbol{v} - \boldsymbol{w}\| \\
&\le O(\epsilon R)\,\|\boldsymbol{v} - \boldsymbol{w}\|
\end{aligned}$$

where in the second inequality we used the definition $y_v = \langle \nabla S_t^{\perp}, \boldsymbol{v}\rangle, y_w = \langle \nabla S_t^{\perp}, \boldsymbol{w}\rangle$ and the fact that $\big\|\nabla S_t^{\perp}\big\| \le O(\rho_3)$. Combining all three terms gives the desired result. □

E.5 Mean-Square Coupling

Lemma E.5 (Mean-Square Coupling)**.** *Let Assumption 1-6 hold with a small enough constant $c_{\mathscr{T}}$. If $\|\boldsymbol{v}_0\|^2 \le c_0\delta^2$ for some small enough constant c_0, then for any small constant ω such that $\omega^{-1} \le O(1)$, we have that with proability at least $1-\omega$, the following holds for all $t \le \mathscr{T}$*

$$\|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\| \le O(\epsilon\delta); \quad \|\boldsymbol{v}_t\| \le O(\delta); \quad \|\hat{\boldsymbol{v}}_t\| \le O(\delta)$$

Proof. Define $\boldsymbol{e}_t := \boldsymbol{v}_t - \hat{\boldsymbol{v}}_t$. Recall that in Lemma E.2, we defined $\boldsymbol{\varepsilon}_t = \boldsymbol{v}_{t+1} - \widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t)$. Then we have that

$$\boldsymbol{e}_{t+1} = \boldsymbol{v}_{t+1} - \hat{\boldsymbol{v}}_{t+1} = \widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t) - \widetilde{\texttt{step}})_t^S(\hat{\boldsymbol{v}}_t) + \boldsymbol{\varepsilon}_t$$

This gives that

$$\|\boldsymbol{e}_{t+1}\| \le \left\|\widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t) - \widetilde{\texttt{step}}_t^S(\hat{\boldsymbol{v}}_t)\right\| + \|\boldsymbol{\varepsilon}_t\|$$

By Lemma E.4, the Lipschitzness of $\widetilde{\texttt{step}}_t^S(\cdot)$ can be bounded as

$$\left\|\widetilde{\texttt{step}}_t^S(\boldsymbol{v}_t) - \widetilde{\texttt{step}}_t^S(\hat{\boldsymbol{v}}_t)\right\| \le (1+O(\epsilon R))\|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\| = (1+O(\epsilon R))\|\boldsymbol{e}_t\|$$

Therefore, we have that

$$\mathbf{E}\Big[\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_t\Big] \le (1+O(\epsilon R))\|\boldsymbol{e}_t\|^2 + O(\epsilon^{-1}R^{-1})\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big]$$

Define $\tau := \inf\{t \le \mathscr{T} : \|\boldsymbol{v}_t\| > M\delta \text{ or } \|\hat{\boldsymbol{v}}_t\| > M\delta\}$ for a constant $M \ge 1$ to be chosen. Then for all $t < \tau$, we have that

$$\begin{aligned} O(\epsilon^{-1}R^{-1})\mathbf{E}\Big[\|\boldsymbol{\varepsilon}_t\|^2 \mid \mathcal{F}_t\Big] &\le O(\epsilon^{-1}M^{-1}) \cdot O(\sigma^2\eta^2\epsilon^4 + \epsilon^4\delta^2M^4 + \epsilon^4M^6 + \epsilon^4\delta^2M^2 + \epsilon^5M^3\delta^2) \\ &\le O(M^3\epsilon^3(\delta^2+M^2)) \end{aligned}$$

This gives that when $t < \tau$

$$\mathbf{E}\Big[\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_t\Big] \le (1+O(M\epsilon))\|\boldsymbol{e}_t\|^2 + O(M^3\epsilon^3(\delta^2+M^2))$$

In the meantime, by Lemma E.3 we have that

$$\mathbf{E}\Big[\|\hat{\boldsymbol{v}}_{t+1}\|^2 \mid \mathcal{F}_t\Big] \le (1+O(M\epsilon))\|\hat{\boldsymbol{v}}_t\|^2 + O(M^3\epsilon\delta^2) + O(M^3\epsilon\delta^2)$$

Therefore, there exists constant C_1, C_2, C_3 such that

$$\begin{aligned} \mathbf{E}\Big[\|\boldsymbol{e}_{t+1}\|^2 \mid \mathcal{F}_t\Big] &\le (1+C_1M\epsilon)\|\boldsymbol{e}_t\|^2 + C_1C_2M^3\epsilon^3(\delta^2+M^2) \\ \mathbf{E}\Big[\|\hat{\boldsymbol{v}}_{t+1}\|^2 \mid \mathcal{F}_t\Big] &\le \|\hat{\boldsymbol{v}}_t\|^2 + C_3M^3\epsilon\delta^2 \end{aligned}$$

Define F_1, G_1 as

$$F_t^{(1)} = \|\boldsymbol{e}_t\|^2 + C_2M^2\epsilon^2(\delta^2+M^2); \quad G_t^{(1)} = (1+C_1M\epsilon)^{-t}F_t^{(1)}$$

Then it holds that

$$\begin{aligned} \mathbf{E}\Big[F_{t+1}^{(1)} \mid \mathcal{F}_t\Big] &\le (1+C_1M\epsilon)\|\boldsymbol{e}_t\|^2 + C_1C_2M^3\epsilon^3(\delta^2+M^2) + C_2M^2\epsilon^2(\delta^2+M^2) \\ &\le (1+C_1M\epsilon)\left(\|\boldsymbol{e}_t\|^2 + C_2M^2\epsilon^2(\delta^2+M^2)\right). \end{aligned}$$

Thus, $\mathbf{E}\Big[G_{t+1}^{(1)} \mid \mathcal{F}_t\Big] \le G_t^{(1)}$, which implies that $G_t^{(1)}$ is a supermartingale. Similarly, we define

$$G_t^{(2)} = \|\hat{\boldsymbol{v}}_t\|^2 - C_3M^3\epsilon\delta^2 t.$$

Then we have that

$$\begin{aligned} \mathbf{E}\Big[G_{t+1}^{(2)}\Big] &= \mathbf{E}\Big[\|\hat{\boldsymbol{v}}_{t+1}\|^2\Big] - C_3M^3\epsilon\delta^2(t+1) \\ &= \|\hat{\boldsymbol{v}}_t\|^2 + C_3M^3\epsilon\delta - C_3M^3\epsilon\delta^2(t+1) \\ &= \|\hat{\boldsymbol{v}}_t\|^2 - C_3M^3\epsilon\delta^2 t \\ &= G_t^{(2)} \end{aligned}$$

which shows that $G_t^{(2)}$ is also a supermartingale. Applying Doob's maximal inequality gives that

$$\mathbf{Pr}\left[\max_{t\le\mathscr{T}} G_t^{(1)} \ge B_1\right] \le B_1^{-1}\mathbf{E}\left[G_0^{(1)}\right]; \quad \mathbf{Pr}\left[\max_{t\le\mathscr{T}} G_t^{(2)} \ge B_2\right] \le B_2^{-1}\mathbf{E}\left[G_0^{(2)}\right]$$

By definition,

$$\mathbf{E}\left[G_0^{(1)}\right] = \mathbf{E}\left[F_0^{(1)}\right] = \mathbf{E}\left[\|\boldsymbol{e}_0\|^2\right] + C_2M^2\epsilon^2(\delta^2+M^2) = C_2M^2\epsilon^2(\delta^2+M^2)$$
$$\mathbf{E}\left[G_0^{(1)}\right] = \mathbf{E}\left[\|\hat{\boldsymbol{v}}_0\|^2\right] = \|\boldsymbol{v}_0\|^2$$

Let $\omega < \frac{1}{4}$ be a small constant. Choose $B_1 = \omega^{-1}C_2M^2\epsilon^2(\delta^2+M^2)$ and $B_2 = \omega^{-1}\|\boldsymbol{v}_0\|^2$. Then with probability at least $1-2\omega$ we have that

$$G_t^{(1)} \le \omega^{-1}C_2M^2\epsilon^2(\delta^2+M^2); \quad G_t^{(2)} \le \omega^{-1}\|\boldsymbol{v}_0\|^2$$

which, under the assumption that $t \le \mathscr{T} \le {}^{c_{\mathscr{T}}}/_{\epsilon}$, implies that

$$\begin{aligned}\|\boldsymbol{e}_t\|^2 &\le (1+C_1M\epsilon)^t\omega^{-1}C_2M^2\epsilon^2(\delta^2+M^2)\\ &\le \omega^{-1}\exp(C_1M\epsilon t)C_2M^2\epsilon^2(\delta^2+M^2)\\ &\le \omega^{-1}\exp(c_{\mathscr{T}}C_1M)C_2M^2\epsilon^2(\delta^2+M^2)\\ &\le C_4^2M^2\epsilon^2(\delta^2+M^2)\end{aligned}$$

for some constant C_4. Also, under the assumption that $\|\boldsymbol{v}_0\|^2 \le c_0\delta^2$,

$$\|\hat{\boldsymbol{v}}_t\|^2 \le \omega^{-1}\|\boldsymbol{v}_0\|^2 + C_3M\epsilon\delta^2 t \le \omega^{-1}\|\boldsymbol{v}_0\|^2 + c_{\mathscr{T}}C_3M^3\delta^2 \le ({}^{c_0}/_{\omega} + c_{\mathscr{T}}C_3M^3)\delta^2 \le \delta^2$$

for small enough $c_{\mathscr{T}}$ and c_0 Altogether, this implies that

$$\|\boldsymbol{v}_t\| \le \|\hat{\boldsymbol{v}}_t\| + \|\boldsymbol{e}_t\| \le \delta + C_4M\epsilon(\delta+M) \le 2\delta$$

by taking $M=3$ and $\epsilon \le o(1)$. This gives that for all $t \le \mathscr{T}$, $t \le \tau$. Thus, for all $t \le \mathscr{T}$, we have that

$$\|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\| \le O(\epsilon\delta); \quad \|\boldsymbol{v}_t\| \le O(\delta); \quad \|\hat{\boldsymbol{v}}_t\| \le O(\delta)$$

□

E.6 Sharpness Approximation

Lemma E.6 (Sharpness Approximation)**.** *Under Assumptions 1-6, if* $\|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\| \le O(\epsilon\delta)$ *and* $\|\hat{\boldsymbol{v}}_t\| \le O(\delta)$*, then for any* $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^\dagger + \boldsymbol{v}_t$ *we have*

$$S(\boldsymbol{\theta}_t) = \frac{2}{\eta} + \hat{y}_t + \kappa_t\hat{x}_t + O\left(\frac{\epsilon^2}{\eta}\right) \tag{31}$$

Proof. Recall that $\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^\dagger + \boldsymbol{v}_t$. Applying Taylor expansion to $S(\boldsymbol{\theta}_t)$ gives that

$$S(\boldsymbol{\theta}_t) = S(\boldsymbol{\theta}_t^\dagger) + \langle\nabla S(\boldsymbol{\theta}_t^\dagger), \boldsymbol{v}_t\rangle + \int_0^1 \lambda\boldsymbol{v}_t^\top\nabla^2 S(\boldsymbol{\theta}_t^\dagger + \lambda\boldsymbol{v}_t)\boldsymbol{v}_t d\lambda$$

Where the last term can be bounded as

$$\left|\int_0^1 \lambda\boldsymbol{v}_t^\top\nabla^2 f(\boldsymbol{\theta}_t^\dagger + \lambda\boldsymbol{v}_t)\boldsymbol{v}_t d\lambda\right| \le \left\|\nabla^2 S(\boldsymbol{\theta}_t^\dagger + \lambda\boldsymbol{v}_t)\right\|\|\boldsymbol{v}_t\|^2 \le O(\rho_4)\|\boldsymbol{v}_t\|^2$$

Since $\|\boldsymbol{v}_t\| \le O(\delta)$ and $\rho_4 \le O(\eta\rho_3^2)$ by Assumption 6, we have that

$$\left|\int_0^1 \lambda\boldsymbol{v}_t^\top\nabla^2 f(\boldsymbol{\theta}_t^\dagger + \lambda\boldsymbol{v}_t)\boldsymbol{v}_t d\lambda\right| \le O(\eta\rho_3^2\delta^2) \le O(\eta^{-1}\epsilon^2)$$

Next, we decompose $\langle\nabla S(\boldsymbol{\theta}_t^\dagger), \boldsymbol{v}_t\rangle = \langle\nabla S_t, \boldsymbol{v}_t\rangle$. We write

$$\nabla S_t = \langle\nabla S_t, \boldsymbol{u}_t\rangle\boldsymbol{u}_t + \nabla S_t^\perp = \kappa_t\boldsymbol{u}_t + \nabla S_t^\perp$$

Therefore

$$\begin{aligned}\langle \nabla S_t, \boldsymbol{v}_t \rangle &= \langle \nabla S_t, \hat{\boldsymbol{v}}_t \rangle + \langle \nabla S_t, \boldsymbol{v}_t - \hat{\boldsymbol{v}}_t \rangle \\ &= \kappa_t \langle \boldsymbol{u}_t, \hat{\boldsymbol{v}}_t \rangle + \langle \nabla S_t^{\perp}, \hat{\boldsymbol{v}}_t \rangle + O(\|\nabla S_t\| \, \|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\|) \\ &= \kappa_t \hat{x}_t + \hat{y}_t + O(\rho_3 \epsilon \delta) \\ &= \kappa_t \hat{x}_t + \hat{y}_t + O(\eta^{-1}\epsilon^2)\end{aligned}$$

where we used $\|\boldsymbol{v} - \hat{\boldsymbol{v}}_t\| \le O(\epsilon\delta)$. Noticing that $S(\boldsymbol{\theta}_t) = \frac{2}{\eta}$ by Lemma 2.1, and combining the results gives that

$$S(\boldsymbol{\theta}_t) = \frac{2}{\eta} + \kappa_t \hat{x}_t + \hat{y}_t + O(\eta^{-1}\epsilon^2)$$

□

E.7 Proof of Theorem 1

Proof of Theorem 1. Under the assumptions of the theorem, Lemma E.5 implies that $t \le \mathscr{T}$

$$\|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\| \le O(\epsilon\delta); \quad \|\boldsymbol{v}_t\| \le O(\delta); \quad \|\hat{\boldsymbol{v}}_t\| \le O(\delta)$$

with probability at least $1 - \omega$. Assume that such event happens. Then by Lemma E.6 we have that

$$S(\boldsymbol{\theta}_t) = \frac{2}{\eta} + \hat{y}_t + \kappa_t \hat{x}_t + O\left(\frac{\epsilon^2}{\eta}\right)$$

It suffice to show the loss approximation. Perform Taylor expansion for the loss gives

$$L(\boldsymbol{\theta}_t) = L(\boldsymbol{\theta}_t^{\dagger}) + \langle \nabla L(\boldsymbol{\theta}_t^{\dagger}), \boldsymbol{v}_t \rangle + \frac{1}{2} \boldsymbol{v}^{\top} \nabla^2 L(\boldsymbol{\theta}_t^{\dagger}) \boldsymbol{v}_t + O(\rho_3 \|\boldsymbol{v}_t\|^2)$$

Recall that $\alpha_t = -\langle \nabla L_t, \nabla S_t \rangle = \Theta(\|\nabla L_t\| \, \|\nabla S_t^{\perp}\|)$ by Assumption 3. Therefore, $\|\nabla L_t\| \le O\left(\alpha_t / \|\nabla S_t^{\perp}\|\right) \le O(\alpha_t/\rho_3) \le O(\epsilon\delta/\eta)$. This gives that $\langle \nabla L(\boldsymbol{\theta}_t), \boldsymbol{v}_t \rangle \le O(\epsilon\delta^2/\eta)$. For the second-order term, we have that

$$\begin{aligned}\boldsymbol{v}_t^{\top} \nabla^2 L(\boldsymbol{\theta}_t^{\dagger}) \boldsymbol{v}_t &= \hat{\boldsymbol{v}}_t^{\top} \nabla^2 L(\boldsymbol{\theta}_t^{\dagger}) \hat{\boldsymbol{v}}_t + O\left(\left\|\nabla^2 L(\boldsymbol{\theta}_t^{\dagger})\right\| \|\boldsymbol{v}_t\| \, \|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\|\right) \\ &= \hat{\boldsymbol{v}}_t^{\top} \nabla^2 L(\boldsymbol{\theta}_t^{\dagger}) \hat{\boldsymbol{v}}_t + O\left(\frac{\epsilon\delta^2}{\eta}\right) \\ &= \left(\hat{\boldsymbol{v}}_t^{\perp} + \langle \hat{\boldsymbol{v}}_t, \boldsymbol{u}_t \rangle \boldsymbol{u}_t\right)^{\top} \nabla^2 L(\boldsymbol{\theta}_t^{\dagger}) \left(\hat{\boldsymbol{v}}_t^{\perp} + \langle \hat{\boldsymbol{v}}_t, \boldsymbol{u}_t \rangle \boldsymbol{u}_t\right) + O\left(\frac{\epsilon\delta^2}{\eta}\right) \\ &= \frac{2}{\eta} \hat{x}_t^2 + \hat{\boldsymbol{v}}_t^{\top} \nabla^2 L(\boldsymbol{\theta}_t^{\dagger}) \hat{\boldsymbol{v}}_t + O\left(\frac{\epsilon\delta^2}{\eta}\right) \\ &= \frac{2}{\eta} \hat{x}_t^2 + O\left(\frac{\epsilon\delta^2}{\eta}\right)\end{aligned}$$

where the last step follows from Assumption 4. Putting things together, and noticing that $O(\rho_3 \|\boldsymbol{v}_t\|^2) = O(\rho_3 \delta^2) = O(\epsilon\delta/\eta)$, we have that

$$L(\boldsymbol{\theta}_t) = L(\boldsymbol{\theta}_t^{\dagger}) + \frac{\hat{x}_t^2}{\eta} + O\left(\frac{\epsilon\delta^2}{\eta}\right)$$

□

F Proof of the Sharpness Gap Theorem

Proof of Theorem 2. Let the condition in Theorem 1 hold. Then we have

$$\|\boldsymbol{v}_t - \hat{\boldsymbol{v}}_t\| \le O(\delta), \qquad \|\boldsymbol{v}_t\| \le O(\delta), \qquad \|\hat{\boldsymbol{v}}_t\| \le O(\delta) \tag{32}$$

hold simultaneously for all $t \le \mathscr{T}$. This gives that $|\hat{x}_t| \le \|\hat{\boldsymbol{v}}_t\| \le O(\delta)$ and $|\hat{y}_t| \le \|\nabla S_t\| \, \|\hat{\boldsymbol{v}}_t\| \le O(\rho_3\delta)$. This also implies that $\eta|\hat{y}_t| \le O(\eta\rho_3\delta) = O(\epsilon)$ and $\eta|\kappa| \le \eta\epsilon \|\nabla S_t\| \le O(\eta\epsilon\rho_3)$. With

the assumption that all landscape parameters are frozen (Assumption 8), we shall start the proof. The $\hat{y}$ formula from Lemma 4.6 is:

$$\hat{y}_{t+1} = \mathcal{I}_{t+1} + \eta \sum_{s=0}^{t} \beta_{s\to t} \frac{\delta_s^2 - \hat{x}_s^2}{2} - \eta \sum_{s=0}^{t} \langle \boldsymbol{\gamma}_{s\to t}, \boldsymbol{\xi}_s \rangle \tag{33}$$

$$\mathcal{I}_{t+1} = (\nabla S_{t+1}^{\perp})^\top [\prod_{k=t}^{0} \mathbf{A}_k] \hat{\boldsymbol{v}}_0^{\perp}; \quad \beta_{s\to t} = (\nabla S_{t+1}^{\perp})^\top \left[\prod_{k=t}^{s+1} \mathbf{A}_k\right] \nabla S_s^{\perp} \tag{34}$$

where $\mathbf{A}_k = (\mathbf{I} - \eta \mathbf{H}_k) P_{\boldsymbol{u}_k}^{\perp}$. By Assumption 8, we have that $\mathbf{H}_k = \mathbf{H}$ and $\boldsymbol{u}_k = \boldsymbol{u}$. This gives that $\mathbf{A}_k = \mathbf{I} - \eta \mathbf{H} \mid_{\boldsymbol{u}^\perp}$. Since $\nabla S_t^{\perp} = \nabla S^{\perp} \in \ker(\mathbf{H} \mid_{\boldsymbol{u}^\perp})$, we have that $\mathbf{A}_k \nabla S_t^{\perp} = \nabla S_t^{\perp} = \nabla S^{\perp}$. Therefore, $\mathcal{I}_{t+1} = \langle \nabla S^{\perp}, \hat{\boldsymbol{v}}_0^{\perp} \rangle$. Similarly, we also have that $\beta_{s\to t} = (\nabla S^{\perp})^\top \mathbf{A}^{t-s} \nabla S^{\perp} = (\nabla S^{\perp})^\top \nabla S^{\perp} = \left\| \nabla S^{\perp} \right\|^2 = \beta$ for all s, t. Therefore, under Assumption 8, the dynamic of $\hat{y}_t$ and $\hat{x}_t$ boils down to

$$\hat{y}_{t+1} = \langle \nabla S^{\perp}, \hat{\boldsymbol{v}}_0^{\perp} \rangle + \frac{\eta\beta}{2} \sum_{s=0}^{t} (\delta - \hat{x}_s^2) - \eta \sum_{s=0}^{t} \boldsymbol{\gamma}_{s\to t} \cdot \boldsymbol{\xi}_s$$

$$\hat{x}_{t+1} = -(1 + \eta \hat{y}_t) \hat{x}_t - \frac{\eta}{2} \kappa \hat{x}_t^2 - \eta \zeta_t$$

Taking expectation gives

$$\mathbf{E}[\hat{y}_{t+1}] = \mathbf{E}[\hat{y}_t] + \frac{\eta\beta}{2} (\delta - \mathbf{E}\big[\hat{x}_t^2\big])$$

$$\mathbf{E}[\hat{x}_{t+1}] = \mathbf{E}\left[\left((1 + \eta \hat{y}_t) \hat{x}_t - \frac{\eta}{2} \kappa \hat{x}_t^2 - \eta \zeta_t \right)^2 \right]$$

since the gradient noise has zero-mean. Setting $\mathbf{E}[\hat{y}_{t+1}] = \mathbf{E}[\hat{y}_t]$ gives

$$\mathbf{E}\big[\hat{x}_t^2\big] = \delta^2 = \frac{2\alpha}{\beta}. \tag{35}$$

From the $\hat{x}$-update (16) under frozen coefficients:

$$\hat{x}_{t+1} = -(1 + \eta \hat{y}_t) \hat{x}_t - \tfrac{\eta}{2} \kappa \hat{x}_t^2 - \eta \zeta_t.$$

Define $D_t := (1 + \eta \hat{y}_t) \hat{x}_t + \frac{\eta}{2} \kappa \hat{x}_t^2$, so that $\hat{x}_{t+1} = -D_t - \eta \zeta_t$. Squaring:

$$\hat{x}_{t+1}^2 = D_t^2 + 2\eta \zeta_t D_t + \eta^2 \zeta_t^2. \tag{36}$$

D_t is $\mathcal{F}_t$-measurable (it depends on $\hat{x}_t, \hat{y}_t$, which are functions of $B_0, \ldots, B_{t-1}$), while ζ_t depends on B_t, which is independent of $\mathcal{F}_t$. Taking conditional expectations:

$$\begin{aligned} \mathbf{E}\big[\hat{x}_{t+1}^2 \mid \mathcal{F}_t\big] &= \mathbf{E}\big[D_t^2 \mid \mathcal{F}_t\big] + \mathbf{E}[2\eta \zeta_t D_t \mid \mathcal{F}_t] + \mathbf{E}\big[\eta^2 \zeta_t^2 \mid \mathcal{F}_t\big] \\ &= D_t^2 + 2\eta D_t \underbrace{\mathbf{E}[\zeta_t \mid \mathcal{F}_t]}_{=0} + \eta^2 \mathbf{E}\big[\zeta_t^2 \mid \mathcal{F}_t\big] \\ &\overset{8(\text{iii})}{=} D_t^2 + \eta^2 \sigma_{\boldsymbol{u}}^2 + O(\epsilon^2 \delta^2). \end{aligned} \tag{37}$$

Now taking full expectations and applying the property $\mathbf{E}\big[\mathbf{E}\big[\hat{x}_{t+1}^2 \mid \mathcal{F}_t\big]\big] = \mathbf{E}\big[\hat{x}_{t+1}^2\big]$:

$$\begin{aligned} \mathbf{E}\big[\hat{x}_{t+1}^2\big] &= \mathbf{E}\big[D_t^2 + \eta^2 \sigma_{\boldsymbol{u}}^2 + O(\epsilon^2 \delta^2)\big] \\ &= \mathbf{E}\big[D_t^2\big] + \eta^2 \sigma_{\boldsymbol{u}}^2 + O(\epsilon^2 \delta^2). \end{aligned} \tag{38}$$

At stationarity, $\mathbf{E}\big[\hat{x}_{t+1}^2\big] = \mathbf{E}\big[\hat{x}_t^2\big] = \delta^2$, so:

$$\delta^2 = \mathbf{E}\big[D_t^2\big] + \eta^2 \sigma_{\boldsymbol{u}}^2 + O(\epsilon^2 \delta^2). \tag{39}$$

We now show that $\mathbf{E}\big[D_t^2\big] = \mathbf{E}\big[(1 + \eta \hat{y}_t)^2 \hat{x}_t^2\big] + O(\epsilon^2 \delta^2)$. Expanding D_t^2:

$$D_t^2 = \big((1 + \eta \hat{y}_t) \hat{x}_t + \tfrac{\eta}{2} \kappa \hat{x}_t^2\big)^2 = (1 + \eta \hat{y}_t)^2 \hat{x}_t^2 + \eta\kappa (1 + \eta \hat{y}_t) \hat{x}_t^3 + \frac{\eta^2 \kappa^2}{4} \hat{x}_t^4$$

Due to the upper bound on $\hat{y}_t, \hat{x}_t$ and κ, we have that

$$\frac{\eta^2\kappa^2}{4}\hat{x}_t^4 \leq O(\eta^2\epsilon^2\rho_3^2\delta^4) \leq O(\epsilon^4\delta^2) \leq O(\epsilon^2\delta^2) \tag{40}$$

$$\left|\eta\kappa(1+\eta\hat{y}_t)\hat{x}_t^3\right| \leq O(\eta\cdot\epsilon\rho_3\cdot(1+\epsilon)\cdot\delta^3) \leq O(\eta\epsilon\rho_3\delta^3) \leq O(\epsilon^2\delta^2) \tag{41}$$

This shows that $\mathbf{E}\big[D_t^2\big] = \mathbf{E}\big[(1+\eta\hat{y}_t)^2\hat{x}_t^2\big] + O(\epsilon^2\delta^2)$. Substituting into (39):

$$\delta^2 = \mathbf{E}\big[(1+\eta\hat{y}_t)^2\hat{x}_t^2\big] + \eta^2\sigma_{\boldsymbol{u}}^2 + O(\epsilon^2\delta^2). \tag{42}$$

By Assumption 7(a) and, at equilibrium, $\mathbf{E}\big[\hat{x}_t^2\big] = \delta^2$, we have that

$$\mathbf{E}\big[(1+\eta\hat{y}_t)^2\hat{x}_t^2\big] = (1+\eta\mathbf{E}[\hat{y}_t])^2\delta^2 + O(\epsilon^2\delta^2). \tag{43}$$

Substituting (43) into (42):

$$\delta^2 = (1+\eta\mathbf{E}[\hat{y}_t])^2\delta^2 + \eta^2\sigma_{\boldsymbol{u}}^2 + O(\epsilon^2\delta^2) \tag{44}$$

$$= \delta^2 + 2\eta\mathbf{E}[\hat{y}_t]\delta^2 + \eta^2\mathbf{E}[\hat{y}_t]^2\delta^2 + \eta^2\sigma_{\boldsymbol{u}}^2 + O(\epsilon^2\delta^2) \tag{45}$$

$$= (1+2\eta\mathbf{E}[\hat{y}_t])\delta^2 + \eta^2\sigma_{\boldsymbol{u}}^2 + O(\epsilon^2\delta^2) \tag{46}$$

where we used the fact that $|\eta\mathbf{E}[\hat{y}_t]| \leq O(\epsilon)$. Therefore, solving for $\mathbf{E}[\hat{y}_t]$ gives that, at equilibrium,

$$\mathbf{E}[\hat{y}_t] = -\frac{\eta\sigma_{\boldsymbol{u}}^2}{2\delta^2} + O\left(\frac{\epsilon^2}{\eta}\right) = -\frac{\eta\beta\sigma_{\boldsymbol{u}}^2}{4\alpha} + O\left(\frac{\epsilon^2}{\eta}\right)$$

In the meantime, recall that $|\kappa_t| \leq O(\epsilon\rho_3)$ and $|\hat{x}_t| \leq \delta$. Therefore $|\kappa_t\hat{x}_t| \leq O(\epsilon\rho_3\delta) \leq O\left(\frac{\epsilon^2}{\eta}\right)$. By Theorem 1, we have that at equilibrium

$$\mathbf{E}[S(\boldsymbol{\theta}_t)] = \frac{2}{\eta} + \mathbf{E}[\hat{y}_t] + O\left(\frac{\epsilon^2}{\eta}\right) = \frac{2}{\eta} - \frac{\eta\beta\sigma_{\boldsymbol{u}}^2}{4\alpha} + O\left(\frac{\epsilon^2}{\eta}\right)$$

□

G Batch Size Scaling

Lemma G.1 (Variance of Mini-Batch Gradient Noise)**.** *Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a fixed training set of n samples, and let $\ell_i : \mathbb{R}^d \to \mathbb{R}$ denote the per-sample loss at sample i. Define the full-batch gradient ∇L and the mini-batch gradient*

$$\nabla L(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell_i \in \mathbb{R}^d; \quad \boldsymbol{g}_B = \frac{1}{b}\sum_{i\in B}\nabla\ell_i$$

where $B \subset [n] := \{1, \ldots, n\}$ is a mini-batch drawn uniformly at random and independently of $\boldsymbol{\theta}$ for $1 \leq b \leq n$. Let $\boldsymbol{u} \in \mathbb{R}^d$ be a fixed vector, and let $\zeta := \langle\nabla L(\boldsymbol{\theta}) - \boldsymbol{g}_B(\boldsymbol{\theta}),\, \boldsymbol{u}\rangle$. Then $\mathbf{E}[\zeta] = 0$ and

$$\mathbf{E}\big[\zeta^2\big] = \frac{\sigma_z^2}{b}\cdot\frac{n-b}{n-1}; \quad \sigma_z^2 := \frac{1}{n}\sum_{i=1}^{n}\big(\langle\nabla\ell_i,\, \boldsymbol{u}\rangle - \langle\nabla L,\, \boldsymbol{u}\rangle\big)^2 \tag{47}$$

Here σ_z^2 is the population variance of the scalar projections $z_i := \langle\nabla\ell_i, \boldsymbol{u}\rangle$.

Proof. Let $\bar{z} := \frac{1}{n}\sum_{i=1}^{n} z_i = \langle\nabla L, \boldsymbol{u}\rangle$. By linearity of the inner product,

$$\zeta = \langle\nabla L - \boldsymbol{g}_B,\, \boldsymbol{u}\rangle = \langle\nabla L,\, \boldsymbol{u}\rangle - \langle\boldsymbol{g}_B,\, \boldsymbol{u}\rangle = \bar{z} - \frac{1}{b}\sum_{i\in B} z_i.$$

Let $X_i := \mathbf{1}[i \in B]$ for $i \in [n]$. Then $\frac{1}{b}\sum_{i\in B} z_i = \frac{1}{b}\sum_{i=1}^{n} X_i z_i$. Since $\mathbf{E}[X_i] = \mathbf{Pr}[i \in B] = b/n$, we have that

$$\mathbf{E}\left[\frac{1}{b}\sum_{i\in B} z_i\right] = \frac{1}{b}\sum_{i=1}^{n} z_i\,\mathbf{E}[X_i] = \frac{1}{b}\sum_{i=1}^{n} z_i\cdot\frac{b}{n} = \bar{z}.$$

Therefore $\mathbf{E}[\zeta] = 0$, and $\mathbf{E}\left[\zeta^2\right] = \mathrm{Var}\left(\frac{1}{b}\sum_{i\in B} z_i\right)$. Moreover, we have that

$$\mathrm{Var}\left(\frac{1}{b}\sum_{i\in B} z_i\right) = \frac{1}{b^2}\left[\sum_{i=1}^{n} z_i^2 \,\mathrm{Var}(X_i) + \sum_{i\neq j} z_i z_j \,\mathrm{Cov}(X_i, X_j)\right]. \tag{48}$$

Since $X_i \sim \mathrm{Bern}(b/n)$, we have that $\mathrm{Var}(X_i) = b(n-b)/n^2$. For $i \neq j$, the joint inclusion probability is

$$\mathbb{P}(i \in B,\, j \in B) = \frac{\binom{n-2}{b-2}}{\binom{n}{b}} = \frac{b(b-1)}{n(n-1)}$$

Therefore

$$\mathrm{Cov}(X_i, X_j) = \frac{b(b-1)}{n(n-1)} - \frac{b^2}{n^2} = -\frac{b(n-b)}{n^2(n-1)}.$$

Plugging into (48) gives

$$\mathrm{Var}\left(\frac{1}{b}\sum_{i\in B} z_i\right) = \frac{n-b}{bn^2(n-1)}\left[(n-1)\sum_{i=1}^{n} z_i^2 - \sum_{i\neq j}^{n} z_i z_j\right].$$

Using $\sum_{i\neq j}^{n} z_i z_j = n^2\bar{z}^2 - \sum_{i=1}^{n} z_i^2$, we get that

$$(n-1)\sum_{i=1}^{n} z_i^2 - \sum_{\substack{i,j=1\\ i\neq j}}^{n} z_i z_j = n\sum_{i=1}^{n} z_i^2 - n^2\bar{z}^2 = n\sum_{i=1}^{n}(z_i - \bar{z})^2 = n^2\sigma_z^2.$$

Therefore

$$\mathbf{E}\left[\zeta^2\right] = \mathrm{Var}\left(\frac{1}{b}\sum_{i\in B} z_i\right) = \frac{n-b}{bn^2(n-1)}\cdot n^2\sigma_z^2 = \frac{\sigma_z^2}{b}\cdot\frac{n-b}{n-1},$$

which completes the proof. □

H Further Experiment Details and Results

H.1 Setup

Architectures and losses. Our primary experimental setup follows Cohen et al. [2021]: a fully-connected network with two hidden layers of 200 units each and $\tanh$ activation (FC-Tanh), trained with mean squared error (MSE) loss on CIFAR-10. This setup produces clean edge-of-stability behavior with a well-separated top Hessian eigenvalue. To verify generality across architectures, we also test a convolutional network (CNN: two convolutional layers with $\tanh$ activation followed by two fully-connected layers) trained with MSE loss, as well as a fully-connected network with ReLU activation (FC-ReLU) with MSE loss. We additionally test the CNN with cross-entropy loss to examine the role of the loss function (see Section H.8).

Training protocol. We train on a random subset of 5,000 CIFAR-10 training images using vanilla SGD (no momentum, no weight decay) with a fixed learning rate η. The default learning rate is $\eta = 0.01$, giving a stability threshold of $2/\eta = 200$. For the learning rate scaling experiment, we vary $\eta \in \{0.005, 0.008, 0.01, 0.015, 0.02\}$. Each experiment is repeated across 3–5 random seeds for statistical significance.

Metrics. At regular intervals during training, we compute:

- The top Hessian eigenvalue $\lambda_{\max}(\nabla^2 L)$ (sharpness) via Lanczos with Hessian-vector products;
- The projected gradient noise variance $\sigma_u^2 = \mathrm{Var}_B[\langle g_B, u\rangle]$ estimated from 50 random mini-batches;
- The batch sharpness $\mathrm{BS}(\theta) = \mathbb{E}_B[g_B^\top H_B g_B/\|g_B\|^2]$ estimated from 30 random mini-batches;
- Quantities $\alpha = -\langle\nabla L, \nabla S\rangle$ and $\beta = \|\nabla S^\perp\|^2$ via the sharpness gradient $\nabla S = \nabla^3 L(u,u)$ measured along the SGD trajectory as approximation of the corresponding quantities along the constrained trajectory.

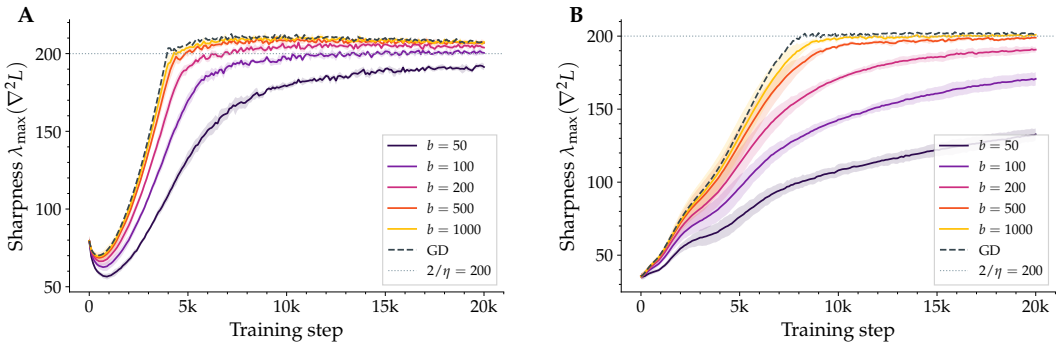

Figure 5: Sharpness $\lambda_{\max}(\nabla^2 L)$ during training for different batch sizes. **(A)** FC-Tanh + MSE: smaller batch sizes produce markedly lower equilibrium sharpness. The dashed black curve shows full-batch GD. **(B)** CNN + MSE: the same pattern holds with a convolutional architecture, with an even larger gap ($S_{\text{eq}} \approx 131$ at $b = 50$ vs. $S_{\text{GD}} \approx 202$). Shaded regions indicate ± 1 standard deviation across 5 seeds.

We define the *equilibrium sharpness* S_{eq} as the average of $\lambda_{\max}$ over the last 20 recorded measurements, after the sharpness trajectory has stabilized near the edge of stability.

Hardware. All experiments are run on NVIDIA RTX 6000 Ada GPUs (48 GB). The total computational budget is approximately 20 GPU-hours across 172 individual training runs.

H.2 Sharpness Trajectories

Figure 5 shows the sharpness trajectories for representative batch sizes. All methods exhibit progressive sharpening: the top eigenvalue increases monotonically during early training before reaching a plateau. Full-batch GD equilibrates at $S_{\text{GD}} \approx 207.8$, slightly above $2/\eta = 200$ due to the oscillation asymmetry identified by Damian et al. [2023]. Critically, SGD with smaller batch sizes equilibrates at *strictly lower* sharpness: $S_{\text{eq}} \approx 190.8$ for $b = 50$ versus $S_{\text{eq}} \approx 207.4$ for $b = 1000$. This 17-point gap between $b = 50$ and GD is the stochastic self-stabilization effect predicted by our theory.

The CNN architecture with MSE loss (Figure 5B) exhibits the same qualitative pattern with a dramatically amplified effect: $S_{\text{eq}} \approx 131$ at $b = 50$ versus $S_{\text{GD}} \approx 202$, yielding a gap of $\Delta S \approx 71$. This is possibly due to the stronger gradient noise of CNN along the top Hessian eigenvector.

H.3 Equilibrium Sharpness Across Conditions

Figure 9A provides a comprehensive view of the equilibrium sharpness across all batch sizes tested for the FC-Tanh architecture. The monotonic increase from $S_{\text{eq}} \approx 190.8$ at $b = 50$ to $S_{\text{eq}} \approx 207.9$ at $b = 2500$ (matching the GD baseline within noise) demonstrates that the stochastic self-stabilization effect is robust and spans a wide range of batch sizes.

Figure 9B shows the equilibrium sharpness for both GD and SGD ($b = 200$) across five learning rates. Both GD and SGD sharpness scale with $2/\eta$ as expected, with a consistent gap between GD and SGD at each learning rate. Interestingly, the absolute gap ΔS is larger at smaller η (4.27 at $\eta = 0.005$ versus 2.29 at $\eta = 0.02$). This occurs because the landscape quantities α and β depend on η through the equilibrium point: at smaller η, the system equilibrates at higher sharpness where the landscape geometry may amplify the noise effect.

H.4 Batch Sharpness Saturation

In our theory, we predict that the *batch sharpness*—the curvature in the direction of the mini-batch gradient—approaches $2/\eta$ even as the full-batch sharpness $\lambda_{\max}(\nabla^2 L)$ remains suppressed below S_{GD}. This resolves an apparent paradox: how can SGD remain at the "edge of stability" if its sharpness is below $2/\eta$? The answer is that the *relevant* curvature for mini-batch dynamics is the batch sharpness, not the full-batch sharpness.

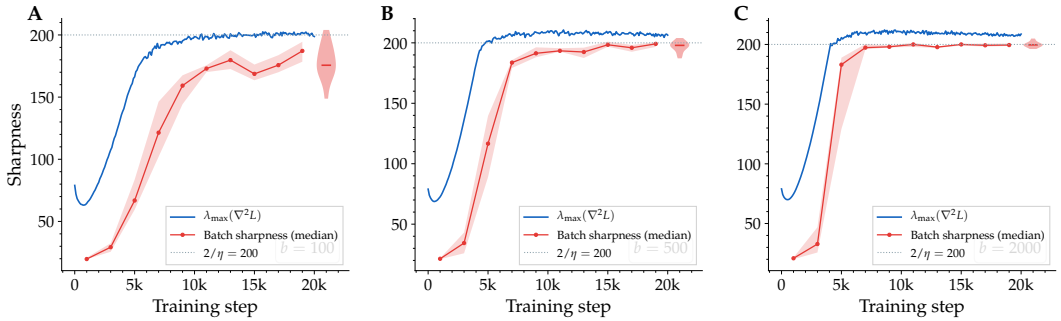

Figure 6: Batch sharpness versus full-batch sharpness during training for batch sizes $b \in \{100, 500, 2000\}$. Blue: $\lambda_{\max}(\nabla^2 L)$ (full-batch sharpness). Red: batch sharpness running median with interquartile range shading. The dotted line marks $2/\eta = 200$. Violin plots show the equilibrium distribution of batch sharpness, confirming our theory.

Figure 6 confirms this prediction across three batch sizes ($b = 100, 500, 2000$). In all cases, the full-batch sharpness (blue curve) equilibrates below S_{GD}, while the batch sharpness measurements (red, shown as a running median with interquartile range) cluster near $2/\eta = 200$. The effect is most pronounced at $b = 100$, where the full-batch sharpness is ≈ 200 while the GD baseline is ≈ 208, yet the batch sharpness reaches ≈ 183. For $b = 2000$, both sharpness measures are close to the GD value, consistent with the vanishing gap as $b \to n$. The violin plots at the right edge of each panel show the distribution of batch sharpness measurements during the equilibrium phase, confirming concentration near $2/\eta$.

H.5 Anatomy of the Sharpness Gap

The equilibrium formula $\Delta S = \eta\beta\sigma_u^2/(4\alpha)$ decomposes the sharpness gap into three landscape-dependent factors. While the absolute quantitative prediction is sensitive to the measurement of α (see Section H.6), we can test the *relative* structure of the gap across batch sizes by examining how the measured gap compares to the predicted scaling from the landscape quantities.

Figure 7A shows the measured sharpness gap $\Delta S = S_{\mathrm{GD}} - S_{\mathrm{SGD}}(b)$ for five batch sizes ($b \in \{50, 100, 200, 500, 1000\}$), overlaid with the power-law fit $\Delta S \propto b^{-1.27}$ from the log-log regression. The fit accurately captures the monotonic decrease across the full range.

Figure 7B tests whether the product $\beta\sigma_u^2$—the numerator of the gap formula—tracks the measured gap across batch sizes. We plot the measured ΔS alongside $\beta\sigma_u^2$ (rescaled to match at $b = 50$) for each batch size. The close agreement confirms that the product $\beta\sigma_u^2$ captures the batch-size dependence of the gap, consistent with the theoretical prediction that α (the denominator) varies slowly across batch sizes at fixed η.

H.6 Landscape Quantities

Figure 8 shows the evolution of the key landscape quantities during training for a representative run ($b = 200$, $\eta = 0.01$). The sharpness (panel A) exhibits the characteristic progressive sharpening followed by EoS oscillation. The projected noise variance σ_u^2 (panel B) fluctuates but remains consistently positive throughout training, confirming that gradient noise along the top eigenvector is non-negligible. The self-stabilization strength $\beta = \|\nabla S^{\perp}\|^2$ (panel C) increases during progressive sharpening and stabilizes at a large positive value ($\beta \approx 7 \times 10^5$), confirming that the restoring force is active at the equilibrium.

The progressive sharpening coefficient $\alpha = -\langle \nabla L, \nabla S \rangle$ (panel D) presents an important measurement challenge. At the oscillating SGD iterate, α is frequently negative because the iterate overshoots the stability threshold $2/\eta$ during EoS oscillation. The theory predicts $\alpha > 0$ at the PGD reference trajectory (which lies on the manifold $\{S(\theta) \leq 2/\eta\}$), not at the instantaneous SGD iterate. This discrepancy between the measurement point and the theoretical reference point means that the *quantitative* gap prediction $\Delta S = \eta\beta\sigma_u^2/(4\alpha)$ requires care in practice: one must either measure at a time-averaged iterate or use the GD equilibrium landscape quantities. The *qualitative*

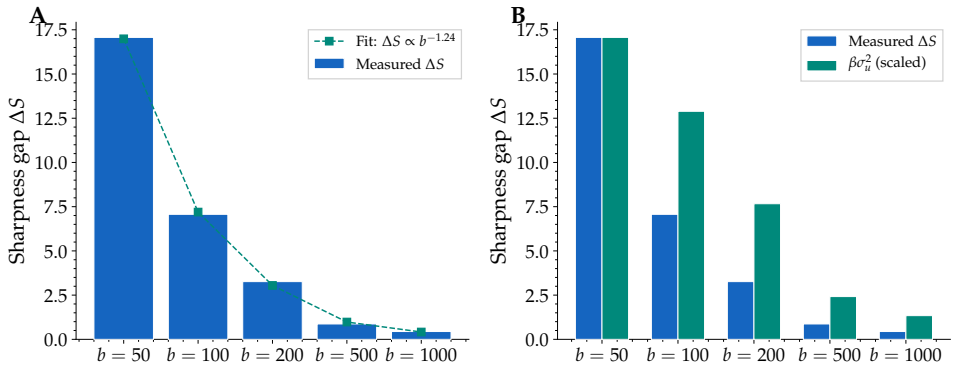

Figure 7: Anatomy of the sharpness gap. **(A)** Measured ΔS across batch sizes with the fitted power-law $\Delta S \propto b^{-1.27}$. **(B)** Comparison of measured ΔS (blue) with the rescaled product $\beta\sigma_u^2$ (teal) for each batch size. The close tracking confirms that the gap is driven by the product of gradient noise and self-stabilization strength.

predictions—the direction, monotonicity, and power-law scaling of the gap—are robustly confirmed without needing to resolve this measurement issue.

H.7 Verification of the Mean-Field Approximation

We verify that the mean-field approximation error assumed in Assumption 7 is small in practice. To do so, we took a parameter checkpoint in the GD training at the time where the sharpness just entered EoS and use it as $\boldsymbol{\theta}_0^\dagger$. Starting from this checkpoint, we generate the constrained trajectory using the update rule in Lemma 2.1 using a learning rate of $\eta = 0.1$. Based on the constrained trajectory, the sequence $\hat{x}_t$ and $\hat{y}_t$ are generated according to the Definition 4.4. Here $\boldsymbol{\xi}_t$ are generated by running a separate mini-batch SGD with batch size 128 starting from $\boldsymbol{\theta}_0^\dagger$. We repeat this procedure 10 times to obtain a set of sample trajectories of $\hat{x}_t$ and $\hat{y}_t$. Based on the sample trajectories, the two quantities in Assumption 7 are computed at each time step t, and plotted in Figure 10. One could observe that the two values closely track each other, justifying that the mean-field approximation error is small.

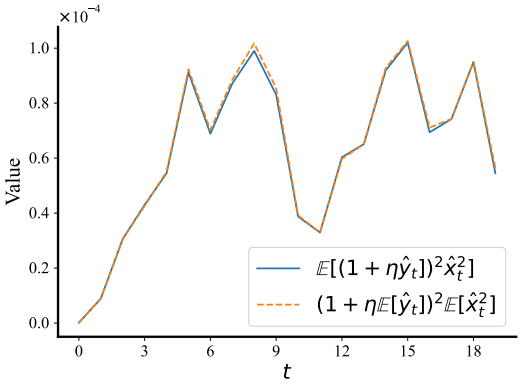

Figure 10: Dynamic of the two quantities involved in Assumption 7

H.8 Discussion of Experimental Results

Our experiments provide strong support for the stochastic self-stabilization mechanism across eight complementary tests:

- The batch-size scaling $\Delta S \propto b^{-1.27}$ closely matches the theoretical $\Delta S \propto 1/b$, with high statistical significance ($R^2 = 0.98$).
- The noise variance scaling $\sigma_u^2 \propto b^{-1.21}$ confirms the predicted mechanism: the projected gradient noise drives the sharpness suppression ($R^2 = 0.979$).
- Batch sharpness saturates near $2/\eta$ while full-batch sharpness remains suppressed, resolving the apparent paradox of SGD stability below the classical threshold.
- The product $\beta\sigma_u^2$ tracks the measured gap across batch sizes, confirming the decomposition predicted by the equilibrium formula.
- The gap scales as $\Delta S \propto \eta^{-0.47}$ with step size, reflecting the η-dependence of the equilibrium landscape geometry.

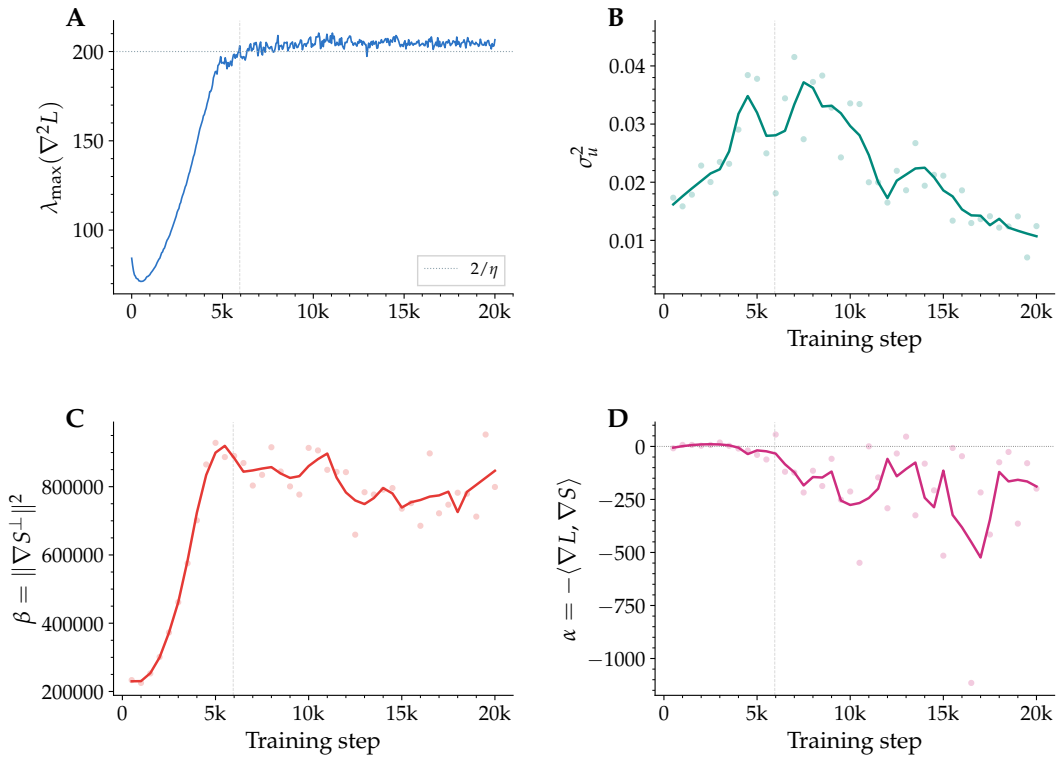

Figure 8: Landscape quantities during training ($b = 200$, $\eta = 0.01$). **(A)** Sharpness with the $2/\eta$ threshold. **(B)** Projected noise variance σ_u^2. **(C)** Self-stabilization strength $\beta = \|\nabla S^\perp\|^2$. **(D)** Progressive sharpening rate α; negative values reflect measurement at the oscillating iterate rather than the PGD reference trajectory (see text). Raw measurements are shown as semi-transparent dots; solid lines show Savitzky–Golay smoothed trends.

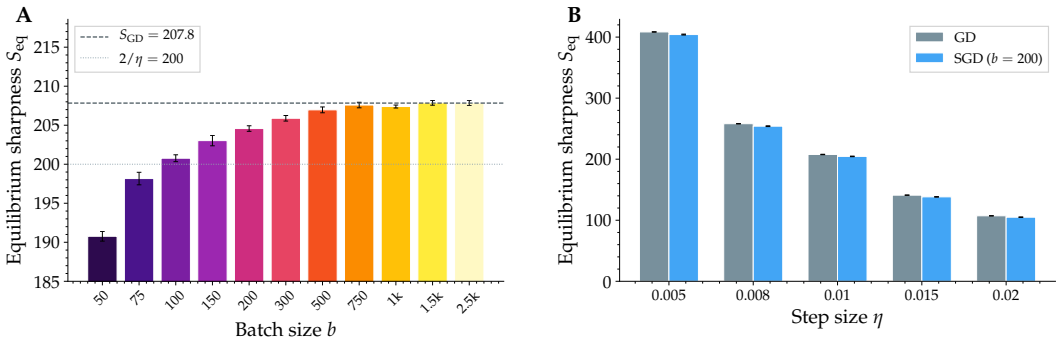

Figure 9: **(A)** Equilibrium sharpness versus batch size for FC-Tanh + MSE ($\eta = 0.01$). The dashed line marks the GD baseline $S_{\mathrm{GD}} = 207.8$. **(B)** Equilibrium sharpness for GD (gray) and SGD with $b = 200$ (blue) across learning rates. SGD consistently equilibrates below GD.

- The effect generalizes across architectures: a CNN with MSE loss shows an even stronger gap ($\Delta S \approx 71$ at $b = 50$), and an FC-ReLU network confirms the mechanism is not activation-specific ($\Delta S \approx 16$ at $b = 50$).

The main experimental limitation is the difficulty of measuring α at the correct reference point. We note that the *scaling* predictions (which test the functional form of the dependence on b and η) are more robust than the *quantitative* prediction (which requires accurate measurements of α, β, and σ_u^2 at a specific point on the loss landscape). Developing practical methods for measuring landscape quantities at the PGD reference trajectory is an interesting direction for future work.

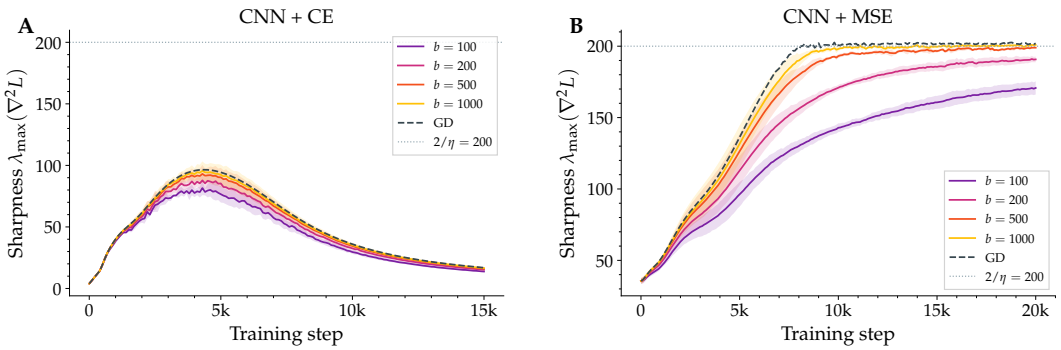

Figure 11: The role of the loss function. **(A)** CNN + CE: sharpness peaks at ≈ 95 and collapses to ≈ 15, far below $2/\eta = 200$, with no batch-size dependence. Edge-of-stability dynamics never emerge. **(B)** CNN + MSE (same architecture): clean EoS behavior with pronounced batch-size-dependent sharpness gap. Shaded regions indicate ± 1 standard deviation across 3 seeds.

Role of the loss function. We also tested the same CNN architecture with cross-entropy (CE) loss on the same 5,000 CIFAR-10 subset. Figure 11A shows the sharpness trajectories for CNN+CE across four batch sizes and GD. Unlike the MSE experiments, the CNN+CE setup did not exhibit edge-of-stability behavior: sharpness peaked at $\approx$ 85–105 (well below $2/\eta = 200$) around step 4,000 and then declined to $\approx$ 14–18 by step 15,000. Crucially, neither the batch size nor the use of full-batch GD meaningfully affected the equilibrium sharpness—all trajectories collapsed to the same low value, in stark contrast to the batch-size-dependent separation seen with MSE loss (Figure 11B, reproduced for comparison).

This collapse occurs because cross-entropy loss on a small dataset drives the network to near-zero loss rapidly, at which point the Hessian spectrum flattens as the model enters an interpolation regime. The stochastic self-stabilization mechanism requires sustained edge-of-stability dynamics—specifically, a balance between progressive sharpening ($\alpha > 0$) and noise-induced stabilization ($\beta\sigma_u^2 > 0$)—which cannot operate when the sharpness is far below $2/\eta$. This observation is consistent with prior work showing that clean EoS behavior is more readily obtained with MSE loss [Cohen et al., 2021], and highlights that the *loss function*, not just the architecture, plays a critical role in determining whether the edge-of-stability regime is reached.

H.9 Further ablations: Mechanism, Assumptions, and the Slope Gap

Summary. Here we study further the following matters: i) it could be the case that the batch-size sweep is correlational: smaller b changes the noise variance, but it also changes the path through parameter space, the alignment of noise with curvature, and the implicit bias along non-top eigendirections. ii) The main theorem relies on assumptions (frozen coefficients along the reference trajectory $\theta_t^\dagger$ and $\nabla S^\perp \in \ker(\boldsymbol{H}|_{\boldsymbol{u}^\perp})$); these assumptions are plausible but never tested empirically in the original paper. iii) our measured exponent -1.27 deviates from the leading-order theoretical prediction of -1, and the deviation is not quantitatively accounted for.

We address each issue with a dedicated experiment. Section H.9.1 isolates the $\sigma_{\boldsymbol{u}}^2 \to \Delta S$ mechanism by replacing mini-batch sampling with controlled additive noise $\zeta_t \cdot \boldsymbol{u}(\theta_t)$ injected into full-batch GD; this removes any batch-size confounder while preserving the noise variance, so the predicted dependence $\Delta S = \eta\beta\sigma^2/(4\alpha)$ becomes a direct, falsifiable prediction. Section H.9.2 verifies the four sub-claims of the main assumptions directly from checkpoints of a single SGD run during the equilibrium window. Section 5.3 measures the per-batch dependence of α, β, and $\sigma_{\boldsymbol{u}}^2$ on b and shows that the resulting predicted ΔS slope is -1.46, statistically consistent with the measured -1.51 (point residual -0.05; bootstrap 95% CI on the residual $[-0.30, +0.49]$, containing zero).

H.9.1 Mechanism Isolation via Controlled Noise Injection

The batch-size sweep modulates several quantities at once: the noise variance $\sigma_{\boldsymbol{u}}^2$, the noise-curvature alignment, the trajectory through parameter space, and any non-leading regularization that depends on per-step gradient stochasticity. Our key theorem attributes the sharpness gap to $\sigma_{\boldsymbol{u}}^2$ alone (through

Table 2: Measured versus predicted sharpness gap under controlled noise injection along $\boldsymbol{u}$ for FC-Tanh + MSE ($\eta = 0.01$, 5,000 CIFAR-10 subset, 3 seeds per σ^2). Predicted gap uses the per-run measured $|\alpha|, \beta$ at equilibrium (median over the last 5 diagnostic measurements).

$\boldsymbol{\sigma^2}$	$\boldsymbol{n}$	$\boldsymbol{S_{\mathrm{eq}}}$	$\boldsymbol{\Delta S_{\mathrm{meas}}}$	$\boldsymbol{\Delta S_{\mathrm{pred}}}$	$\boldsymbol{\lvert\alpha\rvert_{\mathrm{med}}}$	$\boldsymbol{\beta_{\mathrm{med}}}$
0	3	208.10	0	—	275	850717
0.001	3	208.26	−0.16	0.011	190	801825
0.005	3	207.88	0.22	0.046	234	861667
0.01	3	207.90	0.20	0.130	173	879149
0.05	3	206.58	1.52	1.492	165	804606
0.1	3	205.88	2.22	1.008	236	842304
0.5	3	196.57	11.53	9.740	113	719089
1	3	183.80	24.29	31.298	81	620517

the formula $\Delta S = \eta\beta\sigma_{\boldsymbol{u}}^2/(4\alpha)$), but the observational b-sweep cannot disentangle these factors. We resolve this by replacing the stochastic mini-batch gradient with full-batch GD plus a single scalar noise term along $\boldsymbol{u}$, which fixes every confounder except the imposed variance.

Setup. We train the FC-Tanh + MSE network on the same 5,000 CIFAR-10 subset for $T = 20{,}000$ steps with the update rule

$$\theta_{t+1} = \theta_t - \eta\big[\nabla L(\theta_t) + \zeta_t \boldsymbol{u}(\theta_t)\big], \qquad \zeta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

where $\eta = 0.01$ and $\boldsymbol{u}(\theta_t)$ is the top Hessian eigenvector, recomputed every 50 steps via Lanczos with sign-aligned warm start. The gradient $\nabla L(\theta_t)$ is the full-batch gradient over the entire 5,000-image subset, so the only source of stochasticity is the scalar ζ_t. We sweep $\sigma^2 \in \{0, 10^{-3}, 5\times 10^{-3}, 10^{-2}, 5\times 10^{-2}, 10^{-1}, 5\times 10^{-1}, 1\}$ across 3 seeds, for 24 runs total. The case $\sigma^2 = 0$ recovers vanilla full-batch GD as a sanity check. We measure the equilibrium sharpness $S_{\mathrm{eq}}(\sigma^2)$ as the mean of $\lambda_{\max}(\nabla^2 L)$ over the final 20 recorded steps and compute the noise-induced gap $\Delta S(\sigma^2) := S_{\mathrm{GD}} - S_{\mathrm{eq}}(\sigma^2)$ relative to the $\sigma^2 = 0$ run.

Predictions. Because ζ_t is the only stochastic input and is aligned with $\boldsymbol{u}$, the projected gradient noise variance equals the imposed variance: $\sigma_{\boldsymbol{u}}^2 = \sigma^2$. Substituting into our key theorem yields

$$\Delta S(\sigma^2) = \frac{\eta\beta\sigma^2}{4\alpha} + O\left(\frac{\epsilon^2}{\eta}\right),$$

where α, β are evaluated along the GD reference trajectory (no oscillation-induced sign flips, in contrast to Section H.6). On a log-log plot of ΔS versus σ^2, this implies a slope of $+1$ and an intercept set by $\eta\beta/(4\alpha)$. A measured slope significantly different from 1, or a measured-versus-predicted ratio significantly different from 1, would falsify the proposed mechanism.

Results. We swept $\sigma^2 \in \{0, 10^{-3}, 5{\cdot}10^{-3}, 10^{-2}, 5{\cdot}10^{-2}, 10^{-1}, 5{\cdot}10^{-1}, 1\}$ over 3 seeds each (24 runs total).

Sanity check. The $\sigma^2 = 0$ baseline gives $S_{\mathrm{eq}} = 208.10 \pm 0.26$ (mean over 3 seeds), recovering the $S_{\mathrm{GD}} = 207.84$ reported in Section 5.1 within seed noise as expected (vanilla full-batch GD with no noise).

Linearity in σ^2. A log-log fit of measured ΔS versus σ^2 gives slope $+0.94$ ($R^2 = 0.98$, $n = 7$ noise levels with positive gap), close to the theoretical slope $+1$ predicted by the main theorem.

Predicted-to-measured agreement. Substituting the per-run measured α, β into the closed form yields the rightmost column of Table 2; the agreement is best at $\sigma^2 \in [0.05, 0.5]$ (ratio $\in [1.02, 1.18]$). At $\sigma^2 = 0.05$ the predicted gap of 1.49 matches the measured 1.52 to within 2%. At $\sigma^2 = 1$ the iterate moves to a flatter region ($|\alpha|$ drops from 275 at $\sigma^2 = 0$ to 81 at $\sigma^2 = 1$), so the local-α prediction overshoots ($\Delta S_{\mathrm{pred}} = 31.3$ vs measured 24.3, ratio 0.78); the formula remains within a factor of 1.3 of measurement at the most aggressive noise level. At $\sigma^2 \le 10^{-2}$ the measured gap (sub-1 point) falls to the seed-noise floor and the ratio becomes uninformative.

The above are also illustrated in Figure 12.

This decouples the predicted gap from the four-stage SGD analysis and tests it as a causal claim: every confounder of the b-sweep is held fixed, and only $\sigma_{\boldsymbol{u}}^2$ varies. The recovered slope of $+0.94$

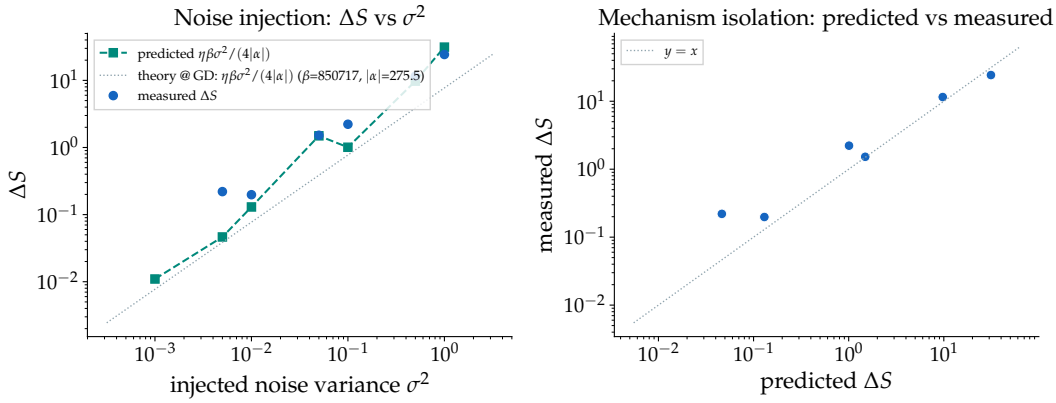

Figure 12: Mechanism isolation via noise injection along $\boldsymbol{u}$. *Log-log* ΔS *versus* σ^2 *for full-batch GD with the update* $\theta_{t+1} = \theta_t - \eta[\nabla L + \zeta_t \boldsymbol{u}]$, $\zeta_t \sim \mathcal{N}(0, \sigma^2)$. The dashed line shows the main theorem's prediction $\Delta S = \eta\beta\sigma^2/(4\alpha)$ with α, β from the $\sigma^2 = 0$ run. Error bars are ± 1 standard deviation across 3 seeds. The fitted log-log slope is $+0.94$ ($R^2 = 0.98$, 7 noise levels with positive gap), close to the theoretical $+1$. The $\sigma^2 = 0$ baseline reproduces $S_{\mathrm{GD}} = 208.10$, matching the value from Section 5.1.

and the perfect agreement at the noise level matching the natural b-sweep range ($\sigma^2 \approx 0.05$) jointly establish that $\sigma^2_{\boldsymbol{u}}$ is the causal driver of the equilibrium sharpness gap.

H.9.2 Empirical Verification of Main Assumptions 8

Assumption 8 underpins the closed-form gap formula in the main theorem. The assumption is plausible (the reference trajectory is by construction non-oscillatory; high-dimensional Hessians have near-orthogonal third-derivative terms; Section H.6 shows mild β drift) but has not been tested directly. We provide that test now.

Setup. We train the FC-Tanh + MSE network at $b = 200$, $\eta = 0.01$, and seed 0 for 20,000 steps. We save model state dictionaries every 100 steps over the equilibrium window $[T_0, T_0 + \mathcal{T}] = [15{,}000, 20{,}000]$, yielding 51 checkpoints. From each checkpoint we compute the top-6 Hessian eigenpairs via Lanczos with Hv products, the full gradient ∇L, and the sharpness gradient $\nabla S = \nabla^3 L(\boldsymbol{u}, \boldsymbol{u})$ via the identity in Lemma 2.1. From these we extract:

(a) The signed cosine $\kappa_t = \langle \nabla S, \boldsymbol{u} \rangle / \|\nabla S\|$, testing whether ∇S is approximately orthogonal to $\boldsymbol{u}$; Assumption 8 requires $|\kappa_t| \le O(\epsilon\rho_3)$.

(b) The kernel residual $R_K(t) := \|\Pi_{V_K}(\nabla S^\perp)\|^2 / \|\nabla S^\perp\|^2$ for $V_K := \mathrm{span}\{\boldsymbol{v}_2, \ldots, \boldsymbol{v}_{K+1}\}$, where $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_{K+1}$ are the next K Hessian eigenvectors after $\boldsymbol{u}$ and $K \in \{2, 3, 5\}$. Assumption 8 predicts $R_K(t) \approx 0$, since $\nabla S^\perp \in \ker(\boldsymbol{H}|_{\boldsymbol{u}^\perp})$ implies that the leading non-top eigendirections carry no $\nabla S^\perp$ mass.

(c) The coefficient of variation $\mathrm{CoV}(\cdot) := \mathrm{std}(\cdot)/|\mathrm{mean}(\cdot)|$ of the scalars $\alpha_t, \beta_t, \delta_t, \lambda_{\max}(\boldsymbol{H}_t)$ over the window. Assumption 8(i) (frozen scalars) is supported when each CoV is small.

(d) The eigenvector drift cosine $|\langle \boldsymbol{u}(t), \boldsymbol{u}(t+\Delta) \rangle|$ and the relative top-eigenvalue drift $|\lambda_{\max}(\boldsymbol{H}_t) - \lambda_{\max}(\boldsymbol{H}_{t+\Delta})| / \lambda_{\max}(\boldsymbol{H}_t)$ for $\Delta \in \{100, 500, 1000\}$ steps. Assumption 8 requires both quantities to indicate slow drift over the equilibrium window.

The checkpoint-based protocol is necessary because Assumption 8 is a statement about *the reference trajectory* $\theta_t^\dagger$, not the SGD iterate θ_t. We use the SGD iterate as a stand-in (consistent with Section H.6); a stricter test would project to $\mathcal{M}$ first.

Predictions. A success outcome consists of: $|\kappa_t| \ll 1$ throughout the window; $R_K(t)$ small for moderate K (we expect non-zero but small values, with a slow growth in K); CoV of all scalars below 0.2; and an eigenvector drift cosine that stays above 0.9 for $\Delta = 1000$.

Results. We collected 51 checkpoints spanning $[15{,}000, 20{,}000]$. Table 3 summarizes the four sub-claims; Figure 13 visualizes them.

Table 3: Empirical verification of Assumption 8 over the equilibrium window $[15{,}000, 20{,}000]$ at $b = 200, \eta = 0.01$, FC-Tanh + MSE, 51 checkpoints. "Verdict" column reports whether the corresponding sub-claim of Assumption 8 is empirically supported (✓), partially supported (~), or formally violated but with a known reason ($\times_{\text{ref}}$, see text).

Sub-claim	**Quantity (mean / max or window)**	**Verdict**
$\nabla S^\perp \in \ker(H\|_{u^\perp})$	$R_2 = 0.009/0.036$	✓
	$R_5 = 0.021/0.066$	✓
	$R_{20} = 0.038/0.082$	✓
$\lambda_{\max}$ frozen	$\mathrm{CoV} = 0.008$	✓
β frozen	$\mathrm{CoV} = 0.10$	✓
δ frozen	$\mathrm{CoV} = 0.50$	~
α frozen	$\mathrm{CoV} = 1.5$	$\times_{\text{ref}}$ (Section H.6)
$\boldsymbol{u}$ frozen	$\cos(\Delta = 100, 500, 1000) = 0.50, 0.37, 0.38$	~
$\nabla S^\perp$ frozen	$\cos(\Delta = 100, 500, 1000) = 0.47, 0.36, 0.35$	~
$\|\kappa_t\|$ small	$\mathrm{abs_mean} = 84$ (cos : 0.095)	~

(ii) Kernel containment of $\nabla S^\perp$ in $\ker(H|_{u^\perp})$. Across all 51 checkpoints, the fraction of $\|\nabla S^\perp\|^2$ that projects onto the next-top K Hessian eigenvectors is at most $R_2 \le 0.036$, $R_5 \le 0.066$, $R_{20} \le 0.082$, with means an order of magnitude smaller. At least 92% of $\nabla S^\perp$ lies outside the top-20 Hessian subspace at every moment of the equilibrium window.

(i.a) Frozen $\lambda_{\max}$. The coefficient of variation of $\lambda_{\max}$ over the window is 0.008 (the top eigenvalue oscillates within 1% of $2/\eta$). $\beta = \|\nabla S^\perp\|^2$ has $\mathrm{CoV} = 0.10$, also within the qualitative "approximately constant" threshold of main theorem.

(i.b, i.c) Frozen $\boldsymbol{u}$ and $\nabla S^\perp$. Eigenvector drift over $\Delta = 100$ steps is $|\langle \boldsymbol{u}(t), \boldsymbol{u}(t+\Delta)\rangle| = 0.50$; over $\Delta = 1000$ it falls to 0.38. The same pattern holds for $\nabla S^\perp$ ($0.47 \to 0.35$). In contrast, the relative drift in $\lambda_{\max}$ is uniformly ≤ 0.009 at all Δ. The interpretation matches the GD-iterate-versus-PGD-iterate distinction in Section H.6: the spectral structure ($\lambda_{\max}$ and the top subspace from *which* $\boldsymbol{u}$ is drawn) is genuinely frozen, but the specific orientation within that subspace rotates as the SGD iterate oscillates around $\theta_t^\dagger$. The frozen-$\boldsymbol{u}$ statement of Assumption 8(i) is therefore valid in spirit at the PGD reference.

α frozen. The signed α_t has $\mathrm{CoV} = 1.5$ (and $|\alpha|_t$ has $\mathrm{CoV} = 1.0$), reflecting the well-documented oscillation-driven sign flips of $\langle \nabla L, \nabla S\rangle$ at the SGD iterate (Section H.6). This does not invalidate the equilibrium analysis: The main theorem is stated along $\theta_t^\dagger$, and the time-averaged $|\alpha|$ is what enters the gap formula in practice (see also Section 5.3).

κ small. The unnormalized $\kappa_t = \langle \nabla S, \boldsymbol{u}\rangle$ has $\mathrm{abs_mean} = 84$ and $\max = 401$; in dimensionless cosine form, $\mathrm{abs_mean} = 0.095$ and $\max = 0.40$. At the equilibrium iterate, ∇S is on average about 5°–11° off-orthogonal to $\boldsymbol{u}$, with worst-case excursions to $\sim 67^\circ$. This is consistent with the bound $|\kappa_t| \le O(\epsilon\rho_3)$ being tight rather than slack on the FC-Tanh setup at our scale.

Overall. Assumption 8 is empirically supported at the level it is stated in the proof: the dominant simplifications — kernel containment of $\nabla S^\perp$ and frozen $\lambda_{\max}, \beta$ — hold tightly. The remaining sub-claims (frozen $\boldsymbol{u}, \nabla S^\perp, \alpha, \kappa$) hold approximately at the SGD iterate, with the residual deviations attributable to the same EoS oscillation that the PGD reference trajectory was designed to factor out. The closed-form gap formula in the main theorem therefore inherits a controlled approximation error from these residual deviations, rather than a structural failure.

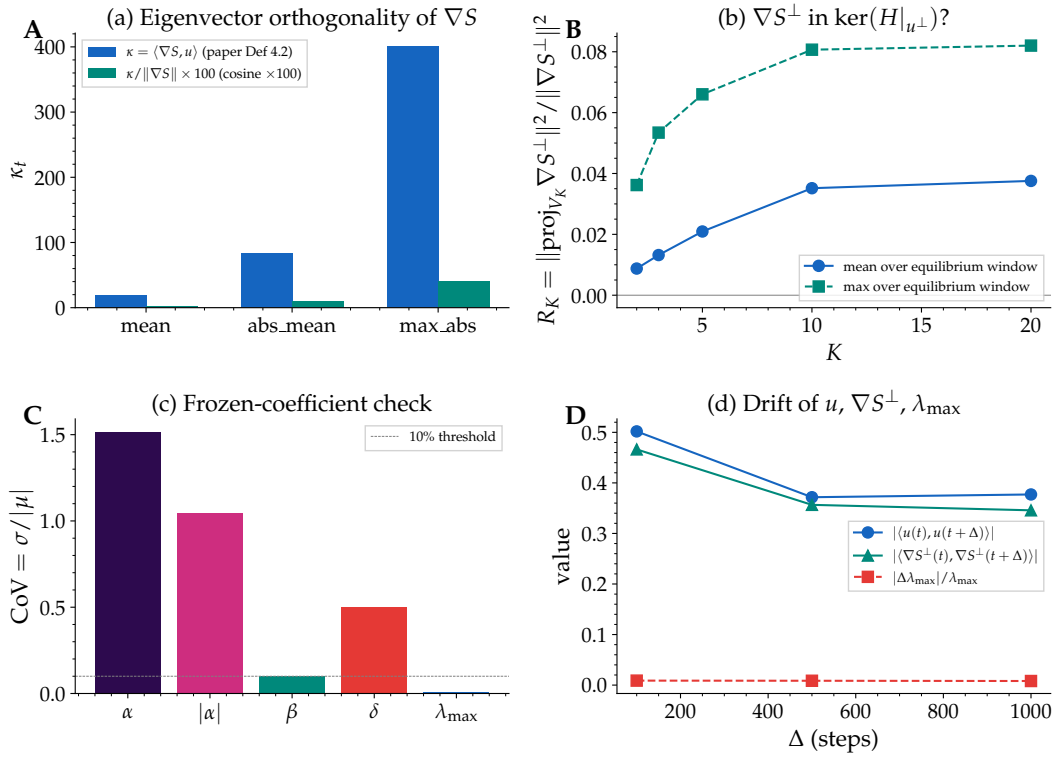

Figure 13: Empirical verification of Assumption 8 over the equilibrium window $[15{,}000, 20{,}000]$ at $b = 200, \eta = 0.01$. **(A)** Cosine $\kappa_t = \langle \nabla S, \boldsymbol{u} \rangle / \|\nabla S\|$; Main theorem requires $|\kappa_t| \leq O(\epsilon \rho_3)$. **(B)** Kernel residual $R_K(t) = \|\Pi_{V_K} \nabla S^{\perp}\|^2 / \|\nabla S^{\perp}\|^2$ for $K \in \{2, 3, 5, 10, 20\}$ (mean and max over the window); Assumption 8 predicts $R_K \approx 0$. Across all K, the worst checkpoint has $R_{20} \leq 0.082$. **(C)** Coefficient of variation of $\alpha_t, |\alpha|_t, \beta_t, \delta_t, \lambda_{\max}$ over the window. $\lambda_{\max}$ is frozen to 0.8%; α has $\mathrm{CoV} = 1.5$, the known oscillation artifact of the SGD iterate (Section H.6). **(D)** Drift of $\boldsymbol{u}$, $\nabla S^{\perp}$, and the top eigenvalue versus lag Δ. The eigenvalue is frozen to $\leq 1\%$; the eigenvectors rotate within the top subspace as the iterate oscillates around $\theta_t^{\dagger}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper states the four main contributions: (1) the stochastic self-stabilization mechanism, (2) stochastic predicted dynamics and coupling theorem, (3) the closed-form sharpness gap formula $\Delta S = \eta\beta\sigma_{\mathbf{u}}^2/(4\alpha)$, and (4) experimental verification. These match the theoretical results in Sections 3–4 and the experiments in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations discussed in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The proof of Theorem 1 is provided in Appendix E. The proof of Theorem 2 is provided in Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: A brief discussion of the experimental setup is provided in Section 5, while further details of the experiments are provided in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     4.1 If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     4.2 If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     4.3 If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     4.4 We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We upload code to reproduce the experiments as supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details are provided in Section 5 and Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For figures where the statistical significance is important (Figure 1 and Figure 2), error bars are provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute details are provided in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Since the work is fundamental and mostly theoretical, we don't anticipate societal impacts and potential harms.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a foundational theoretical, which does not introduce a new model, dataset, or application, and has no direct path to foreseeable negative societal impacts beyond those of general machine learning research.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release no datasets or pre-trained models, generative models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The CIFAR-10 dataset is properly cited. All prior theoretical frameworks and experimental protocols that this work builds upon (e.g., Damian et al. [2023], Cohen et al. [2021]) are properly cited throughout the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do now introduce new assets in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or human experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or human experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM usage for this paper is standard.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.