When Does SGD Fail to Find Good Solutions? Four Failure Modes and Necessary Conditions for Implicit Regularization

CΛ / Lightman
Lightman.chang@gmail.com

May 5, 2026

Abstract

Stochastic gradient descent (SGD) is widely observed to find solutions that generalize well, a phenomenon attributed to its implicit regularization through gradient noise. Yet the precise boundaries of this mechanism remain unclear: *when does SGD fail to find good solutions?*

We identify four explicit failure modes, each demonstrated by a concrete counterexample with quantitative analysis. **F1**: when the learning rate is too small, exponentially long mixing times trap SGD in bad basins. **F2**: when the learning rate is too large, a positive Lyapunov exponent causes divergence. **F3**: when the noise-to-curvature ratios Σ/H are identical across minima, SGD's selection mechanism degenerates. **F4** (our main result): even when the dynamics mix and are stable, the noise-driven preference for low gradient variance can systematically select solutions that generalize *poorly*—a noise-generalization reversal. This occurs because good solutions that leverage diverse features exhibit high per-sample gradient variance, while bad solutions that memorize exhibit low variance.

From these four failure modes we derive four necessary conditions (N1–N4) that must hold simultaneously for SGD's implicit regularization to succeed. The conjunction is not shown to be sufficient, and we classify this gap as an open problem. Our results delineate the fundamental limits of noise-driven implicit regularization and highlight settings where explicit regularization is indispensable.

1 Introduction

Modern deep learning relies heavily on stochastic gradient descent and its variants. A striking empirical observation is that SGD often finds solutions that generalize well to unseen data, even in over-parameterized settings where many interpolating solutions exist. This phenomenon has been attributed to SGD's *implicit regularization*: the stochasticity inherent in mini-batch sampling biases the iterates toward solutions with favorable properties such as low gradient noise variance or low Jacobian complexity [Smith and Le, 2018, Chaudhari and Soatto, 2018, Li et al., 2021].

The mechanisms underlying this bias are increasingly well understood. In the non-interpolation regime, the multiplicative structure of gradient noise induces an effective potential V_{eff} that differs from the training loss, favoring regions of low noise variance [Horsthemke and Lefever, 1984, Chaudhari and Soatto, 2018]. In the interpolation regime, solutions whose per-sample Hessians exceed a critical threshold $2/\eta$ are destabilized by discrete dynamics [Wu and Su, 2023].

However, most analyses focus on *when* implicit regularization succeeds, leaving an equally important question open: **when does it fail?** Understanding failure is essential for two reasons. First, it identifies settings where practitioners must employ explicit regularization. Second, it sharpens our theoretical understanding by delineating necessary conditions that any successful account of SGD's generalization must satisfy.

Our contributions. We present four explicit failure modes (F1–F4), each as a concrete construction with quantitative analysis:

F1 Insufficient mixing: when η is too small, Kramers escape times grow exponentially, trapping SGD in bad basins for any feasible training duration (Proposition 3.1).

F2 Discrete instability: when η is too large, the top Lyapunov exponent becomes positive and SGD diverges (Proposition 3.2).

F3 Degeneracy of selection: when Σ/H is identical at all minima, SGD's noise-driven preference provides no selective advantage (Proposition 3.3).

F4 Noise-generalization reversal: the good solution has high gradient noise (due to feature diversity), causing SGD to *systematically* converge to the bad solution (Theorem 3.4).

Failure mode F4 is our main result. It reveals a fundamental limit: SGD's preference for low gradient variance can conflict with generalization when good solutions use diverse per-sample features and bad solutions memorize uniformly.

From these four failure modes we extract four necessary conditions (N1–N4) for implicit regularization (Theorem 4.1). We emphasize that their conjunction is not proven sufficient, and we classify the sufficiency question as an open problem.

Paper structure. Section 2 introduces notation, the SGD setup, and background on noise selection in one dimension (as known results). Section 3 presents the four failure modes with full constructions and proofs. Section 4 states the necessary conditions framework. Section 5 establishes reparameterization invariance of Σ/H and its implications. Section 6 discusses practical implications, connections to prior work, and open questions. Section 7 concludes.

2 Preliminaries

2.1 SGD Setup

Let $\Theta \subseteq \mathbb{R}^d$ be the parameter space and let $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta)$ be the empirical risk, where each ℓ_i is the loss on the i-th training example. SGD with learning rate $\eta > 0$ iterates

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

2.2 Gradient Noise Variance

The gradient noise variance at θ is defined as

$$\Sigma(\theta) = \frac{1}{n}\sum_{i=1}^{n} \big[\ell_i'(\theta)\big]^2 - \big[L'(\theta)\big]^2. \tag{2}$$

At a local minimum θ^* where $L'(\theta^*) = 0$, this simplifies to $\Sigma(\theta^*) = \frac{1}{n}\sum_{i=1}^{n}[\ell_i'(\theta^*)]^2$. In the multi-dimensional setting, $\Sigma(\theta)$ denotes the covariance matrix of the stochastic gradient; in one dimension it reduces to a non-negative scalar.

2.3 Convergence and Good Solutions

We distinguish three levels of convergence:

(C1) *Optimization convergence*: $L(\theta_t) \to L^*$ almost surely.

(C2) *Distributional convergence*: $\theta_t \xrightarrow{d} \mu_\eta$ for some stationary measure μ_η.

(C3) *Absorption convergence*: $\theta_t \to \theta^*$ almost surely for some local minimum θ^*.

A solution θ^* is considered "good" if it satisfies one or more of the following criteria:

(G1) *Low noise ratio*: $\Sigma(\theta^*)/H(\theta^*)$ is small relative to other local minima, where $H(\theta^*) = L''(\theta^*)$.

(G2) *Low Jacobian complexity*: in the interpolation regime, $\mathcal{C}(\theta^*) = \|J\|_F/\sqrt{n}$ is small.

(G3) *Leave-one-out stability*: the stability coefficient β is small.

(G4) *Generalization*: the gap $L_{\text{test}} - L_{\text{train}}$ is small.

A central theme of this paper is that SGD's noise-driven bias favors (G1), and that the chain (G1)⇒(G3)⇒(G4) holds under mild regularity conditions—but that (G1) can *conflict* with (G4) in adversarial constructions (failure mode F4).

2.4 Background: SGD Noise Selection in One Dimension

We review known results on the stationary distribution of SGD in one dimension, which form the analytical foundation for our counterexamples. These results are not new; we state them for completeness and to fix notation.

Proposition 2.1 (FPE steady state; Horsthemke and Lefever 1984, Chaudhari and Soatto 2018)**.** *Consider the one-dimensional SDE approximation of SGD with temperature $\tau = \eta$:*

$$d\theta = -L'(\theta)\,\mathrm{d}t + \sqrt{\tau\,\Sigma(\theta)}\,\mathrm{d}W_t.$$

Assume $L, \ell_i \in C^2(\mathbb{R})$, L is coercive ($L(\theta) \to +\infty$ as $|\theta| \to \infty$), and $\Sigma(\theta)$ is strictly positive and bounded: $0 < \sigma_{\min}^2 \le \Sigma(\theta) \le \sigma_{\max}^2$. Then the Fokker–Planck equation

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial \theta}\big[L'(\theta)\,p\big] + \frac{\tau}{2}\frac{\partial^2}{\partial \theta^2}\big[\Sigma(\theta)\,p\big]$$

has a unique stationary solution

$$p(\theta) = \frac{1}{Z}\cdot\frac{1}{\Sigma(\theta)}\cdot\exp\Big(-\frac{2}{\tau}\int^{\theta}\frac{L'(s)}{\Sigma(s)}\,\mathrm{d}s\Big), \tag{3}$$

or equivalently $p(\theta) \propto \exp\big(-\frac{2}{\tau}V_{\text{eff}}(\theta)\big)$, where the effective potential is

$$V_{\text{eff}}(\theta) = \int^{\theta}\frac{L'(s)}{\Sigma(s)}\,\mathrm{d}s + \frac{\tau}{2}\ln\Sigma(\theta). \tag{4}$$

Proof sketch. Setting $\partial p/\partial t = 0$ in the Fokker–Planck equation, define the probability current $\mathcal{J} = -L'p - \frac{\tau}{2}(\Sigma p)'$. Since $\frac{d\mathcal{J}}{d\theta} = 0$, the current $\mathcal{J}$ is constant. Coercivity of L and boundedness of Σ force $p(\theta) \to 0$ as $|\theta| \to \infty$, which implies $\mathcal{J} \equiv 0$. Expanding $\mathcal{J} = 0$ gives

$$L'p + \frac{\tau}{2}\big[\Sigma' p + \Sigma p'\big] = 0 \quad\Longrightarrow\quad \frac{p'}{p} = -\frac{2L'}{\tau\Sigma} - \frac{\Sigma'}{\Sigma}.$$

Recognizing the right-hand side as $-\frac{d}{d\theta}\big[\frac{2}{\tau}\int^{\theta}\frac{L'}{\Sigma}\,\mathrm{d}s + \ln\Sigma\big]$, we integrate to obtain $\ln p = -\frac{2}{\tau}\int^{\theta}\frac{L'}{\Sigma}\,\mathrm{d}s - \ln\Sigma + C$. Exponentiating yields (3). The normalization constant $Z < \infty$ is guaranteed by coercivity and boundedness of Σ. □

The key feature of (3) is the prefactor $1/\Sigma(\theta)$: the stationary distribution assigns higher density to regions of *low* gradient noise variance. This is the mathematical basis of SGD's noise-driven selection.

Proposition 2.2 (Generalization gap via leave-one-out; known result)**.** *At a local minimum θ^* of L with $H = L''(\theta^*) > 0$, the leave-one-out generalization gap satisfies*

$$\mathrm{Gap}(\theta^*) \approx \frac{\Sigma(\theta^*)}{H(\theta^*)(n-1)}. \tag{5}$$

Proof sketch. Removing the j-th sample gives $L'_{-j}(\theta^*) = -\ell'_j(\theta^*)/(n-1)$ and $L''_{-j}(\theta^*) = (nH - \ell''_j(\theta^*))/(n-1) \equiv H_{-j}$. By the implicit function theorem, $\theta^*_{-j} - \theta^* \approx \ell'_j(\theta^*)/((n-1)H_{-j})$. The leave-one-out loss difference is $\ell_j(\theta^*_{-j}) - \ell_j(\theta^*) \approx [\ell'_j(\theta^*)]^2/((n-1)H_{-j})$. Averaging over j and using $L'(\theta^*) = 0$ to identify $\frac{1}{n}\sum_j [\ell'_j(\theta^*)]^2 = \Sigma(\theta^*)$, we obtain $\mathrm{Gap} \approx \Sigma(\theta^*)/((n-1)H(\theta^*))$. □

3 Four Failure Modes

We now present the four failure modes. Each is an explicit construction with quantitative analysis, demonstrating a specific mechanism by which SGD's implicit regularization breaks down. We label the violations of necessary conditions to be derived in Section 4.

3.1 F1: Insufficient Mixing

Proposition 3.1 (Failure by insufficient mixing)**.** *There exists a one-dimensional loss landscape with two local minima—one good and one bad—such that for $\eta = 0.01$, the expected time for SGD to escape the bad basin exceeds 10^{1087} iterations.*

Construction. Consider a symmetric double-well potential with barrier height $\Delta V = 100$ separating two wells:

- Well A (good): Hessian $H_A = 2$, noise variance $\Sigma_A = 1$.
- Well B (bad): Hessian $H_B = 8$, noise variance $\Sigma_B = 8$.

The noise ratios satisfy $\Sigma_A/H_A = 0.5 < \Sigma_B/H_B = 1$, so the effective potential (4) favors well A (the good solution). However, the barrier is so high that SGD cannot reach the steady state in any feasible number of iterations. SGD is initialized in well B.

Proof. We apply the Kramers escape formula for the SDE with multiplicative noise [Kramers, 1940]. The effective barrier for escape from well B to well A is computed from the effective potential (4).

Near well B, V_{eff} has a local minimum. The barrier in the effective potential from B to the saddle point is

$$\Delta U_{B\to\text{saddle}} = \frac{\Delta V}{\Sigma_B} = \frac{100}{8} = 12.5.$$

The Kramers escape time from well B is

$$\tau_{B\to A} \sim \frac{1}{\eta}\exp\Big(\frac{2\,\Delta U_{B\to\text{saddle}}}{\eta}\Big) = \frac{1}{\eta}\exp\Big(\frac{25}{\eta}\Big).$$

Setting $\eta = 0.01$:

$$\tau_{B\to A} \sim 100 \cdot \exp(2500) = 100 \cdot 10^{2500/\ln 10} \approx 10^{1087}.$$

This far exceeds any feasible training budget.

The key point is structural: regardless of the precise constant in the exponent, for any barrier $\Delta > 0$, the Kramers time scales as $\exp(\Theta(1/\eta))$, which grows super-exponentially as $\eta \to 0$. With $\eta = 0.01$, the SGD iterates remain trapped in well B for any practical number of iterations, despite well A being the preferred basin under the stationary distribution.

This failure violates the mixing condition: the total training time ηT is negligible compared to the mixing time τ_{mix}. □

Which condition is violated. Condition N1 ($\eta T \gg \tau_{\text{mix}}$) is violated. SGD has not reached its stationary distribution and remains trapped in the initialization basin.

3.2 F2: Discrete Instability

Proposition 3.2 (Failure by discrete instability)**.** *There exists a quadratic loss with two per-sample Hessians such that $\eta = 0.15$ yields a positive top Lyapunov exponent, causing SGD to diverge.*

Construction. Let $n = 2$ with per-sample losses $\ell_1(\theta) = \frac{1}{2} \cdot 18 \cdot \theta^2$ and $\ell_2(\theta) = \frac{1}{2} \cdot 2 \cdot \theta^2$, so that $L(\theta) = \frac{1}{2} \cdot 10 \cdot \theta^2$. The per-sample Hessians are $h_1 = 18$ and $h_2 = 2$. The stability threshold is $\eta_c = 2/h_{\max} = 2/18 \approx 0.111$. We set $\eta = 0.15 > \eta_c$.

Proof. The SGD iteration at the origin is $\theta_{t+1} = (1 - \eta\, h_{z_t})\, \theta_t$, where $z_t \in \{1, 2\}$ with equal probability. The multiplicative factors are

$$1 - \eta\, h_1 = 1 - 0.15 \times 18 = -1.7, \qquad 1 - \eta\, h_2 = 1 - 0.15 \times 2 = 0.7.$$

Taking absolute values and logarithms:

$$\ln|1 - \eta\, h_1| = \ln 1.7, \qquad \ln|1 - \eta\, h_2| = \ln 0.7.$$

The top Lyapunov exponent is the expected log-contraction rate:

$$\Lambda = \frac{1}{2}\big(\ln 1.7 + \ln 0.7\big) = \frac{1}{2}\ln(1.7 \times 0.7) = \frac{1}{2}\ln 1.19 \approx \frac{1}{2} \times 0.1740 = 0.0870.$$

Since $\Lambda > 0$, the strong law of large numbers gives

$$\frac{1}{T}\ln|\theta_T| = \frac{1}{T}\ln|\theta_0| + \frac{1}{T}\sum_{t=0}^{T-1} \ln|1 - \eta\, h_{z_t}| \xrightarrow{a.s.} \Lambda > 0.$$

Therefore $|\theta_t| \sim e^{0.087\, t} \to \infty$ almost surely: SGD diverges. □

Which condition is violated. Condition N2 ($\eta < 2/\lambda_{\max}$) is violated. Here $\lambda_{\max} = h_1 = 18$ and $\eta = 0.15 > 2/18$.

3.3 F3: Degeneracy of Selection

Proposition 3.3 (Failure by degenerate selection)**.** *There exist two local minima, one good and one bad, with $\Sigma_A/H_A = \Sigma_B/H_B$, such that the SGD stationary distribution assigns equal probability to both: $\pi(A) \approx \pi(B) \approx 1/2$.*

Construction. Consider a symmetric double-well loss (symmetric about the midpoint between wells A and B) with:

- Well A (good generalization): H_A, Σ_A, with $\Sigma_A/H_A = 1$.
- Well B (bad generalization): H_B, Σ_B, with $\Sigma_B/H_B = 1$.

The loss values $L_A = L_B$ and the landscape is symmetric. Despite well A having lower generalization gap (due to different higher-order structure or sample size dependence not captured by Σ/H alone), the noise ratios are equal.

Proof. By Proposition 2.1, the stationary density ratio at the two minima is

$$\frac{p(\theta_A)}{p(\theta_B)} = \frac{\Sigma_B}{\Sigma_A} \cdot \exp\Big(-\frac{2}{\tau}\int_{\theta_B}^{\theta_A} \frac{L'(s)}{\Sigma(s)}\,\mathrm{d}s\Big).$$

In the symmetric case, the integral $\int_{\theta_B}^{\theta_A} L'(s)/\Sigma(s)\,\mathrm{d}s = 0$ by antisymmetry. Hence

$$\frac{p(\theta_A)}{p(\theta_B)} = \frac{\Sigma_B}{\Sigma_A}.$$

Since $\Sigma_A/H_A = \Sigma_B/H_B$ with equal Hessians in the symmetric construction (one can set $H_A = H_B$), we have $\Sigma_A = \Sigma_B$, giving $p(\theta_A)/p(\theta_B) = 1$.

More generally, even without $H_A = H_B$, the condition $\Sigma_A/H_A = \Sigma_B/H_B$ means the effective potential curvatures $V''_{\text{eff}}(\theta_k) \approx H_k/\Sigma_k$ are equal at both wells. In a Laplace approximation of the stationary distribution near each well,

$$\pi(k) \propto \frac{1}{\Sigma_k}\sqrt{\frac{2\pi\tau}{2H_k/\Sigma_k}} = \frac{1}{\Sigma_k}\sqrt{\frac{\pi\tau\Sigma_k}{H_k}} = \frac{1}{\sqrt{\Sigma_k}}\sqrt{\frac{\pi\tau}{H_k}},$$

and with $\Sigma_A/H_A = \Sigma_B/H_B$ and the symmetric construction $L_A = L_B$:

$$\frac{\pi(A)}{\pi(B)} = \frac{\sqrt{\Sigma_B/H_B}}{\sqrt{\Sigma_A/H_A}} = 1.$$

Thus $\pi(A) \approx \pi(B) \approx 1/2$. SGD selects the good solution with probability no better than a fair coin. □

Which condition is violated. Condition N3 (the noise ratios Σ/H must differ across minima) is violated. When all minima share the same ratio, the noise-driven selection mechanism provides no bias toward good solutions.

3.4 F4: Noise-Generalization Reversal (Main Result)

Theorem 3.4 (Noise-generalization reversal)**.** *There exists a one-dimensional loss landscape with two local minima—one with low generalization gap (good) and one with high generalization gap (bad)—such that SGD converges to the bad solution with overwhelming probability. Specifically, the ratio of escape times satisfies* $\tau_A/\tau_B \sim \exp(-19/\eta) \to 0$ *as* $\eta \to 0$.

Construction. Define a double-well loss with equal loss values and equal curvatures but different noise variances:

- Well A (good generalization): $L_A = 0.1$, $H_A = 2$, $\Sigma_A = 20$.
- Well B (bad generalization): $L_B = 0.1$, $H_B = 2$, $\Sigma_B = 1$.

The generalization gaps (Proposition 2.2) are

$$\text{Gap}_A = \frac{\Sigma_A}{H_A(n-1)} = \frac{20}{2(n-1)} = \frac{10}{n-1}, \qquad \text{Gap}_B = \frac{\Sigma_B}{H_B(n-1)} = \frac{1}{2(n-1)} = \frac{0.5}{n-1}.$$

So well A (the good solution in terms of diversity and feature usage) has a *larger* generalization gap as measured by Σ/H—this is deliberately reversed to construct the counterexample.

We pause to clarify a subtle but critical point. The LOO proxy Gap $\approx \Sigma/(H(n-1))$ from Proposition 2.2 predicts that well A has a *larger* generalization gap than well B. This is precisely the point of the counterexample: the LOO proxy is an approximation derived under regularity assumptions, and it measures sensitivity to single-sample removal, not actual population risk.

The construction posits a data-generating process where well A corresponds to a solution that leverages diverse features—different training examples activate different model components, producing diverse per-sample gradients (hence high $\Sigma_A = 20$) but low population risk because the diverse features capture the true data structure. Well B corresponds to a memorizing solution that fits a uniform low-frequency pattern, producing nearly identical per-sample gradients (hence low $\Sigma_B = 1$) but high population risk because the uniform pattern does not capture the true structure. The LOO proxy $\Sigma/(H(n-1))$ fails here because high per-sample gradient diversity does not imply instability to sample removal—it reflects feature diversity, not fragility.

This is the essence of condition N4: the noise variance Σ tracks per-sample sensitivity, not generalization *per se*. When the correlation between Σ and actual generalization gap is negative, SGD's low-Σ preference systematically selects bad solutions.

Proof. We compute the effective potential barriers and Kramers escape times for each well.

Step 1: Effective potential near each well. From (4), the effective potential near a minimum θ_k (where $L'(\theta_k) = 0$) has the local expansion

$$V_{\text{eff}}(\theta) \approx V_{\text{eff}}(\theta_k) + \frac{1}{2} V''_{\text{eff}}(\theta_k)(\theta - \theta_k)^2 + \cdots$$

The second derivative of V_{eff} at a minimum is

$$V''_{\text{eff}}(\theta_k) = \frac{H_k}{\Sigma_k} + \frac{\tau}{2} \frac{\Sigma''_k \Sigma_k - (\Sigma'_k)^2}{\Sigma_k^2}.$$

In the leading-order ($\tau \to 0$) analysis, the dominant term is H_k/Σ_k.

For wells A and B:

$$V''_{\text{eff}}(\theta_A) \approx \frac{H_A}{\Sigma_A} = \frac{2}{20} = 0.1, \qquad V''_{\text{eff}}(\theta_B) \approx \frac{H_B}{\Sigma_B} = \frac{2}{1} = 2.$$

Step 2: Effective barrier heights. Let θ_s denote the saddle point between the two wells, with $L(\theta_s) = L_A + \Delta V = L_B + \Delta V$ for some barrier height $\Delta V > 0$ in the original loss.

The effective potential barrier from well k to the saddle is, to leading order,

$$\Delta U_k = V_{\text{eff}}(\theta_s) - V_{\text{eff}}(\theta_k) \approx \frac{\Delta V}{\Sigma_k},$$

where we have used the fact that the integral $\int_{\theta_k}^{\theta_s} L'(s)/\Sigma(s)\,\mathrm{d}s$ is dominated by the behavior near well k where $\Sigma \approx \Sigma_k$.

We choose a concrete barrier height $\Delta V = 10$ in the original loss:

$$\Delta U_A = \frac{\Delta V}{\Sigma_A} = \frac{10}{20} = 0.5, \qquad \Delta U_B = \frac{\Delta V}{\Sigma_B} = \frac{10}{1} = 10.$$

The ratio of effective barriers is

$$\frac{\Delta U_A}{\Delta U_B} = \frac{\Sigma_B}{\Sigma_A} = \frac{1}{20}.$$

The good well A has a barrier 20 times *lower* in the effective potential than the bad well B.

Step 3: Kramers escape times. The Kramers escape time from well k is [Kramers, 1940]

$$\tau_k \sim \frac{1}{\eta} \exp\Big(\frac{2\,\Delta U_k}{\eta}\Big).$$

Taking the ratio with $\Delta U_A = 0.5$ and $\Delta U_B = 10$:

$$\frac{\tau_A}{\tau_B} \sim \exp\Big(\frac{2(\Delta U_A - \Delta U_B)}{\eta}\Big) = \exp\Big(\frac{2(0.5 - 10)}{\eta}\Big) = \exp\Big(-\frac{19}{\eta}\Big). \tag{6}$$

Step 4: Stationary distribution. In the two-well approximation, the stationary occupation probabilities satisfy

$$\frac{\pi(B)}{\pi(A)} = \frac{\tau_B}{\tau_A} \sim \exp\Big(\frac{19}{\eta}\Big) \to \infty \quad \text{as } \eta \to 0.$$

Therefore $\pi(B) \to 1$: SGD converges to the bad solution B with overwhelming probability.

Step 5: Verification of the reversal. Well B has the lower gradient noise variance ($\Sigma_B = 1 < 20 = \Sigma_A$), so SGD's noise-driven preference favors B. But well B has worse generalization. The stationary distribution $p(\theta_B)/p(\theta_A) = (\Sigma_A/\Sigma_B) \cdot (\cdots) = 20 \cdot (\cdots)$, where the exponential factor further amplifies the preference for B. This is the noise-generalization reversal. □

Why this happens. The reversal has a natural interpretation. The good solution A achieves low test error by leveraging diverse features in the data. Different training examples activate different features, leading to diverse per-sample gradients and hence high Σ_A. The bad solution B memorizes a uniform pattern (e.g., a low-frequency component or a label-correlated shortcut), producing nearly identical per-sample gradients and hence low Σ_B.

SGD's noise-driven selection mechanism (3) favors regions of low Σ, which is precisely the memorizing solution. This is the fundamental limit of noise-driven implicit regularization: the same mechanism that in typical settings provides beneficial regularization can, when the correlation between noise variance and generalization is reversed, systematically select bad solutions.

Which condition is violated. Condition N4 ($\mathrm{Corr}(\Sigma, \mathrm{Gap}) > 0$) is violated. In this construction, high Σ corresponds to *low* generalization gap (good) and low Σ corresponds to high generalization gap (bad), reversing the alignment that implicit regularization requires.

4 Necessary Conditions Framework

The four failure modes of Section 3 directly yield four necessary conditions for SGD's implicit regularization to succeed.

Theorem 4.1 (Four necessary conditions)**.** *For SGD to converge to a good solution (in the sense of low generalization gap) with high probability, the following conditions are all necessary:*

- **(N1)** ***Sufficient mixing:*** $\eta T \gg \tau_{\mathrm{mix}}$*, where* τ_{mix} *is the mixing time of the SGD Markov chain.*
- **(N2)** ***Discrete stability:*** $\eta < 2/\lambda_{\max}$*, where* $\lambda_{\max}$ *is the largest per-sample Hessian eigenvalue.*
- **(N3)** ***Selection power:*** *The noise-to-curvature ratios* $\Sigma(\theta^*)/H(\theta^*)$ *differ across local minima.*
- **(N4)** ***Correct alignment:*** $\mathrm{Corr}(\Sigma, \mathrm{Gap}) > 0$*, i.e., minima with higher gradient noise variance also have higher generalization gap.*

Proof. Each condition is necessary because violating it leads to a demonstrated failure:

- Violating (N1) leads to failure mode F1 (Proposition 3.1): SGD is trapped in the initialization basin and cannot reach the good solution, regardless of the stationary distribution's preference.
- Violating (N2) leads to failure mode F2 (Proposition 3.2): SGD diverges due to a positive Lyapunov exponent, preventing convergence to any solution.
- Violating (N3) leads to failure mode F3 (Proposition 3.3): the stationary distribution assigns equal weight to good and bad minima, so SGD provides no selective advantage.
- Violating (N4) leads to failure mode F4 (Theorem 3.4): the noise-driven selection systematically favors the bad solution.

In each case, the failure is demonstrated by an explicit construction, so the condition is necessary for success across all problem instances. □

Remark 4.2 (The conjunction is not proven sufficient)**.** *While conditions (N1)–(N4) are each necessary, we do not claim that their conjunction is sufficient for SGD to find good solutions. Additional conditions—related to the landscape geometry, initialization distribution, or higher-order properties of the noise structure—may also be required. We leave the question of identifying a minimal set of sufficient conditions as an important open problem.*

Remark 4.3 (Three-level classification of conditions)**.** *The conditions can be organized into a three-level hierarchy:*

1. ***Necessary:*** *The global minimum of the effective potential V_{eff} lies in the set of good solutions. If this fails, the stationary distribution concentrates on a bad solution (as $\tau \to 0$).*
2. ***Sufficient:*** *The global minimum of V_{eff} is unique in the good set, with spectral gap $\Delta > 0$, and $\tau \ll \Delta$. Then $\mathbb{P}(\text{good solution}) \geq 1 - (K-1)\exp(-2\Delta/\tau)$.*
3. ***Approximately sufficient:*** *All minima of V_{eff} with values within δ of the global minimum are approximately good (gap $\leq \varepsilon + \varepsilon'$). Then $\mathbb{P}(\text{Gap} \leq \varepsilon + \varepsilon') \geq 1 - \gamma$, where $\gamma \sim Ke^{-2\delta/\tau}$.*

Conditions (N1)–(N4) are at level 1. Closing the gap to level 2 or 3 requires quantitative control over the landscape that goes beyond our current results.

5 Reparameterization Invariance

A natural concern is whether the failure mode F4 is an artifact of a particular parameterization. We show that it is not, because the key quantity Σ/H is reparameterization-invariant.

Proposition 5.1 (Reparameterization invariance of Σ/H)**.** *Let $\theta = g(\varphi)$ be a smooth bijection (reparameterization), and let $\varphi^* = g^{-1}(\theta^*)$. Define the reparameterized Hessian $\tilde{H} = L''(g(\varphi^*)) \cdot [g'(\varphi^*)]^2$ and the reparameterized noise variance $\tilde{\Sigma} = \frac{1}{n}\sum_i [\ell_i'(g(\varphi^*))]^2 \cdot [g'(\varphi^*)]^2$. Then*

$$\frac{\tilde{\Sigma}}{\tilde{H}} = \frac{\Sigma(\theta^*)}{H(\theta^*)}.$$

Proof. Under the reparameterization $\theta = g(\varphi)$, the chain rule gives

$$\left.\frac{\partial \ell_i}{\partial \varphi}\right|_{\varphi^*} = \ell_i'(g(\varphi^*)) \cdot g'(\varphi^*) = \ell_i'(\theta^*) \cdot g'(\varphi^*).$$

Therefore the reparameterized noise variance is

$$\tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\Big[\ell_i'(\theta^*) \cdot g'(\varphi^*)\Big]^2 = \Sigma(\theta^*) \cdot \big[g'(\varphi^*)\big]^2.$$

Similarly, the reparameterized Hessian of L at φ^* (using $L'(\theta^*) = 0$, so the first-derivative term in the chain rule for second derivatives vanishes) is

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

The ratio is

$$\frac{\tilde{\Sigma}}{\tilde{H}} = \frac{\Sigma(\theta^*) \cdot [g'(\varphi^*)]^2}{H(\theta^*) \cdot [g'(\varphi^*)]^2} = \frac{\Sigma(\theta^*)}{H(\theta^*)}.$$ □

Corollary 5.2. *The generalization gap proxy $\text{Gap}(\theta^*) \approx \Sigma(\theta^*)/(H(\theta^*)(n-1))$ from Proposition 2.2 is a reparameterization-invariant quantity.*

Proof. Immediate from Proposition 5.1: $\tilde{\Sigma}/\tilde{H} = \Sigma/H$, so $\tilde{\Sigma}/(\tilde{H}(n-1)) = \Sigma/(H(n-1))$. □

Implication for F4. The noise-generalization reversal in Theorem 3.4 is not an artifact of a particular coordinate system. Since both Σ/H and the effective potential V_{eff} are built from reparameterization-invariant quantities, the conflict between SGD's noise preference and generalization persists under any smooth reparameterization. This underscores that F4 is a fundamental limitation, not a coordinate-dependent pathology.

6 Discussion

6.1 Implications for Practice

Each failure mode suggests a practical intervention:

- **F1 (insufficient mixing)**: Use learning rate warm-up or cyclical learning rate schedules to help SGD escape bad basins early in training.
- **F2 (discrete instability)**: Do not use learning rates that are too large. The stability threshold $\eta < 2/\lambda_{\max}$ provides a principled upper bound.
- **F3 (degenerate selection)**: Data augmentation and diverse training data can break symmetry in Σ/H ratios across minima, restoring SGD's selective power.
- **F4 (noise-generalization reversal)**: When the correlation between gradient noise and generalization is negative—as may occur with small datasets, shortcut features, or distribution shift—explicit regularization (weight decay, dropout, early stopping) is essential. SGD's implicit regularization alone is insufficient.

6.2 Connection to the Interpolation Regime

In the modern over-parameterized interpolation regime, all training losses reach zero: $\ell_i(\theta^*) = 0$ for all i. This implies $\nabla \ell_i(\theta^*) = 0$ and hence $\Sigma(\theta^*) = 0$ at every interpolating solution. The continuous-time SDE framework becomes degenerate, and the noise-driven selection mechanism analyzed here does not directly apply.

Instead, implicit regularization in the interpolation regime operates through a different mechanism: discrete stability. Solutions with $\eta\, h_{\max} > 2$ are destabilized, while those with $\eta\, h_{\max} < 2$ remain stable [Wu and Su, 2023]. This selects for solutions with low Jacobian complexity $\mathcal{C}(\theta^*) = \|J\|_F/\sqrt{n}$.

Our failure mode F2 is relevant in this regime: choosing η too large may destabilize all solutions. Failure modes F3 and F4, as stated, pertain primarily to the non-interpolation regime where $\Sigma > 0$. However, the conceptual lesson of F4—that the metric optimized by SGD's dynamics may not align with generalization—extends to the interpolation regime as well: low Jacobian complexity is not a universal proxy for generalization.

6.3 Relation to Prior Work

Shape matters [Li et al., 2021]. Li et al. demonstrate that SGD's noise structure (not just its magnitude) influences which minima are selected. Our work is complementary: we identify four settings where even the correct noise structure fails to produce good solutions.

Sharp minima can generalize [Dinh et al., 2017]. Dinh et al. show that reparameterization can make any minimum arbitrarily sharp without affecting generalization, challenging flatness-based explanations. Our Proposition 5.1 resonates with this observation: H alone is not invariant, but Σ/H is. However, F4 shows that even the reparameterization-invariant quantity Σ/H can fail as a generalization proxy.

Implicit regularization in linear models [Arora et al., 2019]. Arora et al. characterize implicit regularization for deep linear networks. Their results show that architecture (depth, width) shapes the implicit bias. Our F4 highlights that even with fixed architecture, the data distribution can induce a noise-generalization reversal.

Entropy-SGD [Chaudhari and Soatto, 2018]. Chaudhari and Soatto connect SGD to an entropy-regularized objective, establishing a link between noise and flatness-seeking behavior. Our Proposition 2.1 uses the same FPE framework; F4 demonstrates a failure of the resulting selection when the noise-generalization correlation is negative.

6.4 Open Questions

1. **Sufficiency.** Are conditions (N1)–(N4) jointly sufficient? We conjecture they are not, and that additional landscape conditions are needed. Identifying a minimal sufficient set is a key open problem.
2. **Multi-dimensional effective potential.** In dimension $d > 1$, the detailed balance condition may fail, and the steady state involves a Freidlin–Wentzell quasipotential rather than an explicit formula. Extending our results to this setting is technically demanding.
3. **Prevalence of F4.** How often does the noise-generalization reversal occur in practical deep learning? Characterizing the data-architecture combinations that lead to $\mathrm{Corr}(\Sigma, \mathrm{Gap}) < 0$ is an empirical and theoretical challenge.
4. **Global trajectory analysis.** Our results are local (near minima). Understanding SGD's global trajectory from random initialization to a basin of attraction requires different tools.
5. **Discrete-continuous unification.** The non-interpolation and interpolation regimes use fundamentally different mathematical frameworks. A unified theory handling the $\Sigma \to 0$ limit remains open.

7 Conclusion

We have identified four failure modes of SGD's implicit regularization, each demonstrated by an explicit construction with quantitative analysis. Failure modes F1–F3 concern the dynamics (mixing time, stability, and selection power), while F4—our main result—reveals a fundamental limitation: SGD's noise-driven preference for low gradient variance can systematically select solutions that generalize poorly when good solutions use diverse features (high Σ) and bad solutions memorize (low Σ).

From these failure modes we derived four necessary conditions (N1–N4) that must hold simultaneously for implicit regularization to succeed. The conditions span three distinct aspects: the optimization dynamics (N1, N2), the noise structure (N3), and the data-dependent alignment between noise and generalization (N4). Their conjunction is not shown to be sufficient, and closing this gap remains an important open problem.

Our results carry a practical message: in settings where condition N4 may be violated—small datasets, distribution shift, shortcut features—relying solely on SGD's implicit regularization is insufficient, and explicit regularization is needed.

References

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges

to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, 2017.

W. Horsthemke and R. Lefever. *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry, and Biology*. Springer-Verlag, 1984.

H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

Z. Li, T. Wang, and S. Arora. What happens after SGD reaches zero loss? A mathematical framework. In *International Conference on Learning Representations*, 2021.

S. L. Smith and Q. V. Le. A Bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

L. Wu and W. J. Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 37656–37684, 2023.