Discrete Lyapunov Stability of SGD in the Interpolation Regime:
Per-Sample Analysis and Generalization Bounds

CΛ / Lightman
Lightman.chang@gmail.com

Abstract

We study the implicit regularization of stochastic gradient descent (SGD) in the interpolation regime, where the training loss reaches zero and the continuous-time stochastic differential equation (SDE) approximation degenerates because the gradient noise covariance vanishes at every interpolating solution. Working directly with the discrete SGD recursion, we introduce the *per-sample Lyapunov exponent* $\lambda_k = \frac{1}{n}\ln|1-\eta h_k|$, where $h_k = \|J_k\|^2$ is the per-sample curvature and η is the learning rate. Under a Jacobian orthogonality assumption, we establish that every interpolating solution with $h_{\max} > 2/\eta$ is almost surely unstable (positive maximal Lyapunov exponent), while every solution with $h_{\max} < 2/\eta$ is almost surely asymptotically stable. We then derive a closed-loop chain: *SGD stability* $\Rightarrow$ *Jacobian complexity control* ($\mathcal{C} \le \sqrt{2/\eta}$) $\Rightarrow$ *Rademacher complexity bound* $\Rightarrow$ *generalization bound* $O(\rho/\sqrt{\eta n})$. This mechanism is primarily a discrete phenomenon that does not rely on any SDE approximation. We additionally provide a PAC-Bayes bound that eliminates the dependence on the neighborhood radius ρ. Our analysis clarifies the distinction between per-sample stability ($h_{\max} < 2/\eta$) and aggregate spectral conditions such as $\mathrm{tr}(H) \le 2/\eta$, and connects stability-based selection to generalization through an explicit, verifiable complexity measure.

1 Introduction

Modern overparameterized neural networks routinely achieve zero training loss—a regime known as *interpolation*—yet generalize well to unseen data. Understanding why stochastic gradient descent (SGD) selects, among the continuum of interpolating solutions, those that generalize is a central open problem in the theory of deep learning.

Background. A fruitful line of work models SGD through its continuous-time SDE approximation [3, 5]:

$$d\theta = -\nabla L(\theta)\,dt + \sqrt{\tfrac{\eta}{B}\,\Sigma(\theta)}\,dW_t,$$

where $\Sigma(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell_i(\theta)\nabla\ell_i(\theta)^\top - \nabla L(\theta)\nabla L(\theta)^\top$ is the gradient noise covariance. The multiplicative structure of Σ biases the Fokker–Planck stationary distribution toward minima with low noise-to-curvature ratio Σ/H [1]. This provides a satisfying explanation of implicit regularization—*when* $\Sigma > 0$.

In the interpolation regime, however, every interpolating solution θ^* satisfies $\nabla\ell_i(\theta^*) = 0$ for all i, so $\Sigma(\theta^*) = 0$. The SDE approximation degenerates and the Fokker–Planck analysis breaks down entirely.

Known results. Wu and Su [6] proved that SGD cannot converge to interpolating solutions at which $\mathrm{tr}(H) > 2/\eta$, establishing an aggregate spectral condition for instability. Wu, Wang, and Su [7] studied the sharpness reduction phenomenon. Dinh et al. [2] demonstrated that sharpness (Hessian eigenvalues) alone is not reparameterization-invariant and hence cannot serve as a sole predictor of generalization.

The gap. The condition $\mathrm{tr}(H) > 2/\eta$ is an aggregate condition: it sums over all eigenvalues and does not distinguish between having one very large per-sample curvature versus many moderately large ones. Moreover, the connection from instability to generalization has remained informal.

Our contribution. We provide a complete, self-contained analysis that:

(i) introduces the per-sample Lyapunov exponent $\lambda_k = \frac{1}{n}\ln|1-\eta h_k|$ and establishes that stability is governed by $h_{\max} < 2/\eta$ rather than the aggregate $\mathrm{tr}(H) < 2/\eta$;

(ii) proves that stable solutions have controlled Jacobian complexity $\mathcal{C}(\theta^*) \le \sqrt{2/\eta}$;

(iii) derives a Rademacher-based generalization bound of order $O(\rho/\sqrt{\eta n})$ and a PAC-Bayes bound of order $O(\sqrt{d\ln(1/\eta)/n})$;

(iv) assembles these into a closed-loop theorem: SGD stability $\Rightarrow$ complexity control $\Rightarrow$ generalization.

The entire analysis works directly on the discrete SGD recursion and does not invoke any SDE approximation. The key mechanism—linear stability filtering via random matrix products—is primarily a discrete phenomenon.

Paper structure. Section 2 introduces notation and the interpolation setting. Section 3 presents the core definitions (critical learning rate, good/bad solutions, Jacobian complexity). Section 4 states the assumptions. Section 5 presents the main results: two lemmas, six theorems, and one corollary. Section 6 contains the complete proofs. Section 7 discusses reparameterization invariance. Section 8 provides connections to the non-interpolation regime, testable predictions, limitations, and open problems. Section 9 reviews related work.

2 Setup and Notation

Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a training set with $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$. Consider a model $f_\theta \colon \mathcal{X} \to \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$, with $d > n$ (overparameterized regime). Define the per-sample MSE loss and the empirical risk:

$$\ell_i(\theta) = \tfrac{1}{2}\big(f_\theta(x_i) - y_i\big)^2, \qquad L(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_i(\theta).$$

Interpolation manifold. The set of global minimizers of L with zero training loss is

$$\mathcal{M} = \big\{\theta \in \mathbb{R}^d : f_\theta(x_i) = y_i,\ \forall\, i \in [n]\big\}.$$

Since $d > n$, generically $\mathcal{M}$ is a smooth submanifold of dimension at least $d - n$.

Per-sample quantities at $\theta^* \in \mathcal{M}$.

- **Jacobian vectors:** $J_i = \nabla_\theta f_{\theta^*}(x_i) \in \mathbb{R}^d$.
- **Jacobian matrix:** $\mathbf{J} \in \mathbb{R}^{n\times d}$, with the i-th row equal to $J_i^\top$.
- **Per-sample Hessian:** $H_i = J_i J_i^\top$. At $\theta^* \in \mathcal{M}$ the residual $r_i(\theta^*) = f_{\theta^*}(x_i) - y_i = 0$, so $\nabla^2 \ell_i(\theta^*) = H_i$.
- **Per-sample curvature:** $h_i = \|J_i\|^2$.
- **Maximum curvature:** $h_{\max} = \max_{i\in[n]} h_i$.
- **Active subspace:** $\mathcal{S} = \mathrm{span}(J_1, \ldots, J_n)$.

SGD recursion.

$$\theta_{t+1} = \theta_t - \eta \nabla \ell_{z_t}(\theta_t), \qquad z_t \sim \text{Uniform}([n]) \text{ i.i.d.}$$

Population risk. $L_{\text{pop}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_\theta(x), y)]$.

3 Definitions

Definition 1 (Critical learning rate)**.** For an interpolating solution $\theta^* \in \mathcal{M}$, the *critical learning rate* is

$$\eta_c(\theta^*) = \frac{2}{h_{\max}(\theta^*)}.$$

Definition 2 (Good and bad solutions)**.** Given a learning rate $\eta > 0$:

- θ^* is η**-stable** (good, flat) if $\eta < \eta_c(\theta^*)$, equivalently $\eta\, h_{\max} < 2$.
- θ^* is η**-unstable** (bad, sharp) if $\eta > \eta_c(\theta^*)$, equivalently $\eta\, h_{\max} > 2$.

The set of η-stable interpolating solutions is $\mathcal{M}_\eta = \{\theta^* \in \mathcal{M} : h_{\max}(\theta^*) < 2/\eta\}$.

Definition 3 (Jacobian complexity)**.**

$$\mathcal{C}(\theta^*) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} h_i(\theta^*)} = \frac{\|\mathbf{J}(\theta^*)\|_F}{\sqrt{n}}.$$

Remark 4 (Physical intuition)*.* The quantity $h_i = \|\nabla_\theta f_{\theta^*}(x_i)\|^2$ measures the sensitivity of the model output at x_i with respect to the parameters. A good interpolating solution is one where no training sample requires the model to be in a high-sensitivity configuration. A bad interpolating solution has at least one sample x_i for which the model must precisely "align" its parameters to interpolate; such alignment cannot be maintained under the stochastic perturbations of SGD.

4 Assumptions

Assumption 1 (Local smoothness)**.** $f_\theta(x)$ is twice continuously differentiable in θ. In a neighborhood $B(\theta^*, r_0)$ of θ^*,

$$\|\nabla^2_\theta f_\theta(x_i)\|_{\text{op}} \le B_2 \quad \text{for all } i \in [n].$$

Assumption 2 (Jacobian independence)**.** $\{J_1, \ldots, J_n\}$ are linearly independent, so $\dim \mathcal{S} = n$.

Assumption 3 (Jacobian orthogonality)**.** $J_i^\top J_j = 0$ for all $i \neq j$.

Remark 5 (On the orthogonality assumption)*.* Assumption 3 is a technical assumption that enables complete decoupling of the stability analysis into n independent scalar problems. We emphasize that this is a genuine restriction: whitening the kernel matrix $K = \mathbf{J}\mathbf{J}^\top$ diagonalizes K but does *not* make the Jacobian vectors J_i orthogonal in parameter space (it only orthogonalizes them in the n-dimensional output space). Thus the NTK regime does not automatically satisfy Assumption 3. The qualitative conclusion—$\eta h_{\max} > 2$ implies instability—is expected to hold in the general (non-orthogonal) case by Furstenberg's theory of random matrix products [12], but rigorous quantitative bounds in the non-commutative setting remain open (see Section 8). We present fully rigorous proofs under Assumption 3, which has the virtue of yielding exact, closed-form Lyapunov exponents.

5 Main Results

We state the results in logical order. All proofs are deferred to Section 6.

Lemma 6 (Linearization of SGD near an interpolating solution)**.** *Let $\theta^* \in \mathcal{M}$ and suppose Assumption 1 holds. Define $\delta_t = \theta_t - \theta^*$. When $\|\delta_t\| < r_0$,*

$$\delta_{t+1} = A_{z_t}\,\delta_t + R_{z_t}(\delta_t),$$

where $A_i = I - \eta\, J_i J_i^\top$ and $\|R_i(\delta)\| \le C\eta\|\delta\|^2$ for a constant $C > 0$.

Lemma 7 (Spectrum of the rank-one update)**.** *The matrix $A_i = I - \eta J_i J_i^\top$ has eigenvalues:*

- *$1 - \eta h_i$, with eigenvector J_i;*
- *1 with multiplicity $d-1$, corresponding to the eigenspace $J_i^\perp$.*

The operator norm satisfies $\|A_i\|_{\mathrm{op}} = \max(1, |1-\eta h_i|)$. In particular, $|1-\eta h_i| \le 1$ if and only if $0 \le \eta h_i \le 2$.

Theorem 8 (Instability of bad interpolating solutions)**.** *Let $\theta^* \in \mathcal{M}$ and suppose Assumption 1–Assumption 3 hold, with $\eta\, h_{\max} > 2$. Then the maximal Lyapunov exponent of the linearized SGD at θ^* is*

$$\lambda(\eta, \theta^*) = \max_{k\in[n]} \frac{1}{n} \ln|1-\eta h_k| > 0.$$

*Consequently, θ^* is **almost surely unstable** under SGD: for almost every initial perturbation $\delta_0 \neq 0$ with nonzero component in $\mathcal{S}$, the linearized system satisfies $\|\delta_t\| \to \infty$.*

Theorem 9 (Stability of good interpolating solutions)**.** *Let $\theta^* \in \mathcal{M}$ and suppose Assumption 1–Assumption 3 hold, with $\eta\, h_{\max} < 2$. Then all Lyapunov exponents are negative:*

$$\lambda_k = \frac{1}{n}\ln|1-\eta h_k| < 0, \quad \forall\, k \in [n].$$

*The solution θ^* is **almost surely asymptotically stable** under linearized SGD: the active-subspace component decays exponentially,*

$$\|\delta_t|_{\mathcal{S}}\| \le \|\delta_0|_{\mathcal{S}}\| \cdot \exp\Big(\max_k \lambda_k \cdot t + o(t)\Big),$$

while $P_{\mathcal{S}^\perp}\delta_t = P_{\mathcal{S}^\perp}\delta_0$ remains constant.

Theorem 10 (Jacobian complexity controls Rademacher complexity)**.** *Let $\theta^* \in \mathcal{M}$ and suppose Assumption 1 holds. Define*

$$\mathcal{G}_\rho = \big\{x \mapsto f_\theta(x) - f_{\theta^*}(x) : \theta \in B(\theta^*, \rho)\big\}.$$

Then the empirical Rademacher complexity satisfies

$$\hat{\mathcal{R}}_S(\mathcal{G}_\rho) \le \frac{\rho\, \mathcal{C}(\theta^*)}{\sqrt{n}} + \frac{B_2\, \rho^2}{2}. \tag{1}$$

Theorem 11 (Generalization bound)**.** *Let $\theta^* \in \mathcal{M}$ and suppose Assumption 1 holds. Suppose the loss $\ell(\hat{y}, y)$ is L_ℓ-Lipschitz in its first argument and takes values in $[0, \bar{\ell}]$. Then with probability at least $1-\delta$ over the draw of S:*

$$L_{\mathrm{pop}}(\theta^*) \le \underbrace{L(\theta^*)}_{=0} + \frac{2L_\ell\, \rho\, \mathcal{C}(\theta^*)}{\sqrt{n}} + L_\ell\, B_2\, \rho^2 + \bar{\ell}\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{2}$$

Theorem 12 (Closed-loop implicit regularization—Main Theorem)**.** *Let Assumption 1–Assumption 3 hold and $\mathcal{M} \neq \varnothing$. Define $\mathcal{M}_\eta = \{\theta^* \in \mathcal{M} : h_{\max}(\theta^*) < 2/\eta\}$. Consider SGD with learning rate η.*

(I) Selection. *For every interpolating solution in $\mathcal{M} \setminus \mathcal{M}_\eta$ (i.e., with $\eta h_{\max} > 2$), the maximal Lyapunov exponent of the* linearized *SGD satisfies $\lambda = \frac{1}{n}\ln(\eta h_{\max} - 1) > 0$, so these solutions are almost surely unstable under linearized SGD. Conversely, for solutions in $\mathcal{M}_\eta$ (with $\eta h_{\max} < 2$), all Lyapunov exponents in $\mathcal{S}$ are negative and the $\mathcal{S}$-direction perturbation decays exponentially under linearized SGD. The boundary case $\eta h_{\max} = 2$ is degenerate ($\lambda = 0$) and is excluded from both sets.*

(II) Complexity control. *Every solution in $\mathcal{M}_\eta$ satisfies $h_i \leq 2/\eta$ for all i, hence*

$$\mathcal{C}(\theta^*)^2 = \frac{1}{n}\sum_{i=1}^{n} h_i \leq h_{\max} < \frac{2}{\eta}.$$

(III) Generalization. *With probability at least $1-\delta$, any SGD-selected solution $\theta^* \in \mathcal{M}_\eta$ satisfies*

$$L_{\text{pop}}(\theta^*) \leq \frac{2\sqrt{2}\, L_\ell\, \rho}{\sqrt{\eta\, n}} + L_\ell\, B_2\, \rho^2 + \bar{\ell}\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

(IV) Monotonicity. *$\eta_1 < \eta_2 \;\Rightarrow\; \mathcal{M}_{\eta_2} \subseteq \mathcal{M}_{\eta_1} \;\Rightarrow\; \sup_{\mathcal{M}_{\eta_2}} \mathcal{C} \leq \sup_{\mathcal{M}_{\eta_1}} \mathcal{C}$. Increasing η tightens the generalization bound.*

Theorem 13 (PAC-Bayes bound)**.** *Let $\theta^* \in \mathcal{M}$ and suppose Assumption 1 holds. Let the prior be $P = \mathcal{N}(0, \sigma_0^2 I_d)$ and the posterior be $Q = \mathcal{N}(\theta^*, \sigma^2 I_d)$. Suppose $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2 \in [0, \bar{\ell}]$. Then with probability at least $1-\delta$:*

$$\mathbb{E}_{\theta\sim Q}\big[L_{\text{pop}}(\theta)\big] \leq \frac{\sigma^2\, \mathcal{C}(\theta^*)^2}{2} + O(\sigma^4) + \sqrt{\frac{\frac{d}{2}\Big(\frac{\sigma^2}{\sigma_0^2} - 1 - \ln\frac{\sigma^2}{\sigma_0^2}\Big) + \frac{\|\theta^*\|^2}{2\sigma_0^2} + \ln\frac{2\sqrt{n}}{\delta}}{2n}}. \tag{3}$$

Corollary 14 (PAC-Bayes bound for SGD-selected solutions)**.** *Suppose SGD selects $\theta^* \in \mathcal{M}_\eta$, so that $\mathcal{C}(\theta^*)^2 \leq 2/\eta$. Set $\sigma^2 = \alpha\eta$ for a constant $\alpha > 0$. Then:*

$$\mathbb{E}_{\theta\sim Q}[L_{\text{pop}}(\theta)] \leq \alpha + O(\eta^2) + \sqrt{\frac{\frac{d}{2}\Big(\frac{\alpha\eta}{\sigma_0^2} - 1 - \ln\frac{\alpha\eta}{\sigma_0^2}\Big) + \frac{\|\theta^*\|^2}{2\sigma_0^2} + \ln\frac{2\sqrt{n}}{\delta}}{2n}}.$$

In particular, when $\alpha\eta \ll \sigma_0^2$ the KL term is dominated by $\frac{d}{2}\ln\frac{\sigma_0^2}{\alpha\eta}$, and the bound scales as $O\big(\sqrt{d\,\ln(1/\eta)\,/\,n}\big)$.

6 Proofs

6.1 Proof of Lemma 6 (Linearization)

Proof. **Step 1 (Residual expansion).** At θ^*, the residual is $r_i(\theta^*) = f_{\theta^*}(x_i) - y_i = 0$ by the interpolation condition. By Taylor expansion around θ^*:

$$r_i(\theta^* + \delta) = \underbrace{r_i(\theta^*)}_{=0} + J_i^\top \delta + O(\|\delta\|^2) = J_i^\top \delta + O(\|\delta\|^2).$$

Step 2 (Jacobian expansion). Similarly,

$$\nabla_\theta f_{\theta^*+\delta}(x_i) = J_i + \nabla_\theta^2 f_{\theta^*}(x_i)\,\delta + O(\|\delta\|^2) = J_i + O(\|\delta\|).$$

Step 3 (Loss gradient expansion). Since $\nabla \ell_i(\theta) = r_i(\theta) \cdot \nabla_\theta f_\theta(x_i)$, we have

$$\begin{aligned}\nabla \ell_i(\theta^* + \delta) &= \big(J_i^\top \delta + O(\|\delta\|^2)\big)\big(J_i + O(\|\delta\|)\big) \\ &= J_i(J_i^\top \delta) + O(\|\delta\|^2) \\ &= J_i J_i^\top \delta + O(\|\delta\|^2) \\ &= H_i\,\delta + O(\|\delta\|^2).\end{aligned}$$

For the $O(\|\delta\|^2)$ remainder: the cross terms are $(J_i^\top \delta) \cdot \nabla_\theta^2 f_{\theta^*}(x_i)\,\delta$ (of magnitude $\sqrt{h_i}\,\|\delta\| \cdot B_2\,\|\delta\|$) and $\frac{1}{2}(\delta^\top \nabla_\theta^2 f_{\theta^*}(x_i)\,\delta) \cdot J_i$ (of magnitude $\frac{B_2}{2}\|\delta\|^2\sqrt{h_i}$), plus higher-order terms. Therefore $\|\nabla \ell_i(\theta^* + \delta) - H_i\,\delta\| \le C_1\|\delta\|^2$ where $C_1 = B_2\big(\sqrt{h_{\max}} + \frac{B_2}{2}\sqrt{h_{\max}}\big)$.

Step 4 (SGD update).

$$\begin{aligned}\delta_{t+1} &= \delta_t - \eta\,\nabla \ell_{z_t}(\theta^* + \delta_t) \\ &= \delta_t - \eta\,H_{z_t}\,\delta_t - \eta \cdot O(\|\delta_t\|^2) \\ &= (I - \eta\,H_{z_t})\,\delta_t + R_{z_t}(\delta_t),\end{aligned}$$

where $\|R_i(\delta)\| \le \eta\,C_1\,\|\delta\|^2$. Setting $C = C_1$ completes the proof. □

6.2 Proof of Lemma 7 (Rank-one spectrum)

Proof. **Action on J_i:** $A_i\,J_i = (I - \eta\,J_i J_i^\top)\,J_i = J_i - \eta\,\|J_i\|^2\,J_i = (1 - \eta h_i)\,J_i$.

Action on $u \perp J_i$: $A_i\,u = (I - \eta\,J_i J_i^\top)\,u = u - \eta\,(J_i^\top u)\,J_i = u$.

Thus the eigenvalues are $1 - \eta h_i$ (eigenvector J_i) and 1 with multiplicity $d - 1$ (eigenspace $J_i^\perp$).

The operator norm is $\|A_i\|_{\text{op}} = \max(|1 - \eta h_i|, 1)$.

Finally, $|1 - \eta h_i| \le 1 \;\Leftrightarrow\; -1 \le 1 - \eta h_i \le 1 \;\Leftrightarrow\; 0 \le \eta h_i \le 2$. □

6.3 Proof of Theorem 8 (Instability)

Proof. **Step 1 (Coordinate decoupling under orthogonality).** By Assumption 3, $J_i^\top J_j = 0$ for $i \ne j$. Construct an orthonormal basis of $\mathcal{S}$:

$$e_k = \frac{J_k}{\sqrt{h_k}}, \quad k = 1, \dots, n.$$

Then $e_k^\top e_l = \frac{J_k^\top J_l}{\sqrt{h_k\,h_l}} = \delta_{kl}$.

Step 2 (Matrix representation of A_i on $\mathcal{S}$). The (k, l)-entry of A_i restricted to $\mathcal{S}$ in the basis $\{e_k\}$ is

$$(\hat{A}_i)_{kl} = e_k^\top A_i\,e_l = e_k^\top (I - \eta\,J_i J_i^\top)\,e_l = \delta_{kl} - \eta\,(e_k^\top J_i)(J_i^\top e_l).$$

By orthogonality, $e_k^\top J_i = \frac{J_k^\top J_i}{\sqrt{h_k}} = \sqrt{h_i}\,\delta_{ki}$. Therefore

$$(\hat{A}_i)_{kl} = \delta_{kl} - \eta\,h_i\,\delta_{ki}\,\delta_{li}.$$

This means $\hat{A}_i$ is a **diagonal matrix**:

$$(\hat{A}_i)_{kk} = \begin{cases} 1 - \eta h_i, & k = i, \\ 1, & k \ne i. \end{cases}$$

Step 3 (Coordinate evolution). Let $\xi_t \in \mathbb{R}^n$ denote the coordinates of $\delta_t|_{\mathcal{S}}$ in the basis $\{e_k\}$: $(\xi_t)_k = e_k^\top \delta_t$. By the diagonal structure of $\hat{A}_i$:

$$(\xi_{t+1})_k = (\hat{A}_{z_t})_{kk} \cdot (\xi_t)_k = \begin{cases} (1 - \eta h_k)(\xi_t)_k, & z_t = k, \\ (\xi_t)_k, & z_t \ne k. \end{cases}$$

Each coordinate k evolves **independently**: there is no coupling between different coordinates.

Step 4 (Product representation).

$$(\xi_T)_k = (\xi_0)_k \cdot \prod_{t=1}^{T} (\hat{A}_{z_t})_{kk}.$$

Taking logarithms:

$$\ln |(\xi_T)_k| = \ln |(\xi_0)_k| + \sum_{t=1}^{T} X_t^{(k)},$$

where $X_t^{(k)} = \ln |(\hat{A}_{z_t})_{kk}|$.

Step 5 (Distribution of $X_t^{(k)}$). Since $z_t \sim \text{Uniform}([n])$:

$$X_t^{(k)} = \begin{cases} \ln |1 - \eta h_k|, & \text{with probability } 1/n, \\ 0, & \text{with probability } (n-1)/n. \end{cases}$$

The mean is $\mathbb{E}[X_t^{(k)}] = \frac{1}{n} \ln |1 - \eta h_k| \triangleq \lambda_k$.

The variance is $\text{Var}(X_t^{(k)}) = \frac{1}{n} (\ln |1 - \eta h_k|)^2 - \frac{1}{n^2} (\ln |1 - \eta h_k|)^2 = \frac{n-1}{n^2} (\ln |1 - \eta h_k|)^2 < \infty$.

Step 6 (Strong law of large numbers). The sequence $\{X_t^{(k)}\}_{t \geq 1}$ is i.i.d. (since $z_1, z_2, \ldots$ are i.i.d.) with finite mean and variance. By the Kolmogorov strong law of large numbers:

$$\frac{1}{T} \sum_{t=1}^{T} X_t^{(k)} \xrightarrow{T \to \infty} \lambda_k = \frac{1}{n} \ln |1 - \eta h_k| \quad \text{almost surely.}$$

Hence

$$\frac{1}{T} \ln |(\xi_T)_k| \to \lambda_k \quad \text{almost surely.}$$

Step 7 (Sign of the Lyapunov exponent).

Case 1: $\eta h_k < 2$. Then $|1 - \eta h_k| < 1$, so $\ln |1 - \eta h_k| < 0$ and $\lambda_k < 0$.

Case 2: $\eta h_k > 2$. Then $|1 - \eta h_k| = \eta h_k - 1 > 1$, so $\ln |1 - \eta h_k| > 0$ and $\lambda_k > 0$.

Step 8 (Instability conclusion). Let $k^* = \arg\max_k h_k$. By hypothesis, $\eta h_{k^*} = \eta h_{\max} > 2$, so $\lambda_{k^*} > 0$. Provided $(\xi_0)_{k^*} \neq 0$ (i.e., the initial perturbation has a nonzero component in the J_{k^*} direction),

$$|(\xi_T)_{k^*}| = |(\xi_0)_{k^*}| \cdot \exp\big(\lambda_{k^*} \cdot T + o(T)\big) \to \infty \quad \text{almost surely.}$$

Therefore $\|\delta_T\| \geq |(\xi_T)_{k^*}| \to \infty$, and θ^* is almost surely unstable under the linearized SGD.

Step 9 (Maximal Lyapunov exponent). The maximal Lyapunov exponent of the system on $\mathcal{S}$ is

$$\lambda = \max_{k \in [n]} \lambda_k = \max_k \frac{1}{n} \ln |1 - \eta h_k| \geq \lambda_{k^*} = \frac{1}{n} \ln(\eta h_{\max} - 1) > 0. \qquad \square$$

6.4 Proof of Theorem 9 (Stability)

Proof. All steps follow the same structure as the proof of Theorem 8, with the sign reversed.

Steps 1–6: Identical to Steps 1–6 in the proof of Theorem 8. We obtain the per-coordinate Lyapunov exponent $\lambda_k = \frac{1}{n} \ln |1 - \eta h_k|$ for each $k \in [n]$.

Step 7 (All exponents are negative). By hypothesis, $\eta h_k \leq \eta h_{\max} < 2$ for all k. By Lemma 7, $|1 - \eta h_k| < 1$ for every k, so $\lambda_k < 0$ for all $k \in [n]$.

Step 8 (Asymptotic stability). Since $\frac{1}{T} \ln |(\xi_T)_k| \to \lambda_k < 0$ almost surely (by Step 6), we have $|(\xi_T)_k| \to 0$ for every k. Therefore

$$\|\delta_T|_{\mathcal{S}}\| = \sqrt{\sum_{k=1}^{n} (\xi_T)_k^2} \to 0 \quad \text{almost surely.}$$

The decay rate is governed by the slowest coordinate: $\lambda_{\max} = \max_k \lambda_k < 0$, so

$$\|\delta_T|_{\mathcal{S}}\| \le \|\delta_0|_{\mathcal{S}}\| \cdot \exp(\lambda_{\max} T + o(T)).$$

Behavior in $\mathcal{S}^\perp$. For any $v \in \mathcal{S}^\perp$, $A_i\, v = v$ by Lemma 7 (since $v \perp J_i$ for all i). Therefore $P_{\mathcal{S}^\perp}\delta_t = P_{\mathcal{S}^\perp}\delta_0$, which remains constant. □

6.5 Proof of Theorem 10 (Rademacher bound)

Proof. **Step 1 (Taylor expansion).** For $\theta = \theta^* + \delta$ with $\|\delta\| \le \rho$:

$$f_\theta(x_i) - f_{\theta^*}(x_i) = J_i^\top \delta + r_i(\delta),$$

where $r_i(\delta) = \frac{1}{2}\delta^\top \nabla_\theta^2 f_{\tilde\theta}(x_i)\,\delta$ for some $\tilde\theta$ on the segment $[\theta^*, \theta]$, by Taylor's theorem with Lagrange remainder. By Assumption 1, $|r_i(\delta)| \le \frac{B_2}{2}\|\delta\|^2 \le \frac{B_2\,\rho^2}{2}$.

Step 2 (Rademacher decomposition). Let $\sigma_1, \dots, \sigma_n$ be i.i.d. Rademacher random variables.

$$\begin{aligned}\hat{\mathcal{R}}_S(\mathcal{G}_\rho) &= \mathbb{E}_\sigma\left[\sup_{\|\delta\|\le\rho} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(J_i^\top\delta + r_i(\delta)\big)\right] \\ &\le \mathbb{E}_\sigma\left[\sup_{\|\delta\|\le\rho} \frac{1}{n}\sum_{i=1}^n \sigma_i\, J_i^\top\delta\right] + \frac{B_2\,\rho^2}{2}.\end{aligned}$$

The second inequality holds because $\sup_{\|\delta\|\le\rho} \frac{1}{n}\sum_i \sigma_i\big(J_i^\top\delta + r_i(\delta)\big) \le \sup_{\|\delta\|\le\rho} \frac{1}{n}\sum_i \sigma_i\, J_i^\top\delta + \sup_{\|\delta\|\le\rho} \frac{1}{n}\sum_i |\sigma_i\, r_i(\delta)| \le \sup_{\|\delta\|\le\rho} \frac{1}{n}\sum_i \sigma_i\, J_i^\top\delta + \frac{B_2\rho^2}{2}$, where the last term is deterministic since $|r_i(\delta)| \le \frac{B_2\rho^2}{2}$ regardless of σ.

Step 3 (Linear part computation).

$$\sup_{\|\delta\|\le\rho} \frac{1}{n}\sum_{i=1}^n \sigma_i\, J_i^\top\delta = \sup_{\|\delta\|\le\rho} \frac{1}{n}\left(\sum_{i=1}^n \sigma_i\, J_i\right)^\top \delta = \frac{\rho}{n}\left\|\sum_{i=1}^n \sigma_i\, J_i\right\|.$$

The last equality follows from $\sup_{\|\delta\|\le\rho} v^\top\delta = \rho\,\|v\|$, attained at $\delta^* = \rho\, v/\|v\|$.

Step 4 (Expectation estimate via Jensen's inequality). Since $\sqrt{\cdot}$ is concave, by Jensen's inequality:

$$\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i\, J_i\right\|\right] \le \sqrt{\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i\, J_i\right\|^2\right]}.$$

Expanding the squared norm:

$$\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i\, J_i\right\|^2\right] = \mathbb{E}_\sigma\left[\sum_{i=1}^n\sum_{j=1}^n \sigma_i\,\sigma_j\, J_i^\top J_j\right] = \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[\sigma_i\,\sigma_j]\cdot J_i^\top J_j.$$

Since σ_i are i.i.d. Rademacher: $\mathbb{E}[\sigma_i\,\sigma_j] = \delta_{ij}$ (because $\mathbb{E}[\sigma_i^2] = 1$ and $\mathbb{E}[\sigma_i\,\sigma_j] = \mathbb{E}[\sigma_i]\,\mathbb{E}[\sigma_j] = 0$ for $i \ne j$). Therefore

$$\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^n \sigma_i\, J_i\right\|^2\right] = \sum_{i=1}^n \|J_i\|^2 = \sum_{i=1}^n h_i = n\,\mathcal{C}(\theta^*)^2.$$

Hence $\mathbb{E}_\sigma\big[\|\sum_i \sigma_i\, J_i\|\big] \le \sqrt{n}\,\mathcal{C}(\theta^*)$.

Step 5 (Combining).

$$\hat{\mathcal{R}}_S(\mathcal{G}_\rho) \le \frac{\rho}{n}\cdot\sqrt{n}\,\mathcal{C}(\theta^*) + \frac{B_2\,\rho^2}{2} = \frac{\rho\,\mathcal{C}(\theta^*)}{\sqrt{n}} + \frac{B_2\,\rho^2}{2}.$$

□

6.6 Proof of Theorem 11 (Generalization bound)

Proof. **Step 1 (Zero training loss).** Since $\theta^* \in \mathcal{M}$, we have $f_{\theta^*}(x_i) = y_i$ for all i, so $L(\theta^*) = 0$.

Step 2 (Standard Rademacher generalization bound). Define the function class $\mathcal{H} = \{(x, y) \mapsto \ell(f_\theta(x), y) : \theta \in B(\theta^*, \rho)\}$. By the standard uniform convergence result (Mohri, Rostamizadeh, and Talwalkar, Theorem 3.1), with probability at least $1 - \delta$ over the draw of S:

$$\sup_{\theta \in B(\theta^*, \rho)} \left| L_{\text{pop}}(\theta) - L(\theta) \right| \le 2\, \hat{\mathcal{R}}_S(\mathcal{H}) + \bar{\ell} \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Step 3 (Lipschitz contraction). By the Ledoux–Talagrand contraction lemma, the L_ℓ-Lipschitz property of ℓ in its first argument implies

$$\hat{\mathcal{R}}_S(\mathcal{H}) \le L_\ell\, \hat{\mathcal{R}}_S(\mathcal{F}_\rho),$$

where $\mathcal{F}_\rho = \{x \mapsto f_\theta(x) : \theta \in B(\theta^*, \rho)\}$. Adding and subtracting the constant f_{θ^*} does not affect the Rademacher complexity (since $\mathbb{E}_\sigma[\sum_i \sigma_i\, c] = 0$ for any constant c), so $\hat{\mathcal{R}}_S(\mathcal{F}_\rho) = \hat{\mathcal{R}}_S(\mathcal{G}_\rho)$.

Step 4 (Substituting Theorem 10).

$$\begin{aligned} L_{\text{pop}}(\theta^*) &\le L(\theta^*) + 2\, L_\ell\, \hat{\mathcal{R}}_S(\mathcal{G}_\rho) + \bar{\ell} \sqrt{\frac{\ln(2/\delta)}{2n}} \\ &\le 0 + 2\, L_\ell \left(\frac{\rho\, \mathcal{C}(\theta^*)}{\sqrt{n}} + \frac{B_2\, \rho^2}{2} \right) + \bar{\ell} \sqrt{\frac{\ln(2/\delta)}{2n}} \\ &= \frac{2\, L_\ell\, \rho\, \mathcal{C}(\theta^*)}{\sqrt{n}} + L_\ell\, B_2\, \rho^2 + \bar{\ell} \sqrt{\frac{\ln(2/\delta)}{2n}}. \end{aligned}$$

□

6.7 Proof of Theorem 12 (Main Theorem—Closed-loop implicit regularization)

Proof. **Part (I): Selection.** This follows directly from Theorem 8 and Theorem 9. For $\theta^* \in \mathcal{M} \setminus \mathcal{M}_\eta$, we have $h_{\max}(\theta^*) \ge 2/\eta$, so $\eta h_{\max} \ge 2$. When $\eta h_{\max} > 2$, by Theorem 8 the maximal Lyapunov exponent is $\lambda \ge \frac{1}{n} \ln(\eta h_{\max} - 1) > 0$, and SGD escapes almost surely. For $\theta^* \in \mathcal{M}_\eta$, we have $\eta h_{\max} < 2$, and by Theorem 9 all Lyapunov exponents are negative and the $\mathcal{S}$-direction perturbation decays exponentially.

Part (II): Complexity control. For $\theta^* \in \mathcal{M}_\eta$, by definition $h_i \le h_{\max} < 2/\eta$ for every $i \in [n]$. Therefore

$$\mathcal{C}(\theta^*)^2 = \frac{1}{n} \sum_{i=1}^{n} h_i \le h_{\max} < \frac{2}{\eta}.$$

Part (III): Generalization. Substituting $\mathcal{C}(\theta^*) \le \sqrt{2/\eta}$ into (2):

$$L_{\text{pop}}(\theta^*) \le \frac{2\, L_\ell\, \rho\, \sqrt{2/\eta}}{\sqrt{n}} + L_\ell\, B_2\, \rho^2 + \bar{\ell} \sqrt{\frac{\ln(2/\delta)}{2n}} = \frac{2\sqrt{2}\, L_\ell\, \rho}{\sqrt{\eta\, n}} + L_\ell\, B_2\, \rho^2 + \bar{\ell} \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Part (IV): Monotonicity. By definition, $\mathcal{M}_\eta = \{\theta^* \in \mathcal{M} : h_{\max} < 2/\eta\}$. When η increases, $2/\eta$ decreases, making the threshold stricter. Hence $\eta_1 < \eta_2$ implies $\mathcal{M}_{\eta_2} \subseteq \mathcal{M}_{\eta_1}$. Consequently, $\sup_{\theta^* \in \mathcal{M}_{\eta_2}} \mathcal{C}(\theta^*) \le \sup_{\theta^* \in \mathcal{M}_{\eta_1}} \mathcal{C}(\theta^*)$. Moreover, $\sup_{\mathcal{M}_\eta} \mathcal{C} \le \sqrt{2/\eta}$ is monotonically decreasing in η. □

6.8 Proof of Theorem 13 (PAC-Bayes bound)

Proof. **Step 1 (PAC-Bayes inequality).** By the standard PAC-Bayes theorem (McAllester 1999; Catoni 2007), with probability at least $1 - \delta$ over the draw of S:

$$\mathbb{E}_{\theta \sim Q}[L_{\text{pop}}(\theta)] \le \mathbb{E}_{\theta \sim Q}[L(\theta)] + \sqrt{\frac{\text{KL}(Q \| P) + \ln(2\sqrt{n}/\delta)}{2n}}.$$

Step 2 (Computing $\mathbb{E}_Q[L]$). Write $\theta = \theta^* + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. By Taylor expansion at θ^*:

$$f_\theta(x_i) - y_i = J_i^\top \epsilon + \tfrac{1}{2}\epsilon^\top \nabla_\theta^2 f_{\theta^*}(x_i)\,\epsilon + O(\|\epsilon\|^3).$$

Squaring:

$$(f_\theta(x_i) - y_i)^2 = (J_i^\top \epsilon)^2 + 2(J_i^\top \epsilon)\cdot \tfrac{1}{2}\epsilon^\top \nabla_\theta^2 f_{\theta^*}(x_i)\,\epsilon + O(\|\epsilon\|^4).$$

For the cross term: $\mathbb{E}\big[(J_i^\top \epsilon)\cdot \epsilon^\top \nabla^2 f_{\theta^*}(x_i)\,\epsilon\big] = \sum_{a,b,c}(J_i)_a\,(\nabla^2 f)_{bc}\,\mathbb{E}[\epsilon_a\,\epsilon_b\,\epsilon_c] = 0$, since all odd moments of the Gaussian distribution vanish.

For the leading term: $\mathbb{E}[(J_i^\top \epsilon)^2] = \sigma^2\,\|J_i\|^2 = \sigma^2\,h_i$.

Therefore $\mathbb{E}[\ell_i(\theta)] = \frac{\sigma^2 h_i}{2} + O(\sigma^4)$, and

$$\mathbb{E}_Q[L] = \frac{\sigma^2}{2n}\sum_{i=1}^n h_i + O(\sigma^4) = \frac{\sigma^2\,\mathcal{C}(\theta^*)^2}{2} + O(\sigma^4).$$

Step 3 (KL divergence). The KL divergence between two Gaussians is given by the standard formula:

$$\mathrm{KL}\big(\mathcal{N}(\theta^*, \sigma^2 I)\,\|\,\mathcal{N}(0, \sigma_0^2 I)\big) = \frac{d}{2}\left(\frac{\sigma^2}{\sigma_0^2} - 1 - \ln\frac{\sigma^2}{\sigma_0^2}\right) + \frac{\|\theta^*\|^2}{2\sigma_0^2}.$$

Step 4 (Assembly). Substituting Steps 2 and 3 into the PAC-Bayes inequality of Step 1 yields (3). □

6.9 Proof of Corollary 14 (PAC-Bayes for SGD-selected solutions)

Proof. By Theorem 12 (II), $\mathcal{C}(\theta^*)^2 \le 2/\eta$ for any $\theta^* \in \mathcal{M}_\eta$. Setting $\sigma^2 = \alpha\eta$ in Theorem 13:

$$\mathbb{E}_Q[L] = \frac{\alpha\eta}{2}\,\mathcal{C}(\theta^*)^2 + O(\alpha^2\eta^2) \le \frac{\alpha\eta}{2}\cdot\frac{2}{\eta} + O(\eta^2) = \alpha + O(\eta^2).$$

The KL term becomes $\frac{d}{2}\left(\frac{\alpha\eta}{\sigma_0^2} - 1 - \ln\frac{\alpha\eta}{\sigma_0^2}\right) + \frac{\|\theta^*\|^2}{2\sigma_0^2}$. When $\alpha\eta \ll \sigma_0^2$, we have $\frac{\alpha\eta}{\sigma_0^2} \approx 0$ and $-\ln\frac{\alpha\eta}{\sigma_0^2} = \ln\frac{\sigma_0^2}{\alpha\eta} \gg 1$, so the KL term is dominated by $\frac{d}{2}\ln\frac{\sigma_0^2}{\alpha\eta}$. The full bound then scales as $\alpha + O(\eta^2) + O\left(\sqrt{\frac{d\,\ln(1/\eta)}{n}}\right)$. □

7 Reparameterization Invariance

Remark 15 (Reparameterization invariance of Σ/H). The ratio Σ/H (and the interpolation-regime analogue $\mathcal{C}^2/h_{\max}$) is invariant under smooth reparameterization. This distinguishes our complexity measure from flatness (small H) or small noise (Σ) individually, both of which can be altered by reparameterization without changing the function [2].

Proof. Let $\theta = g(\varphi)$ be a smooth bijection with $\varphi^* = g^{-1}(\theta^*)$. In the new parameterization:

$$\tilde{H} = L''(g(\varphi^*))\cdot[g'(\varphi^*)]^2 = H\cdot(g')^2,$$

$$\tilde{\Sigma} = \mathbb{E}_i\big[(\ell_i'(g(\varphi^*)))^2\big]\cdot[g'(\varphi^*)]^2 = \Sigma\cdot(g')^2.$$

Therefore $\tilde{\Sigma}/\tilde{H} = \Sigma\cdot(g')^2/(H\cdot(g')^2) = \Sigma/H$.

In the multivariate case, $\mathrm{tr}(H^{-1}\Sigma)$ is similarly invariant: the Jacobian factors from the change of variables cancel in the product $H^{-1}\Sigma$.

Implication. The generalization gap $\approx \Sigma/(H(n-1))$ in the non-interpolation regime is an intrinsic geometric quantity. In the interpolation regime, the per-sample curvature condition $h_{\max} < 2/\eta$ is coordinate-dependent (since h_i involves the parameter-space norm of J_i), but the *stability criterion* (whether the Lyapunov exponent is positive or negative) is invariant: a reparameterization that changes h_i also changes the effective learning rate in the new coordinates, preserving the sign of λ_k.

8 Discussion

8.1 Connection to the non-interpolation regime

In the non-interpolation regime ($L(\theta^*) > 0$), the gradient noise $\Sigma(\theta^*) > 0$ enables an SDE approximation whose Fokker–Planck stationary distribution $p \propto (1/\Sigma)\exp(-2V_{\text{eff}}/\tau)$ biases SGD toward minima with low Σ. At interpolation ($L(\theta^*) = 0$), $\Sigma = 0$ and the SDE degenerates; the regularization mechanism shifts from first-order noise to the second-order structure $H_i = J_i J_i^\top$ analyzed in this paper. Both mechanisms reflect per-sample sensitivity of f_θ, but through different mathematical tools (continuous SDE vs. discrete random matrix products).

8.2 Testable predictions

Corollary 16 (Testable predictions from the theory)**.** *Under the assumptions of Theorem 12:*

(G1) $h_{\max}$ ***predicts generalization.*** *After training to interpolation, measure* $h_{\max} = \max_i \|\nabla_\theta f_{\theta^*}(x_i)\|^2$. *Models trained with different* η *should exhibit a positive correlation between* $h_{\max}$ *and test error, stronger than the correlation with* $\operatorname{tr}(H)$ *or other flatness measures.*

(G2) ***Critical phase transition.*** *For a fixed model, training loss should transition sharply from zero to nonzero at* $\eta \approx 2/h_{\max}$.

(G3) ***Monotone effect of*** η**.** *In the interpolation regime, increasing* η *should monotonically decrease* $\mathcal{C} = \|\mathbf{J}\|_F/\sqrt{n}$ *(up to the point where SGD fails to converge).*

Conjecture 17 (Beyond the orthogonality assumption)**.** *(G3)* ***Noise structure matters more than magnitude.*** *Replacing SGD noise with isotropic Gaussian noise of the same magnitude should degrade generalization performance, because the structured per-sample filtering mechanism (which depends on the alignment of* J_i *directions) is destroyed.*

(G4) ***F4 observability.*** *In small-data, large-model settings, overfitting solutions may have* lower Σ *than generalizing solutions, because memorization can lead to aligned per-sample gradients.*

8.3 Limitations

We discuss the main limitations of the present analysis.

Jacobian orthogonality (Assumption 3). The complete decoupling of coordinate dynamics relies on the orthogonality $J_i^\top J_j = 0$. While this is approximately satisfied in the NTK regime (after whitening), real neural networks need not operate in this regime. Removing Assumption 3 requires the theory of products of non-commuting random matrices (Furstenberg–Kesten theory), which yields qualitatively similar conclusions but with less explicit constants.

Linearization gap. Lemma 6 establishes that the SGD dynamics near θ^* are well-approximated by the linear system $\delta_{t+1} = A_{z_t}\,\delta_t$ only when $\|\delta_t\|$ is small. The stability conclusions hold in a neighborhood of θ^*; the size of this neighborhood depends on the ratio of the linearization error $C\eta\|\delta\|^2$ to the linear term. A complete quantitative analysis of the basin of attraction requires nonlinear stability theory.

Behavior in $\mathcal{S}^\perp$. The linearized SGD has no force in the orthogonal complement $\mathcal{S}^\perp$: perturbations in this subspace remain constant. The full nonlinear dynamics may drift in $\mathcal{S}^\perp$ due to higher-order effects, but characterizing this drift requires going beyond the linearization.

Global trajectory. Both Theorem 8 and Theorem 9 are local stability results. The global trajectory of SGD from random initialization to an element of $\mathcal{M}_\eta$ involves traversing a complex loss landscape, and its analysis remains open. Our results establish *which* solutions SGD can stably reside at, not *how* it gets there.

8.4 Open problems

1. **Removing the orthogonality assumption.** Establish the instability threshold $\eta h_{\max} > 2$ for general (non-orthogonal) Jacobian configurations using Furstenberg's theory, with quantitative bounds on the Lyapunov exponent.

2. **Global convergence.** Characterize the full trajectory of SGD from random initialization to $\mathcal{M}_\eta$, combining the transient (non-interpolation) phase with the local stability analysis.

3. **Discrete–continuous unification.** Develop a unified mathematical framework that handles the $\Sigma \to 0$ transition between the non-interpolation and interpolation regimes.

4. **Architecture-dependent analysis.** Determine for which architectures and data distributions the condition "low $h_{\max}$ implies good generalization" (the analogue of $\neg F4$ in the notation of Section 8.1) is satisfied.

5. **Mini-batch and momentum.** Extend the per-sample Lyapunov analysis to mini-batch SGD and SGD with momentum, determining how the critical learning rate η_c is modified.

9 Related Work

Implicit regularization of SGD. Smith and Le [5] observed that larger learning rates and smaller batch sizes improve generalization, and proposed the SDE approximation $d\theta = -\nabla L\, dt + \sqrt{(\eta/B)\Sigma}\, dW_t$ to explain this phenomenon. Chaudhari and Soatto [1] analyzed the Fokker–Planck equation of this SDE and showed that the stationary distribution $p \propto (1/\Sigma)\exp(-2V_{\text{eff}}/\tau)$ biases SGD toward minima with low noise covariance.

Edge of stability and sharpness reduction. Wu, Wang, and Su [7] studied the phenomenon of sharpness reduction during training, where the largest eigenvalue of the Hessian tends to decrease toward $2/\eta$. Wu and Su [6] proved that SGD cannot stably converge to solutions where $\operatorname{tr}(H) > 2/\eta$. Our work refines this by showing that the relevant condition is per-sample: $h_{\max} > 2/\eta$ (not the aggregate $\operatorname{tr}(H) > 2/\eta$), and by closing the loop from stability to generalization. The distinction matters because $\operatorname{tr}(H) = \sum_i h_i/n \cdot n = \sum_i h_i$ (for appropriately defined H), and $\operatorname{tr}(H) \le 2/\eta$ is compatible with having one very large h_k compensated by many small ones, whereas our condition is not.

Reparameterization invariance. Dinh et al. [2] demonstrated that Hessian-based sharpness measures are not invariant under reparameterization and constructed explicit examples where "sharp" minima (in terms of Hessian eigenvalues) generalize well. Our Remark 15 shows that the ratio Σ/H (and its interpolation-regime analogue) is reparameterization-invariant, providing a more robust complexity measure.

Li et al. (2021). Li et al. [4] studied the implicit bias of SGD toward flat minima in the context of label noise SGD, providing theoretical evidence that SGD noise structure matters for generalization.

Chang and Khanna (2025). Chang and Khanna [8] studied the relationship between SGD dynamics and generalization in overparameterized models, providing complementary perspectives on the role of stochastic noise.

References

[1] P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *Proceedings of ICLR*, 2018.

[2] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In *Proceedings of ICML*, pages 1019–1028, 2017.

[3] Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of ICML*, pages 2101–2110, 2017.

[4] Z. Li, T. Wang, and S. Arora. What happens after SGD reaches zero loss?—A mathematical framework. In *Proceedings of ICLR*, 2021.

[5] S. L. Smith and Q. V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *Proceedings of ICLR*, 2018.

[6] L. Wu and W. J. Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *Proceedings of ICML*, 2023.

[7] L. Wu, X. Wang, and W. J. Su. Does the sharpness of SGD iterates control generalization? *arXiv preprint arXiv:2207.06680*, 2022.

[8] H. Chang and R. Khanna. Implicit regularization of stochastic gradient descent in overparameterized models. *arXiv preprint*, 2025.

[9] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of COLT*, pages 164–170, 1999.

[10] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. IMS Lecture Notes, vol. 56. Institute of Mathematical Statistics, 2007.

[11] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

[12] H. Furstenberg and H. Kesten. Products of random matrices. *Annals of Mathematical Statistics*, 31(2):457–469, 1960.