FIRST-ORDER STATIONARITY OF REVERSE DIFFUSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent literature has shown a strong connection between optimization and sampling. We develop the corresponding first-order theory for diffusion models. First, the SDE-based reverse-time flows of overdamped and underdamped Langevin diffusions contract relative Fisher divergences at explicit exponential rates whenever the stationary potential of the forward process is strongly convex—a condition on the noising process one chooses, not on the data. This is a unique advantage of SDE-based reverse diffusion, absent in the reverse process based on ODEs. Second, we incorporate discretization and establish averaged first-order stationarity bounds—the sampling analog of averaged gradient-norm guarantees in nonconvex optimization—for samplers of both overdamped and underdamped diffusion models. As in nonconvex optimization, the convexity-free certificate is local: it guarantees score consistency, not global mode weights.

1 INTRODUCTION

Diffusion models generate data in two stages: a forward Markov process gradually corrupts samples from an unknown data distribution ρ_0 until their law is essentially a tractable stationary distribution, and generation simulates this process backward in time, guided by the score functions $\nabla \log \rho_s$ of the forward marginals, learned by (denoising) score matching (Hyvärinen, 2005; Vincent, 2011; Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021). Three approximations separate the implemented sampler from the ideal reverse process—initialization from the reference law rather than the true terminal marginal, time discretization, and score estimation—and a large body of work bounds their combined effect on the *terminal* law, in total variation, KL, or Wasserstein distance.

This paper takes a different, more structural view. Rather than asking how far the final sample is from the data, we ask what the reverse dynamics itself—exact or discretized—does to a *first-order* measure of mismatch. The central objects are the *relative score* and the *relative Fisher divergence*: for positive densities p and q,

$$\mathcal{F}(p \,\|\, q) = \mathbb{E}_p \left\| \nabla \log \frac{p}{q} \right\|^2 .$$

The relative score vanishes identically if and only if $p = q$, so $\mathcal{F}(p \,\|\, q)$ is a natural stationarity measure: it plays the role of the squared gradient norm $\|\nabla f\|^2$, as $\mathrm{KL}(\cdot \,\|\, q)$ plays that of the objective value f. The analogy is not merely formal. The Fokker–Planck equation of overdamped Langevin dynamics is the gradient flow of relative entropy in the Wasserstein geometry (Jordan et al., 1998), and the relative Fisher divergence is exactly its gradient norm. We will show in this work that this is also true for the *reverse* flow: $\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \,\|\, q_t) = -\mathcal{F}(p_t \,\|\, q_t)$ (Corollary 2.2)—the exact analog of the descent identity $\frac{\mathrm{d}}{\mathrm{d}t} f(x_t) = -\|\nabla f(x_t)\|^2$ along a gradient flow. Table 1 records the full dictionary the paper centers around.

Most rows of this dictionary are not new: they underlie the first-order theory of forward Langevin Monte Carlo (Balasubramanian et al., 2022) and Fisher-information dissipation along forward flows (Wibisono, 2026). The contribution of this paper is that the dictionary survives time reversal, where the reference law q_t is no longer a fixed stationary distribution but a moving target set by the noising schedule. We show that this symmetry holds for both overdamped and underdamped reverse-time SDEs, but does not extend to their probability flow ODE counterparts.

Why first-order guarantees? In smooth nonconvex optimization one cannot certify convergence to a global minimum, yet gradient descent still admits the averaged guarantee $\frac{1}{N} \sum_k \|\nabla f(x_k)\|^2 \lesssim$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Smooth optimization	Reverse diffusion sampling
objective $f(x)$	relative entropy $\mathrm{KL}(p \,\|\, q)$
gradient ∇f	relative score $\nabla \log(p/q)$ in $L^2(p)$
squared gradient norm $\|\nabla f\|^2$	Fisher divergence $\mathcal{F}(p \,\|\, q)$
learning rate η	step size h
descent identity $\frac{\mathrm{d}}{\mathrm{d}t} f = -\|\nabla f\|^2$	KL dissipation (Corollary 2.2)
strong convexity of f	strong log-concavity of the forward potential U
$\|\nabla f(x_t)\|^2 \le e^{-2mt} \|\nabla f(x_0)\|^2$	exponential Fisher contraction (Theorems 3.2, 3.4)
Polyak–Łojasiewicz inequality	log-Sobolev inequality for q
$\frac{1}{N}\sum_k \|\nabla f(x_k)\|^2 \lesssim \frac{f(x_0)-f^\star}{\eta N} + O(\eta)$	averaged Fisher bounds (Theorems 4.5, 5.5)
change of variables $x \mapsto \Phi(x)$	probability-flow ODE (Proposition 2.3)

Table 1: The optimization dictionary used throughout the paper.

$\frac{f(x_0)-f^\star}{\eta N} + O(\eta)$ (Ghadimi & Lan, 2013). The sampling situation is parallel: for non-log-concave data, converting score accuracy into KL or total-variation convergence requires functional inequalities, whereas the relative Fisher divergence remains meaningful *without such assumptions*—driving it to zero certifies that the sampler's law is locally consistent with its target, the two scores agreeing on average in L^2. The analogy also makes the limitation transparent: as a stationary point need not be a minimum, a law can have small relative Fisher divergence while misallocating mass between well-separated modes. This limitation parallels the gap between stationarity and global optimality in nonconvex optimization.

1.1 Related work

Convergence theory of diffusion models. Time reversal of diffusions goes back to Anderson (1982) and Haussmann & Pardoux (1986); diffusion models build directly on this theory, including the critically damped kinetic variant of Dockhorn et al. (2022). A growing line of quantitative convergence theory controls initialization, discretization, and score-estimation errors simultaneously for essentially arbitrary data distributions (Lee et al., 2023; Chen et al., 2023b; Jiao et al., 2025).

Connections between sampling and optimization. A complementary line of work draws connections between sampling dynamics and optimization. The classical variational formulation of the Fokker–Planck equation interprets overdamped Langevin dynamics as a Wasserstein gradient flow of the relative entropy (Jordan et al., 1998), motivating the viewpoint of sampling as optimization over the space of probability measures (Wibisono, 2018). Momentum-based sampling admits an even closer analogy with accelerated optimization: Ma et al. (2021) interpreted underdamped Langevin dynamics as an analog of Nesterov acceleration for minimizing the KL divergence, with hypocoercivity playing a role analogous to the coupling between position and momentum in accelerated optimization.

Fisher-divergence contraction. Recent work has also studied the evolution of Fisher divergence along isotropic Fokker–Planck channels. Wibisono (2026) established contraction of Fisher divergence under suitable convexity assumptions. Closest in spirit to our work, Balasubramanian et al. (2022) developed Fisher-information guarantees for forward Langevin Monte Carlo, viewing Fisher divergence as a sampling analog of first-order stationarity in optimization. Our work develops a corresponding perspective for *reverse-time* diffusion, where contraction must be established against a reference that itself evolves.

1.2 Contributions and paper organization

Contraction in continuous reverse flow. Our results form two groups, one about exact reverse flows and one about discrete samplers. If the stationary potential U of the forward diffusion is strongly convex, then the reverse SDE contracts score mismatch exponentially, for both overdamped and underdamped diffusion. The assumptions concern only the forward potential, and no log-Sobolev inequality is imposed on the data: the continuous reverse dynamics forgets initialization error exponentially fast for almost arbitrary data.

Averaged Fisher divergence bound for discretized algorithms. (i) For the frozen-score exponential integrator of the reverse overdamped process—the idealized DDPM-type update—with spatially Lipschitz scores and sub-Gaussian data assumptions, the numerical law $\hat{p}_t$ satisfies $\frac{1}{T}\int_0^T \mathcal{F}(\hat{p}_t \,\|\, q_t)\,\mathrm{d}t \lesssim H_0/T + L_s^2 dh + O(h^2)$ against the exact reverse marginals q_t (Theorem 4.5). (ii) For the critically damped reverse process we prove an averaged bound on the full phase-space relative Fisher divergence (Theorem 5.5), with the sampler compared to the twin chain that pushes the *exact* initialization through the *same* numerical kernels, so that the bound isolates the initialization error.

Organization Section 2 fixes the setting; in Section 3, we give results for continuous reverse flows. In Sections 4 and 5, we give results for discrete samplers of the overdamped and underdamped model separately. Section 6 gives our experiment results. Appendices A–D prove the results of Sections 2–5 in order, and Appendix E collects critical-damping identities. Our work focuses on the process mechanism itself, so we analyze initialization error and discretization error. Throughout the paper we assume we have access to the true score functions.

2 Setting: two reverse diffusions and a first-order criterion

This section introduces the two forward models and their reverse dynamics.

Overdamped model. The forward overdamped Langevin diffusion on $\mathbb{R}^d$ is

$$\mathrm{d}X_s = -\nabla U(X_s)\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W_s, \qquad s \in [0,T], \tag{2.1}$$

with invariant density $p^* \propto e^{-U}$. Its Fokker–Planck equation can be rewritten as the continuity equation $\partial_s \rho_s = \nabla\cdot(\rho_s \nabla \log(\rho_s/p^*))$: the probability mass of ρ_s is transported by the velocity field $-\nabla\log(\rho_s/p^*)$. Running time backward, the exact reverse marginals $q_t := \rho_{T-t}$ therefore satisfy $\partial_t q_t = -\nabla\cdot(q_t v_t^q)$ with *current velocity*

$$v_t^q = \nabla \log \frac{q_t}{p^*}, \tag{2.2}$$

realized by the exact reverse SDE $\mathrm{d}Y_t = [\nabla U(Y_t) + 2\nabla \log q_t(Y_t)]\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\bar{W}_t$ (Anderson, 1982; Haussmann & Pardoux, 1986). If p_t is the law of the *same* reverse SDE initialized from $p_0 \neq q_0$, then $\partial_t p_t = -\nabla\cdot(p_t v_t^p)$ with

$$v_t^p = v_t^q - \nabla r_t, \qquad r_t = \log\frac{p_t}{q_t}. \tag{2.3}$$

Underdamped model. The forward kinetic (underdamped) Langevin diffusion lives on phase space $\mathbb{R}^d \times \mathbb{R}^d$, in variables $z = (x, v)$, with friction $\gamma > 0$ and inverse mass $\xi > 0$:

$$\mathrm{d}X_s = \xi V_s\,\mathrm{d}s, \qquad \mathrm{d}V_s = -\nabla U(X_s)\,\mathrm{d}s - \gamma\xi V_s\,\mathrm{d}s + \sqrt{2\gamma}\,\mathrm{d}W_s, \tag{2.4}$$

with invariant density $p^* \propto \exp(-U(x) - \frac{\xi}{2}\|v\|^2)$. Noise and friction act on the velocity coordinates only. Write $J = \left(\begin{smallmatrix} 0 & I_d \\ -I_d & 0 \end{smallmatrix}\right)$ and $G_\gamma = \left(\begin{smallmatrix} 0 & 0 \\ 0 & \gamma I_d \end{smallmatrix}\right)$, so that the drift of (2.4) is $b_{\mathrm{fwd}}(z) = -(J - G_\gamma)\nabla \log p^* = (\xi v,\, -\nabla U(x) - \gamma\xi v)$, with Jacobian

$$M(x) := \nabla b_{\mathrm{fwd}}(z) = \begin{pmatrix} 0 & \xi I_d \\ -\nabla^2 U(x) & -\gamma\xi I_d \end{pmatrix} \tag{2.5}$$

The Fokker–Planck equation of (2.4) is again a continuity equation, with current velocity $w_s = -J\nabla \log p^* - G_\gamma \nabla \log(\rho_s/p^*)$ (Lemma A.2); hence $q_t := \rho_{T-t}$ obeys $\partial_t q_t = -\nabla\cdot(q_t v_t^q)$ with

$$v_t^q = J\nabla \log p^* + G_\gamma \nabla \log \frac{q_t}{p^*}, \tag{2.6}$$

realized by the exact reverse SDE

$$\mathrm{d}\bar{X}_t = -\xi\bar{V}_t\,\mathrm{d}t, \qquad \mathrm{d}\bar{V}_t = \big[\nabla U(\bar{X}_t) + \gamma\xi\bar{V}_t + 2\gamma\nabla_v \log q_t(\bar{X}_t,\bar{V}_t)\big]\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}\bar{W}_t, \tag{2.7}$$

in which only the *velocity* score $\nabla_v \log q_t$ enters—one reason kinetic diffusion models are attractive in practice. For p_t solving the same reverse SDE from $p_0 \neq q_0$

$$v_t^p = v_t^q - G_\gamma \nabla r_t, \qquad r_t = \log\frac{p_t}{q_t}. \tag{2.8}$$

The same probability flow (2.6) is also realized by the deterministic reverse ODE

$$\mathrm{d}\bar{X}_t = -\xi\bar{V}_t\,\mathrm{d}t, \qquad \mathrm{d}\bar{V}_t = \big[\nabla U(\bar{X}_t) + \gamma\xi\bar{V}_t + \gamma\nabla_v \log q_t(\bar{X}_t,\bar{V}_t)\big]\,\mathrm{d}t \tag{2.9}$$

2.1 DIVERGENCES AND THE EXACT DISSIPATION IDENTITY

This subsection defines the objects of the dictionary of Table 1, and then proves the analog of the descent identities, which holds with no assumption on U or on the data. For smooth positive probability densities p, q on $\mathbb{R}^n$ ($n = d$ or $2d$), write $r = \log(p/q)$, $g = \nabla r$, and define

$$\mathrm{KL}(p \,\|\, q) = \int p \log \frac{p}{q}, \qquad \mathcal{F}(p \,\|\, q) = \int p \, \|g\|^2, \qquad \mathcal{F}_P(p \,\|\, q) = \int p \, \langle g, Pg \rangle , \tag{2.10}$$

for a symmetric positive definite matrix P. In the underdamped model we also use the velocity part $\mathcal{F}_v(p \,\|\, q) = \int p \, \|\nabla_v r\|^2$.

Lemma 2.1 (KL derivative for two continuity equations)**.** *Suppose $\partial_t q_t = -\nabla \cdot (q_t v_t^q)$ and $\partial_t p_t = -\nabla \cdot (p_t v_t^p)$, and let $r_t = \log(p_t/q_t)$. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \,\|\, q_t) = \int p_t \, \nabla r_t \cdot (v_t^p - v_t^q) \tag{2.11}$$

Combining Lemma 2.1 with the velocity differences in (2.3) and (2.8) gives the reverse-diffusion descent identity.

Corollary 2.2 (KL dissipation along the exact reverse flows)**.** *Along the common exact reverse dynamics,*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \| q_t) = -\mathcal{F}(p_t \| q_t) \;\; \textit{(overdamped)}, \qquad \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \| q_t) = -\gamma \, \mathcal{F}_v(p_t \| q_t) \;\; \textit{(underdamped)}.$$

In particular $\mathrm{KL}(p_t \,\|\, q_t)$ is always nonincreasing in both models.

The entire effect of the mismatched initialization is the single extra correction term $v_t^p - v_t^q$. The probability-flow ODE realizes the same marginals q_t from q_0, but transports *every* initial law by the same field v_t^q, so no correction term appears at all.

Proposition 2.3 (No forgetting along the probability-flow ODE)**.** *Let Φ_t be the flow map of the probability-flow ODE of either model, and let $p_t = (\Phi_t)_{\#} p_0$ for an arbitrary initial law p_0, and $q_t = (\Phi_t)_{\#} q_0$ are the exact reverse marginals. Then $r_t = r_0 \circ \Phi_t^{-1}$; hence for all $t \in [0, T]$*

$$\mathrm{KL}(p_t \,\|\, q_t) = \mathrm{KL}(p_0 \,\|\, q_0), \qquad \mathrm{TV}(p_t, q_t) = \mathrm{TV}(p_0, q_0),$$

and likewise for every f-divergence, while the relative score is merely carried along and rescaled by the Jacobian of the flow,

$$\nabla r_t\big(\Phi_t(x)\big) = \big(\nabla \Phi_t(x)\big)^{-\top} \nabla r_0(x), \qquad \mathcal{F}(p_t \,\|\, q_t) = \int p_0 \, \big\| (\nabla \Phi_t)^{-\top} \nabla r_0 \big\|^2.$$

In the dictionary the ODE is a change of variables: the objective and every zeroth-order distance are untouched, and the gradient is transported, not contracted.

Remark 2.4 (Optimization counterpart, and a moving target). Corollary 2.2 has a familiar optimization counterpart. For the gradient flow $\dot{x}_t = -\nabla f(x_t)$ and for the heavy-ball flow $\dot{x}_t = v_t$, $\dot{v}_t = -\nabla f(x_t) - \gamma v_t$, one has $\frac{d}{dt} f(x_t) = -\|\nabla f(x_t)\|^2$ and $\frac{d}{dt}\Big[f(x_t) + \frac{1}{2}\|v_t\|^2\Big] = -\gamma \, \|v_t\|^2$. In Wasserstein space the role of f is played by $\mathrm{KL}(\cdot \,\|\, q_t)$. In reverse diffusion the target itself moves: p_t chases q_t, which travels on a schedule fixed by the forward process; for the exact flows this costs nothing, and Lemma 2.1 shows why. The same terms in v_t^q and v_t^p only make two distributions move in parallel; only relative velocity $v_t^p - v_t^q$ matters, and it is precisely what drives p_t toward q_t. As in nonconvex optimization, a counterpart of the averaged statement $\frac{1}{T}\int_0^T \|\nabla f(x_t)\|^2 \, dt \le \big(f(x_0) - \inf f\big)/T$ can be obtained by integrating Corollary 2.2. Denote the initialization error as $H_0 := \mathrm{KL}(p_0 \,\|\, q_0)$.

Proposition 2.5 (Averaged stationarity of the exact reverse flows)**.** *For every potential U, every data distribution, and every pair of initializations (p_0, q_0),*

$$\frac{1}{T}\int_0^T \mathcal{F}(p_t \,\|\, q_t) \, \mathrm{d}t = \frac{H_0 - \mathrm{KL}(p_T \,\|\, q_T)}{T} \le \frac{H_0}{T} \qquad \textit{(overdamped)}, \tag{2.12}$$

$$\frac{\gamma}{T}\int_0^T \mathcal{F}_v(p_t \,\|\, q_t) \, \mathrm{d}t = \frac{H_0 - \mathrm{KL}(p_T \,\|\, q_T)}{T} \le \frac{H_0}{T} \qquad \textit{(underdamped)}, \tag{2.13}$$

Assuming only $H_0 < +\infty$, the time-averaged Fisher divergence of the overdamped flow converges to 0 as $T \to +\infty$. However, for the underdamped model we only obtain a bound for the divergence in velocity dimensions $\mathcal{F}_v$. The theorem of Section 4 is a discrete-time version of Proposition 2.5 for the overdamped model.

3 Exponential Fisher contraction of the exact reverse flows

In this section we directly study the evolution of $\mathcal{F}(p_t \,\|\, q_t)$, and show that it contracts exponentially fast under strong convexity of the forward potential U. Throughout this section q_t is the exact reverse marginal, p_t solves the *same* exact reverse SDE from an arbitrary initial law p_0, and $r_t = \log(p_t/q_t)$, $g_t = \nabla r_t$, $S_t = \nabla^2 r_t$.

Assumption 3.1 (Strong convexity and smoothness)**.** $U \in C^2(\mathbb{R}^d)$ and there are $0 < m \le L < \infty$ with $mI_d \preceq \nabla^2 U(x) \preceq LI_d$ for all x.

Assumption 3.1 constrains only the potential U chosen by the practitioner (for the standard Gaussian noising process $m = L = 1$).

Overdamped: the ordinary Fisher divergence contracts. The mechanism is a dissipation identity for $\mathcal{F}(t) := \mathcal{F}(p_t \,\|\, q_t)$, proved in Appendix B (Proposition B.3): along the exact reverse flow,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(t) = -2\int p_t \, \|S_t\|_{\mathrm{F}}^2 - 2\int p_t \, g_t^\top \nabla^2 U \, g_t. \tag{3.1}$$

The first term is a negative term and is always favorable; the second is where convexity of U enters. Gronwall's inequality then gives:

Theorem 3.2 (Overdamped Fisher contractivity)**.** *In the overdamped model, if U is convex, then $\mathcal{F}(p_t \,\|\, q_t)$ is nonincreasing; if U is m-strongly convex, then $\mathcal{F}(p_t \,\|\, q_t) \le e^{-2mt}\,\mathcal{F}(p_0 \,\|\, q_0)$ for $0 \le t \le T$.*

Remark 3.3 (The reverse flow inherits forward contractivity)*.* Perhaps surprisingly, identity (3.1) is exactly the same as for the *forward* Langevin dynamics: Forward Fisher contractivity against the invariant law relies on convexity of U alone, and the reverse flow inherits it *without any functional inequality on q_t*. This is a special case, which we want to highlight here for its importance, of Fisher evolution along isotropic Fokker–Planck channels (Wibisono, 2026, Lemma 2). For continuous reverse processes, our technical contribution is the theorem for the underdamped model.

Underdamped: a modified Fisher divergence contracts. The underdamped flow dissipates only the velocity components of the relative score (Corollary 2.2), so position-direction error can be damped only *indirectly*, by rotating into the velocity directions through the Hamiltonian coupling. Following the hypocoercivity paradigm (Villani, 2009), we therefore measure the error in a quadratic form with position–velocity *cross terms*, chosen so that the coupling converts the conserved part of the error into the dissipated part. Concretely, for $\gamma^2\xi > L$ define

$$P_\gamma = \begin{pmatrix} 1 & -\frac{\gamma}{2} \\ -\frac{\gamma}{2} & \frac{\gamma^2}{2} \end{pmatrix}, \qquad \kappa_{\gamma,\xi} = \min\left(\frac{m}{\gamma},\ \gamma\xi - \frac{L}{\gamma}\right) > 0. \tag{3.2}$$

With $P = P_\gamma \otimes I_d$, the associated *modified Fisher divergence* is

$$\mathcal{F}_P(p_t \,\|\, q_t) = \int p_t \, g_t^\top P \, g_t \,\mathrm{d}z = \int p_t \Big(\|\nabla_x r_t\|^2 - \gamma \,\langle \nabla_x r_t, \nabla_v r_t \rangle + \frac{\gamma^2}{2}\|\nabla_v r_t\|^2 \Big)\,\mathrm{d}z. \tag{3.3}$$

Theorem 3.4 (Underdamped modified Fisher contractivity)**.** *Assume the underdamped model* (2.4)*, Assumption 3.1, and $\gamma^2\xi > L$. With $P = P_\gamma \otimes I_d$ as in* (3.2)*,*

$$\mathcal{F}_P(p_t \,\|\, q_t) \le e^{-2t\kappa_{\gamma,\xi}}\,\mathcal{F}_P(p_0 \,\|\, q_0)$$

Since P has eigenvalues bounded away from 0 and ∞, contraction of $\mathcal{F}_P$ is equivalent, up to a fixed conditioning factor, to contraction of $\mathcal{F}$ itself. For the standard Gaussian potential $U(x) = \frac{1}{2}\|x\|^2$ we have $m = L = 1$, and we can select parameters $\gamma = 2$, $\xi = 1$ used by critically-damped diffusion models (Dockhorn et al., 2022).

4 AVERAGED STATIONARITY OF THE DISCRETIZED OVERDAMPED REVERSE

We now account for the error incurred by time discretization. The practically dominant forward process is the standard Ornstein–Uhlenbeck (OU) diffusion,

$$\mathrm{d}Y_s = -Y_s\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}B_s, \qquad Y_0 \sim \rho_0, \tag{4.1}$$

with marginals ρ_s: the overdamped model (2.1) with $U(x) = \frac{1}{2}\|x\|^2$. The sampler we analyze is the *frozen-score exponential integrator*, the idealized form of standard DDPM-type updates: on the uniform grid $t_k = kh$, $h = T/N$, freeze the score at the step start and solve the resulting linear SDE exactly. Its continuous-time interpolation is

$$\mathrm{d}X_t = \{X_t + 2s_{t_k}(X_{t_k})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad t \in [t_k, t_{k+1}], \tag{4.2}$$

which integrates in closed form: for $\tau = t - t_k$,

$$X_t = e^{\tau} X_{t_k} + 2(e^{\tau} - 1)s_{t_k}(X_{t_k}) + \sqrt{e^{2\tau} - 1}\,\zeta, \qquad \zeta \sim N(0, I_d). \tag{4.3}$$

Let $\hat{p}_t$ denote the law of X_t, the numerical process. The object of interest is the time-averaged Fisher divergence $\frac{1}{T}\int_0^T \mathcal{F}(\hat{p}_t \,\|\, q_t)\,\mathrm{d}t$ against the *exact reverse marginal*: the sampling analog of the averaged gradient norm at a randomly stopped iterate (Ghadimi & Lan, 2013), and the reverse-process counterpart of Balasubramanian et al. (2022).

Assumption 4.1 (Score Lipschitzness)**.** There is $L_s \geq 0$ with $\|s_t(x) - s_t(y)\| \leq L_s\|x - y\|$ for every $t \in [0, T]$ and $x, y \in \mathbb{R}^d$.

Assumption 4.2 (Finite initial divergence)**.** The initial laws have positive densities with $H_0 := \mathrm{KL}(\hat{p}_0 \,\|\, q_0) < \infty$.

Assumption 4.3 (Sub-Gaussian data)**.** For $Y_0 \sim \rho_0$ there is a $\lambda > 0$ with $\mathbb{E}\exp\left(\frac{\lambda}{d}\|Y_0\|^2\right) =: M < \infty$.

Set $H_0 := \mathrm{KL}(\hat{p}_0 \,\|\, q_0)$, $\lambda_0 := \min\{\lambda, \frac{d}{4}\}$, and $\mathcal{M} := \frac{d}{\lambda_0}(H_0 + \log M + d)$.

Remark 4.4 (Role of the Sub-Gaussian assumption). Assumption 4.3 serves exactly one purpose in the proof: controlling the second moments of the *numerical* marginals $\hat{p}_t$. The assumption is equivalent to standard definition of sub-Gaussian distributions, normalized by the dimension d.

Theorem 4.5 (Averaged Fisher stationarity, overdamped)**.** *Suppose Assumptions 4.1–4.3 hold. There is a universal constant $c > 0$ such that the following holds. If*

$$h \leq \frac{c}{1 + L_s} \qquad \textit{and} \qquad T\left[L_s^2 dh + L_s^2(1 + L_s)^2\mathcal{M}h^2\right] \leq cd, \tag{4.4}$$

then the numerical law never drifts far from the exact one in relative entropy, $\sup_{0\leq t\leq T}\mathrm{KL}(\hat{p}_t \,\|\, q_t) < H_0 + d$*, and*

$$\frac{1}{T}\int_0^T \mathcal{F}(\hat{p}_t \,\|\, q_t)\,\mathrm{d}t \;\lesssim\; \frac{H_0}{T} + L_s^2 dh + L_s^2(1 + L_s)^2\mathcal{M}h^2. \tag{4.5}$$

Remark 4.6 (Shape of the bound). Bound (4.5) has exactly the shape of the nonconvex optimization guarantee in Table 1: an amortized initialization term H_0/T plus step-size terms. It refines the exact identity (2.12), which it recovers as $h \to 0$.

Remark 4.7 (Score error). Suppose the integrator is run with estimated scores $\hat{s}_{t_k}$ in place of s_{t_k} in (4.2), with $\max_k \mathbb{E}_{\hat{p}_{t_k}}\|\hat{s}_{t_k} - s_{t_k}\|^2 \leq \varepsilon^2$. The only change in the proof is one triangle inequality inside the mismatch term of Proposition C.3, $\mathbb{E}\|\hat{s}_{t_k}(X_{t_k}) - s_t(X_t)\|^2 \leq 2\varepsilon^2 + 2\,\mathbb{E}\|s_{t_k}(X_{t_k}) - s_t(X_t)\|^2$. Consequently, with slight adjustment of the proof, the conclusions of Theorem 4.5 hold with the additional term ε^2 on the right-hand side of (4.5). In practice, however, the estimator is trained under the exact marginals q_{t_k}, so its error is naturally controlled in $L^2(q_{t_k})$ rather than in $L^2(\hat{p}_{t_k})$. Transferring the bound from one measure to the other requires either a uniform bound or a change of measure, which is achievable under a Gaussian-tail assumption on score errors, but we do not pursue here. Appendix F.4 confirms empirically that the mechanism analyzed here persists under a learned score.

5 AVERAGED STATIONARITY OF THE DISCRETIZED CRITICALLY DAMPED REVERSE

We now prove the underdamped counterpart of Theorem 4.5, for the critically damped kinetic OU forward process: the underdamped model (2.4) with $U(x) = \frac{1}{2}\|x\|^2$ and the parameters $\gamma = 2$, $\xi = 1$, and for the frozen-score exponential integrator. One genuine obstacle separates this case from Section 4. Repeating the argument of Section 4 along the frozen-score SDE interpolation is possible, but it could only provide a bound on $\mathcal{F}_v(\hat{p}_t \,\|\, q_t)$, as in Proposition 2.5. To upgrade $\mathcal{F}_v$ to the full $\mathcal{F}$ one wants a hypocoercive Lyapunov functional in the spirit of Ma et al. (2021); but differentiating $\mathcal{F}_P(\hat{p}_t \,\|\, q_t)$ along an SDE interpolation like (4.2) produces derivatives of the score mismatch, which are hard to control.

Our resolution is decomposing the integrator. Since the frozen-score step solves a *linear* SDE, it can be written as a deterministic *score kick* applied at the beginning of the step, followed by the score-free reverse dynamics $\mathrm{d}Y = CY\,\mathrm{d}\tau + 2E\,\mathrm{d}W$ run for the full step (Lemma D.1). Along this interpolation no derivative of the score error ever appears: the kick is a deterministic map, and the score-free leg is an exact linear diffusion on which entropy and twisted-Fisher identities hold in closed form. Therefore, hypocoercive dissipation and discretization decouple completely.

However, this decoupled composition also forces us to tell the story in an indirect way. Instead of bounding $\mathcal{F}(\hat{p}_t \,\|\, q_t)$ directly, we will bound $\mathcal{F}(\hat{p}_t \,\|\, \hat{q}_t)$ for a reference process $\hat{q}_t$, which isolates the initialization error. The discretization is isolated in the difference between $\hat{q}_t$ and q_t, and it is controlled at the level of path measures by the standard Girsanov computation (Chen et al., 2023b).

5.1 THE CRITICALLY DAMPED REVERSE PROCESS AND THE INTEGRATOR

For $z = (x, v) \in \mathbb{R}^{2d}$ set

$$C := \begin{pmatrix} 0 & -I_d \\ I_d & 2I_d \end{pmatrix}, \qquad E := \begin{pmatrix} 0 \\ I_d \end{pmatrix}, \qquad \Pi_v := EE^\top, \qquad D := 4\Pi_v. \tag{5.1}$$

The critically damped kinetic noising process, in the forward clock s and with marginal densities ρ_s, is

$$\mathrm{d}Z_s = -CZ_s\,\mathrm{d}s + 2E\;\mathrm{d}B_s, \qquad \text{i.e.} \qquad \mathrm{d}X_s = V_s\,\mathrm{d}s, \qquad \mathrm{d}V_s = -(X_s + 2V_s)\,\mathrm{d}s + 2\;\mathrm{d}B_s; \tag{5.2}$$

its transition law over time h is $N(e^{-hC}z,\, G_h)$ with $G_h := \int_0^h e^{-rC} D e^{-rC^\top}\,\mathrm{d}r$. In the *increasing* reverse clock $t := T - s$, with $q_t := \rho_{T-t}$, the path-reversed process has dynamics

$$\mathrm{d}Y_t = \big(CY_t + 4E\nabla_v \log q_t(Y_t)\big)\,\mathrm{d}t + 2E\;\mathrm{d}\bar{B}_t, \qquad Y_0 \sim \rho_T \tag{5.3}$$

Fix an increasing reverse-time grid $0 = t_0 < t_1 < \cdots < t_N = T$ with steps $h_i := t_{i+1} - t_i > 0$, not necessarily uniform, and let $s_i := \nabla_v \log q_{t_i}$ be the velocity score at reverse time t_i. The sampler is the frozen-score exponential integrator of (5.3), the kinetic analog of (4.3): freeze s_i at the step start and solve the resulting linear SDE exactly,

$$\hat{Y}_{i+1} = e^{h_i C}\hat{Y}_i + 4\Big(\int_0^{h_i} e^{uC}\,\mathrm{d}u\Big) E s_i(\hat{Y}_i) + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, Q_{h_i}), \tag{5.4}$$

where $Q_h := \int_0^h e^{rC} D e^{rC^\top}\,\mathrm{d}r = e^{hC} G_h e^{hC^\top}$ is the covariance accumulated by the noise over a step of the reverse dynamics. For $h > 0$ define the kick vector and the kick map

$$B_h := 4\int_0^h e^{-rC} E\,\mathrm{d}r = \begin{pmatrix} 4[1 - (1+h)e^{-h}]I_d \\ 4he^{-h} I_d \end{pmatrix}, \qquad \Psi_{i,h}(z) := z + B_h s_i(z), \tag{5.5}$$

and let $\mathcal{R}_\tau$ be the Markov semigroup of the *score-free reverse dynamics*, the linear SDE

$$\mathrm{d}Y_\tau = CY_\tau\,\mathrm{d}\tau + 2E\;\mathrm{d}W_\tau, \qquad \mathcal{R}_\tau :\; y \longmapsto e^{\tau C} y + \mathcal{N}(0, Q_\tau), \tag{5.6}$$

obtained from (5.3) by switching the score off. Therefore the transition kernel K_i of the update (5.4) can be factorized as (Lemma D.1):

$$\mu\mathsf{K}_i = \mathcal{R}_{h_i}\big((\Psi_{i,h_i})_{\#}\mu\big) \qquad \text{for every input law } \mu : \tag{5.7}$$

one step of the integrator is the score kick Ψ_{i,h_i}—the whole score impulse of the step, applied at its start—followed by the score-free reverse flow for the length of the step. All discretization error sits in the deterministic kick, and all stochastic evolution is an exact linear diffusion. We propagate the numerical initialization $\hat{p}_0$ *and* the exact initialization $\hat{q}_0 = \rho_T$ through the same kernels, interpolating within step i by the clock $0 \le \tau \le h_i$ of the score-free leg:

$$\hat{p}_{i+1} = \hat{p}_i \mathsf{K}_i, \qquad \hat{q}_{i+1} = \hat{q}_i \mathsf{K}_i; \qquad \hat{p}_{i,\tau} := \mathcal{R}_\tau\big((\Psi_{i,h_i})_\# \hat{p}_i\big), \qquad \hat{q}_{i,\tau} := \mathcal{R}_\tau\big((\Psi_{i,h_i})_\# \hat{q}_i\big), \tag{5.8}$$

so that $\hat{p}_{i,h_i} = \hat{p}_{i+1}$ and $\hat{q}_{i,h_i} = \hat{q}_{i+1}$. At the grid points these are the laws of the sampler and of its twin; within a step they are the laws, at reverse time $t_i + \tau$, of the process that kicks at t_i and then follows the score-free reverse dynamics.

5.2 Assumptions and the stationarity theorem

Assumption 5.1 (Score Lipschitzness)**.** Every s_i is C^1, and $\|\nabla s_i(z)\|_{\mathrm{op}} \le L_{\mathrm{map}}$ for all $0 \le i < N$ and $z \in \mathbb{R}^{2d}$, for some $L_{\mathrm{map}} \ge 0$.

Assumption 5.2 (Reference mixed Hessian)**.** $\|\nabla_z \nabla_v \log \hat{q}_{i,\tau}(z)\|_{\mathrm{op}} \le L_{\mathrm{ref}}$ for all i, $\tau \in (0, h_i]$, z, for some $L_{\mathrm{ref}} \ge 0$.

Assumption 5.3 (Finite initial divergence)**.** The initial laws have positive densities with $H_0 := \mathrm{KL}(\hat{p}_0 \,\|\, q_0) < \infty$ and $F_0 := \mathcal{F}(\hat{p}_0 \,\|\, q_0) < \infty$.

Remark 5.4 (On the status of Assumption 5.2). Assumption 5.2 is actually an implicit constraint on the data distribution, and does not involve the numerical initialization $\hat{p}_0$. The twin $\hat{q}_{i,\tau}$ is the *exact* initialization ρ_T pushed through the known, fixed kernels—kicks by the velocity scores and explicit linear diffusions—so Assumption 5.2 is an additional regularity property of the data propagated by known dynamics, and it can be explicitly verified for Gaussian ρ_0.

Set $\Lambda := \max\{1, L_{\mathrm{map}}, L_{\mathrm{ref}}\}$.

Theorem 5.5 (Averaged relative-Fisher stationarity, underdamped)**.** *Suppose Assumptions 5.1–5.3 hold and*

$$\max_{0 \le i < N} h_i \le \frac{1}{64\Lambda^2}. \tag{5.9}$$

Then the full phase-space relative score satisfies

$$\frac{1}{T} \sum_{i=0}^{N-1} \int_0^{h_i} \mathcal{F}(\hat{p}_{i,\tau} \,\|\, \hat{q}_{i,\tau}) \,\mathrm{d}\tau \;\lesssim\; \frac{\Lambda^2 H_0 + F_0}{T}. \tag{5.10}$$

Remark 5.6 (Discretization error). Notice that the right-hand side only contains initialization error. The discretization error is inside the difference between $\hat{q}_{i,\tau}$ and q_t. Denote $Q_{0:T}$ the path measure of the exact reverse process and $\hat{Q}_{0:T}$ that of the twin. The standard Girsanov analysis (Theorem 16 of Chen et al. (2023b)) gives $\mathrm{KL}\left(Q_{0:T} \,\|\, \hat{Q}_{0:T}\right) \le \tilde{O}(h)$ at grid times.

6 Numerical experiments

Two experiments with *exact* scores (Appendix F) show initialization error being forgotten exponentially fast in the first-order sense, but not the allocation of mass between well-separated modes.

First, $\rho_0 = 0.3\,N(-20, 1.5^2) + 0.4\,N(0, 1.5^2) + 0.3\,N(20, 1.5^2)$ is noised by (4.1) up to $T = 5$, and both exact reverse dynamics—the SDE and the probability-flow ODE—are started from $p_0 = N(10, 0.5^2)$ instead of $q_0 = \rho_T$ ($H_0 = 49.8$, $\mathcal{F}_0 = 100$). Along the SDE (Figure 1(a)), $\mathcal{F}(p_t \,\|\, q_t)$ contracts at exactly the rate of Theorem 3.2, $\mathrm{KL}(p_t \,\|\, q_t)$ decreases monotonically but flattens once the trajectories have committed to a mode. The terminal relative Fisher divergence $\mathcal{F}(p_t \,\|\, q_T) = 4.5 * 10^{-3}$ is small, but panel (b) shows mode weights $(0.05, 0.25, 0.71)$ instead of $(0.3, 0.4, 0.3)$—locally indistinguishable from the data, globally wrong. The ODE, with the same score, the same marginals and the same initialization, forgets nothing (Proposition 2.3): the terminal law is far outside the data. Panel (c): the time-averaged Fisher divergence of the integrator stays below H_0/T, with a gap to the exact flow decaying like $\tilde{O}(h)$ (Theorem 4.5).

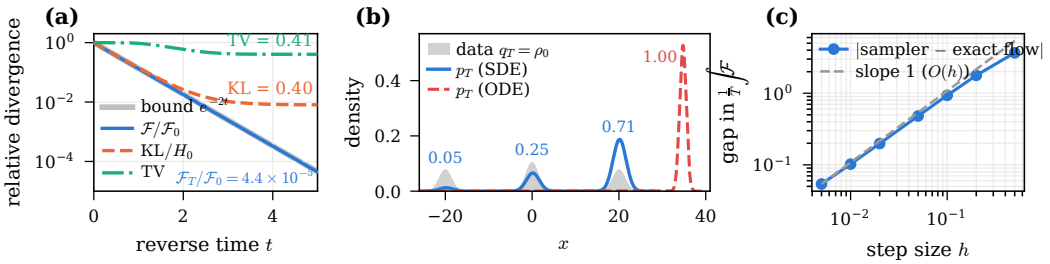


Figure 1: One-dimensional Gaussian mixture with exact scores, wrong initialization $p_0 = N(10, 0.5^2)$. **(a)** $\mathcal{F}/\mathcal{F}_0$ against the bound e^{-2t} of Theorem 3.2, KL/H_0, and TV along the exact reverse SDE. **(b)** Terminal laws of the reverse SDE and of the probability-flow ODE from the same initialization. **(c)** Gap between the time-averaged $\mathcal{F}(\hat{p}_t \,\|\, q_t)$ of the frozen-score integrator and its exact-flow counterpart $\mathcal{F}(p_t \,\|\, q_t)$ versus step size (log–log).

Second, with ρ_0 the empirical law of the $n = 28000$ CelebA-HQ training images (Karras et al., 2018) at 256×256, the OU marginals are kernel density estimates whose bandwidth is the noise level, so the sampler of Section 4 runs with the *exact* score ($T = 5.06$, 150 steps geometric in the noise-to-signal ratio, ending at $s = 0$). Figure 3 shows 64 generated images for three initializations. Every image is close to a training image whatever the initialization—local consistency holds—but which images are produced depends on it: the nearly exact $N(0, I_d)$ returns a mixed draw of the training set, while the mean shifts $\mu = \pm 4\,\mathbf{1}/\sqrt{d}$ return almost exclusively bright, respectively dark, photographs. A pretrained DDPM, which includes the score estimation error and does not purely memorize, shows the same bias from the same initializations and horizon (Appendix F.4).

7 Conclusion

Reverse diffusions admit a first-order theory that parallels smooth optimization line by line. The relative score is the gradient; the SDE reverse flow is its descent dynamics, dissipating KL and contracting Fisher divergences exponentially under strong log-concavity of the forward stationary potential—a condition on the noising process one chooses, not on the data. Under mild assumptions, implementable integrators then achieve *averaged* first-order stationarity at rate H_0/T plus explicit step-size terms.

This also separates the two standard samplers. Only the SDE descends, and the ODE is a change of variables, so it forgets nothing of its initialization. Since the identities of Section 2 hold from any time on, the same holds for any perturbation of the law introduced during sampling: the SDE forgets it in the first-order sense, the ODE carries it to the end. This gives a precise form to the observation that stochastic samplers correct earlier errors (Karras et al., 2022) and to the role of Langevin corrector steps in ODE-based samplers (Song et al., 2021; Chen et al., 2023a).

However, the stationarity guarantee for the underdamped reverse process is indirect. Theorem 5.5 compares the sampler $\hat{p}$ with a twin chain $\hat{q}$ that pushes the *exact* initialization through the *same* numerical kernels, so it isolates the initialization error. This is the price of decoupling the analysis of discretization and hypocoercivity, and Assumption 5.2 is also an artifact of this compromise. Even though we have $\mathrm{KL}\left(Q_{0:T} \,\|\, \hat{Q}_{0:T}\right) \le \tilde{O}(h)$, it does not automatically transfer to a control on $\mathcal{F}(\hat{p}_t \,\|\, q_t)$. Closing this gap—a first-order comparison of the kinetic sampler with the exact reverse process itself—is a natural direction for future work.

As the dictionary predicts, a first-order certificate is local: a law can have small relative Fisher divergence while misallocating mass between well-separated modes, so our guarantees certify local consistency with the target rather than recovery of mode weights—the price of convexity-free guarantees, in sampling as in nonconvex optimization. A log-Sobolev inequality for the data plays the role of the Polyak–Łojasiewicz condition (Polyak, 1963; Karimi et al., 2016): when it holds, the Fisher divergence bounds then convert into KL bounds against the reference laws: the first-order theory recovers zeroth-order convergence exactly when a PL-type inequality holds, as in optimization.

ETHICS STATEMENT

This is a theoretical work on the mathematical foundations of diffusion-based sampling. Our experiments use a synthetic one-dimensional mixture and the public CelebA-HQ face dataset (Karras et al., 2018), which we use for non-commercial research only; because the exact empirical score memorizes its training set, the images in Figure 3 are (up to sub-gray-level noise) images from that dataset. We collect no new data, involve no human subjects, and deploy nothing; we do not foresee direct ethical concerns beyond those generic to improved understanding of generative models.

REPRODUCIBILITY STATEMENT

The theoretical results are self-contained. Every theorem, proposition, corollary, and lemma is stated together with its assumptions; the complete proofs are given in Appendices A–D. The numerical experiments of Section 6 are documented in full in Appendix F. Code reproducing all figures is provided as supplementary material.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims and to assist in the writing of proofs. The strategy of every proof is our own: for each result the authors devised the overall argument—the quantities to control, the decomposition of the error, and the key inequalities. Generative AI tools were then used to fill in some of the intermediate steps, such as routine computations, algebraic manipulations, and the verification of individual inequalities, and to write them up. Additionally, we used generative AI tools to edit the paper for readability, and to create and edit implementation code.

We have not used generative AI tools to develop the theoretical model or conceptual framework of the paper, to formulate mathematical claims, to propose or refine hypotheses, to generate synthetic data sets, or to interpret results; assisting with translation, cleaning and reformatting data sets, and supporting qualitative or thematic data analysis are not applicable to this work.

We have reviewed all AI-assisted work: every AI-assisted proof was checked line by line by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Krishna Balasubramanian, Sinho Chewi, Murat A. Erdogdu, Adil Salim, and Shunshi Zhang. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for Langevin Monte Carlo. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, pp. 2896–2923, 2022.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 68552–68575, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023b.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped Langevin diffusion. In *International Conference on Learning Representations (ICLR)*, 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025.

Ulrich G. Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Yuchen Jiao, Yuchen Zhou, and Gen Li. Optimal convergence analysis of DDPM for general distributions. *arXiv preprint arXiv:2510.27562*, 2025.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations (ICLR)*, 2024.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 795–811, 2016.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 26565–26577, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory (ALT)*, pp. 946–985, 2023.

Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.

Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942–1992, 2021. doi: 10.3150/20-BEJ1297.

Boris T. Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Cédric Villani. *Hypocoercivity*, volume 202(950) of *Memoirs of the American Mathematical Society*. American Mathematical Society, 2009.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pp. 2093–3027. PMLR, 2018.

Andre Wibisono. Mixing time of the proximal sampler in relative Fisher information via strong data processing inequality. *IEEE Transactions on Information Theory*, 72(5):2667–2688, 2026. doi: 10.1109/TIT.2026.3670723.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

STANDING CONVENTIONS FOR THE APPENDICES

Regularity. All probability densities appearing in the appendices are positive and smooth, with enough decay at infinity that the integrals we write are finite, that differentiation under the integral sign is justified, and that the integrations by parts we perform carry no boundary terms. This is the standard qualitative setting for Fisher-information computations (Vempala & Wibisono, 2019; Balasubramanian et al., 2022).

Notation. Universal constants $c, C > 0$ are absolute and may change from line to line; $A \lesssim B$ means $A \le CB$ (in Appendix D, C is the matrix of (5.1)). For a vector field F we write $(\nabla F)_{ij} = \partial_{z_j} F_i$; for a scalar φ, $\nabla^2\varphi$ is the Hessian, and ∇_v, Δ_v act on the velocity block. For the pair of laws under consideration we write $r = \log(\text{tracked law}/\text{reference law})$ and $g = \nabla r$, so that $\mathcal{F}(p \,\|\, q) = \int p \, \|g\|^2$ and $\mathcal{F}_P(p \,\|\, q) = \int p \, \langle g, Pg\rangle$ as in (2.10).

A PROOFS FOR SECTION 2

Both models of Section 2 are instances of a forward diffusion on $\mathbb{R}^n$ with constant diffusion matrix,

$$\mathrm{d}Z_s = b(Z_s)\,\mathrm{d}s + \sqrt{2}\,G^{1/2}\,\mathrm{d}W_s, \qquad s \in [0,T], \tag{A.1}$$

where $b \in C^1(\mathbb{R}^n;\mathbb{R}^n)$ and G is symmetric positive semidefinite: the overdamped model (2.1) has $n = d$, $b = -\nabla U$, $G = I_d$, and the underdamped model (2.4) has $n = 2d$, $b = b_{\mathrm{fwd}}$, $G = G_\gamma$. We write ρ_s for the marginals of (A.1) and treat both models at once.

Lemma A.1 (Current velocities of the forward and reverse flows)**.** *The marginals of* (A.1) *satisfy* $\partial_s\rho_s = -\nabla\cdot(\rho_s w_s)$ *with* $w_s = b - G\nabla\log\rho_s$. *Hence the reverse marginals* $q_t := \rho_{T-t}$ *satisfy* $\partial_t q_t = -\nabla\cdot(q_t v_t^q)$ *with*

$$v_t^q = -b + G\nabla\log q_t, \tag{A.2}$$

and q_t *is the law of the reverse SDE*

$$\mathrm{d}Y_t = \big[-b(Y_t) + 2G\nabla\log q_t(Y_t)\big]\,\mathrm{d}t + \sqrt{2}\,G^{1/2}\,\mathrm{d}\bar{W}_t, \qquad Y_0 \sim q_0. \tag{A.3}$$

If p_t *is the law of* (A.3) *started from* $p_0 \neq q_0$ *instead, then* $\partial_t p_t = -\nabla\cdot(p_t v_t^p)$ *with*

$$v_t^p = v_t^q - G\nabla r_t, \qquad r_t = \log\frac{p_t}{q_t}. \tag{A.4}$$

Proof. The Fokker–Planck equation of (A.1) is

$$\begin{aligned}\partial_s\rho_s &= -\nabla\cdot(\rho_s b) + \Delta(G\rho_s)\\ &= -\nabla\cdot(\rho_s b) + \nabla\cdot(G\nabla\rho_s)\\ &= -\nabla\cdot(\rho_s(b - G\nabla\log\rho_s))\end{aligned}$$

Substituting $s = T - t$ reverses the sign of the velocity, which gives (A.2). The Fokker–Planck equation of (A.3), for the law μ_t of a solution, is

$$\partial_t\mu_t = -\nabla\cdot\big(\mu_t[-b + 2G\nabla\log q_t]\big) + \nabla\cdot(G\nabla\mu_t) = -\nabla\cdot\big(\mu_t[-b + 2G\nabla\log q_t - G\nabla\log\mu_t]\big).$$

For $\mu_t = q_t$ the velocity field in the bracket equals v_t^q, so q_t solves this equation, and by uniqueness for the Fokker–Planck equation the SDE (A.3) started from q_0 has marginals q_t: this is the time-reversal theorem of Anderson (1982); Haussmann & Pardoux (1986). For $\mu_t = p_t$ the bracket equals $v_t^q - G\nabla\log(p_t/q_t)$, which is (A.4). □

With $G = I_d$ and $b = -\nabla U$ one has $-b + G\nabla \log q_t = \nabla \log(q_t/p^*)$, and (A.2)–(A.4) are (2.2) and (2.3). For the underdamped model, the next lemma records the form used in Section 2.

Lemma A.2 (Current velocity, underdamped)**.** *Let ρ_s be the density of* (2.4)*. Then $\partial_s \rho_s = -\nabla \cdot (\rho_s w_s)$ with*

$$w_s = -J\nabla \log p^* - G_\gamma \nabla \log \frac{\rho_s}{p^*}.$$

Consequently $q_t = \rho_{T-t}$ has the current velocity (2.6)*, and a mismatched initialization has the current velocity* (2.8)*.*

Proof. Since $\nabla \log p^*(z) = (-\nabla U(x), -\xi v)$, the drift of (2.4) is $b_{\mathrm{fwd}} = (\xi v, -\nabla U(x) - \gamma \xi v) = -(J - G_\gamma)\nabla \log p^*$, and the noise $\sqrt{2\gamma}\, \mathrm{d}W_s$ on the velocity block is $\sqrt{2}\, G_\gamma^{1/2}\, \mathrm{d}W_s$. So (2.4) is (A.1) with $b = b_{\mathrm{fwd}}$ and $G = G_\gamma$, and by Lemma A.1, we have

$$w_s = b_{\mathrm{fwd}} - G_\gamma \nabla \log \rho_s = -J\nabla \log p^* + G_\gamma \nabla \log p^* - G_\gamma \nabla \log \rho_s = -J\nabla \log p^* - G_\gamma \nabla \log \frac{\rho_s}{p^*}.$$

Then $v_t^q = -w_{T-t}$ is (2.6), and (A.4) with $G = G_\gamma$ is (2.8). □

Lemma 2.1 (KL derivative for two continuity equations)**.** *Suppose $\partial_t q_t = -\nabla \cdot (q_t v_t^q)$ and $\partial_t p_t = -\nabla \cdot (p_t v_t^p)$, and let $r_t = \log(p_t/q_t)$.*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \,\|\, q_t) = \int p_t \, \nabla r_t \cdot (v_t^p - v_t^q)$$

Proof. By integration by parts and $\int \partial_t p_t = 0$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \int p_t \log \frac{p_t}{q_t} &= \int \partial_t p_t \, r_t - \int \frac{p_t}{q_t}\, \partial_t q_t \\
&= -\int \nabla \cdot (p_t v_t^p) r_t + \int \nabla \cdot (q_t v_t^q) \frac{p_t}{q_t} \\
&= \int p_t \, v_t^p \cdot \nabla r_t - \int q_t v_t^q \cdot \nabla (p_t/q_t) \\
&= \int p_t \, v_t^p \cdot \nabla r_t - \int p_t \, v_t^q \cdot \nabla r_t
\end{aligned}$$

□

Corollary 2.2 (KL dissipation along the exact reverse flows)**.** *Along the common exact reverse dynamics,*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \| q_t) = -\mathcal{F}(p_t \| q_t) \ \textit{(overdamped)}, \qquad \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \| q_t) = -\gamma\, \mathcal{F}_v(p_t \| q_t) \ \textit{(underdamped)}.$$

In particular $\mathrm{KL}(p_t \,\|\, q_t)$ is always nonincreasing in both models.

Proof. By (A.4), $v_t^p - v_t^q = -G\nabla r_t$, so Lemma 2.1 gives $\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}(p_t \,\|\, q_t) = -\int p_t \, \nabla r_t^\top G \nabla r_t$. For $G = I_d$ this is $-\mathcal{F}(p_t \,\|\, q_t)$; for $G = G_\gamma$, which acts as γ on the velocity block, it is $-\gamma \int p_t \|\nabla_v r_t\|^2 = -\gamma\, \mathcal{F}_v(p_t \,\|\, q_t)$. □

Proposition 2.3 (No forgetting along the probability-flow ODE)**.** *Let Φ_t be the flow map of the probability-flow ODE of either model, and let $p_t = (\Phi_t)_{\#} p_0$ for an arbitrary initial law p_0, and $q_t = (\Phi_t)_{\#} q_0$ are the exact reverse marginals. Then $r_t = r_0 \circ \Phi_t^{-1}$; hence for all $t \in [0, T]$*

$$\mathrm{KL}(p_t \,\|\, q_t) = \mathrm{KL}(p_0 \,\|\, q_0), \qquad \mathrm{TV}(p_t, q_t) = \mathrm{TV}(p_0, q_0),$$

and likewise for every f-divergence, while the relative score is merely carried along and rescaled by the Jacobian of the flow,

$$\nabla r_t\big(\Phi_t(x)\big) = \big(\nabla \Phi_t(x)\big)^{-\top} \nabla r_0(x), \qquad \mathcal{F}(p_t \,\|\, q_t) = \int p_0 \, \big\|(\nabla \Phi_t)^{-\top} \nabla r_0\big\|^2.$$

Proof. (D.3) with $M = I$ gives the formula for $\mathcal{F}$. For a general f-divergence the same change of variables, $(p_t \circ \Phi_t)\,|\det \nabla\Phi_t| = p_0$ together with $r_t \circ \Phi_t = r_0$, yields

$$\int p_t\, f\Big(\frac{p_t}{q_t}\Big) = \int (p_t \circ \Phi_t)\,|\det \nabla\Phi_t|\, f\big(e^{r_0}\big) = \int p_0\, f\Big(\frac{p_0}{q_0}\Big),$$

which for $f(u) = \log u$ is the KL statement, and for $f(u) = \frac{1}{2}|\frac{1}{u} - 1|$ is the TV statement. □

Proposition 2.5 (Averaged stationarity of the exact reverse flows)**.** *For every potential U, every data distribution, and every pair of initializations (p_0, q_0), with $H_0 = \mathrm{KL}(p_0 \,\|\, q_0)$,*

$$\frac{1}{T}\int_0^T \mathcal{F}(p_t \,\|\, q_t)\,\mathrm{d}t = \frac{H_0 - \mathrm{KL}(p_T \,\|\, q_T)}{T} \le \frac{H_0}{T} \qquad \textit{(overdamped)},$$

$$\frac{\gamma}{T}\int_0^T \mathcal{F}_v(p_t \,\|\, q_t)\,\mathrm{d}t = \frac{H_0 - \mathrm{KL}(p_T \,\|\, q_T)}{T} \le \frac{H_0}{T} \qquad \textit{(underdamped)}.$$

Proof. Integrate the identities of Corollary 2.2 over $[0, T]$ (under the standing conventions $t \mapsto \mathrm{KL}(p_t \,\|\, q_t)$ is absolutely continuous) and use $\mathrm{KL}(p_T \,\|\, q_T) \ge 0$. □

B Proofs for Section 3: exponential contraction

Throughout this appendix q_t is the exact reverse marginal, p_t solves the same exact reverse SDE from an arbitrary initial law p_0, and, as in Section 3, $r_t = \log(p_t/q_t)$, $g_t = \nabla r_t$, $S_t = \nabla^2 r_t$. Both contraction theorems come from one computation in the generic setting of Appendix A: a transport equation for g_t (Lemma B.1) and the dissipation identity it implies for quadratic forms of g_t (Proposition B.2).

B.1 Transport of the relative score and the dissipation identity

In the setting of Lemma A.1, define the reference generator

$$L_t f := \frac{1}{q_t}\nabla\cdot(q_t G\nabla f) = G : \nabla^2 f + a_t \cdot \nabla f, \qquad a_t := G\nabla \log q_t, \tag{B.1}$$

where $G : \nabla^2 f = \mathrm{tr}(G\nabla^2 f)$. It satisfies $\int q_t\, L_t\varphi = 0$ by the divergence theorem and the regularity assumption; when it acts on a vector, it applies componentwise. It also follows the product rule

$$L_t(\varphi\psi) = \varphi L_t\psi + \psi L_t\varphi + 2\nabla\varphi^\top G\nabla\psi. \tag{B.2}$$

Indeed, since $\nabla(\varphi\psi) = \psi\nabla\varphi + \varphi\nabla\psi$ and $\nabla\cdot(\psi V) = \psi\,\nabla\cdot V + \nabla\psi\cdot V$ for any vector field V,

$$\begin{aligned} q_t\, L_t(\varphi\psi) &= \nabla\cdot\big(\psi\, q_t G\nabla\varphi\big) + \nabla\cdot\big(\varphi\, q_t G\nabla\psi\big) \\ &= \psi\,\nabla\cdot(q_t G\nabla\varphi) + \varphi\,\nabla\cdot(q_t G\nabla\psi) + q_t\big(\nabla\psi^\top G\nabla\varphi + \nabla\varphi^\top G\nabla\psi\big), \end{aligned}$$

and the last bracket equals $2\nabla\varphi^\top G\nabla\psi$ because G is symmetric; dividing by q_t yields (B.2).

Lemma B.1 (Transport of the relative density and the relative score)**.** *In the setting of Lemma A.1,*

$$(\partial_t + v_t^q\cdot\nabla)r_t = L_t r_t + g_t^\top G g_t, \tag{B.3}$$

$$(\partial_t + v_t^q\cdot\nabla)g_t = L_t g_t + (\nabla b)^\top g_t + 2S_t G g_t. \tag{B.4}$$

Proof. Let $h_t = p_t/q_t = e^{r_t}$. Since $p_t\nabla r_t = \nabla p_t - p_t\nabla\log q_t = q_t\nabla h_t$, (A.4) gives $p_t v_t^p = h_t q_t v_t^q - q_t G\nabla h_t$, so the continuity equation of p_t can be written as

$$\partial_t(h_t q_t) = -\nabla\cdot(h_t q_t v_t^q) + \nabla\cdot(q_t G\nabla h_t).$$

Expanding the left side and the first term on the right, and canceling $h_t\partial_t q_t = -h_t\nabla\cdot(q_t v_t^q)$, leaves $q_t\partial_t h_t = -q_t v_t^q\cdot\nabla h_t + \nabla\cdot(q_t G\nabla h_t)$, that is,

$$(\partial_t + v_t^q\cdot\nabla)h_t = L_t h_t. \tag{B.5}$$

Now $h_t = e^{r_t}$. Since $\nabla e^r = e^r \nabla r$ and $\nabla^2 e^r = e^r\big(\nabla^2 r + \nabla r\,\nabla r^\top\big)$, we have

$$\begin{aligned} L_t e^r &= G : \nabla^2 e^r + a_t \cdot \nabla e^r \\ &= e^r\big(G : \nabla^2 r + \operatorname{tr}(G\,\nabla r\,\nabla r^\top)\big) + e^r\, a_t \cdot \nabla r \\ &= e^r\big(L_t r + \nabla r^\top G \nabla r\big), \end{aligned}$$

where $\operatorname{tr}(G\,\nabla r\,\nabla r^\top) = \nabla r^\top G \nabla r$. On the left side of (B.5), $(\partial_t + v_t^q \cdot \nabla)e^{r_t} = e^{r_t}(\partial_t + v_t^q \cdot \nabla) r_t$ by the chain rule, so dividing (B.5) by $h_t = e^{r_t}$ yields (B.3).

For (B.4) take the gradient of (B.3). On the left of (B.3), since ∇ and ∂_t commute,

$$\begin{aligned} \nabla\big[(\partial_t + v_t^q \cdot \nabla) r_t\big] &= \partial_t \nabla r_t + (\nabla v_t^q)^\top \nabla r_t + S_t v_t^q \\ &= \partial_t g_t + v_t^q \cdot \nabla g_t + (\nabla v_t^q)^\top g_t \\ &= (\partial_t + v_t^q \cdot \nabla) g_t + (\nabla v_t^q)^\top g_t. \end{aligned}$$

On the right of (B.3), the second-order part $G : \nabla^2$ of L_t has constant coefficients, so it commutes with ∇ and acts on g_t componentwise; and $\nabla(g_t^\top G g_t) = 2 S_t G g_t$, because $\partial_j \sum_{i,k} (g_t)_i G_{ik} (g_t)_k = \sum_{i,k} \big[(S_t)_{ji} G_{ik} (g_t)_k + (g_t)_i G_{ik} (S_t)_{jk}\big] = 2 (S_t G g_t)_j$ for symmetric S_t and G. Hence

$$\begin{aligned} \nabla\big[L_t r_t + g_t^\top G g_t\big] &= \nabla\big(G : \nabla^2 r_t\big) + \nabla(a_t \cdot \nabla r_t) + \nabla\big(g_t^\top G g_t\big) \\ &= G : \nabla^2 g_t + \big[(\nabla a_t)^\top g_t + a_t \cdot \nabla g_t\big] + 2 S_t G g_t \\ &= L_t g_t + (\nabla a_t)^\top g_t + 2 S_t G g_t. \end{aligned}$$

Equating the two expansions and moving $(\nabla v_t^q)^\top g_t$ to the right,

$$(\partial_t + v_t^q \cdot \nabla) g_t = L_t g_t + \big(\nabla(a_t - v_t^q)\big)^\top g_t + 2 S_t G g_t,$$

and by (A.2), $a_t - v_t^q = G \nabla \log q_t - (-b + G \nabla \log q_t) = b$, which is (B.4). □

The last line is the structural point of Section 3: the coefficient $(\nabla b)^\top$ of the zeroth-order term in (B.4) does not depend on q_t, hence not on the data.

Proposition B.2 (Dissipation identity for quadratic forms of the relative score)**.** *In the setting of Lemma A.1, for every constant symmetric matrix P,*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int p_t\, g_t^\top P g_t = \int p_t\, g_t^\top \big(\nabla b\, P + P\,(\nabla b)^\top\big) g_t - 2 \int p_t \operatorname{tr}\big(S_t G S_t P\big). \tag{B.6}$$

Proof. Write $F_t = g_t^\top P g_t$ and $h_t = p_t / q_t$, so that $\int p_t F_t = \int q_t\, h_t F_t$. Since q_t is transported by v_t^q, integrating its continuity equation by parts gives $\frac{\mathrm{d}}{\mathrm{d}t} \int q_t \varphi_t = \int q_t (\partial_t + v_t^q \cdot \nabla)\varphi_t$ for every time-dependent function φ_t. We apply this with $\varphi_t = h_t F_t$. Writing $D := \partial_t + v_t^q \cdot \nabla$ (acting componentwise on vector fields), the product rule, (B.5) and the chain rule for the quadratic form $F_t = g_t^\top P g_t$ give

$$\begin{aligned} D(h_t F_t) &= F_t\, D h_t + h_t\, D F_t \\ &= F_t\, L_t h_t + 2 h_t\, g_t^\top P\, D g_t, \end{aligned}$$

using that P is symmetric. Multiplying by q_t, integrating, and using $q_t h_t = p_t$ in the second term,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int p_t F_t = \int q_t\, (L_t h_t) F_t + 2 \int p_t\, g_t^\top P\, (\partial_t + v_t^q \cdot \nabla) g_t. \tag{B.7}$$

For the first term, the product rule (B.2) and $\int q_t L_t (h_t F_t) = 0$ give

$$\int q_t\, (L_t h_t) F_t = -\int q_t\, h_t L_t F_t - 2 \int q_t\, \nabla h_t^\top G \nabla F_t = -\int p_t\, L_t F_t - 2 \int p_t\, g_t^\top G \nabla F_t,$$

using $q_t \nabla h_t = p_t g_t$. Two derivatives of $F_t = \sum_{i,k} (g_t)_i P_{ik} (g_t)_k$ are needed. First, $\nabla F_t = 2 S_t P g_t$. Second, the product rule (B.2) applied to $(g_t)_i (g_t)_j$, with $\nabla (g_t)_i = \nabla \partial_i r_t = S_t e_i$ and hence $\nabla (g_t)_i^\top G \nabla (g_t)_j = (S_t G S_t)_{ij}$, yields

$$\begin{aligned} L_t F_t &= \sum_{i,j} P_{ij} \Big[(g_t)_i \, L_t (g_t)_j + (g_t)_j \, L_t (g_t)_i + 2 \, (S_t G S_t)_{ij} \Big] \\ &= 2 \, g_t^\top P \, L_t g_t + 2 \operatorname{tr}(S_t G S_t P), \end{aligned}$$

Inserting both into the previous display,

$$\int q_t \, (L_t h_t) F_t = -2 \int p_t \, g_t^\top P L_t g_t - 2 \int p_t \operatorname{tr}(S_t G S_t P) - 4 \int p_t \, g_t^\top G S_t P g_t.$$

For the second term of (B.7), (B.4) shows

$$2 \int p_t \, g_t^\top P (\partial_t + v_t^q \cdot \nabla) g_t = 2 \int p_t \, g_t^\top P L_t g_t + 2 \int p_t \, g_t^\top P (\nabla b)^\top g_t + 4 \int p_t \, g_t^\top P S_t G g_t.$$

Adding the two, the terms in $L_t g_t$ cancel, and so do the two mixed terms, because $g^\top P S G g = g^\top G S P g$ for symmetric P, S, G. What remains is (B.6), after writing $2 g^\top P (\nabla b)^\top g = g^\top (\nabla b \, P + P (\nabla b)^\top) g$. □

B.2 Overdamped

Proposition B.3 (Dissipation identity for the ordinary Fisher divergence)**.** *In the overdamped model, $\mathcal{F}(t) := \mathcal{F}(p_t \,\|\, q_t)$ satisfies* (3.1)*:*

$$\mathcal{F}'(t) = -2 \int p_t \, \|S_t\|_{\mathrm{F}}^2 - 2 \int p_t \, g_t^\top \nabla^2 U \, g_t.$$

Proof. Take $b = -\nabla U$, $G = I_d$ and $P = I_d$ in Proposition B.2: then $\nabla b \, P + P (\nabla b)^\top = -2 \nabla^2 U$ and $\operatorname{tr}(S_t S_t) = \|S_t\|_{\mathrm{F}}^2$. □

Theorem 3.2 (Overdamped Fisher contractivity)**.** *In the overdamped model, if U is convex, then $\mathcal{F}(p_t \,\|\, q_t)$ is nonincreasing; if U is m-strongly convex, then $\mathcal{F}(p_t \,\|\, q_t) \le e^{-2mt} \, \mathcal{F}(p_0 \,\|\, q_0)$ for $0 \le t \le T$.*

Proof. If $\nabla^2 U \succeq 0$, both terms in Proposition B.3 are nonpositive. If $\nabla^2 U \succeq m I_d$, the identity gives $\mathcal{F}'(t) \le -2m \mathcal{F}(t)$, and Gronwall's inequality concludes. □

B.3 Underdamped

In the underdamped model the forward drift is b_{fwd}, whose Jacobian is $M(x)$ in (2.5), and $G = G_\gamma$. Proposition B.2 therefore gives, for every constant symmetric P,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}_P(p_t \,\|\, q_t) = \int p_t \, g_t^\top \big(M(x) P + P M(x)^\top \big) g_t \, \mathrm{d}z - 2 \int p_t \operatorname{tr}\big(S_t G_\gamma S_t P \big) \, \mathrm{d}z. \tag{B.8}$$

The second term is nonpositive whenever $P \succ 0$, since $\operatorname{tr}(S_t G_\gamma S_t P) = \|G_\gamma^{1/2} S_t P^{1/2}\|_{\mathrm{F}}^2$. The first term is where the Hamiltonian coupling acts: $M(x)^\top g_t$ rotates position-gradient error into the velocity directions, and everything comes down to finding a P for which $M(x) P + P M(x)^\top$ is negative definite uniformly in x. The matrix P_γ of (3.2) does this.

Lemma B.4 (Lyapunov inequality for P_γ)**.** *Suppose $\gamma^2 \xi > L$ and let $P = P_\gamma \otimes I_d$ with P_γ, $\kappa_{\gamma,\xi}$ as in* (3.2)*. Under Assumption 3.1,*

$$M(x) P + P M(x)^\top \preceq -2 \kappa_{\gamma,\xi} \, P \qquad \text{for all } x \in \mathbb{R}^d. \tag{B.9}$$

Moreover $\det P_\gamma = \gamma^2/4 > 0$*, so $P_\gamma \succ 0$, with eigenvalues*

$$\lambda_\pm(P_\gamma) = \frac{1}{2}\Big(1 + \frac{\gamma^2}{2}\Big) \pm \frac{1}{2} \sqrt{1 + \frac{\gamma^4}{4}}. \tag{B.10}$$

Proof. Fix x and diagonalize $\nabla^2 U(x) = O^\top \operatorname{diag}(\eta_1, \dots, \eta_d)\, O$ with O orthogonal and $\eta_i \in [m, L]$. Conjugation by $\operatorname{diag}(O, O)$ leaves $P = P_\gamma \otimes I_d$ unchanged, because each $d \times d$ block of P is a multiple of I_d, and turns $M(x)$ into $\left(\begin{smallmatrix} 0 & \xi I_d \\ -\operatorname{diag}(\eta_i) & -\gamma\xi I_d \end{smallmatrix}\right)$, which splits into d two-dimensional systems with matrices

$$M(\eta) := \begin{pmatrix} 0 & \xi \\ -\eta & -\gamma\xi \end{pmatrix}, \qquad \eta \in [m, L].$$

So it suffices to prove $M(\eta)P_\gamma + P_\gamma M(\eta)^\top \preceq -2\kappa_{\gamma,\xi} P_\gamma$ for every $\eta \in [m, L]$.

In the coordinates $(x,\ x + \frac{2}{\gamma}v)$, i.e. after conjugation by $T = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 2/\gamma \end{smallmatrix}\right)$, the matrix $M(\eta)$ becomes

$$A(\eta) := TM(\eta)T^{-1} = \begin{pmatrix} -\dfrac{\gamma\xi}{2} & \dfrac{\gamma\xi}{2} \\ \dfrac{\gamma\xi}{2} - \dfrac{2\eta}{\gamma} & -\dfrac{\gamma\xi}{2} \end{pmatrix}, \qquad \frac{A(\eta) + A(\eta)^\top}{2} = \begin{pmatrix} -\dfrac{\gamma\xi}{2} & s \\ s & -\dfrac{\gamma\xi}{2} \end{pmatrix},$$

with $s = (\gamma^2\xi - 2\eta)/2\gamma$. The symmetric part has eigenvalues $-\gamma\xi/2 \pm |s|$, and according to the sign of s,

$$-\frac{\gamma\xi}{2} + |s| = -\min\left(\frac{\eta}{\gamma},\ \gamma\xi - \frac{\eta}{\gamma}\right) \le -\min\left(\frac{m}{\gamma},\ \gamma\xi - \frac{L}{\gamma}\right) = -\kappa_{\gamma,\xi} \qquad (\eta \in [m, L]),$$

which is negative precisely when $m > 0$ and $\gamma^2\xi > L$. Thus $A(\eta) + A(\eta)^\top \preceq -2\kappa_{\gamma,\xi} I$. Multiplying on the left by $T^\top$ and on the right by T, and writing $Q := T^\top T$, this reads $M(\eta)^\top Q + QM(\eta) \preceq -2\kappa_{\gamma,\xi} Q$; multiplying on both sides by Q^{-1} gives $M(\eta)Q^{-1} + Q^{-1}M(\eta)^\top \preceq -2\kappa_{\gamma,\xi} Q^{-1}$. Finally

$$Q = \begin{pmatrix} 2 & 2/\gamma \\ 2/\gamma & 4/\gamma^2 \end{pmatrix}, \qquad Q^{-1} = \begin{pmatrix} 1 & -\gamma/2 \\ -\gamma/2 & \gamma^2/2 \end{pmatrix} = P_\gamma,$$

whose trace and determinant are $1 + \gamma^2/2$ and $\gamma^2/4$; this implies (B.10). □

Theorem 3.4 (Underdamped modified Fisher contractivity)**.** *Assume the underdamped model* (2.4)*, Assumption 3.1, and* $\gamma^2\xi > L$*. With* $P = P_\gamma \otimes I_d$ *as in* (3.2)*,* $\mathcal{F}_P(p_t \,\|\, q_t) \le e^{-2t\kappa_{\gamma,\xi}}\, \mathcal{F}_P(p_0 \,\|\, q_0)$.

Proof. In (B.8) with $P = P_\gamma \otimes I_d \succ 0$, the second term is nonpositive and, by Lemma B.4, the first is at most $-2\kappa_{\gamma,\xi} \int p_t\, g_t^\top P g_t = -2\kappa_{\gamma,\xi} \mathcal{F}_P(p_t \,\|\, q_t)$. Gronwall's inequality concludes. If U is merely convex ($m = 0$), then we have $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_P \le 0$. □

C Proofs for Section 4: the overdamped sampler

Throughout this appendix, $q_t = \rho_{T-t}$ and $s_t = \nabla \log q_t$ are the exact reverse marginals and scores of the forward OU process (4.1), X_t is the interpolated numerical process (4.2) on the uniform grid $t_k = kh$, $h = T/N$, and $\hat{p}_t$ is its law. Assumption 4.2 supplies the positive densities and the finite initial entropy H_0 required by the standing conventions, and $\lambda_0 = \min\{\lambda, \frac{d}{4}\}$, $\mathcal{M} = \frac{d}{\lambda_0}(H_0 + \log M + d)$ are as in Section 4.

The proof of Theorem 4.5 runs as follows. Freezing the score costs a source term in the entropy dissipation: for $t \in [t_k, t_{k+1}]$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \operatorname{KL}(\hat{p}_t \,\|\, q_t) \le -\frac{1}{2}\mathcal{F}(\hat{p}_t \,\|\, q_t) + 2\,\mathbb{E}\|s_{t_k}(X_{t_k}) - s_t(X_t)\|^2 \tag{C.1}$$

(Proposition C.3). The source term compares the frozen score with the current one. It has a spatial part, from the increment $X_t - X_{t_k}$, which Lipschitzness and the explicit step (4.3) control, and a temporal part, from $s_{t_k} - s_t$ at a fixed point, which we control through the OU semigroup (Lemma C.5; compare Chen et al., 2023b, Lemma 17). Both can be controlled by second moments under the numerical law $\hat{p}_t$. Assumption 4.3 bounds exponential moments of the exact marginals q_t, and the Gibbs variational inequality converts them into second moments of $\hat{p}_t$ as long as $\operatorname{KL}(\hat{p}_t \,\|\, q_t)$ stays bounded (Lemma C.8). Integrating (C.1) with this bound closes the entropy bootstrap and results in our main theorem.

C.1 KL DISSIPATION FOR THE NUMERICAL MARGINAL

Lemma C.1 (Marginal Markov projection). *For $t \in [t_k, t_{k+1}]$ let $\bar{s}_t(x) := \mathbb{E}[s_{t_k}(X_{t_k}) \mid X_t = x]$. Then*

$$\partial_t \hat{p}_t = -\nabla \cdot \big(\hat{p}_t[x + 2\bar{s}_t(x)]\big) + \Delta \hat{p}_t. \tag{C.2}$$

Proof. For a smooth compactly supported φ, Itô's formula applied to (4.2) gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\,\varphi(X_t) &= \mathbb{E}\big[\,\langle \nabla\varphi(X_t), X_t + 2s_{t_k}(X_{t_k})\rangle + \Delta\varphi(X_t)\big] \\ &= \mathbb{E}\big[\,\langle \nabla\varphi(X_t), X_t + 2\bar{s}_t(X_t)\rangle + \Delta\varphi(X_t)\big],\end{aligned}$$

by conditioning on X_t in the frozen-score term. This is the weak form of (C.2). □

Lemma C.2 (Relative entropy derivative for two Fokker–Planck equations). *If $\partial_t \mu_t = -\nabla \cdot (\mu_t b_t^\mu) + \Delta\mu_t$ and $\partial_t \nu_t = -\nabla \cdot (\nu_t b_t^\nu) + \Delta\nu_t$, then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(\mu_t \,\|\, \nu_t) = -\mathcal{F}(\mu_t \,\|\, \nu_t) + \mathbb{E}_{\mu_t}\Big\langle \nabla \log \frac{\mu_t}{\nu_t},\, b_t^\mu - b_t^\nu \Big\rangle. \tag{C.3}$$

Proof. Both equations are continuity equations, with velocities $v_t^\mu = b_t^\mu - \nabla \log \mu_t$ and $v_t^\nu = b_t^\nu - \nabla \log \nu_t$. Thus $v_t^\mu - v_t^\nu = b_t^\mu - b_t^\nu - \nabla \log(\mu_t/\nu_t)$, and Lemma 2.1 completes the proof. □

Proposition C.3 (Master inequality). *For $t \in [t_k, t_{k+1}]$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(\hat{p}_t \,\|\, q_t) \le -\frac{1}{2}\mathcal{F}(\hat{p}_t \,\|\, q_t) + 2\,\mathbb{E}\|s_{t_k}(X_{t_k}) - s_t(X_t)\|^2.$$

Proof. The exact marginal satisfies $\partial_t q_t = -\nabla \cdot (q_t[x + 2s_t]) + \Delta q_t$, the Fokker–Planck equation of the exact reverse SDE (Lemma A.1 with $b = -x$, $G = I_d$). Lemma C.2 with $\mu_t = \hat{p}_t$, $\nu_t = q_t$ and Lemma C.1 give

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(\hat{p}_t \,\|\, q_t) = -\mathcal{F}(\hat{p}_t \,\|\, q_t) + 2\,\mathbb{E}_{\hat{p}_t}\Big\langle \nabla \log \frac{\hat{p}_t}{q_t},\, \bar{s}_t - s_t \Big\rangle \le -\frac{1}{2}\mathcal{F}(\hat{p}_t \,\|\, q_t) + 2\,\mathbb{E}_{\hat{p}_t}\|\bar{s}_t - s_t\|^2,$$

by $2\langle g, e\rangle \le \frac{1}{2}\|g\|^2 + 2\|e\|^2$. Finally $\bar{s}_t(X_t) - s_t(X_t) = \mathbb{E}[s_{t_k}(X_{t_k}) - s_t(X_t) \mid X_t]$, and Jensen's inequality for the conditional expectation bounds $\mathbb{E}\|\bar{s}_t(X_t) - s_t(X_t)\|^2$ by $\mathbb{E}\|s_{t_k}(X_{t_k}) - s_t(X_t)\|^2$. □

C.2 TEMPORAL SCORE VARIATION FROM THE OU SEMIGROUP

Along the forward flow (4.1), $Y_{s+\tau} \overset{d}{=} e^{-\tau} Y_s + \sqrt{1 - e^{-2\tau}}\, G$ with $G \sim N(0, I_d)$ independent of Y_s. Read at $s = T - t$ and $s + \tau = T - t_k$, i.e. at $t = t_k + \tau$, this expresses the earlier reverse-time marginal as a scaled and then smoothed version of the later one:

$$q_{t_k} = \big((aI_d)_{\#} q_t\big) * N(0, \sigma^2 I_d), \qquad a = e^{-\tau}, \quad \sigma^2 = 1 - e^{-2\tau}. \tag{C.4}$$

We first quantify what the smoothing does to a score.

Lemma C.4 (Score perturbation by Gaussian smoothing). *Let π be a smooth positive density on $\mathbb{R}^d$ whose score $u = \nabla \log \pi$ is L_π-Lipschitz, and let $\pi_\sigma = \pi * N(0, \sigma^2 I_d)$. If $L_\pi \sigma^2 \le \frac{1}{2}$, then for every x,*

$$\|\nabla \log \pi_\sigma(x) - u(x)\| \le L_\pi \sqrt{2d\sigma^2} + 2L_\pi \sigma^2\, \|u(x)\|. \tag{C.5}$$

Proof. Fix x and let μ_x be the probability density on $\mathbb{R}^d$ proportional to $\pi(y)\exp(-\|x - y\|^2/2\sigma^2)$, the conditional law of $Y \sim \pi$ given $Y + \sigma G = x$. Since $\nabla_x e^{-\|x-y\|^2/2\sigma^2} = -\nabla_y e^{-\|x-y\|^2/2\sigma^2}$, an integration by parts in y gives

$$\nabla \pi_\sigma(x) \propto \int \pi(y)\, \nabla_x e^{-\|x-y\|^2/2\sigma^2}\, \mathrm{d}y = \int \nabla\pi(y)\, e^{-\|x-y\|^2/2\sigma^2}\, \mathrm{d}y,$$

and dividing by $\pi_\sigma(x)$, with $\nabla\pi = \pi u$, gives $\nabla \log \pi_\sigma(x) = \mathbb{E}_{\mu_x} u(Y)$. Hence $\|\nabla \log \pi_\sigma(x) - u(x)\| \le L_\pi\, \mathbb{E}_{\mu_x}\|Y - x\|$, and it remains to show that μ_x concentrates near x.

Write $\mu_x \propto e^{-V_x}$ with $V_x(y) = -\log \pi(y) + \|x-y\|^2/2\sigma^2$. As $\nabla^2(-\log\pi) = -\nabla u \succeq -L_\pi I_d$, we have $\nabla^2 V_x \succeq (\sigma^{-2} - L_\pi) I_d \succeq (2\sigma^2)^{-1} I_d$, so V_x has a unique minimizer $y_\star$. At $y_\star$, we know

$$-u(y_\star) + (y_\star - x)/\sigma^2 = 0, \quad \text{i.e.} \quad y_\star - x = \sigma^2 u(y_\star),$$

and therefore

$$\|y_\star - x\| \le \sigma^2 \|u(x)\| + L_\pi \sigma^2 \|y_\star - x\|, \qquad \text{hence} \qquad \|y_\star - x\| \le 2\sigma^2 \|u(x)\|.$$

Next we show that μ_x concentrates around $y_\star$. Since $\nabla \mu_x = -\mu_x \nabla V_x$, an integration by parts shows,

$$\mathbb{E}_{\mu_x} \langle Y - y_\star, \nabla V_x(Y)\rangle = -\int \langle y - y_\star, \nabla \mu_x(y)\rangle \, \mathrm{d}y = \int \mu_x(y)\, \nabla \cdot (y - y_\star)\, \mathrm{d}y = d.$$

On the other hand, we have $\nabla^2 V_x \succeq (2\sigma^2)^{-1} I_d$ and $\nabla V_x(y_\star) = 0$, so

$$\langle y - y_\star, \nabla V_x(y)\rangle = \Big\langle y - y_\star, \int_0^1 \nabla^2 V_x\big(y_\star + s(y - y_\star)\big)(y - y_\star)\, \mathrm{d}s \Big\rangle \ge \frac{\|y - y_\star\|^2}{2\sigma^2}.$$

Therefore $\mathbb{E}_{\mu_x}\|Y - y_\star\|^2 \le 2d\sigma^2$. Altogether, by the triangle inequality and Jensen's inequality,

$$\mathbb{E}_{\mu_x}\|Y - x\| \le \big(\mathbb{E}_{\mu_x}\|Y - y_\star\|^2\big)^{1/2} + \|y_\star - x\| \le \sqrt{2d\sigma^2} + 2\sigma^2\|u(x)\|.$$

Substituting this bound into $\|\nabla \log \pi_\sigma(x) - u(x)\| \le L_\pi \, \mathbb{E}_{\mu_x}\|Y - x\|$ completes the proof. □

Lemma C.5 (Temporal OU score perturbation)**.** *Under Assumption 4.1, let $t = t_k + \tau$ with $0 \le \tau \le h$. If $h \le c/(1 + L_s)$ with $c \le \frac{1}{8}$, then for every $x \in \mathbb{R}^d$,*

$$\|s_{t_k}(x) - s_t(x)\|^2 \lesssim L_s^2 d\tau + L_s^2 \tau^2 \|x\|^2 + (1 + L_s)^2 \tau^2 \|s_t(x)\|^2. \tag{C.6}$$

Proof. Let $\pi = (aI_d)_{\#} q_t$ be the scaled density in (C.4) and $u = \nabla \log \pi$. Since $\pi(x) = a^{-d} q_t(x/a)$,

$$u(x) = a^{-1} s_t(x/a), \tag{C.7}$$

so u is L_π-Lipschitz with $L_\pi = L_s a^{-2}$. For $\tau \le h \le c/(1 + L_s)$ we have $a^{-1} = e^\tau \le 2$, $e^\tau - 1 \le 2\tau$, $\sigma^2 = 1 - e^{-2\tau} \le 2\tau$, and $L_\pi \sigma^2 = L_s(e^{2\tau} - 1) \le 4 L_s \tau \le \frac{1}{2}$. By Lemma C.4, we get

$$\|s_{t_k}(x) - u(x)\| \lesssim L_s \sqrt{d\tau} + L_s \tau \|u(x)\|. \tag{C.8}$$

It remains to compare u with s_t. By (C.7), $u(x) - s_t(x) = (a^{-1} - 1) s_t(x/a) + \big(s_t(x/a) - s_t(x)\big)$, and $\|x/a - x\| = (e^\tau - 1)\|x\| \le 2\tau\|x\|$, so by Lipschitzness

$$\|u(x) - s_t(x)\| \le 2\tau\big(\|s_t(x)\| + 2L_s\tau\|x\|\big) + 2L_s\tau\|x\| \lesssim \tau\|s_t(x)\| + L_s\tau\|x\|. \tag{C.9}$$

In particular $\|u(x)\| \lesssim \|s_t(x)\| + L_s\tau\|x\|$. Substituting this into (C.8), adding (C.9), and using $L_s\tau \le 1$,

$$\|s_{t_k}(x) - s_t(x)\| \lesssim L_s\sqrt{d\tau} + L_s\tau\|x\| + (1 + L_s)\tau\|s_t(x)\|,$$

and squaring both sides completes the proof. □

C.3 Second moments of the numerical law

Lemma C.6 (Exponential moments along the OU flow)**.** *Under Assumption 4.3, with $\lambda_0 = \min\{\lambda, \frac{d}{4}\}$,*

$$\mathbb{E}_{q_t} \exp\Big(\frac{\lambda_0}{d}\|X\|^2\Big) \le e^{2\lambda_0} M \qquad \textit{for every } t \in [0, T]. \tag{C.10}$$

Proof. Every q_t is a forward marginal ρ_s, and $Y_s \overset{d}{=} aY_0 + \sigma G$ with $a = e^{-s}$, $\sigma^2 = 1 - e^{-2s}$, and $G \sim N(0, I_d)$ independent of $Y_0 \sim \rho_0$. Conditionally on Y_0, Y_s is Gaussian with mean aY_0 and covariance $\sigma^2 I_d$, and for $Z \sim N(m, \sigma^2 I_d)$ and $2\theta\sigma^2 < 1$,

$$\mathbb{E}e^{\theta\|Z\|^2} = (1 - 2\theta\sigma^2)^{-d/2} \exp\Big(\frac{\theta\|m\|^2}{1 - 2\theta\sigma^2}\Big).$$

Take $\theta = \lambda_0/d \le \frac{1}{4}$ and put $x := 2\theta\sigma^2 \le 2\lambda_0/d \le \frac{1}{2}$. Since $-\log(1 - x) \le 2x$ on $[0, \frac{1}{2}]$, the prefactor is $(1 - x)^{-d/2} \le e^{dx} \le e^{2\lambda_0}$, and $a^2 \le 1 - 2\theta\sigma^2$ because $a^2 = 1 - \sigma^2$ and $2\theta \le 1$. Hence

$$\mathbb{E}e^{(\lambda_0/d)\|Y_s\|^2} \le e^{2\lambda_0}\, \mathbb{E}\exp\Big(\frac{\lambda_0 a^2}{d(1 - 2\theta\sigma^2)}\|Y_0\|^2\Big) \le e^{2\lambda_0}\, \mathbb{E}e^{(\lambda_0/d)\|Y_0\|^2} \le e^{2\lambda_0} M,$$

the last step because $\lambda_0 \le \lambda$. □

Lemma C.7 (Gibbs variational inequality). *For probability measures $Q \ll P$ and every measurable f with $\mathbb{E}_P e^f < \infty$, $\mathbb{E}_Q f \le \mathrm{KL}(Q \,\|\, P) + \log \mathbb{E}_P e^f$.*

Proof. With $\varrho = \mathrm{d}Q/\mathrm{d}P$, apply Jensen's inequality

$$\mathbb{E}_Q f - \mathrm{KL}(Q \,\|\, P) = \mathbb{E}_Q \log(e^f/\varrho) \le \log \mathbb{E}_Q(e^f/\varrho) \le \log \mathbb{E}_P e^f.$$

□

Lemma C.8 (Moment bounds under the KL bootstrap). *Assume Assumptions 4.1 and 4.3. If, at some time t,*

$$\mathrm{KL}(\hat{p}_t \,\|\, q_t) \le H_0 + d, \tag{C.11}$$

then

$$\mathbb{E}_{\hat{p}_t}\|X\|^2 \lesssim \mathcal{M}, \qquad \mathbb{E}_{\hat{p}_t}\|s_t(X)\|^2 \lesssim L_s^2\mathcal{M}, \qquad \mathbb{E}_{\hat{p}_t}\|X + 2s_t(X)\|^2 \lesssim (1+L_s)^2\mathcal{M}. \tag{C.12}$$

Proof. Lemma C.7 with $f(x) = \frac{\lambda_0}{d}\|x\|^2$, Lemma C.6 and (C.11) give

$$\frac{\lambda_0}{d}\,\mathbb{E}_{\hat{p}_t}\|X\|^2 \le \mathrm{KL}(\hat{p}_t \,\|\, q_t) + \log \mathbb{E}_{q_t} e^{(\lambda_0/d)\|X\|^2} \le H_0 + d + \log M + 2\lambda_0 \le \tfrac{3}{2}\,\frac{\lambda_0}{d}\,\mathcal{M}$$

which is the first bound; likewise we have

$$\frac{\lambda_0}{d}\mathbb{E}_{q_t}\|Y\|^2 \le \log \mathbb{E}_{q_t} e^{(\lambda_0/d)\|Y\|^2} \le \log M + 2\lambda_0,$$

so $\mathbb{E}_{q_t}\|Y\|^2 \le \frac{d}{\lambda_0}(\log M + \frac{d}{2}) \le \mathcal{M}$. For the score, $\mathbb{E}_{q_t} s_t(Y) = \int \nabla q_t = 0$, so for every x,

$$\|s_t(x)\| = \big\|\mathbb{E}_{Y\sim q_t}[s_t(x) - s_t(Y)]\big\| \le L_s\,\mathbb{E}_{q_t}\|x - Y\| \le L_s\big(\|x\| + \mathbb{E}_{q_t}\|Y\|\big),$$

hence $\|s_t(x)\|^2 \le 2L_s^2(\|x\|^2 + \mathcal{M})$. Averaging over $x \sim \hat{p}_t$ yields the second bound, and $\|x + 2s_t(x)\|^2 \le 2\|x\|^2 + 8\|s_t(x)\|^2$ the third. □

C.4 One-step score error

Lemma C.9 (Local score mismatch). *Assume Assumptions 4.1 and 4.3, and $h \le c/(1+L_s)$ with c as in Lemma C.5. Let $t = t_k + \tau$ with $0 \le \tau \le h$, and suppose $\mathrm{KL}(\hat{p}_s \,\|\, q_s) \le H_0 + d$ for all $s \le t$. Then*

$$\mathbb{E}\|s_{t_k}(X_{t_k}) - s_t(X_t)\|^2 \lesssim L_s^2 d\tau + L_s^2(1+L_s)^2\mathcal{M}\tau^2. \tag{C.13}$$

Proof. We split

$$\mathbb{E}\|s_{t_k}(X_{t_k}) - s_t(X_t)\|^2 \le 2\,\mathbb{E}\|s_{t_k}(X_{t_k}) - s_{t_k}(X_t)\|^2 + 2\,\mathbb{E}\|s_{t_k}(X_t) - s_t(X_t)\|^2. \tag{C.14}$$

For the first term, by Lipschitzness we have

$$\mathbb{E}\|s_{t_k}(X_{t_k}) - s_{t_k}(X_t)\|^2 \le L_s^2\,\mathbb{E}\|X_t - X_{t_k}\|^2,$$

and by (4.3)

$$X_t - X_{t_k} = (e^\tau - 1)\{X_{t_k} + 2s_{t_k}(X_{t_k})\} + \sqrt{e^{2\tau} - 1}\,\zeta$$

with $\zeta \sim N(0, I_d)$ independent of X_{t_k}, so the cross term has zero mean. Since $e^\tau - 1 \le 2\tau$ and $e^{2\tau} - 1 \le 4\tau$ for $\tau \le h \le \frac{1}{8}$, Lemma C.8 at time t_k implies

$$\mathbb{E}\|X_t - X_{t_k}\|^2 \lesssim (1+L_s)^2\mathcal{M}\tau^2 + d\tau. \tag{C.15}$$

For the second term, take $x = X_t$ in Lemma C.5, average over $X_t \sim \hat{p}_t$, and use Lemma C.8 at time t:

$$\begin{aligned}\mathbb{E}\|s_{t_k}(X_t) - s_t(X_t)\|^2 &\lesssim L_s^2 d\tau + L_s^2\tau^2\,\mathbb{E}\|X_t\|^2 + (1+L_s)^2\tau^2\,\mathbb{E}\|s_t(X_t)\|^2 \\ &\lesssim L_s^2 d\tau + L_s^2(1+L_s)^2\mathcal{M}\tau^2.\end{aligned}$$

Inserting both bounds into (C.14) completes the proof. □

C.5 Proof of Theorem 4.5

Theorem 4.5 (Averaged Fisher stationarity, overdamped)**.** *Suppose Assumptions 4.1–4.3 hold and $H_0 < \infty$. There are universal constants c such that if $h \le c/(1+L_s)$ and $T\big[L_s^2 dh + L_s^2(1+L_s)^2\mathcal{M}h^2\big] \le cd$, then $\sup_{0\le t\le T} \mathrm{KL}(\hat{p}_t \,\|\, q_t) < H_0 + d$ and*

$$\frac{1}{T}\int_0^T \mathcal{F}(\hat{p}_t \,\|\, q_t)\,\mathrm{d}t \;\lesssim\; \frac{H_0}{T} + L_s^2 dh + L_s^2(1+L_s)^2\mathcal{M}h^2.$$

Proof. Write $\mathrm{H}(t) := \mathrm{KL}(\hat{p}_t \,\|\, q_t)$, a continuous function of t under the standing conventions. Whenever $\mathrm{H}(s) \le H_0 + d$ for all $s \le t$, Proposition C.3 and Lemma C.9 give, for $t \in [t_k, t_{k+1}]$ and $\tau = t - t_k$,

$$\mathrm{H}'(t) \le -\frac{1}{2}\mathcal{F}(\hat{p}_t \,\|\, q_t) + C\big[L_s^2 d\tau + L_s^2(1+L_s)^2\mathcal{M}\tau^2\big]. \tag{C.16}$$

The source term integrates to $C[L_s^2 dh^2/2 + L_s^2(1+L_s)^2\mathcal{M}h^3/3]$ over a full step, to less over a partial one, and there are T/h steps. So, for every $t \le T$ up to which (C.16) is in force,

$$\mathrm{H}(t) + \frac{1}{2}\int_0^t \mathcal{F}(\hat{p}_s \,\|\, q_s)\,\mathrm{d}s \le H_0 + CT\big[L_s^2 dh + L_s^2(1+L_s)^2\mathcal{M}h^2\big] \le H_0 + \frac{d}{2}, \tag{C.17}$$

where the last inequality is the second condition of the theorem, once $c \le 1/2C$.

The bootstrap closes by continuity. Let $t_\star := \sup\{t \in [0,T] : \mathrm{H}(s) \le H_0 + d \text{ for all } s \le t\}$; since $\mathrm{H}(0) = H_0$, $t_\star > 0$ and $\mathrm{H} \le H_0 + d$ on $[0, t_\star]$, so (C.17) applies up to $t_\star$ and gives $\mathrm{H}(t_\star) \le H_0 + d/2$. Were $t_\star < T$, H would stay below $H_0 + d$ slightly beyond $t_\star$ by continuity, contradicting its definition. Hence $t_\star = T$, (C.16) holds on all of $[0,T]$. The inequality (C.17) at $t = T$ implies

$$\frac{1}{2}\int_0^T \mathcal{F}(\hat{p}_t \,\|\, q_t)\,\mathrm{d}t \le H_0 + CT\big[L_s^2 dh + L_s^2(1+L_s)^2\mathcal{M}h^2\big].$$

Dividing by T proves (4.5). □

D Proofs for Section 5: the underdamped sampler

Throughout this appendix, $\hat{p}_i$ and $\hat{q}_i$ are the grid laws of the numerical chain started from $\hat{p}_0$ and from the exact initialization $\hat{q}_0 = \rho_T$ respectively, $\hat{p}_{i,\tau}$ and $\hat{q}_{i,\tau}$ are their within-step interpolations (5.8), and C, E, Π_v, D are the matrices of (5.1). Assumption 5.3 supplies the positive densities and the finite H_0, F_0 of the standing conventions; since $\hat{q}_0 = q_0$, $H_0 = \mathrm{KL}(\hat{p}_0 \,\|\, \hat{q}_0)$ and $F_0 = \mathcal{F}(\hat{p}_0 \,\|\, \hat{q}_0)$.

The argument follows the two-factor form (5.7) of the integrator. The kick is a deterministic map applied to both laws, so it preserves relative entropy and transforms the relative score by the inverse transpose of its Jacobian (Lemma D.2). The score-free leg is a linear diffusion, along which relative entropy dissipates the velocity part of the relative score, and a twisted Fisher divergence with a time-dependent metric can dissipate the full relative score, at the price of a term involving the mixed Hessian of the reference law, which Assumption 5.2 controls (Proposition D.3). Appendix D.4 combines the two into a Lyapunov functional, lists the four properties (M1)–(M4) that the moving metric must have, and shows that they make the functional decrease at least as fast as $\mathcal{F}(\hat{p}_{i,\tau} \,\|\, \hat{q}_{i,\tau})$ along each leg and match at the grid points, so that the bound telescopes (Proposition D.4). Appendix D.5 motivates and defines the metric, verifies (M1)–(M3), and completes the proof of the theorem; the remaining property (M4), a 2×2 matrix inequality, is verified in Appendix D.6.

D.1 Reverse dynamics and the integrator

Lemma D.1 (Exact two-factor decomposition of the integrator)**.** *Let K_i be the transition kernel of the update* (5.4)*. Then for every input law μ,*

$$\mu\mathsf{K}_i = \mathcal{R}_{h_i}\big((\Psi_{i,h_i})_{\#}\mu\big), \tag{D.1}$$

with the kick $\Psi_{i,h}$ of (5.5) *and the score-free semigroup $\mathcal{R}_\tau$ of* (5.6)*.*

Proof. Fix z and let $f := 4Es_i(z)$, a constant vector. The frozen-score step from z is the time-h value of the linear SDE $\mathrm{d}X = (CX + f)\,\mathrm{d}t + 2E\,\mathrm{d}\bar{B}$, $X_0 = z$, which by variation of constants is

$$X_h = e^{hC}\Big(z + \int_0^h e^{-rC} f\,\mathrm{d}r\Big) + \int_0^h e^{(h-u)C} 2E\,\mathrm{d}\bar{B}_u = e^{hC}\big(z + B_h s_i(z)\big) + \zeta,$$

where $\zeta \sim \mathcal{N}\big(0, \int_0^h e^{vC} D e^{vC^\top}\,\mathrm{d}v\big) = \mathcal{N}(0, Q_h)$. This is $\mathcal{R}_h$ applied to the kicked point $\Psi_{i,h}(z)$. □

D.2 Two laws under a common map

Lemma D.2 (Common deterministic pushforward). *Let $\mathsf{T}\colon \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ be a C^1 diffeomorphism, and let $\mu^+ = \mathsf{T}_{\#}\mu$, $\nu^+ = \mathsf{T}_{\#}\nu$ for positive densities μ, ν. With $r = \log(\mu/\nu)$ and $r^+ = \log(\mu^+/\nu^+)$,*

$$r^+ \circ \mathsf{T} = r, \qquad \mathrm{KL}(\mu^+ \,\|\, \nu^+) = \mathrm{KL}(\mu \,\|\, \nu), \qquad (\nabla r^+) \circ \mathsf{T} = (\nabla \mathsf{T})^{-\top} \nabla r. \tag{D.2}$$

Consequently, for every symmetric matrix M,

$$\mathcal{F}_M(\mu^+ \,\|\, \nu^+) = \int \mu \Big\langle \nabla r,\, (\nabla \mathsf{T})^{-1} M (\nabla \mathsf{T})^{-\top} \nabla r \Big\rangle. \tag{D.3}$$

Proof. By the change of variables formula, $\mu^+ \circ \mathsf{T} = \mu/|\det \nabla \mathsf{T}|$ and likewise for ν; the Jacobians cancel in the ratio, which is the first identity. The second is obtained by integrating r^+. Differentiating the first by the chain rule gives the third, and inserting the third into $\mathcal{F}_M(\mu^+ \,\|\, \nu^+) = \int \mu\, \langle (\nabla r^+) \circ \mathsf{T}, M\, (\nabla r^+) \circ \mathsf{T} \rangle$ implies (D.3). □

D.3 Identities on the score-free leg

On the score-free leg both interpolated densities in (5.8) solve

$$\partial_\tau \nu_\tau = -\nabla \cdot (\nu_\tau\, Cz) + 2\Delta_v \nu_\tau, \tag{D.4}$$

the Fokker–Planck equation of (5.6). For a pair (μ_τ, ν_τ) of solutions put

$$r_\tau = \log \frac{\mu_\tau}{\nu_\tau}, \qquad g_\tau = \nabla r_\tau, \qquad g_{v,\tau} = E^\top g_\tau = \nabla_v r_\tau,$$

and let $\nabla_z \nabla_v r_\tau$ and $\nabla_z \nabla_v \log \nu_\tau$ denote the $2d \times d$ matrices of mixed second derivatives, the latter being the mixed Hessian of Assumption 5.2. Unlike the exact reverse flow of Appendix B, whose drift contains the reference score, the score-free leg is oblivious to the pair it transports; the price is the last term of (D.6) below, in which the reference law enters through its mixed Hessian.

Proposition D.3 (Entropy and twisted-Fisher identities on the score-free leg). *Let $(\mu_\tau)_\tau$ and $(\nu_\tau)_\tau$ be positive densities solving* (D.4)*. Then*

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \mathrm{KL}(\mu_\tau \,\|\, \nu_\tau) = -2\, \mathcal{F}_v(\mu_\tau \,\|\, \nu_\tau), \tag{D.5}$$

and for every C^1 family $\tau \mapsto M(\tau)$ of symmetric matrices,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau} \mathcal{F}_{M(\tau)}(\mu_\tau \,\|\, \nu_\tau) = &\int \mu_\tau \Big\langle g_\tau, \big(\dot{M}(\tau) - CM(\tau) - M(\tau)C^\top\big) g_\tau \Big\rangle \\ &- 4 \int \mu_\tau\, \mathrm{tr}\Big((\nabla_z \nabla_v r_\tau)^\top M(\tau)\, \nabla_z \nabla_v r_\tau \Big) \\ &+ 8 \int \mu_\tau \Big\langle M(\tau) g_\tau,\, (\nabla_z \nabla_v \log \nu_\tau)\, g_{v,\tau} \Big\rangle. \end{aligned} \tag{D.6}$$

Proof. Equation (D.4) is the continuity equation with velocity field

$$u_\tau^\nu = Cz - 2E \nabla_v \log \nu_\tau, \qquad \text{so that} \qquad u_\tau^\mu - u_\tau^\nu = -2E g_{v,\tau},$$

and Lemma 2.1 implies

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \mathrm{KL}(\mu_\tau \,\|\, \nu_\tau) = \int \mu_\tau\, \langle g_\tau, -2E g_{v,\tau} \rangle = -2 \int \mu_\tau \|g_{v,\tau}\|^2 = -2\, \mathcal{F}_v(\mu_\tau \,\|\, \nu_\tau).$$

For the second identity, divide (D.4) by ν_τ. Using

$$\nabla \cdot (\nu_\tau C z) = \nu_\tau \operatorname{tr} C + Cz \cdot \nabla \nu_\tau, \qquad \frac{\Delta_v \nu_\tau}{\nu_\tau} = \Delta_v \log \nu_\tau + \|\nabla_v \log \nu_\tau\|^2,$$

we get

$$\partial_\tau \log \nu_\tau = -\operatorname{tr} C - Cz \cdot \nabla \log \nu_\tau + 2\Delta_v \log \nu_\tau + 2\|\nabla_v \log \nu_\tau\|^2.$$

Subtracting this from the same equation for μ_τ, and writing $\|\nabla_v \log \mu_\tau\|^2 - \|\nabla_v \log \nu_\tau\|^2 = \|g_{v,\tau}\|^2 + 2\langle g_{v,\tau}, \nabla_v \log \nu_\tau \rangle$,

$$\partial_\tau r_\tau = -Cz \cdot \nabla r_\tau + 2\Delta_v r_\tau + 2\|g_{v,\tau}\|^2 + 4\langle g_{v,\tau}, \nabla_v \log \nu_\tau \rangle . \tag{D.7}$$

Let $S_\tau = \nabla^2 r_\tau$, so that $\nabla_z \nabla_v r_\tau = S_\tau E$, and write $\mathcal{H}_\tau = \nabla_z \nabla_v \log \nu_\tau$. Taking the gradient of (D.7) with

$$\nabla(Cz \cdot \nabla r_\tau) = C^\top g_\tau + S_\tau C z, \qquad \nabla \|g_{v,\tau}\|^2 = 2 S_\tau E g_{v,\tau},$$
$$\nabla \langle g_{v,\tau}, \nabla_v \log \nu_\tau \rangle = S_\tau E \nabla_v \log \nu_\tau + \mathcal{H}_\tau g_{v,\tau},$$

we obtain

$$\begin{aligned} \partial_\tau g_\tau &= -C^\top g_\tau - S_\tau C z + 2\Delta_v g_\tau + 4 S_\tau E\big(g_{v,\tau} + \nabla_v \log \nu_\tau\big) + 4\mathcal{H}_\tau g_{v,\tau} \\ &= -C^\top g_\tau - S_\tau C z + 2\Delta_v g_\tau + 4 S_\tau E \nabla_v \log \mu_\tau + 4\mathcal{H}_\tau g_{v,\tau}. \end{aligned} \tag{D.8}$$

Now differentiate $\mathcal{F}_{M(\tau)}(\mu_\tau \,\|\, \nu_\tau) = \int \mu_\tau \ \langle g_\tau, M(\tau) g_\tau \rangle$. Write $M = M(\tau)$ for brevity. Using the continuity equation of μ_τ for the first term and $\nabla \langle g_\tau, M g_\tau \rangle = 2 S_\tau M g_\tau$,

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \mathcal{F}_{M(\tau)}(\mu_\tau \,\|\, \nu_\tau) = 2 \int \mu_\tau \ \langle S_\tau M g_\tau, u_\tau^\mu \rangle + \int \mu_\tau \ \Big\langle g_\tau, \dot{M} g_\tau \Big\rangle + 2 \int \mu_\tau \ \langle M g_\tau, \partial_\tau g_\tau \rangle .$$

Since $u_\tau^\mu = Cz - 2E\nabla_v \log \mu_\tau$ and S_τ, M are symmetric, the first term equals

$$2 \int \mu_\tau \ \langle M g_\tau, S_\tau C z \rangle - 4 \int \mu_\tau \ \langle M g_\tau, S_\tau E \nabla_v \log \mu_\tau \rangle .$$

In the third term we insert (D.8) and integrate the Laplacian by parts: as $\partial_{v_j} g_\tau = S_\tau e_{d+j}$ and $\partial_{v_j}(\mu_\tau M g_\tau) = \mu_\tau \, \partial_{v_j} \log \mu_\tau \ M g_\tau + \mu_\tau \, M S_\tau e_{d+j}$,

$$4 \int \mu_\tau \ \langle M g_\tau, \Delta_v g_\tau \rangle = -4 \int \mu_\tau \ \langle M g_\tau, S_\tau E \nabla_v \log \mu_\tau \rangle - 4 \int \mu_\tau \operatorname{tr}\big((S_\tau E)^\top M S_\tau E\big).$$

Collecting everything,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau} \mathcal{F}_{M(\tau)}(\mu_\tau \,\|\, \nu_\tau) = & \int \mu_\tau \ \Big\langle g_\tau, \dot{M} g_\tau \Big\rangle - 2 \int \mu_\tau \ \big\langle M g_\tau, C^\top g_\tau \big\rangle + (2-2) \int \mu_\tau \ \langle M g_\tau, S_\tau C z \rangle \\ & + (-4 + 8 - 4) \int \mu_\tau \ \langle M g_\tau, S_\tau E \nabla_v \log \mu_\tau \rangle \\ & - 4 \int \mu_\tau \operatorname{tr}\big((S_\tau E)^\top M S_\tau E\big) + 8 \int \mu_\tau \ \langle M g_\tau, \mathcal{H}_\tau g_{v,\tau} \rangle . \end{aligned}$$

Since $2\big\langle M g_\tau, C^\top g_\tau \big\rangle = \big\langle g_\tau, (CM + MC^\top) g_\tau \big\rangle$, this is (D.6). □

D.4 A Lyapunov functional with a moving metric

Motivation. By (D.5), relative entropy dissipates only the velocity part $\mathcal{F}_v$ of the relative score. To reach the full $\mathcal{F}$ we add, as in the hypocoercive functional of Section 3, a twisted Fisher divergence $\mathcal{F}_M$. Identity (D.6) shows what each ingredient contributes along a leg: the *drift term* $\Big\langle g, (\dot{M} - CM - MC^\top) g \Big\rangle$ is the only one that can dissipate in every direction, including position; the *trace term* has the right sign as soon as $M \succeq 0$; and the *mixed-Hessian term* has no sign, but by Assumption 5.2 it costs only a multiple of $\mathcal{F}_v$, which the entropy pays.

A constant metric does not survive the kick. By (D.3), the kick replaces the metric M by $(\nabla\Psi)^{-1} M (\nabla\Psi)^{-\top}$, which can be larger than M in some directions. We therefore let the metric move within each leg: it starts at a matrix $M(0)$ slightly below a fixed grid metric Θ, low enough that the increase by the kick is absorbed by the decrease in the metric $\Theta - M(0)$, and returns to Θ at the end of the leg, so that consecutive legs match at the grid points. The deficit $\Theta - M(0)$ is a budget that the dissipation along the leg must pay back.

Requirements. Fix a symmetric positive definite matrix Θ (the grid metric, chosen in (D.14)) and a weight $a > 0$. Fix a step i and write $h = h_i$ and $\Psi = \Psi_{i,h}$. A C^1 family $\{M(\tau)\}_{0\le\tau\le h}$ of symmetric $2d \times 2d$ matrices is an *admissible metric* for step i if the following hold.

(M1) *Kick compatibility:* for every $z \in \mathbb{R}^{2d}$,

$$M(0) \preceq \nabla\Psi(z)\,\Theta\,\nabla\Psi(z)^\top.$$

(M2) *Matching at the grid:* $M(h) = \Theta$.

(M3) *Positivity:* $M(\tau) \succeq 0$ for $0 \le \tau \le h$.

(M4) *Dissipation:* for $0 \le \tau \le h$,

$$-\big(\dot M(\tau) - CM(\tau) - M(\tau)C^\top\big) - \frac{1}{4}M(\tau)^2 + (2a - 64\Lambda^2)\Pi_v \succeq I_{2d}.$$

Given an admissible metric for step i, define on the ith leg and at the grid points

$$\begin{aligned} \mathcal{Y}_i(\tau) &:= a\,\mathrm{KL}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau}) + \mathcal{F}_{M(\tau)}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau}), \qquad 0 \le \tau \le h_i, \\ \mathcal{Y}_i^{\mathrm{grid}} &:= a\,\mathrm{KL}(\hat p_i \,\|\, \hat q_i) + \mathcal{F}_\Theta(\hat p_i \,\|\, \hat q_i). \end{aligned} \tag{D.9}$$

(M1) and (M2) make $\mathcal{Y}_i$ start below $\mathcal{Y}_i^{\mathrm{grid}}$ and end at $\mathcal{Y}_{i+1}^{\mathrm{grid}}$, while (M3) and (M4) make it decrease at least as fast as $\mathcal{F}$ in between.

Proposition D.4 (One-step dissipation)**.** *Suppose Ψ_{i,h_i} is a C^1 diffeomorphism and $\{M(\tau)\}$ is an admissible metric for step i. Then, for $0 < \tau < h_i$,*

$$\mathcal{Y}_i'(\tau) \le -\mathcal{F}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau}), \tag{D.10}$$

and consequently

$$\mathcal{Y}_{i+1}^{\mathrm{grid}} + \int_0^{h_i} \mathcal{F}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau})\,\mathrm{d}\tau \le \mathcal{Y}_i^{\mathrm{grid}}. \tag{D.11}$$

If this holds for every $0 \le i < N$ with the same a and Θ, then

$$\sum_{i=0}^{N-1} \int_0^{h_i} \mathcal{F}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau})\,\mathrm{d}\tau + \mathcal{Y}_N^{\mathrm{grid}} \le \mathcal{Y}_0^{\mathrm{grid}} = aH_0 + \mathcal{F}_\Theta(\hat p_0 \,\|\, \hat q_0). \tag{D.12}$$

Proof. Write $h = h_i$, $M = M(\tau)$, and apply Proposition D.3 to $\mu_\tau = \hat p_{i,\tau}$ and $\nu_\tau = \hat q_{i,\tau}$, with

$$g_\tau = \nabla \log \frac{\hat p_{i,\tau}}{\hat q_{i,\tau}}, \qquad g_{v,\tau} = E^\top g_\tau, \qquad \mathcal{H}_\tau = \nabla_z \nabla_v \log \hat q_{i,\tau}.$$

By Assumption 5.2, $\|\mathcal{H}_\tau\|_{\mathrm{op}} \le \Lambda$, so the mixed-Hessian term of (D.6) is bounded pointwise by

$$8\,\langle Mg_\tau, \mathcal{H}_\tau g_{v,\tau}\rangle \le 8\Lambda \|Mg_\tau\|\,\|g_{v,\tau}\| \le \frac{1}{4}\|Mg_\tau\|^2 + 64\Lambda^2\|g_{v,\tau}\|^2.$$

The trace term of (D.6) is nonpositive by (M3). Adding a times (D.5), which contributes $-2a\int \hat p_{i,\tau}\|g_{v,\tau}\|^2$, and writing $\|Mg_\tau\|^2 = \langle g_\tau, M^2 g_\tau\rangle$ and $\|g_{v,\tau}\|^2 = \langle g_\tau, \Pi_v g_\tau\rangle$, we get

$$\mathcal{Y}_i'(\tau) \le \int \hat p_{i,\tau} \left\langle g_\tau, \Big(\dot M - CM - MC^\top + \frac{1}{4}M^2 - (2a - 64\Lambda^2)\Pi_v\Big) g_\tau \right\rangle \mathrm{d}z \le -\int \hat p_{i,\tau}\|g_\tau\|^2$$

by (M4). This is (D.10).

At the start of the leg, $\hat p_{i,0} = (\Psi_{i,h})_\# \hat p_i$ and $\hat q_{i,0} = (\Psi_{i,h})_\# \hat q_i$. Relative entropy is invariant (Lemma D.2), and (M1) is equivalent to $(\nabla\Psi_{i,h})^{-1} M(0) (\nabla\Psi_{i,h})^{-\top} \preceq \Theta$, so (D.3) shows

$$\mathcal{F}_{M(0)}(\hat p_{i,0} \,\|\, \hat q_{i,0}) \le \mathcal{F}_\Theta(\hat p_i \,\|\, \hat q_i), \qquad \text{hence} \qquad \mathcal{Y}_i(0) \le \mathcal{Y}_i^{\mathrm{grid}}.$$

At the end of the leg, $M(h) = \Theta$ by (M2) and $\hat p_{i,h} = \hat p_{i+1}$, $\hat q_{i,h} = \hat q_{i+1}$, so $\mathcal{Y}_i(h) = \mathcal{Y}_{i+1}^{\mathrm{grid}}$. Integrating (D.10) over $[0, h]$ gives (D.11), and summing (D.11) over $i = 0, \dots, N-1$ gives (D.12), since $\hat q_0 = q_0$. □

Theorem 5.5 is thus reduced to constructing, for every step, an admissible metric with $a \lesssim \Lambda^2$ and a grid metric Θ comparable to I_{2d}. This is done in Appendix D.5, with $a = 88\Lambda^2$ and $\Theta \preceq 4I_{2d}$.

Remark D.5 (The velocity part alone). Relative entropy is invariant under the common kick (a diffeomorphism by Lemma D.6) and satisfies (D.5) on every leg, so without any metric

$$2\sum_{i=0}^{N-1}\int_0^{h_i} \mathcal{F}_v(\hat{p}_{i,\tau} \,\|\, \hat{q}_{i,\tau})\,\mathrm{d}\tau = H_0 - \mathrm{KL}(\hat{p}_N \,\|\, \hat{q}_N) \le H_0, \tag{D.13}$$

the discrete counterpart of (2.13). This identity does not use Assumption 5.2; the moving metric is what upgrades $\mathcal{F}_v$ to $\mathcal{F}$.

D.5 Construction of the moving metric

All matrices in this subsection and the next have scalar $d \times d$ blocks, $X = \bar{X} \otimes I_d$ with $\bar{X} \in \mathbb{R}^{2\times 2}$. We identify X with its 2×2 representative $\bar{X}$, write X_{xx}, X_{xv}, X_{vv} for its entries, and note that $X \succeq Y$ if and only if $\bar{X} \succeq \bar{Y}$. Fix one step, of length h, drop the index i, and write $\Psi_h = \Psi_{i,h}$ and $s = s_i$.

Definition. Set

$$\Theta := \begin{pmatrix} \frac{1}{2} & -1 \\ -1 & 3 \end{pmatrix}, \qquad A := C\Theta + \Theta C^\top = \begin{pmatrix} 2 & -\frac{9}{2} \\ -\frac{9}{2} & 10 \end{pmatrix}, \qquad \Theta^{-1} = \begin{pmatrix} 6 & 2 \\ 2 & 1 \end{pmatrix}, \tag{D.14}$$

so that $0 \prec \Theta \preceq 4I_{2d}$, and

$$\begin{aligned} P_h &:= \frac{h}{3}\Theta^2 + \frac{3\Lambda^2}{h} B_h B_h^\top, \\ M_h(\tau) &:= \Big(1 - \frac{\tau}{h}\Big) e^{\tau C}(\Theta - P_h) e^{\tau C^\top} + \frac{\tau}{h}\, e^{(\tau-h)C}\Theta e^{(\tau-h)C^\top}, \qquad 0 \le \tau \le h. \end{aligned} \tag{D.15}$$

Thus

$$M_h(0) = \Theta - P_h, \qquad M_h(h) = \Theta, \tag{D.16}$$

and, since $e^{-\tau C} M_h(\tau) e^{-\tau C^\top}$ is affine in τ with derivative $\frac{1}{h}\big(e^{-hC}\Theta e^{-hC^\top} - \Theta + P_h\big)$,

$$\dot{M}_h(\tau) - C M_h(\tau) - M_h(\tau) C^\top = \frac{1}{h}\, e^{\tau C}\big(e^{-hC}\Theta e^{-hC^\top} - \Theta + P_h\big) e^{\tau C^\top}. \tag{D.17}$$

Verification. Requirement (M2) is (D.16). Requirements (M1) and (M3) are checked in Lemma D.6, and (M4) in Lemma D.7.

Lemma D.6 (The kick and the budget P_h: (M1) and (M3))**.** *If $h \le (64\Lambda^2)^{-1}$, then Ψ_h is a global C^1 diffeomorphism,*

$$\Theta - P_h \succeq \frac{1}{6}\Theta, \qquad M_h(\tau) \succ 0 \quad (0 \le \tau \le h),$$

and, for every $z \in \mathbb{R}^{2d}$,

$$M_h(0) \preceq \nabla\Psi_h(z)\,\Theta\,\nabla\Psi_h(z)^\top. \tag{D.18}$$

Consequently, for all positive densities μ, ν,

$$\mathcal{F}_{M_h(0)}\big((\Psi_h)_\#\mu \,\big\|\, (\Psi_h)_\#\nu\big) \le \mathcal{F}_\Theta(\mu \,\|\, \nu). \tag{D.19}$$

Proof. Diffeomorphism. By (5.5), $B_h = (\beta_x, \beta_v)^\top$ with

$$\beta_x = 4\big[1 - (1+h)e^{-h}\big] \in [0, 2h^2], \qquad \beta_v = 4he^{-h} \in [0, 4h],$$

where we used $1 - (1+h)e^{-h} \le h^2/2$. Hence $\|B_h\|_{\mathrm{op}} \le 4h\sqrt{1 + h^2/4}$. As $h \le (64\Lambda^2)^{-1}$ and $\Lambda \ge 1$, we have $\Lambda h \le \frac{1}{64}$, so $\Lambda\|B_h\|_{\mathrm{op}} < 1$. Thus $z \mapsto y - B_h s(z)$ is a contraction for every y, so $\Psi_h(z) = z + B_h s(z)$ is a bijection; its Jacobian $I_{2d} + B_h \nabla s(z)$ is invertible, and by the inverse function theorem Ψ_h is a C^1 diffeomorphism.

Kick compatibility. Put $\Gamma = B_h \nabla s(z)$, so that $\nabla\Psi_h = I_{2d} + \Gamma$ and

$$(I_{2d} + \Gamma)\Theta(I_{2d} + \Gamma)^\top = \Theta + \Gamma\Theta + \Theta\Gamma^\top + \Gamma\Theta\Gamma^\top.$$

The last term is positive semidefinite, and for every g, using $\|\Gamma^\top g\| = \|(\nabla s)^\top B_h^\top g\| \le \Lambda \|B_h^\top g\|$,

$$\big|\,\langle g, (\Gamma\Theta + \Theta\Gamma^\top) g\rangle\,\big| = 2\big|\,\langle \Theta g, \Gamma^\top g\rangle\,\big| \le \frac{h}{3}\|\Theta g\|^2 + \frac{3}{h}\|\Gamma^\top g\|^2 \le \langle g, P_h g\rangle\,.$$

Hence $(I_{2d}+\Gamma)\Theta(I_{2d}+\Gamma)^\top \succeq \Theta - P_h = M_h(0)$, which is (D.18). It is equivalent to $(\nabla\Psi_h)^{-1} M_h(0) (\nabla\Psi_h)^{-\top} \preceq \Theta$, and (D.19) follows from (D.3).

Positivity. For $\Theta - P_h \succeq \frac{1}{6}\Theta$ it suffices that $\|\Theta^{-1/2} P_h \Theta^{-1/2}\|_{\mathrm{op}} \le \frac{5}{6}$. Using $\|\Theta\|_{\mathrm{op}} \le 4$, Θ^{-1} from (D.14), and the bounds on β_x, β_v,

$$\begin{aligned}\|\Theta^{-1/2} P_h \Theta^{-1/2}\|_{\mathrm{op}} &\le \frac{h}{3}\|\Theta\|_{\mathrm{op}} + \frac{3\Lambda^2}{h} B_h^\top \Theta^{-1} B_h = \frac{h}{3}\|\Theta\|_{\mathrm{op}} + \frac{3\Lambda^2}{h}\big(6\beta_x^2 + 4\beta_x\beta_v + \beta_v^2\big) \\ &\le \frac{4h}{3} + 48\Lambda^2 h\Big(1 + 2h + \frac{3}{2}h^2\Big) \le \frac{1}{48} + \frac{3}{4}\Big(1 + \frac{1}{32} + \frac{3}{8192}\Big) < \frac{5}{6}.\end{aligned}$$

Finally, $M_h(\tau)$ is a convex combination of

$$e^{\tau C}(\Theta - P_h) e^{\tau C^\top} \succeq \frac{1}{6} e^{\tau C}\Theta e^{\tau C^\top} \succ 0 \qquad \text{and} \qquad e^{(\tau-h)C}\Theta e^{(\tau-h)C^\top} \succ 0,$$

so $M_h(\tau) \succ 0$. □

Lemma D.7 (Dissipation certificate: (M4))**.** *For $h \le (64\Lambda^2)^{-1}$ and $0 \le \tau \le h$,*

$$-\big(\dot M_h(\tau) - C M_h(\tau) - M_h(\tau) C^\top\big) - \frac{1}{4} M_h(\tau)^2 + 112\Lambda^2 \Pi_v \succeq I_{2d}. \tag{D.20}$$

Proof of Theorem 5.5. We now collect the pieces.

Theorem 5.5 (Averaged relative-Fisher stationarity, underdamped)**.** *Suppose Assumptions 5.1–5.3 hold and $\max_{0\le i<N} h_i \le (64\Lambda^2)^{-1}$. Then*

$$\frac{1}{T}\sum_{i=0}^{N-1} \int_0^{h_i} \mathcal{F}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau})\,\mathrm{d}\tau \;\lesssim\; \frac{\Lambda^2 H_0 + F_0}{T}.$$

Proof. Fix i. By Lemma D.6, Ψ_{i,h_i} is a C^1 diffeomorphism and M_{h_i} satisfies (M1) and (M3); (M2) is (D.16); and by Lemma D.7, M_{h_i} satisfies (M4) with $2a - 64\Lambda^2 = 112\Lambda^2$, that is, with $a = 88\Lambda^2$. Proposition D.4 therefore applies to every step, and (D.12) together with $\Theta \preceq 4 I_{2d}$ gives

$$\sum_{i=0}^{N-1} \int_0^{h_i} \mathcal{F}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau})\,\mathrm{d}\tau + \mathcal{Y}_N^{\mathrm{grid}} \le 88\Lambda^2 H_0 + \mathcal{F}_\Theta(\hat p_0 \,\|\, \hat q_0) \le 88\Lambda^2 H_0 + 4F_0.$$

Since $\mathcal{Y}_N^{\mathrm{grid}} \ge 0$ and $\sum_i h_i = T$,

$$\frac{1}{T}\sum_{i=0}^{N-1} \int_0^{h_i} \mathcal{F}(\hat p_{i,\tau} \,\|\, \hat q_{i,\tau})\,\mathrm{d}\tau \le \frac{88\Lambda^2 H_0 + 4F_0}{T}. \tag{D.21}$$

□

D.6 Proof of the dissipation certificate

Lemma D.7 (Dissipation certificate: (M4))**.** *For $h \le (64\Lambda^2)^{-1}$ and $0 \le \tau \le h$,*

$$-\big(\dot M_h(\tau) - C M_h(\tau) - M_h(\tau) C^\top\big) - \frac{1}{4} M_h(\tau)^2 + 112\Lambda^2 \Pi_v \succeq I_{2d}.$$

Proof. Throughout, $h \le (64\Lambda^2)^{-1}$ with $\Lambda \ge 1$, so that $h \le \frac{1}{64}$, and $0 \le \tau \le h$, $u := \tau/h \in [0,1]$, $R_t := e^{tC}$. We work with 2×2 representatives. The proof has five steps: an exact formula for the left side of (D.20) (Step 1), elementary bounds on its ingredients (Steps 2–4), and the estimate of its three entries (Step 5).

Step 1: the matrix to be bounded. For a symmetric matrix X write $X_t := R_t X R_t^\top$; then $\frac{\mathrm{d}}{\mathrm{d}t}X_t = R_t(CX + XC^\top)R_t^\top$, and in particular $\frac{\mathrm{d}}{\mathrm{d}t}\Theta_t = A_t$. Put

$$K_t := R_t \Theta^2 R_t^\top, \qquad b = \begin{pmatrix} b_x \\ b_v \end{pmatrix} := R_\tau B_h, \qquad \bar{A}_\tau := \frac{1}{h}\int_{\tau-h}^{\tau} A_t \,\mathrm{d}t = \frac{1}{h}\big(\Theta_\tau - \Theta_{\tau-h}\big).$$

By (D.17) and the definition of P_h,

$$-\big(\dot{M}_h(\tau) - CM_h(\tau) - M_h(\tau)C^\top\big) = \frac{1}{h}\big(\Theta_\tau - \Theta_{\tau-h}\big) - \frac{1}{h}\,R_\tau P_h R_\tau^\top = \bar{A}_\tau - \frac{1}{3}K_\tau - \frac{3\Lambda^2}{h^2}bb^\top.$$

The left side of (D.20) is therefore

$$\mathcal{D}_h(\tau) := \bar{A}_\tau - \frac{1}{3}K_\tau - \frac{3\Lambda^2}{h^2}bb^\top - \frac{1}{4}M_h(\tau)^2 + 112\Lambda^2\Pi_v. \tag{D.22}$$

We shall prove

$$\big(\mathcal{D}_h(\tau)\big)_{xx} \ge \frac{19}{16}, \qquad \big|\big(\mathcal{D}_h(\tau)\big)_{xv}\big| \le \frac{17}{5}, \qquad \big(\mathcal{D}_h(\tau)\big)_{vv} \ge 64\Lambda^2. \tag{D.23}$$

This yields (D.20). Indeed, the diagonal entries of $\mathcal{D}_h(\tau) - I_2$ are at least $\frac{3}{16}$ and $64\Lambda^2 - 1 \ge 63$, so

$$\det\big(\mathcal{D}_h(\tau) - I_2\big) \ge \frac{3}{16}\cdot 63 - \Big(\frac{17}{5}\Big)^2 = \frac{101}{400} > 0,$$

and a symmetric 2×2 matrix with positive diagonal and positive determinant is positive definite.

Step 2: scalar bounds on Θ_t, A_t, K_t. By (E.1), $R_t = e^t\left(\begin{smallmatrix} 1-t & -t \\ t & 1+t \end{smallmatrix}\right)$, and direct multiplication shows

$$\Theta_t = e^{2t}\begin{pmatrix} \frac{1}{2}+t+\frac{3}{2}t^2 & -1-\frac{5}{2}t-\frac{3}{2}t^2 \\ -1-\frac{5}{2}t-\frac{3}{2}t^2 & 3+4t+\frac{3}{2}t^2 \end{pmatrix}, \qquad A_t = e^{2t}\begin{pmatrix} 2+5t+3t^2 & -\frac{9}{2}-8t-3t^2 \\ -\frac{9}{2}-8t-3t^2 & 10+11t+3t^2 \end{pmatrix},$$
$$K_t = e^{2t}\begin{pmatrix} \frac{5}{4}+\frac{9}{2}t+\frac{17}{4}t^2 & -\frac{7}{2}-\frac{35}{4}t-\frac{17}{4}t^2 \\ -\frac{7}{2}-\frac{35}{4}t-\frac{17}{4}t^2 & 10+13t+\frac{17}{4}t^2 \end{pmatrix}.$$

Every entry of these three matrices, and every trace, is $\pm f(t)$ with

$$f(t) = e^{2t}(\alpha + \beta t + \gamma t^2), \qquad \alpha \ge \tfrac{1}{2}, \quad \beta \ge 0, \quad 0 \le \gamma \le \tfrac{17}{2}.$$

For such f,

$$f'(t) = e^{2t}\big[(2\alpha+\beta) + 2(\beta+\gamma)t + 2\gamma t^2\big], \qquad f''(t) = e^{2t}\big[(4\alpha+4\beta+2\gamma) + 4(\beta+2\gamma)t + 4\gamma t^2\big],$$

and both are positive on $|t| \le \frac{1}{64}$, because there the constant coefficient dominates the linear one. So f is increasing and convex on $|t| \le \frac{1}{64}$, and lies above its tangent at 0 and below its chords:

$$f(0) + f'(0)\,t \;\le\; f(t) \;\le\; \begin{cases} f(0), & -\frac{1}{64} \le t \le 0, \\ f(0) + f'(\frac{1}{64})\,t, & 0 \le t \le \frac{1}{64}, \end{cases} \qquad \text{and} \qquad f(t) \le f(\tfrac{1}{64}). \tag{D.24}$$

We apply (D.24) with the following constants, computed from the formulas above and $e^{1/32} < 1.032$ (recall $\frac{\mathrm{d}}{\mathrm{d}t}\Theta_t = A_t$):

- $f'(0) = 9,\ 31,\ \frac{9}{2},\ 12,\ 17$ for $(A_t)_{xx}$, $(A_t)_{vv}$, $-(\Theta_t)_{xv}$, $\operatorname{tr}\Theta_t$, $-(A_t)_{xv}$;
- $f'(\frac{1}{64}) \le \frac{11}{5},\ \frac{24}{5},\ \frac{21}{2},\ 18,\ 8,\ 35$ for $(\Theta_t)_{xx}$, $-(\Theta_t)_{xv}$, $(\Theta_t)_{vv}$, $-(A_t)_{xv}$, $(K_t)_{xx}$, $(K_t)_{vv}$
- $f(\frac{1}{64}) \le \frac{15}{4},\ \frac{19}{5},\ 12$ for $\operatorname{tr}\Theta_t$, $-(K_t)_{xv}$, $\operatorname{tr}K_t$

The result is, writing $t^+ = \max\{t, 0\}$: for $|t| \le \frac{1}{64}$,

$$(A_t)_{xx} \ge 2 + 9t, \qquad (A_t)_{vv} \ge 10 + 31t, \qquad -\frac{9}{2} - 18\,t^+ \le (A_t)_{xv} < 0; \tag{D.25}$$

for $0 \le t \le \frac{1}{64}$,

$$0 < (\Theta_t)_{xx} \le \frac{1}{2} + \frac{11}{5}t, \qquad -1 - \frac{24}{5}t \le (\Theta_t)_{xv} \le -1, \qquad (\Theta_t)_{vv} \le 3 + \frac{21}{2}t,$$
$$\frac{7}{2} \le \operatorname{tr}\Theta_t \le \frac{15}{4}, \qquad (K_t)_{xx} \le \frac{5}{4} + 8t, \qquad -\frac{19}{5} \le (K_t)_{xv} \le -\frac{7}{2}, \tag{D.26}$$
$$(K_t)_{vv} \le 10 + 35t, \qquad \operatorname{tr} K_t \le 12;$$

and for $-\frac{1}{64} \le t \le 0$,

$$0 < (\Theta_t)_{xx} \le \frac{1}{2}, \qquad -1 \le (\Theta_t)_{xv} \le -1 + \frac{9}{2}|t|,$$
$$(\Theta_t)_{vv} \le 3, \qquad \frac{7}{2} - 12|t| \le \operatorname{tr}\Theta_t \le \frac{7}{2}. \tag{D.27}$$

In particular, for $-\frac{1}{64} \le t \le 0$,

$$(\Theta_t^2)_{xx} = (\Theta_t)_{xx}^2 + (\Theta_t)_{xv}^2 \le \frac{1}{4} + 1 = \frac{5}{4}.$$

Step 3: the kick vector. Since $B_h = 4\int_0^h R_{-r}E\,\mathrm{d}r$ and $R_tE = e^t\binom{-t}{1+t}$,

$$b = R_\tau B_h = 4\int_0^h R_{\tau-r}E\,\mathrm{d}r = 4\int_{\tau-h}^{\tau} e^t\begin{pmatrix}-t\\1+t\end{pmatrix}\mathrm{d}t.$$

The interval $[\tau-h,\tau] \subset [-h,h]$ has length h, so $\int_{\tau-h}^{\tau}|t|\,\mathrm{d}t \le \frac{h^2}{2}$; and $0 < e^t(1+t) \le (1+h)e^h$ there. Hence

$$|b_x| \le 2h^2e^h, \qquad 0 < b_v \le 4h(1+h)e^h.$$

With $\Lambda^2 h \le \frac{1}{64}$, $h \le \frac{1}{64}$ and $(1+h)^2e^{2h} \le e^{4h} \le 1 + 4he^{1/16}$, this gives

$$\frac{3\Lambda^2}{h^2}b_x^2 \le 12\Lambda^2h^2e^{2h} \le \frac{h}{5}, \qquad \frac{3\Lambda^2}{h^2}|b_xb_v| \le 24\Lambda^2h(1+h)e^{2h} \le \frac{2}{5},$$
$$\frac{3\Lambda^2}{h^2}b_v^2 \le 48\Lambda^2(1+h)^2e^{2h} \le 48\Lambda^2 + 3e^{1/16} \le 48\Lambda^2 + \frac{16}{5}, \tag{D.28}$$
$$\frac{3\Lambda^2}{h}\big(b_x^2 + b_v^2\big) \le \frac{3}{16}h^2e^{2h} + \frac{3}{4}e^{1/16} \le \frac{4}{5}.$$

Only the size of b_x is used; it has no fixed sign.

Step 4: the moving metric. Put $W := R_\tau(\Theta - P_h)R_\tau^\top$. By (D.15),

$$M_h(\tau) = (1-u)W + u\,\Theta_{\tau-h}, \qquad W = \Theta_\tau - \frac{h}{3}K_\tau - \frac{3\Lambda^2}{h}bb^\top, \tag{D.29}$$

with $\tau \in [0,h]$ and $\tau - h \in [-h,0]$.

Bounds on W. By Lemma D.6, $W \succeq \frac{1}{6}\Theta_\tau \succ 0$, so $W_{xx}, W_{vv} > 0$. The two matrices subtracted from Θ_τ in (D.29) are positive semidefinite, so the diagonal entries of W are at most those of Θ_τ. For the off-diagonal entry, (D.26) gives $-\frac{h}{3}(K_\tau)_{xv} \in \big[\frac{7}{6}h, \frac{19}{15}h\big]$, and (D.28) gives $\frac{3\Lambda^2}{h}|b_xb_v| \le \frac{2}{5}h < \frac{7}{6}h$. For the trace we use $\operatorname{tr}K_\tau \le 12$ and the last bound of (D.28). Altogether,

$$0 < W_{xx} \le \frac{1}{2} + \frac{11}{5}\tau, \qquad 0 < W_{vv} \le 3 + \frac{21}{2}\tau, \qquad -1 - \frac{24}{5}\tau \le W_{xv} \le -1 + \frac{5}{3}h,$$
$$\operatorname{tr}W \ge \frac{7}{2} - 4h - \frac{4}{5} \ge \frac{21}{8}. \tag{D.30}$$

Bounds on M_h. Since $M_h(\tau)$ is a convex combination of W and $\Theta_{\tau-h}$, (D.30) and (D.27) (at $t = \tau - h$, so $|t| \le h$) give

$$-1 - \frac{24}{5}h \le \big(M_h(\tau)\big)_{xv} \le -1 + \frac{9}{2}h,$$
$$0 < \big(M_h(\tau)\big)_{vv} \le 3 + \frac{21}{2}h, \qquad \frac{21}{8} \le \operatorname{tr}M_h(\tau) \le \frac{15}{4}. \tag{D.31}$$

Here $(M_h)_{vv} > 0$ because $M_h \succ 0$, and $\operatorname{tr} M_h \le \max\{\operatorname{tr}\Theta_\tau, \operatorname{tr}\Theta_{\tau-h}\} \le \frac{15}{4}$.

Bounds on M_h^2. For a symmetric 2×2 matrix X,

$$(X^2)_{xx} = X_{xx}^2 + X_{xv}^2, \qquad (X^2)_{xv} = X_{xv}\operatorname{tr} X, \qquad (X^2)_{vv} = X_{xv}^2 + X_{vv}^2.$$

With (D.31) and $h \le \frac{1}{64}$, the last two give

$$\begin{aligned} \frac{21}{8}\Big(1-\frac{9}{2}h\Big) &\le -\big(M_h(\tau)^2\big)_{xv} \le \frac{43}{40}\cdot\frac{15}{4}, \\ \big(M_h(\tau)^2\big)_{vv} &\le \Big(\frac{43}{40}\Big)^2 + \Big(3+\frac{21}{128}\Big)^2 \le \frac{56}{5}. \end{aligned} \tag{D.32}$$

For the position entry we use the convexity of $X \mapsto X^2$,

$$M_h(\tau)^2 \preceq (1-u)W^2 + u\,\Theta_{\tau-h}^2 \qquad \big(\text{the difference is } u(1-u)(W-\Theta_{\tau-h})^2\big),$$

together with $(\Theta_{\tau-h}^2)_{xx} \le \frac{5}{4}$ and, by (D.30) (note $|W_{xv}| \le 1 + \frac{24}{5}\tau$),

$$(W^2)_{xx} \le \Big(\frac{1}{2}+\frac{11}{5}\tau\Big)^2 + \Big(1+\frac{24}{5}\tau\Big)^2 = \frac{5}{4} + \frac{59}{5}\tau + \frac{697}{25}\tau^2 \le \frac{5}{4} + \frac{25}{2}\tau,$$

where $\frac{697}{25}\tau^2 \le \frac{697}{1600}\tau$. Hence

$$\big(M_h(\tau)^2\big)_{xx} \le (1-u)\Big(\frac{5}{4}+\frac{25}{2}\tau\Big) + \frac{5}{4}u \le \frac{5}{4} + \frac{25}{2}\tau. \tag{D.33}$$

Step 5: the three entries. We insert the bounds of Steps 2–4 into (D.22). For $\bar{A}_\tau$ we use (D.25) and the averages

$$\frac{1}{h}\int_{\tau-h}^{\tau} t\,\mathrm{d}t = \tau - \frac{h}{2}, \qquad \frac{1}{h}\int_{\tau-h}^{\tau} t^+\,\mathrm{d}t = \frac{\tau^2}{2h} \le \frac{h}{2}.$$

Position entry. By (D.25), (D.26), (D.28) and (D.33),

$$\begin{aligned} \big(\mathcal{D}_h(\tau)\big)_{xx} &\ge 2 + 9\Big(\tau - \frac{h}{2}\Big) - \frac{1}{3}\Big(\frac{5}{4}+8\tau\Big) - \frac{h}{5} - \frac{1}{4}\Big(\frac{5}{4}+\frac{25}{2}\tau\Big) \\ &= \frac{61}{48} + \frac{77}{24}\tau - \frac{47}{10}h \;\ge\; \frac{61}{48} - \frac{47}{640} = \frac{2299}{1920} > \frac{19}{16}. \end{aligned}$$

Velocity entry. By (D.25), (D.26), (D.28) and (D.32),

$$\begin{aligned} \big(\mathcal{D}_h(\tau)\big)_{vv} &\ge 10 + 31\Big(\tau - \frac{h}{2}\Big) - \frac{1}{3}(10+35\tau) - \Big(48\Lambda^2 + \frac{16}{5}\Big) - \frac{1}{4}\cdot\frac{56}{5} + 112\Lambda^2 \\ &= 64\Lambda^2 + \frac{2}{3} + \frac{58}{3}\tau - \frac{31}{2}h \;\ge\; 64\Lambda^2 + \frac{2}{3} - \frac{31}{128} > 64\Lambda^2. \end{aligned}$$

Off-diagonal entry. From below, by (D.25), (D.26), (D.28) and (D.32),

$$\begin{aligned} \big(\mathcal{D}_h(\tau)\big)_{xv} &\ge \Big(-\frac{9}{2} - 9h\Big) + \frac{7}{6} - \frac{2}{5} + \frac{1}{4}\cdot\frac{21}{8}\Big(1-\frac{9}{2}h\Big) \\ &= -\frac{1477}{480} - \frac{765}{64}h \;\ge\; -\frac{1477}{480} - \frac{3}{16} = -\frac{1567}{480} > -\frac{17}{5}. \end{aligned}$$

From above, $(\bar{A}_\tau)_{xv} < 0$ by (D.25), so the same bounds give

$$\big(\mathcal{D}_h(\tau)\big)_{xv} \le \frac{19}{15} + \frac{2}{5} + \frac{1}{4}\cdot\frac{43}{40}\cdot\frac{15}{4} = \frac{5}{3} + \frac{129}{128} < \frac{17}{5}.$$

This proves (D.23), and hence the lemma. □

E ELEMENTARY CRITICAL-DAMPING FORMULAS

The estimates of Section 5 and Appendix D use only the following explicit identities for the matrix C of (5.1). Writing $C = I_{2d} + K_0$ with $K_0^2 = 0$ (critical damping makes $C - I_{2d}$ nilpotent),

$$e^{tC} = e^t \begin{pmatrix} (1-t)I_d & -tI_d \\ tI_d & (1+t)I_d \end{pmatrix}, \qquad e^{-tC} = e^{-t} \begin{pmatrix} (1+t)I_d & tI_d \\ -tI_d & (1-t)I_d \end{pmatrix}. \tag{E.1}$$

The singular values of C are $1+\sqrt{2}$ and $\sqrt{2}-1$, each repeated d times. Finally,

$$e^{tC}e^{tC^\top} = e^{2t} \begin{pmatrix} (1-2t+2t^2)I_d & -2t^2 I_d \\ -2t^2 I_d & (1+2t+2t^2)I_d \end{pmatrix}. \tag{E.2}$$

These formulas verify all block identities used in Section 5.1 and in Appendix D and show explicitly how velocity expansion is transferred to position through the critically damped coupling.

F NUMERICAL EXPERIMENTS: DETAILS

This appendix documents the experiments of Section 6. Both use scores that are available in closed form, so that initialization and discretization are the only error sources, exactly as in the theory; nothing is trained until Section F.4.

F.1 THE INITIALIZATION ERROR EVOLVES BY THE FORWARD SEMIGROUP

The one-dimensional experiment computes the law of the exact reverse flow started from the wrong initialization *without simulating it*, through the following consequence of the transport equation (B.5). Let P_t^{fwd} denote the Markov semigroup of the generic forward diffusion (A.1), $P_t^{\mathrm{fwd}} f(z) = \mathbb{E}[f(Z_t) \mid Z_0 = z]$.

Lemma F.1 (Semigroup representation of the initialization error)**.** *In the setting of Lemma A.1, $p_t = q_t\, P_t^{\mathrm{fwd}}(p_0/q_0)$ for all $t \in [0,T]$. In particular, for the OU process* (4.1)*, $p_t(x) = q_t(x)\,\mathbb{E}\big[(p_0/q_0)\big(e^{-t}x + \sqrt{1-e^{-2t}}\,Z\big)\big]$ with $Z \sim N(0, I_d)$.*

Proof. By (B.5), $h_t = p_t/q_t$ satisfies $\partial_t h_t = -v_t^q \cdot \nabla h_t + L_t h_t$. Inserting $v_t^q = -b + G\nabla \log q_t$ and $L_t h = G : \nabla^2 h + G\nabla \log q_t \cdot \nabla h$, the two terms containing $\nabla \log q_t$ cancel, and

$$\partial_t h_t = b \cdot \nabla h_t + G : \nabla^2 h_t,$$

the backward Kolmogorov equation of (A.1), whose solution with initial datum h_0 is $P_t^{\mathrm{fwd}} h_0$. (Probabilistically: under the exact reverse path measure, $\frac{\mathrm{d}p_t}{\mathrm{d}q_t}(y) = \mathbb{E}[h_0(Y_0) \mid Y_t = y]$, which by time reversal and the Markov property of the forward process equals $\mathbb{E}[h_0(Z_T) \mid Z_{T-t} = y] = P_t^{\mathrm{fwd}} h_0(y)$.) For the OU process, $P_t^{\mathrm{fwd}} f(x) = \mathbb{E} f\big(e^{-t}x + \sqrt{1-e^{-2t}}\,Z\big)$. □

F.2 ONE-DIMENSIONAL GAUSSIAN MIXTURE

Setting. The data law is $\rho_0 = \sum_{i=1}^3 w_i N(\mu_i, \sigma^2)$ with $w = (0.3, 0.4, 0.3)$, $\mu = (-20, 0, 20)$, $\sigma = 1.5$; the forward process is the OU diffusion (4.1) on $[0,T]$ with $T = 5$, so $q_t = \rho_{T-t} = \sum_i w_i N(e^{-s}\mu_i, e^{-2s}\sigma^2 + 1 - e^{-2s})$, $s = T - t$, and $s_t = \nabla \log q_t$ are explicit. The uniform score Lipschitz constant of Assumption 4.1 is $L_s = \sup_{t,x} |s_t'(x)| \approx 20$. The wrong initialization is $p_0 = N(10, 0.5^2)$; the exact initialization $q_0 = \rho_T$ is a near-standard Gaussian, so both $H_0 = \mathrm{KL}(p_0 \,\|\, q_0) = 49.8$ and $\mathcal{F}_0 = 100$ are large.

Exact reverse overdamped flow. By Lemma F.1, $p_t = q_t P_t^{\mathrm{fwd}} h_0$ with $h_0 = p_0/q_0$, and the relative score has the posterior-mean form $\nabla \log P_t^{\mathrm{fwd}} h_0(x) = \frac{e^{-t}}{1-e^{-2t}} \mathbb{E}_{\nu_x}[Y - e^{-t}x]$ with $\nu_x(y) \propto N(y; e^{-t}x, 1-e^{-2t}) h_0(y)$, which is evaluated using numerical quadrature algorithms. The mass of p_T in the three basins (split at the midpoints between the modes) is $(0.05, 0.25, 0.71)$, against $(0.3, 0.4, 0.3)$ for ρ_0.

Discretized sampler. The frozen-score exponential integrator (4.3) is run on uniform grids with $N = T/h \in \{10, 25, 50, 100, 250, 500, 1000\}$ steps. Its law $\hat{p}_t$ is propagated on a spatial grid by applying the Gaussian transition kernel of (4.3) exactly (a dense kernel–vector product), including at intermediate times $t \in (t_k, t_{k+1})$, and $\frac{1}{T}\int_0^T \mathcal{F}(\hat{p}_t \,\|\, q_t)\,\mathrm{d}t$ is computed by four-point Gauss–Legendre quadrature in time within every step. No trajectories are sampled. The gap of the time-averaged Fisher divergence decays like $O(h)$ (Figure 1(c)), matching the $O(h)$ discretization term of Theorem 4.5.

Critically damped flow. Figure 2 repeats the whole experiment for the kinetic model of Section 5 ($\gamma = 2$, $\xi = 1$), with data law $\rho_0 \otimes N(0, 1)$ on phase space and the same wrong position law, $p_0 = N(10, 0.5^2) \otimes N(0, 1)$. Lemma F.1 applies verbatim with the kinetic forward semigroup, whose transition law is Gaussian with mean $e^{-tC}z$ and covariance G_t (Section 5.1). In panel (a), the modified Fisher divergence $\mathcal{F}_P$ with $P = P_2$ decays faster than the guaranteed e^{-t}. The terminal TV is larger than in the overdamped case because the critically damped forward process contracts positions only like $(1 + s)e^{-s}$, so at $T = 5$ it has mixed less and the mode weights are decided earlier in reverse time: panel (b) shows the terminal position marginals—p_T puts mass $(0.00, 0.00, 1.00)$ on the three basins, against $(0.3, 0.4, 0.3)$ for ρ_0, i.e. almost everything on the mode nearest the initialization—while the marginal relative score remains small, the same locally-consistent, globally-reweighted picture as in Figure 1(b).

Kinetic exponential integrator. We also ran the frozen-score exponential integrator (5.4) itself, in the two-factor form (5.7), on uniform grids with $h \in \{0.1, 0.05, 0.025, 0.0125, 0.00625\}$ and $t_N = T$: the sampler $\hat{p}$ (wrong initialization) and its twin $\hat{q}$ (exact initialization ρ_T) of Theorem 5.5 are propagated through the same kernels on the 561×281 phase-space grid, and $\int_0^h \mathcal{F}(\hat{p}_{i,\tau} \,\|\, \hat{q}_{i,\tau})\,\mathrm{d}\tau$ is accumulated by four-point Gauss–Legendre quadrature along the within-step interpolation (5.8), exactly the left-hand sides of the theorem. The time average of $\mathcal{F}(\hat{p} \,\|\, \hat{q})$ carries almost no discretization term—the divergence $\mathcal{F}(\hat{p} \,\|\, \hat{q})$ measures initialization error alone, as the decomposition of Section 5 intends.

Although we do not prove this directly, the sampler is indeed empirically first-order stationary with respect to the exact marginals themselves. For the comparison with Figure 1(c) we evaluate $\frac{1}{T}\int_0^T \mathcal{F}(\hat{p}_t \,\|\, q_t)\,\mathrm{d}t$ against the analytic scores along the frozen-score SDE interpolation of the scheme, the kinetic analog of (4.2). Its gap to the exact-flow value decays like $O(h)$ (Figure 2(d)).

F.3 CelebA-HQ with the exact empirical score

Setting. Let $x_1, \dots, x_n \in [-1, 1]^d$ be the $n = 28000$ training images of CelebA-HQ (Karras et al., 2018) at 256×256 resolution, and $\rho_0 = \frac{1}{n}\sum_i \delta_{x_i}$. The OU marginals are $\rho_s = \frac{1}{n}\sum_i N(a_s x_i, b_s^2 I_d)$ with $a_s = e^{-s}$, $b_s^2 = 1 - e^{-2s}$—a Gaussian kernel density estimate whose bandwidth is the noise level—and

$$\nabla \log \rho_s(y) = \frac{a_s \bar{x}_s(y) - y}{b_s^2}, \quad \bar{x}_s(y) = \sum_i w_i^s(y)\, x_i, \quad w^s(y) = \mathrm{softmax}_i\Big(\frac{2a_s\langle y, x_i\rangle - a_s^2\|x_i\|^2}{2b_s^2}\Big)$$

(the common term $\|y\|^2$ is dropped from the softmax), where $\bar{x}_s(y) = \mathbb{E}[X_0 \mid X_s = y]$ is the posterior mean. One score evaluation is one matrix product with the $n \times d$ data matrix. Since the data are exactly memorized, every trajectory ends at a training image, and what the initialization can change is *which* training images are produced.

That the samples are training images is not a shortcoming of the experiment but the setting of the paper, which assumes the true score throughout: for a finite training set the true score is the empirical one above, and it is also the exact minimizer of the denoising score-matching objective, so a diffusion model whose network fitted its objective perfectly would memorize; the generalization of trained models comes from the inductive bias of the network, not from the objective (Yoon et al., 2023; Gu et al., 2025; Kadkhodaie et al., 2024; Li et al., 2024). Using the exact score is therefore the only way to run the object analyzed in Sections 3–4 on real data with the score error removed, leaving initialization and discretization as the only error sources. The phenomenon shown here is moreover not tied to memorization: Appendix F.4 shows the same phenomenon with a learned score that does not memorize.

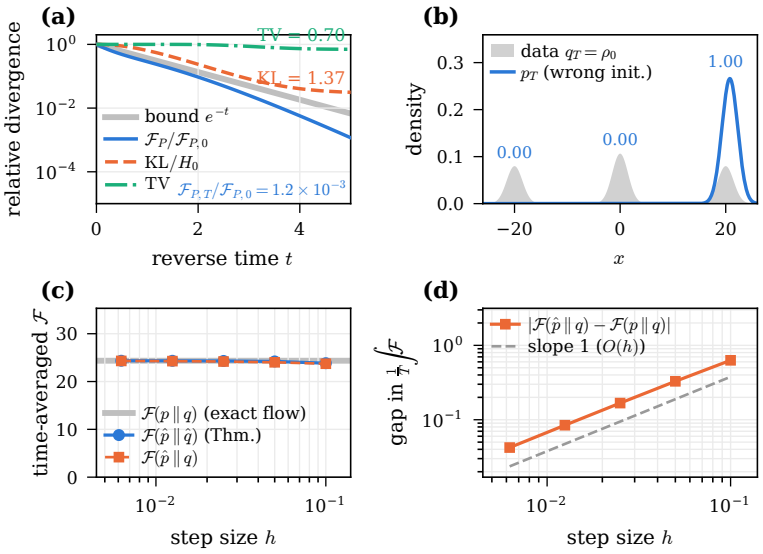


Figure 2: The critically damped experiment ($\gamma = 2$, $\xi = 1$), presented as in Figure 1. **(a)** Along the exact reverse flow, the modified Fisher divergence $\mathcal{F}_P$ contracts faster than the guaranteed e^{-t}, KL decreases monotonically, and TV plateaus. **(b)** Terminal position marginal. **(c)** The frozen-score exponential integrator (5.4): the time-averaged $\mathcal{F}(\hat{p} \,\|\, \hat{q})$ of Theorem 5.5 (circles) and the time-averaged $\mathcal{F}(\hat{p}_t \,\|\, q_t)$ against the exact marginals along the frozen-score SDE interpolation (squares) versus step size, with the exact-flow value $\frac{1}{T}\int_0^T \mathcal{F}(p_t \,\|\, q_t)\,\mathrm{d}t$ (gray). **(d)** Gap between the time-averaged $\mathcal{F}(\hat{p}_t \,\|\, q_t)$ of (c) and its exact-flow counterpart $\mathcal{F}(p_t \,\|\, q_t)$ versus step size (log–log; dashed: slope one), as in Figure 1(c).

Sampler. We use the frozen-score exponential integrator (4.3) on a grid of forward times $s_0 = T > s_1 > \cdots > s_{K-1} = s_{\min} > s_K = 0$ that is geometric in the noise-to-signal ratio $\sigma_s = \sqrt{e^{2s} - 1}$ (Karras et al., 2022) between $s_0 = T$ and $s_{K-1} = s_{\min}$, followed by one last step to $s = 0$ ($K = 150$, $s_{\min} = 1 \times 10^{-5}$, $T = 5.06$). The horizon is $T = 5.06$, the horizon of the DDPM noise schedule used in Appendix F.4, so that the two experiments run the same three initial laws over the same time span; at this T the standard initialization is close to exact ($\mathrm{KL}(N(0, I_d) \,\|\, \rho_T) = 0.30$ nats) and serves as the control. Each initialization is run for 64 samples.

Initializations. (i) *standard*: $\hat{p}_0 = N(0, I_d)$, the initialization of every practical sampler; (ii) *mean shift*: $\hat{p}_0 = N(\mu, I_d)$ with $\mu = \beta\,\mathbf{1}/\sqrt{d}$, a uniform brightness offset of $\beta/\sqrt{d}$ per pixel, $\beta = \pm 4$—the same three initial laws as in the learned-score experiment of Appendix F.4. H_0 and $\mathcal{F}_0$ are estimated by Monte Carlo from the closed-form densities.

Results. Figure 3 shows the 64 generated images of each run. Every one of them is close to a training image (root-mean-square distance to the nearest training image below one gray level, i.e. the noise of the last step): local consistency holds regardless of the initialization. Which training images are produced does depend on the initialization. The standard initialization $N(0, I_d)$, nearly exact at this horizon ($H_0 = 0.30$ nats), is the control: it returns a mixed draw of the training set—dark and light backgrounds, hair colors and lighting conditions (mean pixel value -0.14 on the $[-1, 1]$ scale, mean per-image pixel standard deviation 0.51). The mean shifts ($H_0 = 5.03$ resp. 6.78 nats) move the output to opposite ends of the training set: $\beta = -4$ returns almost exclusively dark photographs—dim lighting, dark backgrounds, hair and skin (mean pixel value -0.43)—and $\beta = +4$ almost exclusively bright ones—light backgrounds, blond hair, pale skin, over-exposed shots. The initialization error is not negligible but shrinks along the flow in the first-order sense, yet it decides the palette of the output: small relative Fisher divergence certifies local consistency, not the mode weights.



$N(0, I_d)$ $N(-4\,\mathbf{1}/\sqrt{d}, I_d)$ $N(4\,\mathbf{1}/\sqrt{d}, I_d)$

Figure 3: CelebA-HQ with the exact empirical score: 64 images generated by the sampler of Section 4 from the standard initialization and the two mean shifts, $T = 5.06$, $n = 28000$ training images. Every image is a training image; the initialization decides which ones are produced.



$N(0, I)$ (mean pixel -0.09) $N(\mu, I)$, $\mu = -4\,\mathbf{1}/\sqrt{d}$ (-0.56) $N(\mu, I)$, $\mu = 4\,\mathbf{1}/\sqrt{d}$ $(+0.50)$

Figure 4: Pretrained CelebA-HQ DDPM (Ho et al., 2020), 64 samples of the 250-step ancestral sampler from the same three initializations of the initial noise as in Figure 3. The mean shifts along the all-ones direction nevertheless bias them towards dark ($\beta < 0$) or bright ($\beta > 0$) photographs, exactly as with the exact score.

F.4 The same effect with a learned score

To check that this phenomenon persists in the presence of score estimation errors, we repeat the mean-shift experiment with the pretrained CelebA-HQ 256×256 DDPM of Ho et al. (2020) (`google/ddpm-celebahq-256`), sampled with its ancestral sampler (250 steps)—the discrete predecessor of the integrator of Section 4—whose noise schedule corresponds to the OU horizon $T = 5.06$. Figure 4 shows 64 samples from the standard initialization $N(0, I_d)$ and from $N(\mu, I_d)$ with $\mu = \beta\,\mathbf{1}/\sqrt{d}$ for $\beta = -4$ and $+4$—the same three initial laws as in Figure 3. The mean pixel value of the samples moves from -0.09 to -0.56 and $+0.50$ respectively: a positive shift makes the learned model return bright photographs and a negative shift dark ones. The bias now acts on the cluster of bright (dark) photographs rather than on individual training images, as one expects from a score that has learned the data distribution rather than the training set; the mechanism—the mass allocated to the clusters is decided by the tilt the initialization induces, while every sample remains a clean face—is the same.