# Minima Selection in Stochastic Optimization: A Long-Time, Small-Stepsize Perspective

Xiang Li[†] Zebang Shen[†] Ya-Ping Hsieh[†] Niao He[†]

**Abstract**

Nonconvex stochastic optimization often admits many global minimizers, and an optimizer may do more than find a point of small loss: it also selects among minima. We study this selection through the small-stepsize stationary distribution of a diffusion approximation to stochastic algorithms, under the assumption that the set of global minimizers is a compact manifold $\mathcal{M}$. Assuming a local manifold WKB representation $p^\varepsilon(x) = c^\varepsilon(x)\exp(-V(x)/\varepsilon)$ near $\mathcal{M}$, we identify the limiting stationary measure induced on $\mathcal{M}$ as the stepsize tends to zero. This limiting bias separates into three contributions: the intrinsic geometry of the minima manifold, a normal-curvature term determined by the large-deviation quasi-potential, and a prefactor that captures the remaining redistribution of mass along the manifold. Our main technical step is a reduction of the stationary Fokker–Planck equation to the minimizer manifold. After substituting the WKB ansatz, we obtain an exact identity whose leading orders yield a Hamilton–Jacobi equation for $V$, a transport equation for the prefactor, and finally a closed elliptic PDE on $\mathcal{M}$ for the leading manifold prefactor. This replaces the ambient stationary problem by a lower-dimensional equation on the space where selection takes place.

## 1 Introduction

Modern stochastic optimization methods often drive the training loss close to zero, especially in over-parameterized models (Allen-Zhu, Y. Li, and Song 2019; Du et al. 2019), yet the solutions they return can generalize quite differently. For example, stochastic gradient descent (SGD) may select solutions that differ from those found by deterministic or adaptive variants (Keskar et al. 2016; C. Zhang et al. 2017). This suggests that an optimizer does more than locate a point with small loss: when many minimizers are available, it also places a bias on which minimizers are selected. Understanding this optimizer-dependent bias is therefore part of the broader problem of explaining why different training procedures lead to different generalization behavior.

We study this question through the diffusion approximation of stochastic gradient methods. Formally, consider the optimization problem

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f$ is smooth and possibly nonconvex. The update rule of SGD (Robbins and Monro 1951) reads

$$x_{k+1} = x_k - \varepsilon g(x_k;\xi_k), \qquad \mathbb{E}_\xi\left[g(x;\xi)\right] = \nabla f(x), \tag{1}$$

where $\varepsilon$ is the stepsize. A natural continuous-time approximation of this discrete algorithm is the diffusion (Mandt, M. D. Hoffman, and D. M. Blei 2015; Mandt, M. Hoffman, and D. Blei 2016):

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2\varepsilon D(X_t)}\,dW_t, \tag{2}$$

[†]Department of Computer Science, ETH Zurich, Switzerland. xiang.li@inf.ethz.ch, zebang.shen@inf.ethz.ch, yaping.hsieh@inf.ethz.ch, niao.he@inf.ethz.ch.

where $D$ is the covariance matrix of the stochastic gradient $g$.

To understand the bias induced by the diffusion Equation (2), we study its stationary behavior in the small-stepsize regime. Let $\mu^\varepsilon$ denote a stationary measure of the diffusion. Regions that carry more stationary mass are then precisely the regions preferred by the noisy dynamics. The family $\{\mu^\varepsilon\}_{\varepsilon>0}$ therefore encodes how minima selection depends on the stepsize. As $\varepsilon \to 0$, these measures concentrate on the global minima of $f$. Assuming $\arg\min f$ is a compact manifold $\mathcal{M}$, the remaining question is how this limiting mass is distributed along $\mathcal{M}$. Put differently, once the dynamics has identified the correct set of minima, what determines which parts of that set carry more stationary weight?

Our main result gives an explicit description of this limiting measure. Under a local WKB description of the stationary density in $\mathcal{M}$, the induced measure on $\mathcal{M}$ is shaped by three effects:

- the intrinsic geometry of the minima manifold, represented by the volume element $d\mathcal{M}$;
- a noise-modified normal sharpness term determined by the Hessian of a quasi-potential $V$, itself determined by the loss $f$ and the noise covariance $D$; in the reversible Langevin case, where $V = f$, this reduces to the usual Hessian-based notion of sharpness;
- a prefactor $c_0$ that captures the remaining redistribution of mass along $\mathcal{M}$.

The last term is nontrivial in general and is determined by an elliptic PDE posed directly on $\mathcal{M}$, so the remaining selection bias can be studied intrinsically on the manifold of minimizers itself.

Our contributions are summarized as follows:

1. Under a local manifold WKB ansatz for the stationary density of the diffusion Equation (2) near the manifold of global minimizers $\mathcal{M}$, we derive the limiting stationary measure on $\mathcal{M}$. The limiting mass factors into three components: the intrinsic volume element on $\mathcal{M}$, a noise-modified normal curvature term $\det(\partial_r^2 V)^{-1/2}$, and a prefactor $c_0$.
2. We show that the quasi-potential $V$ governing the exponential scale of the stationary density is characterized locally by the stationary Fokker–Planck equation through a Hamilton–Jacobi equation. Differentiating this relation along the normal directions yields a Riccati equation for the normal Hessian of $V$, which determines the curvature term in the limiting measure.
3. We derive a closed equation for the residual manifold-selection factor $c_0$. First, $c_0$ satisfies a transport equation in a neighborhood of $\mathcal{M}$. Then, by restricting the stationary WKB identity to $\mathcal{M}$ and eliminating normal derivatives, we obtain a second-order elliptic PDE posed directly on the minimizer manifold. This reduces the ambient stationary problem to a lower-dimensional equation on the true selection space.

## 1.1 Related work

**Large deviation principles (LDP) and invariant measures.** Small-noise asymptotics for invariant measures of diffusions are classical. Freidlin–Wentzell theory and the invariant-density analyses of (Sheu 1986; Mikami 1990; Freidlin and Wentzell 2012) identify the quasi-potential as the object governing the exponential scale of stationary densities. These works motivate the WKB form $p^\varepsilon(x) \asymp \exp(-V(x)/\varepsilon)$, but the clean classical asymptotics are typically formulated for isolated stable equilibria rather than for a compact manifold of minimizers. In stochastic optimization, Azizian et al. (2024) derive a LDP of the long-run distribution of SGD across connected components of critical points, and Bajovic, Jakovetic, and Kar (2023) establish large-deviation rates for SGD in the strongly convex regime. Our focus is complementary: LDP results identify the exponential scale of the stationary density through the quasi-potential, but they do not determine the subexponential prefactor. We study this finer structure and show how it affects the limiting stationary measure on the minimizer manifold.

**Escape dynamics and manifold dynamics via SDEs.** Another line of work studies minima selection through transient dynamics near minima. Diffusion-based analyses connect SGD noise to escape from sharp minima or saddles (Hu et al. 2019; Xie, Sato, and Sugiyama 2020; Mori et al. 2022; Ibayashi and Imaizumi 2023). Closer to our geometric setting, Z. Li, T. Wang, and Arora (2022) study the regime after SGD reaches zero loss and derive an effective SDE for the dynamics near a manifold of minimizers. Our contribution differs in that we directly analyze the stationary measure induced on the minimizer manifold itself, obtaining an explicit limiting density formula together with a closed elliptic PDE on the manifold.

**Stability and implicit regularization.** A separate line of work studies optimizer bias through stability and modified objectives. For full-batch GD, Cohen et al. (2021) observe the edge-of-stability regime, i.e., GD converges to minima with a maximum Hessian eigenvalue close to $2/\varepsilon$. For SGD, stability analyses relate accessible minima to curvature, batch size, and noise structure (L. Wu, C. Ma, et al. 2018; L. Wu, M. Wang, and Su 2022). A complementary viewpoint interprets algorithmic bias through modified objectives: Damian, T. Ma, and Lee (2021) show that label-noise SGD converges toward stationary points of an explicit sharpness-penalized regularized loss, and L. Zhang et al. (2025) show that zeroth-order optimization implicitly favors minima with small Hessian trace through a smoothed objective.

# 2 Geometric Setup and Small-Stepsize Asymptotics

This section introduces the small-stepsize setting used throughout the paper and the local WKB framework that underlies our analysis. We assume that the set of global minimizers of the objective forms a compact manifold $\mathcal{M}$, and we study how the stationary distribution of Equation (2) concentrates near $\mathcal{M}$ and how it distributes mass along that manifold.

## 2.1 Geometry Near the Minimizer Manifold

We begin with the geometric structure of $f$ and its minimizer set.

**Assumption 2.1** (Single-Valley Manifold Objective)**.** *Let $f \in C^4(\mathbb{R}^d)$ with $\min_{x\in\mathbb{R}^d} f = 0$. Assume:*

*1. there exists a connected compact $C^4$ manifold $\mathcal{M} \subset \mathbb{R}^d$ of dimension $0 < n < d$ such that*

$$\{x \in \mathbb{R}^d \mid \nabla f(x) = 0\} = \arg\min f = \mathcal{M};$$

*2. $\nabla^2 f$ is non-degenerate in the normal directions on $\mathcal{M}$: for every $x \in \mathcal{M}$ and every nonzero $v \in \mathcal{N}_x\mathcal{M}$,*

$$v^\intercal \nabla^2 f(x) v > 0;$$

*3. $f$ is coercive:*

$$f(x) \to +\infty \qquad \text{as} \quad \|x\| \to \infty.$$

Assumption 2.1 says that every critical point is globally minimizing, so the only stationary set of the gradient flow is the manifold $\mathcal{M}$. The normal nondegeneracy condition gives quadratic confinement away from $\mathcal{M}$, while coercivity prevents escape to infinity. Assumptions of this type are compatible with several settings that arise in optimization. In over-parameterized models, Laurent and Brecht (2018) and Luo, C. Wu, and Lee (2018) suggest that all local minima are global, while Nguyen (2019) establish connectedness of the set of global minima in certain neural-network models. On the geometric side, Rebjock and Boumal (2025) show that minimizers of $C^2$ Polyak–Lojasiewicz functions form a submanifold.

To analyze the stationary density near $\mathcal{M}$, we introduce local tangent-normal coordinates. Write

$$x(u, r) = m(u) + N(u)r,$$

where $u \in \mathbb{R}^n$ is a tangential coordinate, $r \in \mathbb{R}^{d-n}$ is a normal coordinate, $m(u) \in \mathcal{M}$, and the columns of $N(u)$ form an orthonormal basis of the normal space. The induced volume measure on the manifold is denoted by $d\mathcal{M}$. These coordinates are local and are used only inside a tubular neighborhood of $\mathcal{M}$. Once they are fixed, we write functions in coordinates by composition with the parametrization. For example, $f(u)$ means $f(m(u))$, and more generally $f(u, r)$ means $f(x(u, r))$; the intended pullback will always be clear from the context.

We also impose a regularity and ellipticity condition on the diffusion process.

**Assumption 2.2** (Regularity of the Equation (2))**.** *Assume the SDE is well-posed and satisfies the following:*

1. *the diffusion tensor $D \in C^2(\mathbb{R}^d)$ is uniformly positive definite on $\mathbb{R}^d$;*
2. *for each sufficiently small $\varepsilon > 0$, the Equation (2) admits a unique invariant density $p^\varepsilon$, and the associated invariant measures $\{\mu^\varepsilon\}_{\varepsilon>0}$ are tight;*
3. *the quasi-potential $V$ defined below belongs to $C^4$ in a neighborhood of $\mathcal{M}$.*

The first item is the uniform ellipticity hypothesis on the diffusion, the second ensures that the stationary family is well defined, and the third is the local smoothness assumption needed later when differentiating the WKB identities near $\mathcal{M}$.

## 2.2 Quasi-Potential and Manifold WKB Ansatz

Let $p^\varepsilon = d\mu^\varepsilon/dx$ denote the stationary density of Equation (2). To study the small-stepsize behavior of $p^\varepsilon$, we first introduce the quasi-potential associated with the minimizer manifold $\mathcal{M}$:

**Definition 2.1** (Quasi-potential)**.** The quasi-potential associated with $\mathcal{M}$ is defined by

$$V(x) := \frac{1}{4} \inf_{\substack{x_0 \in \mathcal{M},\, T \\ \psi(0)=x_0,\ \psi(T)=x}} \int_0^T \Big(\dot\psi(t) + \nabla f(\psi(t))\Big)^\intercal D(\psi(t))^{-1} \Big(\dot\psi(t) + \nabla f(\psi(t))\Big)\, dt. \tag{3}$$

This quantity is the least Freidlin–Wentzell action needed to travel from some point on the minimizer manifold to the point $x$. Under our assumptions,

$$\arg\min V = \arg\min f = \mathcal{M},$$

and $V$ is comparable to the objective $f$ up to multiplicative constants; see Lemma A.2 in Appendix A. In particular, $V$ plays the role of the large-deviation rate function for the stationary density, which satisfies the following large-deviation principle (LDP).

**Theorem 2.1** (Stationary-density LDP, Corollary of Mou et al. (2025, Theorem A))**.** *Assume Assumptions 2.1 and 2.2. Then, for every $x \in \mathbb{R}^d$,*

$$\lim_{\varepsilon \to 0} \varepsilon \log p^\varepsilon(x) = -V(x).$$

The large-deviation principle identifies the exponential scale of the stationary density. This is the motivation for the manifold WKB ansatz: once the leading exponential part is fixed by $V$, the remaining question is how the sub-exponential prefactor redistributes mass along the minimizer manifold. To capture that finer structure, we assume a local WKB representation in a fixed tubular neighborhood of $\mathcal{M}$.

**Assumption 2.3** (Manifold WKB ansatz)**.** *There exists a tubular neighborhood $U$ of $\mathcal{M}$, with compact closure, such that the stationary density admits the representation*

$$p^\varepsilon(x) \propto c^\varepsilon(x) \exp\left(-\frac{V(x)}{\varepsilon}\right), \qquad x \in U,$$

*where*

$$c^\varepsilon(x) = c_0(x) + r^\varepsilon(x),$$

*with $c_0 \in C^2(U)$, $c_0 > 0$ on $U$, and*

$$\|r^\varepsilon\|_{C^2(U)} \to 0 \qquad \text{as } \varepsilon \to 0.$$

*Remark* 2.1 (Validity of the ansatz)*.* For isolated nondegenerate equilibria, asymptotic expansions of this type are classical (Sheu 1986; Mikami 1990). For a manifold of equilibria, the corresponding justification would be more delicate. In this paper, we take the WKB expansion as a local assumption.

The exponential factor $\exp(-V(x)/\varepsilon)$ has already appeared in the optimization literature in the small-stepsize regime (Hu et al. 2019; Azizian et al. 2024). For minima selection, however, that exponential term describes only the leading concentration toward $\mathcal{M}$. As the next section shows, the induced measure on $\mathcal{M}$ also depends on the prefactor $c^\varepsilon$. Determining the full stationary density is generally difficult, especially for non-reversible diffusions, but the WKB ansatz lets us focus directly on the leading prefactor $c_0$ that controls the limiting distribution on the manifold.

# 3 The Limiting Distribution on the Minima Manifold

This section has two goals. We first derive the measure obtained by collapsing the stationary density onto the minimizer manifold $\mathcal{M}$ as $\varepsilon \to 0$. We then show how the stationary Fokker–Planck equation characterizes the three ingredients appearing in that limiting measure. The proofs are deferred to Section B.

## 3.1 From the WKB Ansatz to the Limiting Measure

**Theorem 3.1** (The limiting distribution)**.** *Let $d\mu^\varepsilon(x) = p^\varepsilon(x)\,dx$ with $p^\varepsilon$ satisfying Assumption 2.3. Then, as $\varepsilon \to 0$, $\mu^\varepsilon$ converges weakly to a measure $\mu_0$ supported on $\mathcal{M}$, and given by*

$$d\mu_0(u) \propto c_0(u)\, \det\big(\partial_r^2 V(u,0)\big)^{-1/2}\, d\mathcal{M}(u).$$

*Here $d\mathcal{M}(u)$ denotes the volume element, and $\partial_r^2 V(u,0)$ is the normal Hessian of the pullback of $V$ in the associated tangent-normal coordinates.*

Theorem 3.1 shows that the limiting measure is determined by the normal behavior of the quasi-potential and by the leading prefactor along $\mathcal{M}$. These factors are not yet explicitly characterized. For the quasi-potential, for example, Equation (3) provides a variational definition, but that formula alone is not directly computable. The theorem should therefore be read as a structural formula: after concentration onto $\mathcal{M}$, the remaining weight at $u \in \mathcal{M}$ is the product of a geometric factor, a normal Laplace factor, and the manifold value of the prefactor.

To characterize these terms more explicitly, we return to the stationary Fokker–Planck equation for $p^\varepsilon$:

$$\operatorname{div}_x\big((\nabla f)\, p^\varepsilon\big) + \varepsilon \sum_{i,j} \partial^2_{x_i,x_j}\big([D]_{i,j} p^\varepsilon\big) = 0.$$

After substituting the WKB ansatz into this equation, we obtain a formal expansion in powers of $\varepsilon$ of the schematic form

$$\frac{1}{\varepsilon}c_0\mathcal{H}[V] - \mathcal{L}_1 c_0 + \varepsilon\big(\mathcal{L}_2 c_0 + \cdots\big) = 0,$$

where $\mathcal{H}[V]$, $\mathcal{L}_1 c_0$, and $\mathcal{L}_2 c_0$ denote the successive coefficients at orders $1/\varepsilon$, 1, and $\varepsilon$, respectively. Their explicit forms are given in Lemma B.1. The next three subsections follow the formal WKB strategy: we characterize $V$ and $c_0$ by setting the coefficients at orders $1/\varepsilon$, 1, and $\varepsilon$ to zero. Under the WKB ansatz, this formal step is justified by the regularity of $c^\varepsilon$ and the smoothness of the coefficients on the fixed neighborhood $U$. Indeed, once we divide the stationary equation by the relevant power of $\varepsilon$ and let $\varepsilon \to 0$, all higher-order terms vanish, so the coefficient at the order under consideration must be zero. In this way, the $1/\varepsilon$ equation determines the quasi-potential $V$ and, in particular, its normal Hessian, while the next two equations govern the leading prefactor $c_0$.

## 3.2 Hamilton–Jacobi Characterization of the Quasi-Potential

We begin with the $1/\varepsilon$ equation in this expansion, which determines the quasi-potential $V$.

**Lemma 3.1** (Hamilton–Jacobi equation and normal Riccati relation)**.** *Assume Assumptions 2.1 to 2.3. Then $V$ satisfies*

$$-\langle \partial_x V, \partial_x f\rangle + \langle \partial_x V, D\partial_x V\rangle = 0 \qquad \text{on } U. \tag{4}$$

*Furthermore, differentiating Equation (4) in the normal directions along $\mathcal{M}$ yields*

$$\partial_r^2 V\, \partial_r^2 f + \partial_r^2 f\, \partial_r^2 V = 2\partial_r^2 V\, N^\intercal D N\, \partial_r^2 V \qquad \text{on } \mathcal{M}.$$

For the limiting measure, the normal Riccati relation characterizes the Hessian $\partial_r^2 V$ along $\mathcal{M}$, and therefore the normal-curvature factor $\det(\partial_r^2 V)^{-1/2}$ appearing in Theorem 3.1. Intuitively, $\partial_r^2 V$ is obtained from the normal Hessian of $f$ after weighting by the inverse diffusion in the normal directions. This is most transparent in one dimension, or more generally when the normal eigendirections of $\partial_r^2 f$ and $N^\intercal D N$ are aligned: the Riccati equation then decouples mode by mode, and each normal curvature of $V$ is determined by the corresponding curvature of $f$ scaled by the diffusion in that direction.

## 3.3 Transport Equation for the Leading Prefactor

Having characterized the quasi-potential through the $1/\varepsilon$ equation, we next turn to the coefficient of order 1. This yields the first equation satisfied by the leading prefactor $c_0$.

**Lemma 3.2** (Transport equation for $c_0$)**.** *Assume Assumptions 2.1 to 2.3. Then $c_0$ satisfies*

$$\langle \partial_x c_0, -\partial_x f + 2D\partial_x V\rangle + c_0\Big(\mathrm{Tr}(-\partial_x^2 f + D\partial_x^2 V) + 2\langle \mathrm{div}_x\, D, \partial_x V\rangle\Big) = 0 \qquad \text{on } U. \tag{5}$$

*Define*

$$F(x) := \mathrm{Tr}(-\partial_x^2 f + D\partial_x^2 V)(x) + 2\langle \mathrm{div}_x\, D(x), \partial_x V(x)\rangle.$$

*If $\psi$ solves the characteristic equation*

$$\dot\psi(s) = -\partial_x f(\psi(s)) + 2D(\psi(s))\partial_x V(\psi(s)), \qquad \psi(0) = x,$$

*then we have the path-integral representation*

$$c_0(x) = c_0(\bar x)\exp\left(-\int_{-\infty}^0 F(\psi(s))\,ds\right) \quad \text{where} \quad \bar x = \lim_{s\to-\infty}\psi(s) \in \mathcal{M}.$$

This transport equation governs how the leading prefactor extends away from the manifold along the characteristic drift $-\nabla f + 2D\nabla V$: once the value of $c_0$ is known on $\mathcal{M}$, the equation transports it into the surrounding neighborhood. This order alone does not fully determine $c_0$, because it still requires the boundary value of $c_0$ on the stable set of the effective drift, namely the manifold $\mathcal{M}$. Additional information is therefore needed, which is why we continue to the next order.

The direct restriction of the transport equation to $\mathcal{M}$ is trivial because both $\nabla f$ and $\nabla V$ vanish there. However, differentiating the transport equation in the normal directions produces relations for $\partial_r c_0$, $\partial^2_{u,r} c_0$, and $\partial^2_r c_0$, and these are exactly the ingredients needed to eliminate the normal derivatives in the final manifold equation.

*Remark* 3.1. In the classical WKB setting, where the stable set consists of a single point, Bouchet and Reygner (2016) and related works effectively stop at this stage, because the boundary value of $c_0$ at that point is absorbed into the overall normalization. In our setting, by contrast, the stable set is the manifold $\mathcal{M}$, so the restriction of $c_0$ to $\mathcal{M}$ remains a genuine unknown function. This is precisely what distinguishes the present manifold problem from the usual single-equilibrium WKB analysis.

### 3.4 Elliptic Closure on the Manifold

The coefficient of order $\varepsilon$ in the WKB expansion supplies exactly the missing information from the previous subsection. Restricted to $\mathcal{M}$, it produces an identity involving second derivatives of $c_0$.

**Lemma 3.3** (Second-order equation restricted to the manifold)**.** *Assume Assumptions 2.1 to 2.3. Then*

$$\left( D : \partial_x^2 c_0 + \sum_{i,j} c_0 \, \partial^2_{x_i,x_j} [D]_{i,j} + 2\langle \partial_x c_0, \operatorname{div}_x D \rangle \right) \Big|_{\mathcal{M}} = 0. \tag{6}$$

Equation (6) is not yet a closed PDE for $c_0$ in manifold coordinates, because it contains mixed and normal second derivatives of $c_0$. This is precisely where the previous steps enter. The transport equation and its normal derivatives express $\partial_r c_0$, $\partial^2_{u,r} c_0$, and $\partial^2_r c_0$ in terms of tangential derivatives. Substituting these relations into Equation (6) eliminates the normal directions and leaves a genuine elliptic equation on $\mathcal{M}$.

**Theorem 3.2** (Elliptic PDE for the manifold prefactor)**.** *Assume Assumptions 2.1 to 2.3. Then there exists a second-order elliptic operator $\mathcal{L}_{\text{eff}}$ on $\mathcal{M}$ such that*

$$\mathcal{L}_{\text{eff}} c_0 = 0 \qquad \text{on } \mathcal{M}.$$

*In local coordinates $u$ on $\mathcal{M}$, this operator has the form*

$$\mathcal{L}_{\text{eff}} c_0 = \operatorname{Tr}\big(\mathbf{A}(u) g(u)^{-1} \partial_u^2 c_0\big) + \beta(u)^{\intercal} \partial_u c_0 + \gamma(u) c_0,$$

*where the coefficients are given in the appendix.*

Theorem 3.2 is the main structural reduction of the paper. It turns the open transport problem for $c_0$ into a closed equation posed directly on the manifold of minimizers. In this way, the three-factor formula from Theorem 3.1 becomes effective: the geometric term is explicit, the normal-curvature term is characterized through the Riccati relation, and the remaining tangential redistribution is determined by the elliptic equation for $c_0$.

In the standard approach, one first solves the stationary Fokker–Planck equation for $p^\varepsilon$ and only then lets $\varepsilon \to 0$ to recover the limiting distribution. For non-reversible dynamics, this is known to be difficult because it requires solving a stationary PDE in the full ambient dimension $d$. In our framework, we bypass that step. Rather than finding $p^\varepsilon$ first, we derive an equation directly for the limiting object that controls minima selection. The resulting PDE for $c_0$ is posed on the manifold $\mathcal{M}$, whose dimension is $n = \dim(\mathcal{M}) < d$. In this sense, the problem is effectively reduced from the ambient space to the lower-dimensional space where selection actually takes place.

### 3.5 Langevin Dynamics as an Example

The Langevin case provides the cleanest benchmark for the general theory. The limiting measure for isotropic noise is classical (Hwang 1980; Wojtowytsch 2021). Here we show that the same conclusion follows directly from our general framework, without first using the explicit stationary density. When the diffusion is isotropic, $D(x) \equiv I$, the quasi-potential coincides with the objective:

$$V(x) = f(x),$$

as proved in Lemma A.1. The transport equation also simplifies to

$$\langle \partial_x c_0, \partial_x f \rangle = 0,$$

so $c_0$ is constant along the reverse-gradient characteristics. Equivalently, if $\bar{x} \in \mathcal{M}$ is the limit of the backward characteristic through $x$, then

$$c_0(x) = c_0(\bar{x}).$$

In this sense, the transport equation does not create any additional off-manifold reweighting beyond the value of $c_0$ already prescribed on $\mathcal{M}$.

The manifold equation from Theorem 3.2 then simplifies to

$$\Delta_{\mathcal{M}} c_0 - \frac{1}{2} \big\langle \nabla_{\mathcal{M}} c_0, \nabla_{\mathcal{M}} \log \det\big(\partial_r^2 f\big) \big\rangle_g = 0.$$

This is the weighted Laplace equation

$$\Delta_{\mathcal{M}}^{\nu} c_0 = 0,$$

associated with the measure

$$\nu(du) \propto \det\big(\partial_r^2 f(u,0)\big)^{-1/2} \, d\mathcal{M}(u).$$

Because $\mathcal{M}$ is connected and compact, the only stationary solutions are constants. Therefore the prefactor does not introduce any additional redistribution along the manifold, and the limiting measure reduces to

$$d\mu_0(u) \propto \det\big(\partial_r^2 f(u,0)\big)^{-1/2} \, d\mathcal{M}(u).$$

This recovers the inverse-Hessian weighting for Langevin dynamics. In particular, isotropic noise favors minima that are flat in the normal directions, in the sense of having a smaller determinant of the normal Hessian, and also favors regions of the minima manifold that have larger intrinsic volume. The key contrast with the general case is that the prefactor is now comparatively simple: minima selection is fully described by the intrinsic geometry of $\mathcal{M}$ together with the normal Hessian of the loss. For anisotropic or state-dependent noise, by contrast, the elliptic equation for $c_0$ is genuinely nontrivial and contributes an additional selection mechanism.

## 4 Discussion: A Time-Scale Picture

The previous section identifies the limiting stationary measure on $\mathcal{M}$ and reduces the tangential selection mechanism to the elliptic equation $\mathcal{L}_{\text{eff}} c_0 = 0$. Although our results are stationary, they suggest a heuristic dynamical picture for how the asymptotic bias may emerge.

**The $O(1)$ scale.** On the fast time scale, the deterministic drift $-\nabla f$ dominates. The process is therefore expected to relax toward a neighborhood of the minima manifold $\mathcal{M}$. At this stage, the main effect is to identify the stable set; the finer weighting along $\mathcal{M}$ has not yet appeared.

**The $O(1/\varepsilon)$ scale.** On the next scale, the factor $\exp(-V/\varepsilon)$ should already govern the distribution. In this regime, the mass is sharply concentrated in the normal directions, so the large-deviation profile is essentially in place. What may still evolve is the sub-exponential weight, namely the redistribution of mass along $\mathcal{M}$ represented in our stationary formula by the prefactor $c_0$.

**The $O(1/\varepsilon^2)$ scale.** On a slower tangential scale, it is natural to associate the manifold operator $\mathcal{L}_{\text{eff}}$ with the equilibration of that prefactor. From this viewpoint, the stationary equation $\mathcal{L}_{\text{eff}} c_0 = 0$ can be read as the equilibrium condition for an effective dynamics on $\mathcal{M}$.

In short, the picture is: fast relaxation toward $\mathcal{M}$, formation of the normal quasi-potential profile, and slow equilibration of mass along the manifold.

# 5 Conclusion

We studied minima selection for small-stepsize diffusions when the set of global minimizers is a compact manifold $\mathcal{M}$. Under a local manifold WKB ansatz, the limiting stationary measure on $\mathcal{M}$ takes the form

$$d\mu_0(u) \propto c_0(u) \det\big(\partial_r^2 V(u,0)\big)^{-1/2} d\mathcal{M}(u).$$

This formula separates three contributions to minima selection: the intrinsic geometry of $\mathcal{M}$, the normal curvature of the quasi-potential, and a prefactor that describes the remaining redistribution of mass along the manifold. In particular, the limiting bias is not determined by sharpness alone: geometry and the manifold prefactor are genuine additional effects. The main technical step is a reduction of the stationary Fokker–Planck equation to the minimizer manifold. After substituting the WKB ansatz, we obtain an exact identity whose leading orders yield the Hamilton–Jacobi equation for $V$, a transport equation for the prefactor, and finally a closed elliptic PDE on $\mathcal{M}$ for $c_0$. In this way, the original stationary problem in the ambient space is reduced to a lower-dimensional equation on the space where selection actually takes place.

# References


Robbins, Herbert and Sutton Monro (1951). "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407 (cit. on p. 1).

Hwang, Chii-Ruey (1980). "Laplace's method revisited: weak convergence of probability measures". In: *The Annals of Probability*, pp. 1177–1182 (cit. on pp. 8, 16).

Sheu, Shuenn-Jyi (1986). "Asymptotic behavior of the invariant density of a diffusion Markov process with small diffusion". In: *SIAM journal on mathematical analysis* 17.2, pp. 451–460 (cit. on pp. 2, 5).

Mikami, Toshio (1990). "Asymptotic analysis of invariant density of randomly perturbed dynamical systems". In: *The Annals of Probability*, pp. 524–536 (cit. on pp. 2, 5).

Freidlin, Mark I. and Alexander D. Wentzell (2012). *Random Perturbations of Dynamical Systems*. Vol. 260. Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg (cit. on p. 2).

Mandt, Stephan, Matthew D Hoffman, and David M Blei (2015). "Continuous-time limit of stochastic gradient descent revisited". In: *OPT workshop, NIPS* (cit. on p. 1).

Bouchet, Freddy and Julien Reygner (2016). "Generalisation of the Eyring–Kramers transition rate formula to irreversible diffusion processes". In: *Annales Henri Poincaré*. Vol. 17. 12. Springer, pp. 3499–3532 (cit. on p. 7).

Keskar, Nitish Shirish, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang (2016). "On large-batch training for deep learning: Generalization gap and sharp minima". In: *arXiv preprint arXiv:1609.04836* (cit. on p. 1).

Mandt, Stephan, Matthew Hoffman, and David Blei (2016). "A variational analysis of stochastic gradient algorithms". In: *International Conference on Machine Learning* (cit. on p. 1).

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2017). "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations* (cit. on p. 1).

Laurent, Thomas and James Brecht (2018). "Deep linear networks with arbitrary loss: All local minima are global". In: *International conference on machine learning*. PMLR, pp. 2902–2907 (cit. on p. 3).

Luo, Jiajun, Chenwei Wu, and Jason D Lee (2018). "No spurious local minima in a two hidden unit Relu network". In: *International Conference on Learning Representations* (cit. on p. 3).

Wu, Lei, Chao Ma, et al. (2018). "How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective". In: *Advances in Neural Information Processing Systems* 31 (cit. on p. 3).

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song (2019). "A convergence theory for deep learning via over-parameterization". In: *International conference on machine learning*. PMLR, pp. 242–252 (cit. on p. 1).

Du, Simon, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai (2019). "Gradient descent finds global minima of deep neural networks". In: *International conference on machine learning*. PMLR, pp. 1675–1685 (cit. on p. 1).

Hu, Wenqing, Chris Junchi Li, Lei Li, and Jian-Guo Liu (2019). "On the diffusion approximation of nonconvex stochastic gradient descent". In: *Annals of Mathematical Sciences and Applications* 4.1, pp. 3–32 (cit. on pp. 3, 5).

Nguyen, Quynh (2019). "On connected sublevel sets in deep learning". In: *International conference on machine learning*. PMLR, pp. 4790–4799 (cit. on p. 3).

Xie, Zeke, Issei Sato, and Masashi Sugiyama (2020). "A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima". In: *arXiv preprint arXiv:2002.03495* (cit. on p. 3).

Cohen, Jeremy, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar (2021). "Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability". In: *International Conference on Learning Representations* (cit. on p. 3).

Damian, Alex, Tengyu Ma, and Jason D Lee (2021). "Label noise sgd provably prefers flat global minimizers". In: *Advances in Neural Information Processing Systems* 34, pp. 27449–27461 (cit. on p. 3).

Wojtowytsch, Stephan (2021). "Stochastic gradient descent with noise of machine learning type. Part II: Continuous time analysis". In: (cit. on p. 8).

Li, Zhiyuan, Tianhao Wang, and Sanjeev Arora (2022). "What Happens after SGD Reaches Zero Loss?–A Mathematical Framework". In: *International Conference on Learning Representations* (cit. on p. 3).

Mori, Takashi, Liu Ziyin, Kangqiao Liu, and Masahito Ueda (2022). "Power-law escape rate of SGD". In: *International Conference on Machine Learning*. PMLR, pp. 15959–15975 (cit. on p. 3).

Wu, Lei, Mingze Wang, and Weijie Su (2022). "The alignment property of SGD noise and how it helps select flat minima: A stability analysis". In: *Advances in Neural Information Processing Systems* 35, pp. 4680–4693 (cit. on p. 3).

Bajovic, Dragana, Dusan Jakovetic, and Soummya Kar (2023). "Large deviations rates for stochastic gradient descent with strongly convex functions". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 10095–10111 (cit. on p. 2).

Ibayashi, Hikaru and Masaaki Imaizumi (2023). "Why does SGD prefer flat minima?: Through the lens of dynamical systems". In: *When Machine Learning meets Dynamical Systems: Theory and Applications* (cit. on p. 3).

Azizian, Waïss, Franck Iutzeler, Jerome Malick, and Panayotis Mertikopoulos (2024). "What is the Long-Run Distribution of Stochastic Gradient Descent? A Large Deviations Analysis". In: *International Conference on Machine Learning*. PMLR, pp. 2168–2229 (cit. on pp. 2, 5).

Mou, Chenchen, Weiwei Qi, Zhongwei Shen, and Yingfei Yi (2025). "Quasi-potential of stationary and quasi-stationary densities: existence, regularity, and applications". In: *arXiv preprint arXiv:2506.17546* (cit. on pp. 4, 11, 13, 14).

Rebjock, Quentin and Nicolas Boumal (2025). "Fast convergence to non-isolated minima: four equivalent conditions for C 2 functions: Q. Rebjock, N. Boumal". In: *Mathematical Programming* 213.1, pp. 151–199 (cit. on p. 3).

Zhang, Liang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, Michael Muehlebach, and Niao He (2025). "Zeroth-Order Optimization Finds Flat Minima". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (cit. on p. 3).

# A Auxiliary Results on the Quasi-Potential and Invariant Density

This appendix collects the auxiliary facts on the quasi-potential and the invariant density used in the main text.

## A.1 Basic Properties of the Quasi-Potential

**Corollary A.1** (Hamilton–Jacobi equation)**.** *Under Assumptions 2.1 and 2.2, the quasi-potential satisfies*

$$\langle D(x)\nabla V(x), \nabla V(x)\rangle - \langle \nabla f(x), \nabla V(x)\rangle = 0.$$

*Proof.* This is a direct consequence of Mou et al. (2025, Theorem A and Lemma 1.1), which identify the limit of the logarithmic transform of the stationary density with the quasi-potential and show that this limit is a viscosity solution of the corresponding first-order Hamilton–Jacobi equation. □

**Lemma A.1** (Identity-diffusion benchmark)**.** *If $D \equiv I$, then the quasi-potential is*

$$V(x) = f(x).$$

*Proof.* In this case the quasi-potential is

$$V(x) = \inf_{\substack{\phi, T_1, T_2 \\ \phi(T_1)\in\mathcal{M},\ \phi(T_2)=x}} \frac{1}{4}\int_{T_1}^{T_2} \|\dot\phi_t + \nabla f(\phi_t)\|^2\, dt.$$

For any admissible path $\phi$, expand the square:

$$\frac{1}{4}\int_{T_1}^{T_2} \|\dot\phi_t - \nabla f(\phi_t)\|^2\, dt + \int_{T_1}^{T_2} \langle \dot\phi_t, \nabla f(\phi_t)\rangle\, dt.$$

By the chain rule,

$$\int_{T_1}^{T_2} \langle \dot{\phi}_t, \nabla f(\phi_t) \rangle \, dt = f(\phi(T_2)) - f(\phi(T_1)) = f(x),$$

because $f = 0$ on $\mathcal{M}$ by Assumption 2.1. Hence every admissible path has action at least $f(x)$, and therefore

$$V(x) \geq f(x).$$

For the reverse inequality, let $y_x(t)$ solve the gradient flow

$$\dot{y}_x(t) = -\nabla f(y_x(t)), \qquad y_x(0) = x.$$

Since $f$ is coercive and has no critical points outside $\mathcal{M}$, $y_x(t)$ converges to some point of $\mathcal{M}$ as $t \to \infty$. For each $T > 0$, let $\pi(y_x(T)) \in \mathcal{M}$ be a nearest-point projection, and define a path $\phi_T : [-1, T] \to \mathbb{R}^d$ by

$$\dot{\phi}_T(t) = \begin{cases} y_x(T) - \pi(y_x(T)), & -1 \leq t \leq 0, \\ \nabla f(\phi_T(t)), & 0 \leq t \leq T, \end{cases}$$

with initial condition $\phi_T(-1) = \pi(y_x(T))$. Then

$$\phi_T(T) = x, \qquad \phi_T(-1) \in \mathcal{M},$$

so $\phi_T$ is admissible for the definition of $V(x)$.

We estimate the action of the two pieces separately. On $[0, T]$ we have

$$\begin{aligned} \frac{1}{4} \int_0^T \|\dot{\phi}_T(t) + \nabla f(\phi_T(t))\|^2 \, dt &= \int_0^T \|\nabla f(\phi_T(t))\|^2 \, dt \\ &= \int_0^T \frac{d}{dt} f(\phi_T(t)) \, dt \\ &= f(x) - f(y_x(T)). \end{aligned}$$

On the initial segment $[-1, 0]$, the path length is $\|y_x(T) - \pi(y_x(T))\|$, which tends to zero as $T \to \infty$. Since $\nabla f$ vanishes on $\mathcal{M}$ and is $C^1$, it follows that

$$\sup_{-1 \leq t \leq 0} \|\nabla f(\phi_T(t))\| \to 0,$$

and therefore the action of this segment also tends to zero. Consequently, the total action of $\phi_T$ converges to $f(x)$ as $T \to \infty$, because $f(y_x(T)) \to 0$ on $\mathcal{M}$. Hence

$$V(x) \leq f(x).$$

Combining the two inequalities yields $V(x) = f(x)$. □

**Lemma A.2.** *Under Assumptions 2.1 and 2.2, the quasi-potential satisfies*

$$\frac{1}{\lambda_{\max}} f(x) \leq V(x) \leq \frac{1}{\lambda_{\min}} f(x).$$

*Proof.* Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the uniform ellipticity constants from Assumption 2.2. By uniform ellipticity,

$$\frac{1}{\lambda_{\max}} \|v\|^2 \leq \langle D(x)^{-1} v, v \rangle \leq \frac{1}{\lambda_{\min}} \|v\|^2.$$

Let

$$I_D(\phi) := \frac{1}{4}\int_{T_1}^{T_2} \langle D(\phi_t)^{-1}\big(\dot\phi_t + \nabla f(\phi_t)\big), \dot\phi_t + \nabla f(\phi_t)\rangle\, dt$$

be the action defining $V$, and let

$$I_I(\phi) := \frac{1}{4}\int_{T_1}^{T_2} \|\dot\phi_t + \nabla f(\phi_t)\|^2\, dt$$

be the corresponding action in the identity-diffusion case. For every absolutely continuous path $\phi$ we therefore have

$$\frac{1}{\lambda_{\max}} I_I(\phi) \le I_D(\phi) \le \frac{1}{\lambda_{\min}} I_I(\phi).$$

Taking the infimum over all admissible paths from $\mathcal{M}$ to $x$ gives

$$\frac{1}{\lambda_{\max}} V_I(x) \le V(x) \le \frac{1}{\lambda_{\min}} V_I(x),$$

where $V_I$ denotes the quasi-potential for $D \equiv I$. By Lemma A.1, $V_I(x) = f(x)$, and the claim follows. □

**Lemma A.3.** *For every $x \in \mathcal{M}$, the Hessians $\nabla^2 V(x)$ and $\nabla^2 f(x)$ have the same nullspace.*

*Proof.* By Lemma A.2, there exist constants $c_1, c_2 > 0$ such that

$$c_1 f(y) \le V(y) \le c_2 f(y) \qquad \text{for } y \text{ near } x.$$

Let $v \in \mathbb{R}^d$ and expand both functions at $x \in \mathcal{M}$ along the line $x + \delta v$:

$$V(x+\delta v) = \frac{\delta^2}{2} v^\intercal \nabla^2 V(x) v + o(\delta^2),$$
$$f(x+\delta v) = \frac{\delta^2}{2} v^\intercal \nabla^2 f(x) v + o(\delta^2).$$

If $v^\intercal \nabla^2 f(x) v = 0$, then the upper bound $V \le c_2 f$ implies $v^\intercal \nabla^2 V(x) v = 0$. If instead $v^\intercal \nabla^2 f(x) v > 0$, then the lower bound $V \ge c_1 f$ implies $v^\intercal \nabla^2 V(x) v > 0$. Hence the two Hessians have the same nullspace. □

## A.2 Density LDP via Mou et al. (2025)

**Proposition A.1.** *Under Assumption 2.1, the manifold $\mathcal{M}$ is the maximal attractor of the deterministic flow*

$$\dot x = -\nabla f(x)$$

*on $\mathbb{R}^d$, and it is an equivalence class for the Freidlin–Wentzell action.*

*Proof.* Since $f$ is coercive and

$$\{x \in \mathbb{R}^d : \nabla f(x) = 0\} = \mathcal{M},$$

the function $f$ is a strict Lyapunov function for the deterministic flow away from $\mathcal{M}$:

$$\frac{d}{dt} f(x_t) = -\|\nabla f(x_t)\|^2 < 0 \qquad \text{for } x_t \notin \mathcal{M}.$$

Therefore every forward orbit is precompact and approaches $\mathcal{M}$, so $\mathcal{M}$ is the maximal attractor.

To show that $\mathcal{M}$ is an equivalence class, fix $x, y \in \mathcal{M}$. Because $\mathcal{M}$ is connected and compact, there exists a piecewise $C^1$ path $\phi : [0,1] \to \mathcal{M}$ from $x$ to $y$. Along $\mathcal{M}$ one has $\nabla f = 0$, so if we traverse this path in time $T > 0$, its Freidlin–Wentzell action is bounded by $C/T$ for some constant $C$. Letting $T \to \infty$ shows that the quasi-potential from $x$ to $y$ is zero, and similarly from $y$ to $x$. Hence $\mathcal{M}$ is an equivalence class. □

**Theorem A.1** (Proof of Theorem 2.1 via Mou et al. (2025))**.** *Assume Assumptions 2.1 and 2.2. Then, for every $\alpha \in (0,1)$,*

$$\varepsilon \log p^\varepsilon = -V \qquad \text{in } C^\alpha(\mathbb{R}^d).$$

*In particular, for every $x \in \mathbb{R}^d$,*

$$\lim_{\varepsilon \to 0} \varepsilon \log p^\varepsilon(x) = -V(x).$$

*Proof.* We apply Mou et al. (2025, Theorem A) with

$$\Omega = \mathbb{R}^d, \qquad b = -\nabla f, \qquad A = 2D.$$

Their assumptions (H1)–(H3) hold in our setting:

1. By Proposition A.1, $\mathcal{M}$ is the maximal attractor of the deterministic flow.
2. By Assumption 2.2, the stationary densities exist and the associated invariant measures are tight.
3. By Proposition A.1, $\mathcal{M}$ is an equivalence class.

Their conclusion is stated on $\Omega_V = \{x \in \Omega : V(x) < \rho_V\}$, where for unbounded domains the convention at the beginning of their Subsection 2.2 interprets $x \to \partial\Omega$ as $\|x\| \to \infty$. In our case, Lemma A.2 and coercivity of $f$ imply

$$V(x) \to \infty \qquad \text{as } \|x\| \to \infty,$$

hence $\rho_V = \infty$ and therefore $\Omega_V = \mathbb{R}^d$.

Finally, we match scalings. Their small-noise operator is

$$\frac{\delta^2}{2} \nabla \cdot (\nabla \cdot (Au)) - \nabla \cdot (bu),$$

while our stationary Fokker–Planck operator is

$$\varepsilon \nabla \cdot (\nabla \cdot (Dp^\varepsilon)) - \nabla \cdot ((-\nabla f) p^\varepsilon).$$

With $A = 2D$ and $\delta^2/2 = \varepsilon$, the two coincide. Therefore their conclusion

$$\frac{\delta^2}{2} \log u_\delta = -V$$

becomes precisely

$$\varepsilon \log p^\varepsilon = -V.$$

□

# B Proofs for Section 3

This appendix merges the former coordinate appendix with proofs of the results from Section 3.

## B.1 Tangent-Normal Coordinates Near the Minima Manifold

Let $m(u)$ be a local parametrization of $\mathcal{M}$, and let the columns of $N(u)$ form an orthonormal basis of the normal space $\mathcal{N}_{m(u)}\mathcal{M}$. In a tubular neighborhood of $\mathcal{M}$ we write

$$x = x(u, r) \coloneqq m(u) + N(u)r,$$

where $u \in \mathbb{R}^n$ is tangential and $r \in \mathbb{R}^{d-n}$ is normal. Denote by $J(u, r)$ the Jacobian matrix of this coordinate map and by

$$G(u, r) \coloneqq J(u, r)^\intercal J(u, r)$$

the induced metric tensor in $(u, r)$ coordinates.

The following elementary facts are used repeatedly.

1. The ambient volume element becomes

$$dx = \det J(u, r)\, du\, dr.$$

2. At $r = 0$, the tangent and normal blocks decouple:

$$G(u, 0) = \begin{bmatrix} g(u) & 0 \\ 0 & I \end{bmatrix}, \qquad G(u, 0)^{-1} = \begin{bmatrix} g(u)^{-1} & 0 \\ 0 & I \end{bmatrix},$$

   where $g(u)$ is the intrinsic metric tensor on $\mathcal{M}$.

3. Since $f$ and $V$ both vanish on $\mathcal{M}$, their tangential derivatives vanish there:

$$\partial_u f(u, 0) = \partial_u V(u, 0) = 0, \qquad \partial^2_{u,r} f(u, 0) = \partial^2_{u,r} V(u, 0) = 0,$$

   and their only nontrivial Hessian blocks on $\mathcal{M}$ are the normal ones.

## B.2 Proof of Theorem 3.1

*Proof of Theorem 3.1.* Let

$$p^\varepsilon(x) = Z_\varepsilon^{-1} c^\varepsilon(x) \exp\left(-\frac{V(x)}{\varepsilon}\right) \qquad \text{on } U,$$

where $Z_\varepsilon$ is the normalization constant. Fix a bounded continuous test function $\varphi$.

By Theorem 2.1, the stationary density satisfies

$$\varepsilon \log p^\varepsilon(x) \to -V(x) \qquad \text{for every } x \in \mathbb{R}^d.$$

Since $V$ vanishes only on $\mathcal{M}$ and $\overline{U}$ is a compact tubular neighborhood of $\mathcal{M}$, there exists $\delta_U > 0$ such that

$$V(x) \geq \delta_U \qquad \text{for all } x \in \mathbb{R}^d \setminus U.$$

Hence

$$\int_{\mathbb{R}^d \setminus U} p^\varepsilon(x)\,dx \to 0,$$

so it suffices to analyze the contribution from $U$.

Write

$$N_\varepsilon(\varphi) \coloneqq \int_U \varphi(x) c^\varepsilon(x) \exp\left(-\frac{V(x)}{\varepsilon}\right) dx,$$
$$\widetilde{N}_\varepsilon(\varphi) \coloneqq \int_U \varphi(x) c_0(x) \exp\left(-\frac{V(x)}{\varepsilon}\right) dx.$$

Since $c^\varepsilon \to c_0$ uniformly on the compact set $U$ and $c_0 > 0$ on $U$, there exists $m_0 > 0$ such that $c_0 \geq m_0$ on $U$. Therefore

$$|N_\varepsilon(\varphi) - \widetilde{N}_\varepsilon(\varphi)| \leq \|\varphi\|_{L^\infty(U)} \|c^\varepsilon - c_0\|_{L^\infty(U)} \int_U \exp\left(-\frac{V(x)}{\varepsilon}\right) dx$$
$$\leq \frac{\|\varphi\|_{L^\infty(U)}}{m_0} \|c^\varepsilon - c_0\|_{L^\infty(U)}\, \widetilde{N}_\varepsilon(1).$$

Hence

$$\frac{N_\varepsilon(\varphi)}{N_\varepsilon(1)} - \frac{\widetilde{N}_\varepsilon(\varphi)}{\widetilde{N}_\varepsilon(1)} \to 0.$$

So it is enough to compute the weak limit using the fixed weight $c_0(x)\,dx$.

We now apply the argument of Hwang (1980, Theorem 3.1) directly. Cover the compact manifold $\mathcal{M}$ by a finite tubular atlas and choose a subordinate partition of unity. It is enough to work in one chart, with coordinates $x = x(u, r)$ from the previous subsection. In these coordinates define

$$H(u, r) \coloneqq V(u, r),$$
$$f(u, r) \coloneqq c_0(u, r) \det J(u, r).$$

Then:

1. $\mathcal{M}$ is compact and $H(u, 0) = 0$;
2. $r = 0$ is the unique minimizer of $r \mapsto H(u, r)$ for each $u$;
3. by continuity of $V$ and compactness, after shrinking the tube if necessary there exists $\hat{\varepsilon} > 0$ such that
$$\inf_{u,\ \hat{\varepsilon} \leq \|r\| \leq \varepsilon_0} H(u, r) > 0;$$
4. by Lemma A.3 and the normal nondegeneracy of $\nabla^2 f$ from Assumption 2.1, the normal Hessian $\partial_r^2 H(u, 0) = \partial_r^2 V(u, 0)$ is positive definite, and by continuity its smallest eigenvalue is bounded uniformly away from zero on the tube;
5. $f(u, r) = c_0(u, r) \det J(u, r)$ is $C^1$ and bounded on the compact coordinate chart.

These are exactly the hypotheses used in Hwang (1980, Theorem 3.1), specialized to a single compact manifold of minimizers. Therefore

$$\frac{\widetilde{N}_\varepsilon(\varphi)}{\widetilde{N}_\varepsilon(1)} \to \frac{\int_{\mathcal{M}} \varphi(u) c_0(u) \det\left(\partial_r^2 V(u, 0)\right)^{-1/2} d\mathcal{M}(u)}{\int_{\mathcal{M}} c_0(u) \det(\partial_r^2 V(u, 0))^{-1/2}\, d\mathcal{M}(u)}.$$

Combining this with the reduction from $N_\varepsilon$ to $\widetilde{N}_\varepsilon$ and the negligible mass outside $U$ yields

$$\int_{\mathbb{R}^d} \varphi(x)\, d\mu^\varepsilon(x) \to \frac{\int_{\mathcal{M}} \varphi(u) c_0(u) \det\big(\partial_r^2 V(u,0)\big)^{-1/2}\, d\mathcal{M}(u)}{\int_{\mathcal{M}} c_0(u) \det(\partial_r^2 V(u,0))^{-1/2}\, d\mathcal{M}(u)}.$$

Since this holds for every bounded continuous $\varphi$, the measures $\mu^\varepsilon$ converge weakly to a measure $\mu_0$ supported on $\mathcal{M}$ with density

$$d\mu_0(u) \propto c_0(u) \det\big(\partial_r^2 V(u,0)\big)^{-1/2}\, d\mathcal{M}(u).$$

□

## B.3 An Exact Identity After WKB Substitution

**Lemma B.1** (Exact identity for the WKB ansatz)**.** *Suppose $p^\varepsilon$ is stationary and satisfies Assumption 2.3 on $U$. Define*

$$\begin{aligned}
\mathcal{H}[V] &:= -\langle \partial_x V, \partial_x f\rangle + \langle \partial_x V, D\partial_x V\rangle,\\
\mathcal{L}_1 c &:= \langle \partial_x c, -\partial_x f + 2D\partial_x V\rangle + c\Big(\mathrm{Tr}(-\partial_x^2 f + D\partial_x^2 V) + 2\langle \mathrm{div}_x\, D, \partial_x V\rangle\Big),\\
\mathcal{L}_2 c &:= D : \partial_x^2 c + \sum_{i,j} c\, \partial^2_{x_i,x_j}[D]_{i,j} + 2\langle \partial_x c, \mathrm{div}_x\, D\rangle.
\end{aligned}$$

*Then*

$$\frac{1}{\varepsilon} c^\varepsilon \mathcal{H}[V] - \mathcal{L}_1 c^\varepsilon + \varepsilon \mathcal{L}_2 c^\varepsilon = 0 \qquad \text{on } U. \tag{7}$$

*Proof.* Substitute

$$p^\varepsilon = c^\varepsilon \exp\left(-\frac{V}{\varepsilon}\right)$$

into the stationary Fokker–Planck equation

$$\mathrm{div}_x\big((\nabla f) p^\varepsilon\big) + \varepsilon \sum_{i,j} \partial^2_{x_i,x_j}\big([D]_{i,j} p^\varepsilon\big) = 0.$$

Expanding derivatives and collecting the contributions of order $1/\varepsilon$, 1, and $\varepsilon$ gives exactly Equation (7). □

## B.4 Proof of Lemma 3.1

*Proof of Lemma 3.1.* By Assumption 2.3, $c^\varepsilon \to c_0$ in $C^2(U)$ and $c_0 > 0$ on the compact set $U$. Hence $c^\varepsilon$ is uniformly bounded in $C^2(U)$ and bounded away from zero for sufficiently small $\varepsilon$. Since the coefficients of $\mathcal{L}_1$ and $\mathcal{L}_2$ are smooth on $U$, we have

$$\varepsilon \mathcal{L}_1 c^\varepsilon \to 0, \qquad \varepsilon^2 \mathcal{L}_2 c^\varepsilon \to 0 \qquad \text{uniformly on } U.$$

Multiplying Equation (7) by $\varepsilon$ and letting $\varepsilon \to 0$ therefore yields

$$\mathcal{H}[V] = 0 \qquad \text{on } U,$$

which is exactly Equation (4).

For the Riccati relation, define

$$H(x) \coloneqq -\langle \partial_x V(x), \partial_x f(x) \rangle + \langle \partial_x V(x), D(x) \partial_x V(x) \rangle.$$

Since Equation (4) states that $H \equiv 0$ on $U$, its Hessian vanishes at every point of $\mathcal{M}$.

Fix $x = m(u) \in \mathcal{M}$ and let $\xi, \eta \in \mathcal{N}_x\mathcal{M}$ be normal vectors. Because $\partial_x f(x) = \partial_x V(x) = 0$, differentiating $H$ twice in the directions $\xi, \eta$ gives

$$\begin{aligned} 0 &= \nabla^2 H(x)[\xi, \eta] \\ &= -\big\langle \partial_x^2 V(x)\xi, \partial_x^2 f(x)\eta \big\rangle - \big\langle \partial_x^2 V(x)\eta, \partial_x^2 f(x)\xi \big\rangle + 2\big\langle \partial_x^2 V(x)\xi, D(x)\partial_x^2 V(x)\eta \big\rangle. \end{aligned}$$

Here all terms involving derivatives of $D$ disappear because they are multiplied by $\partial_x V(x) = 0$. Moreover, since $f$ and $V$ vanish on $\mathcal{M}$, their tangential and mixed Hessian blocks vanish on $\mathcal{M}$, so for normal vectors only the normal-normal blocks remain.

Write $\xi = N(u)a$ and $\eta = N(u)b$ with $a, b \in \mathbb{R}^{d-n}$. Then the previous identity becomes

$$\begin{aligned} 0 = &-a^\intercal \partial_r^2 V(u,0)\, \partial_r^2 f(u,0)\, b - a^\intercal \partial_r^2 f(u,0)\, \partial_r^2 V(u,0)\, b \\ &+ 2a^\intercal \partial_r^2 V(u,0)\, N(u)^\intercal D(u,0) N(u)\, \partial_r^2 V(u,0)\, b. \end{aligned}$$

Since this holds for all $a, b$, we obtain

$$\partial_r^2 V\, \partial_r^2 f + \partial_r^2 f\, \partial_r^2 V = 2\partial_r^2 V\, N^\intercal D N\, \partial_r^2 V \qquad \text{on } \mathcal{M}.$$

□

## B.5 Proof of Lemma 3.2

*Proof of Lemma 3.2.* Since Lemma 3.1 gives Equation (4), the $1/\varepsilon$ term in Equation (7) vanishes and we obtain

$$\mathcal{L}_1 c^\varepsilon = \varepsilon \mathcal{L}_2 c^\varepsilon.$$

The convergence $c^\varepsilon \to c_0$ in $C^2(U)$ implies

$$\mathcal{L}_1 c^\varepsilon \to \mathcal{L}_1 c_0, \qquad \varepsilon \mathcal{L}_2 c^\varepsilon \to 0 \qquad \text{uniformly on } U,$$

so passing to the limit gives Equation (5).

Define

$$\begin{aligned} \mathcal{A}(x) &\coloneqq -\partial_x f(x) + 2D(x)\partial_x V(x), \\ \mathcal{B}(x) &\coloneqq \operatorname{Tr}(-\partial_x^2 f + D\partial_x^2 V)(x) + 2\langle \operatorname{div}_x D(x), \partial_x V(x) \rangle. \end{aligned}$$

Then Equation (5) is

$$\langle \partial_x c_0, \mathcal{A} \rangle + \mathcal{B} c_0 = 0.$$

Along a characteristic $\psi$ satisfying

$$\dot{\psi}(s) = \mathcal{A}(\psi(s)),$$

the chain rule gives

$$\frac{d}{ds} c_0(\psi(s)) = -\mathcal{B}(\psi(s))\, c_0(\psi(s)),$$

hence, since $c_0 > 0$,

$$\frac{d}{ds} \log c_0(\psi(s)) = -\mathcal{B}(\psi(s)).$$

Integrating between $s_0$ and $s_1$ yields

$$c_0(\psi(s_1)) = c_0(\psi(s_0)) \exp\left( - \int_{s_0}^{s_1} \mathcal{B}(\psi(s))\, ds \right).$$

It remains to justify the manifold-anchored representation. On $\mathcal{M}$, $\mathcal{A} = 0$. In the normal directions, let

$$A \coloneqq \partial_r^2 f, \qquad S \coloneqq \partial_r^2 V, \qquad B \coloneqq N^\intercal D N.$$

By Lemma A.3, $\nabla^2 V$ and $\nabla^2 f$ have the same nullspace on $\mathcal{M}$. Since $\nabla^2 f$ is nondegenerate in the normal directions by Assumption 2.1, the normal block $S = \partial_r^2 V$ is invertible on $\mathcal{N}_x\mathcal{M}$. The normal linearization of $\mathcal{A}$ is

$$-A + 2BS.$$

Using the Riccati relation $SA + AS = 2SBS$, we obtain

$$-A + 2BS = S^{-1}AS.$$

Since $A$ is positive definite in the normal directions by Assumption 2.1, the matrix $S^{-1}AS$ has positive spectrum. Thus $\mathcal{M}$ is normally repelling for the forward characteristic flow and normally attracting in backward time.

Finally, on $\mathcal{M}$ we have $\partial_x V = 0$, so

$$\mathcal{B}|_{\mathcal{M}} = \mathrm{Tr}(-\partial_x^2 f + D\partial_x^2 V)\big|_{\mathcal{M}}.$$

Taking the trace of the Riccati relation shows that this vanishes on $\mathcal{M}$. Therefore $\mathcal{B}(\psi(s)) \to 0$ exponentially as $s \to -\infty$, and letting $s_0 \to -\infty$ in the finite-time formula gives

$$c_0(x) = c_0(\bar{x}) \exp\left( - \int_{-\infty}^{0} \mathcal{B}(\psi(s))\, ds \right), \qquad \bar{x} = \lim_{s \to -\infty} \psi(s) \in \mathcal{M}.$$

□

## B.6 Proof of Lemma 3.3

*Proof of Lemma 3.3.* By Lemma 3.1, Equation (4) holds on $U$, so Equation (7) reduces to

$$-\mathcal{L}_1 c^\varepsilon + \varepsilon \mathcal{L}_2 c^\varepsilon = 0.$$

Restrict this identity to $\mathcal{M}$. There, the drift part of $\mathcal{L}_1$ vanishes because $\partial_x f = \partial_x V = 0$ on $\mathcal{M}$. Its remaining zeroth-order coefficient is exactly the function $\mathcal{B}$ from the proof of Lemma 3.2,

$$\mathcal{B}(x) = \mathrm{Tr}(-\partial_x^2 f + D\partial_x^2 V)(x) + 2\langle \mathrm{div}_x\, D(x), \partial_x V(x) \rangle.$$

Since $\partial_x V = 0$ on $\mathcal{M}$, this reduces to the trace term already shown to vanish in the proof of Lemma 3.2. Hence

$$\mathcal{L}_1 c^\varepsilon = 0 \qquad \text{on } \mathcal{M}$$

for every $\varepsilon > 0$, and therefore

$$\mathcal{L}_2 c^\varepsilon = 0 \qquad \text{on } \mathcal{M}.$$

Passing to the limit using $c^\varepsilon \to c_0$ in $C^2(U)$ gives Equation (6). □

## B.7 Proof of Theorem 3.2

*Proof of Theorem 3.2.* We work in tangent-normal coordinates $x = x(u, r)$ near $\mathcal{M}$ and set

$$\begin{aligned}\mathcal{A} &\coloneqq -\partial_x f + 2D\partial_x V,\\ \mathcal{B} &\coloneqq \mathrm{Tr}(-\partial_x^2 f + D\partial_x^2 V) + 2\langle \mathrm{div}_x\, D, \partial_x V\rangle.\end{aligned}$$

Then Equation (5) becomes

$$\langle \partial_x c_0, \mathcal{A}\rangle + \mathcal{B}\, c_0 = 0. \tag{8}$$

For a matrix-valued quantity $M$, we use the block notation

$$\begin{aligned}M^{\perp} &\coloneqq N^{\intercal} M N,\\ M^{\perp,\parallel} &\coloneqq N^{\intercal} M\, \partial_u m,\\ M^{\parallel,\perp} &\coloneqq \partial_u m^{\intercal} M N.\end{aligned}$$

We first rewrite ambient derivatives of $c_0$ in local coordinates. Since

$$\partial_x c_0 = J^{-\intercal}\partial_z c_0, \qquad z = (u, r),$$

differentiating once more gives the standard Hessian transformation formula

$$\partial_x^2 c_0 = J^{-\intercal}\left(\partial_z^2 c_0 - \Big(\sum_k \partial_{z_k} c_0\, \Gamma^k_{i,j}\Big)_{i,j}\right) J^{-1}, \tag{9}$$

where $\Gamma^k_{i,j}$ are the Christoffel symbols of the metric $G$ in the coordinates $(u, r)$. At $r = 0$, these split into intrinsic and extrinsic pieces. Define

$$\begin{aligned}(\mathcal{U}_1)_{\alpha,\beta} &\coloneqq \partial_{u_\gamma} c_0\, \Gamma^{\gamma}_{\alpha,\beta} + \partial_{r_p} c_0\, \mathrm{II}^p_{\alpha,\beta},\\ (\mathcal{U}_2)_{\alpha,p} &\coloneqq \omega^q_{\alpha,p}\, \partial_{r_q} c_0 - g^{\gamma,\beta}\, \mathrm{II}^p_{\alpha,\beta}\, \partial_{u_\gamma} c_0,\end{aligned}$$

where II is the second fundamental form and $\omega$ is the normal connection form of the normal frame $N$. Then Equation (9) becomes, at $r = 0$,

$$\begin{aligned}\partial_x^2 c_0 = \mathcal{K} &+ \partial_u m\, g^{-1}\partial_u^2 c_0\, g^{-1}\partial_u m^{\intercal} + \partial_u m\, g^{-1}\partial^2_{u,r} c_0\, N^{\intercal}\\ &+ N\partial^2_{r,u} c_0\, g^{-1}\partial_u m^{\intercal} + N\partial_r^2 c_0\, N^{\intercal},\end{aligned} \tag{10}$$

with

$$\mathcal{K} = -\partial_u m\, g^{-1}\mathcal{U}_1\, g^{-1}\partial_u m^{\intercal} - \partial_u m\, g^{-1}\mathcal{U}_2\, N^{\intercal} - N\mathcal{U}_2^{\intercal} g^{-1}\partial_u m^{\intercal}. \tag{11}$$

We next solve for the derivatives of $c_0$ in the normal directions. Since $\mathcal{A}|_{\mathcal{M}} = 0 = \mathcal{B}|_{\mathcal{M}}$, differentiating Equation (8) and evaluating on $\mathcal{M}$ gives

$$(\partial_x \mathcal{A})^{\intercal}\partial_x c_0 + c_0 \partial_x \mathcal{B} = 0.$$

Using

$$\partial_x c_0 = \partial_u m\, g^{-1}\partial_u c_0 + N\partial_r c_0,$$

and projecting onto the normal space yields

$$\partial_r c_0 = -\big((\partial_x \mathcal{A})^{\perp}\big)^{-\intercal}\left(\big((\partial_x \mathcal{A})^{\intercal}\big)^{\perp,\parallel} g^{-1}\partial_u c_0 + c_0 N^{\intercal}\partial_x \mathcal{B}\right). \tag{12}$$

On $\mathcal{M}$, we have

$$(\partial_x \mathcal{A})^\perp = -\partial_r^2 f + 2N^\intercal D N\, \partial_r^2 V.$$

The inverse of $\partial_r^2 V$ below is well defined by Lemma A.3, exactly as in the proof of Lemma 3.2. By the Riccati identity from Lemma 3.1,

$$(\partial_x \mathcal{A})^\perp = (\partial_r^2 V)^{-1} \partial_r^2 f\, \partial_r^2 V, \tag{13}$$

hence this matrix is invertible.

Differentiate Equation (8) twice and evaluate on $\mathcal{M}$. This gives

$$\partial_x^2 c_0\, \partial_x \mathcal{A} + (\partial_x \mathcal{A})^\intercal \partial_x^2 c_0 + \mathcal{C} = 0, \tag{14}$$

where

$$\mathcal{C} \coloneqq \sum_i (\partial_{x_i} c_0)\, \partial_x^2 \mathcal{A}_i + \partial_x c_0 (\partial_x \mathcal{B})^\intercal + \partial_x \mathcal{B} (\partial_x c_0)^\intercal + c_0 \partial_x^2 \mathcal{B}. \tag{15}$$

Set

$$\mathcal{J} \coloneqq (\partial_x \mathcal{A})^\intercal \mathcal{K}. \tag{16}$$

Projecting Equation (14) with $N^\intercal(\cdot)\partial_u m$ gives

$$\begin{aligned} \partial_{r,u}^2 c_0 = -\big((\partial_x \mathcal{A})^\perp\big)^{-\intercal} \Big( &(\mathcal{J} + \mathcal{J}^\intercal)^{\perp,\parallel} g^{-1} + \big((\partial_x \mathcal{A})^\intercal\big)^{\perp,\parallel} g^{-1} \partial_u^2 c_0\, g^{-1} \\ &+ \mathcal{C}^{\perp,\parallel} g^{-1} \Big). \end{aligned} \tag{17}$$

Projecting instead with $N^\intercal(\cdot)N$ yields the Sylvester equation

$$\big((\partial_x \mathcal{A})^\perp\big)^\intercal \partial_r^2 c_0 + \partial_r^2 c_0\, (\partial_x \mathcal{A})^\perp = -(\mathcal{J} + \mathcal{J}^\intercal + \mathcal{C})^\perp + \mathcal{E} + \mathcal{E}^\intercal, \tag{18}$$

where

$$\mathcal{E} \coloneqq \big((\partial_x \mathcal{A})^\intercal\big)^{\perp,\parallel} \Big( g^{-1} (\mathcal{J} + \mathcal{J}^\intercal)^{\parallel,\perp} + g^{-1} \partial_u^2 c_0\, g^{-1} (\partial_x \mathcal{A})^{\parallel,\perp} + g^{-1} \mathcal{C}^{\parallel,\perp} \Big) \big((\partial_x \mathcal{A})^\perp\big)^{-1}. \tag{19}$$

By Equation (13), the spectrum of $(\partial_x \mathcal{A})^\perp$ is positive, so the Sylvester operator in Equation (18) is invertible. Thus $\partial_{r,u}^2 c_0$ and $\partial_r^2 c_0$ are explicit linear expressions in $c_0$, $\partial_u c_0$, and $\partial_u^2 c_0$.

We now return to the next-order identity Equation (6). Using Equation (10), we obtain on $\mathcal{M}$

$$D : \partial_x^2 c_0 = \operatorname{Tr}\Big( \widehat{D}\, g^{-1} \partial_u^2 c_0 \Big) + 2 \operatorname{Tr}\big( D\, \partial_u m\, g^{-1} \partial_{u,r}^2 c_0\, N^\intercal \big) + \operatorname{Tr}(D^\perp \partial_r^2 c_0) + \operatorname{Tr}(D\mathcal{K}), \tag{20}$$

where

$$\widehat{D}(u) \coloneqq g(u)^{-1} \partial_u m(u)^\intercal D(m(u))\, \partial_u m(u). \tag{21}$$

To remove the trace term involving $\partial_r^2 c_0$, we use the second derivative of the Hamilton–Jacobi equation. Writing $b = -\partial_x f$, that identity reads

$$\partial_x^2 V\, \partial_x b + (\partial_x b)^\intercal \partial_x^2 V + 2 \partial_x^2 V\, D\, \partial_x^2 V = 0 \qquad \text{on } \mathcal{M}.$$

Multiplying by the Moore–Penrose inverse $(\partial_x^2 V)^\dagger$ and restricting to the normal space gives

$$2D^\perp = -\big(\partial_x b (\partial_x^2 V)^\dagger\big)^\perp - \big((\partial_x^2 V)^\dagger \partial_x b^\intercal\big)^\perp. \tag{22}$$

Combining Equation (22) with the Sylvester equation Equation (18) gives

$$2\,\mathrm{Tr}(D^{\perp}\partial_r^2 c_0) = \mathrm{Tr}\Big(\big((\partial_x^2 V)^{\dagger}\big)^{\perp}\big(\mathcal{E}+\mathcal{E}^{\intercal}-(\mathcal{J}+\mathcal{J}^{\intercal}+\mathcal{C})^{\perp}\big)\Big). \tag{23}$$

Substituting Equation (20) and Equation (23) into Equation (6), we arrive at the explicit identity

$$\begin{aligned} 0 = {} & \mathrm{Tr}\Big(\widehat{D}\,g^{-1}\partial_u^2 c_0\Big) + 2\,\mathrm{Tr}\big(D\,\partial_u m\,g^{-1}\partial_{u,r}^2 c_0\,N^{\intercal}\big) + \mathrm{Tr}(D\mathcal{K}) \\ & + \frac{1}{2}\,\mathrm{Tr}\Big(\big((\partial_x^2 V)^{\dagger}\big)^{\perp}\big(\mathcal{E}+\mathcal{E}^{\intercal}-(\mathcal{J}+\mathcal{J}^{\intercal}+\mathcal{C})^{\perp}\big)\Big) \\ & + \sum_{i,j} c_0\,\partial_{x_i,x_j}^2[D]_{i,j} + 2\langle\partial_x c_0, \mathrm{div}_x\,D\rangle \qquad \text{on } \mathcal{M}. \end{aligned} \tag{24}$$

Every term on the right-hand side is now an explicit linear combination of $c_0$, $\partial_u c_0$, and $\partial_u^2 c_0$, because Equation (12), Equation (17), and Equation (18) have already expressed all normal derivatives of $c_0$ in terms of those quantities.

To rewrite the geometric correction term $\mathcal{K}$ intrinsically, split $\mathcal{U}_1 = \mathcal{U}_{1,1} + \mathcal{U}_{1,2}$ into its pure tangential Christoffel part and its second-fundamental-form part. A direct expansion of Equation (11) then yields

$$\begin{aligned} 0 = {} & \mathrm{div}_{\mathcal{M}}(\widehat{D}\,g\,\nabla_{\mathcal{M}} c_0) - \mathrm{div}_{\mathcal{M}}(c_0\ \mathrm{div}_{\mathcal{M}}\,\widehat{D}) + c_0\big(\nabla_i\,\mathrm{div}_{\mathcal{M}}\,\widehat{D}\big)^i \\ & - \mathrm{Tr}(\widehat{D}\,\mathcal{U}_{1,2}) - \frac{1}{2}\,\mathrm{Tr}\big(N(\partial_r^2 V)^{-1}N^{\intercal}\mathcal{C}\big) \\ & + \sum_{i,j} c_0\,\partial_{x_i,x_j}^2[D]_{i,j} + 2\langle\partial_x c_0, \mathrm{div}_x\,D\rangle \qquad \text{on } \mathcal{M}. \end{aligned} \tag{25}$$

The quantities $\mathcal{U}_{1,2}$ and $\mathcal{C}$ still contain $\partial_r c_0$, but Equation (12) makes those explicit as well. Consequently, after substituting Equation (12), Equation (17), and Equation (18) into Equation (25) and collecting the coefficients of $\partial_u^2 c_0$, $\partial_u c_0$, and $c_0$, we obtain

$$\mathcal{L}_{\mathrm{eff}} c_0 = \mathrm{Tr}\big(\mathbf{A}(u)g(u)^{-1}\partial_u^2 c_0\big) + \beta(u)^{\intercal}\partial_u c_0 + \gamma(u)c_0,$$

with

$$\mathbf{A}(u) = \widehat{D}(u) = g(u)^{-1}\partial_u m(u)^{\intercal} D(m(u))\,\partial_u m(u),$$

and with $\beta$ and $\gamma$ given explicitly by the lower-order terms in Equation (25) after the above substitutions. This is the full coefficient formula promised in the main text.

Finally, we verify ellipticity for the operator exactly as stated in Theorem 3.2. Its principal matrix is

$$\mathbf{A}(u)g(u)^{-1} = g(u)^{-1}\partial_u m(u)^{\intercal} D(m(u))\,\partial_u m(u)\,g(u)^{-1}.$$

Since $m$ is a local parametrization, $g(u) = \partial_u m(u)^{\intercal}\partial_u m(u)$ is symmetric positive definite. Moreover,

$$M(u) \coloneqq \partial_u m(u)^{\intercal} D(m(u))\,\partial_u m(u)$$

is also symmetric positive definite, because for every nonzero $\zeta \in \mathbb{R}^n$,

$$\zeta^{\intercal} M(u)\zeta = \big(\partial_u m(u)\zeta\big)^{\intercal} D(m(u))\big(\partial_u m(u)\zeta\big) > 0.$$

Therefore, for every nonzero $\xi \in \mathbb{R}^n$,

$$\begin{aligned} \xi^{\intercal}\mathbf{A}(u)g(u)^{-1}\xi &= \xi^{\intercal} g(u)^{-1} M(u) g(u)^{-1}\xi \\ &= \big(g(u)^{-1}\xi\big)^{\intercal} M(u)\big(g(u)^{-1}\xi\big) > 0, \end{aligned}$$

because $g(u)^{-1}\xi \neq 0$. Hence the principal symbol is positive definite, and $\mathcal{L}_{\mathrm{eff}}$ is elliptic. □