A Closed-Loop Theorem for SGD's Noise-Driven Selection in the Non-Interpolation Regime

Lightman Chang
Independent Researcher
lightman.chang@gmail.com

Abstract

We study the implicit regularization of stochastic gradient descent (SGD) in the *non-interpolation* regime, where the gradient noise covariance $\Sigma(\theta^*)$ at a local minimum is strictly positive and the continuous-time Fokker–Planck approximation is non-degenerate. Our main result (Theorem 3.1) is a closed-loop theorem with three quantitatively linked components: (i) the SGD stationary measure on a multi-well potential admits an explicit Gibbs-type formula whose relative weights at two minima are governed by the ratios Σ_A/Σ_B and H_A/H_B; (ii) the leave-one-out generalization gap at a quadratic minimum equals $\Sigma(\theta^*)/(H(\theta^*)(n-1))$ up to $O(n^{-2})$; (iii) consequently $\Sigma_A/H_A < \Sigma_B/H_B$ implies that SGD concentrates on the basin of θ_A^* *and* that θ_A^* has a strictly smaller test-loss bound than θ_B^*, with explicit margin. We extend the dynamic part to multivariate parameter spaces via the Freidlin–Wentzell quasipotential, working out a two-dimensional example with state-dependent diffusion in which detailed balance fails but the quasipotential admits a closed-form Hamilton–Jacobi expansion (Theorem 7.4). We prove that the ratio Σ/H, and its multivariate generalization $\mathrm{tr}(H^{-1}\Sigma)$, are invariant under smooth diffeomorphisms of the parameter space (Proposition 8.1), which makes the selection criterion intrinsic. The result is the non-interpolation counterpart of the discrete Lyapunov closed loop established in [1] for the interpolation regime, and its sharpness is exactly delimited by the four failure modes catalogued in [2]; in particular, the implication (iii) is reversed by failure mode F4, and the interplay is made precise.

MSC 2020: 60H10 (primary); 62L20, 60F10, 68T07, 49L25, 60J60, 65C05.

1 Introduction

The problem. Stochastic gradient descent (SGD) is the workhorse optimizer of modern machine learning, and a long line of work has tried to explain why it selects, among the many local minima of a non-convex training loss, those that generalize well to unseen data. In the *non-interpolation regime*—where the training loss does not vanish, so the per-sample gradients $\nabla\ell_i(\theta^*)$ at a local minimum θ^* are not all zero—a popular explanation models SGD by the Itô stochastic differential equation (SDE)

$$d\theta_t \;=\; -\nabla L(\theta_t)\,\mathrm{d}t \;+\; \sqrt{\tau\,\Sigma(\theta_t)}\,dW_t, \qquad \tau \;=\; \eta/B,$$

where η is the learning rate, B is the batch size and $\Sigma(\theta)$ is the per-sample gradient covariance [3, 5, 4, 6]. The multiplicative structure of Σ then biases the Fokker–Planck stationary measure toward minima of low noise covariance. This is appealing as a heuristic, but two questions have remained open at a fully rigorous level:

(Q1) Can the dynamic preference of SGD for a particular local minimum be quantified by an explicit, parameter-free formula in terms of the curvature $H(\theta^*)$ and the noise covariance $\Sigma(\theta^*)$?

(Q2) Does this dynamic preference align with a sample-based notion of generalization—e.g. leave-one-out (LOO) stability—and if so, through which intrinsic geometric quantity?

What we do. We answer (Q1) and (Q2) in the affirmative under explicit assumptions that we state quantitatively. Specifically, we prove a single closed-loop theorem that links three components: the Fokker–Planck stationary distribution (selection), the leave-one-out gap (generalization), and the implication that connects them. All three reduce, in the symmetric two-well one-dimensional case, to the single intrinsic quantity $\Sigma(\theta^*)/H(\theta^*)$. The qualitative picture has been suggested in earlier work [6, 5, 7]; the contribution of the present paper is to make every step quantitative and self-contained, including a multivariate extension via the Freidlin–Wentzell quasipotential and a complete reparameterization invariance proof.

Delta from prior work. The components of Theorem 3.1 are not all new in isolation. *Part I* (the one-dimensional Gibbs-type stationary density of the Fokker–Planck equation with multiplicative noise) is textbook material in stochastic processes [10] and was applied to SGD explicitly by Mandt–Hoffman–Blei [4, §3] and others; the closed-form $p_\tau \propto \Sigma^{-1} \exp(-(2/\tau)U)$ is folklore in this literature. *Part II* (the leave-one-out generalization gap $\Sigma/(H(n-1))$ at a quadratic minimum) is a classical second-order expansion of stability [12] and is implicit in optimization-theoretic surveys [15]. *The contribution of the present paper* is to bind these three ingredients together into a single closed loop: stationary mass, LOO gap, and reparameterization invariance through the common intrinsic quantity Σ/H (resp. $\mathrm{tr}(H^{-1}\Sigma)$ in higher dimensions). In particular, we (i) state the loop quantitatively under explicit assumptions, (ii) make the boundary with each of the four failure modes F1–F4 precise (Section 9), and (iii) extend the dynamic part to the multivariate setting where detailed balance fails, via a worked Freidlin–Wentzell quasipotential (Theorem 7.4). The *closed-loop bind* (Parts I+II+III+ reparameterization), not its individual components, is the present paper's contribution.

Position relative to companion papers. This paper is the third in a series. Paper A [1] treats the *interpolation regime*, where every interpolating solution has $\Sigma(\theta^*) = 0$, the SDE approximation degenerates, and a discrete Lyapunov-exponent analysis must replace the Fokker–Planck argument. Paper B [2] catalogues four failure modes (F1: insufficient mixing; F2: discrete instability; F3: degenerate selection; F4: noise–generalization reversal) under which the closed loop breaks. The present paper-C is the positive counterpart in the non-interpolation regime: it states and proves the closed loop precisely under the negation of those four failure modes, and isolates the hypothesis on which the implication between selection and generalization rests, namely the alignment of low Σ/H with low test loss (the negation of F4).

Main results, informally. Let θ_A^* and θ_B^* be two locally quadratic minima of L with curvatures $H_A, H_B > 0$ and noise covariances $\Sigma_A, \Sigma_B > 0$. Under explicit ergodicity assumptions:

- (Theorem 3.1 Part I) The SGD stationary density is $p(\theta) \propto \Sigma(\theta)^{-1} \exp(-(2/\tau)\, U(\theta))$ for an explicit drift potential U, and in the symmetric one-dimensional case $p(\theta_A^*)/p(\theta_B^*) = \Sigma_B/\Sigma_A$.
- (Theorem 3.1 Part II) The LOO generalization gap at a quadratic minimum θ^* satisfies $\mathrm{Gap}_{\mathrm{LOO}}(\theta^*) = \Sigma(\theta^*)/(H(\theta^*)(n-1)) + O(n^{-2})$.
- (Theorem 3.1 Part III) If $\Sigma_A/H_A < \Sigma_B/H_B$, then the SGD-stationary measure concentrates on the basin of θ_A^* in the sense of Theorem 3.1, and θ_A^*'s test-loss bound is strictly smaller than that of θ_B^* with explicit margin $\big(\Sigma_B/H_B - \Sigma_A/H_A\big)/(n-1)$.
- (Theorem 7.4) For a state-dependent multivariate diffusion in which detailed balance fails, the Freidlin–Wentzell quasipotential exists and is the unique viscosity solution of an explicit Hamilton–Jacobi equation. We give a worked 2D example for which the leading correction to the naive potential is computed in closed form.

- (Proposition 8.1) The ratio Σ/H, and its multivariate generalization $\mathrm{tr}(H^{-1}\Sigma)$, are invariant under smooth diffeomorphisms of the parameter space.

Organization. Section 2 fixes the notation and lists the standing assumptions. Section 3 states the main theorem. Sections 4 to 6 contain the three parts of the proof. Section 7 contains the multivariate extension. Section 8 contains the reparameterization invariance proof. Section 9 discusses the boundary with paper B, the relation to paper A, and the open problems.

2 Preliminaries

2.1 The empirical risk and SGD

Let $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$ be a training set drawn i.i.d. from a distribution $\mathcal{D}$. Fix a parameter space $\Theta \subseteq \mathbb{R}^d$ and a per-sample loss $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ that is C^2 in θ. Write

$$L(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell_i(\theta), \qquad \ell_i(\theta) \;:=\; \ell(\theta, z_i).$$

SGD is the recursion

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

We work with constant learning rate $\eta > 0$. The mini-batch case $B \geq 1$ is identical with $\tau := \eta/B$ in place of η throughout.

2.2 Curvature and gradient noise

Definition 2.1 (Curvature and noise covariance)**.** At any $\theta \in \Theta$,

$$H(\theta) \;:=\; \nabla^2 L(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ell_i(\theta),$$

$$\Sigma(\theta) \;:=\; \frac{1}{n}\sum_{i=1}^{n} \big(\nabla \ell_i(\theta) - \nabla L(\theta)\big)\big(\nabla \ell_i(\theta) - \nabla L(\theta)\big)^{\top}.$$

When $d = 1$ both reduce to scalars. At a critical point θ^* ($\nabla L(\theta^*) = 0$), the noise covariance simplifies to $\Sigma(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell_i(\theta^*) \nabla \ell_i(\theta^*)^{\top}$.

Definition 2.2 (Three-level convergence)**.** We distinguish:

(C1) *Loss convergence*: $L(\theta_t) \to L^*$ almost surely, for some L^*.

(C2) *Distributional convergence*: there is a probability measure μ_η on Θ such that the law of θ_t converges weakly to μ_η as $t \to \infty$.

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

In the non-interpolation regime, (C3) does not hold for SGD with constant $\eta > 0$: the iterates fluctuate in a basin rather than converge to a single point. The natural mode of convergence is (C2), and we shall characterize the limit μ_η.

2.3 The Fokker–Planck approximation

For small η, the recursion (1) is approximated [3, 4, 5] by the Itô SDE

$$d\theta_t \;=\; -\nabla L(\theta_t)\,\mathrm{d}t \;+\; \sqrt{\tau\,\Sigma(\theta_t)}\,dW_t, \qquad \tau = \eta/B. \tag{2}$$

The corresponding Fokker–Planck equation (FPE) for the density $p(t,\theta)$ [11, 10] is

$$\partial_t p \;=\; \nabla\cdot\big(\nabla L\, p\big) \;+\; \frac{\tau}{2}\,\nabla\cdot\nabla\cdot\big(\Sigma\, p\big). \tag{3}$$

We take (3) as the working model for the density μ_η and state our assumptions accordingly. The standard estimate that justifies (2)–(3) is that any time-marginal $\mathbb{E}[\phi(\theta_T)]$ for $\phi \in C_b^4$ agrees up to $O(\eta^2)$ with the SDE prediction at horizon $T = N\eta$ uniformly for bounded N [3]. Throughout this paper we therefore identify SGD's long-time behaviour with the stationary measure of (3).

2.4 Standing assumptions

Assumption 1 (Smoothness)**.** $L \in C^3(\Theta)$ and $\ell_i \in C^3(\Theta)$ for $i = 1, \ldots, n$. The C^3 regularity of each ℓ_i is required by the Taylor remainder $O(\left\|\theta^*_{-j} - \theta^*\right\|^3)$ used in the proof of Part II (Lemma 5.2).

Assumption 2 (Coercivity)**.** $L(\theta) \to +\infty$ as $\|\theta\| \to \infty$, with growth at most polynomial: there is $r > 0$ such that $\|\nabla L(\theta)\| \le C(1 + \|\theta\|^r)$.

Assumption 3 (Uniform ellipticity of noise)**.** There are constants $0 < \sigma^2_{\min} \le \sigma^2_{\max} < \infty$ such that for every θ and every unit vector $v \in \mathbb{R}^d$,

$$\sigma^2_{\min} \;\le\; v^\top \Sigma(\theta)\, v \;\le\; \sigma^2_{\max}.$$

Assumption 4 (Two locally quadratic minima)**.** L has exactly two local minima θ^*_A, θ^*_B with strict positive-definite Hessians $H_A := H(\theta^*_A) \succ 0$ and $H_B := H(\theta^*_B) \succ 0$, and exactly one saddle θ_S on the minimum-energy path between them. We write $\Sigma_A := \Sigma(\theta^*_A)$, $\Sigma_B := \Sigma(\theta^*_B)$.

Assumption 5 (Selection hypothesis)**.** $\Sigma_A/H_A < \Sigma_B/H_B$ (in $d = 1$ as scalars; in higher dimensions with the trace generalization of Proposition 8.1).

Assumption 6 (F4-negation: LOO aligns with population)**.** The LOO gap is an unbiased proxy for the population–training gap with no model-misspecification reversal: at both θ^*_A and θ^*_B, $\mathbb{E}[L_{\mathrm{pop}}(\theta^*) - L(\theta^*)] = \mathbb{E}[\mathrm{Gap}_{\mathrm{LOO}}(\theta^*)] + o(1/n)$, and the sign of the difference of population gaps between the two minima agrees with the sign of the difference of LOO gaps. Equivalently, the failure mode F4 of [2] is inactive: low Σ/H corresponds to low population loss rather than memorization under data–architecture misspecification.

Assumption 1–Assumption 3 are mild regularity conditions standard in diffusion theory [10]. Coercivity (Assumption 2) and uniform ellipticity (Assumption 3) together imply that the FPE (3) admits a unique invariant probability density. Assumption 4 restricts attention to the simplest non-trivial landscape. Assumption 5 is the hypothesis on which the third part of the main theorem rests; it is the negation of failure mode F3 of [2] (when the inequality is strict and the alignment with generalization is preserved, F4 is also negated).

2.5 Leave-one-out perturbation

Definition 2.3 (LOO empirical risk and minimizer)**.** For $j \in \{1, \ldots, n\}$ define the leave-one-out empirical risk

$$L_{-j}(\theta) \;:=\; \frac{1}{n-1}\sum_{i \ne j} \ell_i(\theta),$$

and let θ^*_{-j} be its local minimizer in the same basin as θ^*. The LOO generalization gap at θ^* is

$$\mathrm{Gap}_{\mathrm{LOO}}(\theta^*) \;:=\; \frac{1}{n}\sum_{j=1}^{n}\big(\ell_j(\theta^*_{-j}) - \ell_j(\theta^*)\big).$$

A standard consequence of $\mathbb{E}[\mathrm{Gap}_{\mathrm{LOO}}] \approx \mathbb{E}[L_{\mathrm{pop}}(\theta^*) - L(\theta^*)]$ [12, 13] makes $\mathrm{Gap}_{\mathrm{LOO}}$ a sample-based proxy for the population gap.

3 The closed-loop theorem

Theorem 3.1 (Non-interpolation closed loop)**.** *Adopt Assumption 1–Assumption 4, set $\tau = \eta/B$, and consider the Itô SDE* (2)*. Then:*

(I) (Stationary measure.) *The FPE* (3) *admits a unique probability invariant density $p_\tau(\theta)$. In dimension $d = 1$,*

$$p_\tau(\theta) \;=\; \frac{1}{Z_\tau\,\Sigma(\theta)}\,\exp\Big(-\frac{2}{\tau}\,U(\theta)\Big), \quad U(\theta) \;:=\; \int_{\theta_0}^{\theta}\frac{L'(s)}{\Sigma(s)}\,ds, \tag{4}$$

where θ_0 is any reference point and Z_τ normalizes p_τ. Let $B_A, B_B \subset \mathbb{R}$ denote the open basins of attraction of θ^*_A, θ^*_B for $-L'$ (separated by the saddle θ_S), and write the basin masses as $M_\tau(A) := \int_{B_A} p_\tau(\theta)\,d\theta$, $M_\tau(B) := \int_{B_B} p_\tau(\theta)\,d\theta$. Applying Laplace's method to (4) *as $\tau \to 0$ gives*

$$\frac{M_\tau(A)}{M_\tau(B)} \;=\; \sqrt{\frac{\Sigma_B\,H_B}{\Sigma_A\,H_A}}\,\exp\Big(-\tfrac{2}{\tau}\big(U(\theta^*_A) - U(\theta^*_B)\big)\Big)\,(1 + O(\sqrt{\tau})). \tag{5}$$

(The $\sqrt{\Sigma_B/\Sigma_A}$ rather than Σ_B/Σ_A comes from the Gaussian-width factor $\sqrt{\pi\tau\,\Sigma_k/H_k}$ partially cancelling the $1/\Sigma_k$ prefactor at each minimum; see Section 6 and eq. (16)*.) In the symmetric two-well case (L and Σ symmetric about $(\theta^*_A + \theta^*_B)/2$, so that $H_A = H_B$ and $U(\theta^*_A) = U(\theta^*_B)$),*

$$\frac{M_\tau(A)}{M_\tau(B)} \;=\; \sqrt{\frac{\Sigma_B}{\Sigma_A}}\,(1 + O(\sqrt{\tau}))\;\; \big(> 1 \text{ iff } \Sigma_A < \Sigma_B\big). \tag{6}$$

In the general (asymmetric) one-dimensional case the basin-mass ratio is given by (5)*; the corresponding* pointwise *density ratio is*

$$\frac{p_\tau(\theta^*_A)}{p_\tau(\theta^*_B)} \;=\; \frac{\Sigma_B}{\Sigma_A}\,\exp\Big(-\tfrac{2}{\tau}\big(U(\theta^*_A) - U(\theta^*_B)\big)\Big). \tag{7}$$

(II) (LOO gap.) *At a locally quadratic minimum θ^* with $H(\theta^*) > 0$, under Assumption 1,*

$$\mathrm{Gap}_{\mathrm{LOO}}(\theta^*) \;=\; \frac{\Sigma(\theta^*)}{(n-1)\,H(\theta^*)} \;+\; O\Big(\frac{1}{n^2}\Big), \tag{8}$$

where the constant in $O(n^{-2})$ depends only on $\sup_j\big(\|\nabla\ell_j(\theta^*)\|^4 + \|\nabla^3\ell_j(\theta^*)\|\big)$ and $H(\theta^*)^{-1}$.

(III) (Implication.) *Assume in addition Assumption 5, i.e. $\Sigma_A/H_A < \Sigma_B/H_B$, and the equal-loss two-well setting $L(\theta^*_A) = L(\theta^*_B)$ so that the basin-mass formula* (5) *applies. Then*

(i) *if additionally $\Sigma_A H_A < \Sigma_B H_B$ (which holds automatically in the symmetric case $H_A = H_B$, since then Assumption 5 reduces to $\Sigma_A < \Sigma_B$), the stationary mass concentrates on the basin of θ^*_A: for τ sufficiently small, $M_\tau(A)/M_\tau(B) > 1$;*

(ii) *the LOO gap at well A is strictly smaller than at well B:*

$$\mathrm{Gap}_{\mathrm{LOO}}(\theta_B^*) - \mathrm{Gap}_{\mathrm{LOO}}(\theta_A^*) \;=\; \frac{1}{n-1}\Big(\frac{\Sigma_B}{H_B} - \frac{\Sigma_A}{H_A}\Big) \;+\; O(n^{-2}) \;>\; 0; \tag{9}$$

(iii) *under the additional assumption Assumption 6 (which posits, as substantive content, that LOO is an unbiased proxy for the population gap with no F4 reversal in this data–architecture pair), the LOO-based selection coincides with the test-loss-minimizing selection: the test-loss bound at A is below that at B by the same margin (up to $O(n^{-2})$). We emphasize that Assumption 6 is verifiable in principle (e.g. by sample-splitting or held-out evaluation) and is not implied by Parts I or II.*

The remainder of Sections 4 to 6 contains the proof of Theorem 3.1 in three parts.

4 Proof of Theorem 3.1: Part I (stationary measure)

We work in dimension $d = 1$ throughout this section; the multivariate extension is treated separately in Section 7 via the Freidlin–Wentzell quasipotential. Write $L, \Sigma : \mathbb{R} \to \mathbb{R}$ and $L'(\theta) = dL/d\theta$.

4.1 Existence and uniqueness of the stationary density

Lemma 4.1 (Existence and uniqueness)**.** *Under Assumption 1–Assumption 3, the FPE* (3) *has a unique invariant probability density $p_\tau \in C^2(\mathbb{R})$.*

Proof. The drift $-L'$ is continuous; by Assumption 2 there are constants $a, b > 0$ such that $-\theta\, L'(\theta) \le -a\theta^2 + b$ outside a compact set, which is a Lyapunov condition for the SDE (2). Combined with the uniform ellipticity Assumption 3, classical results ([10, Theorem 6.16]; see also [11, Chap. 5]) give the existence of a unique invariant probability measure with a C^2 density p_τ, which solves the stationary FPE $\partial_t p_\tau = 0$ in (3). □

4.2 Closed form of the one-dimensional stationary density

Lemma 4.2 (Stationary density in $d = 1$)**.** *In dimension $d = 1$, the stationary density satisfies*

$$p_\tau(\theta) \;=\; \frac{1}{Z_\tau\, \Sigma(\theta)}\, \exp\Big(-\frac{2}{\tau}\, U(\theta)\Big), \qquad U(\theta) \;=\; \int_{\theta_0}^{\theta} \frac{L'(s)}{\Sigma(s)}\, ds. \tag{10}$$

Equivalently, $p_\tau(\theta) \propto \exp(-(2/\tau) V_{\mathrm{eff}}(\theta))$ with effective potential

$$V_{\mathrm{eff}}(\theta) \;:=\; U(\theta) \;+\; \frac{\tau}{2}\, \ln \Sigma(\theta). \tag{11}$$

Proof. The stationary FPE in $d = 1$ reads

$$0 \;=\; \frac{d}{d\theta}\Big[L'(\theta)\, p(\theta)\Big] + \frac{\tau}{2}\, \frac{d^2}{d\theta^2}\Big[\Sigma(\theta)\, p(\theta)\Big].$$

Define the probability current $J(\theta) := -L'(\theta)\, p(\theta) - \frac{\tau}{2}\, \frac{d}{d\theta}\big[\Sigma(\theta)\, p(\theta)\big]$. Then the FPE is $dJ/d\theta = 0$, so J is a constant. By Assumption 2 and Assumption 3, $p_\tau \in L^1(\mathbb{R})$ with exponential tails, hence $p_\tau(\theta) \to 0$ as $\|\theta\| \to \infty$ and the same holds for $\Sigma\, p_\tau$ and its derivatives. Sending $\theta \to \infty$ along a sequence on which the limit exists forces $J \equiv 0$.

Setting $J \equiv 0$:

$$L'(\theta)\, p(\theta) \;+\; \frac{\tau}{2}\, \Sigma'(\theta)\, p(\theta) \;+\; \frac{\tau}{2}\, \Sigma(\theta)\, p'(\theta) \;=\; 0,$$

so that

$$\frac{p'(\theta)}{p(\theta)} \;=\; -\frac{2L'(\theta)}{\tau\,\Sigma(\theta)} \;-\; \frac{\Sigma'(\theta)}{\Sigma(\theta)} \;=\; \frac{d}{d\theta}\Big[-\frac{2}{\tau}\,U(\theta) - \ln\Sigma(\theta)\Big].$$

Integrating from a reference point θ_0 and exponentiating gives (10); the rewriting $p_\tau \propto \exp(-(2/\tau)V_{\text{eff}})$ follows by absorbing the $\ln\Sigma$ term into the exponent. Normalizability of Z_τ is immediate from the exponential decay of $\exp(-(2/\tau)U)$ implied by Assumption 2 and Assumption 3. □

4.3 Symmetric two-well ratio formula

Lemma 4.3 (Symmetric two-well case)**.** *Suppose, in addition to Assumption 4, that L and Σ are symmetric about $\theta_m := (\theta_A^* + \theta_B^*)/2$, in the sense that $L(\theta_m + u) = L(\theta_m - u)$ and $\Sigma(\theta_m + u) = \Sigma(\theta_m - u)$. Then $U(\theta_A^*) = U(\theta_B^*)$, and*

$$\frac{p_\tau(\theta_A^*)}{p_\tau(\theta_B^*)} \;=\; \frac{\Sigma_B}{\Sigma_A}.$$

Proof. The integrand L'/Σ in U is the ratio of an odd function (L') and an even function (Σ) about θ_m, hence is odd about θ_m. Choose the reference point $\theta_0 = \theta_m$, so that $U(\theta_m) = 0$. Then $U(\theta_m + u) = \int_0^u (L'/\Sigma)(\theta_m + s)\,ds$, which by the change of variables $s \mapsto -s$ and the oddness of the integrand equals $-\int_0^u (L'/\Sigma)(\theta_m - s)\,ds = -U(\theta_m - u)$. Hence U is odd about θ_m:

$$U(\theta_m + u) \;=\; -U(\theta_m - u). \tag{12}$$

Setting $\theta_A^* = \theta_m - r$ and $\theta_B^* = \theta_m + r$ (with $r > 0$), (12) gives

$$U(\theta_A^*) \;=\; -U(\theta_B^*). \tag{13}$$

Now we show in addition that both values equal zero. Symmetry of L gives $L(\theta_A^*) = L(\theta_B^*)$, and since $L'(\theta_A^*) = L'(\theta_B^*) = 0$, the integral $U(\theta_B^*) - U(\theta_A^*) = \int_{\theta_A^*}^{\theta_B^*} L'(s)/\Sigma(s)\,ds$ can be split at the midpoint θ_m:

$$\int_{\theta_A^*}^{\theta_m} \frac{L'(s)}{\Sigma(s)}\,ds \;+\; \int_{\theta_m}^{\theta_B^*} \frac{L'(s)}{\Sigma(s)}\,ds \;=\; -U(\theta_A^*) + U(\theta_B^*) \;=\; 2\,U(\theta_B^*),$$

using $U(\theta_m) = 0$ and (13). But by oddness of L'/Σ about θ_m, the two integrals on the left cancel, so $2U(\theta_B^*) = 0$, hence $U(\theta_A^*) = U(\theta_B^*) = 0$.

Substituting into (10):

$$\frac{p_\tau(\theta_A^*)}{p_\tau(\theta_B^*)} \;=\; \frac{\Sigma_B}{\Sigma_A}\,\exp\Big(-\tfrac{2}{\tau}\big(U(\theta_A^*) - U(\theta_B^*)\big)\Big) \;=\; \frac{\Sigma_B}{\Sigma_A}.$$

□

Lemmas 4.1 to 4.3 together prove Theorem 3.1 Part I. In the asymmetric case the formula (7) is exactly (10) evaluated at the two wells; we record it as a corollary.

Corollary 4.4 (General one-dimensional ratio)**.** *In the general $d = 1$ setting under Assumption 1–Assumption 4,*

$$\frac{p_\tau(\theta_A^*)}{p_\tau(\theta_B^*)} \;=\; \frac{\Sigma_B}{\Sigma_A}\,\exp\Big(-\tfrac{2}{\tau}\big(U(\theta_A^*) - U(\theta_B^*)\big)\Big).$$

5 Proof of Theorem 3.1: Part II (LOO gap)

We now prove the leave-one-out generalization formula (8). We work at a locally quadratic minimum θ^* with $H = H(\theta^*) > 0$ and $\Sigma = \Sigma(\theta^*)$. Throughout this section all $O(\cdot)$ constants depend only on the quantities listed in the statement of Part II.

5.1 Perturbation of the minimizer

Lemma 5.1 (Leave-one-out shift). *For each $j \in \{1, \dots, n\}$ there is a unique local minimizer θ^*_{-j} of L_{-j} in a neighbourhood of θ^*, and*

$$\theta^*_{-j} - \theta^* \;=\; \frac{H_{-j}^{-1}\,\nabla \ell_j(\theta^*)}{n-1} \;+\; O\Big(\frac{\|\nabla \ell_j(\theta^*)\|^2}{n^2}\Big), \qquad H_{-j} \;:=\; \nabla^2 L_{-j}(\theta^*), \tag{14}$$

where in $d = 1$, $H_{-j}^{-1}\nabla \ell_j$ reduces to $\nabla \ell_j / H_{-j}$. Moreover $H_{-j} = H + O(1/n)$.

Proof. Compute

$$\nabla L_{-j}(\theta^*) \;=\; \frac{1}{n-1}\sum_{i \neq j} \nabla \ell_i(\theta^*) \;=\; \frac{n}{n-1}\nabla L(\theta^*) \;-\; \frac{1}{n-1}\nabla \ell_j(\theta^*) \;=\; -\,\frac{\nabla \ell_j(\theta^*)}{n-1},$$

using $\nabla L(\theta^*) = 0$. Similarly,

$$\nabla^2 L_{-j}(\theta^*) \;=\; \frac{n}{n-1}\,H \;-\; \frac{\nabla^2 \ell_j(\theta^*)}{n-1} \;=\; H \;+\; \frac{1}{n-1}\big(H - \nabla^2 \ell_j(\theta^*)\big) \;=\; H + O(1/n).$$

In a neighbourhood of θ^* where $\nabla^2 L_{-j}(\theta) \succ 0$ (which holds for all sufficiently large n by continuity), the implicit function theorem applied to ∇L_{-j} produces a unique solution θ^*_{-j} to $\nabla L_{-j}(\theta^*_{-j}) = 0$ with

$$\theta^*_{-j} - \theta^* \;=\; -\,\big(\nabla^2 L_{-j}(\theta^*)\big)^{-1}\,\nabla L_{-j}(\theta^*) \;+\; O\big(\|\nabla L_{-j}(\theta^*)\|^2\big).$$

Substituting the formulas above gives (14). □

5.2 The pointwise generalization gap

Lemma 5.2 (Pointwise gap). *With the notation of Lemma 5.1,*

$$\ell_j(\theta^*_{-j}) - \ell_j(\theta^*) \;=\; \frac{\|\nabla \ell_j(\theta^*)\|^2}{(n-1)\,H_{-j}} \;+\; O\Big(\frac{\|\nabla \ell_j(\theta^*)\|^3}{n^2}\Big). \tag{15}$$

Proof. Taylor-expand ℓ_j around θ^* to second order:

$$\ell_j(\theta^*_{-j}) - \ell_j(\theta^*) \;=\; \nabla \ell_j(\theta^*)^\top(\theta^*_{-j} - \theta^*) \;+\; \tfrac{1}{2}(\theta^*_{-j} - \theta^*)^\top \nabla^2 \ell_j(\theta^*)(\theta^*_{-j} - \theta^*) \;+\; O(\|\theta^*_{-j} - \theta^*\|^3).$$

Plugging in (14) and using $\|\theta^*_{-j} - \theta^*\| = O(\|H_{-j}^{-1}\|\;\|\nabla \ell_j(\theta^*)\|/n)$ (constants depending on H_{-j}^{-1}, hence on H^{-1}):

$$\begin{aligned}
\nabla \ell_j(\theta^*)^\top(\theta^*_{-j} - \theta^*) \;&=\; \frac{\|\nabla \ell_j(\theta^*)\|^2}{(n-1)\,H_{-j}} \;+\; O\Big(\frac{\|\nabla \ell_j(\theta^*)\|^3}{n^2}\Big),\\
\tfrac{1}{2}(\theta^*_{-j} - \theta^*)^\top \nabla^2 \ell_j(\theta^*)(\theta^*_{-j} - \theta^*) \;&=\; O\Big(\frac{\|\nabla \ell_j(\theta^*)\|^2}{n^2}\Big).
\end{aligned}$$

Combining these and absorbing all lower-order remainders gives (15). □

5.3 Averaging and the noise covariance

Proof of Theorem 3.1 Part II. Average (15) over j:

$$\begin{aligned}
\mathrm{Gap}_{\mathrm{LOO}}(\theta^*) \;&=\; \frac{1}{n}\sum_{j=1}^{n} \frac{\|\nabla \ell_j(\theta^*)\|^2}{(n-1)\,H_{-j}} \;+\; O\Big(\frac{1}{n^2}\Big)\\
&=\; \frac{1}{(n-1)\,H} \;\cdot\; \frac{1}{n}\sum_{j=1}^{n} \|\nabla \ell_j(\theta^*)\|^2 \;+\; O\Big(\frac{1}{n^2}\Big),
\end{aligned}$$

where in the second line we replaced each H_{-j} by H at the cost of an $O(1/n)\cdot\mathbb{E}_j[\|\nabla\ell_j\|^2]/(n-1)$ correction, which is $O(n^{-2})$ uniformly under the standing boundedness assumption $\mathbb{E}_j[\|\nabla\ell_j(\theta^*)\|^2] = O(1)$ (this holds at any minimizer with $\Sigma(\theta^*)$ bounded above by Assumption 3). By Definition 2.1 and $\nabla L(\theta^*) = 0$,

$$\frac{1}{n}\sum_{j=1}^{n}\|\nabla\ell_j(\theta^*)\|^2 \;=\; \operatorname{tr}\Big(\frac{1}{n}\sum_{j=1}^{n}\nabla\ell_j(\theta^*)\nabla\ell_j(\theta^*)^\top\Big) \;=\; \operatorname{tr}\Sigma(\theta^*).$$

In dimension $d = 1$ this is just $\Sigma(\theta^*)$, giving (8). In general dimension the identity becomes $\mathrm{Gap}_{\mathrm{LOO}}(\theta^*) = \operatorname{tr}(H^{-1}\Sigma)/(n-1) + O(n^{-2})$, where the trace formulation uses $\|\nabla\ell_j\|^2_{H^{-1}} = \nabla\ell_j^\top H^{-1}\nabla\ell_j$; this generalization is recorded in Proposition 8.1. □

Remark 5.3 (Higher-order corrections). The $O(n^{-2})$ remainder in (8) arises from the second-order Taylor remainder, the $H_{-j}\mapsto H$ replacement, and the $O(\|\nabla\ell_j\|^2/n^2)$ correction in Lemma 5.1. Each of these is bounded uniformly in n by the data-dependent constants in the statement of Part II.

6 Proof of Theorem 3.1: Part III (implication)

We now combine Parts I and II. Throughout this section we work with the basin masses $M_\tau(A) = \int_{B_A} p_\tau$ and $M_\tau(B) = \int_{B_B} p_\tau$ defined in Part I, and we adopt the equal-loss two-well setting $L(\theta_A^*) = L(\theta_B^*)$ (equivalently, $U(\theta_A^*) = U(\theta_B^*)$, see Lemma 4.3 and its derivation; the asymmetric case can be reduced to this one by shifting L by an additive constant within each basin). Recall Assumption 5: $\Sigma_A/H_A < \Sigma_B/H_B$.

6.1 Selection: SGD concentrates on the basin of θ_A^*

We apply Laplace's method to the integrals $M_\tau(A) = \int_{B_A} p_\tau(\theta)\,d\theta$ and $M_\tau(B) = \int_{B_B} p_\tau(\theta)\,d\theta$ directly. Near the minimum θ_k^* ($k\in\{A,B\}$), expand the effective potential $V_{\mathrm{eff}}(\theta) = U(\theta) + \frac{\tau}{2}\ln\Sigma(\theta)$ to second order: since $L'(\theta_k^*) = 0$, the integrand L'/Σ defining U has $U'(\theta_k^*) = 0$, and $U''(\theta_k^*) = (L''/\Sigma)(\theta_k^*) = H_k/\Sigma_k$. Therefore

$$p_\tau(\theta) \;=\; \frac{1}{Z_\tau\,\Sigma_k}\exp\Big(-\tfrac{2}{\tau}U(\theta_k^*)\Big)\exp\Big(-\tfrac{1}{\tau}\,\tfrac{H_k}{\Sigma_k}\,(\theta-\theta_k^*)^2\Big)(1+O(|\theta-\theta_k^*|))$$

on a neighbourhood of θ_k^*. Integrating against the Gaussian factor over B_k and using the standard Laplace estimate $\int_{-\infty}^{\infty}e^{-\alpha u^2}\,du = \sqrt{\pi/\alpha}$ (the contribution from outside any neighbourhood of θ_k^* is exponentially smaller in τ by Assumption 2 and Assumption 3),

$$M_\tau(k) \;=\; \frac{1}{Z_\tau\,\Sigma_k}\,\exp\Big(-\tfrac{2}{\tau}U(\theta_k^*)\Big)\,\sqrt{\pi\tau\,\Sigma_k/H_k}\,(1+O(\sqrt{\tau})).$$

Taking the ratio at A over B, the prefactor $1/Z_\tau$ cancels and the exponentials combine. Under the equal-loss assumption $U(\theta_A^*) = U(\theta_B^*)$,

$$\frac{M_\tau(A)}{M_\tau(B)} \;=\; \frac{\Sigma_B}{\Sigma_A}\sqrt{\frac{\Sigma_A/H_A}{\Sigma_B/H_B}}\,(1+O(\sqrt{\tau})) \;=\; \sqrt{\frac{\Sigma_B}{\Sigma_A}\,\frac{H_B}{H_A}}\,(1+O(\sqrt{\tau})). \tag{16}$$

Symmetric two-well case ($H_A = H_B$, $\Sigma_A < \Sigma_B$). Assumption 5 reduces to $\Sigma_A < \Sigma_B$, and (16) simplifies to $M_\tau(A)/M_\tau(B) = \sqrt{\Sigma_B/\Sigma_A}\,(1+O(\sqrt{\tau})) > 1$ for τ small.

General equal-loss case. In general, (16) shows that $M_\tau(A)/M_\tau(B) > 1$ iff $\Sigma_B H_B > \Sigma_A H_A$. This condition is *not* implied by Assumption 5 alone (only Σ/H smaller at A); accordingly we adopt for the implication (i) the explicit

$$\Sigma_B\,H_B \;>\; \Sigma_A\,H_A, \tag{17}$$

which holds automatically in the symmetric case ($H_A = H_B$) and otherwise must be checked. Under (17), for τ sufficiently small, $M_\tau(A)/M_\tau(B) > 1$, proving conclusion (i). We note that the predictive content of Part III is unaffected: parts (ii) and (iii) below depend only on Assumption 5, and the only role of (i) is to confirm that SGD's stationary mass is concentrated on the better basin in the symmetric case to which empirical tests of the theorem typically apply.

6.2 Generalization: θ_A^* has smaller LOO gap

Apply Part II at both minima:

$$\begin{aligned} \mathrm{Gap}_{\mathrm{LOO}}(\theta_A^*) &= \frac{\Sigma_A}{(n-1)\,H_A} + O(n^{-2}), \\ \mathrm{Gap}_{\mathrm{LOO}}(\theta_B^*) &= \frac{\Sigma_B}{(n-1)\,H_B} + O(n^{-2}). \end{aligned}$$

Subtracting,

$$\mathrm{Gap}_{\mathrm{LOO}}(\theta_B^*) - \mathrm{Gap}_{\mathrm{LOO}}(\theta_A^*) = \frac{1}{n-1}\Big(\tfrac{\Sigma_B}{H_B} - \tfrac{\Sigma_A}{H_A}\Big) + O(n^{-2}).$$

By Assumption 5, the leading term is strictly positive. This proves (ii) of Part III, with explicit margin $\big(\Sigma_B/H_B - \Sigma_A/H_A\big)/(n-1)$.

6.3 Test loss bound

By the standard $\mathbb{E}[\mathrm{Gap}_{\mathrm{LOO}}] = \mathbb{E}[L_{\mathrm{pop}}(\theta^*) - L(\theta^*)]$ identity [12], the LOO gap is an unbiased estimator of the population–training gap, provided that the data–model interaction does not introduce a systematic reversal between LOO behaviour and population behaviour. We make this explicit by invoking Assumption 6: under the negation of failure mode F4, the sign of the population-gap difference between two minima agrees with the sign of the LOO-gap difference, up to $o(1/n)$. Combined with $L(\theta_A^*) = L(\theta_B^*)$ (equal-loss two wells), the test loss bound at θ_A^* minus that at θ_B^* equals (9) up to $O(n^{-2})$. This proves (iii) of Part III. Without Assumption 6, parts (i) and (ii) still hold, but (iii) need not: this is the precise location at which the closed loop of Theorem 3.1 can be broken by F4, as discussed in Section 9. □

Remark 6.1 (Why this is a closed loop). The implication is genuinely a closed loop: Part I selects the minimum that has lower Σ (in the symmetric case, lower Σ_A/H_A in the asymmetric case via Laplace), Part II shows that lower Σ/H corresponds to a lower LOO gap, and Part III shows that the same intrinsic quantity drives both. The chain

$$\text{SGD stationary measure} \longrightarrow \text{low } \Sigma/H \longrightarrow \text{low LOO gap} \longrightarrow \text{low test loss}$$

is closed by Proposition 8.1 (which guarantees that the intermediate quantity is intrinsic). The chain breaks if and only if one of the failure modes F1–F4 of [2] is active; we make this precise in Section 9.

7 Multivariate extension via Freidlin–Wentzell

In dimension $d > 1$ the FPE (3) need not satisfy detailed balance, and the closed-form Gibbs density (4) need not exist. The correct generalization is to replace $U(\theta)$ by the *Freidlin–Wentzell quasipotential* [8].

7.1 Quasipotential definition and HJ equation

Definition 7.1 (Quasipotential)**.** Fix a stable equilibrium $\theta_0 \in \Theta$ of $-\nabla L$. The Freidlin–Wentzell quasipotential at $\theta \in \Theta$ is

$$V(\theta) \;:=\; \inf\Big\{ \tfrac{1}{2}\int_0^T (\dot\phi + \nabla L(\phi))^\top \, \Sigma(\phi)^{-1} \, (\dot\phi + \nabla L(\phi))\, dt \;:\; T>0,\ \phi(0)=\theta_0,\ \phi(T)=\theta \Big\}. \tag{18}$$

Lemma 7.2 (Hamilton–Jacobi characterization)**.** *Under Assumption 1–Assumption 3, the function V in Definition 7.1 is locally Lipschitz on Θ, and is the unique viscosity solution (in the Lipschitz sense) of the stationary Hamilton–Jacobi equation*

$$-\nabla L(\theta)^\top \nabla V(\theta) \;+\; \tfrac{1}{2}\nabla V(\theta)^\top \Sigma(\theta) \nabla V(\theta) \;=\; 0, \qquad V(\theta_0)=0,\ \ V\ge 0, \tag{19}$$

with the prescribed boundary value at θ_0. Moreover V is C^1 in a neighbourhood of θ_0 (where V admits a smooth quadratic expansion), but globally only Lipschitz regularity is guaranteed—V may fail to be C^1 along the cut locus of the underlying minimum-action geodesic flow.

Proof. Standard Freidlin–Wentzell theory; see [8, Chapter 4] or [9, Theorem 5.7.11]. Local Lipschitz regularity of V is a consequence of the uniform ellipticity of Σ and the smoothness of ∇L. Local C^1 regularity near θ_0 follows from the implicit function theorem applied to the Hamiltonian flow associated with (19) at the stable equilibrium $(\theta_0, 0)$. Global C^1 regularity is *not* guaranteed by Freidlin–Wentzell theory; it requires additional structural conditions on (L, Σ) that are not used in our theorem. □

Proposition 7.3 (Stationary measure with quasipotential)**.** *Under Assumption 1–Assumption 3, the FPE (3) has a unique stationary density p_τ, and as $\tau \to 0$,*

$$-\,\tau \ln p_\tau(\theta) \;\longrightarrow\; 2\,V(\theta) \quad \textit{uniformly on compact subsets of } \Theta,$$

where V is the quasipotential of Definition 7.1. The leading-order ratio between two stable equilibria is

$$\frac{p_\tau(\theta_A^*)}{p_\tau(\theta_B^*)} \;=\; \exp\Big(-\tfrac{2}{\tau}\big(V(\theta_A^*) - V(\theta_B^*)\big)\Big)\,(1+o(1)).$$

Proof. This is the Freidlin–Wentzell large-deviations principle for invariant measures of small-noise diffusions; see [8, Theorem 4.4.1] and [9, Theorem 5.7.12]. Existence and uniqueness of the stationary density follow as in Lemma 4.1. □

7.2 A 2D worked example

We now state and prove Theorem 7.4, in which detailed balance fails but the quasipotential admits a closed-form expansion to fourth order.

Theorem 7.4 (Multivariate extension; 2D worked example)**.** *Let $\Theta = \mathbb{R}^2$, $L(\theta_1, \theta_2) = \frac{1}{2}(\theta_1^2 + \theta_2^2)$, and*

$$\Sigma(\theta) \;=\; \mathrm{diag}\big(1+\theta_2^2,\, 1\big).$$

Then:

(a) *Detailed balance fails: the vector field $b(\theta) := \Sigma(\theta)^{-1}\nabla L(\theta)$ has nonzero curl, namely $\partial_{\theta_2} b_1 - \partial_{\theta_1} b_2 = -2\theta_1\theta_2/(1+\theta_2^2)^2 \neq 0$.*

(b) *The quasipotential V admits a Hamilton–Jacobi expansion*

$$V(\theta_1, \theta_2) \;=\; \theta_1^2 + \theta_2^2 \;-\; \tfrac{1}{2}\,\theta_1^2\theta_2^2 \;+\; O(\|\theta\|^6), \tag{20}$$

which is the unique even C^1 solution of (19) with $V(0) = 0$ and $V \ge 0$ to the displayed order.

(c) *The leading correction $-\frac{1}{2}\theta_1^2\theta_2^2$ is a saddle perturbation: V is reduced along the diagonal where both $|\theta_1|$ and $|\theta_2|$ are large simultaneously. Consequently the stationary measure places more mass along this diagonal direction than the naive Gaussian $\exp(-2L/\tau)$ would predict.*

Proof. (a) Compute

$$b(\theta) \;=\; \Sigma(\theta)^{-1}\,(\theta_1,\theta_2) \;=\; \Big(\tfrac{\theta_1}{1+\theta_2^2},\,\theta_2\Big).$$

Then $\partial_{\theta_2} b_1 = -2\theta_1\theta_2/(1+\theta_2^2)^2$ and $\partial_{\theta_1} b_2 = 0$. Hence $\operatorname{curl} b = -2\theta_1\theta_2/(1+\theta_2^2)^2 \neq 0$ on the open set $\{\theta_1\theta_2 \neq 0\}$, so detailed balance fails.

(b) The HJ equation (19) reads

$$-(\theta_1\partial_{\theta_1}V + \theta_2\partial_{\theta_2}V) \;+\; \tfrac{1}{2}\big((1+\theta_2^2)(\partial_{\theta_1}V)^2 + (\partial_{\theta_2}V)^2\big) \;=\; 0. \tag{21}$$

We seek an even formal power-series solution $V = \sum_{k\geq 1} V_{2k}$, where V_{2k} is homogeneous of degree $2k$. At leading order V_2 satisfies

$$-2V_2 \;+\; \tfrac{1}{2}\big(\|\nabla V_2\|^2\big) \;=\; 0 \;\;\Rightarrow\;\; V_2(\theta) = \theta_1^2 + \theta_2^2,$$

the unique even quadratic such that $V_2(0) = 0$, $V_2 \geq 0$. At order 4, write $V_4(\theta) = a\theta_1^4 + b\theta_1^2\theta_2^2 + c\theta_2^4$. Substituting $V = V_2 + V_4$ into (21) and collecting degree-4 terms requires three contributions:

(i) the drift part $-(\theta_1\partial_1 V_4 + \theta_2\partial_2 V_4) = -(4a\theta_1^4 + 4b\theta_1^2\theta_2^2 + 4c\theta_2^4)$,

(ii) the diffusion cross term $\nabla V_2 \cdot \nabla V_4 = 2\theta_1(4a\theta_1^3 + 2b\theta_1\theta_2^2) + 2\theta_2(2b\theta_1^2\theta_2 + 4c\theta_2^3) = 8a\theta_1^4 + 8b\theta_1^2\theta_2^2 + 8c\theta_2^4$,

(iii) the Σ-correction at order 4 from the $\theta_2^2(\partial_1 V)^2$ term in (21), which at leading order contributes $\frac{1}{2}\theta_2^2(\partial_1 V_2)^2 = \frac{1}{2}\theta_2^2 \cdot 4\theta_1^2 = 2\theta_1^2\theta_2^2$.

(No order-4 contribution arises from $(\partial_2 V_4)^2$ or $(\partial_1 V_4)^2$, since these are of order 6.) Summing the three contributions, (21) at order four becomes

$$-(4a\theta_1^4 + 4b\theta_1^2\theta_2^2 + 4c\theta_2^4) \;+\; (8a\theta_1^4 + 8b\theta_1^2\theta_2^2 + 8c\theta_2^4) \;+\; 2\theta_1^2\theta_2^2 \;=\; 0,$$

which simplifies to

$$4a\,\theta_1^4 \;+\; (4b+2)\,\theta_1^2\theta_2^2 \;+\; 4c\,\theta_2^4 \;=\; 0. \tag{22}$$

Matching coefficients of the three independent monomials gives $a = 0$, $b = -\frac{1}{2}$, $c = 0$, hence

$$V_4(\theta) \;=\; -\tfrac{1}{2}\,\theta_1^2\theta_2^2.$$

This proves (20): $V = \theta_1^2 + \theta_2^2 - \frac{1}{2}\theta_1^2\theta_2^2 + O(\|\theta\|^6)$.

We now verify directly that (20) solves (21) to order four. With $V = \theta_1^2 + \theta_2^2 - \frac{1}{2}\theta_1^2\theta_2^2$, $\partial_1 V = 2\theta_1 - \theta_1\theta_2^2$ and $\partial_2 V = 2\theta_2 - \theta_1^2\theta_2$. Then

$$\begin{aligned}
-\theta_1\partial_1 V - \theta_2\partial_2 V &= -\theta_1(2\theta_1 - \theta_1\theta_2^2) - \theta_2(2\theta_2 - \theta_1^2\theta_2)\\
&= -2(\theta_1^2 + \theta_2^2) + 2\theta_1^2\theta_2^2,\\
\tfrac{1}{2}(1+\theta_2^2)(\partial_1 V)^2 &= \tfrac{1}{2}(1+\theta_2^2)\big(4\theta_1^2 - 4\theta_1^2\theta_2^2 + \theta_1^2\theta_2^4\big)\\
&= 2\theta_1^2 + 2\theta_1^2\theta_2^2 - 2\theta_1^2\theta_2^2 + O(\|\theta\|^6)\\
&= 2\theta_1^2 + O(\|\theta\|^6),\\
\tfrac{1}{2}(\partial_2 V)^2 &= \tfrac{1}{2}(4\theta_2^2 - 4\theta_1^2\theta_2^2 + \theta_1^4\theta_2^2) = 2\theta_2^2 - 2\theta_1^2\theta_2^2 + O(\|\theta\|^6).
\end{aligned}$$

Summing,

$$-2\theta_1^2 - 2\theta_2^2 + 2\theta_1^2\theta_2^2 + 2\theta_1^2 + 2\theta_2^2 - 2\theta_1^2\theta_2^2 + O(\|\theta\|^6) \;=\; O(\|\theta\|^6),$$

confirming (21) to order four. Uniqueness among even C^1 solutions of polynomial form with $V(0) = 0$, $V \geq 0$ follows from the fact that (22) uniquely determines the three coefficients (a, b, c), and analogously at every higher even order the matching conditions are linear with unique solution.

(c) The Gaussian density $\exp(-2L/\tau) = \exp(-(\theta_1^2 + \theta_2^2)/\tau)$ corresponds to $V_{\text{naive}} = \theta_1^2 + \theta_2^2$. The correction $-\frac{1}{2}\theta_1^2\theta_2^2$ in (20) is non-positive, so $V \leq V_{\text{naive}}$, with equality on the axes $\{\theta_1 = 0\}$ and $\{\theta_2 = 0\}$ and strict inequality in the off-axis region. Consequently $\exp(-2V/\tau) \geq \exp(-2V_{\text{naive}}/\tau)$, establishing (c). □

Remark 7.5 (Curl drives the correction). The order-four correction $-\frac{1}{2}\theta_1^2\theta_2^2$ in (20) arises from the same term that produces nonzero curl in part (a). More precisely, the symmetric part of $\Sigma^{-1}\nabla\nabla L$ contributes to V_2 while the antisymmetric part—namely $\operatorname{curl} b$—is the source term for V_4 through the cross-coupling in (21). In dimension $d = 1$ curl vanishes identically, recovering the closed-form (4).

8 Reparameterization invariance

Proposition 8.1 (Reparameterization invariance)**.** *Let $\Phi : \Theta' \to \Theta$ be a C^2 diffeomorphism, write $\theta = \Phi(\phi)$, and let $J_\Phi(\phi)$ be the Jacobian. Define the pulled-back loss $\widetilde{L} := L \circ \Phi$ and the pulled-back per-sample losses $\widetilde{\ell}_i := \ell_i \circ \Phi$. At a critical point $\phi^* := \Phi^{-1}(\theta^*)$ of $\widetilde{L}$:*

(a) *The Hessian transforms as $\widetilde{H}(\phi^*) = J_\Phi(\phi^*)^\top H(\theta^*) J_\Phi(\phi^*)$.*

(b) *The noise covariance transforms as $\widetilde{\Sigma}(\phi^*) = J_\Phi(\phi^*)^\top \Sigma(\theta^*) J_\Phi(\phi^*)$.*

(c) *Consequently $\operatorname{tr}(\widetilde{H}^{-1}\widetilde{\Sigma}) = \operatorname{tr}(H^{-1}\Sigma)$, and in particular in dimension $d = 1$ the scalar ratio Σ/H is invariant.*

Proof. Write $J = J_\Phi(\phi^*)$. At any ϕ, $\widetilde{\ell}_i(\phi) = \ell_i(\Phi(\phi))$, so by the chain rule

$$\nabla_\phi \widetilde{\ell}_i(\phi) \;=\; J_\Phi(\phi)^\top \nabla_\theta \ell_i(\Phi(\phi)).$$

At ϕ^* this becomes $\nabla_\phi \widetilde{\ell}_i(\phi^*) = J^\top \nabla_\theta \ell_i(\theta^*)$. Therefore

$$\widetilde{\Sigma}(\phi^*) \;=\; \frac{1}{n}\sum_{i=1}^n \nabla_\phi \widetilde{\ell}_i(\phi^*) \nabla_\phi \widetilde{\ell}_i(\phi^*)^\top \;=\; J^\top \Big(\frac{1}{n}\sum_{i=1}^n \nabla_\theta \ell_i \nabla_\theta \ell_i^\top\Big) J \;=\; J^\top \Sigma(\theta^*) J,$$

using $\nabla_\theta L(\theta^*) = 0$ to drop the cross term. This proves (b).

For (a), differentiate $\nabla_\phi \widetilde{L}(\phi) = J_\Phi(\phi)^\top \nabla_\theta L(\Phi(\phi))$ once more with respect to ϕ:

$$\nabla_\phi^2 \widetilde{L}(\phi) \;=\; J_\Phi(\phi)^\top \nabla_\theta^2 L(\Phi(\phi))\, J_\Phi(\phi) \;+\; \sum_{a=1}^d \big[\nabla_\theta L(\Phi(\phi))\big]_a \nabla_\phi^2 \Phi_a(\phi).$$

At ϕ^*, $\nabla_\theta L(\theta^*) = 0$, so the second sum vanishes and we obtain $\widetilde{H}(\phi^*) = J^\top H(\theta^*) J$, proving (a).

For (c), with $\widetilde{H} = J^\top H J$ and $\widetilde{\Sigma} = J^\top \Sigma J$, and assuming H, J invertible (so $\widetilde{H}$ is also invertible):

$$\widetilde{H}^{-1}\widetilde{\Sigma} \;=\; (J^\top H J)^{-1}(J^\top \Sigma J) \;=\; J^{-1}H^{-1}J^{-\top} J^\top \Sigma J \;=\; J^{-1}H^{-1}\Sigma\, J.$$

Hence

$$\operatorname{tr}(\widetilde{H}^{-1}\widetilde{\Sigma}) \;=\; \operatorname{tr}(J^{-1}H^{-1}\Sigma\, J) \;=\; \operatorname{tr}(H^{-1}\Sigma\, J\, J^{-1}) \;=\; \operatorname{tr}(H^{-1}\Sigma),$$

using cyclicity of the trace. In dimension $d = 1$, J and H are nonzero scalars and $\widetilde{\Sigma}/\widetilde{H} = (J^2\Sigma)/(J^2 H) = \Sigma/H$. □

Remark 8.2 (The LOO gap is intrinsic). Combined with Theorem 3.1 Part II, Proposition 8.1 (c) implies that, in dimension $d > 1$, the LOO gap formula (8) reads $\mathrm{Gap}_{\mathrm{LOO}}(\theta^*) = \mathrm{tr}(H^{-1}\Sigma)/(n-1) + O(n^{-2})$, and is invariant under reparameterization. In contrast, the curvature alone (e.g. a single eigenvalue or $\det H$) is not invariant [14], and so cannot serve, by itself, as an intrinsic complexity measure. The empirical phenomenon that large-batch SGD tends to find "sharp" minima with worse generalization [19] is consistent with the present picture once curvature is replaced by the reparameterization-invariant quantity $\mathrm{tr}(H^{-1}\Sigma)$.

9 Discussion

9.1 Boundary with paper B's failure modes

The closed-loop Theorem 3.1 fails exactly when one of the four failure modes catalogued in [2] is active, in the following precise correspondence.

F1 (insufficient mixing). The proof of Part I uses the FPE stationary density as a proxy for the SGD long-time average; this requires the mixing time τ_{mix} between the two basins to be much smaller than the runtime T. When $\tau_{\mathrm{mix}} \gg T$ (Kramers regime, $\tau_{\mathrm{mix}} \sim \exp(2\Delta V/\tau)$ for barrier height ΔV), SGD does not reach the stationary measure and Part I's prediction does not apply.

F2 (discrete instability). The SDE approximation (2) is valid only for η below the discrete stability threshold $2/\lambda_{\max}(H)$ [16, 17]; above it the SGD recursion diverges and no stationary measure exists. ([3] provides the SDE-approximation error bound that breaks down in this regime, but the threshold itself originates in classical gradient-descent stability analysis.)

F3 (degenerate selection). When $\Sigma_A/H_A = \Sigma_B/H_B$, Assumption 5 fails: the implication of Part III is vacuous and the test-loss bounds at the two minima coincide to leading order.

F4 (noise–generalization reversal). The implication *(iii)* of Part III rests on the alignment of low Σ/H with low population loss. Paper B exhibits explicit examples in which this alignment is reversed: a memorizing minimum may have $\Sigma_B \ll \Sigma_A$ at a generalizing minimum because all per-sample gradients agree at the memorized solution. When F4 is active, Theorem 3.1 Parts I and II still hold individually, but their composition does not yield a generalization guarantee. In this sense F4 is the one fundamental obstruction; F1–F3 can be removed by tuning of η, B, T, but F4 is a property of the data–architecture pair.

9.2 Relation to paper A (interpolation regime)

Paper A [1] studies the interpolation regime where $\Sigma(\theta^*) = 0$ at every interpolating solution. There the Fokker–Planck argument of Section 4 is vacuous (the diffusion degenerates), and a discrete Lyapunov-exponent analysis takes its place. The quantities that play the role of Σ/H are:

- In paper A, the per-sample curvature $h_{\max} = \max_i \|\nabla_\theta f_{\theta^*}(x_i)\|^2$ and the Jacobian complexity $\mathcal{C} = \|J\|_F/\sqrt{n}$. Stability requires $\eta\, h_{\max} < 2$, and stable solutions have $\mathcal{C} \le \sqrt{2/\eta}$.
- In the present paper C, the noise-to-curvature ratio Σ/H (or $\mathrm{tr}(H^{-1}\Sigma)$).

Both are intrinsic (Proposition 8.1 for the present paper; [1, Sec. 6] for the discrete case), and both deliver $O(\rho/\sqrt{\eta n})$- or $O(1/n)$-type generalization bounds. The unifying picture is that SGD's noise selects against an intrinsic complexity measure—multiplicative noise covariance in the non-interpolation regime, per-sample Jacobian curvature in the interpolation regime—and that this

selection is reflected in sample-based generalization (LOO in the present setting; Rademacher in [1]). The discrete-stability viewpoint in [18] provides an alternative (dynamical-stability-based) treatment of implicit regularization in the same regime addressed by paper A; our paper C complements this by treating the non-interpolation regime via Fokker–Planck.

9.3 Open questions

1. *Higher-order quasipotential.* Theorem 7.4 computes the quasipotential to order $\|\theta\|^4$. Beyond order 4, the HJ equation (21) becomes substantially more complex; a global structural understanding of V in $d = 2$ is open.

2. *Multivariate detailed balance.* Identifying the largest class of (L, Σ) for which detailed balance holds beyond the scalar-isotropic case is open. [6] treats the constant-Σ case; the state-dependent case requires curl conditions of the form $\mathrm{curl}(\Sigma^{-1}\nabla L) = 0$, whose general structure is unknown.

3. *Sharp test-loss bound.* The implication (iii) of Part III gives a relative bound but not an absolute one. Combining Theorem 3.1 with concentration inequalities to obtain an absolute population-loss bound (in expectation or with high probability) would close a remaining gap.

4. *F4 detection.* When does F4 occur in deep networks? Paper B [2] provides constructions but the empirical prevalence is open.

9.4 Testable predictions

Paper-C's theorem suggests the following testable predictions, which can be empirically verified with standard SGD experiments:

1. For two locally quadratic minima with measurable H and Σ, the ratio of stationary basin masses under SGD should match $\sqrt{\Sigma_B/\Sigma_A}$ in the symmetric-well case ($H_A = H_B$), per the basin-mass formula (16); in the asymmetric equal-loss case the prediction generalizes to $\sqrt{(\Sigma_B/\Sigma_A)(H_B/H_A)}$ via Laplace approximation.

2. The LOO gap, which is computable in $O(n)$ time at any local minimum, should match $\Sigma/(H(n-1))$ to leading order. If it does not, F4 is a possible explanation.

3. Reparameterizing the parameter space (e.g. rescaling layers in a neural network) should leave $\mathrm{tr}(H^{-1}\Sigma)$ unchanged at any minimum, even though both H and Σ individually rescale.

References

[1] L. Chang. *Discrete Lyapunov stability of SGD in the interpolation regime: per-sample analysis and generalization bounds.* Companion paper, 2026.

[2] L. Chang. *When does SGD fail to find good solutions? Four failure modes and necessary conditions for implicit regularization.* Companion paper, 2026.

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

[4] S. Mandt, M. D. Hoffman, and D. M. Blei. *Stochastic gradient descent as approximate Bayesian inference.* Journal of Machine Learning Research **18**(134):1–35, 2017.

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

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

[7] J. Z. HaoChen, C. Wei, J. D. Lee, and T. Ma. *Shape matters: understanding the implicit bias of the noise covariance.* In Proceedings of COLT, pp. 2315–2357, 2021.

[8] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*, third edition. Grundlehren der mathematischen Wissenschaften, vol. 260, Springer, 2012.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, second edition. Stochastic Modelling and Applied Probability, vol. 38, Springer, 2010.

[10] G. A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations.* Texts in Applied Mathematics, vol. 60, Springer, 2014.

[11] H. Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*, second edition. Springer Series in Synergetics, vol. 18, Springer, 1996.

[12] O. Bousquet and A. Elisseeff. *Stability and generalization.* Journal of Machine Learning Research **2**:499–526, 2002.

[13] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*, second edition. MIT Press, 2018.

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

[15] L. Bottou, F. E. Curtis, and J. Nocedal. *Optimization methods for large-scale machine learning.* SIAM Review **60**(2):223–311, 2018.

[16] B. T. Polyak. *Introduction to Optimization.* Optimization Software, Inc., New York, 1987.

[17] Y. Nesterov. *Lectures on Convex Optimization*, second edition. Springer Optimization and Its Applications, vol. 137, Springer, 2018.

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

[19] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. *On large-batch training for deep learning: generalization gap and sharp minima.* In Proceedings of ICLR, 2017.