

# WEIGHTED STOCHASTIC DIFFERENTIAL EQUATION TO IMPLEMENT WASSERSTEIN-FISHER-RAO GRADIENT FLOW

Herlock Rahimi

Department of Electrical and Computer Engineering, Yale University

herlock.rahimi@yale.edu

## ABSTRACT

Score-based diffusion models currently constitute the state of the art in continuous generative modeling. These methods are typically formulated via overdamped or underdamped Ornstein–Uhlenbeck-type stochastic differential equations, in which sampling is driven by a combination of deterministic drift and Brownian diffusion, resulting in continuous particle trajectories in the ambient space. While such dynamics enjoy exponential convergence guarantees for strongly log-concave target distributions, it is well known that their mixing rates deteriorate exponentially in the presence of nonconvex or multimodal landscapes, such as double-well potentials. Since many practical generative modeling tasks involve highly non-log-concave target distributions, considerable recent effort has been devoted to developing sampling schemes that improve exploration beyond classical diffusion dynamics.

A promising line of work leverages tools from information geometry to augment diffusion-based samplers with controlled mass reweighting mechanisms. This perspective leads naturally to Wasserstein–Fisher–Rao (WFR) geometries, which couple transport in the sample space with vertical (reaction) dynamics on the space of probability measures. In this work, we formulate such reweighting mechanisms through the introduction of explicit correction terms and show how they can be implemented via weighted stochastic differential equations using the Feynman–Kac representation. Our study provides a preliminary but rigorous investigation of WFR-based sampling dynamics, and aims to clarify their geometric and operator-theoretic structure as a foundation for future theoretical and algorithmic developments.

## 1. INTRODUCTION

Modern score-based diffusion models can be viewed as learning (or approximating) the time-reversed dynamics of a forward noising diffusion, so that generation reduces to sampling by simulating a stochastic process (or its probability-flow ODE counterpart); see, e.g., [38]. This “sampling-first” viewpoint makes the overall quality–compute tradeoff hinge on a classical question in stochastic analysis and MCMC: *how fast does a diffusion (and its discretization) converge to its target distribution?*

When the target density has the form  $\pi(dx) \propto e^{-V(x)}dx$  with  $V$  smooth and (strongly) convex, the associated Langevin dynamics enjoys quantitative convergence rates to equilibrium. At the continuous-time level, a standard route is via functional inequalities (Poincaré/log-Sobolev) and curvature-type criteria, notably the Bakry–Émery framework, which yields exponential decay of suitable divergences under strong convexity/positive curvature assumptions [4, 21, 23]. At the algorithmic level, non-asymptotic guarantees for discretizations such as unadjusted Langevin Monte Carlo (ULA/LMC) are by now well developed in the smooth log-concave setting, including explicit dimension/accuracy scaling

and robustness to gradient error [11, 12, 17, 18]. Moreover, kinetic (underdamped) Langevin diffusions can yield improved complexity bounds over overdamped LMC under comparable regularity/convexity conditions, and several analyses establish accelerated rates in Wasserstein/KL-type metrics [8, 13, 34].

The picture changes drastically once  $V(\cdot)$  is *nonconvex* and the target is multi-modal (e.g., double-well potentials or mixtures), which is the typical regime for many scientific and modern generative-modeling tasks. In such landscapes, Langevin-type dynamics may exhibit *metastability*: trajectories spend exponentially long times trapped near one mode before crossing an energy barrier to another. Sharp asymptotics for transition times are classically captured by Kramers/Eyring–Kramers laws and their refinements, which quantify the barrier-dominated nature of mixing in the low-noise (or low-temperature) regime [19, 6, 25]. This explains, at a mechanism level, why “exponential convergence” in the log-concave case does *not* translate to effective sampling performance in multi-well settings: in the presence of energy barriers, the spectral gap and log-Sobolev constant of the associated Langevin generator typically decay exponentially in the barrier height, implying exponentially large global mixing times [21, 20, 5, 29].

A substantial modern literature addresses this slow-mixing obstruction by changing either (i) the *dynamics* (e.g., kinetic/nonreversible variants, couplings showing contraction under weaker conditions) [15], or (ii) the *effective landscape* (e.g., tempering/replica-exchange ideas designed to move between modes). As one representative example, simulated tempering combined with Langevin updates yields provable improvements for certain mixtures of log-concave components [24]. These developments motivate the central theme of this paper: to systematically enrich the sampling process by incorporating *geometric* degrees of freedom (e.g., reweighting/Fisher–Rao components in Wasserstein–Fisher–Rao-type formulations), aiming to mitigate metastability while preserving a principled continuum description compatible with diffusion-model methodology.

**Brownian Motion is Not Enough to explore the space!** A central modeling choice in diffusion-based sampling (and in score-based generative modeling via SDEs) is the *driving noise*. In the classical overdamped Langevin diffusion,

$$dX_t = -\nabla f(X_t) dt + \sqrt{2} dB_t, \quad (1)$$

the Brownian term  $B_t$  induces increments with typical size  $\|B_t - B_s\| \asymp |t-s|^{1/2}$ . This 1/2-scaling is intimately tied to Gaussianity, finite quadratic variation, and the semimartingale structure that underpins Itô calculus and the classical Fokker–Planck PDE. However, in multimodal landscapes (e.g. double-well potentials), Brownian-driven dynamics can become metastable: barrier crossing is dominated by rare fluctuations whose timescale is exponentially large in the barrier height (cf. Kramers/Eyring–Kramers theory and related metastability results) [19, 7, 26]. This motivates enriching the

noise model to enable more effective exploration and faster inter-mode transport.

A common informal heuristic is: *Brownian paths explore locally (diffusively), whereas heavy-tailed or jump-driven paths can occasionally relocate nonlocally*, potentially reducing barrier-induced trapping. Two canonical generalizations of Brownian motion illustrate the tradeoff between (i) scaling properties that can enhance exploration and (ii) the availability of Itô calculus, and hence writing down the forward/backward equations:

**(A) Fractional Brownian motion (fBm):** Fractional Brownian motion  $(B_t^H)_{t \geq 0}$  with Hurst index  $H \in (0, 1)$  is a centered Gaussian process with covariance

$$\mathbb{E}[B_t^H B_s^H] = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H}), \quad s, t \geq 0, \quad (2)$$

and self-similarity  $B_{ct}^H \stackrel{d}{=} c^H B_t^H$  [31]. For  $H \neq \frac{1}{2}$ , fBm is *not* a semimartingale, so the standard Itô stochastic integral (and Itô formula) is not available; one must use alternative calculi (Young/rough paths, Malliavin–Skorohod integrals, etc.) [31, 22]. In other words, fBm changes scaling (typical increments  $\asymp t^H$ ) while remaining Gaussian, but it breaks the semimartingale foundation of classical diffusion PDEs.

**(B) A general jump diffusion** (or Lévy-driven SDE) can be written as

$$\begin{aligned} dX_t &= b(X_{t-}) dt + \sigma(X_{t-}) dB_t \\ &\quad + \int_{\|z\|<1} \gamma(X_{t-}, z) \tilde{N}(dt, dz) \\ &\quad + \int_{\|z\|\geq 1} \Gamma(X_{t-}, z) N(dt, dz), \end{aligned} \quad (3)$$

where  $N$  is a Poisson random measure on  $\mathbb{R}_+ \times (\mathbb{R}^d \setminus \{0\})$  with intensity  $dt \nu(dz)$ , and  $\tilde{N} := N - dt \nu$  is the compensated measure [3, 32]. For sufficiently smooth test functions  $\varphi$ , one has an Itô formula with jumps, producing both differential and jump correction terms [32, 3].

The infinitesimal generator  $\mathcal{L}$  of a Lévy-driven process is generally *nonlocal*. For a pure Lévy process with triplet  $(b, \Sigma, \nu)$ , acting on smooth  $\varphi$ ,

$$\begin{aligned} \mathcal{L}\varphi(x) &= \langle b, \nabla \varphi(x) \rangle + \frac{1}{2} \text{Tr}(\Sigma \nabla^2 \varphi(x)) \\ &\quad + \int_{\mathbb{R}^d \setminus \{0\}} \left( \varphi(x+z) - \varphi(x) - \langle \nabla \varphi(x), z \rangle \mathbf{1}_{\{\|z\|<1\}} \right) \nu(dz). \end{aligned} \quad (4)$$

When  $\nu$  corresponds to an isotropic  $\alpha$ -stable process,  $\mathcal{L}$  reduces (up to a constant) to the fractional Laplacian  $-(-\Delta)^{\alpha/2}$ , highlighting the precise way in which “ $t^{1/\alpha}$  scaling” induces a *fractional* (nonlocal) diffusion operator [33, 3]. Consequently, the forward equation for densities is a PIDE (often called a nonlocal Fokker–Planck equation), rather than the classical second-order PDE associated with Brownian noise. Explicit derivations and general forms of these equations for Lévy-driven systems (and related stochastic integrals) can be found in [37, 41, 36]. In a multi-well landscape, the bottleneck is the probability of producing a fluctuation large enough to cross the energy barrier. Brownian motion accomplishes this via many small steps (and potentially high rates of diffusion, i.e. higher temperature) making crossing times exponentially sensitive to barrier height. A jump component allows *rare but macroscopic* moves:

a single jump may relocate the state across wells, potentially reducing metastable trapping. This comes at a price: nonlocality in the generator and additional modeling/analysis complexity (e.g. stability of discretizations, defining suitable likelihoods for training, and understanding how scores transform under jump noise). Recent score-based generative modeling work explicitly explores  $\alpha$ -stable Lévy noise as the forward corruption process, precisely to leverage such heavy-tailed jumps [42].

A central practical obstacle for *non-Gaussian* noise designs (e.g. jump–diffusions,  $\alpha$ -stable drivers, or more general Lévy processes) is that their simulation and analysis typically requires a Poisson random measure (or Lévy–Itô decomposition), nonlocal generators, and bespoke discretizations; moreover, many classical identities used in diffusion-model training (e.g. Itô calculus, Girsanov transforms in their simplest form, and PDE duality with second-order elliptic operators) either fail outright or require substantial additional structure and technical overhead. This motivates a complementary route: *retain Brownian motion at the particle level*, but enrich the *law-level* dynamics by allowing *mass reweighting* (and, when needed, resampling/branching).

$$dx_t = v_t(x_t) dt + \sigma_t dB_t, \quad dw_t = \bar{\psi}_t(x_t) dt, \quad (5)$$

In other words,  $x_t$  diffuses, and then we resample different paths w.r.t  $w_t$ , to allow teleportation.

The resulting Wasserstein–Fisher–Rao (WFR) or Hellinger–Kantorovich (HK) metric equips the space of positive measures with a genuine Riemannian structure whose tangent vectors decompose into horizontal (transport) and vertical (reaction) components [27, 28, 10]. Gradient flows in this geometry rigorously capture the coupled dynamics

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t) + \rho_t \bar{\psi}_t, \quad (6)$$

which subsumes weighted Fokker–Planck equations, birth–death samplers, and normalized Feynman–Kac flows as special cases.

Recent work has begun to extend curvature-based analysis (à la Bakry–Émery or Lott–Sturm–Villani) to unbalanced and information-geometric settings. Lower curvature bounds in WFR/HK spaces yield contractivity and stability results for coupled transport–reaction flows, paralleling classical Wasserstein theory but with additional degrees of freedom [30, 16]. These results provide a geometric explanation for why adding Fisher–Rao components can improve mixing: the effective curvature of the sampling manifold is altered, potentially enlarging spectral gaps and reducing geodesic distances between modes.

The perspective adopted in this work is that, while *genuine* jump-driven or non-Gaussian noise generalizations can be powerful tools for enhancing exploration, they often entail a substantial burden both in stochastic calculus and in practical implementation. Weighted stochastic differential equations and their interacting particle system (IPS) realizations, by contrast, provide a pragmatic intermediate approach: they preserve diffusive particle dynamics and thus remain compatible with standard Itô calculus and numerical SDE methods, while enriching the evolution at the *law level* through a controllable vertical (reaction) component. This vertical component admits a natural interpretation in terms of Fisher–Rao geometry and is intrinsically compatible with the Wasserstein–Fisher–Rao (WFR), or equivalently Hellinger–Kantorovich (HK), framework, including branching and resampling interpretations.

The organization of the paper reflects this viewpoint. In Section 2, we formally introduce the Riemannian geometry of the Wasserstein–Fisher–Rao space and review the essential elements of its differential structure. Having established the underlying geometric framework, Section 3 shows that weighted stochastic differential equations induce the corresponding WFR evolution at the level of probability densities, thereby yielding an implementable sampling scheme that realizes Fisher–Rao-type geometric corrections. The proof relies on a Feynman–Kac representation and follows established arguments in the literature. We subsequently present an equivalent formulation in terms of jump or branching processes, further clarifying the relationship between weighted diffusions and nonlocal dynamics.

Both the Feynman–Kac correspondence between weighted SDEs and WFR-type partial differential equations, as well as the equivalence with jump-process formulations, are classical results and can be found in the cited references; accordingly, they are **not claimed as novel contributions** of this work<sup>1</sup>.

## 2. HOW TO BUILD THE GEOMETRIES?

Throughout, we denote by  $\mathcal{P}(\mathbb{R}^d)$  the set of Borel probability measures on  $\mathbb{R}^d$ , and by

$$\mathcal{P}_{\text{ac}}^2(\mathbb{R}^d) := \left\{ \rho \in \mathcal{P}(\mathbb{R}^d) : \rho \ll \text{Leb}, \int_{\mathbb{R}^d} \|x\|^2 \rho(x) dx < \infty \right\},$$

the space of absolutely continuous probability measures with finite second moment. Endowing  $\mathcal{P}_{\text{ac}}^2(\mathbb{R}^d)$  with additional geometric structure allows one to define tangent spaces, variational and gradient flows, geodesics, and curvature-type notions, thereby enabling a differential-geometric treatment of evolution equations on the space of probability measures. Unlike finite-dimensional Euclidean spaces, however, there is no unique canonical geometry on  $\mathcal{P}_{\text{ac}}^2(\mathbb{R}^d)$ . Instead, several distinct—yet deeply interconnected—geometric frameworks arise, depending on which fundamental notions of classical geometry one seeks to generalize.

At a conceptual level, geometric structures on abstract (possibly infinite-dimensional) manifolds may be introduced by generalizing one or more of the following fundamental aspects of Euclidean geometry:

1. *Metric structure*: generalizing straight lines as length-minimizing curves, leading to metric or length spaces and geodesic distances;
2. *Curvature structure*: generalizing angles and second-order variation, allowing one to quantify convexity, contraction, and curvature bounds;
3. *Affine structure*: generalizing straight lines as curves of minimal acceleration, leading to affine connections, parallel transport, and families of distinguished geodesics.

Each of these perspectives yields a different—but mathematically natural—geometry on the space of probability measures. In the sequel, we show how these three viewpoints give rise, respectively, to Wasserstein geometry, Fisher–Rao information geometry,

<sup>1</sup>On the other hand, there is an immediate application: if you want to avoid retraining diffusion models for  $q^1, q^2$ , is there a way to use them to sample from different mixtures of  $q^1$ ? This, in particular for information preserving mixtures, is answered in the last section of Appendix

and connection-based (affine) structures, and how their interaction underpins the Wasserstein–Fisher–Rao framework studied in this work. **(i) Geometry from a distance or divergence.** A classical and conceptually transparent approach begins with a notion of “distance”  $D(\mu, \nu)$  between two probability measures. If  $D$  is a metric, or more generally a divergence with suitable convexity and lower semicontinuity properties, one can define geodesics as curves  $\gamma : [0, 1] \rightarrow \mathcal{P}(\mathbb{R}^d)$  minimizing the length functional

$$L(\gamma) = \sup_{0=t_0 < t_1 < \dots < t_N=1} \sum_{i=0}^{N-1} D(\gamma(t_i), \gamma(t_{i+1})).$$

This viewpoint is inherently variational: the geometry is encoded in the minimizing properties of curves. The Wasserstein space  $(\mathcal{P}^2(\mathbb{R}^d), W_2)$  is the canonical example. Geodesics are displacement interpolations, obtained by pushing forward a measure along optimal transport maps; their variational characterization leads directly to the Otto calculus and second-order differential structure[40].

Another family of geometries arises when  $D$  is a statistical divergence, such as the Kullback–Leibler divergence, squared Hellinger distance, or the Amari  $\alpha$ -divergences [1]. In this setting, geodesics correspond to “interpolations induced by the divergence”, and need not have a transport interpretation. For example, the Fisher–Rao geodesic between densities  $\rho_0$  and  $\rho_1$  is given by the explicit formula

$$\rho_t = ((1-t)\sqrt{\rho_0} + t\sqrt{\rho_1})^2,$$

which is the geodesic associated with the squared Hellinger distance. In these divergence-based geometries, curvature, convexity, and stability properties follow from the analytic structure of the divergence.

**(ii) Geometry from a Riemannian metric (angle-based viewpoint).** A more differential perspective begins by postulating a Riemannian metric on each tangent space  $T_\rho \mathcal{P}_{\text{ac}}^2$ , that is, an inner product (more formally a (0,2)-tensor, or a 2 form)

$$\langle \xi, \eta \rangle_\rho, \quad \xi, \eta \in T_\rho \mathcal{P}_{\text{ac}}^2.$$

This requires specifying a model for the tangent space. In Wasserstein geometry, the tangent vectors are velocity fields  $v$  satisfying the continuity equation  $\partial_t \rho + \nabla \cdot (\rho v) = 0$ , and the Riemannian metric is

$$\langle v_1, v_2 \rangle_\rho := \int_{\mathbb{R}^d} \rho(x) v_1(x) \cdot v_2(x) dx.$$

This makes  $\mathcal{P}_{\text{ac}}^2$  into a formal infinite-dimensional Riemannian manifold where gradient flows of functionals become PDEs. For example, the Fokker–Planck equation is the Wasserstein gradient flow of the free energy functional [2].

In Fisher–Rao geometry, the tangent vectors take the form  $\xi = \dot{\rho}$  with  $\int \xi = 0$  and the metric reads

$$\langle \xi, \eta \rangle_\rho := \int_{\mathbb{R}^d} \frac{\xi(x) \eta(x)}{\rho(x)} dx,$$

inducing a distinct Riemannian structure not based on transport but on multiplicative perturbations. This construction is the infinite-dimensional limit of the classical Fisher information metric on statistical models [1]. Unlike the Wasserstein metric, the Fisher–Rao metric treats mass change, not displacement, as the fundamental mode of variation.

The Riemannian viewpoint provides notions of angles, lengths of curves, Levi–Civita connections, and curvature tensors. It supports a second-order differential calculus on  $\mathcal{P}_{\text{ac}}^2$  that is indispensable for stability analysis, Bochner-type formulas, and geometric interpretations of PDEs.

**(iii) Geometry from affine connections and dualistic structures.** A third pathway, prominent in information geometry, develops geometry starting from an affine connection  $\nabla$  rather than a metric. This notion of building geometry tries to generalize the concept of straightness of lines(henceforth geodesics), rather than the length. This connection determines parallel transport, geodesics, and curvature, and need not arise from any Riemannian metric. In the dualistic formalism of Amari [1], one introduces a pair of torsion-free affine connections  $(\nabla, \nabla^*)$  that are dual with respect to a Riemannian metric  $g$ . Different choices of  $\nabla$  lead to different families of geodesics: mixture geodesics, exponential geodesics, and more generally the  $\alpha$ -geodesics associated with the  $\alpha$ -connections. These geodesics admit simple coordinate expressions: for instance, mixture geodesics are straight lines in the space of densities, while exponential geodesics are straight lines in logarithmic coordinates. In this viewpoint, geodesic structure is primary, and the Riemannian metric appears only as a dualizing object relating  $\nabla$  and  $\nabla^*$ .

This affine-geometric construction generalizes naturally to infinite-dimensional spaces of measures. For instance,  $\alpha$ -connections induce “midpoint operators” defining interpolations between measures that are not derived from minimizing a distance but from preserving affine structure in a chosen coordinate system. These structures play a central role in dual-flat geometries and provide the foundation for Bregman divergences, convex potentials, and natural gradient methods.

**Summary.** The three approaches above are not competing theories but complementary ones. Starting from one viewpoint toward the geometry, one can build other concepts as well, though it is important to have them in our arsenal to use these tools to better understand the geometries

These frameworks collectively demonstrate that the space of probability measures supports a rich family of geometric structures. Choosing one determines the analytic form of PDEs (as gradient flows or geodesic equations), the SDEs representing their particle dynamics, and the curvature properties governing stability and convergence. The remainder of this work builds upon these foundational constructions to examine Wasserstein geometry, Fisher–Rao geometry, and their hybridizations in detail.

## 2.1. Wasserstein Geometry, Otto Calculus

For  $\mu, \nu \in \mathcal{P}_2(\Omega)$ , the 2-Wasserstein distance is defined by

$$W_2^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} \|x - y\|^2 d\pi(x, y),$$

where  $\Pi(\mu, \nu)$  is the set of couplings of  $\mu$  and  $\nu$ . Equipped with  $W_2$ , the space  $(\mathcal{P}_2(\Omega), W_2)$  is a complete separable metric space; in particular, one can speak of absolutely continuous curves, metric derivatives, and so on.

**Definition 1** (Metric derivative). Let  $(X, d)$  be a metric space and let  $(x_t)_{t \in [0,1]}$  be a curve in  $X$ . The *metric derivative* of  $(x_t)$  at time  $t$  (when it exists) is

$$|x'|_t := \lim_{s \rightarrow t} \frac{d(x_s, x_t)}{|s - t|}.$$

A curve is called absolutely continuous if there exists  $m \in L^1([0, 1])$  such that

$$d(x_s, x_t) \leq \int_s^t m(r) dr, \quad 0 \leq s \leq t \leq 1.$$

We are primarily interested in absolutely continuous curves  $(\mu_t)_{t \in [0,1]}$  in  $(\mathcal{P}_2(\Omega), W_2)$ . A basic result (Ambrosio–Gigli–Savaré) says that such curves admit an *Eulerian* description via a continuity equation.

**Definition 2** (Continuity equation). A narrowly continuous curve  $(\mu_t)_{t \in [0,1]} \subset \mathcal{P}_2(\Omega)$  satisfies the continuity equation with (time-dependent) velocity field  $(v_t)_{t \in [0,1]}$  if

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0 \tag{7}$$

in the sense of distributions, i.e. for every  $\varphi \in C_c^\infty(\Omega)$

$$\frac{d}{dt} \int_\Omega \varphi d\mu_t = \int_\Omega \nabla \varphi(x) \cdot v_t(x) d\mu_t(x) \quad \text{for a.e. } t.$$

Formally, one may think of  $(\mu_t)$  as the distribution of a random particle  $X_t$  satisfying the ODE

$$\dot{X}_t = v_t(X_t), \quad X_0 \sim \mu_0,$$

and then  $\mu_t = (\text{Law } X_t)$  solves (7). The pair  $(\mu_t, v_t)$  provides an Eulerian (density/velocity) description, while  $X_t$  gives a Lagrangian (particle/trajecotry) description.

A crucial fact is that, for a given curve  $(\mu_t)$ , the velocity field  $v_t$  solving (7) is not unique: one may add divergence-free fields  $\tilde{v}_t$  with  $\nabla \cdot (\mu_t \tilde{v}_t) = 0$  without changing the evolution of  $\mu_t$ . The standard resolution is to select, at each time, the *minimal kinetic energy* representative.

Define the kinetic energy of  $(\mu_t, v_t)$  on  $[0, 1]$  by

$$\int_0^1 \|v_t\|_{\mu_t}^2 dt := \int_0^1 \int_\Omega \|v_t(x)\|^2 d\mu_t(x) dt.$$

There is a deep link between this functional and the Wasserstein metric: the Benamou–Brenier formula states that

$$W_2^2(\mu_0, \mu_1) = \inf_{(\mu_t, v_t)} \int_0^1 \int_\Omega \|v_t(x)\|^2 d\mu_t(x) dt, \tag{8}$$

where the infimum runs over all narrowly continuous  $(\mu_t)_{t \in [0,1]}$  connecting  $\mu_0$  to  $\mu_1$  and all Borel vector fields  $(v_t)$  solving (7). Minimizers  $(\mu_t, v_t)$  correspond to constant-speed geodesics in  $(\mathcal{P}_2(\Omega), W_2)$ .

This suggests a Riemannian interpretation: at each  $\mu$ , we declare the *tangent space* to be the closure of gradient vector fields in  $L^2(\mu)$ .

**Definition 3** (Wasserstein tangent space and metric). Let  $\mu \in \mathcal{P}_2(\Omega)$  admit a smooth strictly positive density with respect to Lebesgue, which we still denote by  $\mu(x)$ . The (formal) tangent space at  $\mu$  is

$$T_\mu \mathcal{P}_2(\Omega) := \overline{\{\nabla \phi : \phi \in C_c^\infty(\Omega)\}}^{L^2(\mu)},$$

and the Riemannian metric at  $\mu$  is given by the  $L^2(\mu)$  inner product

$$\langle \nabla \phi_1, \nabla \phi_2 \rangle_\mu := \int_\Omega \nabla \phi_1(x) \cdot \nabla \phi_2(x) d\mu(x).$$

We write  $\|v\|_\mu^2 := \langle v, v \rangle_\mu$ .

Under this identification, the minimal-kinetic-energy velocity field  $v_t$  associated with a curve  $(\mu_t)$  is uniquely characterized (up to  $\mu_t$ -null sets), satisfies  $v_t \in T_{\mu_t} \mathcal{P}_2(\Omega)$ , and its norm coincides with the metric derivative:

$$|\mu'|_t(t) = \|v_t\|_{\mu_t} \quad \text{for a.e. } t.$$

At the level of optimal transport maps, if  $T_{\mu_t \rightarrow \mu_{t+h}}$  denotes the optimal map from  $\mu_t$  to  $\mu_{t+h}$  and  $\text{id}$  is the identity map, then formally

$$v_t = \lim_{h \rightarrow 0} \frac{T_{\mu_t \rightarrow \mu_{t+h}} - \text{id}}{h} \quad \text{in } L^2(\mu_t).$$

Let  $F: \mathcal{P}_2(\Omega) \rightarrow \mathbb{R}$  be a functional. Its *first variation* at  $\mu$  (when it exists) is a scalar function  $\frac{\delta F}{\delta \mu}(\mu): \Omega \rightarrow \mathbb{R}$ , defined (up to an additive constant) by

$$\frac{d}{d\varepsilon} \Big|_{\varepsilon=0} F(\mu + \varepsilon\sigma) = \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu)(x) d\sigma(x)$$

for all signed measures  $\sigma$  such that  $\mu + \varepsilon\sigma \in \mathcal{P}_2(\Omega)$  for  $\varepsilon$  small enough. If  $(\mu_t)$  is a curve with velocity field  $v_t$  satisfying the continuity equation, then

$$\begin{aligned} \frac{d}{dt} F(\mu_t) &= \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t)(x) \partial_t \mu_t(x) dx \\ &= - \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t)(x) \nabla \cdot (\mu_t(x) v_t(x)) dx, \end{aligned}$$

and an integration by parts yields

$$\frac{d}{dt} F(\mu_t) = \int_{\Omega} \nabla \left( \frac{\delta F}{\delta \mu}(\mu_t) \right) (x) \cdot v_t(x) d\mu_t(x) = \langle \nabla \frac{\delta F}{\delta \mu}(\mu_t), v_t \rangle_{\mu_t}.$$

**Definition 4** (Wasserstein gradient). The (formal) Wasserstein gradient of  $F$  at  $\mu$  is the element

$$\nabla^{W_2} F(\mu) \in T_{\mu} \mathcal{P}_2(\Omega)$$

characterized by

$$\frac{d}{dt} F(\mu_t) = \langle \nabla^{W_2} F(\mu_t), v_t \rangle_{\mu_t} \quad \text{for all curves } (\mu_t) \text{ with velocity } v_t.$$

From the computation above one obtains

$$\nabla^{W_2} F(\mu) = \nabla \left( \frac{\delta F}{\delta \mu}(\mu) \right).$$

The gradient flow equation associated with  $F$  in the Wasserstein geometry is obtained by following the steepest descent direction:

$$v_t = -\nabla^{W_2} F(\mu_t).$$

Plugging into the continuity equation yields the PDE

$$\partial_t \mu_t = \nabla \cdot \left( \mu_t \nabla \left( \frac{\delta F}{\delta \mu}(\mu_t) \right) \right). \quad (9)$$

Formally,

$$\frac{d}{dt} F(\mu_t) = -\|\nabla^{W_2} F(\mu_t)\|_{\mu_t}^2 \leq 0,$$

so  $F(\mu_t)$  is non-increasing along the flow.

A central sufficient condition for quantitative convergence is (geodesic)  $\lambda$ -convexity of  $F$  in  $(\mathcal{P}_2(\Omega), W_2)$ : if  $F$  is  $\lambda$ -geodesically convex, then the gradient flow is well-posed and  $F(\mu_t) - \inf F$

decays at least exponentially at rate  $2\lambda$ . There are deep connections between  $W_2$  being geodesically convex and  $\mathcal{P}_{ac}^W$  being non-negatively curved as a length space (for more information, check [9]). Before we continue to the next geoemtry, it is worth mentioning briefly about the other concepts, other than the metric, in this Riemannian Manifold. Though not explicitly mentioned in most of the literature, one can construct the connection and covariant derivative over this space, and talk about the Ricci Curvature. But as this is an infinite dimensional Riemannian Manifold (hence called psudo Riemannian), the derivations are not as clean. That's the reason that talking about curvature is much easier if we look into these pdes from a ponit of view of operator theory. We refer the reader to appendix for more details of the Wasserstein geometry, including curvature, connection, and geodesics concepts.

## 2.2. Information Geometry

We now describe the Fisher–Rao metric on the space of positive measures. It is most convenient to start from the embedding  $\mu \mapsto \sqrt{\mu}$  into  $L^2(\Omega)$ . Though it has first been introduced for parametric and hence finite-dimensional probability measures, we refer the reader to [1] for the parametric approach. Here, however, we try to talk more about the infinite-dimensional case.

Assume  $\mu_0, \mu_1 \in \mathcal{M}_+(\Omega)$  admit densities (still denoted by  $\mu_i$ ) with respect to Lebesgue. The Fisher–Rao distance is

$$d_{\text{FR}}^2(\mu_0, \mu_1) := \int_{\Omega} (\sqrt{\mu_0(x)} - \sqrt{\mu_1(x)})^2 dx = \|\sqrt{\mu_0} - \sqrt{\mu_1}\|_{L^2(\Omega)}^2.$$

Thus the mapping  $\mu \mapsto \sqrt{\mu}$  is an isometric embedding of  $(\mathcal{M}_+(\Omega), d_{\text{FR}})$  into the Hilbert space  $L^2(\Omega)$ ; geometrically,  $(\mathcal{M}_+(\Omega), d_{\text{FR}})$  is a (flat) cone.

Consider a smooth curve  $(\mu_t)_{t \in [0,1]} \subset \mathcal{M}_+(\Omega)$  with density  $\mu_t(x)$  and time derivative  $\partial_t \mu_t$ . Differentiating  $\sqrt{\mu_t}$  yields

$$\partial_t \sqrt{\mu_t} = \frac{1}{2} \frac{\partial_t \mu_t}{\sqrt{\mu_t}},$$

and the squared speed in  $L^2$  is

$$\|\partial_t \sqrt{\mu_t}\|_{L^2}^2 = \int_{\Omega} \frac{(\partial_t \mu_t)^2}{4\mu_t} dx.$$

This suggests the following Riemannian metric on  $\mathcal{M}_+(\Omega)$ .

**Definition 5** (Fisher–Rao metric on  $\mathcal{M}_+(\Omega)$ ). The (formal) tangent space at  $\mu \in \mathcal{M}_+(\Omega)$  is

$$T_{\mu} \mathcal{M}_+(\Omega) := \{\dot{\mu} \in L^2_{\text{loc}}(\Omega) : \dot{\mu} \text{ is a signed density}\},$$

with Riemannian metric

$$g_{\mu}^{\text{FR}}(\dot{\mu}, \dot{\mu}) := \int_{\Omega} \frac{\dot{\mu}(x)^2}{4\mu(x)} dx$$

whenever the right-hand side is finite.

It is often convenient to reparametrize tangent vectors multiplicatively. Given a curve  $(\mu_t)$ , define the *reaction rate*

$$\varphi_t(x) := \frac{\partial_t \mu_t(x)}{\mu_t(x)}.$$

Then  $\partial_t \mu_t = \varphi_t \mu_t$ , and the Fisher–Rao metric reads

$$\begin{aligned} g_\mu^{\text{FR}}(\varphi, \varphi) &= g_\mu^{\text{FR}}(\varphi\mu, \varphi\mu) = \int_{\Omega} \frac{\varphi(x)^2 \mu(x)^2}{4\mu(x)} dx \\ &= \frac{1}{4} \int_{\Omega} \varphi(x)^2 d\mu(x). \end{aligned}$$

Up to an inessential constant factor, we may thus identify the tangent space with functions  $\varphi$  and use the inner product

$$\langle \varphi_1, \varphi_2 \rangle_{\mu}^{\text{FR}} := \int_{\Omega} \varphi_1(x) \varphi_2(x) d\mu(x).$$

For *probability* measures, i.e.  $\mu \in \mathcal{P}(\Omega)$ , we additionally impose mass conservation  $\int_{\Omega} \partial_t \mu_t = 0$ , which translates into  $\int_{\Omega} \varphi_t d\mu_t = 0$ . Thus

$$T_{\mu} \mathcal{P}(\Omega) := \left\{ \varphi \in L^2(\mu) : \int_{\Omega} \varphi d\mu = 0 \right\},$$

equipped with the same inner product. Let  $F: \mathcal{M}_+(\Omega) \rightarrow \mathbb{R}$  be a functional with first variation  $\frac{\delta F}{\delta \mu}(\mu)$ . If  $(\mu_t)$  is a curve with  $\partial_t \mu_t = \varphi_t \mu_t$ , then

$$\frac{d}{dt} F(\mu_t) = \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t) \partial_t \mu_t = \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t) \varphi_t d\mu_t.$$

Comparing with the Fisher–Rao inner product, we see that in the unbalanced case

$$\nabla^{\text{FR}} F(\mu) = \frac{\delta F}{\delta \mu}(\mu),$$

as an element of  $T_{\mu} \mathcal{M}_+(\Omega)$  parametrized by  $\varphi$ . In the probabilistic (balanced) setting we must project onto the mean-zero subspace:

$$\nabla^{\text{FR}} F(\mu) = \frac{\delta F}{\delta \mu}(\mu) - \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu) d\mu.$$

The Fisher–Rao gradient flow therefore takes the form

$$\partial_t \mu_t = -\nabla^{\text{FR}} F(\mu_t) \mu_t. \quad (10)$$

In the unbalanced case this reduces to  $\partial_t \mu_t = -\frac{\delta F}{\delta \mu}(\mu_t) \mu_t$ , i.e. a pointwise exponential decay or growth driven by the first variation. In the balanced case, we subtract the spatial mean to keep the total mass equal to one.

### 2.3. Hybrid Geometry, Hellinger–Kantorovich, Wasserstein–Fisher–Rao

The Wasserstein geometry models pure transport of mass; Fisher–Rao models pure creation/annihilation (or, in the probability case, reweighting). In many applications, both mechanisms occur simultaneously: mass is transported in space and its intensity changes in time. The WFR (or Hellinger–Kantorovich) geometry is a natural Riemannian structure on measures that combines both effects in a single metric framework.

Let  $(\mu_t)_{t \in [0,1]}$  be a curve in  $\mathcal{M}_+(\Omega)$ , the cone of finite positive measures on  $\Omega$ . We now consider pairs  $(\varphi_t, v_t)$ , where

$$\varphi_t: \Omega \rightarrow \mathbb{R} \quad v_t: \Omega \rightarrow \mathbb{R}^d \quad (11)$$

coupled through the *continuity equation with reaction*

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = \varphi_t \mu_t. \quad (12)$$

When  $\varphi_t \equiv 0$  this reduces to pure transport (the continuity equation of Wasserstein geometry); when  $v_t \equiv 0$  it reduces to the Fisher–Rao evolution (pure reaction).

In analogy with the Benamou–Brenier formulation (8), one defines a kinetic action functional combining both transport and reaction costs:

$$\mathcal{A}((\mu_t, \varphi_t, v_t)_{t \in [0,1]}) := \int_0^1 \int_{\Omega} (\|v_t(x)\|^2 + \varphi_t(x)^2) d\mu_t(x) dt.$$

The WFR (Hellinger–Kantorovich) distance  $d_{\text{WFR}}$  between two measures is obtained by minimizing this action over all triples  $(\mu_t, \varphi_t, v_t)$  connecting  $\mu_0$  to  $\mu_1$  and satisfying (12):

$$\begin{aligned} d_{\text{WFR}}^2(\mu_0, \mu_1) &:= \inf \left\{ \mathcal{A}((\mu_t, \varphi_t, v_t)_{t \in [0,1]}): (\mu_t, \varphi_t, v_t) \right. \\ &\quad \left. \text{satisfy (12), } \mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1 \right\}. \end{aligned}$$

This dynamic formulation yields a metric on  $\mathcal{M}_+(\Omega)$  that coincides with  $W_2$  when only transport is allowed (i.e.  $\varphi_t \equiv 0$  is enforced), and with  $d_{\text{FR}}$  when only reaction is allowed (i.e.  $v_t \equiv 0$  is enforced). Static (Kantorovich-type) formulations and cone representations are also available in the literature on unbalanced optimal transport, but we emphasize here the Riemannian differential structure suggested by the dynamic formulation.

As in the Wasserstein setting, the pair  $(\varphi_t, v_t)$  satisfying (12) is not unique. There is a gauge freedom (adding divergence-free  $v_t$  and adjusting  $\varphi_t$  appropriately) that leaves  $\partial_t \mu_t$  unchanged. One can again show that there is a unique “minimal norm” representative in the sense of the action  $\mathcal{A}$ , and that it can be parametrized by a scalar potential  $\phi_t: \Omega \rightarrow \mathbb{R}$ , with

$$\varphi_t = \phi_t, \quad v_t = \nabla \phi_t.$$

Formally, this identifies the tangent directions at  $\mu$  with potentials  $\phi$  modulo  $\mu$ -almost everywhere constants, in complete analogy with the Otto calculus where gradients  $\nabla \phi$  represent transport directions.

This leads to the following (formal) tangent space (see, for instance, Chapters 6 and 7 of [9]):

**Definition 6** (Wasserstein–Fisher–Rao tangent space). For  $\mu \in \mathcal{M}_+(\Omega)$ , the formal WFR tangent space is

$$T_{\mu}^{\text{WFR}} \mathcal{M}_+(\Omega) := \overline{\{(\phi, \nabla \phi) : \phi \in C_c^{\infty}(\Omega)\}}^{L^2(\mu)},$$

equipped with the inner product

$$\begin{aligned} \langle (\phi_1, \nabla \phi_1), (\phi_2, \nabla \phi_2) \rangle_{\mu}^{\text{WFR}} &:= \int_{\Omega} (\phi_1(x) \phi_2(x) + \nabla \phi_1(x) \cdot \nabla \phi_2(x)) d\mu(x), \quad (13) \end{aligned}$$

and associated norm

$$\|(\phi, \nabla \phi)\|_{\mu}^2 = \int_{\Omega} (\phi(x)^2 + \|\nabla \phi(x)\|^2) d\mu(x).$$

Thus the WFR metric completes the Wasserstein metric  $\|\nabla \phi\|_{L^2(\mu)}$  by adding the zeroth-order term  $\|\phi\|_{L^2(\mu)}$ , i.e. the full  $H^1$ -type norm of the potential  $\phi$  with respect to  $\mu$ . From the Riemannian viewpoint, this is precisely the inner product induced by the operator

$$\phi \mapsto (I - \Delta_{\mu})\phi, \quad \Delta_{\mu} := \nabla \cdot (\mu \nabla),$$

on potentials, so that the WFR geometry can be seen as an  $H^1$ -Sobolev deformation of the  $L^2$ -based Wasserstein geometry.

Restricting to probability measures  $\mathcal{P}(\Omega) \subset \mathcal{M}_+(\Omega)$ , one imposes the additional constraint  $\int_{\Omega} \phi d\mu = 0$  to ensure conservation of total mass (so that  $\int_{\Omega} \varphi_t d\mu_t = 0$ ).

Let  $F: \mathcal{M}_+(\Omega) \rightarrow \mathbb{R}$  have first variation  $\frac{\delta F}{\delta \mu}(\mu)$ . Consider a curve  $(\mu_t)$  with tangent  $(\varphi_t, v_t)$  satisfying (12). We compute

$$\begin{aligned}\frac{d}{dt}F(\mu_t) &= \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t) \partial_t \mu_t dx \\ &= \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t) (-\nabla \cdot (\mu_t v_t) + \varphi_t \mu_t) dx \\ &= \int_{\Omega} \nabla \left( \frac{\delta F}{\delta \mu}(\mu_t) \right) \cdot v_t d\mu_t + \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t) \varphi_t d\mu_t,\end{aligned}$$

where we integrated by parts in the transport term. If we parametrize  $(\varphi_t, v_t)$  by  $(\phi_t, \nabla \phi_t)$ , then the right-hand side becomes

$$\int_{\Omega} \left( \frac{\delta F}{\delta \mu}(\mu_t) \phi_t + \nabla \left( \frac{\delta F}{\delta \mu}(\mu_t) \right) \cdot \nabla \phi_t \right) d\mu_t,$$

which we recognize as the dual pairing with

$$\left( \frac{\delta F}{\delta \mu}(\mu_t), \nabla \left( \frac{\delta F}{\delta \mu}(\mu_t) \right) \right) \in T_{\mu_t}^{\text{WFR}} \mathcal{M}_+(\Omega).$$

**Definition 7** (WFR gradient). The (formal) WFR gradient of  $F$  at  $\mu$  is the element

$$\nabla^{\text{WFR}} F(\mu) = \left( \frac{\delta F}{\delta \mu}(\mu), \nabla \left( \frac{\delta F}{\delta \mu}(\mu) \right) \right) \in T_{\mu}^{\text{WFR}} \mathcal{M}_+(\Omega).$$

In the probabilistic (balanced) setting we again subtract the mean in the first component to enforce  $\int \phi d\mu = 0$ , i.e.

$$\nabla^{\text{WFR}} F(\mu) = \left( \frac{\delta F}{\delta \mu}(\mu) - \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu) d\mu, \nabla \left( \frac{\delta F}{\delta \mu}(\mu) \right) \right).$$

The WFR gradient flow is defined by following the steepest descent direction:

$$(\varphi_t, v_t) = -\nabla^{\text{WFR}} F(\mu_t).$$

Plugging into (12) gives the PDE

$$\partial_t \mu_t + \nabla \cdot \left( \mu_t \nabla \left( \frac{\delta F}{\delta \mu}(\mu_t) \right) \right) = - \left( \frac{\delta F}{\delta \mu}(\mu_t) - \int_{\Omega} \frac{\delta F}{\delta \mu}(\mu_t) d\mu_t \right) \mu_t. \quad (14)$$

In the unbalanced case one drops the mean-subtraction term. The left-hand side corresponds to Wasserstein-type transport, while the right-hand side corresponds to Fisher–Rao-type reaction. As in the pure Wasserstein and Fisher–Rao cases,

$$\frac{d}{dt} F(\mu_t) = -\|\nabla^{\text{WFR}} F(\mu_t)\|_{\mu_t}^2 \leq 0,$$

so the flow is again a steepest descent of  $F$ , now in the WFR geometry.

We conclude this section by summarizing the three geometric frameworks introduced above—Wasserstein, Fisher–Rao, and Wasserstein–Fisher–Rao (WFR)—highlighting their structural similarities and differences through their tangent spaces, metrics, geodesics, curvature properties, and associated gradient flows. Taken together, these geometries form a coherent hierarchy: Fisher–Rao encodes pure reaction, Wasserstein encodes pure transport, and WFR merges the two in a Sobolev-type structure.

A core theme in the geometric analysis of probability spaces is the intimate relation between *curvature*, *convexity*, and *stability* of associated differential equations.

- In **Wasserstein geometry**, Ricci-like curvature enters through the Bakry–Émery  $\Gamma_2$  condition. If  $F$  is  $\lambda$ -geodesically convex in  $W_2$  (i.e.  $F''(\mu_t) \geq \lambda$  along  $W_2$  geodesics), then the  $W_2$ -gradient flow of  $F$  is contractive:

$$W_2(\mu_t, \nu_t) \leq e^{-\lambda t} W_2(\mu_0, \nu_0),$$

and this corresponds to exponential decay of solutions of Fokker–Planck-type PDEs.

- In **Fisher–Rao geometry**, the metric is flat when lifted to the Hellinger cone, but submanifolds (e.g. exponential families) often exhibit *negative sectional curvature*, reflecting the strict convexity of the log-partition function. This curvature controls the stability of statistical estimators and likelihood flows.
- In **WFR geometry**, curvature arises from a combination of Wasserstein’s (second-order, transport-based) geometry and Fisher–Rao’s (first-order, reaction-based) geometry. The WFR metric induces a generalized convexity notion for functionals: geodesic convexity in WFR implies exponential convergence of the mixed transport–reaction PDE (14).

In all three geometries, convexity of an energy functional  $F$  along the appropriate geodesics determines the long-time behavior of its gradient flow, providing a unified explanation of why Wasserstein, Fisher–Rao, and WFR evolutions exhibit exponential convergence, entropy dissipation, or contractive stability.

Each geometry supports a coherent differential calculus (tangent spaces, gradients, connections, curvature), a natural geodesic structure, and a steepest-descent interpretation of PDEs. The hybrid viewpoint is essential for applications in which both spatial rearrangement and local mass variation play critical roles—ranging from diffusion–reaction systems to generative models and geometric flows.

Now, before going to further analyze these concepts, using the theory of linear operators, we try to come up with an sde for each of the pdes introduced above.

### 3. FROM PDE TO SDE

The previous section endowed spaces of measures with Riemannian structures of Wasserstein, Fisher–Rao, and Wasserstein–Fisher–Rao type, that potentially can help us to sample better, and faster. In the present section we take the complementary viewpoint: starting from a partial differential equation (PDE) for probability densities, we construct weighted stochastic processes on the underlying state space whose laws realize the prescribed evolution. In other words, we systematically pass *from PDE to SDE*, enabling implementation.

Now that we have seen the different pdes provided in the previous section, we provide a corresponding SDE, in this case a weighted sde, to be able to simulate the aforementioned pdes. In general consider:

$$\begin{aligned}\frac{\partial}{\partial t} p_t(x_t) &= -\nabla \cdot (p_t(x_t) v_t(x_t)) + \frac{\sigma_t^2}{2} \Delta p_t(x_t) \\ &+ p_t(x_t) \left( \psi_t(x_t) - \int \psi_t(x_t) p_t(x_t) dx_t \right) \quad (\text{General PDE})\end{aligned}$$

where to sample from  $p_t(x)$ , one first has to sample  $x_t$  via The following SDE

$$dx_t = v_t(x_t) dt + \sigma_t dB_t, \quad dw_t = \bar{\psi}_t(x_t) dt, \quad (15)$$

Geometry	Tangent Space $T_\mu$	Metric $g_\mu$
Wasserstein ( $W_2$ )	$T_\mu = \overline{\{\nabla \phi\}}^{L^2(\mu)}$	$g_\mu(v, w) = \int_{\Omega} v \cdot w \, d\mu$
Fisher–Rao (FR)	$T_\mu = \{\varphi : \int \varphi \, d\mu = 0\}$	$g_\mu(\varphi, \psi) = \int_{\Omega} \varphi \psi \, d\mu$
WFR (HK)	$T_\mu^{\text{WFR}} = \overline{\{(\phi, \nabla \phi)\}}^{L^2(\mu)}$	$g_\mu^{\text{WFR}}((\phi, \nabla \phi), (\psi, \nabla \psi)) = \int_{\Omega} (\phi \psi + \nabla \phi \cdot \nabla \psi) \, d\mu$

**Table 1:** Tangent spaces and Riemannian metrics.

Geometry	Geodesics
Wasserstein ( $W_2$ )	$\partial_t \mu_t + \nabla \cdot (\mu_t \nabla \phi_t) = 0, \quad \partial_t \phi_t + \frac{1}{2}  \nabla \phi_t ^2 = c(t).$ (McCann displacement interpolation: $\mu_t = ((1-t)\text{id} + tT_{\mu_0 \rightarrow \mu_1})\# \mu_0$ )
Fisher–Rao (FR)	$\sqrt{\mu_t} = (1-t)\sqrt{\mu_0} + t\sqrt{\mu_1}; \quad \mu_t = ((1-t)\sqrt{\mu_0} + t\sqrt{\mu_1})^2.$ (Geodesics are great circles in $L^2$ restricted to positive cone.)
WFR (HK)	Minimizers of the action: $\int_0^1 \int_{\Omega} (\phi_t^2 +  \nabla \phi_t ^2) \, d\mu_t \, dt$ subject to $\partial_t \mu_t + \nabla \cdot (\mu_t \nabla \phi_t) = \phi_t \mu_t$ .

**Table 2:** Geodesics in Wasserstein, Fisher–Rao, and WFR geometries.

Geometry	Gradient Flow of $F$
Wasserstein ( $W_2$ )	$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla (\frac{\delta F}{\delta \mu}(\mu_t)))$ (Steepest descent in $W_2$ , e.g. heat flow for $F(\mu) = \int \mu \log \mu$ )
Fisher–Rao (FR)	$\partial_t \mu_t = -\left(\frac{\delta F}{\delta \mu} - \int \frac{\delta F}{\delta \mu} \, d\mu\right) \mu_t$ (Pointwise exponential reweighting.)
WFR (HK)	$\partial_t \mu_t + \nabla \cdot (\mu_t \nabla \frac{\delta F}{\delta \mu}) = -\left(\frac{\delta F}{\delta \mu} - \int \frac{\delta F}{\delta \mu} \, d\mu\right) \mu_t$ (Combined transport + reaction.)

**Table 3:** Gradient flows as steepest descents in each geometry.

and then reweight the obtained samples using  $w_t$ .

In practice, we can account for this difference by sampling

$$i \sim \text{Categorical} \left\{ \frac{\exp(w_T^k)}{\sum_{j=1}^K \exp(w_T^j)} \right\}_{k=1}^K, \quad (16)$$

and returning  $x_T^{(i)}$  as an approximate sample from  $p_T$ . For estimating the expectation of test functions  $\phi$ , we account for the weights by reweighting a collection of  $K$  particles, i.e.,

$$\mathbb{E}_{p_T}[\phi(x)] \approx \sum_{k=1}^K \frac{\exp(w_T^k)}{\sum_j \exp(w_T^j)} \phi(x_T^k). \quad (17)$$

In other words, one can start  $K$  particles  $X_t^i$  each following a drift and diffusion and proceed the spatio movement in a discrete version until interval  $[0, t_1]$ . Along the path, we calculate  $w_{t_1}^i =$

$\exp(\psi_t(X_{t_1}^i))$ , and then, resample the particles with these weights. As proved by the theorem below, this would indeed realizes the pde.

**Remark.** The exploration in space, is still done by the Brownian motion. Though, by introducing this kill/birth process done by weights, we are relocating particles to mimic the teleportation we need. There is a Jump Process interpretation of Weighted SDE above as well discussed later in the text.

**Theorem 8.** *The Weighted SDE, and sampling scheme introduced above, would realize the generalized pde,i.e.*

$$\text{law}(X_t^i) \sim p^t$$

**Proof.** We proceed in two steps, first finding a Kolmogorov backward equation corresponding to evolution under a weighted Feynman–Kac SDE. We then use this identity to derive the expectation estimator.

**Proposition A.1.** For a bounded test function  $\phi : X \rightarrow \mathbb{R}$  and  $p_t$  satisfying Eq. (General PDE), we have

$$\mathbb{E}_{p_T(x_T)}[\phi(x_T)] = \frac{1}{Z_T} \mathbb{E}\left[\exp\left(\int_0^T \psi_s(x_s) ds\right) \phi(x_T)\right] \quad (18)$$

where  $dx_t = v_t(x_t) dt + \sigma_t dB_t$ ,  $x_0 \sim p_0$ .

Here  $Z_T$  is a normalization constant independent of  $x$ . Eq. (18) suggests that the self-normalized importance sampling approximation in Eq. 17 is consistent as  $K \rightarrow \infty$ .

**Proof.** The proof proceeds in three steps. We first derive the backward Kolmogorov equation for appropriate functions, then specify the evolution of the General PDE for the unnormalized density

For a given test function  $\phi(x)$ , consider defining the following function:

$$\begin{aligned} \Phi_T(x, t) &= \mathbb{E}\left[\exp\left(\int_t^T \psi_s(x_s) ds\right) \phi(x_T) \mid x_t = x\right], \\ \Phi_T(x, T) &= \phi(x). \end{aligned} \quad (19)$$

For any  $\tau > t$ , we have

$$\begin{aligned} \Phi_T(x, t) &= \mathbb{E}\left[\exp\left(\int_t^\tau \psi_s(x_s) ds\right) \exp\left(\int_\tau^T \psi_s(x_s) ds\right) \phi(x_T) \mid x_t = x\right] \\ &= \mathbb{E}\left[\exp\left(\int_t^\tau \psi_s(x_s) ds\right) \Phi_T(x_\tau, \tau) \mid x_t = x\right] \end{aligned} \quad (20)$$

To relate  $\Phi_T(x, t)$  and the expected value of  $\Phi_T(x_\tau, \tau)$ , we apply Itô's product rule. For  $\tau \mapsto x_\tau$ :

$$\begin{aligned} d\left(e^{\int_t^\tau \psi_s ds} \Phi_T(x_\tau, \tau)\right) &= e^{\int_t^\tau \psi_s ds} \left( \frac{\partial \Phi_T}{\partial \tau} + \langle v_\tau(x_\tau), \nabla \Phi_T(x_\tau, \tau) \rangle + \frac{\sigma_\tau^2}{2} \Delta \Phi_T(x_\tau, \tau) \right) \\ &\quad + e^{\int_t^\tau \psi_s ds} \sigma_\tau \langle \nabla \Phi_T, dW_\tau \rangle. \end{aligned} \quad (21)$$

Taking expectations removes the martingale term. Thus  $\Phi_T$  satisfies the backward PDE:

$$\begin{aligned} \frac{\partial \Phi_T(x_\tau, \tau)}{\partial \tau} + \langle v_\tau(x_\tau), \nabla \Phi_T(x_\tau, \tau) \rangle \\ + \frac{\sigma_\tau^2}{2} \Delta \Phi_T(x_\tau, \tau) + \psi_\tau(x_\tau) \Phi_T(x_\tau, \tau) = 0. \end{aligned} \quad (22)$$

In practice, we cannot exactly calculate

$$\int \psi_t(x_t) p_t(x_t) dx_t,$$

which appears in Eq. (25) below to ensure normalization.

For now, consider the evolution of the *unnormalized* density

$$\tilde{p}_t(x) = p_t(x) Z_t,$$

for a particular  $v_t, \sigma_t, \psi_t$  and normalization constant  $Z_t$ . With foresight, we define:

$$\frac{\partial}{\partial t} \tilde{p}_t(x_t) = -\nabla \cdot (\tilde{p}_t(x_t) v_t(x_t)) + \frac{\sigma_t^2}{2} \Delta \tilde{p}_t(x_t) + \tilde{p}_t(x_t) \psi_t(x_t). \quad (23)$$

We further define the normalization evolution:

$$\partial_t \log Z_t := \int p_t(x) \psi_t(x) dx. \quad (24)$$

This choice is motivated by the reweighting-only evolution

$$\partial_t p_t^w(x) = p_t^w(x) (\psi_t(x) - \int p_t^w(x) \psi_t(x) dx),$$

which implies:

$$\partial_t \log p_t^w(x) = \psi_t(x) - \int p_t^w(x) \psi_t(x) dx.$$

Writing  $p_t^w(x) = \tilde{p}_t^w(x)/Z_t$  and equating derivatives yields Eq. (24). We now verify that the definitions (23)–(24) are consistent with the original pde:

$$\begin{aligned} \frac{\partial}{\partial t} p_t(x_t) &= -\nabla \cdot (p_t(x_t) v_t(x_t)) + \frac{\sigma_t^2}{2} \Delta p_t(x_t) \\ &\quad + p_t(x_t) \left( \psi_t(x_t) - \int \psi_t(x_t) p_t(x_t) dx_t \right). \end{aligned} \quad (25)$$

Since  $p_t(x) = \tilde{p}_t(x) Z_t^{-1}$ , we compute:

$$\frac{\partial}{\partial t} p_t(x_t) = \frac{\partial}{\partial t} (\tilde{p}_t(x_t) Z_t^{-1}) = Z_t^{-1} \frac{\partial \tilde{p}_t}{\partial t} + \tilde{p}_t(x_t) Z_t^{-1} \partial_t (Z_t^{-1}). \quad (26)$$

Using

$$\partial_t (Z_t^{-1}) = -Z_t^{-1} \partial_t \log Z_t,$$

Eq. (26) becomes:

$$\frac{\partial}{\partial t} p_t(x_t) = Z_t^{-1} \frac{\partial \tilde{p}_t}{\partial t} - \tilde{p}_t(x_t) Z_t^{-1} \partial_t \log Z_t. \quad (27)$$

Inserting Eq. (23):

$$\begin{aligned} \frac{\partial^2}{\partial t^2} p_t(x_t) &= Z_t^{-1} \left( -\nabla \cdot (\tilde{p}_t v_t) + \frac{\sigma_t^2}{2} \Delta \tilde{p}_t + \tilde{p}_t \psi_t \right) \\ &\quad - \tilde{p}_t Z_t^{-1} \int p_t(x) \psi_t(x) dx. \end{aligned} \quad (28)$$

Since  $\nabla Z_t = 0$ , we may move  $Z_t^{-1}$  inside derivatives:

$$\begin{aligned} \frac{\partial}{\partial t} p_t(x_t) &= -\nabla \cdot (p_t(x_t) v_t(x_t)) + \frac{\sigma_t^2}{2} \Delta p_t(x_t) \\ &\quad + p_t(x_t) \psi_t(x_t) - p_t(x_t) \int p_t(x) \psi_t(x) dx. \end{aligned} \quad (29)$$

This matches precisely the original PDE (25), completing the consistency check.

**Expectation Estimation.** Now, we use Eq. (22) to write the total derivative of the following integral under the unnormalized density  $\tilde{p}_t(x)$ :

$$\begin{aligned} \frac{d}{dt} \left[ \int \Phi_T(x, t) \tilde{p}_t(x) dx \right] &= \int \left( \frac{\partial \Phi_T(x, t)}{\partial t} \right) \tilde{p}_t(x) dx \\ &\quad + \int \Phi_T(x, t) \left( \frac{\partial \tilde{p}_t(x)}{\partial t} \right) dx. \end{aligned} \quad (30)$$

Using Eqs. (22) and (23), we obtain

$$\begin{aligned} \frac{d}{dt} \int \Phi_T(x, t) \tilde{p}_t(x) dx &= \\ \int \left( -\langle v_t(x), \nabla \Phi_T(x, t) \rangle - \frac{\sigma_t^2}{2} \Delta \Phi_T(x, t) - \Phi_T(x, t) \psi_t(x) \right) \tilde{p}_t(x) dx \end{aligned}$$

$$+ \int \Phi_T(x, t) \left( -\nabla \cdot (\tilde{p}_t(x) v_t(x)) + \frac{\sigma_t^2}{2} \Delta \tilde{p}_t(x) + \tilde{p}_t(x) \psi_t(x) \right) dx. \quad (31)$$

$$\frac{\partial p_t^{\text{jump}}(x)}{\partial t} = \left( \int \lambda_t(y) J_t(x|y) p_t(y) dy \right) - p_t(x) \lambda_t(x) \quad (39)$$

Integrating by parts, the second line becomes:

$$\begin{aligned} & \frac{d}{dt} \int \Phi_T(x, t) \tilde{p}_t(x) dx \\ &= \int \left( -\langle v_t(x), \nabla \Phi_T(x, t) \rangle - \frac{\sigma_t^2}{2} \Delta \Phi_T(x, t) - \Phi_T(x, t) \psi_t(x) \right) \tilde{dp}_t(x) \\ &+ \int \left( \langle v_t(x), \nabla \Phi_T(x, t) \rangle + \frac{\sigma_t^2}{2} \Delta \Phi_T(x, t) + \Phi_T(x, t) \psi_t(x) \right) \tilde{dp}_t(x) \\ &= 0. \end{aligned}$$

Thus,

$$\frac{d}{dt} \int \Phi_T(x, t) \tilde{p}_t(x) dx = 0. \quad (32)$$

Integrating over  $t \in [0, T]$ , we obtain:

$$\begin{aligned} & \int \Phi_T(x_T, T) \tilde{p}_T(x_T) dx_T - \int \Phi_T(x_0, 0) \tilde{p}_0(x_0) dx_0 = \\ & \int_0^T \frac{d}{dt} \left[ \int \Phi_T(x, t) \tilde{p}_t(x) dx \right] dt = 0. \quad (33) \end{aligned}$$

Hence these two quantities must be equal:

$$\int \Phi_T(x_0, 0) \tilde{p}_0(x_0) dx_0 = \int \Phi_T(x_T, T) \tilde{p}_T(x_T) dx_T. \quad (34)$$

Using  $\tilde{p}_0 = p_0$  and  $Z_0 = 1$ , and the fact that

$$\Phi_T(x_T, T) = \phi(x_T),$$

we have:

$$\int \mathbb{E} \left[ e^{\int_0^T \psi_s(x_s) ds} \phi(x_T) \mid x_0 \right] p_0(x_0) dx_0 = Z_T \int \phi(x_T) p_T(x_T) dx_T. \quad (35)$$

Thus,

$$\frac{1}{Z_T} \mathbb{E} \left[ e^{\int_0^T \psi_s(x_s) ds} \phi(x_T) \right] = \mathbb{E}_{p_T} [\phi(x_T)], \quad (36)$$

which is precisely the identity claimed in Prop. A.1.

In practice, we approximate

$$Z_T \approx \frac{1}{K} \sum_{k=1}^K \exp \left( \int_0^T \psi_s(x_s^{(k)}) ds \right) = \frac{1}{K} \sum_{k=1}^K e^{w_T^{(k)}}, \quad (37)$$

and similarly

$$\mathbb{E} \left[ e^{\int_0^T \psi_s(x_s) ds} \phi(x_T) \right] \approx \frac{1}{K} \sum_{k=1}^K e^{w_T^{(k)}} \phi(x_T^{(k)}). \quad (38)$$

This yields Eq.17 We emphasize that the choice of terminal time  $T$  was arbitrary; the same reasoning applies to any intermediate  $t$ , which implies that samples are correctly weighted for estimating expectations at all intermediate times.

**Jump processes and pure reaction dynamics.** A Markov jump process is determined by a rate function  $\lambda_t(x)$ , which governs the frequency of jump events, and a Markov transition kernel  $J_t(y|x)$ , which is used to sample the next state when a jump occurs. The forward Kolmogorov equation for a jump process is given by

where the two terms can intuitively be seen to measure the inflow and outflow of probability due to jumps.

One could find  $\lambda_t(x)$ ,  $J_t(y|x)$  such that  $p_t^{\text{jump}}$  matches the evolution of  $p_t^w$  in General PDE for a given choice of  $\psi_t$ . For a given  $\psi_t$ , define the jump process rate and transition as

$$\lambda_t(x) = (\psi_t(x) - \mathbb{E}_{p_t} [\psi_t])^-, \quad (40)$$

$$J_t(y|x) = \frac{(\psi_t(y) - \mathbb{E}_{p_t} [\psi_t])^+ p_t(y)}{\int (\psi_t(z) - \mathbb{E}_{p_t} [\psi_t])^+ p_t(z) dz}, \quad (41)$$

where  $(u)^- := \max(0, -u)$  and  $(u)^+ := \max(0, u)$ . Then,

$$\frac{\partial p_t^{\text{jump}}(x)}{\partial t} = \frac{\partial p_t^w(x)}{\partial t} = p_t(x) (\psi_t(x) - \mathbb{E}_{p_t} [\psi_t]). \quad (42)$$

In continuous time and the mean-field limit, this jump process formulation of reweighting corresponds to simulating

$$x_{t+\Delta t} = \begin{cases} x_t & \text{w.p. } 1 - \lambda_t(x_t) dt + o(dt), \\ \sim J_t(y|x_t) & \text{w.p. } \lambda_t(x_t) dt + o(dt). \end{cases} \quad (33)$$

We expect this process to improve the sample population in efficient fashion, since jump events are triggered only in states where  $(\psi_t(x) - \mathbb{E}_{p_t} [\psi_t])^- \geq 0 \iff \psi_t(x) \leq \mathbb{E}_{p_t} [\psi_t]$ , and transitions are more likely to jump to states with high excess weight  $(\psi_t(y) - \mathbb{E}_{p_t} [\psi_t])^+ > 0$ .

**Proof[14].** One way to perform a simulation of the reweighting equation is to rewrite it as a jump process. We recall the definition of the Markov generator of a jump process. Let  $W_t(x, y) = \lambda_t(x) J_t(y|x)$  where  $J_t(y|x)$  is normalized. Then

$$J_t^{(W)}[\phi](x) = \int (\phi(y) - \phi(x)) \lambda_t(x) J_t(y|x) dy, \quad (43)$$

and the adjoint generator satisfies

$$J_t^{*(W)}[p_t](x) = \left( \int \lambda_t(y) J_t(x|y) p_t(y) dy \right) - p_t(x) \lambda_t(x). \quad (44)$$

**Proof.**

$$\begin{aligned} \int \phi(x) J_t^*[p_t](x) dx &= \int J_t[\phi](x) p_t(x) dx \\ &= \int \int (\phi(y) - \phi(x)) \lambda_t(x) J_t(y|x) p_t(y) dy p_t(x) dx \\ &= \int \int \phi(y) \lambda_t(x) J_t(y|x) p_t(x) dy dx \\ &\quad - \int \int \phi(x) \lambda_t(x) J_t(y|x) p_t(x) dy dx \\ &= \int \int \phi(x) \lambda_t(y) J_t(x|y) p_t(y) dx dy \\ &\quad - \int \phi(x) \lambda_t(x) p_t(x) \left( \int J_t(y|x) dy \right) dx \\ &= \int \phi(x) \left[ \int \lambda_t(y) J_t(x|y) p_t(y) dy - p_t(x) \lambda_t(x) \right] dx, \end{aligned}$$

proving Eq. (44).

We aim to construct a jump process whose adjoint generator matches the reweighting generator:

$$J_t^{*(W)}[p_t](x) = L_t^{*(g)}[p_t](x). \quad (45)$$

Define the positive/negative parts:

$$(u)_- := \max(0, -u), \quad (u)_+ := \max(0, u), \quad (u)_+ - (u)_- = u.$$

Following Angeli et al. (2019) and Del Moral (2013), we define:

$$\begin{aligned} \lambda_t(x) &= (\psi_t(x) - \mathbb{E}_{p_t}[\psi_t])_-, \\ J_t(y|x) &= \frac{(\psi_t(y) - \mathbb{E}_{p_t}[\psi_t])_+ p_t(y)}{\int (\psi_t(z) - \mathbb{E}_{p_t}[\psi_t])_+ p_t(z) dz}. \end{aligned} \quad (46)$$

For the choice of  $\lambda_t$  and  $J_t$  in (46), the adjoint generator satisfies

$$J_t^{*(W)}[p_t](x) = L_t^{*(g)}[p_t](x) = p_t(x) \left( \psi_t(x) - \int \psi_t(x) p_t(x) dx \right). \quad (47)$$

**Proof.** Expanding Eq. (44):

$$\begin{aligned} J_t^{*(W)}[p_t](x) &= \int (\psi_t(y) - \mathbb{E}_{p_t}[\psi_t])_- \frac{(\psi_t(x) - \mathbb{E}_{p_t}[\psi_t])_+ p_t(y)}{\int (\psi_t(z) - \mathbb{E}_{p_t}[\psi_t])_+ p_t(z) dz} dy \\ &\quad - p_t(x) (\psi_t(x) - \mathbb{E}_{p_t}[\psi_t])_-. \end{aligned} \quad (48)$$

Using the identity

$$\int (\psi_t(z) - \mu)_+ p_t(z) dz = \int (\psi_t(z) - \mu)_- p_t(z) dz$$

(which follows from  $\mathbb{E}_{p_t}[\psi_t - \mu] = 0$ ), and splitting cases  $\psi_t(x) \geq \mu$  or not, one obtains:

$$J_t^{*(W)}[p_t](x) = p_t(x) (\psi_t(x) - \mathbb{E}_{p_t}[\psi_t]) = L_t^{*(g)}[p_t](x).$$

#### 4. CONCLUSION AND FUTURE DIRECTIONS

This work presents a foundational investigation into the role of information-geometric structure in diffusion-based sampling, with particular emphasis on Wasserstein–Fisher–Rao (WFR) geometry as a unifying framework for transport and reweighting mechanisms. By interpreting diffusion, weighted diffusion, and reaction dynamics through a geometric and operator-theoretic lens, we demonstrate how classical Ornstein–Uhlenbeck–type sampling processes can be systematically augmented without abandoning their underlying stochastic calculus. Beyond its immediate constructions, this study opens several mathematically substantive directions for future research, including (but not limited to) the following:

- **Spectral and semigroup analysis of corrected generators.**

A comprehensive understanding of diffusion–reweighting and diffusion–jump dynamics requires the spectral analysis of the associated (typically unbounded and non-selfadjoint) generators. Tools from semigroup theory, most notably the Hille–Yosida theorem and its extensions to non-conservative and nonlocal operators, provide a natural framework for characterizing well-posedness, ergodicity, and convergence rates of these corrected processes. Establishing quantitative links between geometric corrections and spectral gaps remains a central open problem.

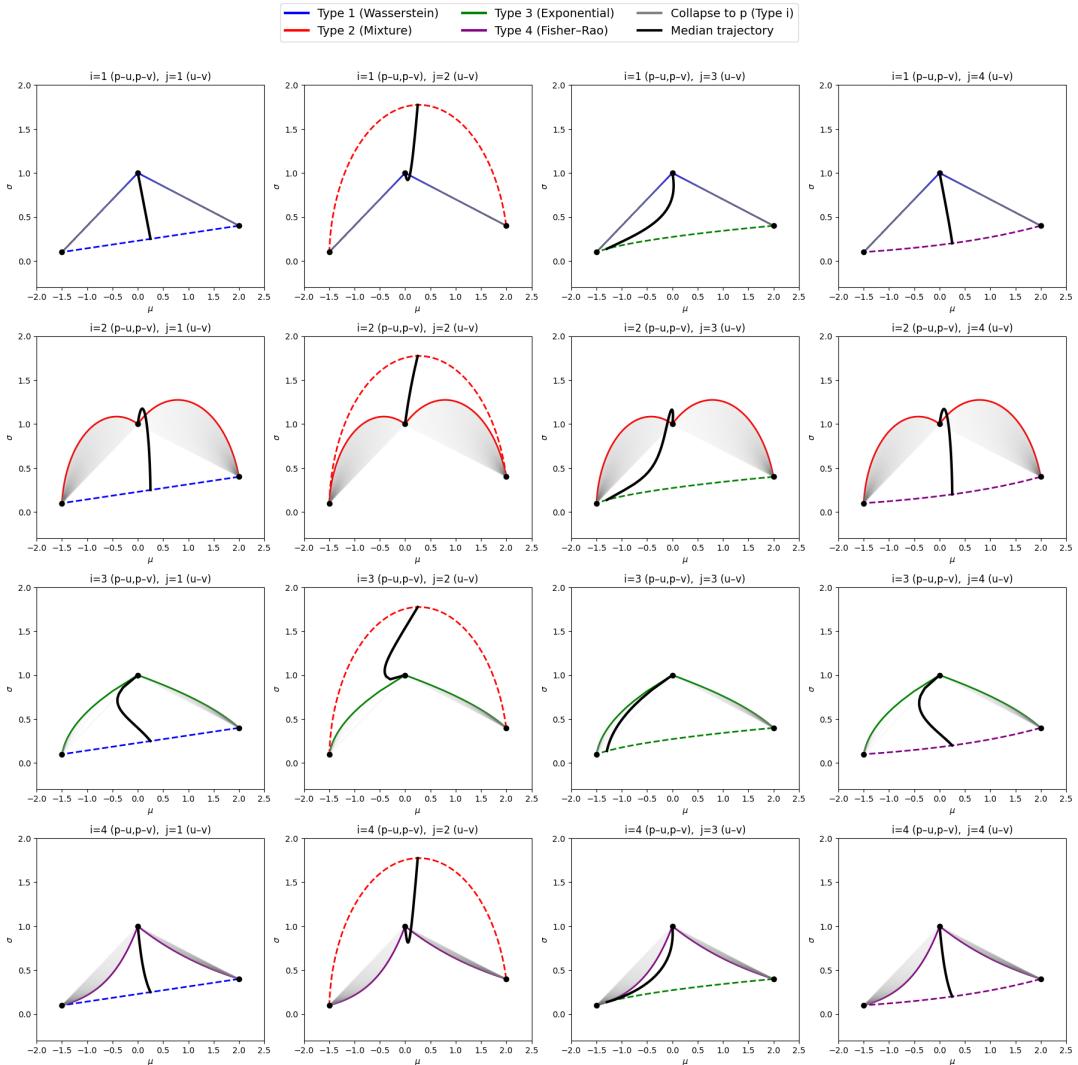
- **Interaction of geometric geodesics and sampling efficiency.**

In classical diffusion models, entropic optimal transport (equivalently, Schrödinger bridges) characterizes the geodesic structure induced by Ornstein–Uhlenbeck dynamics. Within the WFR framework, however, additional geodesic families arise—such as mixture (linear), exponential, and information-preserving interpolations that each encodes a distinct tradeoff between transport and mass variation. Understanding the interaction and possible hybridization of these geodesics is essential for designing sampling schemes capable of traversing mixture distributions and multimodal targets *without* retraining score functions, thereby enabling geometry-driven adaptation at the level of inference rather than model fitting (Look Fig. 1 and Last Appendix).

We believe that further development along these directions will deepen the theoretical foundations of diffusion-based generative modeling and contribute to a principled synthesis of stochastic analysis, optimal transport, and information geometry in modern sampling theory.

#### 5. REFERENCES

- [1] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, Tokyo, 2016.
- [2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2008.
- [3] David Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, 2 edition, 2009.
- [4] Dominique Bakry and Michel Émery. Diffusions hypercontractives. *Séminaire de probabilités de Strasbourg*, 19:177–206, 1985.
- [5] Anton Bovier and Frank den Hollander. *Metastability: A Potential-Theoretic Approach*, volume 351 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2015.
- [6] Anton Bovier, Vincent Gayrard, Anne Klein, et al. Eyring–Kramers law for the underdamped Langevin process. *arXiv preprint*, 2025.
- [7] Anton Bovier, Vincent Gayrard, Anne Klein, et al. Eyring–Kramers law for the underdamped Langevin process. *arXiv preprint*, 2025.
- [8] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323, 2018.
- [9] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, November 2024. arXiv:2407.18163 [math].



**Fig. 1:** Geodesic structure and median trajectories across different geometries in the  $(\mu, \sigma)$  parameter space. Each panel depicts a triangle formed by three distributions  $(p, u, v)$ , where the edges connecting  $p$  to  $u$  and  $p$  to  $v$  are constructed using a fixed geometry indexed by  $i \in \{1, 2, 3, 4\}$ , while the edge connecting  $u$  and  $v$  is generated using a (possibly different) geometry indexed by  $j \in \{1, 2, 3, 4\}$ . Type 1 corresponds to Wasserstein geodesics (blue), Type 2 to linear mixture geodesics (red), Type 3 to exponential geodesics (green), and Type 4 to Fisher-Rao geodesics (purple). Gray segments indicate collapse trajectories toward the reference distribution  $p$  under the geometry indexed by  $i$ , and the black curve denotes the induced median trajectory between  $u$  and  $v$ , obtained by projecting the  $u-v$  geodesic through  $p$  according to the pair  $(i, j)$ . Dashed curves represent the direct  $u-v$  geodesic in geometry  $j$ , while solid colored curves illustrate the lifted or corrected paths resulting from the interaction between the two geometric structures. The collection of panels highlights how mismatches between transport and information-geometric structures modify both geodesic shapes and the resulting median trajectories.

Component	PDE Contribution	Generator $L_t$ acting on $\phi$
<b>Drift</b>	$\partial_t p_t = -\nabla \cdot (p_t v_t)$	$L_t^{(v)}[\phi](x) = \langle \nabla \phi(x), v_t(x) \rangle$
<b>Diffusion</b>	$\partial_t p_t = \frac{\sigma_t^2}{2} \Delta p_t$	$L_t^{(\sigma)}[\phi](x) = \frac{\sigma_t^2}{2} \Delta \phi(x)$
<b>Reweighting</b>	$\partial_t p_t = p_t(x) \left( \psi_t(x) - \int \psi_t \, dp_t \right)$	$L_t^{(g,p)}[\phi](x) = \phi(x) \left( \psi_t(x) - \int \psi_t(x) p_t(x) \, dx \right)$

**Table 4:** Decomposition of the pde into drift, diffusion, and reweighting components, showing the corresponding PDE contribution, infinitesimal generator acting on test functions.

- [10] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11), 2018.
- [11] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.
- [12] Arnak S. Dalalyan and Armen G. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.
- [13] Arnak S. Dalalyan and Laurent Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.
- [14] Pierre Del Moral. *Mean Field Simulation for Monte Carlo Integration*, volume 126 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 2013.
- [15] Andreas Eberle, Arnaud Guillin, and Raphaël Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019.
- [16] Matthias Erbar and Jan Maas. Ricci curvature of finite markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, 2012.
- [17] Murat A. Erdogdu, Rasa Hosseinzadeh, and Matthew S. Zhang. On the convergence of langevin monte carlo. *arXiv preprint*, 2020.
- [18] Murat A. Erdogdu, Rasa Hosseinzadeh, and Matthew S. Zhang. Convergence of langevin monte carlo in chi-squared and rényi divergence. *arXiv preprint*, 2021.
- [19] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62(2):251–341, 1990.
- [20] Bernard Helffer and Francis Nier. *Hypoelliptic Estimates and Spectral Theory for Fokker–Planck Operators and Witten Laplacians*, volume 1862 of *Lecture Notes in Mathematics*. Springer, 2005.
- [21] Richard Holley and Daniel Stroock. Logarithmic sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5–6):1159–1194, 1987.
- [22] Robert Jarrow, Philip Protter, and Kazuhiro Shimbo. No arbitrage without semimartingales. *The Annals of Applied Probability*, 19(2):596–616, 2009.
- [23] Michel Ledoux. The geometry of markov diffusion generators. *Annales de la Faculté des Sciences de Toulouse*, 9(2):305–366, 2000.
- [24] Holden Lee, Maxim Raginsky, and Santosh Vempala. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [25] Jaehoon Lee, Seok-Kyun Seo, et al. Mixing of metastable diffusion processes with Gibbs invariant distribution. *Electronic Journal of Probability*, 30:1–52, 2025.
- [26] Jaehoon Lee, Seok-Kyun Seo, et al. Mixing of metastable diffusion processes with Gibbs invariant distribution. *Electronic Journal of Probability*, 30:1–52, 2025.
- [27] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 2016.
- [28] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 2018.
- [29] Laurent Miclo. Sur l’inégalité de Poincaré pour les diffusions. *Annales de l’IHP Probabilités et statistiques*, 32(1):1–36, 1995.
- [30] Andrea Mondino et al. Curvature-dimension bounds for unbalanced transport. *Journal of Functional Analysis*, 2023.
- [31] David Nualart. Stochastic calculus with respect to fractional Brownian motion. *Annales de la Faculté des Sciences de Toulouse*, 15(1):63–114, 2006.
- [32] Philip E. Protter. *Stochastic Integration and Differential Equations*. Springer, 2 edition, 2005.
- [33] Ken-Iti Sato. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 1999.
- [34] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

- [35] Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alán Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-kac correctors in diffusion: Annealing, guidance, and product of experts, 2025.
- [36] Xu Sun, Jinqiao Duan, et al. Derivation of Fokker–Planck equations for stochastic dynamical systems driven by non-Gaussian Lévy processes. *Journal of Mathematical Analysis and Applications*, 2017.
- [37] Xu Sun, Xiaofan Li, and Yun Zheng. Fokker–planck equations for Marcus stochastic differential equations driven by Lévy processes. *arXiv preprint*, 2016.
- [38] Wenpin Tang and Hanyang Zhao. Score-based diffusion models via stochastic differential equations. *Statistics Surveys*, 19:28–64, 2025.
- [39] Andrea Testa, Søren Hauberg, Tamim Asfour, and Leonel Rozo. Contact wasserstein geodesics for non-conservative schrödinger bridges, 2025.
- [40] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.
- [41] Y. Xu, J. Duan, et al. Fokker–planck equations for nonlinear dynamical systems driven by non-Gaussian Lévy processes. *Journal of Mathematical Physics*, 53(7):072701, 2012.
- [42] Eunbi Yoon, Sungwoo Cho, Seungjae Lee, et al. Score-based generative models with Lévy processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

## A. MORE ON WASSERSTEIN GEOMETRY

Once  $(T_\mu \mathcal{P}_2, \langle \cdot, \cdot \rangle_\mu)$  is identified with the closure of gradients, one may introduce a formal Levi–Civita connection  $\nabla^{W_2}$  on vector fields along curves. For smooth potentials  $\phi_t, \psi_t$ , with corresponding velocity fields  $v_t = \nabla \phi_t, w_t = \nabla \psi_t$ , the covariant derivative along  $\mu_t$  is

$$\frac{D}{dt} w_t := \Pi_{\mu_t} (\partial_t w_t + (v_t \cdot \nabla) w_t),$$

where  $\Pi_{\mu_t}$  denotes projection onto  $T_{\mu_t} \mathcal{P}_2$ , i.e. projection onto gradient vector fields in  $L^2(\mu_t)$ . Explicitly, any vector field can be Helmholtz-decomposed as

$$Z = \nabla \zeta + Z^\perp, \quad Z^\perp \text{ } \mu\text{-divergence-free},$$

and  $\Pi_\mu Z = \nabla \zeta$ . Thus the covariant derivative simplifies to

$$\frac{D}{dt} w_t = \nabla \pi_t, \quad \text{where } \Delta_{\mu_t} \pi_t = \nabla \cdot (\mu_t (\partial_t w_t + (v_t \cdot \nabla) w_t)),$$

with  $\Delta_\mu := \nabla \cdot (\mu \nabla)$ .

This connection is metric-compatible and torsion-free (formally), hence the unique Levi–Civita connection of the  $W_2$  Riemannian structure.

A curve  $(\mu_t, v_t = \nabla \phi_t)$  is a geodesic iff its velocity field is parallel transported:

$$\frac{D}{dt} v_t = 0.$$

Using the covariant derivative above, this yields

$$\partial_t \nabla \phi_t + (\nabla \phi_t \cdot \nabla) \nabla \phi_t = \nabla \pi_t$$

for some scalar pressure  $\pi_t$  that enforces the gradient constraint. Taking divergence w.r.t.  $\mu_t$  shows that  $\pi_t$  ensures  $\nabla \phi_t$  remains in the tangent space. In Lagrangian coordinates where  $X_t$  satisfies  $\dot{X}_t = \nabla \phi_t(X_t)$ , the geodesic equation becomes the classical Hamilton–Jacobi equation

$$\partial_t \phi_t + \frac{1}{2} |\nabla \phi_t|^2 = c(t),$$

with  $c(t)$  a time-dependent scalar.

For a functional  $F$ , the Wasserstein Hessian acts on  $v = \nabla \phi$  as

$$\text{Hess}_{W_2} F(\mu)[v] = \nabla \left( \partial_v \frac{\delta F}{\delta \mu}(\mu) \right) - \nabla v : \nabla^2 \frac{\delta F}{\delta \mu}(\mu),$$

where  $\partial_v$  denotes directional derivative and  $A : B := \text{Tr}(A^T B)$ . This enters the second variation inequality governing geodesic convexity.

The curvature tensor  $R(u, v)w$  is defined via

$$R(u, v)w := \frac{D}{ds} \frac{D}{dt} w - \frac{D}{dt} \frac{D}{ds} w$$

for smooth two-parameter families. In Otto’s geometry this yields a non-trivial curvature operator. For instance, along directions  $\nabla \phi, \nabla \psi$ , one obtains the sectional curvature

$$K(\phi, \psi) = \frac{1}{\|\nabla \phi\|_\mu^2 \|\nabla \psi\|_\mu^2 - \langle \nabla \phi, \nabla \psi \rangle_\mu^2} \int_{\Omega} \mu \text{Tr}[(\nabla^2 \phi)(\nabla^2 \psi) - (\nabla^2 \psi)(\nabla^2 \phi)] dx, \quad (49)$$

which is in general indefinite.

An alternative formulation uses the  $\Gamma-\Gamma_2$  calculus:

$$\Gamma(f) = |\nabla f|^2, \quad \Gamma_2(f) := \frac{1}{2} (L\Gamma(f) - 2\Gamma(f, Lf)),$$

where  $L$  is the generator of the gradient flow of a functional  $F$ . A curvature-dimension condition  $\Gamma_2(f) \geq \lambda\Gamma(f)$  corresponds exactly to  $\lambda$ -geodesic convexity of  $F$  in Wasserstein space.

The same geometry arises from several equivalent constructions:

- *Metric approach:* length structure induced by  $W_2$  and Benamou–Brenier.
- *Variational approach:* dynamic optimal transport as a kinetic action minimization.
- *Lagrangian approach:* geodesics obtained by pushing forward  $\mu_0$  along McCann’s displacement interpolation  $T_t = (1-t)\text{id} + tT$ .
- *Eulerian approach:* tangent bundle defined by continuity equations and minimal-energy velocity fields.
- *Connection-based approach:* Levi–Civita connection defined from Helmholtz projection; geodesics satisfy parallel transport.

All these constructions are mathematically equivalent and jointly define the full Riemannian structure of  $(\mathcal{P}_2(\Omega), W_2)$ .

To argue the latter geoemtry is more general than the wasserstein one, you can look intot eh following set of calculations that show, gradient flow in the latter can express former gradient flows:

## B. MORE ON INFORMATION GEOMETRY

Let  $(X, \mathcal{F}, \lambda)$  be a measurable space with reference measure  $\lambda$ , and let

$$M = \{p_\theta : \theta \in \Theta\} \subset \mathcal{P}(X)$$

be a smooth  $d$ -dimensional statistical model, where  $\Theta \subset \mathbb{R}^d$  is open and

$$p_\theta(x) = p(x; \theta)$$

is a smooth family of probability densities with respect to  $\lambda$ . The *score* is

$$\partial_i \log p_\theta(x) := \frac{\partial}{\partial \theta^i} \log p(x; \theta), \quad i = 1, \dots, d.$$

The *Fisher information metric* on  $M$  is defined by

$$g_{ij}(\theta) := \int_X \partial_i \log p_\theta(x) \partial_j \log p_\theta(x) p_\theta(x) d\lambda(x),$$

turning  $(M, g)$  into a finite-dimensional Riemannian manifold. When the FR metric on  $\mathcal{P}(X)$  is defined as above, its pullback along the embedding

$$\Theta \ni \theta \mapsto p_\theta \in \mathcal{P}(X)$$

coincides with the Fisher information matrix  $g_{ij}(\theta)$ ; in other words, the infinite-dimensional Fisher–Rao metric is compatible with the classical information-geometric construction on parametric families.

Historically, this Riemannian structure was introduced independently by Rao and others and later characterized by Chentsov as the

*unique* (up to a constant factor) Riemannian metric on the simplex that is invariant under Markov morphisms and can be obtained as the quadratic term in the Taylor expansion of any standard  $f$ -divergence between nearby models.

A distinctive feature of information geometry is that the Riemanian manifold  $(M, g)$  is enhanced by a pair of torsion-free affine connections  $(\nabla, \nabla^*)$  that are *dual* with respect to  $g$ , or more generally by a one-parameter family  $\{\nabla^{(\alpha)}\}_{\alpha \in \mathbb{R}}$  of  $\alpha$ -connections which are mutually dual.

Let  $\nabla$  and  $\nabla^*$  be torsion-free affine connections on  $M$  with Christoffel symbols  $\Gamma_{ij}^k$  and  $(\Gamma^*)_{ij}^k$  in local coordinates  $\theta$ . They are said to be dual with respect to  $g$  if for all vector fields  $X, Y, Z$  on  $M$ ,

$$X(g(Y, Z)) = g(\nabla_X Y, Z) + g(Y, \nabla_X^* Z). \quad (50)$$

Equivalently, in coordinates this reads

$$\partial_i g_{jk} = \Gamma_{ij}^\ell g_{\ell k} + (\Gamma^*)_{ik}^\ell g_{j\ell}. \quad (51)$$

The Levi–Civita connection  $\nabla^{(0)}$  of  $(M, g)$  is the unique torsion-free *metric* connection satisfying

$$\partial_i g_{jk} = \Gamma_{ij}^{(0)\ell} g_{\ell k} + \Gamma_{ik}^{(0)\ell} g_{j\ell}, \quad (52)$$

so it corresponds to the self-dual case  $\nabla = \nabla^* = \nabla^{(0)}$ .

In information geometry one introduces the totally symmetric *Amari–Chentsov tensor*

$$T_{ijk}(\theta) := \int_X \partial_i \log p_\theta(x) \partial_j \log p_\theta(x) \partial_k \log p_\theta(x) p_\theta(x) d\lambda(x),$$

and its  $(1, 2)$ -version

$$T_{ij}^k := g^{k\ell} T_{ij\ell}.$$

The  $\alpha$ -connections are then defined by

$$\Gamma_{ij}^{(\alpha)k} = \Gamma_{ij}^{(0)k} + \frac{\alpha}{2} T_{ij}^k, \quad (53)$$

and one checks that  $\nabla^{(\alpha)}$  and  $\nabla^{(-\alpha)}$  are dual with respect to  $g$  in the sense of (50). The three most important cases are (mixture connection), (Levi–Civita / Fisher–Rao connection) (exponential connection).

**Definition 9** (Statistical manifold). A *statistical manifold* in the sense of information geometry is a quadruple

$$(M, g, \nabla, \nabla^*)$$

consisting of a smooth manifold  $M$ , a Riemannian metric  $g$ , and a pair of torsion-free affine connections  $(\nabla, \nabla^*)$  that are dual with respect to  $g$  as in (50). The special case  $(M, g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$  is called an  $\alpha$ -statistical manifold.

Geometrically, the dualistic structure  $(g, \nabla, \nabla^*)$  allows one to decompose curvature and convexity into “primal” and “dual” contributions, and it underlies many higher-order asymptotic results in statistics (efficiency, bias, and so on).

Given a connection  $\nabla^{(\alpha)}$ , a curve  $\theta(t)$  in parameter space is a  $\nabla^{(\alpha)}$ -geodesic if its velocity vector field is parallel along the curve:

$$\nabla_{\dot{\theta}(t)}^{(\alpha)} \dot{\theta}(t) = 0,$$

or, in local coordinates,

$$\ddot{\theta}^k(t) + \Gamma_{ij}^{(\alpha)k}(\theta(t)) \dot{\theta}^i(t) \dot{\theta}^j(t) = 0. \quad (54)$$

In many important models (in particular exponential families), the  $\alpha = \pm 1, 0$  connections have especially simple geodesics when expressed in suitable affine coordinates.

*Mixture geodesics* ( $\alpha = -1$ ). In mixture coordinates, a  $(-1)$ -geodesic between densities  $\rho_0, \rho_1$  is simply the linear interpolation

$$\rho_t^{\text{mix}} = (1-t)\rho_0 + t\rho_1, \quad t \in [0, 1].$$

On a finite-dimensional exponential family, mixture geodesics correspond to affine lines in the expectation-parameter coordinates. In this sense, the mixture connection encodes the affine structure of the simplex under convex combination.

*Exponential geodesics* ( $\alpha = +1$ ). In natural (canonical) parameters, a  $(+1)$ -geodesic is linear in the log-density. At the level of densities this gives the normalized geometric interpolation

$$\rho_t^{\text{exp}} = \frac{\rho_0^{1-t} \rho_1^t}{\int_X \rho_0^{1-t}(x) \rho_1^t(x) d\lambda(x)}, \quad t \in [0, 1].$$

On an exponential family, exponential geodesics are affine lines in the natural parameter  $\theta$ , and their image in density space is obtained by exponentiating and normalizing.

*Fisher–Rao / Levi–Civita geodesics* ( $\alpha = 0$ ). For the Fisher–Rao connection, geodesics are the usual Riemannian geodesics of  $(M, g)$ . In the infinite-dimensional ambient space  $(\mathcal{M}_+(\Omega), d_{\text{FR}})$ , the map  $\mu \mapsto \sqrt{\mu}$  identifies the manifold with a convex cone in the Hilbert space  $L^2(\Omega)$ . The FR geodesic between  $\mu_0$  and  $\mu_1$  is simply the image of the straight line between  $\sqrt{\mu_0}$  and  $\sqrt{\mu_1}$  in  $L^2(\Omega)$ , restricted to the cone:

$$\sqrt{\mu_t} = (1-t)\sqrt{\mu_0} + t\sqrt{\mu_1}, \quad \mu_t = ((1-t)\sqrt{\mu_0} + t\sqrt{\mu_1})^2.$$

On a normalized probability simplex, this corresponds to great-circle arcs on the unit sphere of  $L^2$ , and one can compute curvature tensors explicitly for specific parametric families (e.g. location-scale Gaussians) to see that many statistical manifolds have negative sectional curvature even though the ambient FR cone is flat; this is the source of the “negative curvature of statistical manifolds” versus “flat Fisher–Rao cone” contrast.

The Fisher–Rao construction above is inherently infinite-dimensional:  $\mathcal{P}(\Omega)$  (or  $\mathcal{M}_+(\Omega)$ ) is modelled on suitable function spaces (e.g.  $L^2$  or Orlicz spaces), and the map  $\mu \mapsto \sqrt{\mu}$  realizes it as a (nonlinear) submanifold of  $L^2(\Omega)$ . In this setting, tangent vectors are identified with square-integrable functions  $\varphi$  with zero mean, and the FR inner product is simply

$$\langle \varphi_1, \varphi_2 \rangle_{\mu}^{\text{FR}} = \int_{\Omega} \varphi_1(x) \varphi_2(x) d\mu(x).$$

Restricting this infinite-dimensional structure to a finite-dimensional statistical model  $M = \{p_\theta\}$  recovers the Fisher information metric, and the dualistic structure  $(g, \nabla^{(\alpha)}, \nabla^{(-\alpha)})$  lifts to suitable infinite-dimensional settings.

One rigorous approach, due to Pistone and Sempi, constructs an infinite-dimensional exponential statistical manifold on the space of all probability measures equivalent to a fixed reference measure, with charts given by centered log-densities and with a natural extension of the FR metric and  $\alpha$ -connections. This provides a genuine Banach (or Fréchet) manifold structure on  $\mathcal{P}(\Omega)$  in which mixture and exponential geodesics, as well as FR geodesics, can be treated

on the same footing, and where the dualistic information-geometric calculus extends beyond parametric models.

In summary, information geometry equips spaces of probability measures with a Riemannian metric (Fisher–Rao), a family of dual affine connections ( $\alpha$ -connections), and the associated mixture, exponential, and FR geodesics. This dualistic structure is compatible with both finite-dimensional parametric models and infinite-dimensional manifolds of measures, and it will be crucial later when we relate curvature, spectral properties of Markov semigroups, and stability of PDE/SDE flows.

### C. MORE ON HK GEOMETRY

From the Riemannian point of view,  $(\mathcal{M}_+(\Omega), d_{\text{WFR}})$  is a formal infinite-dimensional Riemannian manifold whose tangent space at  $\mu$  is identified with potentials  $\phi$  modulo constants, endowed with the inner product

$$\langle \phi_1, \phi_2 \rangle_{\mu}^{\text{WFR}} = \int_{\Omega} (\phi_1 \phi_2 + \nabla \phi_1 \cdot \nabla \phi_2) d\mu.$$

The associated Levi–Civita connection is obtained by projecting the time derivative of a time-dependent potential plus its convective derivative (arising from  $v_t = \nabla \phi_t$ ) back onto this tangent space, exactly as in the Otto calculus but with an additional reaction component. Geodesics in this geometry are characterized variationally as minimizers of the action  $\mathcal{A}$ , and infinitesimally by the condition that their velocity potentials are parallel transported by this connection.

Geometrically, the WFR (Hellinger–Kantorovich) space can also be represented as a cone over a suitable Wasserstein base space: informally, one can rewrite a positive measure as a “mass” (radial variable) times a probability distribution (angular variable). The radial part evolves according to a Fisher–Rao-type dynamics, while the angular part follows a Wasserstein-type displacement. In this sense, purely radial geodesics reproduce Fisher–Rao geodesics, purely angular geodesics reproduce Wasserstein geodesics, and general WFR geodesics involve simultaneous changes in both mass and spatial distribution.

## D. MARKOV SEMIGROUPS, SPECTRAL GAPS, AND CURVATURE VIA CARRÉ DU CHAMP

### D.1. Markov semigroups and infinitesimal generators

Let  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$  be a filtered probability space and  $(X_t)_{t \geq 0}$  a time-homogeneous Markov process on  $\mathbb{R}^d$ . Its Markov semigroup  $(P_t)_{t \geq 0}$  acts on suitable test functions  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$  by

$$(P_t \varphi)(x) := \mathbb{E}[\varphi(X_t) | X_0 = x]. \quad (55)$$

The semigroup property  $P_{t+s} = P_t P_s$  follows from the Markov property.

The (infinitesimal) generator  $\mathcal{L}$  is defined on its domain  $\mathcal{D}(\mathcal{L})$  by the strong limit

$$\mathcal{L}\varphi := \lim_{t \downarrow 0} \frac{P_t \varphi - \varphi}{t}, \quad \varphi \in \mathcal{D}(\mathcal{L}). \quad (56)$$

Formally one may write  $P_t = \exp(t\mathcal{L})$  and the backward Kolmogorov equation becomes

$$\partial_t(P_t \varphi) = \mathcal{L}(P_t \varphi) = P_t(\mathcal{L}\varphi), \quad P_0 \varphi = \varphi. \quad (57)$$

When the law of  $X_t$  has a density  $p_t$  w.r.t. Lebesgue measure, the forward evolution is governed by the adjoint  $\mathcal{L}^*$  (in the  $\langle \cdot, \cdot \rangle_{L^2(dx)}$  pairing):

$$\partial_t p_t = \mathcal{L}^* p_t, \quad \int \varphi(x) \mathcal{L}^* p(x) dx = \int (\mathcal{L}\varphi)(x) p(x) dx. \quad (58)$$

## D.2. The OU/Langevin diffusion as the canonical example

A guiding example throughout this paper is the Langevin (Ornstein–Uhlenbeck as the quadratic case) diffusion

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t, \quad (59)$$

for a potential  $V : \mathbb{R}^d \rightarrow \mathbb{R}$ . Its generator acting on smooth compactly supported  $\varphi$  is

$$\mathcal{L}\varphi = \Delta\varphi - \langle \nabla V, \nabla\varphi \rangle, \quad (60)$$

and the invariant probability measure (when normalizable) is

$$\pi(dx) \propto e^{-V(x)} dx. \quad (61)$$

For the OU process,  $V(x) = \frac{1}{2}\|x\|^2$  and  $\pi$  is Gaussian.

It is often advantageous to switch from Lebesgue to the invariant measure and work in  $L^2(\pi)$ . In particular, rather than evolving the Lebesgue-density  $p_t$ , one may evolve the relative density  $\rho_t := p_t/\pi$ , in which case (under reversibility assumptions made below) the forward equation takes the simple form

$$\partial_t \rho_t = \mathcal{L}\rho_t. \quad (62)$$

## D.3. Reversibility, symmetry, and Dirichlet forms

A probability measure  $\pi$  is stationary for  $(P_t)$  if  $\int P_t \varphi d\pi = \int \varphi d\pi$  for all  $t \geq 0$ . The semigroup is *reversible* w.r.t.  $\pi$  if

$$\int f(P_t g) d\pi = \int (P_t f) g d\pi, \quad \forall f, g \in L^2(\pi), \forall t \geq 0. \quad (63)$$

Equivalently,  $P_t$  (and hence  $\mathcal{L}$ ) is self-adjoint in  $L^2(\pi)$ , and one expects a real spectrum. A basic consequence is the  $L^2(\pi)$  contraction inequality

$$\|P_t f\|_{L^2(\pi)} \leq \|f\|_{L^2(\pi)}, \quad t \geq 0, \quad (64)$$

which can be obtained from Jensen and stationarity.

Given a (symmetric) generator  $\mathcal{L}$ , the *carré du champ* operator  $\Gamma$  is defined by

$$\Gamma(f, g) := \frac{1}{2} \left( \mathcal{L}(fg) - f \mathcal{L}g - g \mathcal{L}f \right), \quad \Gamma(f) := \Gamma(f, f). \quad (65)$$

The associated Dirichlet energy is

$$\mathcal{E}(f, g) := \int \Gamma(f, g) d\pi. \quad (66)$$

Under reversibility, one has the fundamental integration-by-parts identity

$$\int f(-\mathcal{L})g d\pi = \int \Gamma(f, g) d\pi = \mathcal{E}(f, g), \quad (67)$$

so that  $-\mathcal{L}$  is positive semidefinite on  $L^2(\pi)$ .

For the Langevin generator (60), a direct computation yields the canonical identity

$$\Gamma(f) = \|\nabla f\|^2. \quad (68)$$

## D.4. Spectral gap, Poincaré inequality, and mixing rates

The spectral gap of  $-\mathcal{L}$  (in  $L^2(\pi)$ ) governs exponential decay to equilibrium in  $L^2$ . A standard sufficient condition is the Poincaré inequality: there exists  $\lambda_{\text{PI}} > 0$  such that

$$\text{Var}_\pi(f) := \int \left( f - \int f d\pi \right)^2 d\pi \leq \frac{1}{\lambda_{\text{PI}}} \int \Gamma(f) d\pi, \quad \forall f \in \mathcal{D}(\mathcal{E}). \quad (69)$$

When (69) holds, one obtains exponential decay of variance along the semigroup:

$$\text{Var}_\pi(P_t f) \leq e^{-2\lambda_{\text{PI}}t} \text{Var}_\pi(f), \quad t \geq 0, \quad (70)$$

which is a precise, operator-theoretic notion of *mixing rate*.

A stronger inequality is the logarithmic Sobolev inequality (LSI): there exists  $C_{\text{LSI}} > 0$  such that for all densities  $\mu$  w.r.t.  $\pi$ ,

$$\text{KL}(\mu\|\pi) \leq C_{\text{LSI}} \text{FI}(\mu\|\pi), \quad \text{FI}(\mu\|\pi) := \mathbb{E}_\mu [\|\nabla \log(\mu/\pi)\|^2]. \quad (71)$$

LSI implies exponential decay of relative entropy (and hence, under mild conditions, convergence in total variation).

## D.5. Iterated carré du champ and curvature-dimension

The *iterated carré du champ*  $\Gamma_2$  is defined by

$$\Gamma_2(f, g) := \frac{1}{2} \left( \mathcal{L}\Gamma(f, g) - \Gamma(f, \mathcal{L}g) - \Gamma(g, \mathcal{L}f) \right), \quad \Gamma_2(f) := \Gamma_2(f, f). \quad (72)$$

Conceptually,  $\Gamma$  appears when differentiating Lyapunov functionals (e.g.  $\chi^2$  or KL) once in time, while  $\Gamma_2$  appears upon differentiating a second time, and it is the correct analytic object that encodes *curvature* information of the diffusion (in the Bakry–Émery sense).

A diffusion Markov semigroup is said to satisfy the Bakry–Émery criterion with constant  $\alpha > 0$  if

$$\Gamma_2(f) \geq \alpha \Gamma(f), \quad \forall f. \quad (73)$$

This is also written as the curvature-dimension condition  $\text{CD}(\alpha, \infty)$ . In the Langevin case,  $\text{CD}(\alpha, \infty)$  is equivalent to  $\alpha$ -strong convexity of  $V$ :

$$\nabla^2 V(x) \succeq \alpha I \iff \text{CD}(\alpha, \infty). \quad (74)$$

The significance for sampling is that  $\text{CD}(\alpha, \infty)$  yields functional inequalities (in particular, LSI) with constants controlled by  $1/\alpha$ , hence quantitative convergence rates.

## D.6. How Fisher–Rao reweighting can affect the spectral gap: a precise operator viewpoint

In this subsection we make precise, at the level of generators and quadratic forms, the sense in which adding a Fisher–Rao (FR) *reaction/reweighting* term on top of an Ornstein–Uhlenbeck (OU) diffusion can change the spectral properties that govern convergence rates. We emphasize from the outset that the resulting dynamics is no longer a linear Markov semigroup on densities; rather, it is a Feynman–Kac (FK) evolution with normalization, or equivalently a mean-field interacting particle system. Consequently, the classical notion of *spectral gap of a Markov generator* does not apply verbatim. Nevertheless, there is a natural and rigorous way to measure the additional dissipation induced by the FR term via bilinear forms and variance decay identities.

Let  $\pi$  denote the invariant Gaussian measure of the OU process  $dX_t = -\alpha X_t dt + \sqrt{2} dB_t$  (with  $\alpha > 0$ ), and let  $\mathcal{L}$  be its (symmetric) generator on  $L^2(\pi)$ ,

$$\mathcal{L}f = \Delta f - \alpha \langle x, \nabla f \rangle.$$

The OU semigroup  $(P_t)_{t \geq 0}$  is reversible w.r.t.  $\pi$ , and the spectral gap inequality (Poincaré inequality) asserts that there exists  $\lambda_{\text{OU}} > 0$  such that

$$\text{Var}_\pi(f) \leq \frac{1}{\lambda_{\text{OU}}} \mathcal{E}(f, f), \quad \mathcal{E}(f, f) = \int \Gamma(f) d\pi = \int \|\nabla f\|^2 d\pi, \quad (75)$$

for all  $f$  in the Dirichlet domain with  $\int f d\pi = 0$ . Equivalently,

$$\int f(-\mathcal{L})f d\pi \geq \lambda_{\text{OU}} \int f^2 d\pi, \quad f \perp 1, \quad (76)$$

which yields exponential  $L^2(\pi)$  convergence:  $\|P_t f\|_{L^2(\pi)} \leq e^{-\lambda_{\text{OU}} t} \|f\|_{L^2(\pi)}$ .

Let  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  be a measurable potential (in applications,  $g$  may depend on  $t$  and on the current density through a corrector; here we freeze  $g$  to isolate the spectral mechanism). Consider the *normalized* FK evolution of densities w.r.t. Lebesgue,

$$\partial_t p_t = \mathcal{L}^* p_t + p_t \left( g - \mathbb{E}_{p_t}[g] \right), \quad \mathbb{E}_{p_t}[g] = \int g(x) p_t(x) dx, \quad (77)$$

which is exactly the OU transport-diffusion plus an FR-type reweighting term. The centering by  $\mathbb{E}_{p_t}[g]$  enforces mass conservation  $\int p_t = 1$ . This dynamics is no longer linear in  $p_t$ , hence it does not define a Markov semigroup on densities.

A convenient way to compare with the OU spectral picture is to switch to the *relative density*  $\rho_t := \frac{p_t}{\pi}$ , for which the baseline OU evolution becomes  $\partial_t \rho_t = \mathcal{L} \rho_t$  (cf. the semigroup preliminaries). In these coordinates, (77) reads

$$\partial_t \rho_t = \mathcal{L} \rho_t + \rho_t \left( g - \mathbb{E}_{\rho_t \pi}[g] \right), \quad \mathbb{E}_{\rho_t \pi}[g] = \int g \rho_t d\pi. \quad (78)$$

A natural  $L^2(\pi)$  distance from equilibrium is the  $\chi^2$  functional  $\chi^2(p_t \|\pi) = \int (\rho_t - 1)^2 d\pi = \text{Var}_\pi(\rho_t)$  (since  $\int \rho_t d\pi = 1$ ). Differentiating along (78) and using the integration-by-parts identity  $\int f \mathcal{L} f d\pi = -\int \Gamma(f) d\pi$ , we obtain the exact identity

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \text{Var}_\pi(\rho_t) &= \int (\rho_t - 1) \partial_t \rho_t d\pi \\ &= \underbrace{\int (\rho_t - 1) \mathcal{L} \rho_t d\pi}_{-\int \Gamma(\rho_t) d\pi} + \int (\rho_t - 1) \rho_t \left( g - \mathbb{E}_{\rho_t \pi}[g] \right) d\pi \\ &= -\int \Gamma(\rho_t) d\pi + \int \rho_t (\rho_t - 1) g d\pi - \mathbb{E}_{\rho_t \pi}[g] \int \rho_t (\rho_t - 1) d\pi. \end{aligned}$$

Since  $\int \rho_t (\rho_t - 1) d\pi = \text{Var}_\pi(\rho_t)$ , the last term is explicit:

$$\frac{1}{2} \frac{d}{dt} \text{Var}_\pi(\rho_t) = -\int \Gamma(\rho_t) d\pi + \int \rho_t (\rho_t - 1) g d\pi - \mathbb{E}_{\rho_t \pi}[g] \text{Var}_\pi(\rho_t) \quad (79)$$

Equation (79) is rigorous and shows precisely how the reweighting term alters the baseline OU dissipation  $-\int \Gamma(\rho_t) d\pi$  by two additional contributions, both governed by correlations between  $\rho_t$  and  $g$ .

To connect (79) to spectral gaps, we linearize the flow around equilibrium. Let  $\rho_t = 1 + \varepsilon h_t$  with  $\int h_t d\pi = 0$  and  $\varepsilon \ll 1$ . Expanding (78) to first order in  $\varepsilon$  yields the linearized evolution

$$\partial_t h_t = \mathcal{L} h_t + \left( g - \mathbb{E}_\pi[g] \right) h_t, \quad (80)$$

since the centering term contributes only through  $\mathbb{E}_\pi[g]$  at leading order. Define the (generally non-symmetric) linear operator

$$\mathcal{A} := \mathcal{L} + (g - \mathbb{E}_\pi[g]) \text{Id}. \quad (81)$$

In the reversible OU setting,  $\mathcal{L}$  is self-adjoint on  $L^2(\pi)$ , and multiplication by  $(g - \mathbb{E}_\pi[g])$  is also self-adjoint. Thus  $\mathcal{A}$  is self-adjoint on  $L^2(\pi)$  and the linearized dynamics admits an  $L^2(\pi)$  spectral decomposition. In particular, if

$$-\sup \{ \langle h, \mathcal{A}h \rangle_{L^2(\pi)} : \|h\|_{L^2(\pi)} = 1, h \perp 1 \} = \lambda_{\text{eff}} > 0, \quad (82)$$

then the linearized perturbations decay exponentially:  $\|h_t\|_{L^2(\pi)} \leq e^{-\lambda_{\text{eff}} t} \|h_0\|_{L^2(\pi)}$ . Comparing (82) with the OU gap (76), we see that the FR potential effectively modifies the Rayleigh quotient by adding  $\int (g - \mathbb{E}_\pi[g]) h^2 d\pi$ . Hence, whenever the potential is such that

$$\int (g - \mathbb{E}_\pi[g]) h^2 d\pi \leq -c \|h\|_{L^2(\pi)}^2 \quad \text{for all } h \perp 1, \quad (83)$$

for some  $c > 0$ , the effective gap improves to  $\lambda_{\text{eff}} \geq \lambda_{\text{OU}} + c$ . Condition (83) holds, for instance, if  $g - \mathbb{E}_\pi[g] \leq -c \pi\text{-a.s.}$  (a strong sufficient condition), or more generally if  $g$  is negative on the dominant modes of  $-\mathcal{L}$ .

The preceding argument makes two points precise:

1. The FR reweighting term changes the *dissipation identity* (79) by adding terms controlled by correlations with the potential  $g$ .
2. After linearization around equilibrium, the FR term appears as a self-adjoint “killing/anti-killing” perturbation of the OU generator (81), and can therefore increase or decrease the effective spectral gap depending on the sign/structure of  $g$  through the Rayleigh quotient (82).

In particular, the statement “FR improves the spectral gap” is *not* automatic: it requires that the induced potential acts as additional dissipation on the relevant slow modes.

In the applications of this paper,  $\psi_t$  is not an arbitrary fixed potential but is induced by geometric correction terms (e.g. mixture/exponential/OT interpolations) and may depend on the evolving density. This places the resulting dynamics outside the scope of classical linear Markov semigroup theory: the flow is nonlinear, and its implementation relies on interacting particles with resampling/jump mechanisms. A systematic theory connecting such normalized Feynman–Kac evolutions to *quantitative* improvements of mixing rates via spectral gaps (or suitable nonlinear analogues) appears to be underdeveloped. Establishing sharp conditions under which the induced FR/FK potentials enlarge an effective gap and yield provably faster convergence is an important direction for future research.

## E. THREE ELEMENTARY LEMMAS: DRIFT, DIFFUSION, AND FISHER–RAO RATES

In this section we isolate three simple but important identities that clarify how the classical drift and diffusion terms appearing in the Fokker–Planck equation can be re-expressed in terms of (i) a pure continuity equation with a suitable velocity field, and (ii) Fisher–Rao reaction equations of the form

$$\partial_t \mu_t = \psi_t \mu_t,$$

for an explicitly computable rate function  $\psi_t$ . These identities will be the basic algebraic tools that allow us, later on, to express diffusion-model dynamics in Fisher–Rao language and to identify the correct Feynman–Kac weights.

Throughout we work on  $\mathbb{R}^d$  and assume that for each  $t$ ,  $\mu_t$  is a strictly positive, smooth density with sufficient decay at infinity so that all integrations by parts below are justified. We write  $\mu_t(x)$  simply as  $\mu_t$  when no confusion can arise.

### E.1. Diffusion as drift

We first show that pure diffusion can be rewritten exactly as a continuity equation with a suitable (state- and time-dependent) velocity field. This gives a first indication that diffusion is not fundamentally different from drift at the level of the PDE; rather, it is a very specific choice of drift depending on the current density.

**Lemma 1** (Diffusion can be written as drift). *Let  $(\mu_t)_{t \geq 0}$  solve the heat equation*

$$\partial_t \mu_t = \frac{\sigma_t^2}{2} \Delta \mu_t, \quad t \geq 0. \quad (84)$$

Define the vector field

$$v_t(x) := -\frac{\sigma_t^2}{2} \nabla \log \mu_t(x), \quad x \in \mathbb{R}^d. \quad (85)$$

Then  $(\mu_t, v_t)$  satisfy the continuity equation

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0, \quad (86)$$

and conversely, any strictly positive solution of (86) with  $v_t$  given by (85) also solves the heat equation (84).

*Proof.* Using the definition (85), we compute

$$\mu_t(x) v_t(x) = -\frac{\sigma_t^2}{2} \mu_t(x) \nabla \log \mu_t(x) = -\frac{\sigma_t^2}{2} \nabla \mu_t(x),$$

since  $\nabla \log \mu_t = \nabla \mu_t / \mu_t$ . Therefore

$$\nabla \cdot (\mu_t v_t) = -\frac{\sigma_t^2}{2} \nabla \cdot (\nabla \mu_t) = -\frac{\sigma_t^2}{2} \Delta \mu_t.$$

Substituting into the continuity equation  $\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0$  yields

$$\partial_t \mu_t = -\nabla \cdot (\mu_t v_t) = \frac{\sigma_t^2}{2} \Delta \mu_t,$$

which is exactly the heat equation (84). Conversely, if  $\mu_t$  solves (84), setting  $v_t$  as in (85) gives the desired continuity equation (86). This proves the equivalence.  $\square$

**Remark 10.** Lemma 1 shows that diffusion can be seen as a very particular drift, namely the *score-driven* drift  $v_t = -(\sigma_t^2/2) \nabla \log \mu_t$ . In Wasserstein geometry, this is the velocity field corresponding to the  $W_2$ -gradient of the negative entropy functional.

### E.1.1. Drift as Fisher–Rao reaction

We next show that the usual drift term, expressed via the continuity equation, admits an equivalent Fisher–Rao representation

$$\partial_t \mu_t = \psi_t \mu_t,$$

with an explicit rate  $\psi_t$  depending on the drift and on the log-density. This is the precise sense in which a transport equation can be understood through the Fisher–Rao lens.

**Lemma 2** (Drift induces a Fisher–Rao rate). *Let  $(\mu_t)_{t \geq 0}$  solve the continuity equation*

$$\partial_t \mu_t = -\nabla \cdot (\mu_t v_t), \quad t \geq 0, \quad (87)$$

for a given smooth drift field  $v_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ . Assume  $\mu_t > 0$  everywhere. Then

$$\partial_t \mu_t(x) = \psi_t(x) \mu_t(x), \quad \psi_t(x) := -\nabla \cdot v_t(x) - v_t(x) \cdot \nabla \log \mu_t(x). \quad (88)$$

Equivalently, the Fisher–Rao tangent vector associated with the drift is

$$\psi_t = \frac{\partial_t \mu_t}{\mu_t} = -\nabla \cdot v_t - v_t \cdot \nabla \log \mu_t.$$

*Proof.* Expanding the divergence in (87) gives

$$\partial_t \mu_t = -\nabla \cdot (\mu_t v_t) = -v_t \cdot \nabla \mu_t - \mu_t \nabla \cdot v_t.$$

Since  $\mu_t > 0$ , we can divide by  $\mu_t$  and rewrite

$$\frac{\partial_t \mu_t}{\mu_t} = -\frac{v_t \cdot \nabla \mu_t}{\mu_t} - \nabla \cdot v_t = -v_t \cdot \nabla \log \mu_t - \nabla \cdot v_t.$$

Defining  $\psi_t$  by (88), we obtain  $\partial_t \mu_t = \psi_t \mu_t$  as claimed.  $\square$

**Remark 11.** From the Fisher–Rao perspective, the function  $\psi_t$  in (88) is the *instantaneous log-growth rate* of the density at point  $x$  induced by the drift  $v_t$ : it combines a local volume-change term  $-\nabla \cdot v_t$  with a term  $-v_t \cdot \nabla \log \mu_t$  describing the advective change of the log-density along the flow. Thus even a purely transport equation naturally induces a Fisher–Rao tangent vector  $\psi_t$ .

### E.1.2. Diffusion as Fisher–Rao reaction

Finally, we show that pure diffusion admits a Fisher–Rao representation with a specific rate that depends on the local curvature and gradient of the log-density. This makes it clear that, at the level of the Fisher–Rao geometry, diffusion is also simply a particular choice of  $\psi_t$ .

**Lemma 3** (Diffusion induces a Fisher–Rao rate). *Let  $(\mu_t)_{t \geq 0}$  solve the heat equation (84) with  $\mu_t > 0$ . Then*

$$\begin{aligned} \partial_t \mu_t(x) &= \psi_t(x) \mu_t(x), \\ \psi_t(x) &:= \frac{\sigma_t^2}{2} \left( \Delta \log \mu_t(x) + \|\nabla \log \mu_t(x)\|^2 \right). \end{aligned} \quad (89)$$

*Proof.* We use the identity

$$\Delta \mu_t = \nabla \cdot (\nabla \mu_t) = \nabla \cdot (\mu_t \nabla \log \mu_t).$$

Expanding the last divergence yields

$$\Delta \mu_t = \mu_t \Delta \log \mu_t + \nabla \mu_t \cdot \nabla \log \mu_t.$$

Using  $\nabla \mu_t = \mu_t \nabla \log \mu_t$ , we further obtain

$$\Delta \mu_t = \mu_t \Delta \log \mu_t + \mu_t \|\nabla \log \mu_t\|^2 = \mu_t \left( \Delta \log \mu_t + \|\nabla \log \mu_t\|^2 \right).$$

Substituting into the heat equation (84), we get

$$\partial_t \mu_t = \frac{\sigma_t^2}{2} \Delta \mu_t = \frac{\sigma_t^2}{2} \mu_t \left( \Delta \log \mu_t + \|\nabla \log \mu_t\|^2 \right) = \psi_t \mu_t,$$

with  $\psi_t$  as in (89). This is exactly the desired Fisher–Rao representation.  $\square$

**Remark 12.** Lemma 3 shows that diffusion corresponds to a Fisher–Rao tangent vector  $\psi_t$  that depends on the Laplacian and the squared norm of the score  $\nabla \log \mu_t$ . In particular, in regions where the density is locally log-concave and sharply peaked, the rate can be strongly positive or negative depending on the balance between  $\Delta \log \mu_t$  and  $\|\nabla \log \mu_t\|^2$ . From a geometric point of view, this identifies diffusion as a particular direction in the Fisher–Rao tangent space.

## E.2. Beyond Wasserstein and Fisher–Rao: “Gradient flows” for other geometries

Let  $\mathcal{P}$  be a smooth statistical manifold of strictly positive densities  $p$  on  $\mathbb{R}^d$  (or on a finite set), and let  $\mathcal{F} : \mathcal{P} \rightarrow \mathbb{R}$  be a sufficiently smooth functional. Its first variation is the (equivalence class of) functions  $\frac{\delta \mathcal{F}}{\delta p}$  characterized by

$$\frac{d}{d\epsilon} \Big|_{\epsilon=0} \mathcal{F}(p + \epsilon \dot{p}) = \int \frac{\delta \mathcal{F}}{\delta p}(x) \dot{p}(x) dx. \quad (90)$$

A *gradient flow* requires more structure than an affine connection: it requires a *metric* (or, more generally, an Onsager operator). Connections (mixture/exponential) determine *geodesics / straightness*, while the metric determines *steepest descent*.

Given a (weak) Riemannian metric  $g_p(\cdot, \cdot)$  on  $T_p \mathcal{P}$ , the gradient  $\text{grad}_g \mathcal{F}(p) \in T_p \mathcal{P}$  is defined by

$$g_p(\text{grad}_g \mathcal{F}(p), \xi) = d\mathcal{F}(p)[\xi] = \int \frac{\delta \mathcal{F}}{\delta p} \xi, \quad \forall \xi \in T_p \mathcal{P}. \quad (91)$$

The *gradient flow* is then

$$\partial_t p_t = -\text{grad}_g \mathcal{F}(p_t). \quad (92)$$

Equivalently, in the “Onsager” form, one specifies a positive operator  $K(p) : (T_p \mathcal{P})^* \rightarrow T_p \mathcal{P}$  and writes

$$\partial_t p_t = -K(p_t) \frac{\delta \mathcal{F}}{\delta p}(p_t). \quad (93)$$

Different geometries correspond to different choices of  $g$  or  $K$ .

### E.2.1. Wasserstein (Otto) gradient flow (baseline)

For the 2-Wasserstein metric, the Onsager operator is  $K_W(p)\phi = \nabla \cdot (p \nabla \phi)$ , and the gradient flow reads

$$\partial_t p_t = \nabla \cdot \left( p_t \nabla \frac{\delta \mathcal{F}}{\delta p}(p_t) \right). \quad (94)$$

### E.2.2. Fisher–Rao gradient flow (baseline)

For the Fisher–Rao metric on densities (mass-preserving version), one obtains

$$\partial_t p_t = -p_t \left( \frac{\delta \mathcal{F}}{\delta p}(p_t) - \mathbb{E}_{p_t} \left[ \frac{\delta \mathcal{F}}{\delta p}(p_t) \right] \right), \quad (95)$$

where the centering enforces  $\int p_t = 1$ .

### E.2.3. Mixture vs exponential “gradient flows” in information geometry

The *mixture* ( $m$ ) and *exponential* ( $e$ ) geometries are primarily *affine structures* (dual connections) on a manifold equipped with the Fisher metric. A connection alone does not define a gradient flow. However, on a *dually-flat* manifold (e.g. an exponential family), the Fisher metric is Hessian:

$$g = \nabla^2 \psi(\theta) \quad \text{in } e\text{-coordinates } \theta, \quad g = \nabla^2 \phi(\eta) \quad \text{in } m\text{-coordinates } \eta,$$

with  $\eta = \nabla \psi(\theta)$  and  $\phi$  the Legendre dual of  $\psi$ . Thus one gets a canonical notion of “steepest descent” that is most naturally written in coordinates as a *natural gradient / mirror flow*.

Suppose  $p = p_\theta$  is parametrized by  $e$ -affine coordinates  $\theta$  (natural parameters). Then the Fisher metric is  $G(\theta) = [\frac{\delta \mathcal{F}}{\delta p}(\theta)]$  and the Fisher steepest descent is

$$\dot{\theta}_t = -G(\theta_t)^{-1} \nabla_\theta \mathcal{F}(\theta_t). \quad (96)$$

Geodesics of the  $e$ -connection are straight lines in  $\theta$ ; hence (96) is the canonical “gradient flow compatible with exponential geodesics” (steepest descent measured by the Fisher metric but expressed in the  $e$ -affine chart).

If the same manifold is described in  $m$ -affine coordinates  $\eta$  (expectation parameters), then

$$\dot{\eta}_t = -G^*(\eta_t)^{-1} \nabla_\eta \mathcal{F}(\eta_t), \quad (97)$$

where  $G^*(\eta) = \nabla^2 \phi(\eta)$  is the Fisher metric in  $\eta$ -coordinates. Geodesics of the  $m$ -connection are straight lines in  $\eta$ .

A convenient coordinate-free expression uses a convex potential  $\psi$  generating a Bregman divergence (dually flat geometry). In primal coordinates  $x$  one may write

$$\begin{aligned} \frac{d}{dt} \nabla \psi(x_t) &= -\nabla \mathcal{F}(x_t), & \text{equivalently} \\ \dot{x}_t &= -(\nabla^2 \psi(x_t))^{-1} \nabla \mathcal{F}(x_t). \end{aligned} \quad (98)$$

On the probability simplex, taking  $\psi(p) = \sum_i p_i \log p_i$  yields the KL/Shahshahani geometry and gives the replicator-type gradient flow

$$\dot{p}_i = -p_i \left( \partial_{p_i} \mathcal{F}(p) - \sum_j p_j \partial_{p_j} \mathcal{F}(p) \right),$$

which is precisely the Fisher natural gradient on the simplex.

**Takeaway.** Mixture/exponential “gradient flows” are best understood as *Fisher-metric natural gradients written in the  $m$ - or  $e$ -affine charts* (or, equivalently, mirror flows generated by the associated Bregman divergence).

#### E.2.4. Entropic OT / Schrödinger (EOT) gradient flows

The entropic OT (Schrödinger) geometry can be characterized by the dynamic constraint

$$\partial_t p + \nabla \cdot (pv) = \frac{\varepsilon}{2} \Delta p, \quad (99)$$

and the action  $\int_0^1 \int \frac{1}{2} \|v_t(x)\|^2 p_t(x) dx dt$ . This defines an ‘‘entropic Wasserstein’’ (Schrödinger) norm on tangent vectors  $\dot{p}$  via the infimum over velocities  $v$  satisfying (99). In Onsager form, the induced operator is the same transport operator as Wasserstein, but now coupled to the diffusion constraint (99).

Formally, the steepest descent of  $\mathcal{F}$  under this geometry yields the viscous transport PDE

$$\partial_t p_t = \nabla \cdot \left( p_t \nabla \frac{\delta \mathcal{F}}{\delta p}(p_t) \right) + \frac{\varepsilon}{2} \Delta p_t, \quad (100)$$

i.e. the Wasserstein gradient flow (94) augmented by a diffusion term of strength  $\varepsilon/2$ . Equivalently, (100) is the Fokker–Planck equation of the SDE

$$dX_t = -\nabla \frac{\delta \mathcal{F}}{\delta p}(p_t)(X_t) dt + \sqrt{\varepsilon} dB_t, \quad \mathcal{L}(X_t) = p_t. \quad (101)$$

Thus, relative to Otto calculus, EOT gradient flows can be interpreted as ‘‘Wasserstein steepest descent under an entropic (Brownian) reference’’.

## F. SAMPLING FROM GEOMETRIC MIXTURES WITH REWEIGHTING

In the previous sections, we developed a geometric viewpoint on probability distributions and established a precise correspondence between evolution equations on the space of probability measures and stochastic differential equations on the underlying state space. In particular, we showed how transport, reweighting, and hybrid dynamics can be simulated using weighted stochastic differential equations arising from Feynman–Kac type partial differential equations.

This is already a great tool that could be applied to for faster implementation, as it lets samples go over the energy barriers by getting resampled. For performance analysis, look at chapter 6 [9]. More interestingly, having built geometries, one may ask does this geometric view help us sample from different notions of geodesics? In this section, following the steps of [35], we try to illustrate usage of weighted sde to accomplish sampling from mixture of probabilities we already can run a score-based diffusion model on.

Let  $\rho$  be a reference distribution from which sampling is tractable(i.e. white noise), and let

$$X^{(0)} \sim \rho.$$

Starting from  $\rho$ , suppose we are given two pretrained score-based diffusion models targeting distributions  $q^1$  and  $q^2$ , respectively. We assume that these models induce marginal distributions  $\{q_t^i\}_{t \in [0,1], i \in \{1,2\}}$  satisfying

$$q_0^1 = q_0^2 = \rho, \quad q_1^i = q^i,$$

and that both evolutions follow the *same score-based diffusion dynamics*, differing only in their learned score functions, i.e. for  $i$ , the marginal density satisfies the Fokker–Planck equation

$$\frac{\partial q_t^i}{\partial t} = -\nabla \cdot (q_t^i(-f_t + \sigma_t^2 \nabla \log q_t^i)) + \frac{\sigma_t^2}{2} \Delta q_t^i, \quad (102)$$

If you look into it from the perspective of the geometry, the intermediate probability distributions  $q_t$ ’s are nothing but the geodesic that connects the two probability distributions in the Shrodinger Bridge, or the **Entropic Optimal Transport Geometry**[39]. With the previous section in hand, this pde for probability distributions admits the stochastic representation

$$dX_t^i = (-f_t(X_t^i) + \sigma_t^2 \nabla \log q_t^i(X_t^i)) dt + \sigma_t dB_t, \quad (103)$$

where  $(B_t)_{t \geq 0}$  is a standard Brownian motion. and score functions  $s_i^t(x) := \nabla \log q_t^i(x)$ , which allows us to work directly at the level of stochastic dynamics and to apply the conversion rules between PDEs and SDEs developed earlier.

Our goal is to sample from distributions obtained by different interpolations between  $q^1$  and  $q^2$  using the machinery of Hellinger-Kantrovich(Wasserstein-Fisher-Rao) Geometry. We consider the following four canonical interpolations:

$${}^{\text{Mix}}\pi_\beta := \beta q^1 + (1 - \beta) q^2, \quad (104)$$

$$\log^{\text{Ex}}\pi_\beta := \beta \log q^1 + (1 - \beta) \log q^2 - \log \phi(\beta), \quad (105)$$

$$\sqrt{\text{FR}}\pi_\beta := \beta \sqrt{q^1} + (1 - \beta) \sqrt{q^2}, \quad (106)$$

corresponding respectively to the Wasserstein, mixture, exponential, and Fisher–Rao geodesics.

*Can we use our knowledge of sampling from  $q^i$  to sample from these interpolations?*

A natural heuristic is to define, for each  $t \in [0, 1]$ , an interpolated distribution  $\pi_\beta^t$  by applying the same geometric mixture operation to  $q_t^1$  and  $q_t^2$ . By construction,

$$\pi_\beta^0 = \rho, \quad \pi_\beta^1 = \pi_\beta.$$

At the level of stochastic dynamics, one might attempt to simulate a diffusion driven by an interpolated score  $s_\pi^t$  constructed from  $s_\mu^t$  and  $s_\nu^t$ . However, as shown in the score-based diffusion literature, such heuristic score interpolation does *not* in general reproduce the prescribed marginal evolution, and fails to sample from  $\pi_\beta$  at terminal time except in special cases.

To obtain correct sampling dynamics, we adopt the Weighted SDE introduced in Section 2 and in [35]. For a given geometric interpolation, we:

1. derive the evolution equation satisfied by  $\pi_\beta^t$ ,
2. identify the transport and diffusion terms corresponding to a score-based drift

$$v_\beta^t(x) = -f_t(x) + \sigma_t^2 \nabla \log \pi_\beta^t(x),$$

3. and collect all remaining terms into a multiplicative correction  $\psi_\beta^t(x)$ .

This yields a Feynman–Kac type PDE of the form

$$\partial_t \pi_\beta^t = -\nabla \cdot (\pi_\beta^t v_\beta^t) + \frac{\sigma_t^2}{2} \Delta \pi_\beta^t + \pi_\beta^t \psi_\beta^t, \quad (107)$$

which admits a stochastic representation via a weighted score-based diffusion, together with standard resampling schemes.

In the following subsections, we apply this construction to each of the four geometric interpolations. For each case, we derive the explicit correction term  $\psi_\beta^t$  and the corresponding weighted SDE whose terminal law is exactly  $\pi_\beta$ .

### F.1. Linear (Convex) Mixture Interpolation: linear PDE closure, but score-only sampling still needs ratio tracking

Fix  $\beta \in [0, 1]$  and define

$$p_{t,\beta}^{\text{mix}}(x) := (1 - \beta) q_t^1(x) + \beta q_t^2(x). \quad (108)$$

A common pitfall is to treat (108) as “nonlinear” because it contains the score  $s_t^i = \nabla \log q_t^i$ . However, the identity  $q \nabla \log q = \nabla q$  shows that (108) is equivalent to a *linear* Fokker–Planck equation in divergence form.

Because (108) is linear in  $q$ , convex mixtures are closed.

Unlike the exponential (geometric average) interpolation, the mixture score cannot be expressed solely from  $s_t^1, s_t^2$  without knowing the local density ratio  $q_t^2/q_t^1$ .

**Lemma 4** (Mixture score as a *ratio-weighted* score average). *For  $p_{t,\beta}^{\text{mix}} = (1 - \beta)q_t^1 + \beta q_t^2$ ,*

$$\nabla \log p_{t,\beta}^{\text{mix}}(x) = \omega_{t,\beta}^1(x) s_t^1(x) + \omega_{t,\beta}^2(x) s_t^2(x), \quad (109)$$

where the state-dependent weights are

$$\begin{aligned} \omega_{t,\beta}^1(x) &:= \frac{(1 - \beta)q_t^1(x)}{(1 - \beta)q_t^1(x) + \beta q_t^2(x)}, \\ \omega_{t,\beta}^2(x) &:= \frac{\beta q_t^2(x)}{(1 - \beta)q_t^1(x) + \beta q_t^2(x)}, \\ \omega_{t,\beta}^1 + \omega_{t,\beta}^2 &= 1. \end{aligned} \quad (110)$$

Equivalently, writing the log-ratio  $\ell_t(x) := \log \frac{q_t^2(x)}{q_t^1(x)}$ ,

$$\omega_{t,\beta}^2(x) = \frac{\beta e^{\ell_t(x)}}{(1 - \beta) + \beta e^{\ell_t(x)}}, \quad \omega_{t,\beta}^1(x) = 1 - \omega_{t,\beta}^2(x). \quad (111)$$

*Proof.*  $\nabla p = (1 - \beta)\nabla q^1 + \beta\nabla q^2 = (1 - \beta)q^1 s^1 + \beta q^2 s^2$ , then divide by  $p$ .  $\square$

Even though  $p_{t,\beta}^{\text{mix}}$  admits an *unweighted* SDE representation

$$dX_t = (-f_t(X_t) + \sigma_t^2 \nabla \log p_{t,\beta}^{\text{mix}}(X_t)) dt + \sigma_t dW_t,$$

one cannot implement this drift from access to  $(s_t^1, s_t^2)$  alone unless one also tracks  $\ell_t(X_t)$  (or equivalently the weights  $\omega_{t,\beta}^i$ ).

#### F.1.1. (III) Score-only implementable sampler: isotropic drift from scores + auxiliary ratio tracking; no FKC potential

The correct fix for “score-only” sampling is not an FKC potential (it is indeed zero for the linear FP operator), but rather an *auxiliary dynamics* to recover the ratio weights in Lemma 4.

**Lemma 5** (Closed PDE for the log-ratio  $\ell_t = \log(q_t^2/q_t^1)$ ). *Under (128), the log-ratio  $\ell_t(x) = \log \frac{q_t^2(x)}{q_t^1(x)}$  satisfies*

$$\begin{aligned} \partial_t \ell_t &= \langle f_t, \nabla \ell_t \rangle + \frac{\sigma_t^2}{2} \Delta \ell_t \\ &- \frac{\sigma_t^2}{2} \left( \nabla \cdot (s_t^2 - s_t^1) + \|s_t^2\|^2 - \|s_t^1\|^2 \right), \quad \nabla \ell_t = s_t^2 - s_t^1. \end{aligned} \quad (112)$$

*Proof.* Apply the identity (141) (proved in the earlier section) to  $q_t^1$  and  $q_t^2$  and subtract; use  $\nabla \ell_t = s_t^2 - s_t^1$  and  $\Delta \ell_t = \nabla \cdot (s_t^2 - s_t^1)$ .  $\square$

**Proposition 1** (Score-only mixture sampler (no FKC potential, but ratio tracking)). *Assume we can query  $s_t^1, s_t^2$  and (optionally)  $\nabla \cdot s_t^1, \nabla \cdot s_t^2$  (e.g. via Hutchinson estimators). Fix  $\beta \in [0, 1]$ . Define the mixture score by*

$$\begin{aligned} s_{t,\beta}^{\text{mix}}(x) &:= \nabla \log p_{t,\beta}^{\text{mix}}(x) = \omega_{t,\beta}^1(x) s_t^1(x) + \omega_{t,\beta}^2(x) s_t^2(x), \\ \omega_{t,\beta}^2(x) &= \frac{\beta e^{\ell_t(x)}}{(1 - \beta) + \beta e^{\ell_t(x)}}, \end{aligned} \quad (113)$$

with  $\ell_t$  tracked along the particle trajectory via Itô applied to Lemma 5. Then the (unweighted) SDE

$$dX_t = \left( -f_t(X_t) + \sigma_t^2 s_{t,\beta}^{\text{mix}}(X_t) \right) dt + \sigma_t dW_t \quad (114)$$

has marginal law  $p_{t,\beta}^{\text{mix}}$  (formally, in the ideal continuous-time limit). Moreover, there is no Fisher–Rao/Feynman–Kac correction:

$$g_{t,\beta}^{\text{mix}} \equiv 0, \quad \bar{g}_{t,\beta}^{\text{mix}} \equiv 0, \quad (115)$$

because  $p_{t,\beta}^{\text{mix}}$  already satisfies the linear FP operator (115).

At any fixed  $t$ , sampling  $p_{t,\beta}^{\text{mix}}$  is trivial: draw  $Z \sim \text{Bernoulli}(\beta)$  and sample from  $q_t^2$  if  $Z = 1$ , else from  $q_t^1$ . This produces the correct *mixture at time  $t$*  but does not yield a single guided reverse-time SDE whose entire marginal path is  $\{p_{t,\beta}^{\text{mix}}\}_t$  without additional coupling (such as (114) with ratio tracking).

### F.2. Geometric Average (Classifier-Free Guidance) and Feynman–Kac Correctors

Fix  $\beta \in \mathbb{R}$  and define the (normalized) *geometric average* marginal

$$\begin{aligned} p_{t,\beta}^{\text{geo}}(x) &:= \frac{1}{Z_t(\beta)} (q_t^1(x))^{1-\beta} (q_t^2(x))^\beta, \\ Z_t(\beta) &:= \int (q_t^1)^{1-\beta} (q_t^2)^\beta dx. \end{aligned} \quad (116)$$

This is the same target family considered in Eq. (15) of [35].

A widely-used heuristic (classifier-free guidance) is to simulate the reverse-time SDE with the mixed score

$$\nabla \log p_{t,\beta}^{\text{geo}}(x) = (1 - \beta) \nabla \log q_t^1(x) + \beta \nabla \log q_t^2(x), \quad (117)$$

but in general this does *not* reproduce the prescribed marginals  $\{p_{t,\beta}^{\text{geo}}\}_{t \in [0,1]}$  without correction (see the discussion around Eq. (16) in [35]).

The Feynman–Kac Corrector (FKC) methodology proceeds by: (i) differentiating the target path (116) in time, and (ii) rearranging the resulting PDE to isolate a transport–diffusion operator corresponding to an SDE drift involving  $\nabla \log p_{t,\beta}^{\text{geo}}$ , while (iii) collecting the leftover terms into a *reweighting potential* that is simulated through particle weights (Feynman–Kac). This is precisely the program outlined around Eq. (17)–(18) in [35].

**Proposition 2** (Geometric average via FKC; cf. Prop. 3.1 in [35]). *Assume  $q_t^1, q_t^2$  satisfy (128)–(129). Let  $p_{t,\beta}^{\text{geo}}$  be defined by (116). Define the guided drift*

$$\begin{aligned} v_{t,\beta}(x) &:= -f_t(x) + \sigma_t^2 \nabla \log p_{t,\beta}^{\text{geo}}(x) = -f_t(x) \\ &+ \sigma_t^2 ((1 - \beta) \nabla \log q_t^1(x) + \beta \nabla \log q_t^2(x)). \end{aligned} \quad (118)$$

Then  $p_{t,\beta}^{\text{geo}}$  solves a Feynman–Kac PDE of the form

$$\partial_t p_{t,\beta}^{\text{geo}} = -\nabla \cdot (p_{t,\beta}^{\text{geo}} v_{t,\beta}) + \frac{\sigma_t^2}{2} \Delta p_{t,\beta}^{\text{geo}} + \bar{g}_{t,\beta} p_{t,\beta}^{\text{geo}}, \quad (119)$$

where  $\bar{g}_{t,\beta}(x) := g_{t,\beta}(x) - \int g_{t,\beta}(y) p_{t,\beta}^{\text{geo}}(y) dy$  is the centered potential ensuring conservation of mass, and the (uncentered) potential admits the explicit expression

$$g_{t,\beta}(x) = \frac{\sigma_t^2}{2} \beta(\beta-1) \|\nabla \log q_t^1(x) - \nabla \log q_t^2(x)\|^2. \quad (120)$$

Consequently, (119) can be simulated by the weighted SDE

$$dX_t = v_{t,\beta}(X_t) dt + \sigma_t dW_t, \quad dw_t = \bar{g}_{t,\beta}(X_t) dt, \quad (121)$$

and resampling/reweighting (e.g. SNIS / SMC) yields samples whose law matches  $p_{t,\beta}^{\text{geo}}$  in the large-particle limit. Moreover, the weight evolution may equivalently be written (up to centering) as

$$dw_t = \frac{\sigma_t^2}{2} \beta(\beta-1) \|\nabla \log q_t^1(X_t) - \nabla \log q_t^2(X_t)\|^2 dt, \quad (122)$$

which matches Eq. (19) in [35].

*Proof.* We give a rigorous calculation under the standing smoothness/decay assumptions. Let the unnormalized density be

$$\tilde{p}_{t,\beta}(x) := (q_t^1(x))^{1-\beta} (q_t^2(x))^\beta, \quad p_{t,\beta}^{\text{geo}} = \tilde{p}_{t,\beta}/Z_t(\beta).$$

Differentiating  $\tilde{p}_{t,\beta}$  yields

$$\partial_t \tilde{p}_{t,\beta} = \tilde{p}_{t,\beta} \left( (1-\beta) \partial_t \log q_t^1 + \beta \partial_t \log q_t^2 \right). \quad (123)$$

From (129), using  $\partial_t \log q = (\partial_t q)/q$  and the identity  $\frac{\Delta q}{q} = \Delta \log q + \|\nabla \log q\|^2$ , we obtain

$$\partial_t \log q_t^i = -\frac{1}{q_t^i} \nabla \cdot (q_t^i (-f_t + \sigma_t^2 \nabla \log q_t^i)) \quad (124)$$

$$+ \frac{\sigma_t^2}{2} \left( \Delta \log q_t^i + \|\nabla \log q_t^i\|^2 \right)$$

$$= -\nabla \cdot (-f_t + \sigma_t^2 \nabla \log q_t^i) \quad (125)$$

$$- \langle \nabla \log q_t^i, -f_t + \sigma_t^2 \nabla \log q_t^i \rangle \quad (126)$$

$$+ \frac{\sigma_t^2}{2} \Delta \log q_t^i + \frac{\sigma_t^2}{2} \|\nabla \log q_t^i\|^2. \quad (127)$$

Insert (127) into (123) and regroup the resulting terms into (a) a divergence/transport part, (b) a Laplacian part, and (c) a residual scalar potential. The transport and diffusion parts can be shown to match precisely the operator

$$-\nabla \cdot (\tilde{p}_{t,\beta} v_{t,\beta}) + \frac{\sigma_t^2}{2} \Delta \tilde{p}_{t,\beta}, \quad v_{t,\beta} := -f_t + \sigma_t^2 \nabla \log \tilde{p}_{t,\beta},$$

where  $\nabla \log \tilde{p}_{t,\beta} = (1-\beta) \nabla \log q_t^1 + \beta \nabla \log q_t^2$  coincides with (117). The remaining scalar terms simplify (by cancellation of like terms and completing the square) to the explicit potential (120). Concretely, the only non-transport/diffusion contribution is

$$\frac{\sigma_t^2}{2} \left( \beta(\beta-1) \|\nabla \log q_t^1 - \nabla \log q_t^2\|^2 \right) \tilde{p}_{t,\beta},$$

which yields the unnormalized Feynman–Kac PDE for  $\tilde{p}_{t,\beta}$ .

Finally, normalization introduces a centering term: differentiating  $p_{t,\beta}^{\text{geo}} = \tilde{p}_{t,\beta}/Z_t(\beta)$  gives

$$\partial_t p_{t,\beta}^{\text{geo}} = \frac{1}{Z_t} \partial_t \tilde{p}_{t,\beta} - \frac{\dot{Z}_t}{Z_t} p_{t,\beta}^{\text{geo}} = \left( \dots \right) + (g_{t,\beta} - \mathbb{E}_{p_{t,\beta}^{\text{geo}}} [g_{t,\beta}]) p_{t,\beta}^{\text{geo}},$$

which is exactly (119) with  $\bar{g}_{t,\beta}$ . The weighted SDE representation (121) is the standard Feynman–Kac simulation rule for (119); the explicit drift/weight forms (118) and (122) match Prop. 3.1 / Eq. (19) in [35].  $\square$

### F.3. Fisher–Rao (Hellinger) interpolation via Feynman–Kac correctors: drift/weights from scores only

Fix  $\beta \in [0, 1]$ . we seek a *weighted SDE* (drift, diffusion, and Fisher–Rao potential) that generates samples from the Fisher–Rao/Hellinger interpolation

$$\pi_{t,\beta}^{\text{FR}}(dx) \propto ((1-\beta) \sqrt{q_t^1(x)} + \beta \sqrt{q_t^2(x)})^2 dx$$

at each time  $t$ , in the large-particle limit (SMC/Feynman–Kac), exactly in the spirit of the Feynman–Kac Corrector (FKC) methodology (e.g. [35]).

Assume  $q_t^1, q_t^2$  are strictly positive and  $C^{1,2}$  in  $(t, x)$ , and each solves the same score-based Fokker–Planck equation (cf. Eq. (14a) in [35]):

$$\partial_t q_t^i = -\nabla \cdot (q_t^i (-f_t + \sigma_t^2 s_t^i)) + \frac{\sigma_t^2}{2} \Delta q_t^i, \quad i \in \{1, 2\}, \quad (128)$$

with associated reverse-time denoising SDE (Eq. (14b) in [35])

$$dX_t^i = \left( -f_t(X_t^i) + \sigma_t^2 s_t^i(X_t^i) \right) dt + \sigma_t dW_t. \quad (129)$$

#### F.3.1. Hellinger mixture, pointwise mixture weights, and the guided score

Define the *unnormalized* Hellinger (FR) mixture

$$\begin{aligned} \tilde{p}_{t,\beta}^{\text{FR}}(x) &:= \left( (1-\beta) \sqrt{q_t^1(x)} + \beta \sqrt{q_t^2(x)} \right)^2, \\ p_{t,\beta}^{\text{FR}} &:= \tilde{p}_{t,\beta}^{\text{FR}} / Z_t^{\text{FR}}(\beta). \end{aligned} \quad (130)$$

Introduce the *pointwise* mixing coefficients

$$\begin{aligned} \alpha_{t,\beta}^1(x) &:= \frac{(1-\beta) \sqrt{q_t^1(x)}}{(1-\beta) \sqrt{q_t^1(x)} + \beta \sqrt{q_t^2(x)}}, \\ \alpha_{t,\beta}^2(x) &:= \frac{\beta \sqrt{q_t^2(x)}}{(1-\beta) \sqrt{q_t^1(x)} + \beta \sqrt{q_t^2(x)}}, \end{aligned} \quad (131)$$

so that  $\alpha_{t,\beta}^1(x) + \alpha_{t,\beta}^2(x) = 1$  and

$$\nabla \log \tilde{p}_{t,\beta}^{\text{FR}}(x) = \alpha_{t,\beta}^1(x) s_t^1(x) + \alpha_{t,\beta}^2(x) s_t^2(x),$$

$$\nabla \log p_{t,\beta}^{\text{FR}} = \nabla \log \tilde{p}_{t,\beta}^{\text{FR}}. \quad (132)$$

Unlike the geometric average case (classifier-free guidance), the mixing weights  $\alpha_{t,\beta}^i(x)$  are *state dependent*; hence the drift cannot be written as a *fixed* linear combination of scores. However, the weights can be reconstructed along particle trajectories from the *log-density ratio*

$$\ell_t(x) := \log \frac{q_t^2(x)}{q_t^1(x)}. \quad (133)$$

Indeed,

$$\alpha_{t,\beta}^2(x) = \frac{\beta e^{\ell_t(x)/2}}{(1-\beta) + \beta e^{\ell_t(x)/2}}, \quad \alpha_{t,\beta}^1(x) = 1 - \alpha_{t,\beta}^2(x). \quad (134)$$

Therefore, to implement the FR interpolation using *scores only*, it suffices to track  $\ell_t(X_t)$  along the sampling SDE.

### F.3.2. A closed-form FKC potential depending only on scores and FR weights

We now state the analog of Prop. 3.1 (geometric average) from [35], but for Fisher–Rao/Hellinger interpolation.

**Proposition 3** (Hellinger interpolation via FKC: explicit drift and potential). *Assume  $q_t^1, q_t^2$  satisfy (128) and are  $C^{1,2}$  with sufficient decay for integrations by parts. Fix  $\beta \in [0, 1]$ , and define  $p_{t,\beta}^{\text{FR}}$  by (130). Let the guided score be*

$$s_{t,\beta}^{\text{FR}}(x) := \nabla \log p_{t,\beta}^{\text{FR}}(x) = \alpha_{t,\beta}^1(x) s_t^1(x) + \alpha_{t,\beta}^2(x) s_t^2(x), \quad (135)$$

and define the guided drift

$$v_{t,\beta}^{\text{FR}}(x) := -f_t(x) + \sigma_t^2 s_{t,\beta}^{\text{FR}}(x). \quad (136)$$

Then  $p_{t,\beta}^{\text{FR}}$  solves the Feynman–Kac PDE

$$\partial_t p_{t,\beta}^{\text{FR}} = -\nabla \cdot (p_{t,\beta}^{\text{FR}} v_{t,\beta}^{\text{FR}}) + \frac{\sigma_t^2}{2} \Delta p_{t,\beta}^{\text{FR}} + \bar{g}_{t,\beta}^{\text{FR}} p_{t,\beta}^{\text{FR}}, \quad (137)$$

where  $\bar{g}_{t,\beta}^{\text{FR}}(x) := g_{t,\beta}^{\text{FR}}(x) - \mathbb{E}_{p_{t,\beta}^{\text{FR}}} [g_{t,\beta}^{\text{FR}}]$  centers the potential, and the uncentered potential admits the explicit form

$$g_{t,\beta}^{\text{FR}}(x) = -\frac{\sigma_t^2}{4} \alpha_{t,\beta}^1(x) \alpha_{t,\beta}^2(x) \|s_t^1(x) - s_t^2(x)\|^2. \quad (138)$$

Consequently, (137) can be simulated by the weighted SDE

$$dX_t = v_{t,\beta}^{\text{FR}}(X_t) dt + \sigma_t dW_t, \quad dw_t = \bar{g}_{t,\beta}^{\text{FR}}(X_t) dt, \quad (139)$$

together with standard Feynman–Kac reweighting/resampling. In the large-particle limit, the empirical measure of particles converges to  $p_{t,\beta}^{\text{FR}}$ .

*Proof (calculation; cancellations leaving scores only).* Work with the unnormalized density  $\tilde{p} = \tilde{p}_{t,\beta}^{\text{FR}}$  and write  $r_i := \sqrt{q_i^i}$  so that  $\tilde{p} = m^2$  with  $m := (1-\beta)r_1 + \beta r_2$ . Define  $u_i := \log r_i = \frac{1}{2} \log q_i^i$ . Then  $\log \tilde{p} = 2 \log m$  is a log-sum-exp in  $(u_1, u_2)$ .

*Step 1 (guided score).* Differentiating  $\log \tilde{p} = 2 \log m$  in space yields

$$\nabla \log \tilde{p} = 2 \frac{\nabla m}{m} = 2 \frac{(1-\beta)r_1 \nabla u_1 + \beta r_2 \nabla u_2}{(1-\beta)r_1 + \beta r_2} = \alpha^1 s^1 + \alpha^2 s^2,$$

where we used  $\nabla u_i = \frac{1}{2} s^i$  and the definitions (131). This gives (135) and thus (136).

*Step 2 (time derivative of  $\log \tilde{p}$ ).* Since  $\log m$  is log-sum-exp,  $\partial_t \log m$  is the same convex combination of  $\partial_t u_i$ :

$$\begin{aligned} \partial_t \log \tilde{p} &= 2 \partial_t \log m = \\ &2(\alpha^1 \partial_t u_1 + \alpha^2 \partial_t u_2) = \alpha^1 \partial_t \log q^1 + \alpha^2 \partial_t \log q^2. \end{aligned} \quad (140)$$

From (128) and the identity  $\Delta q/q = \Delta \log q + \|\nabla \log q\|^2$  (with  $\Delta \log q = \nabla \cdot s$ ), one checks (as in the standard score-based expansions in [35]) that for each  $i$ ,

$$\partial_t \log q_t^i = \nabla \cdot f_t + \langle s_t^i, f_t \rangle - \frac{\sigma_t^2}{2} (\nabla \cdot s_t^i + \|s_t^i\|^2). \quad (141)$$

Hence

$$\begin{aligned} \partial_t \log \tilde{p} &= \nabla \cdot f_t + \langle s, f_t \rangle \\ &- \frac{\sigma_t^2}{2} (\alpha^1 (\nabla \cdot s^1 + \|s^1\|^2) + \alpha^2 (\nabla \cdot s^2 + \|s^2\|^2)), \end{aligned} \quad (142)$$

where  $s = \alpha^1 s^1 + \alpha^2 s^2$ .

*Step 3 (compare to the “naive” score-based PDE).* If  $\tilde{p}$  evolved purely under the guided score drift  $v = -f + \sigma^2 s$ , then (exactly as in the derivation of (141)) we would have

$$\partial_t \log \tilde{p} = \nabla \cdot f_t + \langle s, f_t \rangle - \frac{\sigma_t^2}{2} (\nabla \cdot s + \|s\|^2) \quad (\text{no corrector}). \quad (143)$$

Subtracting (143) from (142) yields the residual

$$g = -\frac{\sigma_t^2}{2} (\alpha^1 (\nabla \cdot s^1 + \|s^1\|^2) + \alpha^2 (\nabla \cdot s^2 + \|s^2\|^2) - (\nabla \cdot s + \|s\|^2)).$$

Now use two identities valid for the log-sum-exp mixture:

$$\nabla \cdot s = \alpha^1 \nabla \cdot s^1 + \alpha^2 \nabla \cdot s^2 + \frac{1}{2} \alpha^1 \alpha^2 \|s^1 - s^2\|^2, \quad (144)$$

$$\|s\|^2 = \alpha^1 \|s^1\|^2 + \alpha^2 \|s^2\|^2 - \alpha^1 \alpha^2 \|s^1 - s^2\|^2. \quad (145)$$

Combining (144)–(145) gives

$$\begin{aligned} \alpha^1 (\nabla \cdot s^1 + \|s^1\|^2) + \alpha^2 (\nabla \cdot s^2 + \|s^2\|^2) \\ - (\nabla \cdot s + \|s\|^2) = \frac{1}{2} \alpha^1 \alpha^2 \|s^1 - s^2\|^2, \end{aligned} \quad (146)$$

hence

$$g = -\frac{\sigma_t^2}{4} \alpha^1 \alpha^2 \|s^1 - s^2\|^2,$$

which is (138). Finally, normalizing  $\tilde{p}$  to  $p = \tilde{p}/Z_t$  introduces the standard centering term  $\bar{g} = g - \mathbb{E}_p[g]$ , yielding (137).

The weighted SDE (139) is the standard Feynman–Kac simulation rule for (137) (as in [35]).  $\square$

The potential (138) is (up to constants) the *pointwise variance* of the two scores under the FR weights:

$$\alpha^1 \alpha^2 \|s^1 - s^2\|^2 = \alpha^1 \|s^1\|^2 + \alpha^2 \|s^2\|^2 - \|\alpha^1 s^1 + \alpha^2 s^2\|^2.$$

Thus the corrector punishes regions where the two models disagree strongly about the local score, and the penalty is maximal where the FR mixture is ambiguous (i.e.  $\alpha^1 \approx \alpha^2 \approx \frac{1}{2}$ ).

**How to compute the FR weights  $\alpha^i$  from scores only: an auxiliary log-ratio SDE**

Proposition 3 expresses the drift and potential using  $s_t^1, s_t^2$  and the weights  $\alpha^i(x)$ , which depend on the (unknown) density ratio  $\ell_t(x) = \log(q_t^2/q_t^1)$ . We now show how to track  $\ell_t$  along particle trajectories using only score-accessible quantities. The key observation is that  $\ell_t$  satisfies a *closed* PDE involving only the scores and their divergences.

**Lemma 6** (PDE for the log-density ratio). Let  $\ell_t(x) = \log(\frac{q_t^2(x)}{q_t^1(x)})$ . Under (128),  $\ell_t$  satisfies

$$\partial_t \ell_t = \langle f_t, \nabla \ell_t \rangle + \frac{\sigma_t^2}{2} \Delta \ell_t - \frac{\sigma_t^2}{2} \left( \nabla \cdot (s_t^2 - s_t^1) + \|s_t^2\|^2 - \|s_t^1\|^2 \right). \quad (147)$$

Equivalently, using  $\Delta \ell_t = \nabla \cdot (s_t^2 - s_t^1)$ ,

$$\partial_t \ell_t = \langle f_t, \nabla \ell_t \rangle + \frac{\sigma_t^2}{2} \Delta \ell_t - \frac{\sigma_t^2}{2} \left( \Delta \ell_t + \|s_t^2\|^2 - \|s_t^1\|^2 \right). \quad (148)$$

*Proof.* Subtract (141) for  $i = 1$  from the same identity for  $i = 2$ . The  $\nabla \cdot f_t$  terms cancel, and the remaining terms can be written in the form (147) using  $\nabla \ell_t = s_t^2 - s_t^1$  and  $\Delta \ell_t = \nabla \cdot (s_t^2 - s_t^1)$ .  $\square$

Let  $X_t$  solve the FR-guided SDE in (139). Applying Itô's formula to  $\ell_t(X_t)$ , one obtains an SDE for the running log-ratio along the particle:

$$d\ell_t(X_t) = \left( \partial_t \ell_t + \langle v_{t,\beta}^{\text{FR}}, \nabla \ell_t \rangle + \frac{\sigma_t^2}{2} \Delta \ell_t \right) dt + \sigma_t \langle \nabla \ell_t, dW_t \rangle. \quad (149)$$

Using  $\nabla \ell_t = s_t^2 - s_t^1$  and the PDE (147), this drift can be written entirely in terms of  $(f_t, \sigma_t)$ , the scores  $s_t^1, s_t^2$ , and their divergences  $\nabla \cdot s_t^1, \nabla \cdot s_t^2$  (which are standard to estimate in score models via Hutchinson trace estimators). In particular, once  $\ell_t(X_t)$  is tracked, the FR weights  $\alpha_{t,\beta}^i(X_t)$  are obtained from (134).

**Practical procedure (conceptual).** Starting from pure noise at  $t = 1$ :

1. Initialize particles  $X_1^{(k)} \sim \mathcal{N}(0, I)$  and initialize auxiliary  $\ell_1^{(k)}$  (e.g. 0 if the two models share the same terminal noise law).
2. For  $t \downarrow 0$  (discretize time), at each particle compute:

$$s_t^1(X_t^{(k)}), \quad s_t^2(X_t^{(k)}), \\ \alpha_{t,\beta}^2(X_t^{(k)}) = \frac{\beta e^{\ell_t^{(k)}/2}}{(1-\beta) + \beta e^{\ell_t^{(k)}/2}}, \quad \alpha^1 = 1 - \alpha^2, \quad (150)$$

then form the guided score  $s_{t,\beta}^{\text{FR}} = \alpha^1 s^1 + \alpha^2 s^2$  and drift  $v_{t,\beta}^{\text{FR}} = -f_t + \sigma_t^2 s_{t,\beta}^{\text{FR}}$ .

3. Evolve the state particle by Euler–Maruyama (reverse-time convention)

$$X_{t-\Delta t} = X_t + v_{t,\beta}^{\text{FR}}(X_t) \Delta t + \sigma_t \sqrt{\Delta t} \xi, \quad \xi \sim \mathcal{N}(0, I).$$

4. Evolve the auxiliary log-ratio  $\ell_t$  using (149) with  $\nabla \ell_t = s^2 - s^1$  and  $\Delta \ell_t = \nabla \cdot (s^2 - s^1)$ .

5. Update weights using the FR potential

$$dw_t = \bar{g}_{t,\beta}^{\text{FR}}(X_t) dt, \quad g_{t,\beta}^{\text{FR}}(X_t) = -\frac{\sigma_t^2}{4} \alpha^1 \alpha^2 \|s^1 - s^2\|^2,$$

and resample as needed.

This yields samples from  $\pi_{0,\beta}^{\text{FR}}$  in the Feynman–Kac / SMC limit.

The drift (136) and the potential (138) are expressed using:

$$f_t, \sigma_t, s_t^1, s_t^2,$$

No evaluation of  $p_{t,\beta}^{\text{FR}}$  (or its normalization constant) is required. The only additional ingredient beyond score queries is  $\nabla \cdot s_t^i$  (i.e.  $\Delta \log q_t^i$ ), which is standard in modern diffusion implementations (Hutchinson estimators) and is unavoidable if one insists on an *exact* ratio-tracking mechanism.

### F.3.3. Comparison to the exponential interpolation (classifier-free guidance + FKC)

For the geometric average  $p_{t,\beta}^{\text{geo}} \propto (q_t^1)^{1-\beta} (q_t^2)^\beta$ , the guided score is *exactly* the constant convex combination  $(1-\beta)s^1 + \beta s^2$ , and the FKC potential collapses to the simple closed-form discrepancy  $\frac{\sigma_t^2}{2} \beta(\beta-1) \|s^1 - s^2\|^2$  (Prop. 3.1 in [35]). For Fisher–Rao/Hellinger interpolation, the same structure persists: the corrector is again a negative score-discrepancy penalty, but now it is *weighted pointwise* by  $\alpha^1 \alpha^2$ , reflecting the state-dependent uncertainty of the Hellinger mixture.

Concept	OU specialization (and the general semigroup identities it instantiates)
OU SDE	$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t, \quad V(x) = \frac{\alpha}{2} \ x\ ^2$ ( $\alpha > 0$ ), so $\nabla V(x) = \alpha x$ and $dX_t = -\alpha X_t dt + \sqrt{2} dB_t$ .
Markov semigroup $(P_t)_{t \geq 0}$	$(P_t f)(x) := \mathbb{E}[f(X_t)   X_0 = x]$ (time-homogeneous; semigroup property $P_s P_t = P_{s+t}$ ).
Infinitesimal generator $\mathcal{L}$	$\mathcal{L}f := \lim_{t \downarrow 0} \frac{P_t f - f}{t}$ (on the appropriate domain).
Explicit generator (OU/Langevin form)	For $f \in C^2(\mathbb{R}^d)$ , $(\mathcal{L}f)(x) = \Delta f(x) - \langle \nabla V(x), \nabla f(x) \rangle = \Delta f(x) - \alpha \langle x, \nabla f(x) \rangle.$ (This is exactly the Langevin generator formula specialized to quadratic $V$ .)
Kolmogorov (semigroup) evolution for observables	For all $t \geq 0$ , $\partial_t P_t f = \mathcal{L} P_t f = P_t \mathcal{L} f,$ i.e. $u_t := P_t f$ solves $\partial_t u_t = \mathcal{L} u_t$ .
Forward (Fokker–Planck) equation	Writing $\pi_t$ for the law/density of $X_t$ (w.r.t. Lebesgue), $\partial_t \pi_t = \mathcal{L}^* \pi_t,$ where $\mathcal{L}^*$ is the $L^2(\mathfrak{m})$ -adjoint of $\mathcal{L}$ .
Stationary distribution $\pi$	$\pi$ is stationary iff $\mathcal{L}^* \pi = 0$ . For OU with $V(x) = \frac{\alpha}{2} \ x\ ^2$ , $\pi(x) \propto \exp(-V(x)) = \exp\left(-\frac{\alpha}{2} \ x\ ^2\right), \quad \text{i.e. } \pi = \mathcal{N}(0, \alpha^{-1} I_d).$ The OU/Langevin semigroup is reversible w.r.t. $\pi$ , so $\mathcal{L}$ is symmetric on $L^2(\pi)$ .
Reversibility (w.r.t. $\pi$ )	
Carré du champ $\Gamma$ (your $T$ )	Define $\Gamma(f, g) := \frac{1}{2} (\mathcal{L}(fg) - f \mathcal{L}g - g \mathcal{L}f).$ For the diffusion generator above (in particular OU), $\Gamma(f, g) = \langle \nabla f, \nabla g \rangle, \quad \Gamma(f, f) = \ \nabla f\ ^2.$
Dirichlet energy / Dirichlet form $\mathcal{E}$	$\mathcal{E}(f, g) := \int \Gamma(f, g) d\pi, \quad \mathcal{E}(f, f) = \int \ \nabla f\ ^2 d\pi.$
Fundamental integration-by-parts identity	For reversible semigroups, $\int f (-\mathcal{L})g d\pi = \int \Gamma(f, g) d\pi = \mathcal{E}(f, g), \quad \text{hence } -\mathcal{L} \succeq 0.$
Iterated carré du champ $\Gamma_2$ (your $T_2$ )	Define $\Gamma_2(f, g) := \frac{1}{2} (\mathcal{L}\Gamma(f, g) - \Gamma(f, \mathcal{L}g) - \Gamma(g, \mathcal{L}f)).$ (For OU, $\Gamma_2$ can be computed explicitly and is the object controlling curvature-dimension/Bakry–Émery estimates, but the defining identity above is the canonical “reader convenience” form.)

**Table 5:** Ornstein–Uhlenbeck (OU) process as the quadratic-potential Langevin diffusion: semigroup  $P_t$ , generator  $\mathcal{L}$ , forward/backward PDEs, and the quadratic forms  $T = \Gamma$ ,  $T_2 = \Gamma_2$ , and Dirichlet energy  $\mathcal{E}$ .