PDEs for machine learning: an optimal-transport-oriented review

Gabriel Peyré

CNRS and DMA, École normale supérieure, PSL University, Paris, France

gabriel.peyre@ens.fr

Abstract. Machine learning increasingly relies on evolutions of probability measures. Training dynamics of wide neural networks, interacting particles, diffusion models, flow matching, attention layers and generative maps can all be described through transport, continuity equations, variational flows or stochastic PDE limits. This review develops an optimal-transport-oriented account of these PDE tools. The emphasis is not on PDE theory in isolation, but on the mechanisms by which PDEs organize learning algorithms: mass conservation, least-action principles, Wasserstein gradients, Fokker–Planck equations, McKean–Vlasov limits and transport-based generative dynamics. The document is designed as a standalone PDE/ML review extracted and reorganized from the larger OT4ML project. It first gives a compact primer on Monge and Kantorovich optimal transport, Brenier maps and Wasserstein metric geometry. It then develops the material most directly relevant to PDE methods in machine learning: dynamic optimal transport, Wasserstein gradient flows, mean-field training dynamics, diffusion and flow-matching models, transportation views of transformers, and Gaussian closures. The goal is to expose a coherent mathematical language in which ML algorithms become evolutions on spaces of measures.

Keywords. optimal transport; partial differential equations; machine learning; Wasserstein gradient flows; diffusion models; flow matching

The source code for the full OT4ML project, including figure-generation notebooks, is available at `gpeyre/ot4ml`. Most computational figures were produced with the Python Optimal Transport (POT) library Flamary et al. (2021).

Guide to the Literature and Scope

This review uses optimal transport as the entry point to PDE methods for machine learning. The mathematical foundations of OT are covered in Villani's monographs Villani (2003, 2009), Santambrogio's applied text Santambrogio (2015), the probabilistic treatment of Rachev and Rüschendorf Rachev and Rüschendorf (1998a,b), and the computational account of Peyré and Cuturi Peyré and Cuturi (2019). Gradient flows in metric spaces are developed systematically by Ambrosio, Gigli and Savaré Ambrosio et al. (2006), while the dynamic Benamou–Brenier viewpoint originates in Benamou and Brenier (2000).

Reference map. The review repeatedly returns to a small set of pillars. Static OT and Wasserstein geometry follow the classical line from Monge and Kantorovich to Brenier, McCann and Villani Monge (1781); Kantorovich (1942); Brenier (1991); McCann (1997); Villani (2009). Dynamic OT and gradient flows use the Benamou–Brenier formula, the JKO scheme and Otto's formal Riemannian calculus Benamou and Brenier (2000); Benamou et al. (2016); Otto (2001); Ambrosio et al. (2006). Machine-learning dynamics enter through mean-field neural-network limits and particle descriptions Chizat and Bach (2018); Mei et al. (2018); Rotskoff and Vanden-Eijnden (2022). Generative modelling is organized around score-based diffusion, probability-flow ODEs, flow matching, rectified flows and stochastic interpolants Song et al. (2021); Lipman et al. (2023); Liu et al. (2023); Albergo et al. (2025).

For machine learning, the relevant PDE themes include Fokker–Planck and McKean–Vlasov limits of particle systems, JKO schemes, entropy and interaction-energy flows, mean-field neural-network training, diffusion models, score-based generative modeling, flow matching and stochastic interpolants. The corresponding references are recalled locally in the text, so that each citation is tied to the mathematical mechanism being discussed.

The scope is deliberately selective. The primer only includes the OT material required later: push-forwards, Monge maps, Kantorovich couplings, Brenier's theorem, Wasserstein distances and geodesics. Algorithmic OT, duality, Sinkhorn convergence and generalized OT are treated in the full OT4ML manuscript; here they are invoked only when they support PDE or ML dynamics.

Reading map. The four sections move from static OT notation to dynamic least-action principles, then to Wasserstein gradient flows and particle limits, and finally to transportation dynamics for generative models and continuous-depth architectures.

Contents

1. A Primer on Optimal Transport

This primer fixes the optimal-transport notation used throughout the review. The goal is not to reproduce the full OT4ML development, but to keep the minimal objects needed for PDE methods: push-forward maps, Monge and Kantorovich formulations, Brenier maps, Wasserstein distances and displacement geodesics. The figures are used as a visual glossary: each one is placed next to the definition or theorem whose geometry it makes concrete.

1.1. Measures and Push-Forwards

Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish metric spaces, meaning complete separable metric spaces. This assumption is the natural background for probability theory because it gives enough measurable structure for disintegration, weak convergence and compactness arguments while still covering Euclidean spaces, manifolds, discrete spaces and many path spaces.

Weak convergence and moment classes. The PDE constructions below use both weak convergence of laws and the stronger topology induced by transport. A sequence $(\alpha_n)_n \subset \mathcal{P}(\mathcal{X})$ converges weakly to α, written $\alpha_n \rightharpoonup \alpha$, when

$$\int g \, \mathrm{d}\alpha_n \longrightarrow \int g \, \mathrm{d}\alpha$$

for every bounded continuous function g. After fixing $x_0 \in \mathcal{X}$, define

$$\mathcal{P}_p(\mathcal{X}) := \left\{ \alpha \in \mathcal{P}(\mathcal{X}) : \int d(x, x_0)^p \, \mathrm{d}\alpha(x) < +\infty \right\}. \tag{1.1}$$

This definition does not depend on the choice of x_0. The finite p-moment assumption is exactly what makes the p-Wasserstein distance finite. On a Polish space, $\mathcal{P}_p(\mathcal{X})$ is itself Polish for $\mathcal{W}_p$, and

$$\mathcal{W}_p(\alpha_n, \alpha) \longrightarrow 0 \quad \Longleftrightarrow \quad \alpha_n \rightharpoonup \alpha \qquad \text{and} \qquad \int d(x, x_0)^p \, \mathrm{d}\alpha_n(x) \longrightarrow \int d(x, x_0)^p \, \mathrm{d}\alpha(x). \tag{1.2}$$

These standard facts explain why Wasserstein compactness arguments combine tightness with moment bounds Villani (2009); Ambrosio et al. (2006).

Conditional laws and disintegration. Flow matching repeatedly conditions a latent trajectory on its current position, so the underlying measure-theoretic operation should be explicit. If $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ has first marginal α, the disintegration theorem on Polish spaces provides a measurable family $(\pi_x)_{x \in \mathcal{X}} \subset \mathcal{P}(\mathcal{Y})$ such that

$$\int_{\mathcal{X} \times \mathcal{Y}} h(x, y) \, \mathrm{d}\pi(x, y) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} h(x, y) \, \mathrm{d}\pi_x(y) \right) \mathrm{d}\alpha(x) \tag{1.3}$$

for every bounded measurable h. The family is unique for α-almost every x and is the conditional law of Y given $X = x$. In particular, for an integrable vector-valued function $q(x, y)$,

$$\mathbb{E}[q(X, Y) \mid X = x] = \int q(x, y) \, \mathrm{d}\pi_x(y).$$

This is the conditional expectation that later turns latent path velocities into Eulerian vector fields.

Definition 1.1 (Discrete measure)**.** A discrete probability measure on $\mathcal{X}$ is

$$\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}, \qquad a_i \geqslant 0, \qquad \sum_{i=1}^{n} a_i = 1. \tag{1.4}$$

If $a_i = 1/n$, this is an empirical measure.

Definition 1.2 (Push-forward)**.** Let $T : \mathcal{X} \to \mathcal{Y}$ be measurable and let $\alpha \in \mathcal{P}(\mathcal{X})$. The push-forward $T_\sharp \alpha \in \mathcal{P}(\mathcal{Y})$ is defined by

$$(T_\sharp \alpha)(B) = \alpha(T^{-1}(B)) \tag{1.5}$$

for all Borel sets $B \subset \mathcal{Y}$. Equivalently, for all bounded measurable $g : \mathcal{Y} \to \mathbb{R}$,

$$\int_{\mathcal{Y}} g(y) \, \mathrm{d}(T_\sharp \alpha)(y) = \int_{\mathcal{X}} g(T(x)) \, \mathrm{d}\alpha(x). \tag{1.6}$$

If X is a random variable with law α, then $T_\sharp \alpha$ is the law of $T(X)$. This simple identity is the bridge between deterministic maps, particle systems and measure-valued PDEs.

Proposition 1.3 (Push-forward formula for densities)**.** *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, let T be a diffeomorphism, and let $\alpha = \rho_\alpha(x)\mathrm{d}x$. Then $\beta = T_\sharp \alpha$ has density*

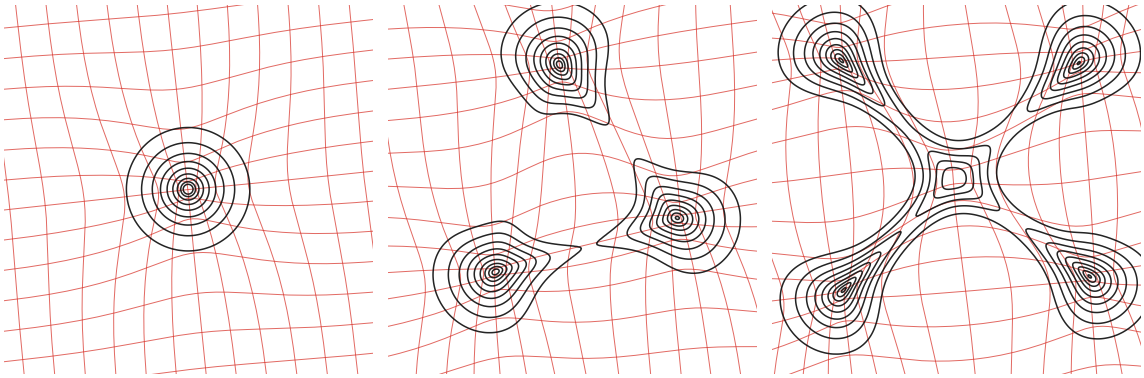
$$\rho_\beta(y) = \rho_\alpha(T^{-1}(y)) \left| \det \nabla T(T^{-1}(y)) \right|^{-1}. \tag{1.7}$$

Proof. Apply the change-of-variables formula to (1.6). □

Figure 1.1 makes the determinant factor visible. This is the first geometric warning behind transport PDEs: moving particles changes both their positions and the local volume element, and several local compressions create several density bumps.

1.2. Monge Problems

Monge's formulation asks for a deterministic map transporting one law onto another. It is geometrically direct and is the natural language for generative maps, but it is not convex and can be infeasible because maps cannot split mass.



one compression bump three compression bumps five compression bumps

Figure 1.1: Jacobian determinant in the density push-forward formula. The three panels show smooth diffeomorphisms that strongly compress a uniform source grid around one, three, and five centers. The deformed grid is drawn as a dense red mesh, and the pushed density is shown only through black level sets of $|\det \nabla T|^{-1}$. This is the geometric content of the Jacobian factor in (1.7), and later becomes the divergence structure of continuity equations.

1.2.1. General Measures

Given $\alpha \in \mathcal{P}(\mathcal{X})$, $\beta \in \mathcal{P}(\mathcal{Y})$ and a cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the Monge problem is

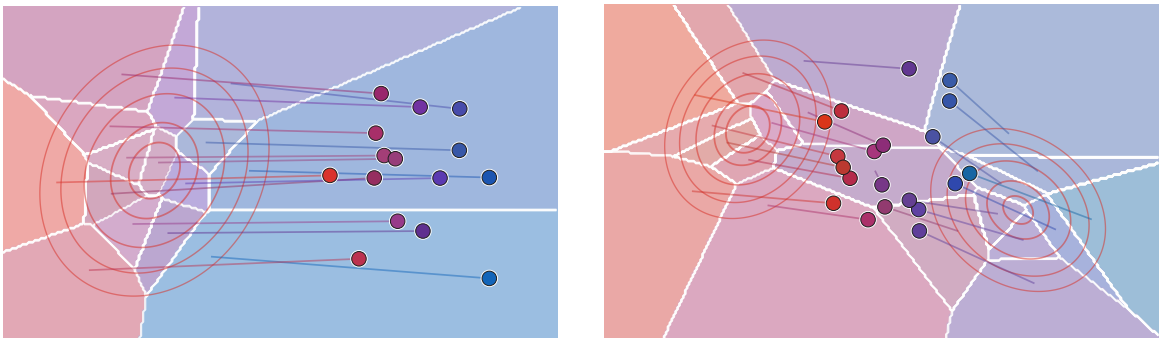
$$\mathcal{M}_c(\alpha, \beta) \coloneqq \inf_T \left\{ \int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\alpha(x) \;;\; T_\sharp \alpha = \beta \right\}. \tag{1.8}$$

The constraint $T_\sharp \alpha = \beta$ means that the endpoint law is imposed exactly. If $\alpha = \delta_x$ and β is not a Dirac mass, the feasible set is empty; this is the simplest mass-splitting obstruction.

Proposition 1.4 (Existence of transport maps from atomless sources)**.** *Let α and β be Borel probability measures on Polish spaces, and assume that α is atomless. Then there exists a measurable map T such that $T_\sharp \alpha = \beta$.*

Proof. Use a standard measure-isomorphism theorem to identify the atomless probability space generated by α with Lebesgue measure on $[0, 1]$, modulo null sets. A generalized quantile map sends Lebesgue measure on $[0, 1]$ to β after a Borel parametrization of the target space. Composing both maps gives a transport map. □

A useful asymmetric case is semi-discrete transport: the source has a density, the target is finitely supported, and a Monge map partitions space into cells. Figure 1.2 gives the cell picture before the fully discrete assignment problem. It also explains why reversing source and target would usually be impossible: finitely many source atoms cannot be mapped deterministically onto a continuous density.



Gaussian source to 15 atoms two-mode source to 20 atoms

Figure 1.2: Semi-discrete Monge maps. The red contours show the continuous source density. The colored regions extend numerical Laguerre cells over the whole displayed domain, with approximately equal source mass, and the circular atoms form the discrete target. Faint segments connect cell barycenters to their images under the piecewise-constant Monge map, making the map-as-partition viewpoint explicit.

1.2.2. Discrete Monge Problem

For equal-size empirical measures

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \qquad \beta = \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j},$$

and distinct target points, a feasible Monge map is a permutation. The discrete problem is therefore the assignment problem

$$\min_{\sigma\in\mathrm{Perm}(n)} \frac{1}{n}\sum_{i=1}^{n} C_{i,\sigma(i)}, \qquad C_{ij} = c(x_i, y_j). \tag{1.9}$$

Proposition 1.5 (Empirical Monge maps and matchings)**.** *Assume that the source atoms x_i are distinct and that $\alpha = n^{-1}\sum_i \delta_{x_i}$, $\beta = n^{-1}\sum_j \delta_{y_j}$. If the y_j are distinct and $T_\sharp\alpha = \beta$, then there exists $\sigma \in \mathrm{Perm}(n)$ such that $T(x_i) = y_{\sigma(i)}$ for all i.*

Proof. For any target atom z, the identity $T_\sharp\alpha = \beta$ gives

$$\beta(\{z\}) = \alpha(T^{-1}(\{z\})) = \frac{1}{n}\#\{i : T(x_i) = z\}.$$

Each target atom has mass $1/n$, so it receives exactly one source atom. □

In one dimension, convex costs have a special monotonicity structure: after sorting the source and target points, the equal-rank matching is optimal. This order structure is exceptional and does not survive in higher-dimensional point clouds.

The next two figures isolate the role of order on the line. Figure 1.3 shows equal-rank matching for convex costs, while Figure 1.4 contrasts it with a concave-cost regime where crossings become favorable.

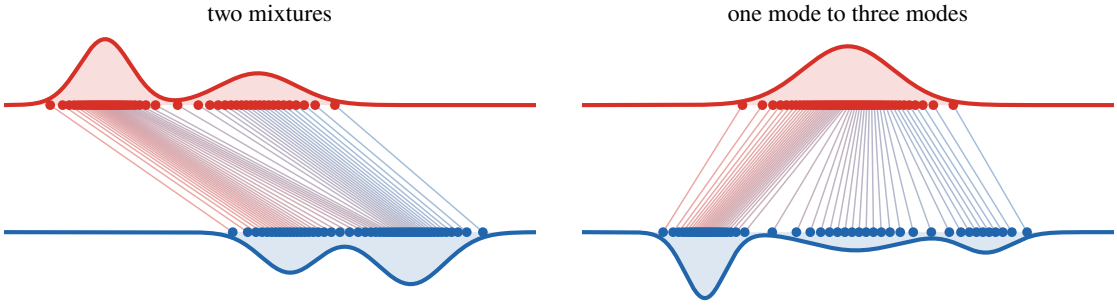


Figure 1.3: One-dimensional optimal matching by quantile sorting. The red and blue curves are smooth laws used to generate equal-weight empirical measures; the dots are inverse-CDF samples at common quantile levels. For convex costs on the line, the monotone assignment connects equal ranks.

The line also has a useful periodic variant. On the circle, one first chooses a cut, unfolds both measures into an interval and then applies the monotone one-dimensional assignment. Optimizing over the cut restores the circular geometry; see Figure 1.5. This is still an ordered one-dimensional construction, but the order is cyclic rather than linear.

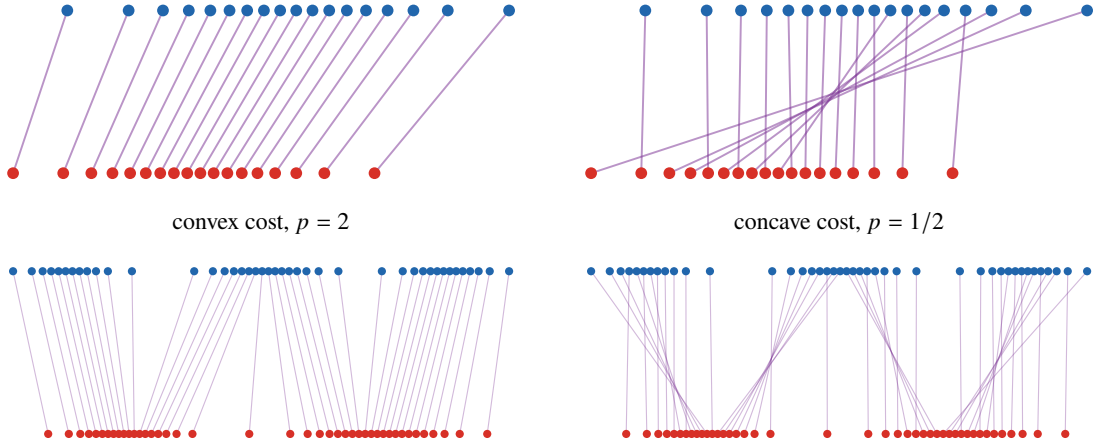
In two dimensions there is no global sorting rule. Figure 1.6 keeps the same two point clouds and changes only the exponent in the cost, showing that the optimal permutation is already a genuinely geometric object.

The permutation viewpoint also assumes equal cardinalities and equal weights. Figure 1.7 previews the Kantorovich matrix formulation: as soon as masses or resolutions differ, transport must allow splitting and merging.

Rational weights give a precise bridge between assignments and transport matrices. If $a_i = k_i/N$ and $b_j = \ell_j/N$, one may duplicate x_i exactly k_i times and y_j exactly ℓ_j times, solve a uniform assignment with N particles on each side, and collapse the result back to the original supports. Figure 1.8 shows how repeated duplicated matches appear as several transport segments attached to the same displayed atom.

1.3. Kantorovich Problems

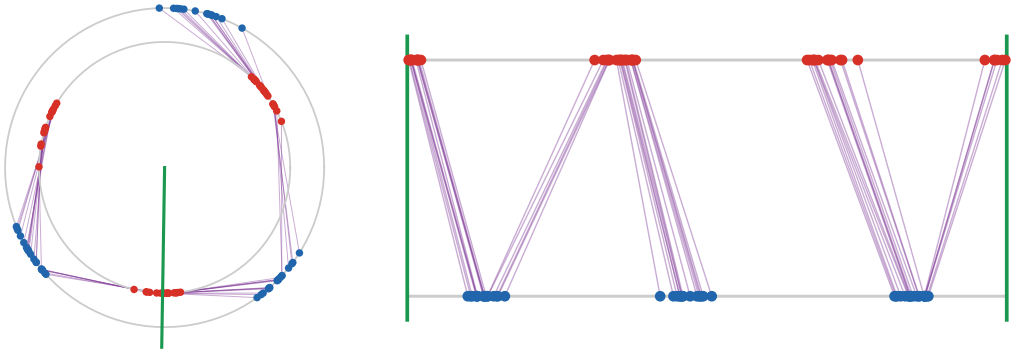
Kantorovich's relaxation replaces deterministic maps by joint probability measures. It is the convex formulation used throughout the rest of the review, and it is the object whose dynamic counterpart becomes the Benamou–Brenier PDE.



convex cost, $p = 2$ concave cost, $p = 1/2$

two-mode to three-mode, $p = 2$ two-mode to three-mode, $p = 1/2$

Figure 1.4: One-dimensional assignments for costs $c_p(x, y) = |x - y|^p$. For the convex quadratic cost, equal ranks are matched and the segments do not cross. For the concave cost, the optimum creates long crossing exchanges. Thus monotone rearrangement is a convex-cost phenomenon, not a generic property of all transport costs.



circular matching unfolded interval

Figure 1.5: Optimal transport on the circle by cutting and unfolding. Red and blue atoms live on two copies of the circle, purple segments show the optimal cyclic matching, and the green radius marks the chosen cut. Once the circle is opened at this angle, the same assignment becomes a monotone one-dimensional matching on the interval.

1.3.1. General Measures

Definition 1.6 (Marginals of a joint measure)**.** Let $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and let $P_{\mathcal{X}}(x, y) = x$, $P_{\mathcal{Y}}(x, y) = y$. Its marginals are $(P_{\mathcal{X}})_\sharp \pi$ and $(P_{\mathcal{Y}})_\sharp \pi$.

Definition 1.7 (Couplings)**.** The set of couplings between $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{Y})$ is

$$\mathcal{U}(\alpha, \beta) := \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \; ; \; (P_{\mathcal{X}})_\sharp \pi = \alpha, \; (P_{\mathcal{Y}})_\sharp \pi = \beta\}. \tag{1.10}$$

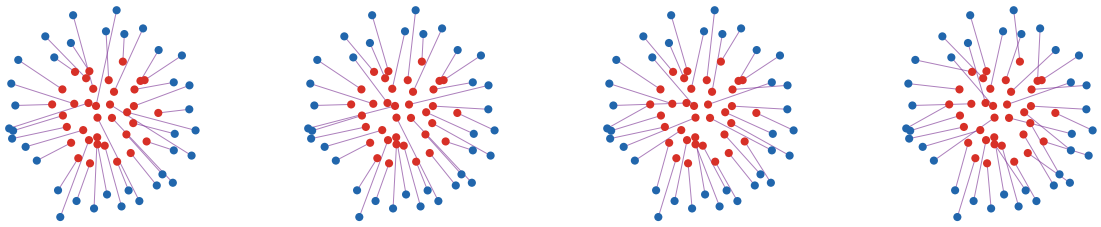
The set is never empty, since it contains the tensor product $\alpha \otimes \beta$. The Kantorovich problem is

$$\mathcal{L}_c(\alpha, \beta) := \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\pi(x, y). \tag{1.11}$$

If $T_\sharp \alpha = \beta$, then $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is a coupling, so Kantorovich is a relaxation of Monge.

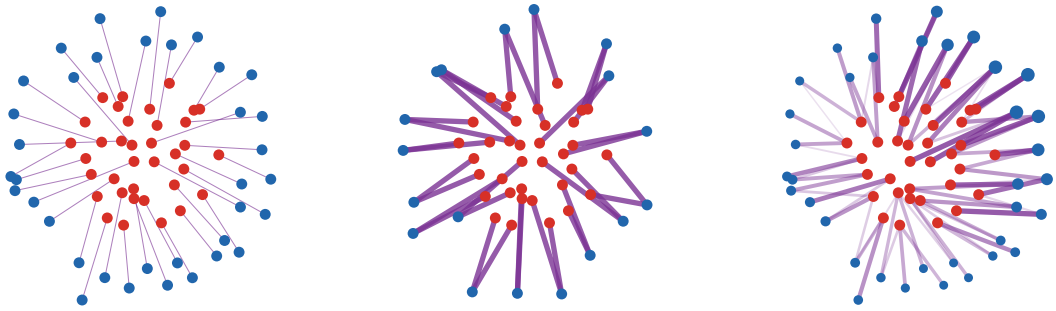
Figure 1.9 gives a first visual dictionary for couplings: a graph coupling is deterministic, the tensor product is independent and diffuse, and an optimal relaxed plan selects the dependence that lowers the cost. This is the finite-dimensional picture to keep in mind when the same symbol π later denotes a dynamic plan or a law on paths.

Proposition 1.8 (Existence on compact spaces)**.** *If $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces and c is continuous, then the infimum in (1.11) is attained.*



$\|x-y\|^{1/2}$ $\|x-y\|$ $\|x-y\|^2$ $\|x-y\|^6$

Figure 1.6: Optimal assignments between the same two planar point clouds for four powers of the Euclidean distance. The feasible permutations are unchanged, but the selected matching changes with the cost geometry: small powers tolerate long exchanges, whereas larger powers suppress the longest edges.



$n = m$ $m < n$ nonuniform target weights

Figure 1.7: From assignments to transport plans, using the same disk-to-annulus geometry as Figure 1.6. In the balanced equal-weight case, each source atom is matched to one target atom. With fewer target atoms, or with strongly nonuniform target weights, the coupling must merge or split mass; segment thickness and opacity encode transported mass.

Proof. The set $\mathcal{U}(\alpha,\beta)$ is tight and closed for weak convergence, hence compact. The functional $\pi \mapsto \int c\,\mathrm{d}\pi$ is continuous on this compact set. □

Dual potentials. The dual formulation exposes the scalar potentials that later reappear in Hamilton–Jacobi equations, entropic transport and adversarial objectives. Under the preceding compactness assumptions,

$$\mathcal{L}_c(\alpha,\beta) = \sup_{f,g}\left\{\int_{\mathcal{X}} f\,\mathrm{d}\alpha + \int_{\mathcal{Y}} g\,\mathrm{d}\beta \;:\; f(x)+g(y)\leqslant c(x,y)\right\}, \tag{1.12}$$

where the supremum may be taken over continuous potentials. Weak duality follows by integrating the pointwise constraint against any coupling. Strong duality is the infinite-dimensional linear-programming duality theorem for transport Kantorovich (1942); Villani (2009). At optimality, complementary slackness gives $f(x)+g(y)=c(x,y)$ on the support of an optimal plan.

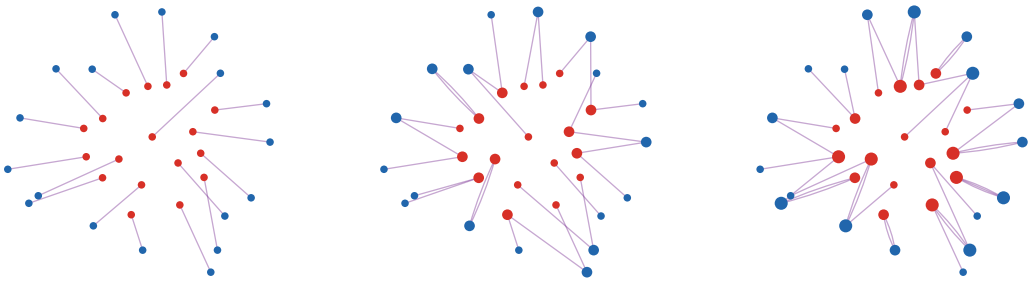
1.3.2. Discrete Kantorovich Problem

For discrete measures $\alpha=\sum_i a_i\delta_{x_i}$ and $\beta=\sum_j b_j\delta_{y_j}$, a coupling is a nonnegative matrix with prescribed row and column sums.

Definition 1.9 (Discrete couplings and mass conservation)**.** For weights $\mathrm{a}\in\Sigma_n$ and $\mathrm{b}\in\Sigma_m$,

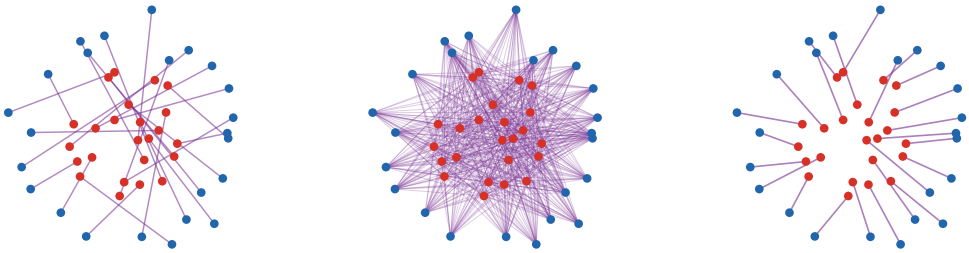
$$\mathrm{U}(\mathrm{a},\mathrm{b}) \coloneqq \left\{\mathrm{P}\in\mathbb{R}_+^{n\times m}\;;\;\mathrm{P}\mathbb{1}_m=\mathrm{a},\;\mathrm{P}^\top\mathbb{1}_n=\mathrm{b}\right\}. \tag{1.13}$$

Definition 1.10 (Discrete product coupling)**.** The product, or trivial, coupling is $(\mathrm{a}\otimes\mathrm{b})_{ij}=a_ib_j$.



uniform multiplicities binary multiplicities ternary multiplicities

Figure 1.8: Rational weights as duplicated uniform matchings. The red and blue locations are displayed once, while disk areas encode integer multiplicities. After duplication, the uniform assignment may contain several matched copies of the same displayed point; collapsing these copies gives the integer count matrix underlying a rational-weight Kantorovich plan.



deterministic graph product plan $\alpha \otimes \beta$ optimal plan

Figure 1.9: Discrete couplings represented as straight transport segments. A deterministic graph coupling sends each source to one destination, the product plan spreads every source mass over all targets, and the optimal Kantorovich plan minimizes the quadratic cost while being allowed to split mass. Line width and opacity encode the transported mass.

Given $C_{ij} = c(x_i, y_j)$, the discrete Kantorovich value is the linear program

$$\mathrm{L_C}(\mathrm{a}, \mathrm{b}) \coloneqq \min_{\mathrm{P} \in \mathrm{U}(\mathrm{a},\mathrm{b})} \sum_{i,j} C_{ij} P_{ij}. \tag{1.14}$$

When $m = n$ and $\mathrm{a} = \mathrm{b} = \mathbb{1}_n/n$, the Birkhoff–von Neumann theorem implies that this relaxation admits an optimal permutation plan, so it is tight for the balanced assignment problem. For nonuniform weights or unequal cardinalities, mass splitting is essential.

The Birkhoff theorem is also a useful mental model for linear programming over couplings: a bistochastic matrix can be moved along cycles in its positive-support bipartite graph until a permutation component is exposed. Figure 1.10 displays this cycle mechanism in the finite case.

Two complementary pictures of the discrete relaxation are useful. Figure 1.11 keeps the physical segments visible, while Figure 1.12 shows the same idea as a matrix with prescribed row and column sums.

One-dimensional transport by quantiles. The line is the exceptional setting where the optimal coupling is explicit without any density assumption. For $\alpha \in \mathcal{P}(\mathbb{R})$, let

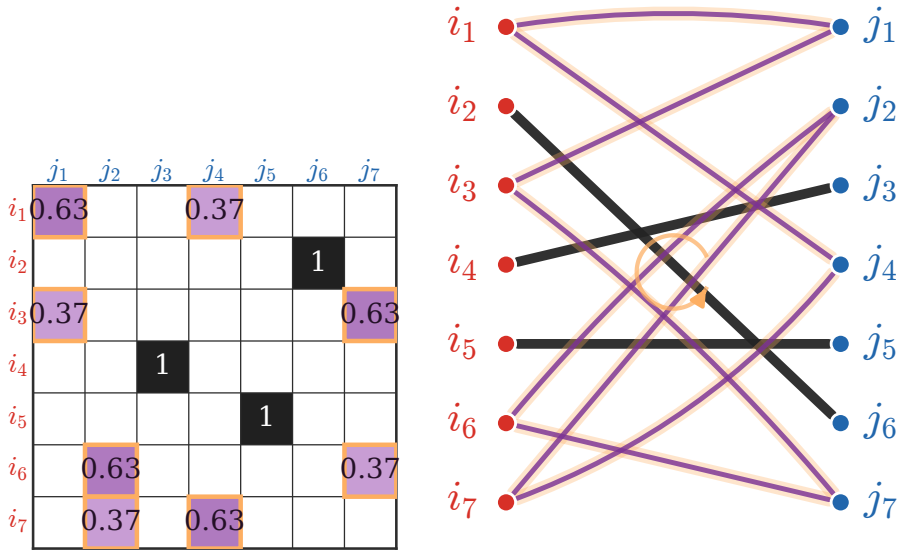
$$F_\alpha(x) = \alpha((-\infty, x]), \qquad Q_\alpha(s) = \inf\{x : F_\alpha(x) \geqslant s\}, \quad s \in (0,1).$$

For every $p \geqslant 1$ and $\alpha, \beta \in \mathcal{P}_p(\mathbb{R})$, the monotone coupling

$$\pi = (Q_\alpha, Q_\beta)_\sharp \mathcal{U}_{(0,1)} \tag{1.15}$$

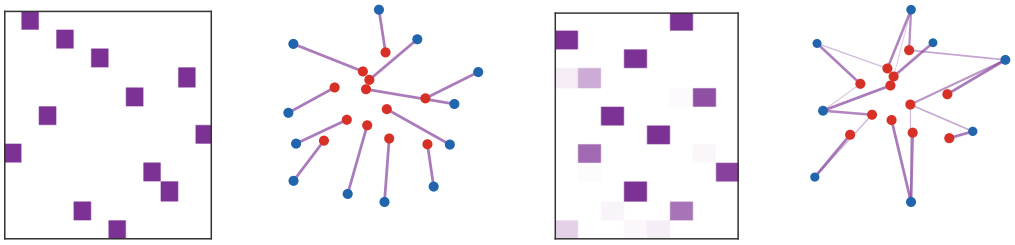
is optimal for $|x-y|^p$, and therefore

$$\mathcal{W}_p^p(\alpha, \beta) = \int_0^1 |Q_\alpha(s) - Q_\beta(s)|^p \, \mathrm{d}s. \tag{1.16}$$



bistochastic matrix support graph and cycle

Figure 1.10: Cycle certificate in the Birkhoff–von Neumann proof. The matrix is bistochastic but not a permutation matrix. Its positive support defines a bipartite graph; a fractional cycle permits an alternating add-subtract perturbation that preserves row and column sums and pushes at least one entry to zero. Repeating this exposes permutation matrices as extreme points.



permutation matrix assignment segments splitting matrix splitting segments

Figure 1.11: From permutation matrices to splitting couplings. When two empirical measures have the same number of atoms and uniform weights, an optimal plan can be a permutation matrix. Once target masses are nonuniform, admissible couplings may split one source among several targets or merge several sources into one target.

If α has no atoms, the coupling is induced by the monotone map $T = Q_\beta \circ F_\alpha$. The proof is the uncrossing argument: convexity of $r \mapsto |r|^p$ shows that swapping two crossed pairs cannot increase the cost, and approximation by empirical measures yields the general statement. Thus one-dimensional Wasserstein geometry is linearized by quantile functions, a fact used below for geodesics, scalar gradient flows and numerical illustrations.

1.4. Brenier Maps and Quadratic OT

The quadratic Euclidean cost is the central case for Wasserstein PDEs. Brenier's theorem says that the optimal coupling is induced by a gradient of a convex potential whenever the source is sufficiently diffuse.

Theorem 1.11 (Brenier)**.** *Let $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ and assume that α is absolutely continuous with respect to Lebesgue measure. For the cost $c(x, y) = \|x - y\|^2$, there exists a convex function φ such that*

$$T = \nabla\varphi, \qquad T_\sharp \alpha = \beta,$$

and T is the unique optimal Monge map α-almost everywhere. The associated optimal Kantorovich plan is unique and equals $(\mathrm{Id}, T)_\sharp \alpha$.

Thus quadratic OT is not merely a linear program over couplings; under absolute continuity it is a nonlinear elliptic problem for a convex potential. When $\alpha = \rho_\alpha \mathrm{d}x$ and $\beta = \rho_\beta \mathrm{d}y$ are smooth positive densities and $T = \nabla\varphi$

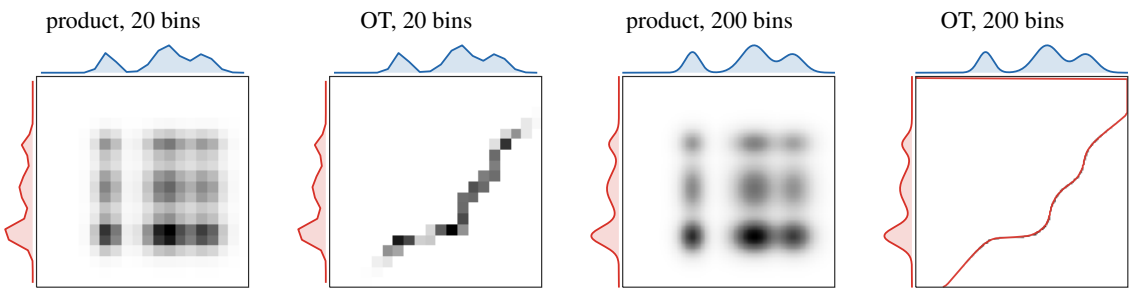


Figure 1.12: Coupling matrices with their prescribed marginals. The central grayscale image displays P_{ij}; the red curve on the left is the source marginal a, and the blue curve on top is the target marginal b. The independent product plan is diffuse, whereas the one-dimensional optimal plan concentrates near the monotone quantile correspondence.

is smooth, the change-of-variables formula gives the Monge–Ampère equation

$$\rho_\alpha(x) = \rho_\beta(\nabla\varphi(x)) \det(\nabla^2\varphi(x)). \tag{1.17}$$

Figure 1.13 should be read as a geometric stress test rather than a regularity theorem. It displays the particle interpolation induced by a quadratic assignment from a convex disk to a connected nonconvex target, making visible why regularity assumptions on the target geometry matter.

Figure 1.13: Empirical quadratic OT interpolation from a disk to a connected nonconvex two-disk domain. A dense farthest-point sample of the disk is matched to a dense farthest-point sample of two disks connected by a thin rectangle. The panels display $(1-t)x_i + tT_N(x_i)$ for the empirical optimal assignment T_N; colors are inherited from the horizontal coordinate of x_i in the initial disk, so transported material regions can be tracked.

The Gaussian case is the main closed-form model reused later for Gaussian closures of flows.

Proposition 1.12 (Gaussian $\mathcal{W}_2$ formula and Bures covariance term)**.** *Let $\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$ with positive definite covariance matrices. Then the quadratic optimal map is affine,*

$$T(x) = m_\beta + A(x - m_\alpha), \qquad A = \Sigma_\alpha^{-1/2}(\Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2})^{1/2}\Sigma_\alpha^{-1/2},$$

and

$$\mathcal{W}_2^2(\alpha,\beta) = \|m_\alpha - m_\beta\|^2 + \operatorname{tr}\left(\Sigma_\alpha + \Sigma_\beta - 2(\Sigma_\alpha^{1/2}\Sigma_\beta\Sigma_\alpha^{1/2})^{1/2}\right). \tag{1.18}$$

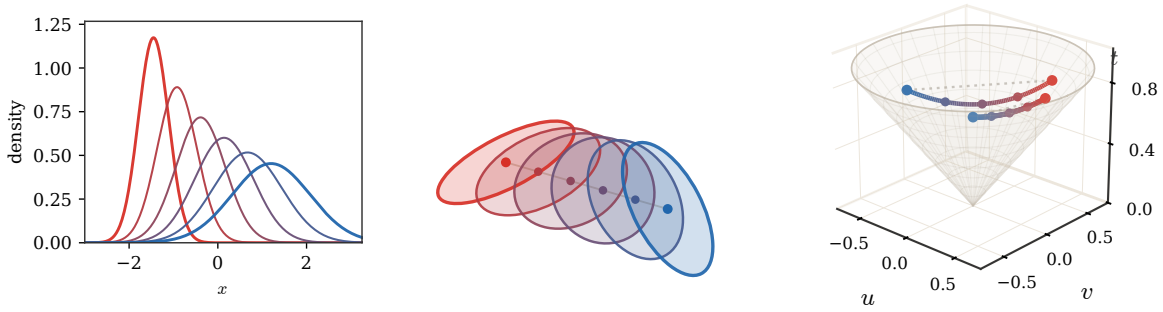
Proof. The affine map above sends the first Gaussian to the second and has symmetric positive definite linear part. It is therefore the gradient of a convex quadratic potential, so Brenier's theorem gives optimality. Substituting the affine displacement in the quadratic transport cost gives the formula. □

Figure 1.14 turns the closed form into three complementary pictures: densities in one dimension, covariance ellipses in two dimensions, and a curve inside the cone of positive semidefinite matrices.

1.5. Wasserstein Distances and Geodesics

This subsection gives the metric language used throughout the PDE-oriented parts.

Metric structure. The metric structure used in the PDE sections is obtained by taking the p-th root of the Kantorovich cost for a ground metric.



one-dimensional densities covariance ellipses covariance cone

Figure 1.14: Gaussian $\mathcal{W}_2$ geodesics. In one dimension, the mean and standard deviation interpolate linearly. In two dimensions, the affine Brenier map transports covariance ellipses along the Bures covariance geometry appearing in (1.18). The cone panel represents 2×2 positive semidefinite covariances and shows that covariance interpolation is not generally Euclidean in matrix entries.

Definition 1.13 (Wasserstein distance)**.** Let $(\mathcal{X}, d)$ be a metric space and $p \geqslant 1$. For $\alpha, \beta \in \mathcal{P}_p(\mathcal{X})$,

$$\mathcal{W}_p(\alpha, \beta) := \left(\inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, \mathrm{d}\pi(x,y) \right)^{1/p}. \tag{1.19}$$

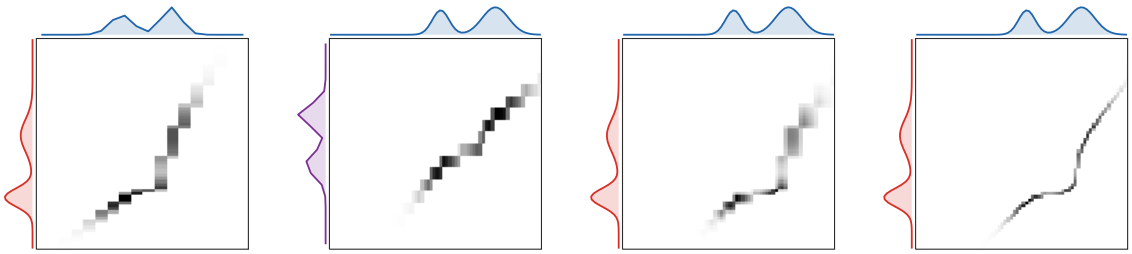
Proposition 1.14 (Metric property of the Wasserstein distance)**.** *If $(\mathcal{X}, d)$ is a metric space, then $\mathcal{W}_p$ is a distance on $\mathcal{P}_p(\mathcal{X})$.*

Proof. Nonnegativity and symmetry are immediate. If $\mathcal{W}_p(\alpha, \beta) = 0$, the optimal mass can only lie on the diagonal, hence $\alpha = \beta$. The triangle inequality follows from the gluing lemma: glue an optimal coupling between α and β with one between β and γ to obtain a joint law of (X, Y, Z); then apply Minkowski's inequality to $d(X, Z) \leqslant d(X, Y) + d(Y, Z)$. □

The gluing step is abstract but elementary in the discrete case. If $P \in \mathrm{U}(\mathrm{a}, \mathrm{b})$ and $Q \in \mathrm{U}(\mathrm{b}, \mathrm{c})$, then

$$R = P \, \mathrm{diag}(\mathrm{b})^{-1} Q$$

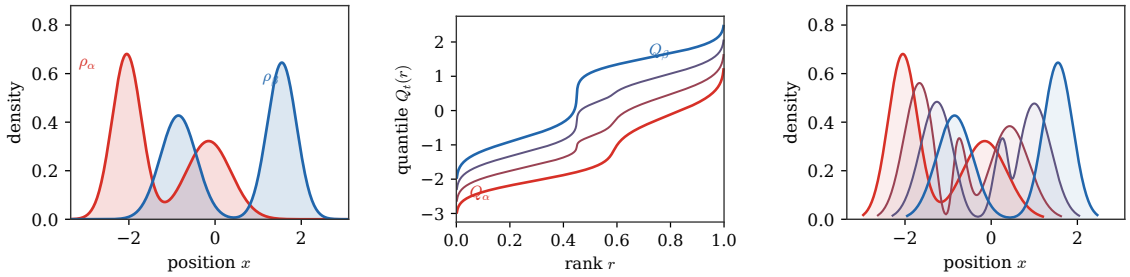
is a feasible coupling between a and c whenever all entries of b are positive. Figure 1.15 visualizes this composition of plans through an intermediate marginal.



a to b b to c glued a to c direct a to c

Figure 1.15: Discrete gluing lemma in matrix form. The first two coupling matrices share the intermediate marginal b. Multiplying through this common marginal gives a feasible glued plan between a and c, which is the coupling used in the triangle-inequality proof. The last panel shows the direct optimal coupling for comparison.

In one dimension, the geodesic can be read directly in quantile coordinates. Figure 1.16 is the static prototype for the Lagrangian views of one-dimensional PDEs used later in the dynamic and gradient-flow sections.



endpoint densities quantile interpolation geodesic densities

Figure 1.16: One-dimensional Wasserstein geodesic by quantiles. The optimal coupling matches equal quantile levels, so the interpolation is linear in inverse-CDF coordinates: $Q_t = (1-t)Q_0 + tQ_1$. The right panel shows the corresponding density path.

Interpolation induced by an optimal plan. For $\mathcal{X} = \mathbb{R}^d$ and $p = 2$, an optimal plan also defines a constant-speed geodesic. If π is optimal between α_0 and α_1, set

Definition 1.15 ($\mathcal{W}_2$ geodesic induced by an optimal plan)**.**

$$\alpha_t = ((1-t)P_0 + tP_1)_\sharp \pi, \qquad P_0(x,y) = x, \quad P_1(x,y) = y. \tag{1.20}$$

When $\pi = (\mathrm{Id}, T)_\sharp \alpha_0$ is induced by a Brenier map, this reduces to McCann's displacement interpolation

$$\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha_0.$$

Proposition 1.16 (Optimal-plan interpolation is a $\mathcal{W}_2$ geodesic)**.** *Let π be an optimal coupling between $\alpha_0, \alpha_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and define α_t by (1.20). Then*

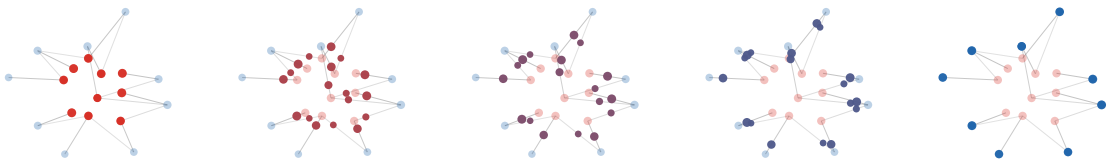
$$\mathcal{W}_2(\alpha_s, \alpha_t) = |t-s|\,\mathcal{W}_2(\alpha_0, \alpha_1), \qquad 0 \leqslant s, t \leqslant 1.$$

Proof. Coupling the particles at times s and t through the same endpoint pair $(x, y) \sim \pi$ gives

$$\mathcal{W}_2^2(\alpha_s, \alpha_t) \leqslant (t-s)^2 \int \|x-y\|^2 \,\mathrm{d}\pi = (t-s)^2\, \mathcal{W}_2^2(\alpha_0, \alpha_1).$$

The reverse inequality follows from the triangle inequality applied to the three pieces $[0, s]$, $[s, t]$ and $[t, 1]$. □

The last two primer figures should be read together. Figure 1.17 illustrates the definition when the plan is a genuine coupling rather than a single-valued map: each active pair (x, y) follows a straight segment. Figure 1.18 then shows the Monge/Brenier special case on two silhouettes.



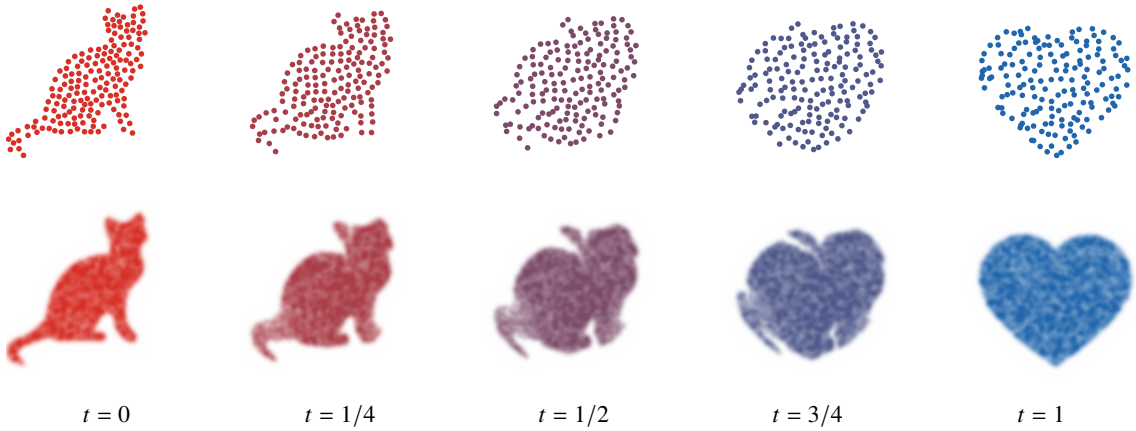
$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

Figure 1.17: Interpolation induced by a Kantorovich plan. Each nonzero entry of an optimal discrete coupling transports mass along the segment $(1-t)x_i + ty_j$. This construction is the static plan-level ancestor of the dynamic Benamou–Brenier viewpoint developed in Section 2.

This geodesic viewpoint is the static counterpart of the Benamou–Brenier formulation in Section 2. It is also the geometry behind the Wasserstein gradient flows of Section 3.

1.6. Functionals and Discrepancies on Measures

The metric specifies how a law is allowed to move; a functional specifies what the evolution tries to decrease. The following compact dictionary fixes the objects used repeatedly in the gradient-flow and generative-model sections.



$t = 0$ $t = 1/4$ $t = 1/2$ $t = 3/4$ $t = 1$

Figure 1.18: McCann displacement interpolation between two silhouette measures. The first row displays transported particles along $T_t(x) = (1-t)x + tT(x)$; the second row renders kernel-smoothed densities from a denser transported cloud. This is the map-induced version of the plan interpolation above and the geometric path used later by Wasserstein gradient-flow arguments.

First variations and scores. Let $f : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$. A first variation $\delta f(\alpha)$ is a function, defined up to an additive constant, such that for signed perturbations η of zero total mass,

$$f(\alpha + \varepsilon\eta) = f(\alpha) + \varepsilon \int \delta f(\alpha)(x)\, \mathrm{d}\eta(x) + o(\varepsilon). \tag{1.21}$$

When $\alpha = \rho\, \mathrm{d}x$ has a positive differentiable density, its score is $\nabla \log \rho$. The spatial gradient $\nabla \delta f(\alpha)$, rather than the additive normalization of $\delta f(\alpha)$, determines the Wasserstein velocity in Section 3.

Relative entropy and kernel discrepancies. If $\alpha \ll \beta$, the relative entropy is

$$\mathrm{KL}(\alpha|\beta) = \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\alpha, \tag{1.22}$$

and it is $+\infty$ otherwise. For smooth positive densities, its first variation is $\log(\rho_\alpha/\rho_\beta)$ up to a constant. If k is a positive-definite kernel with reproducing-kernel Hilbert space $\mathcal{H}_k$, the maximum mean discrepancy is

$$\mathrm{MMD}_k^2(\alpha,\beta) = \left\| \int k(x,\cdot)\, \mathrm{d}(\alpha-\beta)(x) \right\|_{\mathcal{H}_k}^2 = \iint k(x,y)\, \mathrm{d}(\alpha-\beta)(x) \mathrm{d}(\alpha-\beta)(y). \tag{1.23}$$

Its first variation with respect to α is $2\int k(x,y)\, \mathrm{d}(\alpha-\beta)(y)$. Entropy produces local score-driven diffusion, whereas MMD produces a nonlocal interaction field; this contrast is central in the particle examples below.

Sliced Wasserstein discrepancy. For $\theta \in \mathbb{S}^{d-1}$, set $P_\theta(x) = \langle \theta, x \rangle$ and let σ be the uniform probability measure on the sphere. The sliced quadratic Wasserstein distance is

$$\mathrm{SW}_2^2(\alpha,\beta) = \int_{\mathbb{S}^{d-1}} \mathcal{W}_2^2\big((P_\theta)_\sharp \alpha, (P_\theta)_\sharp \beta\big) \mathrm{d}\sigma(\theta). \tag{1.24}$$

Each integrand is computed by the one-dimensional quantile formula (1.16). This makes SW_2 computationally attractive, while its Wasserstein gradient still defines a genuinely multidimensional flow through the average of projected one-dimensional velocities.

2. Dynamic Optimal Transport

Optimal transport becomes especially useful for machine learning once distances between probability laws are written as actions of moving mass. This section is synchronized with the full OT4ML manuscript and keeps the material needed later in the survey: continuity equations, the Benamou–Brenier formula, path-space viewpoints, and generalized dynamic distances that later generate PDE flows.

2.1. Evolutions over the Space of Measures

We start with the continuity equation because it is the common language for particles, densities and weak measure evolutions. It also makes precise which velocity fields actually move mass.

Lagrangian and Eulerian descriptions. We consider the evolution $t \mapsto \alpha_t \in \mathcal{P}(\mathbb{R}^d)$. Such an evolution can be described in a "Lagrangian" way as the advection of particles along a (time-dependent) vector field $v_t(x)$ in $\mathbb{R}^d$. At the particle level, this advection is governed by

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = v_t(x(t)), \tag{2.1}$$

and we write T_t for the associated flow map, so that $T_t(x(0)) = x(t)$. The advected measure is then $\alpha_t = (T_t)_\sharp \alpha_0$. For discrete measures, $\alpha_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$, meaning each $x_i(t)$ solves (2.1).

In the Eulerian description, the same motion is written directly on the evolving measure: the particle ODE becomes the PDE

$$\frac{\partial \alpha_t}{\partial t} + \mathrm{div}(v_t \alpha_t) = 0. \tag{2.2}$$

This PDE is often referred to as the advection equation, the continuity equation, or Liouville's equation when operating over a phase space. It is only a classical PDE when α_t has a smooth density. For general measures, and in particular for empirical measures, it is understood in the weak sense: for any smooth test function $(t,x) \mapsto \varphi(t,x)$ compactly supported in time,

$$\int_0^1 \int_{\mathbb{R}^d} (\partial_t \varphi(t,x) + \langle v_t(x),\, \nabla_x \varphi(t,x)\rangle)\, \mathrm{d}\alpha_t(x)\mathrm{d}t = 0. \tag{2.3}$$

This equation is obtained from (2.2) by integration by parts. Hence, for smooth positive densities, the classical and weak formulations are equivalent; the weak formulation is useful because it still makes sense for discrete measures whose particles evolve according to (2.1).

Proposition 2.1 (Lagrangian flows solve the continuity equation)**.** *Consider a smooth flow $T_t : \mathbb{R}^d \to \mathbb{R}^d$ and define $\alpha_t = (T_t)_\sharp \alpha_0$. Define the Eulerian velocity field by*

$$v_t(T_t(y)) = \partial_t T_t(y).$$

Then (α_t, v_t) solves the continuity equation in the weak sense (2.3). In particular, if $\alpha_0 = \frac{1}{n}\sum_i \delta_{x_i(0)}$ is empirical, then $\alpha_t = \frac{1}{n}\sum_i \delta_{x_i(t)}$ is empirical as well, with particle velocities $\dot{x}_i(t) = v_t(x_i(t))$.

Proof. Let $\varphi(t,x)$ be a smooth test function vanishing at $t = 0$ and $t = 1$. Since $\alpha_t = (T_t)_\sharp \alpha_0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(t,x)\mathrm{d}\alpha_t(x) = \frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(t,T_t(y))\mathrm{d}\alpha_0(y).$$

The chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(t,T_t(y))\mathrm{d}\alpha_0(y) = \int (\partial_t\varphi(t,T_t(y)) + \langle \nabla_x\varphi(t,T_t(y)),\, \partial_t T_t(y)\rangle)\, \mathrm{d}\alpha_0(y).$$

Using the definition of v_t and the push-forward relation, this equals

$$\int (\partial_t\varphi(t,x) + \langle \nabla_x\varphi(t,x),\, v_t(x)\rangle)\, \mathrm{d}\alpha_t(x).$$

Integrating in time and using the boundary values of φ gives (2.3). □

From measure evolutions to vector fields. For a given evolution $(\alpha_t)_t$, there are typically infinitely many velocity fields v_t satisfying

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t v_t) = 0. \tag{2.4}$$

This non-uniqueness comes from the kernel of the weighted divergence. The linear space of vector fields that leave a measure α invariant is

$$\mathcal{H}_\alpha = \{v \;;\; \mathrm{div}(\alpha v) = 0\}.$$

It is usually non-trivial: for instance, if α is an isotropic Gaussian, $\mathcal{H}_\alpha$ contains rotational vector fields generated by anti-symmetric matrices.

Dacorogna–Moser inversion. Reconstructing particles from an observed density evolution is therefore ill-posed. A simple choice, introduced by Dacorogna and Moser Dacorogna and Moser (1990), is to impose that the flux $\alpha_t v_t$ is a gradient field, leading formally to

$$v_t = -\frac{1}{\alpha_t}\nabla\Delta^{-1}(\partial_t\alpha_t), \tag{2.5}$$

with suitable boundary conditions, for instance vanishing at infinity. This formula is useful conceptually but delicate when α_t vanishes, and it does not generally produce a gradient velocity field.

The classical Dacorogna–Moser construction uses the linear density path. If $\alpha_i = \rho_i\,\mathrm{d}x$ are smooth positive densities with the same total mass on a bounded domain Ω, set

$$\alpha_t = (1-t)\alpha_0 + t\alpha_1 = \rho_t\,\mathrm{d}x, \qquad \rho_t = (1-t)\rho_0 + t\rho_1.$$

Choose a time-independent flux w satisfying

$$\operatorname{div} w = \rho_0 - \rho_1, \qquad w\cdot n = 0 \quad \text{on } \partial\Omega,$$

for instance $w = -\nabla\varphi$ with $\Delta\varphi = \rho_1 - \rho_0$ and Neumann boundary condition. Then

$$v_t = \frac{w}{\rho_t}$$

satisfies $\partial_t\rho_t + \operatorname{div}(\rho_t v_t) = 0$. The flow $\partial_t T_t = v_t\circ T_t$, $T_0 = \mathrm{Id}$, therefore transports $\rho_0\mathrm{d}x$ onto $\rho_t\mathrm{d}x$, and T_1 solves the prescribed-Jacobian problem $\rho_1(T_1(x))\det(\nabla T_1(x)) = \rho_0(x)$.

Least-square inversion and gradient structure. A more robust choice, used implicitly in flow matching, optimal transport and Wasserstein gradient flows, is to select among all admissible velocities the one with smallest kinetic energy:

$$\min_v \frac{1}{2}\int_0^1\int_{\mathbb{R}^d}\|v_t(x)\|^2\,\mathrm{d}\alpha_t(x)\mathrm{d}t \quad \text{subject to} \quad \partial_t\alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{2.6}$$

Proposition 2.2 (Least-square velocities are gradients)**.** *Assume that $\alpha_t = \rho_t\,\mathrm{d}x$ is a smooth positive density curve and that boundary terms vanish. The minimizer of (2.6), if it exists, is a gradient field*

$$v_t = \nabla\varphi_t,$$

where φ_t, unique up to an additive constant, solves the weighted Poisson equation

$$-\operatorname{div}(\rho_t\nabla\varphi_t) = \partial_t\rho_t, \qquad v_t = -\nabla\Delta_{\alpha_t}^{-1}(\partial_t\alpha_t), \quad \Delta_{\alpha_t}\varphi = \operatorname{div}(\alpha_t\nabla\varphi). \tag{2.7}$$

Proof. Introduce a Lagrange multiplier φ_t for the continuity equation. The constrained problem has the formal saddle formulation

$$\min_v\max_\varphi \int_0^1\left[\frac{1}{2}\int_{\mathbb{R}^d}\|v_t(x)\|^2\,\mathrm{d}\alpha_t(x) + \int_{\mathbb{R}^d}\varphi_t(x)\big(\operatorname{div}(\alpha_t v_t)(x) + \partial_t\alpha_t(x)\big)\mathrm{d}x\right]\mathrm{d}t.$$

Integrating by parts in the divergence term gives, for each t,

$$\int\left(\frac{1}{2}\|v_t\|^2 - \langle\nabla\varphi_t,\, v_t\rangle\right)\mathrm{d}\alpha_t + \int\varphi_t\,\partial_t\alpha_t.$$

The pointwise minimizer in v_t is therefore $v_t = \nabla\varphi_t$. Substituting this into the constraint $\partial_t\rho_t + \operatorname{div}(\rho_t v_t) = 0$ gives the weighted Poisson equation in (2.7). The inverse notation is just a shorthand for solving this equation on zero-mean right-hand sides, modulo additive constants. □

In general this inversion is still computationally demanding, but special choices of $(\alpha_t)_t$ lead to simpler formulas; this is the mechanism exploited later by flow matching in Section 4.1.

2.2. Benamou–Brenier dynamic formulation of OT

The dynamic formulation identifies $\mathcal{W}_2$ with the kinetic energy of the cheapest continuity-equation path. It is the point where OT becomes a least-action principle.

Benamou–Brenier formulation. Instead of assuming that a whole curve $(\alpha_t)_{t\in[0,1]}$ is prescribed, one only fixes its endpoints α_0 and α_1 and minimizes the least-square energy (2.6). The theorem of Benamou and Brenier states that this geodesic energy is exactly the squared Wasserstein distance Benamou and Brenier (2000).

Theorem 2.3 (Benamou–Brenier). *For probability measures $\alpha_0, \alpha_1 \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathcal{W}_2^2(\alpha_0,\alpha_1) = \inf_{(\alpha_t,v_t)} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\alpha_t(x)\mathrm{d}t, \tag{2.8}$$

where the infimum is over (α_t, v_t) solving $\partial_t\alpha_t + \nabla\cdot(\alpha_t v_t) = 0$ with $\alpha_{t=0} = \alpha_0$ and $\alpha_{t=1} = \alpha_1$. If α_0 has a density and T is the optimal Monge map $T_\sharp\alpha_0 = \alpha_1$, the minimizer is

$$\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp\alpha_0, \qquad v_t\big((1-t)x + tT(x)\big) = T(x) - x. \tag{2.9}$$

Proof. For the inequality "dynamic $\leqslant$ static", assume first that a Monge map T exists and define (α_t, v_t) by (2.9). Since the Lagrangian velocity $T(x) - x$ is independent of t,

$$\int_0^1 \int \|v_t\|^2 \,\mathrm{d}\alpha_t\mathrm{d}t = \int \|T(x) - x\|^2 \,\mathrm{d}\alpha_0(x),$$

so the dynamic cost is no larger than the static Monge cost. Without a Monge map, the same construction is made with an optimal coupling π: sample $(X, Y) \sim \pi$ and move along the straight path $\gamma_{X,Y}(t) = (1-t)X + tY$. This path measure has action $\int \|x-y\|^2\mathrm{d}\pi(x,y)$; projecting path velocities onto their conditional mean at time t gives an admissible Eulerian velocity with no larger action, so the dynamic value is no larger than the Kantorovich value.

Conversely, for a smooth deterministic path, take the flow T_t defined by $\dot{T}_t = v_t \circ T_t$ and $T_0 = \mathrm{Id}$. Then $\alpha_t = (T_t)_\sharp\alpha_0$ and $(T_1)_\sharp\alpha_0 = \alpha_1$. Jensen's inequality gives

$$\|T_1(x) - x\|^2 \leqslant \int_0^1 \|v_t(T_t(x))\|^2\mathrm{d}t.$$

After integration with respect to α_0, the Monge cost is bounded above by the dynamic action. For general finite-energy solutions of the continuity equation, the superposition principle lifts the curve to a probability measure on absolutely continuous paths; applying Jensen's inequality pathwise gives a coupling of the endpoints whose quadratic cost is no larger than the action. Thus the Kantorovich value is bounded above by the dynamic value. □

Convex moment-based reformulation. Although (2.8) is not jointly convex in (α_t, v_t), it becomes convex after replacing velocities by the momentum measure $m_t = v_t\alpha_t$ and using the perspective action. In the absolutely continuous case $\alpha_t = \rho_t\,\mathrm{d}x$ and $m_t(x) = \rho_t(x)v_t(x)$, this reads

$$\mathcal{W}_2^2(\alpha_0,\alpha_1) = \inf_{\substack{\partial_t\rho_t + \mathrm{div}\, m_t = 0 \\ \rho_{t=0}\mathrm{d}x=\alpha_0,\ \rho_{t=1}\mathrm{d}x=\alpha_1}} \int_0^1 \int_{\mathbb{R}^d} \frac{\|m_t(x)\|^2}{\rho_t(x)}\,\mathrm{d}x\,\mathrm{d}t, \tag{2.10}$$

with the usual convention that the integrand is 0 when $(\rho_t, m_t) = (0,0)$ and $+\infty$ when $\rho_t = 0$ but $m_t \neq 0$. For singular endpoints or curves, the same statement is interpreted with vector-valued momentum measures and the corresponding recession convention. This convex reformulation enables geodesic interpolation by convex optimization once the domain is discretized.

Dual Hamilton–Jacobi formulation. The momentum formulation also has a useful dual. It turns the least-action problem into a Hamilton–Jacobi subsolution inequality for a scalar potential, with equality on the part of space-time actually visited by the optimal curve. This is the dual dynamic formulation already present in the original Benamou–Brenier perspective Benamou and Brenier (2000); see also the standard accounts Villani (2003); Santambrogio (2015). The constants below follow the convention of (2.10), where the kinetic energy has no factor $1/2$.

Proposition 2.4 (Dual Benamou–Brenier problem). *Assume, for simplicity, that the densities are smooth, compactly supported, and that boundary terms vanish. Then the convex dynamic value (2.10) has the dual formulation*

$$\mathcal{W}_2^2(\alpha_0,\alpha_1) = \sup_{\varphi} \left\{ \int_{\mathbb{R}^d} \varphi_1 \,\mathrm{d}\alpha_1 - \int_{\mathbb{R}^d} \varphi_0 \,\mathrm{d}\alpha_0 \ ;\ \partial_t\varphi_t + \frac{1}{4}\|\nabla\varphi_t\|^2 \leqslant 0 \right\}, \tag{2.11}$$

where the supremum is over smooth test potentials $\varphi : [0,1] \times \mathbb{R}^d \to \mathbb{R}$. *In the general metric-measure statement, the same formula is understood after relaxation:* φ *is a Hamilton–Jacobi subsolution, for instance in the viscosity sense, and finite-dimensional discretizations follow directly from Fenchel–Rockafellar duality.*

If (ρ, m) *and* φ *are smooth primal and dual optimizers, then*

$$m_t = \frac{\rho_t}{2}\nabla\varphi_t, \qquad \partial_t\varphi_t + \frac{1}{4}\|\nabla\varphi_t\|^2 = 0 \quad \text{on } \{\rho_t > 0\}. \tag{2.12}$$

Equivalently, the optimal Eulerian velocity is $v_t = m_t/\rho_t = \nabla\varphi_t/2$.

Proof. Let (ρ, m) satisfy the continuity equation and let φ be smooth. Multiplying the constraint by φ and integrating by parts gives

$$\int_{\mathbb{R}^d} \varphi_1 \,\mathrm{d}\alpha_1 - \int_{\mathbb{R}^d} \varphi_0 \,\mathrm{d}\alpha_0 = \int_0^1 \int_{\mathbb{R}^d} \left(\rho_t\, \partial_t\varphi_t + \langle m_t,\, \nabla\varphi_t\rangle\right) \mathrm{d}x\mathrm{d}t.$$

If $\partial_t\varphi_t + \|\nabla\varphi_t\|^2/4 \leqslant 0$, Young's inequality yields, pointwise,

$$\rho\, \partial_t\varphi + \langle m,\, \nabla\varphi\rangle \leqslant -\frac{\rho}{4}\|\nabla\varphi\|^2 + \langle m,\, \nabla\varphi\rangle \leqslant \frac{\|m\|^2}{\rho},$$

with the same perspective convention when $\rho = 0$. Hence the dual objective of every feasible φ is no larger than the action of every feasible primal pair.

Conversely, introduce φ as a Lagrange multiplier for $\partial_t\rho + \operatorname{div} m = 0$. After integration by parts, and up to the fixed endpoint contribution in (2.11), the pointwise minimization over (ρ, m) contains

$$\frac{\|m\|^2}{\rho} - \rho\, \partial_t\varphi - \langle m,\, \nabla\varphi\rangle.$$

For fixed $\rho > 0$, its minimum over m is attained at $m = \rho\nabla\varphi/2$ and equals $-\rho(\partial_t\varphi + \|\nabla\varphi\|^2/4)$. The infimum over $\rho \geqslant 0$ is finite exactly when the Hamilton–Jacobi inequality holds, in which case the pointwise infimum is 0. Fenchel–Rockafellar duality then gives no duality gap. Equality in the two pointwise inequalities gives (2.12). □

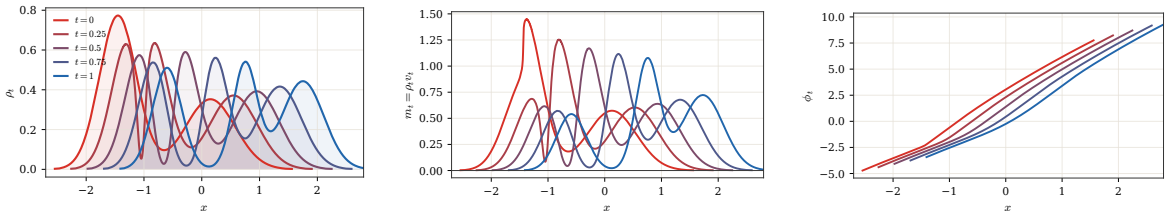
This also recovers the static Kantorovich inequality from a dynamic principle. If γ is any smooth curve with $\gamma(0) = x$ and $\gamma(1) = y$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi_t(\gamma(t)) = \partial_t\varphi_t(\gamma(t)) + \langle\nabla\varphi_t(\gamma(t)),\, \dot\gamma(t)\rangle \leqslant \|\dot\gamma(t)\|^2,$$

where the last inequality uses $\partial_t\varphi + \|\nabla\varphi\|^2/4 \leqslant 0$ and Young's inequality. Integrating and minimizing over curves gives

$$\varphi_1(y) - \varphi_0(x) \leqslant \|x - y\|^2.$$

Thus (φ_0, φ_1) is a feasible static Kantorovich dual pair for the quadratic cost. At optimality the inequality is saturated on the endpoint pairs connected by the primal characteristics, which is the Eulerian version of complementary slackness.



density ρ_t momentum m_t dual potential φ_t

Figure 2.1: One-dimensional Benamou–Brenier primal and dual solutions. The endpoints are Gaussian mixtures and the solution is computed from monotone quantile interpolation. The left and middle panels show the primal density and momentum $m_t = \rho_t v_t$. The right panel shows the dual Hamilton–Jacobi potential, normalized up to an additive constant. Along the active transported mass, the notebook checks the primal–dual identities $m_t = \rho_t\partial_x\varphi_t/2$ and $\partial_t\varphi_t + |\partial_x\varphi_t|^2/4 = 0$.

Proximal splitting. The convex momentum formulation also explains the original Benamou–Brenier solver. Papadakis, Peyré and Oudet Papadakis et al. (2014) showed that, after discretization, the ALG2 scheme is a Douglas–Rachford splitting, equivalently ADMM on the Fenchel–Rockafellar dual. Suppressing discretization indices, write $U = (\rho, m)$ and introduce the perspective integrand

$$J(\rho, m) = \begin{cases} \|m\|^2/\rho, & \rho > 0, \\ 0, & (\rho, m) = (0, 0), \\ +\infty, & \text{otherwise.} \end{cases}$$

Let

$$\mathcal{F}(U) = \int_0^1 \int_{\mathbb{R}^d} J(\rho_t(x), m_t(x))\,\mathrm{d}x\,\mathrm{d}t, \qquad C = \{(\rho, m)\ ;\ \partial_t \rho + \operatorname{div} m = 0,\ \rho_0 \mathrm{d}x = \alpha_0,\ \rho_1 \mathrm{d}x = \alpha_1\},$$

and let $\mathcal{G} = \iota_C$ be the indicator of this affine continuity constraint. The convex Benamou–Brenier problem is therefore

$$\min_U \mathcal{F}(U) + \mathcal{G}(U).$$

On the resulting Hilbert space of unknowns, or formally for an L^2-type product structure, the two proximal operators are

$$\operatorname{prox}_{\tau\mathcal{F}}(\bar{U}) = \operatorname{argmin}_U \frac{1}{2}\|U - \bar{U}\|_{\mathrm{H}}^2 + \tau\mathcal{F}(U), \qquad \operatorname{prox}_{\tau\mathcal{G}}(\bar{U}) = \operatorname{argmin}_{U \in C} \frac{1}{2}\|U - \bar{U}\|_{\mathrm{H}}^2.$$

For the standard product metric, the first map is local in (t, x): it is the proximal operator of the convex perspective J. The second map is the orthogonal projection onto the affine set defined by the divergence equation and endpoint constraints. Douglas–Rachford then alternates these two simple operations:

$$\begin{aligned} U^{k+1} &= \operatorname{prox}_{\tau\mathcal{F}}(Z^k), \\ \widetilde{U}^{k+1} &= \operatorname{prox}_{\tau\mathcal{G}}(2U^{k+1} - Z^k), \\ Z^{k+1} &= Z^k + \widetilde{U}^{k+1} - U^{k+1}. \end{aligned}$$

Equivalently one may swap the roles of $\mathcal{F}$ and $\mathcal{G}$. At convergence, the two shadow sequences U^k and $\widetilde{U}^k$ agree and give a minimizer of the convex dynamic problem. This viewpoint is useful because it separates the nonlinear but pointwise perspective proximal step from the global but linear projection onto the continuity equation Papadakis et al. (2014); Eckstein and Bertsekas (1992).

Algorithm 2.1 Douglas–Rachford for dynamic Benamou–Brenier

Input: Functionals $\mathcal{F}, \mathcal{G} = \iota_C$, proximal parameter $\tau > 0$, initial field Z^0.
Output: Discrete density-momentum field $U^\star$.
For $k = 0, 1, \ldots$ **do**:
  $U^{k+1} = \operatorname{prox}_{\tau\mathcal{F}}(Z^k)$.
  Project reflected point: $\widetilde{U}^{k+1} = \operatorname{prox}_{\tau\mathcal{G}}(2U^{k+1} - Z^k) = \operatorname{Proj}_C(2U^{k+1} - Z^k)$.
  Update $Z^{k+1} = Z^k + \widetilde{U}^{k+1} - U^{k+1}$.
  If $\|U^{k+1} - \widetilde{U}^{k+1}\| \leqslant \mathrm{tol}$ **then**:
    Return U^{k+1}.

Path-space formulation. Let $\mathcal{S} = C([0, 1]; \mathbb{R}^d)$ be the space of continuous paths endowed with the uniform topology. For $t \in [0, 1]$ define the evaluation map

$$P_t : \mathcal{S} \to \mathbb{R}^d, \qquad P_t(\gamma) = \gamma(t).$$

The Benamou–Brenier cost admits the equivalent formulation

$$\mathcal{W}_2^2(\alpha_0, \alpha_1) = \inf_{M \in \mathcal{P}(\mathcal{S})} \left\{ \int_{\mathcal{S}} \int_0^1 \|\dot{\gamma}(t)\|^2 \mathrm{d}t\,\mathrm{d}M(\gamma)\ ;\ (P_0)_\sharp M = \alpha_0,\ (P_1)_\sharp M = \alpha_1 \right\}.$$

density sequence midpoint velocities

Figure 2.2: Benamou–Brenier geodesic between two sampled silhouettes. A discrete quadratic OT plan between finely subsampled cat and two-disks point clouds induces the McCann interpolation $Z_t = (1-t)X + tY$, which is the Lagrangian realization of the least-action solution. The left panel renders local color images of the smaller-bandwidth kernel-smoothed densities with enough padding to include the full silhouettes. The right panel overlays shortened velocity arrows centered at evenly subsampled midpoint particles $Z_{1/2}$; each displayed arrow runs in data coordinates from a source-side tail to a target-side head along the matched characteristic direction $Y - X$, but is not drawn as the full endpoint segment from X to Y.

If α_0 has a density, the minimizer M^* is unique. Its time marginals reproduce the optimal curve: $\alpha_t = (P_t)_\sharp M^*$ for all t. Furthermore, for a.e. t, denoting $Q_t(\gamma) = \dot\gamma(t)$ on absolutely continuous paths, the conditional law of the velocity is deterministic:

$$(P_t, Q_t)_\sharp M^*(\mathrm{d}x, \mathrm{d}q) = \alpha_t(\mathrm{d}x)\delta_{v_t^*(x)}(\mathrm{d}q),$$

where v_t^* is the optimal velocity field in the Benamou–Brenier formulation. Hence M^* concentrates on straight-line geodesics and, for a.e. t, assigns exactly one direction at α_t-a.e. spatial point.

2.3. Generalized Dynamic Wasserstein Distances

The quadratic Benamou–Brenier formula is only one instance of a broader fixed-mass dynamic language. The goal of this section is to define a large family of geodesic-like distances on spaces of probability measures by modifying the action minimized in the Benamou–Brenier formula. The objects introduced here are metric: they specify admissible curves, tangent variables and path energies. All descent constructions are postponed to Section 3.5, where these distances are used to generate gradient-flow PDE models; the reaction–transport case, where mass may vary, is separated in Section 2.5.

Path actions. The common starting point is an instantaneous cost for moving a measure with a prescribed Eulerian velocity. Integrating this cost along continuity-equation curves produces the associated endpoint geometry.

In the mass-preserving Euclidean setting, the basic input is an instantaneous action $\mathbb{A}(\alpha, w)$, where α is the current measure and w is an admissible velocity representative. When this action is normalized as a squared infinitesimal speed, it generates the length-space value

$$\mathsf{D}_{\mathbb{A}}^2(\alpha_0, \alpha_1) = \inf_{\alpha_t, v_t} \left\{ \int_0^1 \mathbb{A}(\alpha_t, v_t)\,\mathrm{d}t : \partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0,\ \alpha_{t=0} = \alpha_0,\ \alpha_{t=1} = \alpha_1 \right\}. \tag{2.13}$$

Equivalently, one may quotient by velocity fields that induce the same first-order variation of the measure. The formula (2.13) should be read as a dynamic definition of the distance, not as a property automatically satisfied by an arbitrary discrepancy. Some standard distances, such as $\mathcal{W}_p$, are first written with a p-homogeneous action and then squared by taking a constant-speed parametrization; this normalization is made explicit below. Different choices of $\mathbb{A}$ change the resulting geometry; Section 3.5 later reuses these choices when dynamics are introduced.

Quadratic, or Riemannian, tangent actions. A particularly transparent case occurs when $w \mapsto \mathbb{A}(\alpha, w)$ is quadratic. For simplicity, take admissible velocities in $L^2(\alpha; \mathbb{R}^d)$; in some applications this Hilbert space is replaced by a closed subspace encoding additional constraints. If the polarization of $\mathbb{A}$ is represented by a positive self-adjoint operator $Q_\alpha : L^2(\alpha; \mathbb{R}^d) \to L^2(\alpha; \mathbb{R}^d)$,

$$\mathbb{A}(\alpha; w, z) = \langle Q_\alpha w, z\rangle_{L^2(\alpha)}, \qquad \mathbb{A}(\alpha, w) = \langle Q_\alpha w, w\rangle_{L^2(\alpha)},$$

and if this quadratic form is nondegenerate after quotienting velocities that induce the same measure variation, then the least-action distance generated by this tensor is

$$\mathsf{D}_Q^2(\alpha_0, \alpha_1) = \mathsf{D}_{\mathbb{A}}^2(\alpha_0, \alpha_1) = \inf_{\substack{\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0 \\ \alpha_{t=0} = \alpha_0,\ \alpha_{t=1} = \alpha_1}} \int_0^1 \left\langle Q_{\alpha_t} v_t, v_t \right\rangle_{L^2(\alpha_t)} \mathrm{d}t. \tag{2.14}$$

The usual $\mathcal{W}_2$ geometry corresponds to $Q_\alpha = \mathrm{Id}$ in this simplified notation. Thus Q_α records how the chosen geometry deforms the Euclidean $L^2(\alpha)$ tangent norm: no deformation for $\mathcal{W}_2$, and a nontrivial tensor for generalized Riemannian geometries. In Section 3.5, the same tensor is reused as a preconditioner for metric descent.

Local velocity actions. Many dynamic distances are local with respect to a reference measure λ. Write $\alpha = a\lambda$. A velocity action is specified by a pointwise integrand

$$A : [0,+\infty) \times \mathbb{R}^d \to [0,+\infty], \qquad (a,w) \mapsto A(a,w),$$

where $a \in \mathbb{R}_+$ is a density value and $w \in \mathbb{R}^d$ is a velocity value, and defines

$$\mathbb{A}(\alpha, w) = \int A\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}(x), w(x)\right) \mathrm{d}\lambda(x). \tag{2.15}$$

For a fixed reference λ, this covers density-dependent mobilities and congestion constraints. If, in addition, A is positively 1-homogeneous in its first variable, $A(\eta a, w) = \eta A(a,w)$ for $\eta \geqslant 0$, then the same formula is intrinsic: replacing λ by another dominating measure gives the same value. The usual Benamou–Brenier action is the model case

$$A_2(a,w) = a\|w\|^2, \qquad \mathbb{A}(\alpha,w) = \int \|w\|^2 \mathrm{d}\alpha. \tag{2.16}$$

Here $a = \mathrm{d}\alpha/\mathrm{d}\lambda$ is the density of transported mass and w is the Eulerian velocity. The factor a says that moving twice as much mass with the same velocity costs twice as much.

Homogeneous momentum actions. The same action can be written in momentum variables, and this is the form in which convexity and metric properties are easiest to read. Set $\omega = \alpha w$, so that ω is a vector-valued measure. When the local description is written with the same reference λ, so that $\alpha = a\lambda$ and $\omega = m\lambda$, the pointwise momentum perspective is

$$J_A(a,m) \coloneqq \begin{cases} A(a, m/a), & a > 0, \\ 0, & a = 0 \text{ and } m = 0, \\ +\infty, & a = 0 \text{ and } m \neq 0, \end{cases} \tag{2.17}$$

and the measure action relative to λ is

$$\mathbb{J}_{A,\lambda}(\alpha,\omega) \coloneqq \int J_A\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}, \frac{\mathrm{d}\omega}{\mathrm{d}\lambda}\right) \mathrm{d}\lambda, \tag{2.18}$$

with value $+\infty$ if α or the total variation $|\omega|$ is not absolutely continuous with respect to λ. This zero-density convention is the lower-semicontinuous one for the superlinear actions used below; other growths use the corresponding recession extension. If A is positively 1-homogeneous in a, then J_A is jointly 1-homogeneous: $J_A(\eta a, \eta m) = \eta J_A(a,m)$. In that intrinsic case the value of $\mathbb{J}_{A,\lambda}$ is independent of the dominating reference measure, and we write simply $\mathbb{J}_A$. For $A_2(a,w) = a\|w\|^2$, one recovers the quadratic perspective

$$J_2(a,m) = \frac{\|m\|^2}{a},$$

which is the integrand used in the convex Benamou–Brenier formulation (2.10). In this case $\omega = \alpha w$; if $\alpha = \rho \mathrm{d}x$ and $\omega = m\mathrm{d}x$, then $m = \rho w$, and $J_2(\rho, m) = \rho\|w\|^2$.

Proposition 2.5 (Concave mobilities give convex momentum actions)**.** *Let $I \subset [0,+\infty)$ be a convex interval, let $\theta : I \to [0,+\infty)$ be concave, and let $L : \mathbb{R}^d \to [0,+\infty]$ be convex with $L(0) = 0$. Define, on the set where $\theta(a) > 0$,*

$$J_{\theta,L}(a,m) \coloneqq \theta(a) L\left(\frac{m}{\theta(a)}\right), \qquad A_{\theta,L}(a,w) \coloneqq J_{\theta,L}(a,aw) = \theta(a) L\left(\frac{aw}{\theta(a)}\right), \tag{2.19}$$

and extend $J_{\theta,L}$ to the boundary by lower semicontinuity. Then $J_{\theta,L}$ is convex in (a,m). In particular, this single construction contains

$$A(a,w) = aL(w) \quad \text{by taking } \theta(a) = a,$$

and the concave-mobility quadratic action

$$A(a,w) = \frac{a^2\|w\|^2}{\theta(a)} \quad \text{by taking } L(u) = \|u\|^2.$$

Proof. The perspective $P(s,m) = sL(m/s)$ of a convex function is convex on $s > 0$. Since $L(0) = 0$, it is also nonincreasing in the scalar s for fixed m: if $s_2 \geqslant s_1 > 0$, then

$$L(m/s_2) = L\big((s_1/s_2)(m/s_1)\big) \leqslant (s_1/s_2)L(m/s_1),$$

hence $P(s_2,m) \leqslant P(s_1,m)$. Let $a_\zeta = (1-\zeta)a_0 + \zeta a_1$ and $m_\zeta = (1-\zeta)m_0 + \zeta m_1$. By concavity of θ,

$$\theta(a_\zeta) \geqslant (1-\zeta)\theta(a_0) + \zeta\theta(a_1).$$

Using the monotonicity in s, and then the convexity of P, gives

$$J_{\theta,L}(a_\zeta, m_\zeta) = P(\theta(a_\zeta), m_\zeta) \leqslant P((1-\zeta)\theta(a_0) + \zeta\theta(a_1), m_\zeta) \leqslant (1-\zeta)J_{\theta,L}(a_0,m_0) + \zeta J_{\theta,L}(a_1,m_1).$$

The boundary cases follow by the chosen lower-semicontinuous extension. □

The next proposition isolates the additional assumptions under which the momentum formulation generated by A defines a genuine path metric rather than only a variational principle.

Proposition 2.6 (Homogeneous dynamic actions define distances)**.** *Fix a reference measure λ, omitted from the notation only in the intrinsic case where $\mathbb{J}_{A,\lambda}$ does not depend on λ. Assume that the momentum perspective J_A defined in (2.17) is lower semicontinuous, convex in (a,m), even in m, and satisfies $J_A(a,0) = 0$. Assume moreover that for some $r > 1$*

$$J_A(a,\xi m) = |\xi|^r J_A(a,m) \qquad \text{for all admissible } a,\ \text{all } m,\ \text{and all } \xi \in \mathbb{R},$$

and that $J_A(a,m) = 0$ if and only if $m = 0$. Equivalently, the evenness, homogeneity and nondegeneracy assumptions translate, away from zero density, into

$$A(a,-w) = A(a,w), \qquad A(a,\xi w) = |\xi|^r A(a,w), \qquad A(a,w) = 0 \Longleftrightarrow w = 0 \qquad (a > 0).$$

Define, on every fixed-mass class,

$$\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) \coloneqq \inf_{\substack{\partial_t \alpha_t + \operatorname{div} \omega_t = 0 \\ \alpha_{t=0} = \alpha_0,\ \alpha_{t=1} = \alpha_1}} \left(\int_0^1 \mathbb{J}_{A,\lambda}(\alpha_t,\omega_t)\,\mathrm{d}t \right)^{1/r}.$$

After the usual lower-semicontinuous relaxation assuming that zero-action values are attained, $\mathsf{D}_{A,\lambda}$ is an extended distance on each finite-action component: it is symmetric, satisfies the triangle inequality, and $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) = 0$ only when $\alpha_0 = \alpha_1$. In the intrinsic case, we write D_A.

Proof. The constant curve $\alpha_t = \alpha_0$, $\omega_t = 0$, gives $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_0) = 0$. Conversely, if $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) = 0$, the relaxed zero-action minimizer satisfies

$$\int_0^1 \mathbb{J}_{A,\lambda}(\alpha_t,\omega_t)\,\mathrm{d}t = 0.$$

Since $\mathbb{J}_{A,\lambda} \geqslant 0$, this implies $\omega_t = 0$ for a.e. t. The continuity equation then gives $\partial_t \alpha_t = 0$ in the sense of distributions, so $\alpha_0 = \alpha_1$.

Symmetry follows by time reversal. If $(\alpha_t,\omega_t)_{t\in[0,1]}$ connects α_0 to α_1, then

$$\widetilde{\alpha}_t = \alpha_{1-t}, \qquad \widetilde{\omega}_t = -\omega_{1-t}$$

connects α_1 to α_0, satisfies the same continuity equation, and has the same action because J_A is even in m.

It remains to prove the triangle inequality. Let (α^1,ω^1) connect α_0 to α_1 with action E_1, and let (α^2,ω^2) connect α_1 to α_2 with action E_2. For $\tau \in (0,1)$, concatenate the two curves after the time changes $s = t/\tau$ and $s = (t-\tau)/(1-\tau)$:

$$(\alpha_t,\omega_t) = \begin{cases} \big(\alpha^1_{t/\tau},\ \tau^{-1}\omega^1_{t/\tau}\big), & 0 \leqslant t \leqslant \tau, \\ \big(\alpha^2_{(t-\tau)/(1-\tau)},\ (1-\tau)^{-1}\omega^2_{(t-\tau)/(1-\tau)}\big), & \tau \leqslant t \leqslant 1. \end{cases}$$

The rescaled curve is admissible from α_0 to α_2. By r-homogeneity in the momentum variable, its action is

$$\tau^{1-r}E_1 + (1-\tau)^{1-r}E_2.$$

Writing $X = E_1^{1/r}$ and $Y = E_2^{1/r}$, the optimal allocation is $\tau = X/(X+Y)$, with the evident limiting convention if $X + Y = 0$, and gives

$$\inf_{\tau\in(0,1)} \left\{\tau^{1-r}E_1 + (1-\tau)^{1-r}E_2\right\} = (X+Y)^r.$$

Hence $\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_2) \leqslant E_1^{1/r} + E_2^{1/r}$. Taking E_1 and E_2 arbitrarily close to the two optimal action values yields

$$\mathsf{D}_{A,\lambda}(\alpha_0,\alpha_2) \leqslant \mathsf{D}_{A,\lambda}(\alpha_0,\alpha_1) + \mathsf{D}_{A,\lambda}(\alpha_1,\alpha_2),$$

with the usual extended-real convention when one of the distances is infinite. □

Without homogeneity or nondegeneracy, the same momentum action remains useful as a variational principle, but its r-th root need not define a distance.

Example 2.7 (Wasserstein-p action). The usual $\mathcal{W}_p$ distances correspond to changing only the homogeneity of the Benamou–Brenier action. The case $p = 2$ is the original formula Benamou and Brenier (2000); for $1 < p < +\infty$, the same dynamic viewpoint gives the standard description of absolutely continuous curves in Wasserstein spaces Ambrosio et al. (2006); Villani (2003); Santambrogio (2015). The specific objects to insert in the general framework above are

$$A_p(a,w) = a\|w\|^p, \qquad J_p(a,m) = \begin{cases} \|m\|^p/a^{p-1}, & a > 0, \\ 0, & (a,m) = (0,0), \\ +\infty, & a = 0,\ m \neq 0, \end{cases}$$

With $A = A_p$, Proposition 2.6 gives the usual identity $\mathsf{D}_{A_p} = \mathcal{W}_p$. The corresponding squared length-space normalization is

$$\mathbb{A}_p(\alpha,w) = \left(\int \|w\|^p \mathrm{d}\alpha\right)^{2/p}.$$

This is the squared version of the p-homogeneous action: minimizing $\int_0^1 \mathbb{A}_p(\alpha_t,v_t)\mathrm{d}t$ gives $\mathcal{W}_p^2$ after constant-speed reparametrization. Thus A_p, J_p denote the local p-homogeneous velocity and momentum densities, whereas $\mathbb{A}_p$ denotes the squared tangent action used in the length formulation. The endpoint $p = 1$ can be treated separately: the perspective becomes $J_1(a,m) = \|m\|$, and the dynamic problem collapses to Beckmann's formulation of $\mathcal{W}_1$ Beckmann (1952).

Concave-mobility actions. One can instead keep a quadratic momentum action and change the mobility. Dolbeault, Nazaret and Savaré introduced this construction as a class of generalized transport distances adapted to nonlinear diffusion Dolbeault et al. (2009). Let $I \subset [0,+\infty)$ be a convex interval and let $\theta : I \to [0,+\infty)$ be concave. Define the extended momentum density

$$J_\theta(a,m) := \begin{cases} \|m\|^2/\theta(a), & \theta(a) > 0, \\ 0, & \theta(a) = 0 \text{ and } m = 0, \\ +\infty, & \theta(a) = 0 \text{ and } m \neq 0. \end{cases}$$

The corresponding velocity action is

$$A_\theta(a,w) = J_\theta(a,aw) = \frac{a^2\|w\|^2}{\theta(a)}.$$

The convexity of J_θ is the special case $L(u) = \|u\|^2$ of Proposition 2.5. This is why concavity of the mobility, rather than convexity, is the structural condition that makes the continuity-equation formulation convex.

Fix now a reference measure λ. If $\alpha = a\lambda$, the induced squared tangent action is

$$\mathbb{A}_{\theta,\lambda}(\alpha,w) = \int A_\theta(a(x),w(x))\,\mathrm{d}\lambda(x) = \int \frac{a(x)^2}{\theta(a(x))}\|w(x)\|^2\,\mathrm{d}\lambda(x),$$

and it is set to $+\infty$ when $\alpha \not\ll \lambda$. Equivalently, on the set where $\theta(a(x)) > 0$,

$$\mathbb{A}_{\theta,\lambda}(\alpha,w) = \int \frac{a(x)}{\theta(a(x))}\|w(x)\|^2\,\mathrm{d}\alpha(x).$$

Hence this is a local Riemannian case, in the sense of (2.14), whenever the multiplier $a(x)/\theta(a(x))$ is finite and positive. The associated tensor is the multiplication operator

$$\big(Q_{\theta,\lambda,\alpha} w\big)(x) = \frac{a(x)}{\theta(a(x))} w(x), \qquad \alpha = a\lambda, \tag{2.20}$$

defined α-a.e. on the set where $\theta(a) > 0$. Except for linear mobilities $\theta(a) = ca$, and in particular the normalized case $\theta(a) = a$ which recovers $\mathcal{W}_2$, the pointwise velocity action $A_\theta(a, w)$ is not positively 1-homogeneous in the density variable a. Consequently the construction is not intrinsic under a change of λ: the resulting distance depends on the chosen reference measure and is finite only between endpoints that can be joined by a finite-action curve with $\alpha_t \ll \lambda$.

The associated value is therefore written

$$\mathsf{W}_{\theta,\lambda}(\alpha_0, \alpha_1) \coloneqq \mathsf{D}_{A_\theta,\lambda}(\alpha_0, \alpha_1),$$

where the subscript λ recalls that the action is measured through the density $a = \mathrm{d}\alpha/\mathrm{d}\lambda$. Here $\mathsf{D}_{A_\theta,\lambda}$ denotes the general path value (2.13) with the action $\mathbb{A} = \mathbb{A}_{\theta,\lambda}$.

Proposition 2.8 (Concave-mobility dynamic distances)**.** *For a fixed reference measure λ, and with the lower-semicontinuous extension of J_θ above, $\mathsf{W}_{\theta,\lambda} = \mathsf{D}_{A_\theta,\lambda}$ is an extended distance on each finite-action component contained in $\{\alpha : \alpha \ll \lambda\}$. In particular, whenever the relevant component has finite pairwise distances, $\mathsf{W}_{\theta,\lambda}$ is a genuine metric there.*

Proof. Proposition 2.5, applied with $L(u) = \|u\|^2$, gives the lower-semicontinuous convex momentum density J_θ. It is even in m, satisfies $J_\theta(a, 0) = 0$, and obeys

$$J_\theta(a, \xi m) = |\xi|^2 J_\theta(a, m).$$

Moreover $J_\theta(a, m) = 0$ if and only if $m = 0$, with the boundary convention used in its definition. For the fixed reference λ, the hypotheses of Proposition 2.6 therefore hold with $r = 2$. That proposition gives symmetry, separation and the triangle inequality for $\mathsf{D}_{A_\theta,\lambda}$, hence for $\mathsf{W}_{\theta,\lambda}$. □

The choice $\theta(a) = a$ recovers $\mathcal{W}_2$. Other choices encode different geometry: $\theta(a) = a^\gamma$ with $0 < \gamma \leqslant 1$ changes the cost of moving dilute mass, while $\theta(a) = a(1 - a/M)$ on $[0, M]$ models a volume-filling or exclusion effect, since mobility vanishes both at vacuum and at the maximal density. The distance is comparable with $\mathcal{W}_2$ on classes where $\theta(a)$ is bounded above and below by positive multiples of a; otherwise zero-mobility barriers can make some pairs infinitely far apart. Its use as a geometry for nonlinear diffusion is discussed in Section 3.5.

Dynamic spectral Wasserstein distances. Static spectral Wasserstein distances penalize a coupling through the covariance of its displacement. A dynamic version keeps the continuity equation but replaces the pointwise kinetic energy by a gauge of the whole velocity covariance. The resulting action is nonlocal in space: velocity directions are charged globally through their covariance, rather than independently at each point.

Let γ be a monotone spectral gauge on $\mathbb{S}^d_+$. For a probability measure α and a velocity field $v \in L^2(\alpha; \mathbb{R}^d)$, define the spectral tangent action

$$\mathbb{A}_\gamma(\alpha, v) \coloneqq \gamma\left(\int v(x) v(x)^\top \mathrm{d}\alpha(x) \right). \tag{2.21}$$

The trace gauge gives the usual Wasserstein tangent action, while the operator gauge $\gamma(M) = \lambda_{\max}(M)$ charges only the largest directional velocity variance. The associated distance is the general path length (2.13) with $\mathbb{A} = \mathbb{A}_\gamma$:

$$\mathsf{W}^2_{\gamma,\mathrm{dyn}} \coloneqq \mathsf{D}^2_{\mathbb{A}_\gamma}. \tag{2.22}$$

In density–momentum variables, this corresponds to the measure action

$$\mathbb{J}_\gamma(\alpha, \omega) = \gamma\left(\int \left(\frac{\mathrm{d}\omega}{\mathrm{d}\alpha}\right) \left(\frac{\mathrm{d}\omega}{\mathrm{d}\alpha}\right)^\top \mathrm{d}\alpha \right),$$

or, when $\alpha = \rho\,\mathrm{d}x$ and $\omega = m\,\mathrm{d}x$,

$$\mathbb{J}_\gamma(\rho, m) = \gamma\left(\int \frac{m(x) m(x)^\top}{\rho(x)}\,\mathrm{d}x \right).$$

This functional is convex in the density–momentum fields (ρ, m) by the matrix perspective, together with the monotonicity and convexity of γ. It is nevertheless not, in general, obtained by integrating a pointwise action density, because the covariance is computed globally before applying γ. It becomes local only for linear spectral gauges. For instance, if $\gamma(M) = \mathrm{tr}(GM)$ with $G \succeq 0$, then the velocity and momentum densities are

$$A_{\mathrm{lin}}(a, w) = a\, w^\top G w, \qquad J_{\mathrm{lin}}(a, m) = \frac{m^\top G m}{a},$$

and the trace gauge, $G = \mathrm{Id}$ in $\gamma(M) = \mathrm{tr}(GM)$, recovers the usual Benamou–Brenier action.

The following result, in the form used for normalized spectral flows in Peyré (2026), shows that this dynamic construction is not merely infinitesimal: it exactly recovers the static displacement-covariance formulation.

Proposition 2.9 (Static/dynamic equality for spectral Wasserstein)**.** *Let γ be a monotone spectral gauge and let $\alpha_0, \alpha_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Then*

$$\mathsf{W}^2_{\gamma,\mathrm{dyn}}(\alpha_0, \alpha_1) = \mathcal{W}^2_\gamma(\alpha_0, \alpha_1).$$

In particular, $\mathcal{W}_\gamma$ defines a distance on $\mathcal{P}_2(\mathbb{R}^d)$.

Proof. We first prove $\mathsf{W}^2_{\gamma,\mathrm{dyn}} \leqslant \mathcal{W}^2_\gamma$. Let $\pi \in \mathcal{U}(\alpha_0, \alpha_1)$ and let $(X, Y) \sim \pi$. Set $Z_t = (1-t)X + tY$ and $\alpha_t = (Z_t)_\sharp \pi$. Define the Eulerian velocity as the conditional mean

$$v_t(z) \coloneqq \mathbb{E}[Y - X \mid Z_t = z].$$

Then (α_t, v_t) solves the continuity equation. If

$$M_\pi \coloneqq \int (x-y)(x-y)^\top \mathrm{d}\pi(x,y), \qquad C_t \coloneqq \int v_t(z) v_t(z)^\top \mathrm{d}\alpha_t(z),$$

conditional Jensen gives $C_t \preceq M_\pi$. Indeed, for every direction $u \in \mathbb{R}^d$,

$$u^\top C_t u = \mathbb{E}\big[\mathbb{E}[\langle u, Y - X\rangle \mid Z_t]^2\big] \leqslant \mathbb{E}[(\langle u, Y - X\rangle)^2] = u^\top M_\pi u.$$

By the Loewner monotonicity of γ,

$$\int_0^1 \gamma(C_t)\mathrm{d}t \leqslant \gamma(M_\pi).$$

Taking the infimum over π gives the first inequality.

Conversely, let (α_t, v_t) be an admissible finite-action competitor. Since γ is a finite positive gauge on the finite-dimensional cone $\mathbb{S}^d_+$, it is equivalent to the trace on this cone; hence finite spectral action implies the usual finite kinetic energy needed for the superposition principle. Thus there exists a probability law η on absolutely continuous paths $\omega_\cdot$ such that $(e_t)_\sharp \eta = \alpha_t$ and $\dot\omega_t = v_t(\omega_t)$ for a.e. t. Let $\pi = (e_0, e_1)_\sharp \eta$ be the induced endpoint coupling. With $C_t = \int v_t v_t^\top \mathrm{d}\alpha_t$, Jensen's inequality along each path gives

$$(\omega_1 - \omega_0)(\omega_1 - \omega_0)^\top = \left(\int_0^1 \dot\omega_t \mathrm{d}t\right)\left(\int_0^1 \dot\omega_t \mathrm{d}t\right)^\top \preceq \int_0^1 \dot\omega_t \dot\omega_t^\top \mathrm{d}t.$$

After integration over η,

$$M_\pi \preceq \int_0^1 C_t \mathrm{d}t.$$

Monotonicity of γ, followed by convexity, yields

$$\gamma(M_\pi) \leqslant \gamma\left(\int_0^1 C_t \mathrm{d}t\right) \leqslant \int_0^1 \gamma(C_t)\mathrm{d}t.$$

Since π is admissible for the static problem, $\mathcal{W}^2_\gamma(\alpha_0, \alpha_1)$ is bounded by the action of the dynamic competitor. Taking the infimum over all competitors gives the reverse inequality.

The crucial assumption is the monotonicity of γ: both parts of the proof produce Loewner-order comparisons between covariance matrices, and these comparisons imply inequalities between actions only for monotone gauges. □

The use of this geometry for normalized flows, including the operator-gauge connection with Muon-type normalization, is developed in Section 3.5.

Kernelized Benamou–Brenier distances. A different way to deform the Benamou–Brenier geometry is to keep the local continuity equation but to measure velocities in a reproducing-kernel Hilbert space rather than in $L^2(\alpha)$. This construction is motivated by Stein variational gradient descent, studied later in Section 4.2: the kernel makes the velocity field smooth and computable from particles, at the price of defining a much more restrictive transport geometry.

Let k be a positive definite kernel on $\mathbb{R}^d$ with scalar RKHS $\mathcal{H}_k$. The vector-valued RKHS is

$$\mathcal{H}_k^d := \mathcal{H}_k \times \cdots \times \mathcal{H}_k, \qquad \|v\|_{\mathcal{H}_k^d}^2 := \sum_{\ell=1}^{d} \|v_\ell\|_{\mathcal{H}_k}^2.$$

This is the vector-valued analogue of the scalar RKHS norm used for maximum mean discrepancies. The specific tangent action is

$$\mathbb{A}_k(\alpha, v) := \|v\|_{\mathcal{H}_k^d}^2, \qquad \mathcal{W}_k^2 := \mathsf{D}_{\mathbb{A}_k}^2, \tag{2.23}$$

where the general length construction is understood with the restricted admissible class $v_t \in \mathcal{H}_k^d$. The action itself is independent of α; the measure only enters through the continuity equation $\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0$, which says how the common smooth velocity field moves all particles. This type of Stein geometry was introduced in the analysis of SVGD by Liu and Wang Liu and Wang (2016); Liu (2017) and later developed geometrically in Duncan et al. (2023); Nüsken and Renger (2023). The important caveat is that the admissible tangent space is the smooth RKHS class, not the whole Wasserstein tangent space.

Proposition 2.10 (Kernelized dynamic distance)**.** *The quantity $\mathcal{W}_k$ defined by (2.23) is an extended distance on each finite-action component of probability measures. More precisely, it is symmetric, satisfies the triangle inequality, and $\mathcal{W}_k(\alpha_0, \alpha_1) = 0$ implies $\alpha_0 = \alpha_1$. It can take the value $+\infty$ between different components.*

Proof. The proof is the standard length-space argument for a quadratic action. The constant curve gives zero self-distance. If $\mathcal{W}_k(\alpha_0, \alpha_1) = 0$, a zero-action relaxed curve has $v_t = 0$ for a.e. t, hence the continuity equation gives $\partial_t \alpha_t = 0$ and $\alpha_0 = \alpha_1$. Symmetry follows by time reversal and by replacing v_t with $-v_{1-t}$. For the triangle inequality, concatenate an almost optimal curve from α_0 to α_1 of action E_1 with one from α_1 to α_2 of action E_2. Compressing the first curve into a time interval of length τ multiplies its action by $1/\tau$, and compressing the second into length $1-\tau$ multiplies its action by $1/(1-\tau)$. Optimizing over $\tau \in (0,1)$ gives $(\sqrt{E_1} + \sqrt{E_2})^2$, and taking infima yields the triangle inequality. □

One should read $\mathcal{W}_k$ as an extended distance on finite-action components, not as a replacement for $\mathcal{W}_2$ on all of $\mathcal{P}_2(\mathbb{R}^d)$. A useful sufficient condition for finiteness is that the endpoints lie on the same RKHS-flow orbit: if there exists $v \in L^2([0,1]; \mathcal{H}_k^d)$ whose flow map Φ_t solves $\dot{\Phi}_t = v_t \circ \Phi_t$ and satisfies $\alpha_1 = (\Phi_1)_\sharp \alpha_0$, then $\mathcal{W}_k^2(\alpha_0, \alpha_1) \leqslant \int_0^1 \|v_t\|_{\mathcal{H}_k^d}^2 \mathrm{d}t$. In particular, for strictly positive definite kernels, two discrete measures with the same weights and distinct moving support points are at finite distance whenever their atoms can be connected by noncolliding smooth paths, because RKHS interpolation constructs vector fields realizing the prescribed atom velocities along the paths.

The same condition also explains the limitation. If k is smooth enough that $\mathcal{H}_k^d$ embeds into Lipschitz vector fields, finite-action curves are induced by flows of regular maps. Atomic measures therefore remain atomic with the same number of atoms along such curves, so a Dirac mass cannot be transported at finite kernelized action to a measure with a density. This lack of splitting is precisely what makes the geometry useful for deterministic particle methods, and also what makes it a nontrivial extended object rather than a full probability metric.

2.4. Nonlocal Wasserstein Distances

The local dynamic distances above transport mass through a vector field on the base space and a classical continuity equation. Nonlocal geometries use a different tangent model: the elementary motion is an exchange across an edge or a jump from x to y. The common data are a reversible kernel K, a symmetric edge or jump measure J, a pairwise increment $\bar{\nabla}$, and a pair-space action $\mathbb{A}_K$. The tangent variable is therefore attached to pairs of points, not to a single point x, so these constructions are not simply obtained by choosing another pointwise local action-density $A(\rho(x), m(x))$.

There are two complementary versions. On a finite state space, the goal is to put a Wasserstein-like geometry on the probability simplex so that the entropy gradient flow is exactly a prescribed reversible Markov chain Maas (2011); Mielke (2013); Chow et al. (2012). On a continuum space, the same edge calculus becomes a jump calculus over $\mathcal{X} \times \mathcal{X}$, which models nonlocal motion, heavy-tailed jumps, and fractional-type diffusion; this is the

construction of Erbar Erbar (2014), building on nonlinear mobilities Dolbeault et al. (2009), with subsequent metric and asymptotic refinements Slepčev and Warren (2022).

In both settings the canonical mobility for entropy is the logarithmic mean

$$\theta(a,b) := \begin{cases} \dfrac{a-b}{\log a - \log b}, & a \neq b, \\ a, & a = b, \end{cases} \tag{2.24}$$

with the usual lower-semicontinuous extension at $a = 0$ or $b = 0$. It appears because $\theta(a,b)(\log a - \log b) = a - b$, which is the edge-wise chain rule identifying entropy-driven flows with the underlying reversible Markov or jump dynamics.

Continuum jump kernels. The continuum version replaces graph edges by a symmetric measure on pairs. Its action is still quadratic, but the tangent variable is an antisymmetric jump velocity $v(x,y)$, and the mobility depends simultaneously on the two endpoint densities $\rho(x)$ and $\rho(y)$. It is best viewed as a convex action on the pair space $\mathcal{X} \times \mathcal{X}$, rather than as an integral of independent costs attached to single base points x.

Let $(\mathcal{X}, \mathfrak{m})$ be a reference measure space, and let $K(x,\cdot)$ be a nonnegative measure on $\mathcal{X}$ for each $x \in \mathcal{X}$, possibly of infinite total mass. We write this kernel as $K(x,\mathrm{d}y)$ to emphasize that the integration variable is y. The pair measure J on $\mathcal{X} \times \mathcal{X}$ is defined by testing against nonnegative measurable functions Φ:

$$\int_{\mathcal{X}\times\mathcal{X}} \Phi(x,y)\, \mathsf{J}(\mathrm{d}x,\mathrm{d}y) := \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \Phi(x,y)\, K(x,\mathrm{d}y) \right) \mathfrak{m}(\mathrm{d}x). \tag{2.25}$$

The reversibility assumption is precisely that this measure J is symmetric, i.e. invariant under $(x,y) \mapsto (y,x)$. For a density $\rho = \mathrm{d}\alpha/\mathrm{d}\mathfrak{m}$, write

$$\bar{\nabla}\varphi(x,y) := \varphi(y) - \varphi(x)$$

for the nonlocal gradient, and use the logarithmic mean θ defined in (2.24). A curve $\alpha_t = \rho_t \mathfrak{m}$ driven by an antisymmetric velocity $v_t(x,y) = -v_t(y,x)$ satisfies the nonlocal continuity equation if, for all test functions φ,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi \,\mathrm{d}\alpha_t = \frac{1}{2} \iint \bar{\nabla}\varphi(x,y)\, v_t(x,y)\, \theta(\rho_t(x), \rho_t(y))\, \mathsf{J}(\mathrm{d}x,\mathrm{d}y). \tag{2.26}$$

The corresponding pair-space tangent action is

$$\mathbb{A}_K(\alpha, v) := \frac{1}{2} \iint |v(x,y)|^2 \theta(\rho(x), \rho(y))\, \mathsf{J}(\mathrm{d}x,\mathrm{d}y),$$

for $\alpha = \rho\mathfrak{m}$. This is the nonlocal analogue of a tangent action; here v is not a vector field on $\mathcal{X}$ but an antisymmetric velocity on pairs (x,y). The nonlocal transport distance is

$$\mathcal{W}_K^2(\alpha_0, \alpha_1) := \inf_{\rho_t, v_t} \int_0^1 \mathbb{A}_K(\alpha_t, v_t)\, \mathrm{d}t, \tag{2.27}$$

where the infimum is over all curves solving (2.26) with endpoints α_0, α_1.

Proposition 2.11 (Metric properties of nonlocal transport)**.** *Under the regularity and irreducibility assumptions of Erbar (2014), $\mathcal{W}_K$ is an extended distance on the set of probability measures absolutely continuous with respect to $\mathfrak{m}$, and finite-distance pairs are connected by constant-speed geodesics.*

Proof. We use the analytic compactness and lower-semicontinuity theorem of Erbar (2014) for the logarithmic-mean action. In particular, action-bounded sequences of admissible curves are compact for the narrow topology, the weak nonlocal continuity equation is closed under this convergence, and the action is lower semicontinuous. These facts are the nonlocal analogues of the compactness theorem used in the Benamou–Brenier formulation.

Nonnegativity is immediate from the definition of $\mathbb{A}_K(\alpha, v)$. If $\alpha_0 = \alpha_1$, the constant curve $\rho_t = \rho_0$, $v_t = 0$, is admissible and has zero action, hence $\mathcal{W}_K(\alpha_0, \alpha_0) = 0$.

Symmetry follows by time reversal. If (ρ_t, v_t) transports α_0 to α_1, set

$$\tilde{\rho}_t = \rho_{1-t}, \qquad \tilde{v}_t = -v_{1-t}.$$

For every test function φ, differentiating $\int \varphi \tilde{\rho}_t \,\mathrm{d}\mathfrak{m}$ and using (2.26) shows that $(\tilde{\rho}_t, \tilde{v}_t)$ solves the same continuity equation from α_1 to α_0. Since the action is quadratic in v, $\mathbb{A}_K(\tilde{\alpha}_t, \tilde{v}_t) = \mathbb{A}_K(\alpha_{1-t}, v_{1-t})$, so the action is unchanged.

For the triangle inequality, let (ρ_t^0, v_t^0) connect α_0 to α_1 with action A_0, and let (ρ_t^1, v_t^1) connect α_1 to α_2 with action A_1. For $0 < \zeta < 1$, concatenate the two curves after the rescaling

$$(\rho_t, v_t) = \begin{cases} \big(\rho^0_{t/\zeta},\ \zeta^{-1} v^0_{t/\zeta}\big), & 0 \leqslant t \leqslant \zeta, \\ \big(\rho^1_{(t-\zeta)/(1-\zeta)},\ (1-\zeta)^{-1} v^1_{(t-\zeta)/(1-\zeta)}\big), & \zeta < t \leqslant 1. \end{cases}$$

The factors ζ^{-1} and $(1-\zeta)^{-1}$ are exactly those needed for the weak continuity equation after changing time. Since $v \mapsto \mathbb{A}_K(\alpha, v)$ is quadratic,

$$\int_0^1 \mathbb{A}_K(\alpha_t, v_t)\,\mathrm{d}t = \frac{A_0}{\zeta} + \frac{A_1}{1-\zeta}.$$

Optimizing in ζ, or taking $\zeta = \sqrt{A_0}/(\sqrt{A_0} + \sqrt{A_1})$ when both actions are positive, gives a concatenated action $(\sqrt{A_0} + \sqrt{A_1})^2$. Taking infima over the two curves yields

$$\mathcal{W}_K(\alpha_0, \alpha_2) \leqslant \mathcal{W}_K(\alpha_0, \alpha_1) + \mathcal{W}_K(\alpha_1, \alpha_2).$$

It remains to justify definiteness. If $\mathcal{W}_K(\alpha_0, \alpha_1) = 0$, choose admissible curves with actions tending to zero. By the compactness and lower-semicontinuity theorem quoted above, a subsequence converges to an admissible curve (ρ_t, v_t) of zero action. Hence $v_t = 0$ for $\theta(\rho_t(x), \rho_t(y))\mathsf{J}(\mathrm{d}x, \mathrm{d}y)\mathrm{d}t$-a.e. (t, x, y). Substituting in the weak continuity equation gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi\,\mathrm{d}\alpha_t = 0$$

for every admissible test function φ. The irreducibility/separation assumption in Erbar (2014) ensures that these test functions determine the measure, so α_t is constant and $\alpha_0 = \alpha_1$.

Finally, if $\mathcal{W}_K(\alpha_0, \alpha_1) < +\infty$, the same direct-method compactness applied to a minimizing sequence gives a minimizer of (2.27). Reparametrizing this minimizing curve by metric arclength gives a constant-speed curve. Equivalently, for every $0 \leqslant s < t \leqslant 1$, the restriction of the minimizer to $[s, t]$, rescaled to unit time, is admissible between α_s and α_t; the triangle inequality and equality of the total minimal action force

$$\mathcal{W}_K(\alpha_s, \alpha_t) = (t-s)\mathcal{W}_K(\alpha_0, \alpha_1)$$

after this constant-speed parametrization. Thus finite-distance pairs are joined by constant-speed geodesics. The consequences for entropy dynamics and fractional PDE examples are developed in Section 3.6. □

Discrete Wasserstein distances on Markov chains. The finite-state version keeps the same pair-space philosophy, but with a finite graph of admissible exchanges. It is not the naive Euclidean metric on the simplex. The key idea, introduced by Maas and independently developed in related forms by Mielke and by Chow–Huang–Li–Zhou, is to use the transition graph of a reversible Markov chain to define both the admissible directions and the mobility of the mass Maas (2011); Mielke (2013); Chow et al. (2012). The resulting distance is the finite-dimensional counterpart of the Wasserstein geometry behind the heat equation, and it is also the geometric structure used in numerical work on discrete metric-measure spaces Erbar and Maas (2014); Erbar et al. (2020). Its entropy interpretation is stated later in Proposition 3.48.

Let $\mathcal{X} = \{1, \dots, n\}$ and let $K = (K_{ij})$ denote the off-diagonal transition rates of an irreducible continuous-time Markov chain which is reversible with respect to a probability vector π, so that $\pi_i K_{ij} = \pi_j K_{ji}$ for $i \neq j$. We write

$$\mathsf{J}_{ij} := \pi_i K_{ij} = \pi_j K_{ji}$$

for the symmetric edge measure, the finite counterpart of the jump measure used above. The transported object is a mass histogram

$$\Sigma_n := \left\{ a \in \mathbb{R}^n_+ : \sum_i a_i = 1 \right\}.$$

Relative densities only enter as auxiliary variables with respect to the invariant law:

$$\rho_i(a) := \frac{a_i}{\pi_i}, \qquad a_i = \pi_i \rho_i(a).$$

The logarithmic mean θ defined in (2.24) is the mobility selected so that the entropy calculus later recovers exactly the Markov evolution; see Proposition 3.48. The identity $\theta(a, b)(\log a - \log b) = a - b$ converts entropy gradients into density differences along graph edges. For a potential $\psi \in \mathbb{R}^n$, set

$$\bar{\nabla}\psi(i, j) := \psi_j - \psi_i.$$

The finite nonlocal divergence is encoded by the density Onsager operator and by its mass form

$$(\mathcal{K}_\rho \psi)_i \coloneqq \sum_j K_{ij}\theta(\rho_i,\rho_j)(\psi_i-\psi_j), \qquad (\mathcal{L}_a\psi)_i \coloneqq \pi_i(\mathcal{K}_{\rho(a)}\psi)_i. \tag{2.28}$$

Its associated tangent action is

$$\mathbb{A}_K(a,\psi) \coloneqq \frac{1}{2}\sum_{i,j} \mathsf{J}_{ij}\theta(\rho_i(a),\rho_j(a))(\bar{\nabla}\psi(i,j))^2. \tag{2.29}$$

This is the finite-state squared tangent action; the tangent variable is the potential ψ, or equivalently the induced edge flux, rather than an ambient Euclidean vector field. The discrete transport distance is the least action

$$\mathcal{W}_K^2(a_0,a_1) \coloneqq \inf_{a_t,\psi_t} \int_0^1 \mathbb{A}_K(a_t,\psi_t)\,\mathrm{d}t, \qquad \dot{a}_t + \mathcal{L}_{a_t}\psi_t = 0, \tag{2.30}$$

with endpoint conditions $a_{t=0} = a_0$ and $a_{t=1} = a_1$. Equivalently, one can write the same formula in edge-flux variables, exactly as in a finite-volume discretization: the flux is only allowed along edges where $K_{ij} > 0$, and the denominator in the kinetic energy is the logarithmic mean of the two relative endpoint densities $\rho_i(a) = a_i/\pi_i$.

The first nontrivial finite Markov geometries already show how the logarithmic mean bends the simplex. In both examples below, take the uniform random walk on the complete neighbor graph, i.e. every pair of distinct points is declared neighboring. Thus $\pi_i = 1/n$, and $K_{ij} = 1/(n-1)$ for $i \neq j$.

Example 2.12 (Two-point complete graph). On Σ_2, write $a_0 = (r, 1-r)$ and $a_1 = (s, 1-s)$. Since there is only one edge, the discrete dynamic problem reduces to the scalar Riemannian length

$$\mathcal{W}_K(a_0,a_1) = \left| \int_s^r \frac{\mathrm{d}u}{\sqrt{\theta(u,1-u)}} \right|, \qquad 0 < r,s < 1. \tag{2.31}$$

This formula is closed but not Euclidean: the logarithmic mean changes the cost of moving mass depending on the current split between the two points.

Example 2.13 (Three-point complete graph). On Σ_3, the complete-neighbor graph is a triangle. For $a \in \mathrm{int}(\Sigma_3)$, set

$$\Theta_{ij}(a) \coloneqq \frac{1}{2}\theta(a_i,a_j), \qquad 1 \leqslant i < j \leqslant 3.$$

For a fixed a, write $\Theta_{ij} = \Theta_{ij}(a)$. For a tangent vector $h \in \mathbb{R}^3$ with $h_1 + h_2 + h_3 = 0$, orient the edges as $1 \to 2$, $1 \to 3$, $2 \to 3$. The squared norm induced by the discrete Wasserstein metric is

$$\|h\|_a^2 = \min_{m_{12},m_{13},m_{23}} \left\{ \frac{m_{12}^2}{\Theta_{12}} + \frac{m_{13}^2}{\Theta_{13}} + \frac{m_{23}^2}{\Theta_{23}} \right\}, \tag{2.32}$$

subject to

$$h_1 + m_{12} + m_{13} = 0, \qquad h_2 - m_{12} + m_{23} = 0, \qquad h_3 - m_{13} - m_{23} = 0.$$

Eliminating the three edge fluxes gives an explicit formula. With $D = \Theta_{12}^{-1} + \Theta_{13}^{-1} + \Theta_{23}^{-1}$,

$$m_{12}^* = \frac{h_2/\Theta_{23} - h_1/\Theta_{13}}{D}, \qquad m_{13}^* = -h_1 - m_{12}^*, \qquad m_{23}^* = m_{12}^* - h_2,$$

and $\|h\|_a^2$ is obtained by inserting these values in (2.32). Therefore

$$\mathcal{W}_K^2(a_0,a_1) = \inf_{a_t \in \mathrm{int}(\Sigma_3)} \int_0^1 \|\dot{a}_t\|_{a_t}^2\,\mathrm{d}t, \qquad a_{t=0} = a_0, \quad a_{t=1} = a_1. \tag{2.33}$$

Thus the three-state distance is an explicit two-dimensional Riemannian geodesic problem on the open triangle. The formula is simple enough to compute directly, but it already shows the main difference with Euclidean geometry on the simplex: the local metric depends nonlinearly on the current density through logarithmic edge mobilities.

Figure 2.3 visualizes these small-dimensional geometries and compares them with the ordinary Wasserstein distance associated with the $0/1$ ground metric, for which W_2^2 is exactly the total variation distance.

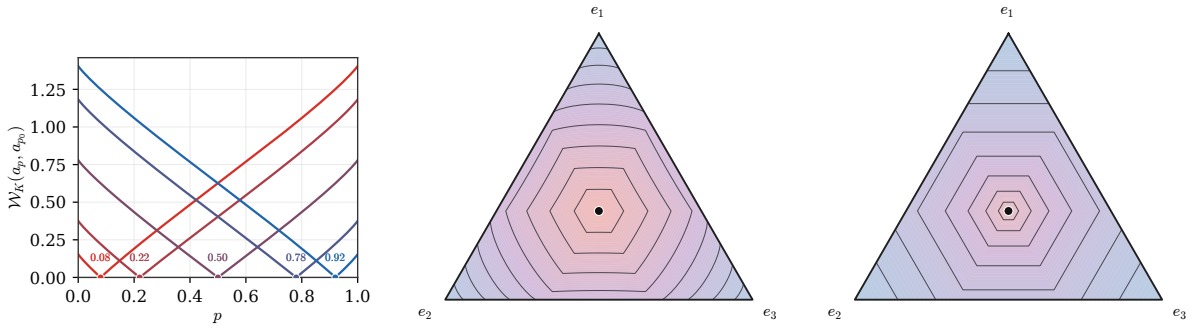


Figure 2.3: Discrete Wasserstein distances on small Markov-chain simplices. Left: the closed-form profiles $r \mapsto \mathcal{W}_K(a_r, a_{r_0})$, with $a_r = (r, 1-r)$, for several anchors r_0 on Σ_2. Middle: numerical level sets of $\mathcal{W}_K(a, \bar{a})$ on Σ_3, where $\bar{a} = (1/3, 1/3, 1/3)$, computed from the local Riemannian norm induced by the complete-neighbor Markov chain. Right: the corresponding level sets for the ordinary W_2 distance with $d(i, j) = 1$ for $i \neq j$, i.e. $W_2^2(a, \bar{a}) = \|a - \bar{a}\|_{\mathrm{TV}}$.

2.5. Dynamic Unbalanced Wasserstein Distances

Balanced dynamic distances keep the total mass fixed: their tangent vectors are transport velocities or fluxes satisfying a continuity equation. Unbalanced distances use a different tangent model. A tangent vector now has both a spatial component and a reaction component, so mass can move, disappear and reappear. This subsection isolates this reaction–transport geometry before its use for gradient flows in Section 3.7.

Balance equation and tangent variables. Unbalanced dynamic transport is obtained by allowing mass to be created and destroyed along the path. The continuity equation is replaced by a balance equation, and an admissible tangent direction is a pair (m, s): the flux m transports mass in space, while the source s changes the amount of mass locally. This dynamic formulation underlies the Hellinger–Kantorovich and Wasserstein–Fisher–Rao metrics Liero et al. (2016); Chizat et al. (2018b); its equivalence with static entropy-transport and cone formulations is developed in Liero et al. (2018); Chizat et al. (2018a). A representative quadratic balance equation is

$$\partial_t \rho_t + \nabla \cdot m_t = s_t, \qquad \int_0^1 \int \left(\frac{|m_t|^2}{\rho_t} + \kappa^2 \frac{s_t^2}{\rho_t} \right) \mathrm{d}x \, \mathrm{d}t,$$

with the usual perspective convention: zero flux and zero source through zero density cost nothing, whereas nonzero flux or source through zero density has infinite cost.

Reaction–transport action. The action attaches one price to displacement and another to local growth. This is the same notation as for balanced dynamic distances, except that the pointwise action now has three variables. On the velocity side, for a mass density $a \geqslant 0$, a velocity $w \in \mathbb{R}^d$, and a relative growth rate $g \in \mathbb{R}$, define

$$A_\kappa(a, w, g) := a\big(\|w\|^2 + \kappa^2 g^2\big). \tag{2.34}$$

Thus, writing $m_t = \rho_t v_t$ and $s_t = \rho_t g_t$, the smooth action is $\int_0^1 \int A_\kappa(\rho_t, v_t, g_t) \mathrm{d}x \, \mathrm{d}t$ under $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = g_t \rho_t$. The parameter κ fixes the relative cost of reaction and transport; changing it rescales the radial/angular balance in the associated cone metric.

For the convex measure formulation, one passes to the flux and source variables $m = aw$ and $r = ag$. The corresponding three-variable perspective is

$$J_\kappa(a, m, r) := \begin{cases} \dfrac{\|m\|^2 + \kappa^2 r^2}{a}, & a > 0, \\ 0, & a = 0,\ m = 0,\ r = 0, \\ +\infty, & a = 0,\ (m, r) \neq (0, 0). \end{cases} \tag{2.35}$$

Equivalently, $J_\kappa(a, m, r) = A_\kappa(a, m/a, r/a)$ when $a > 0$. For measure-valued triples, write α for the transported measure, ω for the vector-valued flux measure, and σ for the signed source measure. If λ dominates α, $|\omega|$, and $|\sigma|$, define

$$\mathbb{J}_\kappa(\alpha, \omega, \sigma) := \int J_\kappa\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}, \frac{\mathrm{d}\omega}{\mathrm{d}\lambda}, \frac{\mathrm{d}\sigma}{\mathrm{d}\lambda} \right) \mathrm{d}\lambda. \tag{2.36}$$

Since J_κ is convex and positively one-homogeneous in (a, m, r), this definition is independent of the dominating measure λ. Finite action forces both the flux and source to be absolutely continuous with respect to the transported mass.

Static and dynamic viewpoints. The dynamic balance-equation formula is not a separate object from static unbalanced OT: it is the least-action representation of the same cone distance. The next proposition records this identification and fixes the normalization used later.

Proposition 2.14 (Static/dynamic equivalence for unbalanced OT)**.** *Fix the action above and let* CW_κ *be the corresponding static cone value, with the cone metric normalized to the same growth scale* κ*. For nonnegative finite measures* α_0, α_1 *on* $\mathbb{R}^d$*, the dynamic value*

$$\mathrm{WFR}_\kappa^2(\alpha_0, \alpha_1) := \inf_{\substack{\partial_t \alpha_t + \nabla\cdot\omega_t = \sigma_t \\ \alpha_{t=0}=\alpha_0,\ \alpha_{t=1}=\alpha_1}} \int_0^1 \mathbb{J}_\kappa(\alpha_t, \omega_t, \sigma_t)\,\mathrm{d}t \tag{2.37}$$

equals the static cone formulation $\mathrm{CW}_\kappa(\alpha_0, \alpha_1)$*. Hence the static unbalanced problem and the balance-equation least-action problem define the same geodesic distance.*

Proof. The cone construction turns variation of mass into radial motion and spatial transport into angular motion on $\mathfrak{C}[\mathbb{R}^d]$. Applying the Benamou–Brenier theorem on the cone to the lifted endpoint measures gives a dynamic least-action problem on $\mathfrak{C}[\mathbb{R}^d]$ whose static value is the cone value CW_κ. This is the standard static/dynamic identification for the Hellinger–Kantorovich and Wasserstein–Fisher–Rao metrics Liero et al. (2016, 2018); Chizat et al. (2018b,a).

Projecting a cone curve back to the base space with the squared cone radius as weight produces a measure curve α_t, a spatial flux measure ω_t, and a source measure σ_t satisfying the balance equation. With the matching normalization of the cone metric, the cone kinetic energy decomposes exactly into the three-variable perspective action $\mathbb{J}_\kappa$ in (2.37). Conversely, any finite-action triple $(\alpha_t, \omega_t, \sigma_t)$ can be lifted to a cone curve whose radial velocity realizes the growth term and whose spatial velocity realizes the transport term, with the same action after relaxation. The two infima are therefore equal; lower semicontinuity gives the general finite-measure statement from the smooth positive case. □

Balanced versus unbalanced geodesics. The distinction is visible already for one-dimensional mixtures with mismatched modal masses. Balanced transport must physically move the excess mass, whereas unbalanced transport can trade transport against reaction.

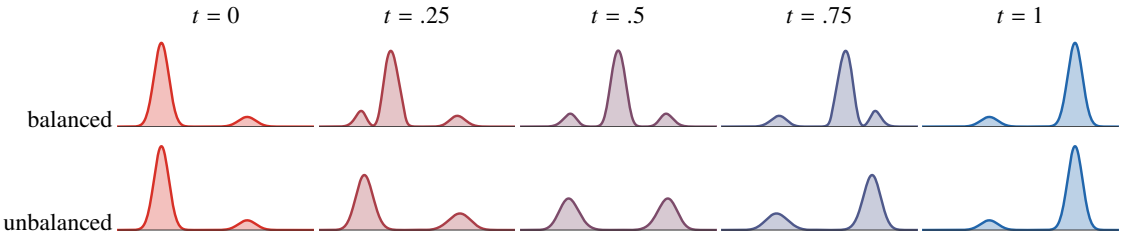


Figure 2.4: Balanced and unbalanced Sinkhorn-barycenter interpolations between two one-dimensional Gaussian mixtures with swapped modal masses. The balanced row conserves total mass, so excess mass from the dominant left mode must move along the line toward the dominant right target mode, producing transient mass in the middle. The unbalanced row uses KL-relaxed marginal constraints; mass can be attenuated near overrepresented modes and recreated near underrepresented modes, giving a reaction–transport interpolation closer to the Wasserstein–Fisher–Rao intuition.

3. Wasserstein Gradient Flows

Once dynamic transport gives a metric on measures, one can run gradient descent directly on probability laws. This section follows the current OT4ML treatment while emphasizing the PDE/ML mechanisms: JKO schemes, Wasserstein gradients, geodesic convexity, functional inequalities, mean-field neural-network training, generalized transport geometries, conditional ResNets and inertial particle flows.

3.1. Minimizing Movements and Wasserstein Gradients

This first section explains how a variational implicit-Euler step on measures gives rise, in the small-step limit, to a continuity equation driven by the Wasserstein gradient of the energy.

We now consider a function $f(\alpha)$ and seek a minimizing evolution $(\alpha_t)_t$. The general strategy of minimizing movement over a metric space is to construct a discrete-time evolution using an implicit Euler scheme:

$$\alpha_{t+\tau} \in \operatorname*{argmin}_{\alpha\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \frac{1}{2\tau} \mathcal{W}_2(\alpha_t, \alpha)^2 + f(\alpha) \right\}. \tag{3.1}$$

Euclidean gradient flows. If we restrict (3.1) to finite dimensions and assume $\alpha_t = \delta_{x(t)}$ and $\alpha = \delta_x$ (single Dirac measures), this matches the implicit Euler scheme:

$$x(t+\tau) \in \operatorname*{argmin}_x \left\{ \frac{1}{2\tau}\|x - x(t)\|^2 + h(x) \right\},$$

where $h(x) = f(\delta_x)$. When h is differentiable and the minimizer is unique, the solution is characterized by the implicit Euler formula

$$x(t+\tau) = (\mathrm{Id} + \tau\nabla h)^{-1}(x(t)).$$

In contrast, the explicit Euler scheme is:

$$x(t+\tau) = (\mathrm{Id} - \tau\nabla h)(x(t)) = x(t) - \tau\nabla h(x(t)).$$

Both schemes converge as $\tau \to 0$ to:

$$\dot{x}(t) = -\nabla h(x(t)). \tag{3.2}$$

Wasserstein gradient formula. The implicit Euler scheme has the advantage that it does not require h or f to be smooth. For f, this is crucial to handle evolutions over measures that may have densities, atoms or other singular parts.

As $\tau \to 0$, under certain conditions on f, (3.1) defines a continuous evolution $t \mapsto \alpha_t$. As discussed earlier, this evolution can be described as a Lagrangian evolution (2.1). We use the following first-variation convention: for any $\beta \in \mathcal{P}(\mathbb{R}^d)$ and the signed zero-mass perturbation $\eta = \beta - \alpha$,

$$f((1-\tau)\alpha + \tau\beta) = f(\alpha + \tau\eta) = f(\alpha) + \tau \int [\delta f(\alpha)](x)\mathrm{d}\eta(x) + o(\tau).$$

The key infinitesimal object is the vector field that represents this differential in the Wasserstein metric.

Definition 3.1 (Wasserstein gradient)**.** Assume that f admits a smooth first variation $\delta f(\alpha)$. In the smooth formal calculus on $\mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein gradient of f at α is the gradient vector field

$$\nabla_W f(\alpha) = \nabla_x \delta f(\alpha).$$

The associated formal gradient flow is the continuity equation

$$\frac{\partial \alpha_t}{\partial t} + \mathrm{div}(-\nabla_W f(\alpha_t)\alpha_t) = 0. \tag{3.3}$$

The following proposition explains why this vector field is the Riemannian gradient for the $L^2(\alpha)$ metric on velocities.

Proposition 3.2 (Formal Wasserstein gradient)**.** *Assume that f admits a smooth first variation $\delta f(\alpha)$ and that α has a smooth positive density. Let v be a sufficiently regular velocity field, compactly supported or satisfying a no-flux boundary condition. For perturbations generated by $(\mathrm{Id} + \tau v)_\sharp \alpha$, the differential of f is*

$$\frac{\mathrm{d}}{\mathrm{d}\tau}_{|\tau=0} f\big((\mathrm{Id} + \tau v)_\sharp \alpha\big) = \int \langle \nabla \delta f(\alpha)(x),\, v(x)\rangle \mathrm{d}\alpha(x).$$

Hence, for the Riemannian metric $\|v\|^2_{L^2(\alpha)} = \int \|v\|^2 \mathrm{d}\alpha$, the Wasserstein gradient is the vector field

$$\nabla_W f(\alpha) = \nabla \delta f(\alpha).$$

Proof. The push-forward expansion gives, in the sense of distributions,

$$(\mathrm{Id} + \tau v)_\sharp \alpha = \alpha - \tau\,\mathrm{div}(\alpha v) + o(\tau).$$

Using the definition of the first variation,

$$f((\mathrm{Id} + \tau v)_\sharp \alpha) = f(\alpha) - \tau \int \delta f(\alpha)\,\mathrm{div}(\alpha v)\mathrm{d}x + o(\tau).$$

An integration by parts, with either compact support or vanishing boundary flux, gives

$$-\int \delta f(\alpha)\,\mathrm{div}(\alpha v)\mathrm{d}x = \int \langle \nabla \delta f(\alpha),\, v\rangle \mathrm{d}\alpha.$$

By definition of the Riesz representative for the $L^2(\alpha)$ metric, this representative is $\nabla \delta f(\alpha)$. □

The Wasserstein gradient-flow viewpoint already appears in John D. Lafferty's PhD work, published as "The Density Manifold and Configuration Space Quantization", under the name "density manifold". It was then systematically developed by Otto, who exposed the formal Riemannian structure of this space Otto (2001). Rigorous metric-space treatments and numerical JKO schemes can be found in Ambrosio et al. (2006); Benamou et al. (2016); Peyré (2015); Gallouët and Monsaingeon (2017).

From the JKO step to the velocity field. A first-order expansion of the JKO step explains why (3.3) uses the vector field $\nabla_{\mathcal{W}} f(\alpha)$. This is a formal local argument: assume that α_t is absolutely continuous and that, for the nearby competitors selected by the JKO step, the optimal map from α_t can be written as $\mathrm{Id} + \tau v$. For these optimal displacement representatives,

$$\mathcal{W}_2^2\big(\alpha_t, (\mathrm{Id} + \tau v)_\sharp \alpha_t\big) = \tau^2 \|v\|^2_{L^2(\alpha_t)}.$$

The local JKO objective therefore reads

$$\min_v \frac{\tau}{2}\|v\|^2_{L^2(\alpha_t)} + f\big((\mathrm{Id} + \tau v)_\sharp \alpha_t\big),$$

where the minimization is restricted to these representatives. The two expansions

$$\begin{aligned}(\mathrm{Id} + \tau v)_\sharp \alpha_t &= \alpha_t - \tau \,\mathrm{div}(v\alpha_t) + o(\tau),\\ f\big((\mathrm{Id} + \tau v)_\sharp \alpha_t\big) &= f(\alpha_t) + \tau \int \langle \nabla_x \delta f(\alpha_t)(x),\, v(x)\rangle \mathrm{d}\alpha_t(x) + o(\tau)\end{aligned}$$

give the first-order problem

$$\min_v f(\alpha_t) + \tau \int \left[\frac{1}{2}\|v(x)\|^2 + \langle \nabla_{\mathcal{W}} f(\alpha_t)(x),\, v(x)\rangle\right] \mathrm{d}\alpha_t(x) + o(\tau).$$

Its pointwise minimizer is $v = -\nabla_{\mathcal{W}} f(\alpha_t)$, which gives the velocity in the continuity equation. A rigorous passage from the discrete scheme to a curve of maximal slope uses compactness and lower-semicontinuity rather than this formal parametrization; see Benamou et al. (2016); Ambrosio et al. (2006).

Metric derivative and curves of maximal slope. The same descent principle admits a coordinate-free formulation, which is the right language for limits of JKO schemes. Let $(\mathcal{X}, d)$ be a metric space and let $x : [0,T] \to \mathcal{X}$ be an absolutely continuous curve. Its metric derivative is

$$|\dot{x}_t| := \lim_{h\to 0} \frac{d(x_{t+h}, x_t)}{|h|}$$

for a.e. t. Equivalently, $|\dot{x}_t|$ is the smallest $g \in L^1_{\mathrm{loc}}(0,T)$ such that $d(x_s, x_t) \leqslant \int_s^t g(r)\mathrm{d}r$ for all $0 < s < t < T$. If $f : \mathcal{X} \to (-\infty, +\infty]$ is lower semicontinuous, its local metric slope at a point x with $f(x) < +\infty$ is

$$|\partial f|(x) := \limsup_{\substack{y\to x\\ y\neq x}} \frac{(f(x) - f(y))_+}{d(x,y)}.$$

When this slope is a strong upper gradient for f, meaning that $|f(y_t) - f(y_s)| \leqslant \int_s^t |\partial f|(y_r)|\dot{y}_r|\mathrm{d}r$ along absolutely continuous curves y_r, an absolutely continuous curve x_t is called a curve of maximal slope if it satisfies the energy-dissipation inequality

$$f(x_t) + \frac{1}{2}\int_s^t |\dot{x}_r|^2 \mathrm{d}r + \frac{1}{2}\int_s^t |\partial f|^2(x_r)\mathrm{d}r \leqslant f(x_s), \qquad 0 \leqslant s < t.$$

This is the metric-space formulation of gradient descent Ambrosio et al. (2006); the inequality is stable under weak limits, while in smooth Hilbert or Wasserstein settings it is usually saturated. In $\mathcal{X} = \mathcal{P}_2(\mathbb{R}^d)$ with $d = \mathcal{W}_2$, an absolutely continuous curve admits a continuity-equation velocity v_t and satisfies $|\dot{\alpha}_t|_{\mathcal{W}_2} \leqslant \|v_t\|_{L^2(\alpha_t)}$, with equality for the minimal velocity. For smooth energies, $|\partial f|(\alpha_t) = \|\nabla_{\mathcal{W}} f(\alpha_t)\|_{L^2(\alpha_t)}$, so the maximal-slope condition reduces to the velocity-field equation $v_t = -\nabla_{\mathcal{W}} f(\alpha_t)$ derived above. The energy consequences of this viewpoint are used below in Section 3.2.

We now detail examples of such Wasserstein gradient flows.

Example 3.3 (Application to single-cell gradient-flow models). Instead of only coupling successive snapshots, one may posit that a latent population law evolves by

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0, \qquad v_t = -\nabla_{\mathcal{W}} f(\alpha_t) = -\nabla \delta f(\alpha_t).$$

Here $\delta f(\alpha_t)$ is the first variation, with the convention introduced above. For example, if $\alpha = \rho\,\mathrm{d}x$, the energy $f(\alpha) = \int V \mathrm{d}\alpha + \sigma^2 \int \rho \log \rho\,\mathrm{d}x$ gives a drift toward a potential landscape V together with diffusion of strength σ^2. In single-cell modeling, V, interaction terms or a neural approximation of v_t can be fitted from unpaired population snapshots. This links Waddington-landscape intuition to the mathematical language of dynamic OT, Fokker–Planck equations and flow matching Tong et al. (2020); Lavenant et al. (2021); Klein et al. (2023).

Figure 3.1 shows the same minimizing-movement construction both in density coordinates and through the motion of quantiles.

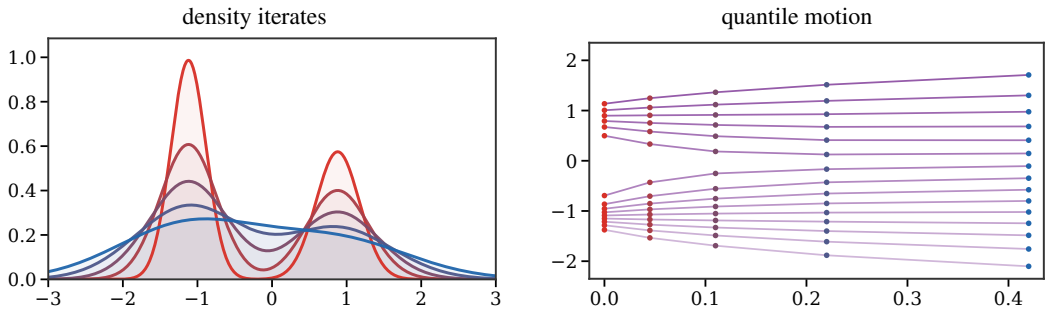


Figure 3.1: JKO minimizing movements for the entropy flow in one dimension. The left panel displays successive implicit-Euler minimizers for the heat equation, colored from red to blue. The right panel tracks inverse CDF values $Q_t(s) = F_t^{-1}(s)$ for selected probability levels s, giving a Lagrangian view of the proximal movement in Wasserstein space.

Discrete evolutions. If f admits a first variation whose spatial gradient is well defined at empirical measures, the characteristic form of (3.3) preserves the number and weights of the atoms. For equal weights, write $\alpha_t = \frac{1}{n}\sum_i \delta_{x_i(t)}$. The particles $X(t) := (x_i(t))_i$ evolve according to the coupled ODEs

$$\dot{x}_i(t) = -n \nabla_{x_i} F(X(t)), \tag{3.4}$$

where $F(X) := f\left(\frac{1}{n}\sum_i \delta_{x_i}\right)$ and the factor n comes from the empirical Wasserstein metric $\frac{1}{n}\sum_i \|\dot{x}_i\|^2$.

Linear functionals. The first benchmark is a functional whose first variation does not depend on the current measure.

Example 3.4 (Linear potentials generate independent particles). Take a linear functional

$$f(\alpha) = \int h(x) \mathrm{d}\alpha(x). \tag{3.5}$$

Then $\delta f(\alpha) = h$ is independent of α, and the flow (3.3) becomes

$$\frac{\partial \alpha_t}{\partial t} + \operatorname{div}(-\nabla h \alpha_t) = 0.$$

This implies particles move independently according to the usual gradient flow (3.2).

Shannon neg-entropy. Entropy gives the opposite benchmark: the flow is no longer a deterministic push-forward of particles, but a diffusion of density.

Example 3.5 (Entropy generates heat and porous-medium flows). The canonical density-dependent functional is the Shannon neg-entropy

$$f(\alpha) = \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x)\right) \mathrm{d}\alpha(x). \tag{3.6}$$

If $\alpha = \rho \mathrm{d}x$, then $\delta f(\alpha) = \log \rho + 1$, up to an immaterial additive constant, so $\nabla_{\mathcal{W}} f(\alpha) = \nabla \log \rho$ is the score.

The flow (3.3) becomes the heat equation

$$\partial_t \rho_t = \Delta \rho_t .$$

Other entropy functionals lead to nonlinear diffusion equations; finite-volume and particle discretizations are discussed in Carrillo et al. (2015); Gianazza et al. (2009); Maas (2011); Erbar (2010). For a generalized entropy

$$f(\rho \mathrm{d}x) = \int g(\rho(x))\mathrm{d}x, \tag{3.7}$$

with a scalar convex function g, one obtains nonlinear diffusions in the smooth-density regime:

$$\frac{\partial \rho_t}{\partial t} = \Delta(P(\rho_t)),$$

where the pressure P satisfies $P'(s) = sg''(s)$. For example, $g(s) = s\log(s)$ gives $P(s) = s$ and recovers (3.6), while $g(s) = s^m/(m-1)$, $m > 1$, gives $P(s) = s^m$ up to an additive constant and yields the porous-medium equation.

Remark 3.6 (Two gradient-flow interpretations of the heat equation). The heat equation already illustrates that a PDE does not determine a unique gradient-flow structure. Write $\alpha = \rho \mathrm{d}x$ on a flat domain, with periodic or no-flux boundary conditions. In the Hilbert geometry of densities, the squared distance and the Dirichlet energy are

$$d^2_{L^2}(\alpha,\beta) = \int |\rho_\alpha(x) - \rho_\beta(x)|^2 \mathrm{d}x, \qquad \mathcal{D}(\alpha) = \frac{1}{2}\int \|\nabla \rho(x)\|^2 \mathrm{d}x,$$

so $\delta\mathcal{D}/\delta\rho = -\Delta\rho$ and the L^2 gradient flow $\partial_t\rho_t = -\delta\mathcal{D}/\delta\rho(\rho_t)$ gives $\partial_t\rho_t = \Delta\rho_t$. This viewpoint says that heat flow decreases oscillations and regularizes the density. In Wasserstein geometry, the same equation is instead the $\mathcal{W}_2$ gradient flow of the Shannon entropy $\int \rho\log\rho \mathrm{d}x$, so the driving mechanism is entropic spreading. The example is a useful warning: explaining a dynamics as a gradient flow requires specifying both an energy and a metric, and different pairs (f, d) may produce the same PDE.

Figure 3.2 contrasts the spreading induced by the linear heat equation with the compactly supported, density-dependent propagation of two porous-medium flows.

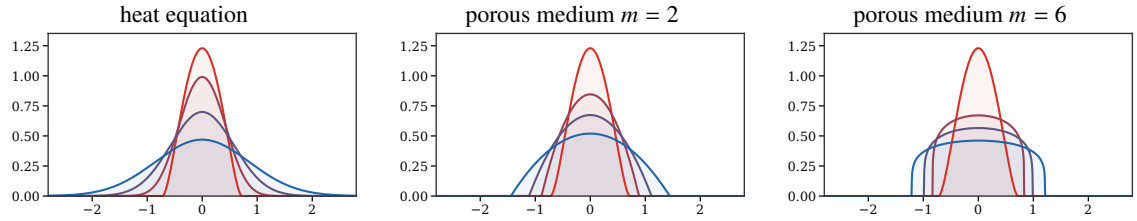


Figure 3.2: Entropy-driven Wasserstein gradient flows from the same compact initial density. The heat flow is generated by Shannon entropy $g(\rho) = \rho\log\rho$ and instantly develops Gaussian tails. The porous-medium flows use the power entropy $g(\rho) = \rho^m/(m-1)$, hence $\partial_t\rho = \Delta(\rho^m)$: the middle panel has $m = 2$, while the right panel has the stronger nonlinearity $m = 6$, i.e. $\partial_t\rho = \Delta(\rho^6)$. Larger powers diffuse mainly where the density is high, producing a flatter core and a sharper compact free boundary.

Interaction energies. In a similar spirit, to obtain nonlinear evolutions without requiring the measure to have density, one can consider

$$f(\alpha) := \iint k(x,y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y). \tag{3.8}$$

For a symmetric kernel k:

$$\delta f(\alpha)(x) = 2\int k(x,y)\mathrm{d}\alpha(y), \quad \nabla_{\mathcal{W}} f(\alpha)(x) = 2\int \nabla_x k(x,y)\mathrm{d}\alpha(y).$$

For $\alpha_0 = \frac{1}{n}\sum_i \delta_{x_i}$, the flow (3.3) implies particles $(x_i(t))_i$ obey:

$$\dot{x}_i(t) = -\frac{2}{n}\sum_j \nabla k(x_i(t), x_j(t)).$$

If k is positive definite, or more generally conditionally positive definite on signed measures of zero total mass as for the energy-distance kernel $k(x, y) = -\|x - y\|$, and one minimizes the squared kernel discrepancy to a teacher distribution β, then

$$\|\alpha - \beta\|_k^2 = \iint k \mathrm{d}\alpha \mathrm{d}\alpha - 2 \int \left(\int k(x, y) \mathrm{d}\beta(y) \right) \mathrm{d}\alpha(x) + \text{constant}.$$

Thus MMD-type training energies are exactly an interaction energy plus a linear potential; the teacher distribution appears through the potential $x \mapsto -2 \int k(x, y) \mathrm{d}\beta(y)$. The corresponding empirical Wasserstein gradient flow is

$$\dot{x}_i(t) = -\frac{2}{n} \sum_j \nabla_x k(x_i(t), x_j(t)) + 2 \int \nabla_x k(x_i(t), y) \mathrm{d}\beta(y).$$

The first term is a kernelized self-interaction, while the second is the attraction induced by the continuous teacher kernel mean. At the continuum level, characteristic positive-definite kernels, and the Euclidean energy-distance kernel on probability measures, have β as the unique minimizer of $\|\alpha - \beta\|_k^2$. For a finite number of particles, however, the flow can only form a kernelized quadrature of β, and small particle systems may cover the target modes poorly. Figure 3.3 illustrates this finite-particle effect.

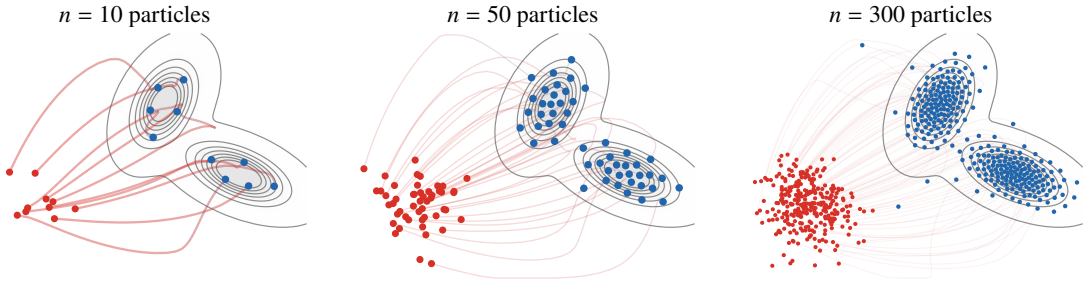


Figure 3.3: Particle count in an energy-distance Wasserstein particle flow. The squared MMD-type discrepancy uses the energy-distance kernel $k(x, y) = -\|x - y\|$ and targets a smooth two-Gaussian teacher distribution. The teacher itself is shown only through true density contours, while red dots are a compact shifted Gaussian initialization placed away from the target, red-to-blue curves show a thinned subset of particle trajectories, and blue dots show the stabilized long-time particles. With too few particles, the empirical measure forms a sparse kernelized quadrature and may under-cover the target modes; increasing n makes the particle cloud approximate the continuous target geometry more faithfully.

The preceding particle flow minimizes a squared kernel discrepancy. Figure 3.4 contrasts this with flows generated by three discrepancies toward the same target. For the energy distance and $\mathcal{W}_2$, the energy is the unsquared distance, so the Wasserstein velocity is normalized by the current discrepancy value away from the target. The KL case is different: it is a relative-entropy gradient flow and therefore follows the Fokker–Planck diffusion toward the target. The comparison keeps the endpoint fixed and makes the geometry visible through the transient densities.

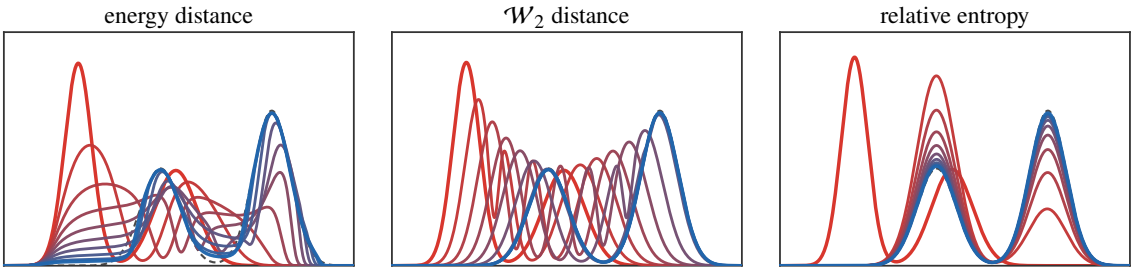


Figure 3.4: Wasserstein gradient flows of three discrepancies toward the same one-dimensional Gaussian-mixture target, shown as a dashed black density. The source density is also a Gaussian mixture. Red-to-blue curves are snapshots of α_t. The energy-distance and $\mathcal{W}_2$ panels descend the unsquared distances and use many equal-weight quantile particles: in one dimension the energy distance is obtained from $\mathrm{ED}^2 = 2 \int |F_\alpha - F_\beta|^2 \mathrm{d}x$, while $\mathcal{W}_2$ uses monotone quantile matching. The relative-entropy panel solves the reversible Fokker–Planck equation $\partial_t \rho = \partial_x(\rho_\beta \, \partial_x(\rho/\rho_\beta))$ by an implicit finite-volume scheme, so that ρ_β is the discrete stationary density.

Figures 3.5 and 3.6 complement this target-matching example: the first isolates three self-interaction regimes, while the second compares the trajectory geometries produced by several discrepancies.

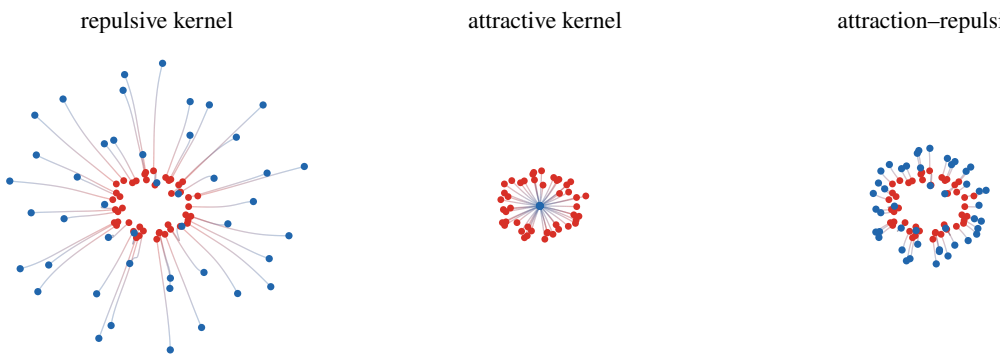


Figure 3.5: Interaction-energy particle flows for three choices of k. A positive Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ produces short-range repulsion under Wasserstein descent; changing its sign produces attraction and collapse; adding a quadratic long-range attraction to the repulsive kernel yields a balanced attraction–repulsion dynamics. The curves use arclength-based red-to-blue coloring along a longer integration of the coupled particle ODE (3.4).

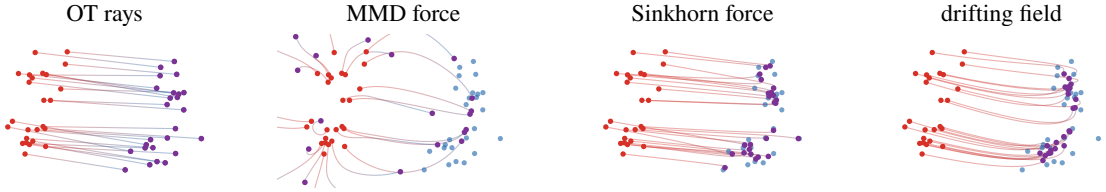


Figure 3.6: Particle trajectories induced by different discrepancy geometries. The red particles and blue target cloud are the same in all panels. Straight OT displacement produces rays from an optimal matching; an MMD-type witness field gives smoother nonlocal forces; the Sinkhorn-divergence force is an entropic, debiased transport attraction; and the normalized drifting field combines attraction to data with self-repulsion. The figure is qualitative: it compares geometric behavior, not solver performance.

Stochastic particles and McKean–Vlasov limits. More generally, deterministic particle flows have stochastic counterparts, where Brownian noise at the particle level becomes an entropy term at the level of measures. If the drift does not depend on the empirical measure, each particle evolves independently according to

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_t,$$

and the one-particle law $\alpha_t = \rho_t \mathrm{d}x$ directly satisfies the linear Fokker–Planck equation

$$\partial_t \rho_t = -\operatorname{div}(b\rho_t) + \sigma^2 \Delta \rho_t.$$

Example 3.7 (Langevin drift as a free-energy flow). If $b = -\nabla V$, this linear Fokker–Planck equation is the $\mathcal{W}_2$ gradient flow of the free energy

$$\rho \mapsto \int V\rho \,\mathrm{d}x + \sigma^2 \int \rho \log \rho \,\mathrm{d}x.$$

The mean-field case is different: the drift is recomputed from the current empirical distribution of all particles,

$$\mathrm{d}X_i^n(t) = b(X_i^n(t), \alpha_t^n)\mathrm{d}t + \sqrt{2}\sigma \mathrm{d}B_i(t), \qquad \alpha_t^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^n(t)}.$$

For finite n, the empirical law α_t^n is itself random. Under suitable Lipschitz, growth and chaotic-initialization assumptions, propagation of chaos states that finitely many particles become asymptotically independent as $n \to \infty$, all with the same deterministic law $\rho_t \mathrm{d}x$; equivalently, the empirical measure α_t^n converges in probability to this law. The limiting density solves the nonlinear Fokker–Planck, or McKean–Vlasov, equation

$$\partial_t \rho_t = -\operatorname{div}\left(b(x,\rho_t)\rho_t\right) + \sigma^2 \Delta \rho_t.$$

When the interaction drift has variational form

$$b(x,\rho) = -\nabla \frac{\delta \mathcal{E}}{\delta \rho}(x),$$

this PDE is the Wasserstein gradient flow of the entropy-regularized energy

$$\mathcal{E}(\rho) + \sigma^2 \int \rho \log \rho \, \mathrm{d}x.$$

For a prescribed target $\beta = \rho_\beta \mathrm{d}x$, take $b = \sigma^2 \nabla \log \rho_\beta$. The SDE, the Fokker–Planck equation, and the deterministic continuity equation with score velocity are then three representations of the same gradient flow:

$$\partial_t \rho_t = \sigma^2 \operatorname{div}\left(\rho_t \nabla \log \frac{\rho_t}{\rho_\beta}\right), \qquad v_t = \sigma^2 \left(\nabla \log \rho_\beta - \nabla \log \rho_t\right). \tag{3.9}$$

Figure 3.7 compares numerical realizations of precisely these three descriptions.

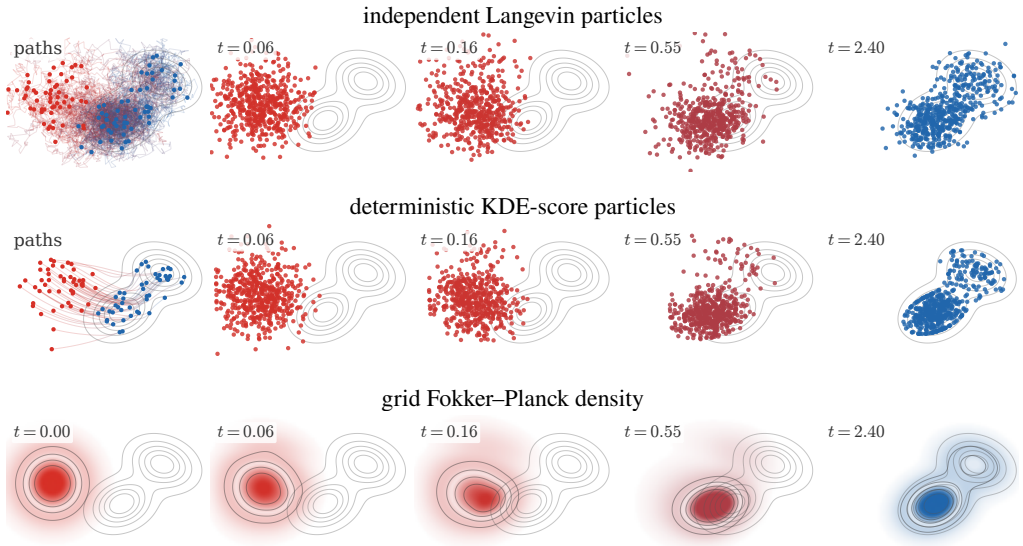


Figure 3.7: Three numerical representations of an entropy-regularized Wasserstein gradient flow. The target β is a two-Gaussian density shifted to the right of an initially isotropic Gaussian density. The first row simulates independent Langevin particles and displays a thinned set of trajectories in the left panel. The second row evolves many deterministic particles with the score velocity in (3.9), estimating $\nabla \log \rho_t$ by a kernel-density estimator; only representative trajectories and particle subsets are displayed. The third row solves the corresponding Fokker–Planck equation on a grid, starting from the initial density in the left panel. The remaining columns use front-loaded times, so that the onset of the flow and the later deformation toward a bimodal law are both visible.

Gradient flows under density constraints. The same minimizing-movement formalism also handles nonsmooth energies. A prototypical example is a linear energy, generated by a potential, with a hard upper bound on the density,

$$f_\kappa(\rho \mathrm{d}x) = \int_\Omega h(x)\rho(x)\mathrm{d}x + \iota_{\mathcal{K}_\kappa}(\rho \mathrm{d}x), \qquad \mathcal{K}_\kappa = \{\rho \mathrm{d}x \in \mathcal{P}_2(\Omega) \; ; \; 0 \leqslant \rho \leqslant \kappa \text{ a.e.}\}, \tag{3.10}$$

where $\Omega \subset \mathbb{R}^d$ is a convex domain and $\mathcal{K}_\kappa$ is assumed nonempty; if Ω has finite volume, this requires $\kappa|\Omega| \geqslant 1$. The indicator term has no classical first variation; it contributes the normal cone to $\mathcal{K}_\kappa$ in the Wasserstein first-order condition. The JKO step becomes

$$\alpha^{k+1} \in \operatorname*{argmin}_{\alpha \in \mathcal{K}_\kappa} \frac{1}{2\tau} \mathcal{W}_2^2(\alpha^k, \alpha) + \int_\Omega h \, \mathrm{d}\alpha. \tag{3.11}$$

This is the mathematical mechanism behind macroscopic crowd-motion models with congestion, where the desired velocity $-\nabla h$ is projected onto velocities that do not increase density beyond the maximal value κ Maury et al. (2010); Santambrogio (2018).

The hard cap is not especially benign numerically: the projection in the JKO step and the pressure field below can be difficult to compute accurately. Its advantage is geometric. Proposition 3.14 in Section 3.2 shows that $\mathcal{K}_\kappa$ is geodesically convex, so the indicator of the constraint is harmless for the usual geodesic-convex convergence guarantees.

The formal PDE contains a pressure field p_t, the Lagrange multiplier of the constraint. With no-flux boundary condition $\rho_t \nabla(h + p_t) \cdot n = 0$ on $\partial\Omega$, the constrained gradient flow is the complementarity system

$$\partial_t \rho_t = \operatorname{div}(\rho_t \nabla(h + p_t)), \qquad 0 \leqslant \rho_t \leqslant \kappa, \qquad p_t \geqslant 0, \qquad p_t(\kappa - \rho_t) = 0. \tag{3.12}$$

The pressure vanishes in the unsaturated region and acts only where $\rho_t = \kappa$, pushing mass away from congested zones so that the density cap remains satisfied.

Figure 3.8 shows how decreasing the cap converts free concentration into a saturated region whose width is fixed by mass conservation.

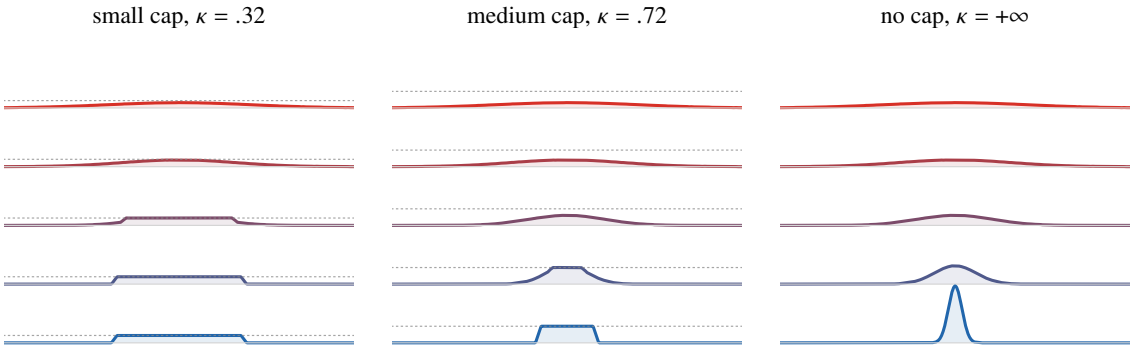


Figure 3.8: Density-constrained gradient flow under a quadratic attraction. The one-dimensional flow is computed in quantile variables by explicit Lagrangian descent followed by the isotonic projection enforcing $\partial_s q \geqslant 1/\kappa$. Each panel stacks five times vertically, colored from red to blue, using the same vertical density scale in all three panels. The dashed gray line marks the maximal density κ for the constrained runs. A small cap forces the mass to form a wide saturated block, a medium cap allows a narrower congested region, and the unconstrained flow concentrates freely near the minimizer of h.

Multi-species gradient flows. Several applications involve several densities living on the same physical space: chemical species, color channels, cell populations or competing phases. Let $\Omega \subset \mathbb{R}^d$ be the physical domain and assume that the component masses $\mathbf{m} = (m_1, \dots, m_p)$ are positive and fixed. A convenient notation is

$$\mathcal{P}_{2,\mathbf{m}}(\Omega; \mathbb{R}_+^p) = \left\{ \alpha = (\alpha_1, \dots, \alpha_p) \;;\; \alpha_i \in \mathcal{M}_+(\Omega),\ \alpha_i(\Omega) = m_i,\ \int_\Omega \|x\|^2 \mathrm{d}\alpha_i(x) < +\infty \right\}. \tag{3.13}$$

Since the components are finite measures rather than necessarily probabilities, set

$$\mathcal{W}_{2,\oplus}^2(\alpha, \eta) = \sum_{i=1}^p m_i \, \mathcal{W}_2^2\left(\frac{\alpha_i}{m_i}, \frac{\eta_i}{m_i}\right). \tag{3.14}$$

This is the diagonal, independent-channel geometry; cross-effects may still enter through the energy or through constraints. In a more general vector-valued dynamic transport language, this is the special case of a diagonal mobility. The corresponding JKO step is

$$\alpha^{k+1} \in \operatorname*{argmin}_{\alpha \in \mathcal{P}_{2,\mathbf{m}}(\Omega; \mathbb{R}_+^p)} \frac{1}{2\tau} \mathcal{W}_{2,\oplus}^2(\alpha^k, \alpha) + f(\alpha). \tag{3.15}$$

For smooth densities and first variations $\varphi_i = \delta f / \delta \rho_i$, this yields

$$\partial_t \rho_i = \operatorname{div}(\rho_i \nabla \varphi_i), \qquad \varphi_i = \frac{\delta f}{\delta \rho_i}, \qquad i = 1, \dots, p. \tag{3.16}$$

The equations are independent only if f is separable. Otherwise the coupling enters through the first variations φ_i, not through the metric tensor. Carlier, Chizat and Laborde Carlier et al. (2024) use displacement smoothness of entropic OT to prove well-posedness of Wasserstein gradient flows that include multi-species systems. Congestion and shared-density variants are closely related to crowd and traffic models Maury et al. (2010); Carlier et al. (2008); Santambrogio (2018).

A second useful model imposes a pointwise composition constraint. Given a fixed reference density $\beta = b\mathrm{d}x$ with total mass $\sum_i m_i$, set

$$C_\beta = \left\{ \alpha \in \mathcal{P}_{2,\mathbf{m}}(\Omega; \mathbb{R}_+^p) \;;\; \sum_{i=1}^p \alpha_i = \beta \right\}. \tag{3.17}$$

The constrained JKO step is

$$\alpha^{k+1} \in \operatorname*{argmin}_{\alpha \in C_\beta} \frac{1}{2\tau} \mathcal{W}^2_{2,\oplus}(\alpha^k, \alpha) + f(\alpha). \tag{3.18}$$

In the smooth-density regime, the formal constrained flow has a common pressure λ_t:

$$\partial_t \rho_i = \operatorname{div}\left(\rho_i \nabla(\varphi_i - \lambda_t)\right), \qquad \operatorname{div}(b\nabla\lambda_t) = \operatorname{div}\left(\sum_{i=1}^{p} \rho_i \nabla\varphi_i\right). \tag{3.19}$$

For the separable Shannon entropy $f(\alpha) = \sum_i \int_\Omega \rho_i \log \rho_i \mathrm{d}x$, this becomes

$$\partial_t \rho_i = \Delta\rho_i - \operatorname{div}(\rho_i \nabla\lambda_t), \qquad \operatorname{div}(b\nabla\lambda_t) = \Delta b. \tag{3.20}$$

When b is constant, the pressure can be chosen constant and the system reduces to independent heat equations which nevertheless preserve the pointwise sum $\sum_i \rho_i = b$. This product geometry should be contrasted with vector-valued Benamou–Brenier distances, where the metric itself can couple the components through a non-diagonal mobility.

Figure 3.9 visualizes this constant-total-density case for two, three, and five species.

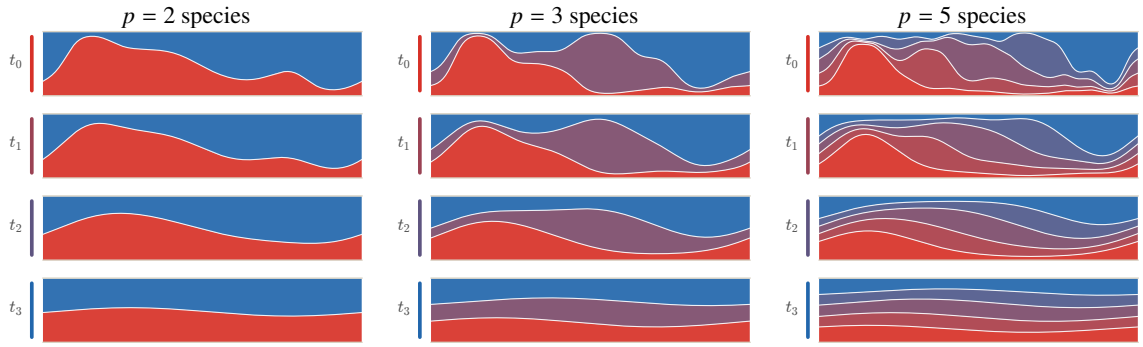


Figure 3.9: Multi-species entropy flow under the shared-density constraint $\sum_i \rho_i = 1$ on the periodic interval. Each row is a time snapshot of the stacked densities, with the species shown as colored bands and time increasing from top to bottom. For a constant total density, the pressure in (3.20) is constant, so each component follows the heat equation while the sum of the bands remains exactly flat.

3.2. Geodesic Convexity and Convergence

Geodesic convexity is the convexity notion adapted to Wasserstein geometry. It is the condition that turns the formal gradient-flow calculus into a convergence theory.

Geodesics and convexity. A constant-speed $\mathcal{W}_2$ geodesic between α_0 and α_1 is obtained, as in Definition 1.15, from any optimal coupling $\pi^\star \in \mathcal{U}(\alpha_0, \alpha_1)$ by the McCann interpolation

$$\alpha_t = ((1-t)P_0 + tP_1)_\sharp \pi^\star, \qquad t \in [0,1],$$

where $P_0(x,y) = x$ and $P_1(x,y) = y$. If the optimal plan is induced by a Brenier map T, this reduces to $((1-t)\mathrm{Id} + tT)_\sharp \alpha_0$. The coupling formula is important because geodesics exist even when no Monge map exists, for instance when a Dirac mass must split.

Definition 3.8 (Geodesic convexity)**.** A functional f on $\mathcal{P}_2(\mathbb{R}^d)$ is geodesically convex if for every $\mathcal{W}_2$ geodesic $(\alpha_t)_t$,

$$f(\alpha_t) \leqslant (1-t) f(\alpha_0) + t f(\alpha_1).$$

It is λ-geodesically convex if the right-hand side is improved by $-\frac{\lambda}{2} t(1-t)\, \mathcal{W}_2^2(\alpha_0, \alpha_1)$.

Proposition 3.9 (Geodesic convexity of linear and quadratic energies)**.** *The following formal statements hold on $\mathcal{P}_2(\mathbb{R}^d)$.*

1. *If h is convex, then the linear energy $\alpha \mapsto \int h \mathrm{d}\alpha$ is geodesically convex; if h is λ-strongly convex, it is λ-geodesically convex. Conversely, if this linear energy is geodesically convex for all Dirac endpoints, then h is convex.*

2. *More generally, let $k \geqslant 2$ and let $H_k : (\mathbb{R}^d)^k \to \mathbb{R} \cup \{+\infty\}$ be convex. Then the polynomial energy*

$$\alpha \mapsto \int_{(\mathbb{R}^d)^k} H_k(x_1, \dots, x_k)\, \mathrm{d}\alpha(x_1) \cdots \mathrm{d}\alpha(x_k)$$

is geodesically convex whenever the integral is well defined. If H_k is λ-strongly convex on the product space, this energy is $k\lambda$-geodesically convex. In particular, the quadratic interaction

$$\alpha \mapsto \frac{1}{2} \iint W(x, y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$$

is geodesically convex when W is convex on $\mathbb{R}^d \times \mathbb{R}^d$, and it is λ-geodesically convex when W is λ-strongly convex there.

Proof. Let π be an optimal coupling between α_0 and α_1, and set $X_t = (1-t)X_0 + tX_1$ under $(X_0, X_1) \sim \pi$. Convexity of h gives $h(X_t) \leqslant (1-t)h(X_0) + th(X_1)$, and strong convexity gives the additional term $-\frac{\lambda}{2}t(1-t)\|X_0 - X_1\|^2$. Integrating proves the first claim. Necessity follows by testing on Dirac masses: the geodesic between δ_x and δ_y is $\delta_{(1-t)x+ty}$, so geodesic convexity of $\int h\mathrm{d}\alpha$ gives the usual convexity inequality for h.

For the polynomial statement, take k independent copies $(X_0^\ell, X_1^\ell)_{\ell=1}^k$ of the same optimal coupling and define $X_t^\ell = (1-t)X_0^\ell + tX_1^\ell$. Convexity of H_k on the product space gives

$$H_k(X_t^1, \dots, X_t^k) \leqslant (1-t)H_k(X_0^1, \dots, X_0^k) + tH_k(X_1^1, \dots, X_1^k).$$

Integrating over the product coupling proves the claim. If H_k is λ-strongly convex, the additional term is

$$-\frac{\lambda}{2}t(1-t)\sum_{\ell=1}^k \mathbb{E}\|X_0^\ell - X_1^\ell\|^2 = -\frac{k\lambda}{2}t(1-t)\,\mathcal{W}_2^2(\alpha_0, \alpha_1),$$

which proves the stated constant. For the quadratic interaction, the prefactor $1/2$ cancels the factor $k = 2$, leaving the constant λ. □

The polynomial criterion is useful but restrictive for kernel losses. MMD kernels are usually chosen for positive definiteness and statistical smoothing, not for convexity as functions of (x, y). Gaussian and Laplace kernels, for instance, are not convex on $\mathbb{R}^d \times \mathbb{R}^d$, so positive definiteness of k does not imply that $\alpha \mapsto \mathrm{MMD}_k^2(\alpha, \beta)$ is geodesically convex. In fact, such MMD losses are not geodesically convex in general. One-dimensional monotone transport gives a sharper exception for distance-type kernels, because ordered quantiles keep the sign of pairwise differences fixed along geodesics; this is the one-dimensional displacement-convexity mechanism emphasized, for instance, in Santambrogio (2015).

Proposition 3.10 (One-dimensional interactions and energy-distance MMD)**.** *Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be continuous with at most quadratic growth, and define on $\mathcal{P}_2(\mathbb{R})$*

$$\mathcal{I}_\varphi(\alpha) \coloneqq \frac{1}{2} \iint \varphi(|x-y|)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y).$$

Then $\mathcal{I}_\varphi$ is geodesically convex for $\mathcal{W}_2$ on $\mathcal{P}_2(\mathbb{R})$ if and only if φ is convex on $\mathbb{R}_+$.

Fix also $\beta \in \mathcal{P}_2(\mathbb{R})$. For the conditionally positive kernel $k(x, y) = -|x-y|$, the squared MMD, equivalently the squared energy distance,

$$\mathcal{E}_\beta(\alpha) \coloneqq \mathrm{MMD}_k^2(\alpha, \beta) = \iint -|x-y|\,\mathrm{d}(\alpha-\beta)(x)\mathrm{d}(\alpha-\beta)(y),$$

is geodesically convex as a function of α.

Proof. Let Q_0, Q_1 be the quantile functions of α_0, α_1, and let $Q_t = (1-t)Q_0 + tQ_1$ be the quantile function of the one-dimensional $\mathcal{W}_2$ geodesic α_t. For $r > s$, monotonicity gives $Q_i(r) - Q_i(s) \geqslant 0$ for $i = 0, 1$, and hence

$$Q_t(r) - Q_t(s) = (1-t)(Q_0(r) - Q_0(s)) + t(Q_1(r) - Q_1(s)).$$

Using the symmetry of the double integral and the quantile parametrization, one can write

$$\mathcal{I}_\varphi(\alpha_t) = \int_0^1 \int_0^r \varphi(Q_t(r) - Q_t(s))\,\mathrm{d}s\,\mathrm{d}r.$$

Convexity of φ on $\mathbb{R}_+$ gives the desired convexity after integration.

Conversely, test geodesic convexity on two-point measures $\alpha_i = \frac{1}{2}(\delta_0 + \delta_{a_i})$, with $a_i \geqslant 0$. Their monotone geodesic is $\alpha_t = \frac{1}{2}(\delta_0 + \delta_{(1-t)a_0+ta_1})$. Since

$$\mathcal{I}_\varphi(\alpha_t) = \frac{1}{4}\varphi(0) + \frac{1}{4}\varphi((1-t)a_0 + ta_1),$$

and the identical $\frac{1}{4}\varphi(0)$ terms cancel from the geodesic-convexity inequality, one obtains the ordinary convexity inequality for φ on $\mathbb{R}_+$.

For the energy distance, the function $\varphi(r) = -r$ is affine on $\mathbb{R}_+$, so the self-interaction $-\iint |x-y|\mathrm{d}\alpha(x)\mathrm{d}\alpha(y)$ is affine along one-dimensional Wasserstein geodesics even though $z \mapsto -|z|$ is not convex on $\mathbb{R}$. Write Q_β for the quantile function of β. Expanding the MMD and dropping the term depending only on β gives

$$\mathcal{E}_\beta(\alpha) = 2\int_0^1\int_0^1 |Q_\alpha(r) - Q_\beta(s)|\,\mathrm{d}r\,\mathrm{d}s - \int_0^1\int_0^1 |Q_\alpha(r) - Q_\alpha(s)|\,\mathrm{d}r\,\mathrm{d}s + \text{constant}.$$

Here the constant is independent of α. The positive cross-distance term is convex in Q_α, since the absolute value is convex and Q_t is affine in t. The self-distance term is affine along quantile geodesics because

$$\int_0^1\int_0^1 |Q_t(r) - Q_t(s)|\,\mathrm{d}r\,\mathrm{d}s = 2\int_0^1\int_0^r (Q_t(r) - Q_t(s))\,\mathrm{d}s\,\mathrm{d}r.$$

Thus $\alpha \mapsto \mathcal{E}_\beta(\alpha)$ is geodesically convex. □

This one-dimensional result is the mechanism behind the favorable behavior of energy-distance particle flows discussed earlier in this chapter and in recent analyses of distance-kernel MMD flows Duong et al. (2024). It should not be read as a general MMD statement: for a typical positive definite kernel k, the decomposition

$$\mathrm{MMD}_k^2(\alpha,\beta) = \iint k\,\mathrm{d}\alpha\mathrm{d}\alpha - 2\iint k\,\mathrm{d}\alpha\mathrm{d}\beta + \text{constant}$$

contains neither a convex self-interaction nor a convex cross term along Wasserstein geodesics.

Internal energies require a different mechanism: convexity is hidden in the Jacobian determinant of the transport map, not in a pointwise convex integrand along particles. The precise criterion is McCann's displacement-convexity theorem McCann (1997).

Theorem 3.11 (McCann displacement convexity for internal energies)**.** *Let $g : [0,+\infty) \to \mathbb{R} \cup \{+\infty\}$ be convex with $g(0) = 0$, and define*

$$\psi(r) = r^d g(r^{-d}), \qquad r > 0.$$

If ψ is convex and nonincreasing on $(0,+\infty)$, then the internal energy

$$\mathcal{U}_g(\alpha) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} g(\rho(x))\,\mathrm{d}x, & \alpha = \rho\mathrm{d}x, \\ +\infty, & \textit{otherwise}, \end{cases}$$

is geodesically convex on $\mathcal{P}_2(\mathbb{R}^d)$.

Proof. It is enough to prove the inequality for smooth positive compactly supported densities; the general case follows by approximation and lower semicontinuity. Let $T = \nabla\varphi$ be the Brenier map from $\alpha_0 = \rho_0\mathrm{d}x$ to $\alpha_1 = \rho_1\mathrm{d}x$, set $T_t = (1-t)\mathrm{Id} + tT$, and denote

$$J_t(x) = \det\big((1-t)I + t\nabla T(x)\big).$$

The interpolation satisfies $\alpha_t = (T_t)_\sharp\alpha_0$ and

$$\rho_t(T_t(x))J_t(x) = \rho_0(x).$$

By concavity of $\det^{1/d}$ on positive semidefinite matrices,

$$J_t(x)^{1/d} \geqslant (1-t) + t\det(\nabla T(x))^{1/d}.$$

Introduce $s_t(x) = (J_t(x)/\rho_0(x))^{1/d}$. Since $\rho_1(T(x))\det\nabla T(x) = \rho_0(x)$, this gives

$$s_t(x) \geqslant (1-t)\rho_0(x)^{-1/d} + t\rho_1(T(x))^{-1/d}.$$

Using the change of variables $y = T_t(x)$, one obtains

$$\int g(\rho_t(y))\,\mathrm{d}y = \int \rho_0(x)\,\psi(s_t(x))\,\mathrm{d}x.$$

Because ψ is nonincreasing and convex,

$$\psi(s_t(x)) \leqslant (1-t)\psi(\rho_0(x)^{-1/d}) + t\psi(\rho_1(T(x))^{-1/d}).$$

Multiplying by ρ_0 and integrating gives $\mathcal{U}_g(\alpha_t) \leqslant (1-t)\mathcal{U}_g(\alpha_0) + t\mathcal{U}_g(\alpha_1)$. □

The theorem applies to $g(s) = s\log s$, since then $\psi(r) = -d\log r$, to porous-medium powers $g(s) = s^m/(m-1)$, $m > 1$, and more generally to the fast-diffusion range $m \geqslant 1 - 1/d$ when the energy is well defined, with $m = 1$ understood as the entropy limit. The same criterion also clarifies which Csiszár divergences inherit displacement convexity from an internal-energy structure.

Proposition 3.12 (Geodesic convexity of power divergences)**.** *For $m > 0$, define*

$$\varphi_m(r) = \begin{cases} \dfrac{r^m - mr + m - 1}{m(m-1)}, & m \neq 1, \\ r\log r - r + 1, & m = 1. \end{cases}$$

Then the following geodesic-convexity statements hold.

1. *If $\Omega \subset \mathbb{R}^d$ is convex, $\beta = b\,\mathbb{1}_\Omega \mathrm{d}x$, and $m \geqslant 1 - 1/d$, then $\alpha \mapsto \mathcal{D}_{\varphi_m}(\alpha|\beta)$ is geodesically convex on $\mathcal{P}_2(\Omega)$.*
2. *If $\beta = Z^{-1}e^{-V}\mathrm{d}x$ on $\mathbb{R}^d$, V is convex, and $m \geqslant 1$, then $\alpha \mapsto \mathcal{D}_{\varphi_m}(\alpha|\beta)$ is geodesically convex on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$.*
3. *In the KL case $m = 1$, if V is λ-strongly convex, then $\alpha \mapsto \mathrm{KL}(\alpha|\beta) = \mathcal{D}_{\varphi_1}(\alpha|\beta)$ is λ-geodesically convex.*

All statements use the convention that the divergence is $+\infty$ when α is not absolutely continuous with respect to β.

Proof. For the flat reference case, write $\alpha = \rho\mathrm{d}x$. Up to affine terms in ρ, which only add constants on probability measures, the integrand $b\,\varphi_m(\rho/b)$ has non-affine part $b^{1-m}\rho^m/(m(m-1))$ for $m \neq 1$, and $\rho\log\rho$ for $m = 1$. For $m = 1$, McCann's criterion gives $\psi(r) = -d\log r$. For $m \neq 1$, the transformed non-affine part is, up to the irrelevant factor b^{1-m},

$$\psi_m(r) = \frac{r^{d(1-m)}}{m(m-1)}.$$

If $m > 1$, this function is convex and nonincreasing. If $0 < m < 1$, it is convex and nonincreasing exactly when $d(1-m) \leqslant 1$, i.e. $m \geqslant 1 - 1/d$. This proves the first claim by Theorem 3.11.

Assume next that $m > 1$ and V is convex. Up to constants,

$$\mathcal{D}_{\varphi_m}(\alpha|\beta) = \frac{Z^{m-1}}{m(m-1)} \int \rho(x)^m e^{(m-1)V(x)}\,\mathrm{d}x.$$

Let $T = \nabla\varphi$ be the Brenier map from $\alpha_0 = \rho_0\mathrm{d}x$ to α_1, set $T_t = (1-t)\mathrm{Id} + tT$, and write $J_t(x) = \det((1-t)I + t\nabla T(x))$. Along the geodesic $\alpha_t = (T_t)_\sharp\alpha_0$,

$$\int \rho_t^m e^{(m-1)V}\mathrm{d}x = \int \rho_0(x)^m \exp\big((m-1)\big(V(T_t(x)) - \log J_t(x)\big)\big)\,\mathrm{d}x.$$

For each x, the function $t \mapsto V(T_t(x))$ is convex, and $t \mapsto -\log J_t(x)$ is convex because $\log\det$ is concave on positive definite matrices. Since $m - 1 > 0$, the exponential of their sum is convex in t. Integrating gives geodesic convexity. The nonsmooth case follows by approximation.

Finally, for $m = 1$,

$$\mathrm{KL}(\alpha|\beta) = \int \rho\log\rho\,\mathrm{d}x + \int V\mathrm{d}\alpha + \log Z.$$

The entropy term is 0-geodesically convex by Theorem 3.11. The linear potential term is λ-geodesically convex when V is λ-strongly convex, by Proposition 3.9. Taking $\lambda = 0$ also covers the $m = 1$ case in the second claim. □

Remark 3.13 (Flat references and general φ-divergences). The Csiszár φ-divergences behave like internal energies only in special reference geometries. If $\beta = b\,\mathbb{1}_\Omega \mathrm{d}x$ is proportional to Lebesgue measure on a bounded convex domain, then

$$\mathcal{D}_\varphi(\alpha|\beta) = \int_\Omega b\,\varphi(\rho(x)/b)\,\mathrm{d}x, \qquad \alpha = \rho \mathrm{d}x,$$

so McCann's theorem applies whenever the integrand $g(s) = b\,\varphi(s/b)$ satisfies the displacement-convexity condition. For a nonuniform target β, this reduction is no longer available. Proposition 3.12 gives a useful structured extension for the power family: log-concave targets preserve geodesic convexity for $m \geqslant 1$, while the KL endpoint is special in the stronger sense that λ-strong log-concavity gives λ-geodesic convexity. Outside such structured families, a generic φ-divergence is not expected to be geodesically convex.

The Hellinger integrand $\varphi_H(r) = (\sqrt{r} - 1)^2$ illustrates why the qualifier "flat reference" matters. Since $\varphi_H = \varphi_{1/2}/2$, Proposition 3.12 shows that the corresponding flat-reference internal energy is geodesically convex in dimension $d = 1$. The same φ-divergence can nevertheless fail to be geodesically convex once the reference is nonuniform.

A concrete Gaussian test shows the obstruction. Let $\beta = \mathcal{N}(0, 1)$ on $\mathbb{R}$ and $\alpha_m = \mathcal{N}(m, 1)$. The family $m \mapsto \alpha_m$ is a $\mathcal{W}_2$-geodesic family, since the geodesic between α_{m_0} and α_{m_1} is $\alpha_{(1-t)m_0+tm_1}$. Writing

$$r_m(x) = \frac{\mathrm{d}\alpha_m}{\mathrm{d}\beta}(x) = \exp(mx - m^2/2),$$

one obtains

$$\mathcal{D}_{\varphi_H}(\alpha_m|\beta) = \int (\sqrt{r_m} - 1)^2\,\mathrm{d}\beta = 2\left(1 - e^{-m^2/8}\right).$$

Its second derivative is

$$\frac{\mathrm{d}^2}{\mathrm{d}m^2}\mathcal{D}_{\varphi_H}(\alpha_m|\beta) = \frac{1}{2}e^{-m^2/8}\left(1 - \frac{m^2}{4}\right),$$

which is negative for $|m| > 2$. Thus the Hellinger φ-divergence to a Gaussian target is not even 0-geodesically convex. More generally, Gaussian translations reduce the question for any φ to ordinary convexity of

$$m \mapsto \mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\varphi\left(e^{mX-m^2/2}\right)\right],$$

which need not hold. The KL integrand is special because this function is $m^2/2$, up to the usual normalization.

Geodesically convex constraints. Geodesic convexity is also useful for hard constraints. If a feasible set is geodesically convex, then adding its indicator to an energy does not destroy the variational structure used by minimizing movements and convergence arguments. The density cap from (3.10) is the model example: it is numerically delicate, because one must compute a projection or a pressure, but it is geometrically compatible with Wasserstein descent.

Proposition 3.14 (Geodesic convexity of density caps)**.** *Assume that $\Omega \subset \mathbb{R}^d$ is convex and let $\kappa > 0$. The set $\mathcal{K}_\kappa$ in (3.10) is geodesically convex in $\mathcal{P}_2(\Omega)$: if $\alpha_0, \alpha_1 \in \mathcal{K}_\kappa$, then the constant-speed $\mathcal{W}_2$ geodesic $(\alpha_t)_{t\in[0,1]}$ between them also belongs to $\mathcal{K}_\kappa$.*

Proof. Assume first that the endpoints have smooth positive densities $\alpha_i = \rho_i \mathrm{d}x$ with $\rho_i \leqslant \kappa$, and let $T = \nabla\varphi$ be the Brenier map from α_0 to α_1. Since Ω is convex, $T_t = (1 - t)\mathrm{Id} + tT$ maps Ω into Ω. The interpolation is $\alpha_t = (T_t)_\sharp\alpha_0$ and

$$\rho_t(T_t(x)) = \frac{\rho_0(x)}{\det((1 - t)I + t\nabla T(x))}.$$

The concavity of $\det^{1/d}$ on positive semidefinite matrices and the identity $\rho_1(T(x))\det\nabla T(x) = \rho_0(x)$ give

$$\rho_t(T_t(x))^{-1/d} \geqslant (1 - t)\rho_0(x)^{-1/d} + t\rho_1(T(x))^{-1/d} \geqslant \kappa^{-1/d}.$$

Hence $\rho_t \leqslant \kappa$ for all t. The general case follows by approximation and lower semicontinuity of the L^∞ bound, as in McCann's displacement-convexity theory McCann (1997). □

Example 3.15 (Squared Wasserstein distance need not be geodesically convex). The squared distance to a fixed measure is not geodesically convex in general. This is a sharp contrast with Hilbert spaces, where

$x \mapsto \|x-y\|^2$ is convex along line segments. In $\mathcal{P}_2(\mathbb{R}^d)$, $d \geqslant 2$, take

$$\beta = \frac{1}{2}\delta_{(-1/2,-1/2)} + \frac{1}{2}\delta_{(1/2,1/2)}, \quad \alpha_0 = \frac{1}{2}\delta_{(-1,0)} + \frac{1}{2}\delta_{(1,0)}, \quad \alpha_1 = \frac{1}{2}\delta_{(0,-1)} + \frac{1}{2}\delta_{(0,1)}.$$

One optimal coupling between α_0 and α_1 matches $(-1,0)$ with $(0,1)$ and $(1,0)$ with $(0,-1)$. The associated geodesic has midpoint

$$\alpha_{1/2} = \frac{1}{2}\delta_{(-1/2,1/2)} + \frac{1}{2}\delta_{(1/2,-1/2)}.$$

Then

$$\mathcal{W}_2^2(\alpha_0,\beta) = \mathcal{W}_2^2(\alpha_1,\beta) = \frac{1}{2}, \qquad \mathcal{W}_2^2(\alpha_{1/2},\beta) = 1,$$

which violates geodesic convexity. Thus objectives built from squared Wasserstein distances can be non-convex in the Wasserstein geometry itself. This is one reason why optimal quantization, which minimizes $\mathcal{W}_2^2(\alpha,\nu)$ over m-atomic measures ν, leads to non-convex algorithms such as Lloyd's method.

Remark 3.16 (Target discrepancies are usually semiconcave). The preceding example is a useful warning. Losses of the form $\alpha \mapsto D(\alpha,\beta)$, where β is a fixed target distribution, are rarely geodesically convex in the transport geometry. This is unfortunate for generative AI and generative modeling more broadly: training a distribution to match data by decreasing $\mathcal{W}_2^2(\alpha,\beta)$, $\mathrm{SW}_2^2(\alpha,\beta)$, or similar discrepancies should not be expected to inherit the convex behavior of Euclidean least squares.

The squared Wasserstein loss has the opposite one-sided curvature: it is 2-geodesically semiconcave along $\mathcal{W}_2$ geodesics. More precisely, if $(\alpha_t)_{t\in[0,1]}$ is a constant-speed $\mathcal{W}_2$ geodesic, then

$$\mathcal{W}_2^2(\alpha_t,\beta) \geqslant (1-t)\,\mathcal{W}_2^2(\alpha_0,\beta) + t\,\mathcal{W}_2^2(\alpha_1,\beta) - t(1-t)\,\mathcal{W}_2^2(\alpha_0,\alpha_1).$$

This is the standard 2-geodesic semiconcavity convention for squared distances, often informally called "2-concavity": the graph may fall below the chord, but only by the quadratic correction $t(1-t)\,\mathcal{W}_2^2(\alpha_0,\alpha_1)$. Equivalently, $-\mathcal{W}_2^2(\cdot,\beta)$ is (-2)-geodesically convex in the convention of Definition 3.8.

To prove the estimate, write $\alpha_t = ((1-t)P_0 + tP_1)_\sharp \pi_{01}$, where π_{01} is optimal between α_0 and α_1. Fix t, put $Z = (1-t)X_0 + tX_1$, take an optimal coupling between α_t and β, and glue it with a disintegration of π_{01} over Z. This gives random variables (X_0, X_1, Y) such that (Z,Y) is optimal between α_t and β, while (X_0,X_1) is optimal between α_0 and α_1. Since (X_i,Y) are admissible couplings between α_i and β,

$$\mathcal{W}_2^2(\alpha_i,\beta) \leqslant \mathbb{E}\|X_i - Y\|^2, \qquad i = 0,1.$$

The Euclidean identity

$$(1-t)\|X_0-Y\|^2 + t\|X_1-Y\|^2 = \|Z-Y\|^2 + t(1-t)\|X_0-X_1\|^2$$

then gives the claim after taking expectations.

The sliced loss has the same flavor, but one must keep track of which geometry is used. The projection of a $\mathcal{W}_2$ geodesic need not be the monotone one-dimensional geodesic between the projected endpoint measures. Fix a direction θ and project π_{01} onto the pair $(\langle\theta, X_0\rangle, \langle\theta, X_1\rangle)$. This projected endpoint coupling generates the projected curve, although it need not be optimal. At time t, couple the projected α_t optimally to the projected β, then glue this coupling with a disintegration of the projected endpoint coupling over $(1-t)\langle\theta, X_0\rangle + t\langle\theta, X_1\rangle$. The same Euclidean identity therefore applies, with correction cost $\mathbb{E}|\langle\theta, X_0 - X_1\rangle|^2$. Integrating in θ with the standard spherical normalization gives, along the original $\mathcal{W}_2$ geodesic,

$$\mathrm{SW}_2^2(\alpha_t,\beta) \geqslant (1-t)\,\mathrm{SW}_2^2(\alpha_0,\beta) + t\,\mathrm{SW}_2^2(\alpha_1,\beta) - \frac{t(1-t)}{d}\,\mathcal{W}_2^2(\alpha_0,\alpha_1),$$

because the projection cost satisfies $\int_{\mathbb{S}^{d-1}} |\langle\theta, x_0 - x_1\rangle|^2 \,\mathrm{d}\sigma(\theta) = \|x_0-x_1\|^2/d$ and π_{01} is $\mathcal{W}_2$-optimal. Thus this is a semiconcavity estimate for the sliced discrepancy along $\mathcal{W}_2$ geodesics, with a $\mathcal{W}_2$-controlled correction. If one instead works in the idealized Hilbert embedding by projected quantile functions, straight sliced chords satisfy the exact identity with the last term $\mathrm{SW}_2^2(\alpha_0,\alpha_1)$. The caveat is that such straight chords need not be realized by actual measures on $\mathbb{R}^d$.

Dissipation and finite metric length. Before asking for convergence rates, one should first extract the basic information already contained in energy dissipation. For the classical gradient flow of a smooth function

$h : \mathbb{R}^d \to \mathbb{R}$, $\dot{x}_t = -\nabla h(x_t)$, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}t} h(x_t) = -\|\nabla h(x_t)\|^2 = -\|\dot{x}_t\|^2.$$

Thus, for $0 \leqslant s < t$, Cauchy's inequality gives

$$\int_s^t \|\dot{x}_r\| \,\mathrm{d}r \leqslant \sqrt{t-s} \left(\int_s^t \|\dot{x}_r\|^2 \mathrm{d}r \right)^{1/2} = \sqrt{t-s} \left(h(x_s) - h(x_t) \right)^{1/2}. \tag{3.21}$$

The Wasserstein analogue uses the metric derivative $|\dot{\alpha}_t|_{\mathcal{W}_2}$ of an absolutely continuous curve. The following estimate isolates the only ingredient needed at this stage: energy dissipation controls the metric length of the path.

Proposition 3.17 (Finite metric length from dissipation)**.** *Let $(\alpha_r)_{r\in[0,T]}$ be an absolutely continuous curve in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$. Assume that, for all $0 \leqslant s < t \leqslant T$, it satisfies the exact energy-dissipation identity*

$$f(\alpha_s) - f(\alpha_t) = \int_s^t |\dot{\alpha}_r|^2_{\mathcal{W}_2} \mathrm{d}r.$$

Then

$$\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big) = \int_s^t |\dot{\alpha}_r|_{\mathcal{W}_2} \,\mathrm{d}r \leqslant \sqrt{t-s} \left(f(\alpha_s) - f(\alpha_t) \right)^{1/2}. \tag{3.22}$$

In the smooth Wasserstein gradient-flow setting with velocity $v_r = -\nabla_{\mathcal{W}} f(\alpha_r)$, the dissipation identity reads equivalently

$$f(\alpha_s) - f(\alpha_t) = \int_s^t |\dot{\alpha}_r|^2_{\mathcal{W}_2} \mathrm{d}r = \int_s^t \int \|v_r(x)\|^2 \mathrm{d}\alpha_r(x) \mathrm{d}r.$$

If one assumes only the metric energy-dissipation inequality for a curve of maximal slope, then the same length estimate holds with the right-hand side multiplied by $\sqrt{2}$.

Proof. Cauchy's inequality gives

$$\int_s^t |\dot{\alpha}_r|_{\mathcal{W}_2} \,\mathrm{d}r \leqslant \sqrt{t-s} \left(\int_s^t |\dot{\alpha}_r|^2_{\mathcal{W}_2} \mathrm{d}r \right)^{1/2}.$$

The exact energy-dissipation identity substitutes the last integral by $f(\alpha_s) - f(\alpha_t)$, which proves (3.22). In the metric theory of curves of maximal slope Ambrosio et al. (2006); Santambrogio (2015), the energy-dissipation inequality gives instead

$$\frac{1}{2} \int_s^t |\dot{\alpha}_r|^2_{\mathcal{W}_2} \mathrm{d}r \leqslant f(\alpha_s) - f(\alpha_t),$$

and the same argument yields the additional factor $\sqrt{2}$. □

Proposition 3.17 is weaker than a convergence rate: it does not identify a minimizer and it gives no decay law. It does, however, rule out escape to infinity in finite metric time whenever the energy is bounded below on finite time intervals. Indeed the trajectory segment remains in the $\mathcal{W}_2$-ball centered at α_s with radius given by (3.22). It also gives a concrete modulus of continuity. If $m_T \leqslant f(\alpha_r)$ for $0 \leqslant r \leqslant T$, then for $0 \leqslant s < t \leqslant T$,

$$\mathcal{W}_2(\alpha_s, \alpha_t) \leqslant \mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big) \leqslant \sqrt{f(\alpha_0) - m_T}\, \sqrt{t-s},$$

with an additional factor $\sqrt{2}$ under the energy-dissipation inequality. Thus an energy-dissipating trajectory is $1/2$-Hölder continuous as a curve in Wasserstein space on every time interval where the energy remains bounded from below. Non-coercivity of the objective means that sublevel sets need not be compact, but an energy-dissipating curve still cannot travel an infinite $\mathcal{W}_2$-distance in finite time without an infinite energy drop. The convex estimate below adds a first quantitative conclusion under geodesic convexity; the PL viewpoint then gives sharper coercive rates toward minimizers.

Proposition 3.18 (Energy decay for convex Wasserstein flows)**.** *Assume formally that f is geodesically convex, admits a smooth first variation, and has a minimizer $\alpha^\star$. Let $(\alpha_t)_t$ be a smooth solution of the Wasserstein gradient flow*

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t v_t) = 0, \qquad v_t = -\nabla_{\mathcal{W}} f(\alpha_t).$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\int \|\nabla_{\mathcal{W}} f(\alpha_t)(x)\|^2 \,\mathrm{d}\alpha_t(x) \leqslant 0.$$

If T_t is the optimal map from α_t to $\alpha^\star$, then

$$f(\alpha_t) - f(\alpha^\star) \leqslant -\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\mathcal{W}_2^2(\alpha_t, \alpha^\star),$$

and consequently

$$f(\alpha_t) - f(\alpha^\star) \leqslant \frac{\mathcal{W}_2^2(\alpha_0, \alpha^\star)}{2t}.$$

Proof. The chain rule and Proposition 3.2 give

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = \int \langle \nabla_{\mathcal{W}} f(\alpha_t)(x),\, v_t(x)\rangle \,\mathrm{d}\alpha_t(x) = -\int \|\nabla_{\mathcal{W}} f(\alpha_t)(x)\|^2 \,\mathrm{d}\alpha_t(x).$$

Geodesic convexity along the geodesic $((1-s)\mathrm{Id} + sT_t)_\sharp \alpha_t$ gives

$$f(\alpha^\star) - f(\alpha_t) \geqslant \int \langle \nabla_{\mathcal{W}} f(\alpha_t)(x),\, T_t(x) - x\rangle \,\mathrm{d}\alpha_t(x).$$

Since $v_t = -\nabla_{\mathcal{W}} f(\alpha_t)$, this reads

$$f(\alpha_t) - f(\alpha^\star) \leqslant \int \langle v_t(x),\, T_t(x) - x\rangle \,\mathrm{d}\alpha_t(x).$$

The standard first-variation formula for the squared Wasserstein distance gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\mathcal{W}_2^2(\alpha_t, \alpha^\star) = \int \langle x - T_t(x),\, v_t(x)\rangle \,\mathrm{d}\alpha_t(x),$$

which proves the differential inequality. Integrating it from 0 to t and using the monotonicity of $s \mapsto f(\alpha_s)$ gives

$$t\big(f(\alpha_t) - f(\alpha^\star)\big) \leqslant \int_0^t \big(f(\alpha_s) - f(\alpha^\star)\big)\,\mathrm{d}s \leqslant \frac{1}{2}\mathcal{W}_2^2(\alpha_0, \alpha^\star).$$

□

Convergence: the Wasserstein–PL viewpoint. In general, analyzing (3.3) is delicate. Geodesic convexity gives the familiar convex-gradient-flow picture, but rates are driven more directly by a first-order coercivity inequality. The relevant quantity is the squared Wasserstein slope

$$|\partial f|^2(\alpha) := \int \|\nabla_{\mathcal{W}} f(\alpha)(x)\|^2 \,\mathrm{d}\alpha(x)$$

in the smooth formal setting. Along a smooth gradient flow, this is both the squared metric speed and the energy-dissipation term. The terminology mirrors the finite-dimensional Polyak–Łojasiewicz condition introduced by Polyak Polyak (1963) and now standard in nonconvex optimization Karimi et al. (2016). It is also part of the broader family of Łojasiewicz and Kurdyka–Łojasiewicz gradient inequalities used to prove convergence of dissipative dynamics Attouch et al. (2010); Hauer and Mazón (2019); Dello Schiavo et al. (2024). The Wasserstein version simply replaces the Euclidean gradient norm by the metric slope associated with $\mathcal{W}_2$.

Definition 3.19 (Wasserstein–Polyak–Łojasiewicz inequality)**.** Let $f_\star := \inf_{\alpha\in\mathcal{P}_2(\mathbb{R}^d)} f(\alpha) > -\infty$. The functional f satisfies the Wasserstein–PL inequality with constant $\kappa > 0$ if

$$|\partial f|^2(\alpha) \geqslant 2\kappa\big(f(\alpha) - f_\star\big) \tag{3.23}$$

for every α in the domain of f. In the smooth Otto notation this reads

$$\int \|\nabla_{\mathcal{W}} f(\alpha)\|^2 \,\mathrm{d}\alpha \geqslant 2\kappa\big(f(\alpha) - f_\star\big).$$

Theorem 3.20 (Wasserstein–PL convergence)**.** *Assume that $f_\star > -\infty$, that f satisfies the Wasserstein–PL inequality (3.23) with constant $\kappa > 0$, and that $(\alpha_t)_{t\geqslant 0}$ is a global Wasserstein gradient flow satisfying the full energy-dissipation identity*

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -|\dot{\alpha}_t|^2 = -|\partial f|^2(\alpha_t) \qquad \text{for a.e. } t > 0. \tag{3.24}$$

Set $E(t) \coloneqq f(\alpha_t) - f_\star$. Then, for $0 \leqslant s \leqslant t$,

$$E(t) \leqslant e^{-2\kappa(t-s)} E(s),$$

and the length of the flow segment satisfies

$$\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big) \leqslant \sqrt{\frac{2}{\kappa}}\big(\sqrt{E(s)} - \sqrt{E(t)}\big).$$

If, in addition, f is lower semicontinuous for $\mathcal{W}_2$-convergence, then α_t converges in $\mathcal{W}_2$ to a minimizer $\alpha_\infty \in \operatorname{Argmin} f$, and

$$\mathcal{W}_2(\alpha_t, \alpha_\infty) \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(t)} \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(0)}\, e^{-\kappa t}.$$

In particular,

$$\mathrm{dist}_{\mathcal{W}_2}(\alpha_t, \operatorname{Argmin} f) \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(0)}\, e^{-\kappa t}.$$

Proof. The energy estimate is immediate from (3.24) and (3.23):

$$E'(t) = -|\partial f|^2(\alpha_t) \leqslant -2\kappa E(t)$$

for almost every t. Gronwall's lemma gives $E(t) \leqslant e^{-2\kappa(t-s)} E(s)$.

The distance estimate uses the same two identities in a slightly different way. On the set where $E(r) > 0$, the PL inequality gives

$$|\partial f|(\alpha_r) \geqslant \sqrt{2\kappa E(r)}.$$

Using the full dissipation identity and $|\dot{\alpha}_r| = |\partial f|(\alpha_r)$,

$$\begin{aligned}
\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_r)_{r\in[s,t]}\big) &= \int_s^t |\dot{\alpha}_r|\,\mathrm{d}r = \int_s^t |\partial f|(\alpha_r)\,\mathrm{d}r \\
&= \int_s^t \frac{|\partial f|^2(\alpha_r)}{|\partial f|(\alpha_r)}\,\mathrm{d}r \leqslant \frac{1}{\sqrt{2\kappa}} \int_s^t \frac{|\partial f|^2(\alpha_r)}{\sqrt{E(r)}}\,\mathrm{d}r \\
&= \frac{1}{\sqrt{2\kappa}} \int_s^t \frac{-E'(r)}{\sqrt{E(r)}}\,\mathrm{d}r = \sqrt{\frac{2}{\kappa}}\big(\sqrt{E(s)} - \sqrt{E(t)}\big).
\end{aligned}$$

If E vanishes on a subinterval, then the flow is stationary there by (3.24), so the same estimate follows by applying the argument on the part where $E > 0$. Since Wasserstein distance is bounded by metric length, the same upper bound controls $\mathcal{W}_2(\alpha_s, \alpha_t)$.

Letting $t \to +\infty$, the energy estimate gives $E(t) \to 0$, and the length estimate shows that the tail of $(\alpha_t)_t$ is Cauchy. Since $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is complete, there is a limit α_∞. Lower semicontinuity yields

$$f(\alpha_\infty) \leqslant \liminf_{t\to+\infty} f(\alpha_t) = f_\star,$$

hence $\alpha_\infty \in \operatorname{Argmin} f$. Sending the endpoint of the length estimate to $+\infty$ with $s = t$ fixed gives

$$\mathcal{W}_2(\alpha_t, \alpha_\infty) \leqslant \sqrt{\frac{2}{\kappa}}\sqrt{E(t)}.$$

Combining this with the exponential energy decay gives the stated distance rate. The distance-to-the-set bound follows because α_∞ is one minimizer. □

The exponential decay of the energy gap is the direct metric-gradient-flow analogue of the classical PL argument. The tail-length estimate used in Theorem 3.20 is a convenient way to turn this energy decay into a $\mathcal{W}_2$-distance-to-Argmin f bound; this exact formulation was communicated to us by Raphaël Barboni. More refined accounts of KL/PL-type inequalities, global convergence and rates for gradient flows and proximal point sequences in general metric spaces are given by Hauer and Mazón Hauer and Mazón (2019) and by Dello Schiavo, Maas and Pedrotti Dello Schiavo et al. (2024).

Geodesic strong convexity. Strong geodesic convexity is the cleanest geometric assumption behind the Wasserstein–PL inequality: it turns displacement convexity into a quantitative lower bound on the metric slope, so that the abstract convergence theorem above applies automatically.

Proposition 3.21 (Strong geodesic convexity implies Wasserstein–PL)**.** *Assume that f is λ-geodesically convex with $\lambda > 0$, that $f_\star$ is attained at some $\alpha^\star$, and that the Wasserstein slope is finite at α. Then*

$$|\partial f|^2(\alpha) \geqslant 2\lambda\big(f(\alpha) - f_\star\big).$$

Thus positive Wasserstein strong convexity implies the Wasserstein–PL inequality with the same constant.

Proof. Fix α and a minimizer $\alpha^\star$. Set

$$g = f(\alpha) - f_\star, \qquad r = \mathcal{W}_2(\alpha, \alpha^\star), \qquad a = |\partial f|(\alpha).$$

If $r = 0$, then $\alpha = \alpha^\star$ and there is nothing to prove. Otherwise let $(\alpha_s)_{s\in[0,1]}$ be a constant-speed geodesic from α to $\alpha^\star$. By λ-geodesic convexity,

$$f(\alpha_s) \leqslant (1-s)f(\alpha) + sf_\star - \frac{\lambda}{2}s(1-s)r^2.$$

Therefore

$$f(\alpha) - f(\alpha_s) \geqslant sg + \frac{\lambda}{2}s(1-s)r^2.$$

Dividing by $\mathcal{W}_2(\alpha, \alpha_s) = sr$ and letting $s \downarrow 0$ in the definition of the descending metric slope yields

$$a \geqslant \frac{g}{r} + \frac{\lambda}{2}r.$$

Equivalently, $g \leqslant ar - \lambda r^2/2$. Maximizing the right-hand side over $r \geqslant 0$ gives $g \leqslant a^2/(2\lambda)$, which is the desired PL inequality. □

Together, Proposition 3.21 and Theorem 3.20 recover the exponential energy rate $e^{-2\lambda t}$ for positively geodesically convex Wasserstein gradient flows, and also give the distance rate $e^{-\lambda t}$ toward the minimizer selected by the flow. The PL formulation is useful because it separates the first-order energy-dissipation mechanism from the stronger curvature requirement of geodesic convexity.

> **Remark 3.22 (Convergence rates for classical examples).** The examples needed to apply this implication have already been isolated above. Proposition 3.9 gives λ-geodesic convexity for linear energies generated by λ-strongly convex potentials, and for the polynomial and pairwise interaction energies covered there when the corresponding finite-dimensional integrand is strongly convex. Theorem 3.11 gives displacement convexity of internal energies, including entropy-type examples, and Proposition 3.12 adds the important nonflat case $\mathrm{KL}(\cdot|\beta)$ when $\mathrm{d}\beta = Z^{-1}e^{-V}\mathrm{d}x$ and V is strongly convex. Whenever these criteria yield a positive convexity constant, Proposition 3.21 turns it into a Wasserstein–PL inequality, and Theorem 3.20 gives the corresponding linear convergence rates.

Wasserstein Kurdyka–Łojasiewicz inequalities. The PL condition is the quadratic case of a broader slope–energy principle. The classical starting point is the gradient inequality of Łojasiewicz for real-analytic functions Łojasiewicz (1963), later extended by Kurdyka to functions definable in o-minimal structures Kurdyka (1998). Modern nonsmooth optimization uses the same idea as the Kurdyka–Łojasiewicz property to prove convergence of descent algorithms and subgradient flows; see, for instance, Bolte, Daniilidis, Ley and Mazet Bolte et al. (2010) and Attouch, Bolte, Redont and Soubeyran Attouch et al. (2010). For displacement-convex functionals, Bolte and Blanchet Bolte and Blanchet (2016) develop a family of Łojasiewicz-type functional inequalities in Wasserstein space and relate them to convergence of the associated gradient dynamics. In the metric-space setting relevant here, the Euclidean gradient norm is replaced by the metric slope, as in the general theory of curves of maximal slope and recent KL/PL convergence results for metric gradient flows Hauer and Mazón (2019); Dello Schiavo et al. (2024).

The Kurdyka–Łojasiewicz viewpoint replaces the linear relation between the energy gap and the squared slope by a power law. It is useful for degenerate convex landscapes, homogeneous energies, and transport discrepancies whose minimizers form a flat set; in those cases the same energy-dissipation computation still gives convergence, but typically with polynomial rather than exponential rates. We use "KL" in this paragraph for Kurdyka–Łojasiewicz, not for Kullback–Leibler divergence.

Definition 3.23 (Wasserstein Kurdyka–Łojasiewicz inequality)**.** Let $f_\star = \inf f$, let $E(\alpha) = f(\alpha) - f_\star$, and fix $c > 0$, $\theta \in [0, 1)$. We say that f satisfies a power Wasserstein–KL inequality on a set $\mathcal{U} \subset \mathcal{P}_2(\mathbb{R}^d)$ if

$$|\partial f|(\alpha) \geqslant c\, E(\alpha)^\theta, \qquad \alpha \in \mathcal{U}, \quad 0 < E(\alpha) < +\infty.$$

Equivalently, the desingularizing function

$$\psi(s) = \frac{s^{1-\theta}}{c(1-\theta)}$$

satisfies $\psi'(E(\alpha))|\partial f|(\alpha) \geqslant 1$. The Wasserstein–PL inequality (3.23) is the special case $\theta = 1/2$ with $c = \sqrt{2\kappa}$. The regime $\theta > 1/2$ gives sublinear rates, while $\theta = 1/2$ is exactly the exponential PL regime.

Theorem 3.24 (Sublinear convergence under Wasserstein–KL)**.** *Let $(\alpha_t)_{t\geqslant 0}$ be a Wasserstein gradient flow of f satisfying the full energy-dissipation identity (3.24). Assume that $f_\star > -\infty$, that $E(t) = f(\alpha_t) - f_\star$ is finite, and that the trajectory stays in a region where the Wasserstein–KL inequality of Definition 3.23 holds with $\theta \in (1/2, 1)$. If $E(0) > 0$, then*

$$E(t) \leqslant \Big(E(0)^{1-2\theta} + c^2(2\theta - 1)t\Big)^{-1/(2\theta-1)}.$$

Moreover, for every $t \geqslant 0$,

$$\mathrm{Length}_{\mathcal{W}_2}\big((\alpha_s)_{s\geqslant t}\big) \leqslant \frac{E(t)^{1-\theta}}{c(1-\theta)}.$$

If f is lower semicontinuous on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$, then α_t converges to some $\alpha_\infty \in \operatorname{Argmin} f$ and

$$\mathcal{W}_2(\alpha_t, \alpha_\infty) \leqslant \frac{E(t)^{1-\theta}}{c(1-\theta)} = O\Big(t^{-(1-\theta)/(2\theta-1)}\Big).$$

Proof. If E reaches zero at some time, then the flow is already at a global minimizer. Indeed, the integrated energy-dissipation identity forces both $|\dot\alpha_t|$ and $|\partial f|(\alpha_t)$ to vanish afterwards. We therefore argue on intervals where $E > 0$. For a.e. t,

$$-E'(t) = |\partial f|^2(\alpha_t) = |\dot\alpha_t|^2.$$

The Wasserstein–KL inequality yields

$$E'(t) \leqslant -c^2 E(t)^{2\theta}.$$

Since $1 - 2\theta < 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t} E(t)^{1-2\theta} = (1-2\theta)E(t)^{-2\theta}E'(t) \geqslant c^2(2\theta - 1),$$

and integration from 0 to t gives the announced polynomial bound.

For the length estimate, use again $|\dot\alpha_t| = |\partial f|(\alpha_t)$. For $T > t$,

$$\begin{aligned}\int_t^T |\dot\alpha_s|\,\mathrm{d}s = \int_t^T \frac{|\partial f|^2(\alpha_s)}{|\partial f|(\alpha_s)}\,\mathrm{d}s &\leqslant \frac{1}{c}\int_t^T -E'(s)E(s)^{-\theta}\,\mathrm{d}s \\ &= \frac{E(t)^{1-\theta} - E(T)^{1-\theta}}{c(1-\theta)} \leqslant \frac{E(t)^{1-\theta}}{c(1-\theta)}.\end{aligned}$$

Letting $T \to +\infty$ shows that the tail has finite length. Hence $(\alpha_t)_t$ is Cauchy in $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$, which is complete, and converges to some α_∞. Since the energy bound gives $E(t) \to 0$, lower semicontinuity gives $f(\alpha_\infty) = f_\star$. The distance to α_∞ is bounded by the remaining tail length, which proves the last estimate. □

Theorem 3.24 is the Wasserstein analogue of the standard KL-rate argument for curves of maximal slope in metric spaces; see Attouch, Bolte and Svaiter Attouch et al. (2010), Hauer and Mazón Hauer and Mazón (2019), and Dello Schiavo, Maas and Pedrotti Dello Schiavo et al. (2024). The endpoint $\theta = 1/2$ is the PL theorem above. Exponents $\theta > 1/2$ correspond to flatter landscapes and slower polynomial rates; formally, $\theta < 1/2$ would instead lead to finite-time convergence in the scalar comparison inequality.

Proposition 3.25 (Powers of PL energies are KL)**.** *Let $g \geqslant 0$ satisfy $\inf g = 0$ and the Wasserstein–PL inequality*

$$|\partial g|^2(\alpha) \geqslant 2\kappa g(\alpha)$$

on a region where $g > 0$. Fix $r \geqslant 1$ and set $f(\alpha) = g(\alpha)^r$. Whenever the metric-slope chain rule $|\partial(g^r)|(\alpha) = r g(\alpha)^{r-1}|\partial g|(\alpha)$ holds, f satisfies the Wasserstein–KL inequality

$$|\partial f|(\alpha) \geqslant r\sqrt{2\kappa}\, f(\alpha)^{1-1/(2r)}.$$

Thus f has KL exponent $\theta = 1 - 1/(2r)$. For $r = 1$ this is PL; for $r > 1$, Theorem 3.24 gives $f(\alpha_t) = O(t^{-r/(r-1)})$ along the f-gradient flow, under its hypotheses.

Proof. By the chain rule and the PL inequality for g,

$$|\partial f| = r g^{r-1}|\partial g| \geqslant r\sqrt{2\kappa}\, g^{r-1/2} = r\sqrt{2\kappa}\, f^{1-1/(2r)}.$$

The exponent and rate follow by substituting $\theta = 1 - 1/(2r)$ into Theorem 3.24. □

Example 3.26 (Homogeneous Wasserstein–KL examples). Power laws first appear for potential energies. Let V be a smooth potential bounded below, write $V_\star = \inf V$, and set

$$f_V(\alpha) = \int V(x)\,\mathrm{d}\alpha(x), \qquad f_\star = V_\star.$$

At the formal smooth level, its Wasserstein slope is

$$|\partial f_V|^2(\alpha) = \int |\nabla V(x)|^2\,\mathrm{d}\alpha(x).$$

Hence, for exponents $\theta \in [1/2, 1)$, the pointwise KL inequality

$$|\nabla V(x)| \geqslant c\,(V(x) - V_\star)^\theta$$

is equivalent to the Wasserstein–KL inequality

$$|\partial f_V|(\alpha) \geqslant c\,(f_V(\alpha) - f_\star)^\theta$$

holding for all probability measures α. The implication from V to f_V follows from Jensen applied to $(V - V_\star)^{2\theta}$, while the converse follows by testing on Dirac masses $\alpha = \delta_x$.
For $q \geqslant 2$, the choice $V(x) = |x|^q$ gives the moment energy

$$f_q(\alpha) = \int |x|^q\,\mathrm{d}\alpha(x)$$

with minimizer δ_0. At the formal smooth level, its Wasserstein gradient is $\nabla|x|^q = q|x|^{q-2}x$, hence

$$|\partial f_q|^2(\alpha) = q^2 \int |x|^{2q-2}\,\mathrm{d}\alpha(x) \geqslant q^2 f_q(\alpha)^{2(q-1)/q},$$

where the last inequality is Jensen applied to $|x|^q$. Thus f_q satisfies a Wasserstein–KL inequality with $c = q$ and $\theta = (q-1)/q$. The case $q = 2$ is PL and produces exponential contraction, while $q > 2$ gives the polynomial rate $f_q(\alpha_t) = O(t^{-q/(q-2)})$.
The same homogeneity appears for powers of distances. On a geodesic metric space, the function $x \mapsto d(x, x_\star)^q$ has metric slope $q d(x, x_\star)^{q-1}$ away from the minimizer. Consequently, in the matching $\mathcal{W}_q$ geometry, the functional $\alpha \mapsto \mathcal{W}_q^q(\alpha, \beta)$ has formal metric slope $q\,\mathcal{W}_q(\alpha, \beta)^{q-1} = q f(\alpha)^{(q-1)/q}$. In the present $\mathcal{W}_2$ geometry, this applies directly to powers $\mathcal{W}_2^q(\alpha, \beta)$; the notation $\mathcal{W}_q^q$ should be read as the analogous statement for the q-Wasserstein metric.
Homogeneous interaction energies exhibit the same scaling. Define

$$f_q^{\mathrm{int}}(\alpha) = \frac{1}{2}\iint |x - y|^q\,\mathrm{d}\alpha(x)\mathrm{d}\alpha(y), \qquad q \geqslant 2.$$

This energy is minimized exactly by Dirac masses, so the minimizer is not unique. Its first variation is

$$\frac{\delta f_q^{\mathrm{int}}}{\delta \alpha}(x) = \int |x - y|^q\,\mathrm{d}\alpha(y),$$

and its Wasserstein gradient is

$$G_\alpha(x) = q \int |x-y|^{q-2}(x-y)\,\mathrm{d}\alpha(y).$$

Writing $\bar{x}_\alpha = \int x\,\mathrm{d}\alpha(x)$, one has $\int G_\alpha\,\mathrm{d}\alpha = 0$ and, by symmetrization,

$$\int G_\alpha(x)\cdot(x-\bar{x}_\alpha)\,\mathrm{d}\alpha(x) = q f_q^{\mathrm{int}}(\alpha).$$

Moreover, Jensen gives

$$\int |x-\bar{x}_\alpha|^q\,\mathrm{d}\alpha(x) \leqslant \iint |x-y|^q\,\mathrm{d}\alpha(x)\mathrm{d}\alpha(y) = 2 f_q^{\mathrm{int}}(\alpha).$$

Combining this estimate with Hölder and Cauchy's inequality yields the non-sharp but useful bound

$$|\partial f_q^{\mathrm{int}}|(\alpha) = \left(\int |G_\alpha|^2\,\mathrm{d}\alpha\right)^{1/2} \geqslant q\,2^{-1/q} f_q^{\mathrm{int}}(\alpha)^{(q-1)/q}.$$

Thus f_q^{int} satisfies a Wasserstein–KL inequality with exponent $\theta = (q-1)/q$ relative to the manifold of Dirac minimizers. For $q = 2$ this is a PL inequality; for $q > 2$ it gives the same polynomial KL rate as the moment energy.

Convexity and curvature. The same language is not restricted to subsets of $\mathbb{R}^d$. If $(\mathcal{X}, d, \mathfrak{m})$ is a geodesic metric-measure space, $\mathcal{W}_2$ geodesics can be defined by transporting each pair of endpoints along metric geodesics, or more intrinsically by dynamical optimal plans on path space, as discussed in Section 1.5. Given a reference measure $\mathfrak{m}$, the entropy relative to $\mathfrak{m}$ is

$$\mathrm{Ent}_{\mathfrak{m}}(\alpha) := \begin{cases} \displaystyle\int_{\mathcal{X}} \rho\log\rho\,\mathrm{d}\mathfrak{m}, & \text{if } \alpha = \rho\,\mathfrak{m}, \\ +\infty, & \text{otherwise.} \end{cases}$$

On a smooth Riemannian manifold (M, g), the Ricci curvature tensor Ric_g is the trace of the Riemann curvature tensor; the lower bound $\mathrm{Ric}_g \geqslant \lambda g$ means that $\mathrm{Ric}_g(v,v) \geqslant \lambda |v|_g^2$ for every tangent vector v. The fundamental link between curvature and optimal transport is that this tensor lower bound is exactly encoded by geodesic convexity of entropy.

Theorem 3.27 (Ricci curvature and entropy convexity)**.** *Let (M, g) be a smooth compact connected Riemannian manifold without boundary, and let $\mathfrak{m} = \mathrm{vol}_g$. For $\lambda \in \mathbb{R}$, the lower Ricci bound $\mathrm{Ric}_g \geqslant \lambda g$ holds if and only if $\mathrm{Ent}_{\mathfrak{m}}$ is λ-geodesically convex on $(\mathcal{P}_2(M), \mathcal{W}_2)$.*

This equivalence was developed in the smooth Riemannian setting by Cordero-Erausquin, McCann and Schmuckenschläger and by von Renesse and Sturm Cordero-Erausquin et al. (2001); von Renesse and Sturm (2005); it is a central theme of the optimal-transport approach to curvature in Villani's monograph Villani (2009). Lott–Villani and Sturm then used the same entropy-convexity principle to define synthetic lower Ricci curvature bounds on metric-measure spaces Lott and Villani (2009); Sturm (2006a,b). Outside this convex, curvature-controlled regime, such as in the mean-field neural-network example below, the flow may still be informative but its convergence analysis requires problem-specific arguments.

3.3. Functional Inequalities via Optimal Transport

Optimal transport proves inequalities by turning comparison into geometry. One transports a density by a monotone or Brenier map, interpolates along the resulting rays, and converts concavity of a Jacobian determinant or convexity of an energy into an integral estimate. The first two examples below illustrate this geometric mechanism. The remainder concerns entropy, where a functional inequality comparing the energy gap with its squared Wasserstein slope is exactly a Wasserstein–PL inequality and therefore yields convergence rates through Theorem 3.20. We work in the smooth Euclidean setting so that the main calculation remains visible; the standard approximation arguments recover the usual Borel and Sobolev formulations. Systematic accounts include McCann's displacement-convexity principle, the Riemannian interpolation inequalities of Cordero-Erausquin, McCann and Schmuckenschläger, and Villani's treatment of concentration and functional inequalities McCann (1997); Cordero-Erausquin et al. (2001); Villani (2009).

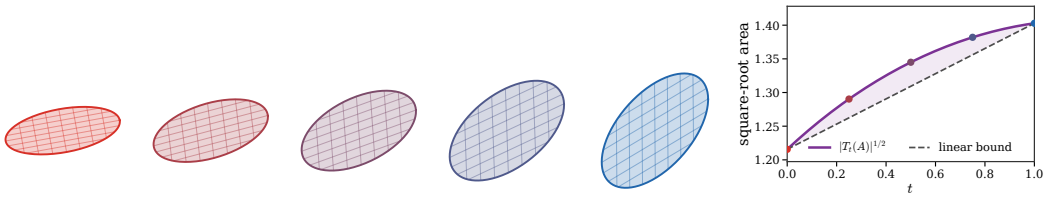


Figure 3.10: Brunn–Minkowski through an affine optimal-transport interpolation between two ellipses. For ellipses, the Brenier map is affine, so the transported support $T_t(A)$ remains an ellipse and its area is computed exactly from the determinant of $(1-t)\mathrm{Id} + tDT$. The right panel shows that $|T_t(A)|^{1/2}$ lies above the linear interpolation of the endpoint square-root areas, which is the two-dimensional determinant concavity behind the proof of Brunn–Minkowski.

Brunn–Minkowski and isoperimetry. The most geometric use of OT is to prove that transporting two uniform measures and looking at the interpolated support increases volume at least concavely. The perimeter inequality is then obtained by differentiating this volume estimate after adding a small ball.

Proposition 3.28 (Brunn–Minkowski and Euclidean isoperimetry)**.** *Let $A, B \subset \mathbb{R}^d$ be bounded Borel sets with positive Lebesgue measure. For $t \in [0,1]$, set*

$$(1-t)A + tB := \{(1-t)x + ty \ : \ x \in A,\ y \in B\}.$$

Then

$$\mathrm{Vol}((1-t)A + tB)^{1/d} \geqslant (1-t)\mathrm{Vol}(A)^{1/d} + t\,\mathrm{Vol}(B)^{1/d}.$$

Consequently, if A is smooth and bounded and $\omega_d = \mathrm{Vol}(B_1)$, then

$$\mathrm{Per}(A) \geqslant d\,\omega_d^{1/d}\mathrm{Vol}(A)^{(d-1)/d}.$$

The constant is sharp, with equality for Euclidean balls.

Proof. We first give the argument for regular sets, the general Borel case following by approximation from inside and outside. Let α and β be the uniform probability measures on A and B, and let $T = \nabla\varphi$ be the Brenier map with $T_\sharp\alpha = \beta$. The interpolating map $T_t = (1-t)\mathrm{Id} + tT$ sends α to a measure α_t supported in $(1-t)A + tB$. At points where T is differentiable,

$$\det DT_t = \det((1-t)\mathrm{Id} + tDT).$$

Since DT is symmetric positive semidefinite and $\det^{1/d}$ is concave on the positive semidefinite cone,

$$\det(DT_t)^{1/d} \geqslant (1-t) + t\det(DT)^{1/d}.$$

The change-of-variables identity for the uniform measures gives $\det(DT) = \mathrm{Vol}(B)/\mathrm{Vol}(A)$ almost everywhere. If ρ_t denotes the density of α_t, a second application of the same formula gives

$$\rho_t(T_t(x))^{-1/d} = \mathrm{Vol}(A)^{1/d}\det(DT_t(x))^{1/d} \geqslant c_t,$$

where $c_t = (1-t)\mathrm{Vol}(A)^{1/d} + t\mathrm{Vol}(B)^{1/d}$. Thus $\rho_t \leqslant c_t^{-d}$ almost everywhere. Since α_t has total mass one and is supported in $(1-t)A + tB$, this gives $\mathrm{Vol}((1-t)A + tB) \geqslant c_t^d$, which is the displayed Brunn–Minkowski inequality.

For isoperimetry, use the homogeneity of Brunn–Minkowski and apply it to A and εB_1. Writing $A + \varepsilon B_1$ for the parallel set gives

$$\mathrm{Vol}(A + \varepsilon B_1)^{1/d} \geqslant \mathrm{Vol}(A)^{1/d} + \varepsilon\omega_d^{1/d}.$$

Using $\mathrm{Vol}(A + \varepsilon B_1) = \mathrm{Vol}(A) + \varepsilon\,\mathrm{Per}(A) + o(\varepsilon)$ and differentiating at $\varepsilon = 0$ yields the claimed perimeter bound. Balls attain equality, so the constant is optimal. □

Figure 3.10 isolates the determinant-concavity step in the special case where the optimal map is affine.

Prékopa–Leindler. The functional analogue of Brunn–Minkowski replaces sets by densities. The transport proof is the same argument with the Jacobian no longer constant.

Proposition 3.29 (Prékopa–Leindler inequality)**.** *Let $u, v, w : \mathbb{R}^d \to [0, +\infty)$ be integrable, and let $t \in [0, 1]$. Assume that*

$$w((1-t)x + ty) \geqslant u(x)^{1-t} v(y)^t \qquad \text{for all } x, y \in \mathbb{R}^d.$$

Then

$$\int_{\mathbb{R}^d} w \geqslant \left(\int_{\mathbb{R}^d} u\right)^{1-t} \left(\int_{\mathbb{R}^d} v\right)^t.$$

Proof. The endpoint cases are immediate, so fix $t \in (0, 1)$. Set $U = \int u$ and $V = \int v$. If $U = 0$ or $V = 0$, the conclusion is trivial; hence assume $U, V > 0$. Let $\alpha = (u/U)\mathrm{d}x$ and $\beta = (v/V)\mathrm{d}x$, and let T be the Brenier map from α to β. Put $T_t = (1-t)\mathrm{Id} + tT$. On a smooth positive regularization, T_t is the gradient of a strictly convex function and is therefore one-to-one almost everywhere. The area formula gives

$$\int w \geqslant \int w(T_t(x)) \det(DT_t(x))\mathrm{d}x.$$

For each eigenvalue $s \geqslant 0$ of DT, weighted arithmetic–geometric mean gives $(1-t) + ts \geqslant s^t$. Hence

$$\det((1-t)\mathrm{Id} + tDT) \geqslant \det(DT)^t.$$

Combining this determinant estimate with the assumption on w yields

$$\int w \geqslant \int u(x)^{1-t} v(T(x))^t \det(DT(x))^t \mathrm{d}x.$$

Since $T_\sharp \alpha = \beta$, the Monge–Ampère identity gives

$$\det(DT(x)) = \frac{u(x)V}{v(T(x))U}$$

almost everywhere. Substitution yields

$$\int w \geqslant \left(\frac{V}{U}\right)^t \int u(x)\mathrm{d}x = U^{1-t}V^t.$$

Approximation removes the smoothness and positivity assumptions. □

Logarithmic Sobolev inequalities. The logarithmic Sobolev inequality, introduced by Gross Gross (1975), turns entropy into a quantitative relaxation principle by controlling it with its dissipation. Let $\beta = \rho_\beta \mathrm{d}x$ be a reference probability measure. For $\alpha = h\beta$, define the relative Fisher information by

$$\mathcal{I}(\alpha|\beta) := \int \|\nabla \log h\|^2 \,\mathrm{d}\alpha = \int \frac{\|\nabla h\|^2}{h}\,\mathrm{d}\beta, \tag{3.25}$$

with the lower-semicontinuous convention $\mathcal{I}(\alpha|\beta) = +\infty$ when $\alpha \not\ll \beta$ or when the displayed integral is not finite. One says that β satisfies a logarithmic Sobolev inequality with constant $\lambda > 0$ if

$$\mathrm{KL}(\alpha|\beta) \leqslant \frac{1}{2\lambda}\mathcal{I}(\alpha|\beta) \qquad \text{for all probability measures } \alpha. \tag{3.26}$$

This is a coercivity property of the reference measure, not a universal inequality. A standard sufficient condition is positive curvature of its potential, in the sense of the Bakry–Emery criterion Bakry and Émery (1985); Bakry et al. (2014). The transport proof below uses the HWI inequality stated later in this section.

Proposition 3.30 (Curvature implies logarithmic Sobolev)**.** *Let $\beta = Z^{-1}e^{-V}\mathrm{d}x$ be a smooth probability measure on $\mathbb{R}^d$. If $\nabla^2 V \geq \lambda \mathrm{Id}$ for some $\lambda > 0$, then β satisfies (3.26) with constant λ.*

Proof. Apply the HWI inequality of Proposition 3.38. Under the curvature assumption, for every absolutely continuous α,

$$\mathrm{KL}(\alpha|\beta) \leqslant \mathcal{W}_2(\alpha,\beta)\sqrt{\mathcal{I}(\alpha|\beta)} - \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$

For fixed $s = \sqrt{\mathcal{I}(\alpha|\beta)}$, maximize the right-hand side over $r = \mathcal{W}_2(\alpha,\beta) \geqslant 0$. The elementary inequality $rs - \lambda r^2/2 \leqslant s^2/(2\lambda)$ gives (3.26). Approximation yields the standard nonsmooth statement. This is the Otto–Villani transport route from the Bakry–Emery curvature bound to logarithmic Sobolev; the original Γ_2 calculus gives the same criterion directly Bakry and Émery (1985); Otto and Villani (2000); Bakry et al. (2014). □

Remark 3.31 (Logarithmic Sobolev as Wasserstein–PL). Under the usual smoothness assumptions, for $f(\alpha) = \mathrm{KL}(\alpha|\beta)$ the Wasserstein gradient is

$$\nabla_{\mathcal{W}} f(\alpha) = \nabla \log \frac{\mathrm{d}\alpha}{\mathrm{d}\beta},$$

and its squared Wasserstein slope is precisely the relative Fisher information,

$$|\partial f|^2(\alpha) = \mathcal{I}(\alpha|\beta).$$

Since the unique minimizer is β and $f(\beta) = 0$, (3.26) is exactly

$$|\partial f|^2(\alpha) \geqslant 2\lambda\big(f(\alpha) - f(\beta)\big),$$

that is, the Wasserstein–PL inequality of Definition 3.19. Thus logarithmic Sobolev is the functional-inequality incarnation of the convergence mechanism studied in Section 3.2, and Theorem 3.20 converts it into exponential decay of both entropy and Wasserstein distance to equilibrium.

Proposition 3.32 (Logarithmic Sobolev inequality implies entropy decay)**.** *Let $\beta = e^{-V}\mathrm{d}x/Z$ be a smooth probability measure on $\mathbb{R}^d$ satisfying the logarithmic Sobolev inequality (3.26) with constant $\lambda > 0$. Let α_t solve the Wasserstein gradient flow of $f(\alpha) = \mathrm{KL}(\alpha|\beta)$, namely*

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V) + \Delta \rho_t, \qquad \alpha_t = \rho_t \mathrm{d}x.$$

If $\mathrm{KL}(\alpha_0|\beta) < +\infty$, then

$$\mathrm{KL}(\alpha_t|\beta) \leqslant e^{-2\lambda t}\,\mathrm{KL}(\alpha_0|\beta).$$

Moreover, if $\alpha_0, \beta \in \mathcal{P}_2(\mathbb{R}^d)$, then

$$\mathcal{W}_2(\alpha_t, \beta) \leqslant \sqrt{\frac{2}{\lambda}\,\mathrm{KL}(\alpha_0|\beta)}\; e^{-\lambda t}.$$

Proof. The first variation of f is $\log(\rho/\rho_\beta)$, up to an additive constant. Along the smooth Wasserstein gradient flow,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(\alpha_t|\beta) = -\int \Big\|\nabla \log \frac{\mathrm{d}\alpha_t}{\mathrm{d}\beta}\Big\|^2 \mathrm{d}\alpha_t = -\mathcal{I}(\alpha_t|\beta).$$

The logarithmic Sobolev inequality bounds the Fisher information below by $2\lambda\,\mathrm{KL}(\alpha_t|\beta)$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{KL}(\alpha_t|\beta) \leqslant -2\lambda\,\mathrm{KL}(\alpha_t|\beta),$$

and Gronwall's lemma gives the entropy estimate. The Wasserstein estimate follows from the length estimate in the Wasserstein–PL theorem: apply Theorem 3.20 to $f = \mathrm{KL}(\cdot|\beta)$, whose unique minimizer is β, and use Remark 3.31. □

Gaussian logarithmic Sobolev inequality. The standard Gaussian is the model case where the optimal constant is known exactly. With the convention of (3.26), $\gamma_d = \mathcal{N}(0, \mathrm{Id})$ satisfies the logarithmic Sobolev inequality with $\lambda = 1$, and this value is sharp. The following transport proof, due to Cordero-Erausquin Cordero-Erausquin (2002), bounds the Jacobian equation for the Brenier map by $\log\det A \leqslant \mathrm{tr}(A - \mathrm{Id})$, then converts the trace term into Fisher information by Gaussian integration by parts.

Proposition 3.33 (Gaussian logarithmic Sobolev inequality)**.** *Let γ_d be the standard Gaussian measure on $\mathbb{R}^d$. If $\rho > 0$ is smooth, has sufficient decay, and $\int \rho\,\mathrm{d}\gamma_d = 1$, then*

$$\mathrm{Ent}_{\gamma_d}(\rho) \coloneqq \int \rho \log \rho\,\mathrm{d}\gamma_d \leqslant \frac{1}{2}\int \frac{\|\nabla\rho\|^2}{\rho}\,\mathrm{d}\gamma_d.$$

Proof. Let $\alpha = \rho\gamma_d$, and let $T = \nabla\varphi$ be the Brenier map with $T_\sharp\alpha = \gamma_d$. Write $T(x) = x + \theta(x)$. The Monge–Ampère equation reads

$$\rho(x)e^{-\|x\|^2/2} = e^{-\|T(x)\|^2/2}\det(DT(x)).$$

Thus

$$\log \rho(x) = -\langle x,\, \theta(x)\rangle - \frac{1}{2}\|\theta(x)\|^2 + \log\det(\mathrm{Id} + D\theta(x)).$$

Since DT is positive semidefinite, $\log\det(\mathrm{Id} + D\theta) \leqslant \operatorname{tr}(D\theta) = \operatorname{div}\theta$. Multiplying by ρ and integrating with respect to γ_d gives

$$\mathrm{Ent}_{\gamma_d}(\rho) \leqslant \int \rho(\operatorname{div}\theta - \langle x,\, \theta\rangle)\,\mathrm{d}\gamma_d - \frac{1}{2}\int \rho\|\theta\|^2\,\mathrm{d}\gamma_d.$$

Gaussian integration by parts gives

$$\int \rho(\operatorname{div}\theta - \langle x,\, \theta\rangle)\,\mathrm{d}\gamma_d = -\int \langle \nabla\rho,\, \theta\rangle\,\mathrm{d}\gamma_d.$$

Completing the square pointwise,

$$-\langle \nabla\rho,\, \theta\rangle - \frac{1}{2}\rho\|\theta\|^2 \leqslant \frac{1}{2}\frac{\|\nabla\rho\|^2}{\rho}.$$

This proves the inequality. The general Sobolev-class statement follows by approximation. □

The constant cannot be improved. For a translation tilt

$$\rho_a(x) = \exp\!\left(\langle a,\, x\rangle - \frac{1}{2}\|a\|^2\right), \qquad a \in \mathbb{R}^d,$$

one has $\rho_a\gamma_d = \mathcal{N}(a, \mathrm{Id})$. A direct computation gives

$$\int \rho_a \log\rho_a\,\mathrm{d}\gamma_d = \frac{1}{2}\|a\|^2, \qquad \frac{1}{2}\int \frac{\|\nabla\rho_a\|^2}{\rho_a}\,\mathrm{d}\gamma_d = \frac{1}{2}\|a\|^2.$$

Thus equality is attained along translations, and the sharp logarithmic Sobolev constant of the standard Gaussian is $\lambda = 1$. Proposition 3.32 then gives the classical exponential convergence of the Ornstein–Uhlenbeck flow.

Poincaré as a linearized Wasserstein–PL inequality. Many quadratic functional inequalities are the infinitesimal form of nonlinear slope inequalities near equilibrium. Assume that β minimizes an energy f, and let ξ be a smooth bounded density modulation with $\int \xi\,\mathrm{d}\beta = 0$. For $|\varepsilon|$ small enough, set

$$\beta_\varepsilon = (1 + \varepsilon\xi)\beta, \qquad \beta_0 = \beta.$$

Define the second-order energy and dissipation forms of f at β in the direction ξ by

$$\mathsf{H}^f_\beta(\xi) := \lim_{\varepsilon\to 0}\frac{2(f(\beta_\varepsilon) - f(\beta))}{\varepsilon^2}, \qquad \mathsf{D}^f_\beta(\xi) := \lim_{\varepsilon\to 0}\frac{|\partial f|^2(\beta_\varepsilon)}{\varepsilon^2},$$

whenever these finite limits exist. The first is the quadratic part of the energy gap and the second is the quadratic part of the squared Wasserstein slope. Their comparison is the linearized Wasserstein–PL inequality.

Proposition 3.34 (Linearizing a Wasserstein–PL inequality)**.** *Let β be a minimizer of an energy f, and assume that f satisfies*

$$|\partial f|^2(\alpha) \geqslant 2\kappa\big(f(\alpha) - f(\beta)\big)$$

in a neighborhood of β. Let ξ be a smooth bounded perturbation with $\int \xi\,\mathrm{d}\beta = 0$, and set $\beta_\varepsilon = (1 + \varepsilon\xi)\beta$. If $\mathsf{H}^f_\beta(\xi)$ and $\mathsf{D}^f_\beta(\xi)$ exist, then

$$\kappa\,\mathsf{H}^f_\beta(\xi) \leqslant \mathsf{D}^f_\beta(\xi).$$

Proof. Apply the PL inequality to β_ε:

$$|\partial f|^2(\beta_\varepsilon) \geqslant 2\kappa\big(f(\beta_\varepsilon) - f(\beta)\big).$$

Divide by ε^2 and let $\varepsilon \to 0$. The definitions of H^f_β and D^f_β give the claim. □

Example 3.35 (Linearized forms for standard discrepancies). The two forms above can be evaluated explicitly under the indicated smoothness and integrability assumptions.

1. Suppose that β has a smooth positive density and that φ is twice continuously differentiable near 1, with $\varphi(1) = 0$ and $\varphi''(1) > 0$. For the Csiszár divergence $f(\alpha) = \mathcal{D}_\varphi(\alpha|\beta)$, one has

$$\mathsf{H}^f_\beta(\xi) = \varphi''(1) \int \xi^2 \,\mathrm{d}\beta, \qquad \mathsf{D}^f_\beta(\xi) = \varphi''(1)^2 \int \|\nabla \xi\|^2 \,\mathrm{d}\beta.$$

   The affine part of φ is immaterial on probability measures, so only the curvature $\varphi''(1)$ enters the linearization. For KL, $\varphi(r) = r \log r - r + 1$, hence $\varphi''(1) = 1$, and the linearized PL inequality becomes the Poincaré inequality.

2. For $f(\alpha) = \mathrm{MMD}^2_k(\alpha, \beta)$, using the convention $\mathrm{MMD}^2_k(\alpha, \beta) = \iint k \,\mathrm{d}(\alpha - \beta)\mathrm{d}(\alpha - \beta)$, assume that k is symmetric, differentiable in its first variable, and positive semidefinite on signed measures of zero total mass, and that the displayed integrals are finite. Then

$$\mathsf{H}^f_\beta(\xi) = 2 \iint k(x, y)\xi(x)\xi(y) \,\mathrm{d}\beta(x)\mathrm{d}\beta(y),$$

   and

$$\mathsf{D}^f_\beta(\xi) = 4 \int \| \int \nabla_x k(x, y)\xi(y) \,\mathrm{d}\beta(y)\|^2 \mathrm{d}\beta(x).$$

   Thus the energy Hessian is the kernel quadratic form, whereas the slope form differentiates the kernel witness spatially.

3. For $f(\alpha) = \mathcal{W}^2_2(\alpha, \beta)$, the calculation is naturally written in the tangent space of $\mathcal{W}_2$. If $\beta = \rho_\beta \mathrm{d}x$ is smooth and positive and the weighted Poisson problem

$$-\nabla \cdot (\rho_\beta \nabla u_\xi) = \xi \rho_\beta,$$

   has a finite-energy solution u_ξ, then

$$\mathsf{H}^f_\beta(\xi) = 2 \int \|\nabla u_\xi\|^2 \,\mathrm{d}\beta, \qquad \mathsf{D}^f_\beta(\xi) = 4 \int \|\nabla u_\xi\|^2 \,\mathrm{d}\beta.$$

   The quadratic form $\int \|\nabla u_\xi\|^2 \mathrm{d}\beta$ is the negative Sobolev norm induced by the linearized Wasserstein metric.

For $f(\alpha) = \mathrm{KL}(\alpha|\beta)$, Proposition 3.34 says precisely that logarithmic Sobolev linearizes to Poincaré. Indeed, for $h_\varepsilon = 1 + \varepsilon\xi$,

$$\mathrm{KL}(h_\varepsilon \beta|\beta) = \frac{\varepsilon^2}{2} \int \xi^2 \,\mathrm{d}\beta + o(\varepsilon^2), \qquad \mathcal{I}(h_\varepsilon \beta|\beta) = \varepsilon^2 \int \|\nabla \xi\|^2 \,\mathrm{d}\beta + o(\varepsilon^2).$$

Proposition 3.36 (Poincaré from logarithmic Sobolev)**.** *Let β be a smooth probability measure on $\mathbb{R}^d$. Assume that β satisfies the logarithmic Sobolev inequality with constant $\lambda > 0$,*

$$\mathrm{KL}(\alpha|\beta) \leqslant \frac{1}{2\lambda} \mathcal{I}(\alpha|\beta), \qquad \mathcal{I}(\alpha|\beta) := \int \|\nabla \log \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\|^2 \,\mathrm{d}\alpha.$$

Then, for every smooth ξ with $\int \xi \,\mathrm{d}\beta = 0$,

$$\lambda \int \xi^2 \,\mathrm{d}\beta \leqslant \int \|\nabla \xi\|^2 \,\mathrm{d}\beta.$$

In particular, if $\mathrm{d}\beta = Z^{-1} e^{-V} \mathrm{d}x$ and $\nabla^2 V \geq \lambda \mathrm{Id}$, so that β is λ-strongly log-concave, then the conclusion holds by the Bakry–Emery criterion Bakry and Émery (1985); Bakry et al. (2014).

Proof. For $h_\varepsilon = 1 + \varepsilon\xi$, with $|\varepsilon|$ small enough, one has $\beta_\varepsilon = h_\varepsilon \beta$ and

$$\mathrm{KL}(\beta_\varepsilon|\beta) = \int h_\varepsilon \log h_\varepsilon \,\mathrm{d}\beta = \frac{\varepsilon^2}{2} \int \xi^2 \,\mathrm{d}\beta + o(\varepsilon^2),$$

because $\int \xi \,\mathrm{d}\beta = 0$. The Fisher information expands as

$$\mathcal{I}(\beta_\varepsilon|\beta) = \int \|\nabla \log h_\varepsilon\|^2 h_\varepsilon \,\mathrm{d}\beta = \varepsilon^2 \int \|\nabla \xi\|^2 \,\mathrm{d}\beta + o(\varepsilon^2).$$

Inserting these expansions in the logarithmic Sobolev inequality and letting $\varepsilon \to 0$ gives the Poincaré inequality. The last assertion is the usual Bakry–Emery implication from λ-convexity of V to the logarithmic Sobolev inequality. □

If $\mathrm{d}\beta = \rho_\beta \mathrm{d}x = Z^{-1} e^{-V} \mathrm{d}x$, the Poincaré inequality is equivalently a spectral gap for the β-weighted Laplacian

$$L_\beta \xi := \Delta \xi - \langle \nabla V,\, \nabla \xi \rangle = \rho_\beta^{-1} \nabla \cdot (\rho_\beta \nabla \xi).$$

Indeed L_β is symmetric in $L^2(\beta)$ and

$$\int \|\nabla \xi\|^2 \,\mathrm{d}\beta = \langle \xi, -L_\beta \xi \rangle_{L^2(\beta)}.$$

Thus Poincaré says that the bottom of the spectrum of $-L_\beta$ on mean-zero functions is at least λ. When the resolvent is compact, this is the first nonzero eigenvalue; on a noncompact space the spectral gap need not be attained by an eigenfunction. This interpretation, classical for reversible Markov diffusion generators Ledoux (2001); Bakry et al. (2014), is the infinitesimal shadow of the nonlinear Wasserstein–PL/log-Sobolev inequality.

This interpretation can be made exact by freezing the Wasserstein mobility at equilibrium. For a mean-zero density perturbation $\dot h$, define

$$\|\dot h\|_{-1,\beta}^2 := \inf_v \left\{ \int \|v\|^2 \,\mathrm{d}\beta \;:\; \dot h = -\rho_\beta^{-1} \nabla \cdot (\rho_\beta v) \right\}.$$

For $E_2(h) = \frac{1}{2}\chi^2(h\beta|\beta) = \frac{1}{2}\int (h-1)^2 \mathrm{d}\beta$, integration by parts shows that its squared slope in this $H^{-1}(\beta)$ geometry is $\int \|\nabla h\|^2 \mathrm{d}\beta$. Poincaré is therefore exactly the PL inequality for E_2 in the linearized Wasserstein geometry. It is not the full $\mathcal{W}_2$-PL inequality for the same energy: the nonlinear Wasserstein mobility is $h\beta$, and the corresponding squared slope is $\int h \|\nabla h\|^2 \mathrm{d}\beta$.

Talagrand's transport-entropy inequality. The distance estimate in Theorem 3.20 already reveals the general principle: when $\mathrm{KL}(\cdot|\beta)$ satisfies logarithmic Sobolev with constant λ, the entropy flow yields the transport-entropy bound $\mathcal{W}_2^2(\alpha,\beta) \leqslant 2\,\mathrm{KL}(\alpha|\beta)/\lambda$. This is the Otto–Villani implication from logarithmic Sobolev to Talagrand's T_2 inequality Otto and Villani (2000). For the standard Gaussian, the same Jacobian estimate used above gives a direct sharp proof of Talagrand's original result Talagrand (1996); Cordero-Erausquin (2002); Villani (2009).

Proposition 3.37 (Gaussian T_2 inequality)**.** *For every $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\frac{1}{2}\mathcal{W}_2^2(\alpha, \gamma_d) \leqslant \mathrm{KL}(\alpha|\gamma_d),$$

with the convention that the right-hand side is $+\infty$ when $\alpha \not\ll \gamma_d$.

Proof. If $\mathrm{KL}(\alpha|\gamma_d) = +\infty$, the claim is immediate. Assume first that $\alpha = \rho\gamma_d$ with ρ smooth and positive. Let $T = \nabla\varphi$ be the Brenier map with $T_\sharp \gamma_d = \alpha$, and write $T(x) = x + \theta(x)$. The change of variables gives

$$\rho(T(x)) e^{-\|T(x)\|^2/2} \det(DT(x)) = e^{-\|x\|^2/2}.$$

Hence

$$\log \rho(T(x)) = \langle x,\, \theta(x)\rangle + \frac{1}{2}\|\theta(x)\|^2 - \log\det(\mathrm{Id} + D\theta(x)).$$

Using again $\log\det(\mathrm{Id} + D\theta) \leqslant \operatorname{div}\theta$ and integrating with respect to γ_d,

$$\mathrm{KL}(\alpha|\gamma_d) \geqslant \frac{1}{2}\int \|\theta\|^2 \,\mathrm{d}\gamma_d + \int (\langle x,\, \theta\rangle - \operatorname{div}\theta)\,\mathrm{d}\gamma_d.$$

The last integral vanishes by Gaussian integration by parts. Since T is optimal,

$$\int \|\theta\|^2 \,\mathrm{d}\gamma_d = \mathcal{W}_2^2(\gamma_d, \alpha),$$

which proves the claim. Approximation gives the general case. Equality holds for translations $\alpha = \mathcal{N}(a, \mathrm{Id})$, so the constant is sharp. □

The HWI bridge. The Gaussian logarithmic Sobolev and transport-entropy inequalities are not isolated facts. The HWI inequality of Otto and Villani Otto and Villani (2000) compares three quantities at once: the entropy H, the Wasserstein distance W, and the Fisher information I. It is one of the cleanest examples where displacement convexity turns into a functional inequality.

Proposition 3.38 (Otto–Villani HWI inequality)**.** *Let $\beta = \rho_\beta \mathrm{d}x = e^{-V}\mathrm{d}x/Z$ be a smooth probability measure in $\mathcal{P}_2(\mathbb{R}^d)$, and assume that $\nabla^2 V \geq \lambda \mathrm{Id}$ for some $\lambda \in \mathbb{R}$. For $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, define the relative Fisher information by*

$$\mathcal{I}(\alpha|\beta) := \int \|\nabla \log \frac{\rho_\alpha}{\rho_\beta}\|^2 \,\mathrm{d}\alpha,$$

when $\alpha = \rho_\alpha \mathrm{d}x$ and the expression is finite, and set $\mathcal{I}(\alpha|\beta) = +\infty$ otherwise. Then

$$\mathrm{KL}(\alpha|\beta) \leqslant \mathcal{W}_2(\alpha,\beta)\sqrt{\mathcal{I}(\alpha|\beta)} - \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$

Proof. If $\mathcal{I}(\alpha|\beta) = +\infty$, there is nothing to prove. Assume first that ρ_α is smooth and positive. Let T be the Brenier map from α to β, set $v = T - \mathrm{Id}$, and let $\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha$. By Theorem 3.11 and Proposition 3.9, $f(\eta) = \mathrm{KL}(\eta|\beta)$ is λ-geodesically convex. Thus the one-dimensional function $e(t) = f(\alpha_t)$ satisfies

$$e(1) \geqslant e(0) + e'(0) + \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$

Since $e(1) = f(\beta) = 0$, it remains to compute the initial derivative. The continuity equation for the geodesic gives $\partial_t \alpha_t + \nabla \cdot (\alpha_t v_t) = 0$ and $v_0 = v$. Hence

$$e'(0) = \int \langle \nabla \log \frac{\rho_\alpha}{\rho_\beta},\, v\rangle \,\mathrm{d}\alpha.$$

Therefore

$$\mathrm{KL}(\alpha|\beta) \leqslant -\int \langle \nabla \log \frac{\rho_\alpha}{\rho_\beta},\, T - \mathrm{Id}\rangle \,\mathrm{d}\alpha - \frac{\lambda}{2}\mathcal{W}_2^2(\alpha,\beta).$$

Cauchy–Schwarz and $\int \|T - \mathrm{Id}\|^2 \mathrm{d}\alpha = \mathcal{W}_2^2(\alpha,\beta)$ give the announced bound. The general case follows by standard approximation and lower semicontinuity of entropy, Fisher information, and $\mathcal{W}_2$. □

When $\lambda > 0$, the HWI inequality is a bridge from curvature to PL. It is not itself a PL inequality because it still contains the distance term $\mathcal{W}_2(\alpha,\beta)$. Optimizing the right-hand side, or simply using Young's inequality $rs - \lambda r^2/2 \leqslant s^2/(2\lambda)$, gives the logarithmic Sobolev estimate

$$\mathrm{KL}(\alpha|\beta) \leqslant \frac{1}{2\lambda}\mathcal{I}(\alpha|\beta).$$

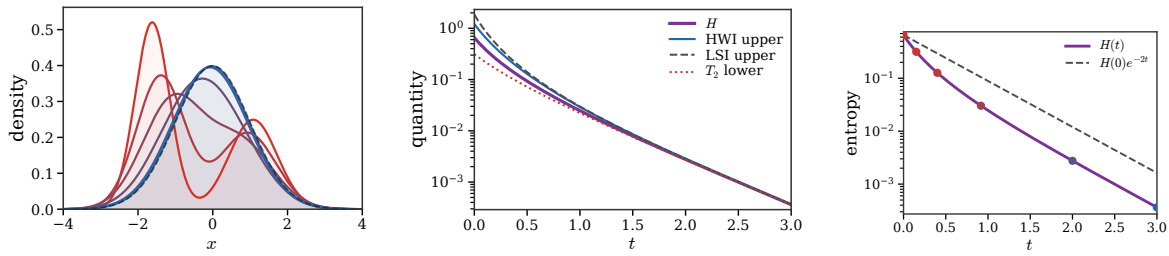
Since $\mathcal{I}(\cdot|\beta) = |\partial\,\mathrm{KL}(\cdot|\beta)|^2$, this is the Wasserstein–PL inequality with constant λ. This is precisely the curvature-to-logarithmic-Sobolev implication used in Proposition 3.30; Proposition 3.32 then converts the same estimate into exponential convergence of the Fokker–Planck flow.

Figure 3.11 shows these inequalities on an exact one-dimensional Ornstein–Uhlenbeck relaxation, where all quantities can be evaluated accurately by grid quadrature and quantile inversion.

3.4. Training Two-Layer MLPs as Wasserstein Flows

Mean-field limits recast the training of wide neural networks as transport of a distribution of neurons. This section shows how the particle ODE of gradient descent becomes a Wasserstein flow in parameter space. This viewpoint, often summarized as treating parameters as interacting particles, was developed in closely related forms by Mei, Montanari and Nguyen Mei et al. (2018), Rotskoff and Vanden-Eijnden Rotskoff and Vanden-Eijnden (2022), and Chizat and Bach Chizat and Bach (2018). These works pass from the finite-width empirical neuron dynamics to a limiting nonlinear PDE for the neuron law.

Wasserstein training of two-layer MLPs. The key modeling step is to forget the ordering of the hidden neurons and to regard the network weights as a probability measure on parameter space. A finite-width network is then an empirical measure, and the population loss becomes a functional of this measure. Gradient descent on the particle positions is precisely the finite-dimensional discretization of a Wasserstein gradient flow for this functional, with



Ornstein–Uhlenbeck densities static inequalities entropy decay

Figure 3.11: Functional inequalities along the Ornstein–Uhlenbeck flow from a one-dimensional Gaussian mixture to the standard Gaussian. The left panel shows the density relaxation, with the Gaussian target as a dashed curve. The middle panel compares $H = \mathrm{KL}(\alpha_t|\gamma_1)$ with the HWI upper bound $W\sqrt{I} - W^2/2$, the logarithmic Sobolev upper bound $I/2$, and the Talagrand lower bound $W^2/2$, where $W = \mathcal{W}_2(\alpha_t, \gamma_1)$ and $I = \mathcal{I}(\alpha_t|\gamma_1)$. The right panel displays the dynamic consequence of logarithmic Sobolev, namely $H(t) \leqslant H(0)e^{-2t}$.

the Wasserstein metric acting on neuron parameters rather than on data samples. This is the sense in which training a two-layer mean-field MLP becomes a transport problem for the law of its neurons.

We use $z \in \mathbb{R}^d$ for the input data and $y \in \mathbb{R}^{d'}$ for the label. A neuron is a particle

$$x = (u, v) \in \mathbb{R}^d \times \mathbb{R}^{d'},$$

where u is the inner weight and v is the outer vector weight. For a scalar nonlinearity σ, define the vector-valued feature

$$\psi(x, z) = v\,\sigma(\langle u,\, z\rangle) \in \mathbb{R}^{d'}.$$

The width-n network and its mean-field version are

$$G_X(z) = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i, z), \qquad G_\alpha(z) = \int \psi(x, z)\mathrm{d}\alpha(x), \qquad \alpha = \frac{1}{n}\sum_i \delta_{x_i}.$$

This formulation removes the artificial ordering of neurons and allows α to be a continuous distribution of infinitely many neurons.

Let ζ be a probability distribution on data-label pairs $(z, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$. The population risk is

$$f(\alpha) = \int \ell(G_\alpha(z), y)\mathrm{d}\zeta(z, y),$$

and the empirical risk is the special case $\zeta = \zeta_N := N^{-1}\sum_{k=1}^{N}\delta_{(z_k,y_k)}$. Since $\alpha \mapsto G_\alpha$ is linear, f is convex as a function of α whenever $\ell(\cdot, y)$ is convex. For the empirical neuron law $\alpha_X = n^{-1}\sum_i \delta_{x_i}$, the Wasserstein metric induces on particles the rescaled metric $n^{-1}\sum_i \|\dot{x}_i\|^2$. The corresponding particle flow is

$$\dot{x}_i = -n\nabla_{x_i}F(X), \qquad F(X) = f\left(\frac{1}{n}\sum_i \delta_{x_i}\right),$$

which is the gradient flow of $F(X) = f(\alpha_X)$ for this Wasserstein particle metric, equivalently Euclidean gradient descent with the time scale multiplied by n. It gives a particle discretization of (3.3).

Assume that ℓ is differentiable in its first variable. The first variation is

$$\delta f(\alpha)(x) = \int \langle \nabla_1 \ell(G_\alpha(z), y),\, \psi(x, z)\rangle \mathrm{d}\zeta(z, y), \tag{3.27}$$

and the Wasserstein gradient in parameter space is

$$\nabla_{\mathcal{W}} f(\alpha)(x) = \nabla_x \delta f(\alpha)(x) = \int [D_x\psi(x, z)]^\top \nabla_1 \ell(G_\alpha(z), y)\mathrm{d}\zeta(z, y).$$

For the squared Euclidean loss $\ell(s, y) = \frac{1}{2}\|s - y\|^2$, the energy is the sum of a quadratic interaction and a linear potential:

$$f(\alpha) = \frac{1}{2}\iint k(x, x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int g(x)\mathrm{d}\alpha(x) + \frac{1}{2}\int \|y\|^2\mathrm{d}\zeta(z, y), \tag{3.28}$$

with

$$k(x,x') = \int \langle \psi(x,z), \psi(x',z)\rangle \mathrm{d}\zeta(z,y), \qquad g(x) = -\int \langle y, \psi(x,z)\rangle \mathrm{d}\zeta(z,y). \tag{3.29}$$

Thus

$$\delta f(\alpha)(x) = \int k(x,x')\mathrm{d}\alpha(x') + g(x), \qquad \nabla_{\mathcal{W}} f(\alpha)(x) = \int \nabla_x k(x,x')\mathrm{d}\alpha(x') + \nabla_x g(x).$$

These kernels are generally not convex in the particle variable, so the geodesic-convex convergence theory above does not apply directly.

Figure 3.12 illustrates the resulting transport of neurons for a homogeneous ReLU model.

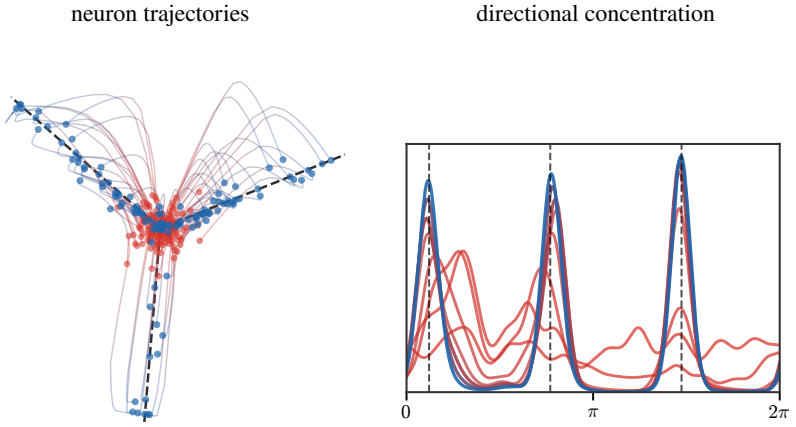


Figure 3.12: Mean-field training of a homogeneous two-layer model as transport in neuron space. The left panel shows the Wasserstein particle gradient flow in the reduced homogeneous coordinates $(|u|v_1, |u|v_2)$, with black dashed rays marking the teacher directions. The right panel shows the weighted angular density along a front-loaded sequence of times, colored from red to blue, so that the early concentration of neuron directions is visible. The display follows the rendering of the auxiliary MLP experiment but keeps only the W_2 flow, not the spectral-flow comparison.

Classical convexity and stationarity. Before using the specific homogeneity mechanism of Chizat and Bach, it is useful to isolate a simpler convex-analytic principle behind many mean-field arguments. Consider an energy of the form

$$f(\alpha) = \frac{1}{2}\iint k(x,x')\mathrm{d}\alpha(x)\mathrm{d}\alpha(x') + \int V(x)\mathrm{d}\alpha(x) + C,$$

on probability measures over a parameter domain. Assume that the quadratic part is convex in the classical sense, namely convex for the affine structure of measures:

$$Q((1-s)\alpha + s\beta) \leqslant (1-s)Q(\alpha) + sQ(\beta), \qquad Q(\alpha) = \frac{1}{2}\iint k\mathrm{d}\alpha\mathrm{d}\alpha.$$

This is ordinary convexity of the functional on the convex set of measures, not displacement convexity along $\mathcal{W}_2$ geodesics.

Proposition 3.39 (Affine convexity and stationary positive densities)**.** *Let $\Omega \subset \mathbb{R}^d$ be connected, let $f = Q + \int V\,\mathrm{d}\alpha + C$ be as above on $\mathcal{P}(\Omega)$, and assume that Q is classically convex. Suppose that a Wasserstein gradient flow α_t of f converges to $\alpha_\infty = \rho_\infty \mathrm{d}x$, where $\rho_\infty > 0$ almost everywhere on Ω, and that $\inf_{t\geqslant 0} f(\alpha_t) > -\infty$. Assume that $\delta f(\alpha_\infty) \in C^1(\Omega)$ and that the squared Wasserstein slope is lower semicontinuous along the convergence, in the sense that*

$$\int_\Omega \|\nabla \delta f(\alpha_\infty)\|^2 \mathrm{d}\alpha_\infty \leqslant \liminf_{n\to\infty} \int_\Omega \|\nabla \delta f(\alpha_{t_n})\|^2 \mathrm{d}\alpha_{t_n}$$

whenever $t_n \to \infty$. Then α_∞ is a global minimizer of f.

Proof. The energy-dissipation identity and the lower bound on f along the convergent trajectory imply

$$\int_0^\infty \int_\Omega \|\nabla \delta f(\alpha_t)\|^2 \mathrm{d}\alpha_t \mathrm{d}t < +\infty.$$

Hence there exists $t_n \to \infty$ for which the inner integral tends to zero. The assumed lower semicontinuity gives

$$\int_\Omega \|\nabla \delta f(\alpha_\infty)\|^2 \mathrm{d}\alpha_\infty = 0.$$

Because $\rho_\infty > 0$ almost everywhere and $\nabla \delta f(\alpha_\infty)$ is continuous, this gradient vanishes throughout Ω. Connectedness then implies that $\delta f(\alpha_\infty)$ is constant. Consequently, for every $\beta \in \mathcal{P}(\Omega)$,

$$\int_\Omega \delta f(\alpha_\infty) \mathrm{d}(\beta - \alpha_\infty) = 0.$$

Classical convexity of f in the affine variable α gives the subgradient inequality

$$f(\beta) \geqslant f(\alpha_\infty) + \int \delta f(\alpha_\infty) \mathrm{d}(\beta - \alpha_\infty) = f(\alpha_\infty).$$

Thus no competitor has smaller energy. Strict positivity is essential to this argument: without it, stationarity controls the first variation only on the support explored by the limiting measure. For square-loss two-layer mean-field models, (3.28) is exactly of this quadratic-plus-linear form, and positive semidefiniteness of the induced kernel k is the classical convexity assumption. □

The convergence theory then separates two regimes. Chizat and Bach prove global convergence for the unregularized, noiseless Wasserstein flow of positively homogeneous models Chizat and Bach (2018). Mei, Montanari and Nguyen prove convergence for the noisy, regularized dynamics Mei et al. (2018): at the PDE level the noise adds a diffusion, or Laplacian, term, so an initially singular neuron law is immediately smoothed and acquires a density. Rotskoff and Vanden-Eijnden emphasize the parameters-as-particles interpretation, long-time convergence and asymptotic error scaling with the network width Rotskoff and Vanden-Eijnden (2022). The following formal statement isolates the noiseless Chizat–Bach mechanism and ignores the technical issues due to ReLU non-smoothness, support propagation and compactness.

Proposition 3.40 (Formal global optimality for two-homogeneous mean-field flows)**.** *Assume that the feature is positively two-homogeneous in the neuron variable,*

$$\psi(\lambda x, z) = \lambda^2 \psi(x, z) \qquad (\lambda > 0),$$

and that $f(\alpha) = J(G_\alpha)$ with J convex and differentiable as a functional of the predictor. Let α be a smooth stationary point of the Wasserstein flow, so that $\nabla_x \delta f(\alpha)(x) = 0$ on $\mathrm{supp}(\alpha)$. Assume also full directional support: for every unit direction ω, the support of α intersects the ray $\{\lambda\omega : \lambda > 0\}$. Then α is a global minimizer of f over the mean-field model class.

Proof. Choose the first-variation representative

$$h_\alpha(x) = \delta f(\alpha)(x) = \langle \nabla J(G_\alpha), \psi(x, \cdot) \rangle_\zeta .$$

Additive constants in first variations do not affect the Wasserstein gradient or first-order inequalities between probability measures; this normalization is useful because it inherits the homogeneity of ψ. By two-homogeneity of ψ, one has

$$h_\alpha(\lambda x) = \lambda^2 h_\alpha(x), \qquad \lambda > 0.$$

Taking $x = 0$ and any $\lambda \neq 1$ in this identity gives $h_\alpha(0) = 0$. We first record that stationarity forces h_α to vanish on $\mathrm{supp}(\alpha)$. Indeed, if $x = r\omega \in \mathrm{supp}(\alpha)$ with $r > 0$ and $\|\omega\| = 1$, then

$$0 = \langle \nabla h_\alpha(r\omega),\, \omega \rangle = \frac{\mathrm{d}}{\mathrm{d}s} h_\alpha(s\omega)\Big|_{s=r} = 2r h_\alpha(\omega),$$

so $h_\alpha(\omega) = 0$, and hence $h_\alpha(x) = r^2 h_\alpha(\omega) = 0$. The case $x = 0$ has already been covered, so $h_\alpha = 0$ on $\mathrm{supp}(\alpha)$ and $\int h_\alpha \, \mathrm{d}\alpha = 0$.

We now argue by contradiction. Suppose that a competitor β satisfies $f(\beta) < f(\alpha)$. Since

$$G_\beta - G_\alpha = \int \psi(x, \cdot) \mathrm{d}(\beta - \alpha)(x),$$

convexity of J gives

$$f(\beta) - f(\alpha) \geqslant \int h_\alpha(x) \mathrm{d}(\beta - \alpha)(x).$$

The left-hand side is negative, while $\int h_\alpha \, \mathrm{d}\alpha = 0$, so $\int h_\alpha \, \mathrm{d}\beta < 0$. Since $h_\alpha(0) = 0$, this strict negativity must occur away from the origin: there exists a nonzero point $x = r\omega$ with $r > 0$, $\|\omega\| = 1$, and $h_\alpha(x) < 0$. Equivalently $h_\alpha(\omega) = h_\alpha(x)/r^2 < 0$. By full directional support, there is $r_\omega > 0$ such that $r_\omega \omega \in \mathrm{supp}(\alpha)$. At this support point,

$$\langle \nabla h_\alpha(r_\omega \omega), \, \omega \rangle = 2 r_\omega h_\alpha(\omega) < 0,$$

which contradicts the stationarity condition $\nabla h_\alpha = 0$ on $\mathrm{supp}(\alpha)$. Thus no competitor has smaller risk.

The rigorous Chizat–Bach theorem replaces the full directional support assumption by propagation and overparameterization hypotheses ensuring that any negative first-variation direction is represented by the evolving support. The same radial-derivative contradiction, applied after this support-propagation step, then rules out non-optimal stationary limits. □

3.5. Generalized Dynamic Wasserstein Flows

This subsection explains how changing the metric geometry changes the associated descent PDE.

Generalized Wasserstein flows. The JKO construction is not tied to the quadratic Wasserstein distance. Once a space of probability measures is equipped with a metric or extended metric d, one can define a minimizing movement by the implicit Euler step

$$\alpha_{t+\tau} \in \operatorname*{argmin}_\alpha \frac{1}{2\tau} d^2(\alpha_t, \alpha) + f(\alpha). \tag{3.30}$$

This is the metric-gradient-flow framework of Ambrosio, Gigli and Savaré Ambrosio et al. (2006); Ambrosio and Gigli (2013): under compactness, coercivity and lower-semicontinuity assumptions adapted to d, the piecewise-constant interpolants of (3.30) can converge, as $\tau \to 0$, to a curve of maximal slope. A nonmetric discrepancy may still be inserted in the same variational update, but the metric theory does not then apply automatically. In the local mass-preserving metrics considered below, a smooth limiting curve is again represented by a continuity equation of the type introduced in (2.2) and used for Wasserstein flows in (3.3),

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t v_{\alpha_t}) = 0, \tag{3.31}$$

but the velocity v_α is selected by the local steepest-descent rule associated with the geometry induced by d, not necessarily by the classical $\mathcal{W}_2$ rule. Unbalanced variants replace this continuity equation by a balance law, while nonlocal variants replace vector fields by jump fluxes.

The metric ingredient needed below is the squared-speed action $\mathbb{A}(\alpha, w)$ introduced in Section 2.3. Its length distance is already defined in (2.13); the present section only uses the local cost in the infinitesimal variational step defined below. Some actions are pointwise integrals, such as ordinary $\mathcal{W}_2$ with $\mathbb{A}(\alpha, w) = \int \|w\|^2 \mathrm{d}\alpha$. Others are built from a homogeneous local density but then squared at the metric level: for $\mathcal{W}_p$, the local densities are $A_p(a, w) = a\|w\|^p$ and $J_p(a, m) = \|m\|^p / a^{p-1}$, while the action paired with the JKO penalty is the squared Finsler speed

$$\mathbb{A}_p(\alpha, w) = \left(\int \|w\|^p \mathrm{d}\alpha \right)^{2/p}.$$

Spectral geometries use the nonlocal covariance action $\mathbb{A}_\gamma$ from (2.21). In the $\mathcal{W}_2$ case, all objects reduce to the familiar Benamou–Brenier ones:

$$A_2(a, w) = a\|w\|^2, \qquad J_2(a, m) = \frac{\|m\|^2}{a}, \qquad \mathsf{D}_{\mathbb{A}} = \mathcal{W}_2 \, .$$

Definition 3.41 (Penalized minimization oracle)**.** Let $\mathbb{A}(\alpha, \cdot)$ be the local squared-speed action at α. For a cotangent field u for which the pairing below is finite, usually $u = \nabla_{\mathcal{W}} f(\alpha)$, the penalized minimization oracle associated with this action is

$$\mathrm{PMO}_{\mathbb{A},\alpha}(u) \in \operatorname*{argmin}_w \left\{ \int \langle u(x), \, w(x) \rangle \mathrm{d}\alpha(x) + \frac{1}{2} \mathbb{A}(\alpha, w) \right\}, \tag{3.32}$$

where the minimization is over admissible finite-action velocity representatives for the chosen geometry. In

Hilbertian examples $u \in L^2(\alpha; \mathbb{R}^d)$; for $\mathcal{W}_p$, the natural dual space is $L^q(\alpha; \mathbb{R}^d)$, with $q = p/(p-1)$.

For an energy f, the formal small-step expansion of (3.30), when the distance is generated by the squared-speed action $\mathbb{A}$ and the JKO displacement admits an admissible finite-action tangent representative, selects the PMO direction

$$v_\alpha = \mathrm{PMO}_{\mathbb{A},\alpha}(\nabla_{\mathcal{W}} f(\alpha)). \tag{3.33}$$

Equivalently, the formal evolution is

$$\partial_t \alpha_t + \mathrm{div}\big(\alpha_t \ \mathrm{PMO}_{\mathbb{A},\alpha_t}(\nabla_{\mathcal{W}} f(\alpha_t))\big) = 0. \tag{3.34}$$

If $w \mapsto \mathbb{A}(\alpha, w)$ has zero-action directions, the minimization in (3.32) is understood on the quotient tangent space, or after projecting onto finite-action directions modulo the kernel of $\mathbb{A}(\alpha, \cdot)$. Along smooth solutions of (3.34), whenever the minimizer is attained and $w \mapsto \mathbb{A}(\alpha, w)$ is 2-homogeneous, one obtains the dissipation identity

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = \langle \nabla_{\mathcal{W}} f(\alpha_t), v_t \rangle_{\alpha_t} = -\mathbb{A}(\alpha_t, v_t), \qquad v_t = \mathrm{PMO}_{\mathbb{A},\alpha_t}(\nabla_{\mathcal{W}} f(\alpha_t)).$$

For $\mathcal{W}_2$, $\mathbb{A}(\alpha, w) = \int \|w\|^2 \mathrm{d}\alpha$ and $\mathrm{PMO}_{\mathbb{A},\alpha}(u) = -u$. Hence (3.34) reduces to the standard Wasserstein gradient-flow equation

$$\partial_t \alpha_t = \mathrm{div}(\alpha_t \nabla_{\mathcal{W}} f(\alpha_t)),$$

which is the convention used throughout the preceding sections.

In the restricted Riemannian case of (2.14), where

$$\mathbb{A}(\alpha, w) = \langle Q_\alpha w, w \rangle_{L^2(\alpha)},$$

the PMO becomes the linear preconditioning rule

$$\mathrm{PMO}_{\mathbb{A},\alpha}(u) = -Q_\alpha^{-1} u, \qquad v_\alpha = -Q_\alpha^{-1} \nabla_{\mathcal{W}} f(\alpha). \tag{3.35}$$

The linear inverse formula is special to Hilbertian actions. A basic non-Hilbertian example is $\mathcal{W}_p$: for $1 < p < +\infty$, the squared Finsler tangent action associated with the distance is

$$\mathbb{A}_p(\alpha, w) = \left(\int \|w\|^p \mathrm{d}\alpha \right)^{2/p}.$$

Writing $q = p/(p-1)$ and $U = \|u\|_{L^q(\alpha)}$, its PMO is the duality-map formula

$$\mathrm{PMO}_{\mathbb{A}_p,\alpha}(u)(x) = -U^{2-q} \|u(x)\|^{q-2} u(x), \tag{3.36}$$

with the conventions that the right-hand side is 0 when $U = 0$, and that $\|u(x)\|^{q-2} u(x) = 0$ when $u(x) = 0$. This is nonlinear in u, and it reduces to $-u$ only for $p = 2$. Endpoint cases such as $p = 1$ are interpreted through subgradients of the dual norm.

The examples below are organized by the same dictionary. Concave-mobility flows use $\mathbb{A}_{\theta,\lambda}(\alpha, v)$ from Section 2.3; spectral flows use $\mathbb{A}_\gamma(\alpha, v)$ from (2.21); kernelized Benamou–Brenier flows use $\mathbb{A}_k(\alpha, v) = \|v\|^2_{\mathcal{H}_k^d}$ with a restricted RKHS tangent space. Nonlocal logarithmic-mean geometries also have quadratic actions, but their tangent variables are edge potentials or jump velocities rather than vector fields in $\mathbb{R}^d$; their flows are treated separately in Section 3.6. WFR, defined by (2.37) and studied separately in Section 3.7, is another quadratic geometry, but its tangent variable is a pair (v, g) combining displacement and mass modulation.

The generalized distances of Section 2.3 are useful precisely because they change what *descent* means. The energy functional can be the same, while the local metric tensor, Finsler norm, mobility, spectral gauge or RKHS constraint changes the PDE selected by the minimizing movement. The rest of this section focuses on local or vector-field mass-preserving geometries. Nonlocal pair-space geometries are separated in Section 3.6, and the transport–reaction case is separated in Section 3.7, because their tangent variables are structurally different.

General mobility flows. The concave-mobility distances of Section 2.3 replace the scalar linear mobility a by a concave mobility $\theta(a)$. For a density field $a = \rho_t(x)$, this changes the Onsager operator of the gradient flow while keeping a continuity equation.

Let λ be Lebesgue measure, write $\alpha = \rho\lambda$, and let $u = \nabla(\delta f/\delta\rho)$. For the velocity action $\mathbb{A}_{\theta,\lambda}$, the PMO introduced in Definition 3.41 is the pointwise minimizer of

$$\int \langle u,\, w\rangle\, \rho \,\mathrm{d}x + \frac{1}{2}\int \frac{\rho}{\theta(\rho)}\|w\|^2 \rho \,\mathrm{d}x,$$

so, wherever $\rho > 0$ and $\theta(\rho) > 0$,

$$\mathrm{PMO}_{\mathbb{A}_{\theta,\lambda},\rho\lambda}(u) = -\frac{\theta(\rho)}{\rho}u. \tag{3.37}$$

Thus the flux ρw is $-\theta(\rho)\nabla(\delta f/\delta\rho)$. For a smooth density ρ_t and an energy f, the formal gradient flow associated with the pointwise momentum action $J_\theta(a,m) = |m|^2/\theta(a)$ is

$$\partial_t \rho_t = \nabla\cdot\left(\theta(\rho_t)\nabla\frac{\delta f}{\delta\rho}(\rho_t)\right). \tag{3.38}$$

If

$$f(\rho) = \int U(\rho(x))\,\mathrm{d}x + \int V(x)\rho(x)\,\mathrm{d}x,$$

then (3.38) becomes

$$\partial_t\rho = \nabla\cdot\big(\theta(\rho)\big(U''(\rho)\nabla\rho + \nabla V\big)\big).$$

Thus $\theta(a) = a$ and $U(a) = a\log a$ gives the usual heat or Fokker–Planck equation, while $\theta(a) = a(1-a/M)$ gives a volume-filling drift–diffusion whose mobility vanishes at saturation. Power internal energies and power mobilities tune nonlinear diffusion: for instance $\theta(a) = a$ and $U(a) = a^m/(m-1)$ yield $\partial_t\rho = \Delta(\rho^m)$, whereas the entropy with $\theta(a) = a^\gamma$, $0 < \gamma < 1$, gives the fast-diffusion-type equation $\partial_t\rho = (1/\gamma)\Delta(\rho^\gamma)$. This viewpoint is useful for nonlinear diffusions, exclusion models, volume-filling models and finite-volume discretizations Dolbeault et al. (2009); vector-valued and dissipative extensions follow the same principle but use richer mobility tensors.

Dynamic spectral Wasserstein flows. The spectral tangent action (2.21), and the dynamic distance (2.22) it generates, normalize the whole velocity covariance through a monotone spectral gauge. Using this geometry for gradient descent replaces the pointwise Wasserstein direction by a globally preconditioned direction. This is close to many large-scale optimizers: normalized gradient methods solve norm-constrained linear minimization problems Pethick et al. (2025); related ideas appear in normalized SGD Murray et al. (2019); Cutkosky and Mehta (2020), tensor-aware optimizers such as Shampoo Gupta et al. (2018), and Muon-type spectral normalizations Jordan et al. (2024); Liu et al. (2025). We now describe the mean-field version developed in Peyré (2026). The trace gauge gives the classical $\mathcal{W}_2$ flow, while the operator gauge $\gamma(M) = \lambda_{\max}(M)$ gives the idealized Muon geometry.

Specializing Definition 3.41 to the spectral action $\mathbb{A}_\gamma$, the descent direction associated with a field $g \in L^2(\alpha;\mathbb{R}^d)$ is simply $\mathrm{PMO}_{\mathbb{A}_\gamma,\alpha}(g)$. The field g should be read as the Wasserstein gradient $g = \nabla_{\mathcal{W}} f(\alpha) = \nabla\delta f(\alpha)$. Since $v \mapsto \mathbb{A}_\gamma(\alpha, v)$ is 2-homogeneous, fixed-speed normalized algorithms keep the same ray as this PMO and only change the scalar normalization; equivalently, they minimize the linear pairing over the unit ball $\mathbb{A}_\gamma(\alpha, w) \leqslant 1$. In the trace case, $\mathbb{A}_{\mathrm{tr}}(\alpha, v) = \int \|v\|^2 \mathrm{d}\alpha$ and $\mathrm{PMO}_{\mathbb{A}_{\mathrm{tr}},\alpha}(g) = -g$. For other gauges, the velocity is globally preconditioned by the covariance of g under α.

Proposition 3.42 (Polar formula for the spectral direction)**.** *Let*

$$S_\alpha(g) \coloneqq \int g(x)g(x)^\top \mathrm{d}\alpha(x).$$

Assume that the inverse-trace problem admits a positive definite minimizer

$$A^\star_\alpha \in \operatorname*{argmin}_{A\in\mathcal{B}_\gamma\cap\mathbb{S}^d_{++}} \mathrm{tr}\Big(A^{-1}S_\alpha(g)\Big).$$

Then

$$\mathrm{PMO}_{\mathbb{A}_\gamma,\alpha}(g)(x) = -(A^\star_\alpha)^{-1}g(x) \tag{3.39}$$

is the spectral PMO direction. If the minimizer lies on the boundary of $\mathcal{B}_\gamma$, the same formula is understood by a limiting argument along positive definite minimizers, or equivalently by replacing the inverse by the corresponding inverse on the range of $S_\alpha(g)$.

Proof. Using the polar formula $\gamma(M) = \sup_{A\in\mathcal{B}_\gamma} \operatorname{tr}(AM)$, the PMO objective (3.32) with $\mathbb{A} = \mathbb{A}_\gamma$ can be written as

$$\sup_{A\in\mathcal{B}_\gamma} \left\{ \int \langle g,\, v\rangle \mathrm{d}\alpha + \frac{1}{2}\int v^\top A v \mathrm{d}\alpha \right\}.$$

For a fixed $A \succ 0$, the quadratic lower bound

$$\int \langle g,\, v\rangle \mathrm{d}\alpha + \frac{1}{2}\int v^\top A v \mathrm{d}\alpha$$

is minimized by $v_A = -A^{-1}g$, with value $-\frac{1}{2}\operatorname{tr}(A^{-1}S_\alpha(g))$. Let $v_\star = -(A^\star_\alpha)^{-1}g$ and

$$M_\star = \int v_\star(x)v_\star(x)^\top \mathrm{d}\alpha(x) = (A^\star_\alpha)^{-1}S_\alpha(g)(A^\star_\alpha)^{-1}.$$

Since $A^\star_\alpha$ minimizes $A \mapsto \operatorname{tr}(A^{-1}S_\alpha(g))$ over the convex polar set, its first-order optimality condition gives

$$\operatorname{tr}(AM_\star) \leqslant \operatorname{tr}(A^\star_\alpha M_\star) \qquad \text{for all } A \in \mathcal{B}_\gamma.$$

Thus $A^\star_\alpha$ is active in the polar representation of $\gamma(M_\star)$, and

$$\gamma(M_\star) = \operatorname{tr}(A^\star_\alpha M_\star) = \operatorname{tr}((A^\star_\alpha)^{-1}S_\alpha(g)).$$

For any v, the polar formula gives

$$\int \langle g,\, v\rangle \mathrm{d}\alpha + \frac{1}{2}\mathbb{A}_\gamma(\alpha, v) \geqslant \int \langle g,\, v\rangle \mathrm{d}\alpha + \frac{1}{2}\int v^\top A^\star_\alpha v \mathrm{d}\alpha,$$

and the right-hand side is minimized at $v_\star$. The activity identity above shows that equality holds at $v_\star$, hence $v_\star$ is the spectral PMO direction. The boundary case follows by approximation with positive definite matrices in the polar set. □

For the trace gauge, $\mathcal{B}_\gamma = \{A\,;\, 0 \preceq A \preceq \mathrm{Id}\}$ and the minimizer is $A^\star_\alpha = \mathrm{Id}$. For the operator gauge $\gamma(M) = \lambda_{\max}(M)$, one has $\mathcal{B}_\gamma = \{A \succeq 0\,;\, \operatorname{tr}(A) \leqslant 1\}$; if $S_\alpha(g) \succ 0$, then

$$A^\star_\alpha = \frac{S_\alpha(g)^{1/2}}{\operatorname{tr}(S_\alpha(g)^{1/2})}, \qquad \mathrm{PMO}_{\mathbb{A}_\gamma,\alpha}(g)(x) = -\operatorname{tr}(S_\alpha(g)^{1/2})\, S_\alpha(g)^{-1/2}g(x).$$

This is the continuum covariance-normalization formula that becomes the Muon polar factor for empirical measures.

Proposition 2.9 identifies the static spectral distance $\mathcal{W}_\gamma$ with the length distance generated by $\mathbb{A}_\gamma$. The infinitesimal descent direction is therefore defined directly by the general PMO problem (3.32) specialized to this action. This action-based statement avoids an important overinterpretation: static/dynamic equality alone does not imply $\mathcal{W}^2_\gamma(\alpha, (\mathrm{Id} + \tau v)_\sharp\alpha) = \tau^2\mathbb{A}_\gamma(\alpha, v) + o(\tau^2)$ for every smooth field v, because v need not be the minimal finite-action representative of the induced tangent perturbation. Whenever a minimizing-movement scheme for $\mathcal{W}_\gamma$ converges to a smooth curve whose metric derivative is represented by this minimal action, it produces the same evolution described below.

Proposition 3.43 (Formal normalized spectral gradient flow)**.** *Assume that f has a smooth first variation and that the spectral PMO is attained along a smooth curve. The formal gradient flow generated by the dynamic spectral action $\mathbb{A}_\gamma$ solves*

$$\partial_t\alpha_t + \operatorname{div}\Big(\alpha_t\ \mathrm{PMO}_{\mathbb{A}_\gamma,\alpha_t}(\nabla_{\mathcal{W}} f(\alpha_t))\Big) = 0. \tag{3.40}$$

Equivalently, if $A^\star_t$ solves the inverse-trace problem for

$$S_t = \int \nabla_{\mathcal{W}} f(\alpha_t)(x)\nabla_{\mathcal{W}} f(\alpha_t)(x)^\top \mathrm{d}\alpha_t(x),$$

then the velocity is

$$v_t(x) = -(A^\star_t)^{-1}\nabla_{\mathcal{W}} f(\alpha_t)(x).$$

Moreover, along smooth solutions,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\mathbb{A}_\gamma(\alpha_t, v_t).$$

If the minimizing movements (3.30) for $d = \mathcal{W}_\gamma$ converge to a smooth curve with this minimal-action metric derivative, that limit satisfies the same equation.

Proof. The action-based steepest-descent rule (3.33) gives $v_t = \mathrm{PMO}_{\mathbb{A}_\gamma,\alpha_t}(\nabla_W f(\alpha_t))$. Inserting this velocity into the continuity equation (3.31) gives (3.40). Proposition 3.42 gives the explicit inverse-trace formula. Finally, the variation formula gives $\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = \int \langle \nabla_W f(\alpha_t),\, v_t\rangle \mathrm{d}\alpha_t$. Since v_t minimizes a 2-homogeneous penalized objective, the one-dimensional rescaling $s \mapsto s v_t$ has derivative zero at $s = 1$, hence

$$\int \langle \nabla_W f(\alpha_t),\, v_t\rangle \mathrm{d}\alpha_t + \mathbb{A}_\gamma(\alpha_t, v_t) = 0,$$

which proves the dissipation identity. □

Remark 3.44 (Empirical and Muon limits). For empirical measures $\alpha_X = n^{-1}\sum_i \delta_{x_i}$, stack the velocities $v(x_i)$ into a matrix $V \in \mathbb{R}^{n\times d}$ and the gradients $g(x_i)$ into a matrix $G \in \mathbb{R}^{n\times d}$. By 1-homogeneity of γ,

$$\mathbb{A}_\gamma(\alpha_X, v) = \gamma\left(\frac{1}{n} V^\top V\right) = \frac{1}{n}\gamma(V^\top V), \qquad \int \langle g,\, v\rangle \mathrm{d}\alpha_X = \frac{1}{n}\operatorname{tr}(G^\top V).$$

Thus the empirical specialization of the PMO problem (3.32) is, up to the harmless factor $1/n$, the finite-dimensional matrix descent problem. Writing

$$(G_X(t))_i = \nabla_W f(\alpha_{X(t)})(x_i(t)),$$

the empirical flow is

$$\dot X(t) \in \operatorname*{argmin}_{V\in\mathbb{R}^{n\times d}} \left\{ \operatorname{tr}(G_X(t)^\top V) + \frac{1}{2}\gamma(V^\top V)\right\}. \tag{3.41}$$

Here G_X is the gradient for the empirical Wasserstein metric $\|\dot X\|_n^2 = n^{-1}\sum_i \|\dot x_i\|^2$, not the ordinary Euclidean gradient of $f(\alpha_X)$. If one rewrites the same rule with the unweighted matrix convention, it corresponds to the lift $X \mapsto n f(\alpha_X)$; using $X \mapsto f(\alpha_X)$ only rescales time by the factor n. If $G = U\,\mathrm{diag}(\sigma_i) W^\top$ and $\gamma(M) = \lambda_{\max}(M)$, then

$$\operatorname*{argmin}_{V\in\mathbb{R}^{n\times d}} \left\{ \operatorname{tr}(G^\top V) + \frac{1}{2}\lambda_{\max}(V^\top V)\right\} \ni -\|G\|_{S_1} U W^\top = -\operatorname{tr}\Big((G^\top G)^{1/2}\Big)\, G (G^\top G)^{\dagger/2}. \tag{3.42}$$

The ray of this direction is the polar factor $-UW^\top$, and the scalar factor $\|G\|_{S_1}$ can be absorbed into a time or step-size normalization. This is the exact-polar, full-batch, continuous-time idealization of Muon Jordan et al. (2024); Peyré (2026). Practical Muon implementations replace the polar factor by a few Newton–Schulz iterations for GPU efficiency Liu et al. (2025); this polynomial approximation fits the same spectral-normalization philosophy, although adaptive rescaling and finite momentum make the practical algorithm more than a position-only metric gradient flow.

The effect of this spectral normalization is visible already in the homogeneous two-layer ReLU model discussed in Section 3.4. Figure 3.13 reproduces, in the style of the book figures, the numerical comparison from Peyré (2026): the same teacher network, same initialization and same empirical first variation are evolved either by the W_2 particle flow or by the operator-gauge Muon direction.

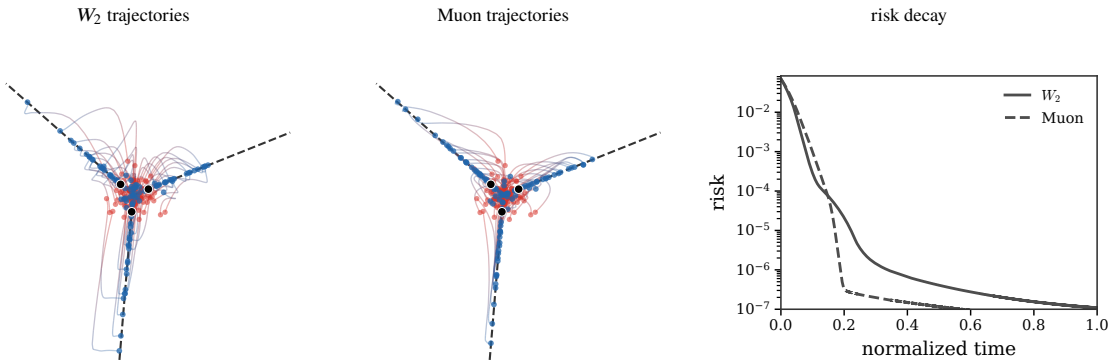


Figure 3.13: Wasserstein versus Muon mean-field training of a homogeneous ReLU model. The first two panels display trajectories in the reduced homogeneous coordinates $(|u|v_1, |u|v_2)$; black dashed rays are the teacher directions. The right panel compares the empirical square-loss decay, with the solid curve for the W_2 flow and the dashed curve for the operator-gauge Muon flow. Both methods use the same particles and first variation; the spectral normalization reaches the low-risk regime earlier in normalized time.

3.6. Nonlocal Wasserstein Flows

The nonlocal distances of Section 2.4 replace the local vector-field tangent model by pairwise exchanges. This separate section records the corresponding gradient-flow interpretation in the same order as the distance constructions: first continuum nonlocal Wasserstein flows and fractional PDEs, then discrete Wasserstein flows on Markov chains. In the continuum jump-kernel case, entropy descent gives nonlocal diffusion and, for power-law kernels, fractional PDEs. On a finite state space, the same logarithmic-mean edge calculus makes a reversible Markov chain exactly the entropy gradient flow of the discrete Wasserstein distance. Both examples are mass-preserving, but they are not local Benamou–Brenier flows: their tangent variables live on pairs or graph edges rather than at individual points.

Nonlocal Wasserstein flows and fractional PDEs. The continuum nonlocal distance $\mathcal{W}_K$ of Section 2.4 replaces local velocities by antisymmetric fluxes across jumps selected by a reversible kernel K. Its logarithmic-mean mobility is chosen so that entropy dissipation produces the Markov generator. This turns jump processes, fractional heat equations and some heavy-tailed stochastic models into gradient flows for nonlocal transport geometries.

Proposition 3.45 (Entropy gradient flow for reversible jump kernels)**.** *Under the regularity and irreducibility assumptions used in Proposition 2.11, the Markov semigroup generated by the closure of*

$$L\psi(x) = \int \big(\psi(y) - \psi(x)\big)K(x, \mathrm{d}y)$$

is the gradient flow, for the distance $\mathcal{W}_K$ defined in (2.27), of the relative entropy

$$\mathrm{Ent}_{\mathfrak{m}}(\alpha) = \int \rho \log \rho \, \mathrm{d}\mathfrak{m}, \qquad \alpha = \rho\mathfrak{m}.$$

Equivalently, in weak form, the entropy flow is

$$\partial_t \rho_t = L\rho_t. \tag{3.43}$$

When K is singular, the integral defining L is understood in the principal-value sense.

Proof idea. The key identity is the logarithmic-mean relation

$$\theta(a, b)(\log a - \log b) = a - b.$$

The first variation of $\mathrm{Ent}_{\mathfrak{m}}$ is $\log \rho + 1$. With the sign convention of (2.26), the steepest descent velocity is

$$v(x, y) = -\bar{\nabla}(\log \rho + 1)(x, y) = \log \rho(x) - \log \rho(y).$$

Hence $\theta(\rho(x), \rho(y))v(x, y) = \rho(x) - \rho(y)$. Substituting this in (2.26) and using the symmetry of J gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi\rho \, \mathrm{d}\mathfrak{m} = \int \varphi(x) \int (\rho(y) - \rho(x))K(x, \mathrm{d}y) \, \mathfrak{m}(\mathrm{d}x),$$

which is the weak form of $\partial_t \rho = L\rho$. The metric properties needed to justify this computation as a gradient-flow identity are the distance and geodesic results recalled in Proposition 2.11. □

The construction becomes especially transparent for translation-invariant kernels. A power-law jump kernel turns the entropy flow into a fractional heat equation.

On $\mathcal{X} = \mathbb{R}^d$, let

$$K_s(x, \mathrm{d}y) = c_{d,s} \frac{\mathrm{d}y}{|x-y|^{d+s}}, \qquad 0 < s < 2.$$

The associated generator is, up to the normalizing constant,

$$L_s\psi(x) = \text{p.v.} \int_{\mathbb{R}^d} \big(\psi(y) - \psi(x)\big) \frac{c_{d,s}}{|x-y|^{d+s}} \,\mathrm{d}y = -(-\Delta)^{s/2}\psi(x),$$

where $(-\Delta)^{s/2}$ is the fractional Laplacian Samko et al. (1993). Hence the $\mathcal{W}_{K_s}$-gradient flow of the entropy is

$$\partial_t \rho_t = -(-\Delta)^{s/2}\rho_t. \tag{3.44}$$

Figure 3.14 illustrates the qualitative effect of lowering s. Classical heat diffusion, $s = 2$, smooths local discontinuities by nearby averaging. Fractional diffusion keeps a sharper memory of localized peaks but immediately creates algebraic long-range tails, because mass can jump over macroscopic distances.

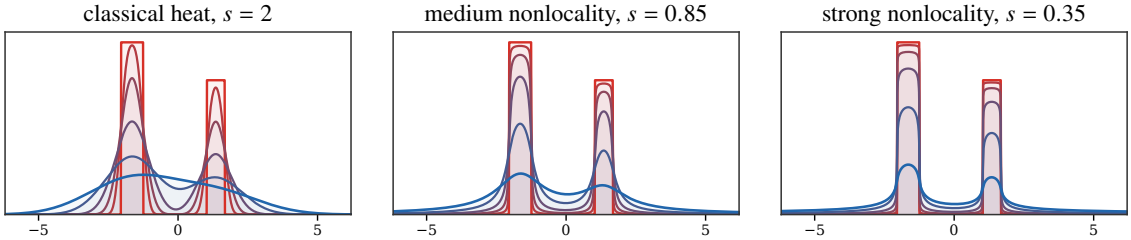


Figure 3.14: Classical and fractional heat flows from two localized indicator bumps. Each panel evolves the same normalized mixture of two intervals by the Fourier multiplier $e^{-t|\xi|^s}$, with time colored from red to blue. The classical heat flow quickly rounds the discontinuities and spreads by local Gaussian averaging. As s decreases, the diffusion becomes increasingly nonlocal: peaks remain more localized near the bumps, while heavier tails appear across the whole displayed window.

More generally, when a target-dependent jump kernel K_β satisfies detailed balance with a desired invariant measure $\beta = \rho_\beta \mathfrak{m}$, the same construction can be applied to $r_t = \mathrm{d}\alpha_t/\mathrm{d}\beta$. The gradient flow of $\mathrm{KL}(\alpha|\beta)$ is then $\partial_t r_t = L_\beta r_t$. This is the nonlocal analogue of the Fokker–Planck gradient flow of relative entropy, and it is one of the motivations behind more recent nonlocal diffusion gradient-flow frameworks Warren (2024).

Remark 3.46 (Lévy SDE viewpoint). The same operators appear probabilistically as generators of jump processes. This gives a complementary interpretation of nonlocal PDEs as laws of stochastic processes with heavy-tailed jumps.

For the kernel K_s, the process with generator L_s is the symmetric s-stable Lévy process $X_t = X_0 + L_t^{(s)}$, whose density solves (3.44) Applebaum (2009). Adding a smooth confining drift gives the fractional Fokker–Planck equation

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V) - \sigma^s (-\Delta)^{s/2}\rho_t, \tag{3.45}$$

which is the forward equation of

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}L_t^{(s)}.$$

Equation (3.45) should be distinguished from the exactly reversible entropy flow above unless the jump kernel is chosen to satisfy detailed balance with the target measure. Both viewpoints, however, share the same nonlocal mechanism: relaxation is driven not only by local drift and Brownian diffusion but also by rare long jumps.

Remark 3.47 (Connection with stochastic gradient dynamics). This nonlocal point of view is useful in machine learning because stochastic optimization noise need not be well approximated by Brownian noise. Classical diffusion approximations model small-step SGD by a Brownian SDE near stationarity Mandt et al. (2017). In deep-network regimes, empirical and theoretical work has shown that gradient noise and even stationary iterates can be heavy-tailed cSimcsekli et al. (2019); Gürbüzbalaban et al. (2021). A coarse continuous-time model for parameters Θ_t is then

$$\mathrm{d}\Theta_t = -\nabla \mathcal{L}(\Theta_t)\,\mathrm{d}t + \sigma\,\mathrm{d}L_t^{(s)},$$

whose law solves a fractional Fokker–Planck equation of the form (3.45). The jumps represent occasional large stochastic-gradient fluctuations, so they can move the dynamics between basins in a way that Brownian approximations tend to understate. This makes nonlocal Wasserstein geometries a useful language for linking heavy-tailed training noise, fractional PDE models, and measure-valued gradient-flow ideas. The correspondence remains a modeling approximation: the effective tail index, anisotropy, minibatch correlations and learning-rate schedule all influence whether a Lévy limit is a faithful description of a concrete training run.

Discrete Wasserstein flows on Markov chains. The finite-state distance $\mathcal{W}_K$ in Section 2.4 is designed so that the reversible Markov chain itself becomes an entropy gradient flow. This is the discrete counterpart of the fact that the heat equation is the Wasserstein gradient flow of Shannon entropy.

For masses $a_i = \pi_i \rho_i$, or equivalently relative densities $\rho_i = a_i/\pi_i$ with respect to the invariant law π, the entropy relative to π is

$$\mathrm{Ent}_\pi(\rho) := \sum_i \pi_i \rho_i \log \rho_i.$$

Applying the general minimizing movement (3.30) with $d = \mathcal{W}_K$ and $f = \mathrm{Ent}_\pi$, the small-τ first-order optimality condition gives

$$\frac{\rho^{k+1} - \rho^k}{\tau} = -\mathcal{K}_{\rho^{k+1}} \log \rho^{k+1} + o(1) = K\rho^{k+1} + o(1).$$

Here $(K\rho)_i = \sum_j K_{ij}(\rho_j - \rho_i)$. Under smoothness and strict positivity, $\rho^{k+1} = \rho^k + O(\tau)$, so the last expression also equals $K\rho^k + O(\tau)$. Thus the discrete Wasserstein geometry is engineered so that the entropy gradient flow is the Markov semigroup; the proposition below makes this identity explicit.

Proposition 3.48 (Entropy gradient flow of a reversible Markov chain)**.** *Let K be reversible with invariant law π. The gradient flow of* Ent_π *for the metric $\mathcal{W}_K$ is the forward equation of the Markov chain,*

$$\dot\rho_i(t) = \sum_j K_{ij}\big(\rho_j(t) - \rho_i(t)\big). \tag{3.46}$$

Equivalently, for the masses $a_i(t) = \pi_i\rho_i(t)$, this is $\dot a_i(t) = \sum_j (a_j(t)K_{ji} - a_i(t)K_{ij})$.

Proof. The first variation of the entropy, with respect to the weighted pairing $\sum_i \pi_i \xi_i \varphi_i$, is $\log \rho_i + 1$. Constants do not contribute to $\mathcal{K}_\rho$, hence the metric gradient-flow equation is

$$\dot\rho = -\mathcal{K}_\rho \log \rho.$$

Using the identity

$$\theta(a,b)(\log a - \log b) = a - b,$$

one obtains, componentwise,

$$\dot\rho_i = -\sum_j K_{ij}\theta(\rho_i,\rho_j)(\log\rho_i - \log\rho_j) = \sum_j K_{ij}(\rho_j - \rho_i),$$

which is the density form of the Kolmogorov forward equation. Multiplying by π_i and using detailed balance gives the equation for the probability masses. □

3.7. Dynamic Unbalanced OT and WFR Flows

The generalized dynamic Wasserstein flows of Section 3.5 are primarily mass-preserving: their tangent vectors are generated by continuity equations and are therefore velocity fields advecting the measure. Unbalanced transport changes the state space from probabilities to positive finite measures, where descent may both move and create or remove mass. The tangent variable is then a pair (v, g): the vector field v displaces mass, while the scalar field g modulates its local intensity. This transport–reaction geometry is the dynamic counterpart of the unbalanced distances of Section 2.5, especially the balance-equation formula (2.37), and it is best treated separately from the purely advective geometries above.

Definition 3.49 (WFR minimizing movement)**.** Let $f : \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ be an energy on positive finite measures and let WFR_κ be the Wasserstein–Fisher–Rao distance with growth scale $\kappa > 0$, defined by the dynamic balance-equation formulation (2.37) and its static equivalence in Proposition 2.14. The unbalanced JKO step is

$$\alpha^{k+1} \in \operatorname*{argmin}_{\alpha \in \mathcal{M}_+(\mathbb{R}^d)} \frac{1}{2\tau}\,\mathrm{WFR}_\kappa^2(\alpha^k, \alpha) + f(\alpha). \tag{3.47}$$

The continuous-time limit, when it exists, is called the WFR_κ gradient flow of f.

Formally, a smooth curve $\alpha_t = \rho_t \mathrm{d}x$ has tangent vectors of the form

$$\partial_t \rho_t + \mathrm{div}(\rho_t v_t) = g_t \rho_t,$$

with squared norm $\int (\|v_t\|^2 + \kappa^2 g_t^2)\rho_t \mathrm{d}x$. The vector field v_t accounts for displacement, while g_t is a Fisher–Rao growth rate. This is the infinitesimal version of the balance-equation action in Section 2.5.

Proposition 3.50 (Formal WFR gradient-flow PDE)**.** *Assume that f has a smooth first variation $\varphi_t = \delta f(\alpha_t)$ and that $\alpha_t = \rho_t \mathrm{d}x$ is smooth and positive. The formal WFR_κ gradient flow of f is*

$$\partial_t \rho_t = \mathrm{div}(\rho_t \nabla \varphi_t) - \frac{1}{\kappa^2} \varphi_t \rho_t. \tag{3.48}$$

Along such a smooth solution,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\int \left(\|\nabla \varphi_t\|^2 + \frac{1}{\kappa^2} \varphi_t^2 \right) \rho_t \, \mathrm{d}x \leqslant 0. \tag{3.49}$$

Proof. For an infinitesimal tangent pair (v, g), the density variation is $\zeta = -\mathrm{div}(\rho v) + g\rho$. The first variation of f is

$$\int \varphi \, \zeta = \int \langle \nabla \varphi, \, v \rangle \rho \, \mathrm{d}x + \int \varphi g \rho \, \mathrm{d}x,$$

after integration by parts. For the inner product

$$\langle (v, g), \, (\tilde{v}, \tilde{g}) \rangle_\rho = \int \langle v, \, \tilde{v} \rangle \rho \, \mathrm{d}x + \kappa^2 \int g \tilde{g} \rho \, \mathrm{d}x,$$

the Riesz representative is $(\nabla \varphi, \varphi/\kappa^2)$. Steepest descent therefore uses $(v, g) = (-\nabla \varphi, -\varphi/\kappa^2)$, which gives (3.48). Substituting this pair in the first-variation formula gives

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\alpha_t) = -\int \left(\|\nabla \varphi_t\|^2 + \frac{1}{\kappa^2} \varphi_t^2 \right) \rho_t \, \mathrm{d}x,$$

which is (3.49). □

The reaction term makes additive constants in the first variation meaningful: adding a constant to φ changes the global growth rate. This is not a defect but the infinitesimal signature of optimizing over positive finite measures, rather than on the probability simplex. The rigorous metric-space construction of Kantorovich–Fisher–Rao and WFR gradient flows is developed in Gallouët and Monsaingeon (2017); Chizat et al. (2018b,a). Chizat's measure-optimization framework relates weight-changing gradient methods to mirror- and Bregman-type descents on positive measures Chizat (2022a,b). In machine learning, birth–death dynamics use the same reaction intuition to let a particle population reweight, remove and create neurons during training Rotskoff et al. (2019).

Example 3.51 (Relative entropy and birth–death relaxation). Let $\beta = b\mathrm{d}x$ and use the finite-measure relative entropy

$$f(\alpha) = \mathrm{KL}(\alpha|\beta) = \int \left(\rho \log \frac{\rho}{b} - \rho + b\right) \mathrm{d}x, \qquad \alpha = \rho \mathrm{d}x.$$

This is the Csiszár divergence on positive measures; it is nonnegative and is minimized, without imposing a mass constraint, at $\alpha = \beta$. Its first variation is $\delta f(\alpha) = \log(\rho/b)$. The balanced $\mathcal{W}_2$ flow, when the total mass is fixed, is the Fokker–Planck equation

$$\partial_t \rho_t = \mathrm{div}\left(\rho_t \nabla \log \frac{\rho_t}{b}\right) = \Delta \rho_t - \mathrm{div}(\rho_t \nabla \log b),$$

whereas the WFR_κ flow is

$$\partial_t \rho_t = \Delta \rho_t - \mathrm{div}(\rho_t \nabla \log b) - \frac{1}{\kappa^2} \rho_t \log \frac{\rho_t}{b}. \tag{3.50}$$

The first two terms transport and diffuse mass, while the last term kills mass where $\rho_t > b$ and creates it where $\rho_t < b$.

Figure 3.15 contrasts this reaction-assisted relaxation with its mass-preserving Wasserstein counterpart.

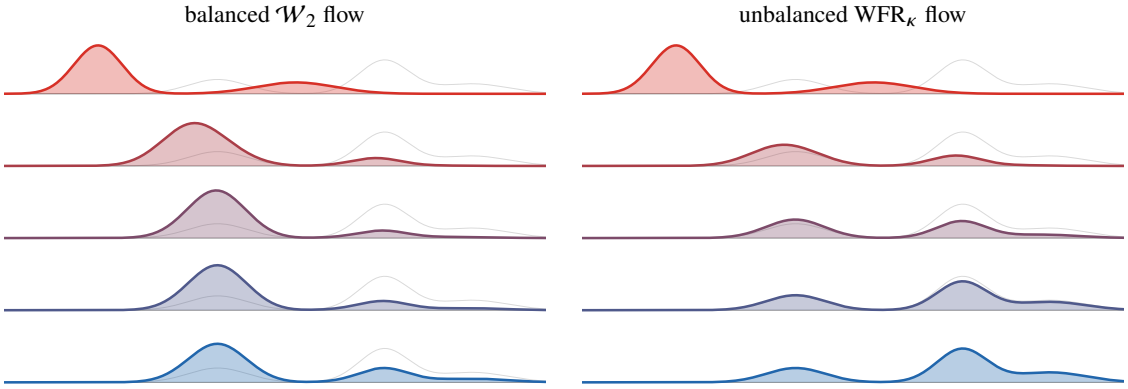


Figure 3.15: Balanced and unbalanced gradient flows of $\mathrm{KL}(\rho|\beta)$ *for one-dimensional Gaussian mixtures.* In each panel the colored density stacks correspond to five increasing times, from red to blue, and the faint gray curve repeated on each row is the target mixture β. The balanced Wasserstein flow is the conservative Fokker–Planck equation, so mass must move through the low-density region between modes. The WFR_κ flow adds the birth–death reaction term in (3.50); it attenuates overrepresented regions and creates missing mass near target modes, reaching the target shape much faster.

3.8. Conditional Wasserstein Training of Infinite ResNets

Residual networks learn small residual updates around the identity He et al. (2016). This makes depth behave like time: very deep residual architectures can be interpreted as discretizations of differential equations, a viewpoint developed in stable architectures, neural ODEs and mean-field optimal-control models of deep learning Haber and Ruthotto (2017); Lu et al. (2018); Chen et al. (2018); E et al. (2019). The conditional-OT formulation of Barboni, Peyré and Vialard Barboni et al. (2024) adds the width mean-field limit to this continuous-depth picture. It belongs to a broader conditional-transport literature, including fibered gradient flows Peszek and Poyato (2023), conditional optimal transport on function spaces Hosseini et al. (2025), conditional Wasserstein distances for Bayesian OT flow matching Chemseddine et al. (2025), and dynamic conditional OT through simulation-free flows Kerrigan et al. (2024). From the viewpoint of Section 3.5, it is a concrete instance of a generalized dynamic Wasserstein flow: the conditional Wasserstein distance has a geodesic dynamic structure obtained by integrating the usual Benamou–Brenier action over the conditioning variable. The point of the construction is to keep track of two independent symmetries: neurons can be relabelled inside each layer, but mass should not move from one depth to another. Conditional Wasserstein geometry implements exactly this rule. It transports neurons fiber by fiber in parameter space, while the depth variable remains fixed.

We follow the three limiting levels of the model. A finite ResNet is first a labelled particle system arranged by layers. Sending the width to infinity turns each layer into a probability law of neurons. Sending the depth step to zero then turns the sequence of layer laws into a conditional measure indexed by depth, on which training becomes a Wasserstein gradient flow.

Finite-depth finite-width ResNets. The starting point is the ordinary architecture: depth is discrete, width is finite, and each layer carries a labelled collection of neurons. Fix a parameter space $\Theta \subset \mathbb{R}^q$, a residual feature $\psi : \Theta \times \mathbb{R}^d \to \mathbb{R}^d$, a data law ζ on input-label pairs (z, y), and a loss ℓ. With L layers, step size $\tau = 1/L$, widths n_r, and parameters $\theta_{r,i} \in \Theta$, a finite ResNet is the composition of the residual updates

$$z_{r+1} = z_r + \tau G_{r,n_r}(z_r), \qquad G_{r,n_r}(z) \coloneqq \frac{1}{n_r}\sum_{i=1}^{n_r} \psi(\theta_{r,i}, z), \qquad z_0 = z, \qquad r = 0, \dots, L-1. \tag{3.51}$$

The associated population loss, with $n = (n_0, \dots, n_{L-1})$, is

$$f_{L,n}((\theta_{r,i})_{r,i}) = \int \ell(z_L^\theta(z), y)\mathrm{d}\zeta(z, y).$$

Equivalently, each layer carries an empirical neuron law

$$\alpha_r = \frac{1}{n_r}\sum_{i=1}^{n_r} \delta_{\theta_{r,i}}, \qquad G_{r,n_r}(z) = \int_\Theta \psi(\theta, z)\mathrm{d}\alpha_r(\theta).$$

Thus a finite ResNet is already a discrete conditional measure on depth and neuron space,

$$\alpha^{L,n}(\mathrm{d}s, \mathrm{d}\theta) = \frac{1}{L}\sum_{r=0}^{L-1} \delta_{s_r}(\mathrm{d}s)\alpha_r(\mathrm{d}\theta), \qquad s_r = r/L,$$

whose marginal on the depth variable is the discrete measure $\lambda_L = L^{-1}\sum_{r=0}^{L-1} \delta_{s_r}$.

Finite-depth mean-field limit. The first mean-field limit removes neuron labels inside each fixed layer. The residual depth grid stays discrete, but the widths tend to infinity. Each empirical law α_r is replaced by an arbitrary probability measure in $\mathcal{P}_2(\Theta)$, and each residual block becomes a mean-field residual block

$$z_{r+1} = z_r + \tau G_{\alpha_r}(z_r), \qquad z_0 = z, \qquad G_{\alpha_r}(z) \coloneqq \int_\Theta \psi(\theta, z)\mathrm{d}\alpha_r(\theta).$$

The loss becomes a functional of the finite family of layer measures,

$$f_L(\alpha_0, \dots, \alpha_{L-1}) = \int \ell(z_L^\alpha(z), y)\mathrm{d}\zeta(z, y).$$

This is the finite-depth conditional Wasserstein setting with condition marginal λ_L: each fiber s_r carries a Wasserstein geometry on neuron laws.

Continuous-depth conditional geometry. The second limit turns the layer index into a conditioning variable. The model is no longer described by one neuron law, nor even by a finite list of them, but by a family of neuron laws indexed by depth. The state is therefore a conditional probability law

$$\alpha(\mathrm{d}s, \mathrm{d}\theta) = \alpha_s(\mathrm{d}\theta)\mathrm{d}s,$$

or, equivalently, a probability measure on $[0, 1] \times \Theta$ whose first marginal is Lebesgue measure. Here the condition is continuous depth, $s \in [0, 1]$, the condition law is $\lambda(\mathrm{d}s) = \mathrm{d}s$, and the fiber space is the neuron parameter space Θ. For two conditional neuron laws $\alpha(\mathrm{d}s, \mathrm{d}\theta) = \alpha_s(\mathrm{d}\theta)\mathrm{d}s$ and $\beta(\mathrm{d}s, \mathrm{d}\theta) = \beta_s(\mathrm{d}\theta)\mathrm{d}s$, the relevant distance is the depthwise quadratic Wasserstein metric

$$\mathcal{W}_{2,\lambda}^2(\alpha, \beta) = \int_0^1 \mathcal{W}_2^2(\alpha_s, \beta_s)\mathrm{d}s. \tag{3.52}$$

It compares neurons only at the same depth: transport in θ is allowed fiber by fiber, while mass is not transported between layers.

Infinite-depth mean-field ResNet. Once depth is continuous, the forward pass is a controlled ODE and the control is the conditional neuron law. A mean-field infinite-depth and infinite-width ResNet is described by a measurable family $s \mapsto \alpha_s \in \mathcal{P}_2(\Theta)$. Each slice is the neuron law of one infinitesimal residual block, and the feature $\psi(\theta, z)$ now acts as a vector field on the state variable. The forward pass is

$$\partial_s z_s = G_{\alpha_s}(z_s) \coloneqq \int_\Theta \psi(\theta, z_s) \mathrm{d}\alpha_s(\theta), \qquad z_0 = z. \tag{3.53}$$

The population risk is

$$f_{\mathrm{Res}}(\alpha) = \int \ell(z_1^\alpha(z), y) \mathrm{d}\zeta(z, y).$$

Under the usual regularity assumptions ensuring well-posedness of the ODE, this defines a functional on the conditional Wasserstein space.

Conditional Wasserstein gradient flow. Training is now steepest descent for an objective on conditional laws. More generally, let (S, λ) be a probability space of conditions and let f be a differentiable functional on measures $\alpha(\mathrm{d}s, \mathrm{d}\theta) = \alpha_s(\mathrm{d}\theta)\lambda(\mathrm{d}s)$ with fixed first marginal λ. For signed perturbations $\xi(\mathrm{d}s, \mathrm{d}\theta) = \xi_s(\mathrm{d}\theta)\lambda(\mathrm{d}s)$ satisfying $\int_\Theta \mathrm{d}\xi_s = 0$ for λ-a.e. s, we use the fiberwise first-variation convention

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f(\alpha + \varepsilon\xi)\bigg|_{\varepsilon=0} = \int_S \int_\Theta \frac{\delta f}{\delta \alpha_s}(\alpha)(\theta) \mathrm{d}\xi_s(\theta) \mathrm{d}\lambda(s).$$

The formal conditional Wasserstein gradient flow is the family of continuity equations

$$\partial_t \alpha_{t,s} + \mathrm{div}_\theta(\alpha_{t,s} v_{t,s}) = 0, \qquad v_{t,s}(\theta) = -\nabla_\theta \frac{\delta f}{\delta \alpha_s}(\alpha_t)(\theta), \tag{3.54}$$

for λ-a.e. condition s, whenever the first variation is differentiable in the parameter variable. In the ResNet case one takes $f = f_{\mathrm{Res}}$ and $S = [0, 1]$. The equation is fiberwise in its transport geometry, but not decoupled as a training dynamics: changing α_s affects the terminal loss through the forward ODE and the corresponding adjoint equation. Thus each depth performs a Wasserstein steepest descent in its own neuron space, while all depths interact through the state trajectory.

Particle discretization and layerwise matching. The finite network is recovered by discretizing both the condition variable and the fiber measures. With the discrete depth marginal λ_L, finite-depth finite-width ResNets are particle approximations of the conditional flow above. If two networks have the same depth grid and the same widths, and if their empirical layer laws are

$$\alpha_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \delta_{\theta_{r,i}}, \qquad \beta_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \delta_{\theta'_{r,i}},$$

then their squared conditional distance is

$$\mathcal{W}^2_{2,\lambda_L}(\alpha^{L,n}, \beta^{L,n}) = \frac{1}{L} \sum_{r=0}^{L-1} \mathcal{W}_2^2(\alpha_r, \beta_r).$$

Keeping neuron labels fixed gives the labelled parameter distance

$$\frac{1}{L} \sum_{r=0}^{L-1} \frac{1}{n_r} \sum_{i=1}^{n_r} \|\theta_{r,i} - \theta'_{r,i}\|^2,$$

which corresponds to a particular feasible coupling and therefore upper bounds the squared conditional Wasserstein distance. Optimizing over permutations inside each equal-weight layer gives the Wasserstein matching term above. In this sense, the shallow mean-field flow is the one-fiber case, whereas infinitely deep ResNets require a continuum of Wasserstein flows indexed by depth and coupled through the network state.

3.9. Second-Order Momentum Flows

The minimizing-movement viewpoint is intrinsically first order: a JKO step produces the next measure by trading off proximity to the previous measure and decrease of the energy. Momentum adds a memory of velocity. The state is therefore no longer only a measure, but a measure together with a tangent, or equivalently phase-space, variable. This makes a direct implicit JKO construction less natural than for first-order flows. We instead use an explicit construction, in the same spirit as optimization and machine-learning practice: momentum reduces zig-zagging across stiff directions, helps iterates keep a coherent direction through shallow valleys, and can accelerate convergence when the geometry is favorable Qian (1999); Sutskever et al. (2013). The goal is to write this inertial idea in a form compatible with Wasserstein particle flows.

Finite-dimensional momentum. Let $F : \mathbb{R}^N \to \mathbb{R}$ be a smooth finite-dimensional objective. Polyak's heavy-ball method Polyak (1964) augments gradient descent with a velocity variable. With step size $h > 0$ and damping $\gamma > 0$, one convenient velocity form is

$$S^{k+1} = (1 - \gamma h)S^k - h\nabla F(X^k), \qquad X^{k+1} = X^k + hS^{k+1}. \tag{3.55}$$

If S^k approximates $\dot{X}(kh)$, this is a semi-implicit, or symplectic-Euler, discretization of the first-order system: the velocity update is explicit, while the position uses the newly updated velocity. The limiting system is

$$\dot{X}(t) = S(t), \qquad \dot{S}(t) = -\gamma S(t) - \nabla F(X(t)), \tag{3.56}$$

or equivalently the damped second-order equation

$$\ddot{X}(t) + \gamma\dot{X}(t) + \nabla F(X(t)) = 0. \tag{3.57}$$

Nesterov's accelerated method Nesterov (1983) uses a closely related extrapolation but evaluates the gradient at a look-ahead point,

$$Y^k = X^k + \theta_k(X^k - X^{k-1}), \qquad X^{k+1} = Y^k - h\nabla F(Y^k), \tag{3.58}$$

with $\theta_k = (k-1)/(k+r-1)$ for the classical convex schedule $r = 3$. Its scaling limit is the time-dependent damping equation

$$\ddot{X}(t) + \frac{r}{t}\dot{X}(t) + \nabla F(X(t)) = 0. \tag{3.59}$$

This ODE interpretation is due to Su, Boyd and Candès Su et al. (2016); see also the variational perspective of Wibisono, Wilson and Jordan Wibisono et al. (2016), and the inertial dynamics with Hessian-driven damping studied by Attouch, Peypouquet and Redont Attouch et al. (2016). High-resolution ODE limits refine this picture by keeping terms that distinguish heavy-ball and Nesterov dynamics beyond their leading continuous-time behavior Shi et al. (2018). For the Wasserstein extension below, the autonomous heavy-ball form (3.56) is the cleanest template.

Empirical Wasserstein lift. We now take a functional f on measures and lift it to particle configurations. For $X = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$, write

$$\alpha_X := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \qquad F(X) := f(\alpha_X). \tag{3.60}$$

This is the same empirical lift as in (3.4), and is the particle viewpoint underlying mean-field analyses of wide neural networks and Wasserstein gradient methods Chizat and Bach (2018); Mei et al. (2018); Rotskoff and Vanden-Eijnden (2022). The important point is that the configuration space is not used with the plain Euclidean metric, but with the empirical metric

$$\|\dot{X}\|_n^2 := \frac{1}{n}\sum_{i=1}^n \|\dot{x}_i\|^2,$$

which is the metric induced by $\mathcal{W}_2$ on equally weighted empirical measures with fixed labels. Moving only x_i to $x_i + \varepsilon\xi$ gives

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} F(x_1, \dots, x_i + \varepsilon\xi, \dots, x_n) = \frac{1}{n}\langle \nabla\delta f(\alpha_X)(x_i), \xi\rangle.$$

Thus, if $\nabla^{(n)}F$ denotes the gradient for the metric $\|\cdot\|_n$, then

$$(\nabla^{(n)}F(X))_i = n\nabla_{x_i}F(X) = \nabla\delta f(\alpha_X)(x_i) = \nabla_{\mathcal{W}} f(\alpha_X)(x_i). \tag{3.61}$$

The first-order Wasserstein particle descent is therefore $\dot{x}_i = v_{\alpha_X}(x_i)$ with

$$v_\alpha(x) := -\nabla_W f(\alpha)(x) = -\nabla \delta f(\alpha)(x). \tag{3.62}$$

Applying the heavy-ball equation to the empirical metric gives the second-order Wasserstein momentum system

$$\dot{x}_i(t) = s_i(t), \qquad \dot{s}_i(t) = -\gamma s_i(t) + v_{\alpha_t}(x_i(t)), \qquad \alpha_t = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}. \tag{3.63}$$

Equivalently,

$$\ddot{x}_i(t) + \gamma \dot{x}_i(t) = v_{\alpha_t}(x_i(t)) = -\nabla_W f(\alpha_t)(x_i(t)).$$

This is an explicit inertial particle method for the measure energy f: the Wasserstein steepest-descent field acts as an acceleration rather than as an instantaneous velocity. The statement is deliberately formal at this level; rigorous mean-field limits require stability and regularity assumptions on v_α, as in propagation-of-chaos arguments for classical Vlasov dynamics and in mean-field analyses of interacting learning systems.

Proposition 3.52 (Discrete mechanical energy dissipation)**.** *Assume that f and the particle solution are smooth enough to justify differentiating the identities above along (3.63). Define*

$$\mathcal{H}_n(X, S) := F(X) + \frac{1}{2n}\sum_{i=1}^n \|s_i\|^2, \qquad S = (s_1, \dots, s_n).$$

Then along (3.63),

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_n(X(t), S(t)) = -\frac{\gamma}{n}\sum_{i=1}^n \|s_i(t)\|^2 \leqslant 0. \tag{3.64}$$

Proof. By (3.61) and the definition $v_\alpha = -\nabla_W f(\alpha)$, one has $\nabla_{x_i} F(X) = -v_{\alpha_X}(x_i)/n$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t}F(X(t)) = \sum_{i=1}^n \langle \nabla_{x_i} F(X(t)),\, s_i(t)\rangle = -\frac{1}{n}\sum_{i=1}^n \langle v_{\alpha_t}(x_i(t)),\, s_i(t)\rangle.$$

On the other hand,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2n}\sum_i \|s_i(t)\|^2 = \frac{1}{n}\sum_i \langle s_i(t),\, -\gamma s_i(t) + v_{\alpha_t}(x_i(t))\rangle.$$

The force terms cancel, leaving (3.64). □

The identity says that friction dissipates exactly the kinetic part of the mechanical energy. In the undamped case $\gamma = 0$, the same computation gives conservation of $\mathcal{H}_n$.

Phase-space formulation. Because momentum is part of the state, the mean-field object is not merely a measure on positions. It is a law on phase space (x, s), whose spatial marginal is the current distribution of particles. Here s is a Lagrangian particle velocity; it should not be confused with the Eulerian momentum measure $m = \rho v$ used in the Benamou–Brenier convex formulation (2.10). The particle ODE then becomes a transport equation in phase space. This is the deterministic analogue of the kinetic mean-field viewpoint used for interacting particle systems; in learning, heavy-ball mean-field limits have recently been used to analyze wide networks trained with momentum Wu et al. (2022).

Proposition 3.53 (Phase-space Liouville formulation)**.** *Let v_α be smooth enough that (3.63) has a classical solution. Define the empirical phase-space measure*

$$\eta_t^n := \frac{1}{n}\sum_{i=1}^n \delta_{(x_i(t), s_i(t))} \in \mathcal{P}(\mathbb{R}^d_x \times \mathbb{R}^d_s), \qquad \alpha_t^n = (\pi_x)_\sharp \eta_t^n.$$

Then η_t^n solves, in the sense of distributions,

$$\partial_t \eta_t^n + \nabla_x \cdot (s\,\eta_t^n) + \nabla_s \cdot \big((-\gamma s + v_{\alpha_t^n}(x))\,\eta_t^n\big) = 0. \tag{3.65}$$

If, as $n \to \infty$, the empirical phase-space laws converge to a smooth density $\eta_t(x, s)$ and the nonlinear fields converge accordingly, the limiting kinetic equation is

$$\partial_t \eta_t + \nabla_x \cdot (s\,\eta_t) + \nabla_s \cdot \big((-\gamma s + v_{\alpha_t}(x))\,\eta_t\big) = 0, \qquad \alpha_t = (\pi_x)_\sharp \eta_t. \tag{3.66}$$

Proof. Let $\varphi \in C_c^\infty(\mathbb{R}^d_x \times \mathbb{R}^d_s)$. Along the particle system,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi \,\mathrm{d}\eta_t^n = \frac{1}{n}\sum_{i=1}^n \left[\langle \nabla_x \varphi(x_i,s_i),\, s_i\rangle + \langle \nabla_s \varphi(x_i,s_i),\, -\gamma s_i + v_{\alpha_t^n}(x_i)\rangle\right].$$

Rewriting the right-hand side as an integral against η_t^n and integrating by parts gives (3.65). Passing formally to the mean-field limit gives (3.66). □

Remark 3.54 (Newtonian limit). If damping is removed, or if inertia dominates friction after rescaling, (3.63) becomes

$$\ddot{x}_i(t) = v_{\alpha_t}(x_i(t)).$$

This is Newton's law with acceleration field v_{α_t}. Thus first-order Wasserstein gradient flow reads the force as a velocity, while second-order momentum flow reads the same force as an acceleration. The price is that the state is no longer just the spatial measure α_t; it is the phase-space law of position and velocity. In this limit the flow is Hamiltonian rather than dissipative whenever the force comes from a smooth interaction energy and $\gamma = 0$.

Example 3.55 (Quadratic energies and gravitational attraction). The quadratic-plus-linear energies

$$f(\alpha) = \frac{1}{2}\iint k(x,y)\,\mathrm{d}\alpha(x)\mathrm{d}\alpha(y) + \int V(x)\,\mathrm{d}\alpha(x), \qquad k(x,y) = k(y,x),$$

are the interaction energies already discussed around (3.8), up to the harmless convention of inserting the factor $1/2$. The force computed there becomes, in the present inertial setting, an acceleration:

$$v_\alpha(x) = -\int \nabla_x k(x,y)\,\mathrm{d}\alpha(y) - \nabla V(x).$$

For the attractive Newtonian kernel in three dimensions, $k(x,y) = -G/\|x-y\|$, understood with a short-distance regularization if needed, this gives

$$v_\alpha(x) = G\int \frac{y-x}{\|x-y\|^3}\,\mathrm{d}\alpha(y) - \nabla V(x),$$

the usual mean-field gravitational acceleration. The phase-space equation (3.66) is then a damped Vlasov-type equation. This kinetic viewpoint is classical in statistical physics: Boltzmann introduced the description of gases through phase-space distributions, the collisionless mean-field limit gives Liouville–Vlasov equations, and adding collision operators leads to the Boltzmann equation Cercignani et al. (1994); Villani (2002); Braun and Hepp (1977).

The numerical illustration below uses the energy-distance kernel

$$k(x,y) = -\|x-y\|, \qquad f(\alpha) = \mathrm{MMD}_k^2(\alpha,\beta) = \iint k(x,y)\,\mathrm{d}(\alpha-\beta)(x)\mathrm{d}(\alpha-\beta)(y).$$

This is the usual squared energy distance between α and β. Away from the diagonal, its Wasserstein velocity is

$$v_\alpha(x) = 2\int \frac{x-y}{\|x-y\|}\,\mathrm{d}(\alpha-\beta)(y).$$

For numerical stability, Figure 3.16 uses the smoothed distance $\sqrt{\|x-y\|^2+\delta^2}$ with a small δ. It compares the first-order flow with the undamped Newton lift $\ddot{x} = v_{\alpha_t}(x) = -\nabla_W f(\alpha_t)(x)$, initialized with zero velocity. The initial measure is a single Gaussian located to the left of a two-Gaussian target mixture. Both the transported measure and the target are discretized by large empirical clouds, and the smooth density views are KDE renderings of these particles. The trajectory panels display only a representative subset of initial particles selected by farthest-point sampling.

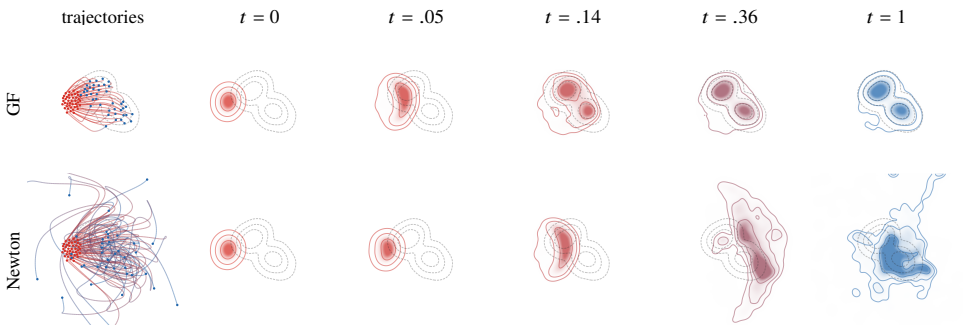


Figure 3.16: First-order and Newton Wasserstein particle flows for the squared MMD with the energy-distance kernel $k(x, y) = -\|x - y\|$. The first row follows the overdamped Wasserstein gradient flow. The second row follows the infinite-momentum Newton dynamics $\ddot{x} = -\nabla_W f(\alpha_t)(x)$, so the same Wasserstein force is read as an acceleration rather than a velocity. Both rows use the same large empirical source cloud and the same empirical two-Gaussian target cloud. Dashed gray contours show a KDE of the target particles, colored panels show KDEs of the evolving transported particles, and the left panels show farthest-point-subsampled trajectories.

Entropy without an empirical energy. The previous interaction example has a genuine value on empirical measures: once the particles are known, the double sum approximates the interaction energy. Entropy gives the complementary situation. Its Wasserstein gradient flow is the simplest diffusion equation, but the entropy itself is finite only on absolutely continuous measures. A particle method must therefore introduce a density, or score, closure before the inertial template can be applied.

Example 3.56 (Entropy-driven inertial flow). Let

$$f(\alpha) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} \rho(x) \log \rho(x)\, \mathrm{d}x, & \text{if } \alpha = \rho \mathrm{d}x, \\ +\infty, & \text{otherwise.} \end{cases}$$

Thus $f(\alpha_X) = +\infty$ for an empirical measure $\alpha_X = n^{-1} \sum_i \delta_{x_i}$, and the finite-dimensional lift $F(X) = f(\alpha_X)$ is not meaningful. For $\alpha = \rho \mathrm{d}x$ with a smooth positive density, however,

$$\delta f(\alpha)(x) = \log \rho(x) + 1, \qquad v_\alpha(x) = -\nabla \log \rho(x).$$

The first-order Wasserstein gradient flow is therefore the heat equation $\partial_t \rho_t = \Delta \rho_t$. If the spatial marginal remains smooth and positive, the corresponding formal second-order phase-space equation is

$$\partial_t \eta_t + \nabla_x \cdot (s\eta_t) + \nabla_s \cdot \big((-\gamma s - \nabla \log \rho_t(x))\eta_t\big) = 0, \qquad \rho_t(x) = \int_{\mathbb{R}^d} \eta_t(x, s)\, \mathrm{d}s. \tag{3.67}$$

This is a kinetic, inertial version of entropy diffusion: the entropy score acts as an acceleration rather than as an instantaneous velocity.

For the numerical illustration, we add the quadratic confinement $V(x) = \|x\|^2/2$. In one dimension the overdamped equation becomes the Ornstein–Uhlenbeck Fokker–Planck equation

$$\partial_t \rho_t = \partial_{xx} \rho_t + \partial_x (x \rho_t),$$

whose stationary law is the centered Gaussian $\mathcal{N}(0, 1)$. The Newton lift is discretized directly in phase space: for a density $\eta_t(x, s)$ with spatial marginal $\rho_t(x) = \int \eta_t(x, s)\, \mathrm{d}s$, it reads

$$\partial_t \eta_t + \partial_x (s\eta_t) + \partial_s \big((-\partial_x \log \rho_t(x) - x)\eta_t\big) = 0.$$

Both equations are solved by finite differences on fixed grids, with a mild grid smoothing only to evaluate the logarithmic derivative in the Newton row. The first row of Figure 3.17 shows only the spatial density of the Wasserstein gradient flow. The second and third rows show, for the Newton equation, respectively the spatial marginal and the full phase-space density; the latter uses white for zero mass and black for high mass.

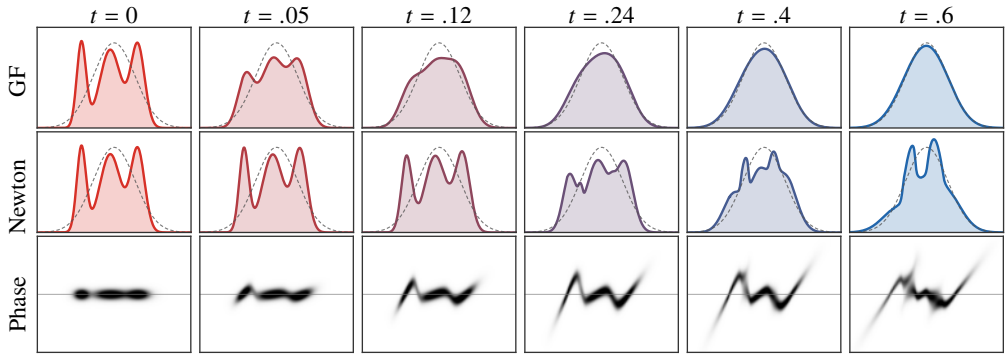


Figure 3.17: Finite-difference entropy and inertial entropy flows in one dimension. The initial spatial density is a three-Gaussian mixture with different widths and overlapping components. The top row solves the confined entropy Wasserstein gradient flow, i.e. the OU/Fokker–Planck equation, whose stationary density is the dashed centered Gaussian. The two lower rows solve the undamped Newton lift on a grid in (x, s), initialized with a narrow centered speed distribution: its spatial marginal is shown in color and the full phase-space density below in grayscale.

This chapter has moved from first-order minimizing movements to several increasingly structured dynamics. The common thread is that an energy alone does not determine an evolution: the metric or action selects the admissible tangent variables and converts the first variation into a velocity, jump flux, growth rate, fiberwise transport, or acceleration. Geodesic convexity and slope inequalities then distinguish the cases with quantitative convergence guarantees from the mean-field and inertial models that require additional problem-specific analysis.

4. Generative Models via Transportation

The preceding gradient-flow calculus is variational. Modern machine-learning models often use the same transportation language more broadly: one may prescribe an interpolation and regress its velocity, fit a one-step generator to a descent field, or view network depth as a continuous transport of token measures. The examples below separate what is genuinely a Wasserstein gradient flow from what is a transportation dynamics with a useful geometric interpretation.

4.1. Generative Models via Flow Matching

Flow matching constructs a generative map by learning the velocity field of an interpolation. The key computational insight is that a constrained continuity-equation problem can be trained by an unconstrained regression.

Generative models aim to build a transportation map T between a reference distribution α (typically an isotropic Gaussian) and the target data distribution β. Since such reference measures are non-atomic, a measurable map with $T_\sharp \alpha = \beta$ exists on standard Borel spaces, for instance by identifying both probability spaces with the unit interval and using a quantile-type rearrangement. This abstract existence statement is much weaker than having an explicit and numerically stable construction of T. Optimal transport is one approach to achieving this, but it is computationally expensive and raises questions about how to estimate it from samples. A different route is to prescribe an interpolation between noise and data, learn its velocity, and obtain T by integrating a time-dependent vector field v_t. This point of view sits at the meeting point of two literatures, surveyed from a transport perspective in Peyré (2025). The diffusion branch builds on score matching Hyvärinen (2005), denoising score matching Vincent (2011), nonequilibrium noising chains Sohl-Dickstein et al. (2015), denoising diffusion probabilistic models Ho et al. (2020), score-based generative modeling Song and Ermon (2019), and the continuous-time score-SDE/probability-flow formulation Song et al. (2021). The deterministic regression branch was introduced, essentially in parallel, under three closely related names: flow matching Lipman et al. (2023), rectified flow Liu et al. (2023), and stochastic interpolants Albergo et al. (2025). In all three cases, the computational object is a velocity field whose regression loss avoids simulating the learned ODE during training. This vector field v_t is obtained by constructing an interpolation α_t and then finding v_t using the least-squares formula (2.7). As we will explain, for a specific class of interpolation (obtained by a parametric push-forward), this v_t can be obtained by avoiding explicitly inverting a Laplacian and instead computing a simple conditional expectation. This conditional expectation can itself be estimated by solving another least-squares problem, but this time unconstrained, making the estimation feasible from finite samples of α and β.

Stochastic interpolant. The word "stochastic" can hide two different levels of randomness. We first use the simpler one: after drawing a latent variable $U \sim \pi$, the path $t \mapsto P_t(U)$ is deterministic and differentiable. The randomness only comes from the initial draw of U; after taking the push-forward law, $\alpha_t = (P_t)_\sharp \pi$ is a deterministic curve of measures and obeys an ordinary continuity equation. This is the setting behind the stochastic-interpolant construction recalled in Remark 4.3, and behind the flow-matching and rectified-flow regressions below. Genuine temporal noise, where the path itself has Brownian fluctuations, is different and is discussed in Remark 4.4.

We assume first that α_t is obtained by pushing a latent distribution $\pi \in \mathcal{P}(\mathbb{R}^{d'})$ through a time-dependent map $P_t : \mathbb{R}^{d'} \to \mathbb{R}^d$; the latent dimension d' may be larger than the data dimension d, i.e.

$$\forall t \in [0,1], \quad \alpha_t := (P_t)_\sharp \pi. \tag{4.1}$$

The basic two-endpoint construction already covers most flow-matching paths used in practice.

Example 4.1 (Linear two-endpoint deterministic interpolants). Set $d' = 2d$, write $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, and choose $P_0(x, y) = x$ and $P_1(x, y) = y$. If π has marginals (α_0, α_1), then $\alpha_t = (P_t)_\sharp \pi$ interpolates between the two endpoint laws. The simplest choices are the independent, or trivial, coupling $\pi = \alpha_0 \otimes \alpha_1$ and the straight path

$$P_t(x, y) = (1 - t)x + ty.$$

With this linear path and an arbitrary coupling π, the regression below is the common core of flow matching and rectified flow: Lipman et al. emphasize conditional probability paths and simulation-free training of continuous normalizing flows, while rectified flow emphasizes straight couplings, reflow, and the possibility of reducing transport costs and discretization error Lipman et al. (2023); Liu et al. (2023).
More complex constructions are possible when sampling from π remains simple. Static auxiliary randomness is still handled by enlarging the latent variable, while Brownian noise leads to the diffusion correction described below; this is the broader stochastic-interpolant viewpoint connecting deterministic flows, probability-flow ODEs and diffusion SDEs Albergo et al. (2025).

If $\pi = \alpha \otimes \beta$ and $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$, $\beta = \frac{1}{m}\sum_j \delta_{y_j}$, then α_t consists of $n \times m$ Dirac masses

$$\alpha_t = \frac{1}{nm} \sum_{i,j} \delta_{P_t(x_i, y_j)}.$$

If $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is a Brenier-type coupling, then $\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha$ is the so-called McCann OT interpolation.

Figure 4.1 separates the effect of the endpoint coupling from the effect of the path joining each paired endpoint.

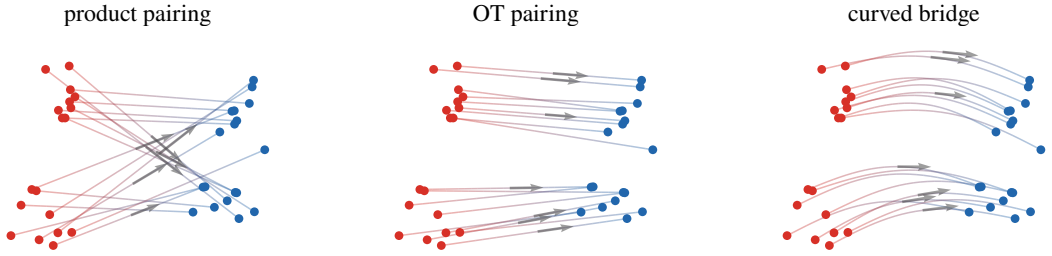


Figure 4.1: Flow matching interpolants between the same empirical source and target measures. A product-style random pairing produces crossing paths, an OT pairing gives direct displacement rays, and a curved bridge changes the path geometry while keeping the same endpoints. Gray arrows mark representative midpoint velocities $\partial_t P_t$.

Flow matching formula. This interpolation is not directly useful for sampling from β, but it can be used to define a flow field v_t so that the Eulerian advection equation (2.2) holds. This flow field is computed by solving an unconstrained least-squares problem, or equivalently, it is a conditional expectation.

Proposition 4.2 (Flow matching vector field)**.** *For each fixed t, assume $\partial_t P_t \in L^2(\pi; \mathbb{R}^d)$. Consider the flow-matching problem over measurable fields $v_t : \mathbb{R}^d \to \mathbb{R}^d$*

$$\min_{v_t} \int_{\mathbb{R}^{d'}} \|v_t(P_t(u)) - [\partial_t P_t](u)\|^2 \, \mathrm{d}\pi(u). \tag{4.2}$$

Its minimizer is characterized α_t-almost everywhere by the conditional expectation

$$v_t(z) = \mathbb{E}_{u\sim\pi}\big([\partial_t P_t](u) \,\big|\, z = P_t(u)\big). \tag{4.3}$$

Then the pair (α_t, v_t) *satisfies the continuity equation* (2.2).

Proof. We first recall the two equivalent ways of writing the interpolated measure. Formally, one may write

$$\alpha_t(z) = \int_{\mathbb{R}^{d'}} \delta(z - P_t(u))\,\mathrm{d}\pi(u),$$

while the rigorous meaning is that, for every smooth test function φ,

$$\int_{\mathbb{R}^d} \varphi(z)\,\mathrm{d}\alpha_t(z) = \int_{\mathbb{R}^{d'}} \varphi(P_t(u))\,\mathrm{d}\pi(u). \tag{4.4}$$

The minimizer in (4.2) is the orthogonal projection in $L^2(\pi;\mathbb{R}^d)$ of the latent velocity $\partial_t P_t(u)$ onto the closed subspace of functions that depend on u only through $P_t(u)$. This projection is the conditional expectation (4.3). Formally, this can be read as

$$v_t(z) = \frac{1}{\alpha_t(z)} \int_{\mathbb{R}^{d'}} \delta(z - P_t(u))[\partial_t P_t](u)\,\mathrm{d}\pi(u),$$

and rigorously it means that, for every smooth test vector field m,

$$\int \langle m(z),\, v_t(z)\rangle\,\mathrm{d}\alpha_t(z) = \int \langle m(P_t(u)),\, [\partial_t P_t](u)\rangle\,\mathrm{d}\pi(u). \tag{4.5}$$

We now prove that this field transports the curve $(\alpha_t)_t$. The weak form of $\partial_t \alpha_t + \mathrm{div}(\alpha_t v_t) = 0$ is that, for every smooth scalar test function φ,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z)\,\mathrm{d}\alpha_t(z) - \int \langle v_t(z),\, \nabla\varphi(z)\rangle\,\mathrm{d}\alpha_t(z) = 0. \tag{4.6}$$

Using (4.4) and differentiating under the integral sign gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \varphi(z)\mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(P_t(u)),\, [\partial_t P_t](u)\rangle \mathrm{d}\pi(u). \tag{4.7}$$

On the other hand, applying (4.5) with $m = \nabla\varphi$ gives

$$\int \langle v_t(z),\, \nabla\varphi(z)\rangle\,\mathrm{d}\alpha_t(z) = \int \langle \nabla\varphi(P_t(u)),\, [\partial_t P_t](u)\rangle \mathrm{d}\pi(u). \tag{4.8}$$

Comparing (4.7) and (4.8) yields (4.6), which is the desired continuity equation. □

The conditional expectation in (4.3) has a simple measure-theoretic meaning. Let $\alpha_t = (P_t)_\sharp \pi$ and define the vector-valued flux measure ω_t on $\mathbb{R}^d$ by

$$\int_{\mathbb{R}^d} \langle \psi(z),\, \mathrm{d}\omega_t(z)\rangle \coloneqq \int_{\mathbb{R}^{d'}} \langle \psi(P_t(u)),\, [\partial_t P_t](u)\rangle \mathrm{d}\pi(u)$$

for every bounded continuous vector field ψ. Since $\alpha_t(A) = 0$ implies $\pi(P_t^{-1}(A)) = 0$, one has $\omega_t \ll \alpha_t$. The Radon–Nikodym decomposition of ω_t with respect to α_t is therefore

$$\mathrm{d}\omega_t(z) = v_t(z)\mathrm{d}\alpha_t(z), \qquad v_t = \frac{\mathrm{d}\omega_t}{\mathrm{d}\alpha_t}.$$

In the language of Lebesgue decomposition, ω_t has only an absolutely continuous part with respect to α_t and no singular part; the conditional expectation is precisely its density. This agrees with the flux notation used in the dynamic formulation of Section 2.2. Equivalently, disintegrating π with respect to the map P_t gives $\pi(\mathrm{d}u) = \pi_{t,z}(\mathrm{d}u)\alpha_t(\mathrm{d}z)$, where $\pi_{t,z}$ is supported on the fiber $\{u : P_t(u) = z\}$, and

$$v_t(z) = \int_{\{P_t(u)=z\}} [\partial_t P_t](u)\mathrm{d}\pi_{t,z}(u).$$

Thus the solution of (4.2) is the conditional expectation of the velocities $\partial_t P_t$: intuitively, $v_t(z)$ is the average velocity of all trajectories passing through z. Numerically, $(x,t) \to v_t(x)$ can be parameterized by a neural network (e.g., a U-Net for vision tasks) and estimated using stochastic gradient descent on the objective in (4.2).

For the exact field v_t, integrating the ODE $\dot{x} = v_t(x)$ defines a transport map T_t. If v_t is regular enough, or more generally if the continuity equation has a unique solution for this velocity, then $(T_t)_\sharp \alpha_0 = \alpha_t$. Thus the same interpolation as (4.1) is represented by a deterministic flow rather than by the original coupling. The sampling procedure first draws $X_0 \sim \alpha$ and then integrates the ODE $\dot{X}_t = v_t(X_t)$ starting with $X_{t=0} = X_0$. In the ideal exact-field limit, the resulting $X_{t=1}$ is distributed according to $\alpha_1 = \beta$.

Remark 4.3 (Static-noise stochastic interpolants). In the terminology of Albergo–Boffi–Vanden-Eijnden Albergo et al. (2025), a stochastic interpolant is not first defined as an SDE. It is an explicit random bridge

$$X_t = I_t(X_0, X_1, Z), \qquad X_0 \sim \alpha_0, \quad X_1 \sim \alpha_1,$$

where Z is an auxiliary random variable, usually Gaussian and independent of the endpoints, and where

$$I_0(x_0, x_1, z) = x_0, \qquad I_1(x_0, x_1, z) = x_1.$$

A typical spatially linear example is

$$X_t = a(t)X_0 + b(t)X_1 + \gamma(t)Z, \qquad \gamma(0) = \gamma(1) = 0.$$

The noise Z is static: conditionally on (X_0, X_1, Z), the path $t \mapsto X_t$ is differentiable. Thus this construction is exactly the previous push-forward framework with $u = (X_0, X_1, Z)$, $\pi = \mathrm{Law}(X_0, X_1, Z)$, and $P_t = I_t$. Its Eulerian velocity is therefore

$$v_t(x) = \mathbb{E}\big[\partial_t I_t(X_0, X_1, Z) \mid X_t = x\big],$$

and the interpolant density satisfies the continuity equation. The associated SDEs in the stochastic-interpolant framework are alternative sampling dynamics having the same one-time marginals; they are not the definition of the interpolant itself.

Remark 4.4 (Brownian realizations of interpolant marginals). One can also represent an interpolating marginal curve by Brownian-in-time dynamics. This is a different construction from the static-noise bridge of Remark 4.3. Let Z_t solve the Itô equation

$$\mathrm{d}Z_t = r_t(U, Z_t)\mathrm{d}t + \Sigma_t(U, Z_t)\mathrm{d}B_t, \qquad \alpha_t = \mathrm{Law}(Z_t),$$

where $U \sim \pi$ is static and B_t is Brownian motion. Define the Eulerian drift and diffusion matrix by conditioning on the observed state,

$$v_t(z) = \mathbb{E}\big[r_t(U, Z_t) \mid Z_t = z\big], \qquad D_t(z) = \mathbb{E}\big[\Sigma_t(U, Z_t)\Sigma_t(U, Z_t)^\top \mid Z_t = z\big].$$

Then, for smooth test functions φ,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi \mathrm{d}\alpha_t = \int \langle \nabla\varphi,\, v_t\rangle \mathrm{d}\alpha_t + \frac{1}{2}\int \mathrm{Tr}\big(D_t\nabla^2\varphi\big)\mathrm{d}\alpha_t,$$

or, in distributional form,

$$\partial_t\alpha_t + \mathrm{div}(\alpha_t v_t) = \frac{1}{2}\sum_{i,j}\partial^2_{ij}\big((D_t)_{ij}\alpha_t\big). \tag{4.9}$$

Thus the natural noisy analogue of (4.2) regresses the instantaneous drift,

$$\min_w \mathbb{E}\big[\|w_t(Z_t) - r_t(U, Z_t)\|^2\big],$$

and learns the drift term of a Fokker–Planck equation, not a pure continuity equation unless the diffusion tensor vanishes. When $\alpha_t = \rho_t \mathrm{d}x$ has a smooth positive density, the same marginal curve can be represented, at least formally, by a probability-flow ODE

$$\partial_t\rho_t + \mathrm{div}(\rho_t\bar{v}_t) = 0, \qquad \bar{v}_t = v_t - \frac{1}{2\rho_t}\mathrm{div}(\rho_t D_t),$$

where the divergence of the matrix field $\rho_t D_t$ is taken row-wise. In the scalar spatially homogeneous case $D_t = \sigma_t^2\mathrm{Id}$, this reduces to

$$\bar{v}_t = v_t - \frac{\sigma_t^2}{2}\nabla\log\rho_t,$$

which is the familiar score correction relating diffusion SDEs to probability-flow ODEs.

Connection with diffusion models. In the special case where $P_t(x, y) = (1-t)x + ty$ is a linear interpolation and $\pi = \alpha \otimes \beta$, the curve α_t is a convolution of rescaled versions of α_0 and α_1. The flow-matching problem (4.2)

becomes

$$\min_{(v_t)_t} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|v_t((1-t)x + ty) - (y-x)\|^2 \, \mathrm{d}\alpha_0(x) \mathrm{d}\alpha_1(y).$$

When one endpoint is an isotropic Gaussian, this construction is closely related to the probability-flow formulation of diffusion models, up to the usual change of time parametrization Song et al. (2021). This is why flow matching can be viewed both as a deterministic alternative to diffusion training and as a common language for diffusion paths, OT-inspired paths, and rectified paths Lipman et al. (2023); Liu et al. (2023); Albergo et al. (2025). The next two propositions are written in the noising direction, from a data law α to a Gaussian; reversing time gives the corresponding sampling flow. They also give an explicit closed form for v_t and show that it is a gradient field. In this setting, v_t is also the solution of the constrained least-squares problem (2.7). The regression (4.2) is computationally simpler because the continuity equation has already been enforced by the chosen interpolant. To prove this, we rely on Tweedie's formula Efron (2011), which expresses the optimal Gaussian denoiser through the score, i.e. the gradient of the log-density.

Proposition 4.5 (Tweedie identity)**.** *Let W be a random vector in $\mathbb{R}^d$ with law $\beta \in \mathcal{P}_1(\mathbb{R}^d)$. For $\sigma > 0$, observe*

$$Z \;=\; W + \sigma\,\varepsilon, \quad \text{where } \varepsilon \sim \mathcal{N}(0, \mathrm{Id}) \text{ is independent of } W.$$

Denote by ρ_σ the smooth positive density of

$$\beta_\sigma \;=\; \beta * \mathcal{N}\big(0, \sigma^2 \mathrm{Id}\big),$$

which is the law of Z. Then the conditional mean admits the following everywhere-defined version:

$$\mathbb{E}\big[\, W \mid Z = z \big] \;=\; z \;+\; \sigma^2\, \nabla \log \rho_\sigma(z) \qquad \text{for all } z \in \mathbb{R}^d.$$

Proof. Let φ_σ be the $\mathcal{N}(0, \sigma^2 \mathrm{Id})$ density. Bayes' rule gives

$$\mathbb{E}[W \mid Z = z] = \frac{1}{\rho_\sigma(z)} \int_{\mathbb{R}^d} w\, \varphi_\sigma(z - w)\, \mathrm{d}\beta(w).$$

Differentiating the Gaussian convolution under the integral sign and using $\nabla_z \varphi_\sigma(z-w) = -\sigma^{-2}(z-w)\, \varphi_\sigma(z-w)$ yields

$$\nabla \rho_\sigma(z) = \int \nabla_z \varphi_\sigma(z - w)\, \mathrm{d}\beta(w) = -\sigma^{-2}\Big(z - \mathbb{E}[W \mid Z = z]\Big)\, \rho_\sigma(z).$$

Rearranging finishes the proof. □

Proposition 4.6 (Gaussian-endpoint flow-matching field)**.** *Let $X \sim \alpha$ and $Y \sim \mathcal{N}(0, \mathrm{Id})$ be independent. For $t \in (0,1)$ set*

$$Z_t \;=\; (1-t)\, X + t\, Y, \qquad \alpha_t = \mathrm{Law}(Z_t) = \rho_t\, \mathrm{d}x.$$

The regression minimizer $v^\star : \mathbb{R}^d \times (0,1) \to \mathbb{R}^d$ of

$$\min_v \int_0^1 \iint_{\mathbb{R}^d \times \mathbb{R}^d} \big| y - x - v\big((1-t)x + ty, t\big)\big|^2 \, \mathrm{d}\alpha(x)\, \mathrm{d}\mathcal{N}(y)\, \mathrm{d}t$$

is

$$v^\star(x,t) = -\frac{1}{1-t}\, x \;-\; \frac{t}{1-t}\, \nabla \log \rho_t(x) \qquad (x \in \mathbb{R}^d,\ t \in (0,1)).$$

In particular, for each $t \in (0,1)$ this field is a gradient field,

$$v^\star(\cdot, t) = -\nabla\left(\frac{\|\cdot\|^2}{2(1-t)} + \frac{t}{1-t} \log \rho_t \right).$$

Proof. Fix $t \in (0,1)$ and write $W = (1-t)X$, $\sigma = t$, so that $Z_t = W + \sigma Y$ matches the setting of Proposition 4.5. Conditional expectations satisfy $v^\star(z,t) = \mathbb{E}[Y - X \mid Z_t = z] = \frac{1}{t}\, \mathbb{E}[Z_t - W \mid Z_t = z] - \frac{1}{1-t}\, \mathbb{E}[W \mid Z_t = z]$. Applying Proposition 4.5 to $\mathbb{E}[W \mid Z_t = z]$ and noting $\mathbb{E}[Y \mid Z_t = z] = -t\, \nabla \log \rho_t(z)$ gives the claimed formula. □

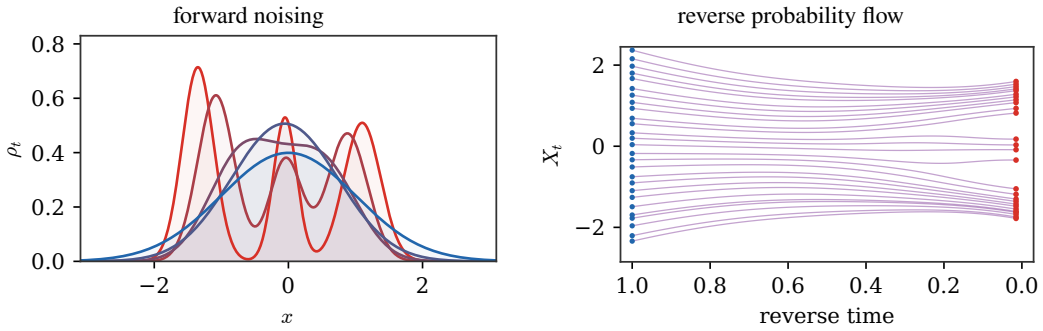


Figure 4.2: One-dimensional diffusion bridge for a Gaussian-mixture data law. The forward path $Z_t = (1-t)X + tY$ smooths the red data density toward a blue Gaussian endpoint. Reversing the probability-flow ODE transports a denser set of blue noise samples back toward the data modes, making the splitting of trajectories across mixture components visible.

Figure 4.2 shows both directions of this construction: the prescribed interpolation noising the data and the same probability-flow ODE integrated backward for sampling.

The same probability-flow intuition is visible in two dimensions. For a discrete data law, or more generally for a Gaussian mixture, the noising density is a Gaussian mixture whose score can be evaluated explicitly. This makes it possible to draw backward trajectories without training a neural network. In Figure 4.3, the Gaussian endpoint has covariance $\sigma^2\mathrm{Id}$ to keep the geometry visible at the scale of the three atoms. For a scalar noising schedule $Z_t = a_t X + b_t Y$, the intermediate law has component centers $a_t c_j$ and covariance $(b_t\sigma)^2\mathrm{Id}$. For the linear bridge, $p_t(z) = \sum_j w_j \mathcal{N}((1-t)c_j, (t\sigma)^2\mathrm{Id})$, with $s_t = \nabla \log p_t$, and the scaled version of Proposition 4.6 gives $v_t(z) = -(z + t\sigma^2 s_t(z))/(1-t)$.

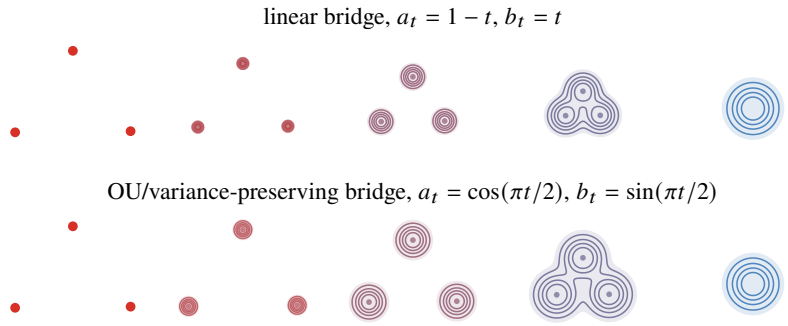


Figure 4.3: Two-dimensional noising paths from three Dirac masses to a single Gaussian. The top row shows the linear interpolation $Z_t = (1-t)X + tY$, whose component centers move linearly toward the origin and whose covariance grows like $(t\sigma)^2\mathrm{Id}$. The bottom row uses the variance-preserving Ornstein–Uhlenbeck coefficients $a_\tau = e^{-\tau}$ and $b_\tau = \sqrt{1-e^{-2\tau}}$, reparametrized by $\tau = -\log\cos(\pi t/2)$ so that $a_t = \cos(\pi t/2)$ and $b_t = \sin(\pi t/2)$. It has the same endpoints but a different speed of contraction and noising.

When is the induced map optimal? Integrating the learned velocity gives a deterministic map from α_0 to α_1, but this map is not automatically the Brenier optimal map. It is optimal only in special cases where the accumulated flow remains the gradient of a convex potential. The Gaussian product-coupling case already shows the precise obstruction: the interpolated covariances are simple, the velocity is affine, but the terminal map can contain a hidden rotational part. This phenomenon, and its extensions to rectified flows and mixtures, is analyzed in depth in Hertrich et al. (2025).

Proposition 4.7 (Gaussian flow matching and optimality)**.** *Let* $\Sigma_0, \Sigma_1 \succ 0$ *and let* $X_0 \sim \mathcal{N}(0, \Sigma_0)$ *and* $X_1 \sim \mathcal{N}(0, \Sigma_1)$ *be independent. Consider the linear flow-matching interpolation*

$$Z_t = (1-t)X_0 + tX_1, \qquad \alpha_t = \mathrm{Law}(Z_t) = \mathcal{N}(0, \Sigma_t),$$

where

$$\Sigma_t = (1-t)^2\Sigma_0 + t^2\Sigma_1. \tag{4.10}$$

Then the exact flow-matching velocity is affine, $v_t(z) = A_t z$, with

$$A_t = \big(t\Sigma_1 - (1-t)\Sigma_0\big)\Sigma_t^{-1}. \tag{4.11}$$

The induced flow map T_t^{FM} from α_0 to α_t is

$$T_t^{\mathrm{FM}} = \Sigma_0^{1/2}\Big((1-t)^2\mathrm{Id} + t^2\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\Big)^{1/2}\Sigma_0^{-1/2}. \tag{4.12}$$

In particular,

$$T_1^{\mathrm{FM}} = \Sigma_0^{1/2}\big(\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\big)^{1/2}\Sigma_0^{-1/2}. \tag{4.13}$$

This terminal map coincides with the quadratic optimal transport map

$$T^{\mathrm{OT}} = \Sigma_0^{-1/2}\big(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2}\big)^{1/2}\Sigma_0^{-1/2} \tag{4.14}$$

if and only if $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$.

Proof. The conditional-expectation formula gives

$$v_t(z) = \mathbb{E}[X_1 - X_0 \mid Z_t = z].$$

Since all variables are jointly Gaussian, this conditional expectation is linear and

$$v_t(z) = \mathrm{Cov}(X_1 - X_0, Z_t)\,\mathrm{Cov}(Z_t)^{-1}z = \big(t\Sigma_1 - (1-t)\Sigma_0\big)\Sigma_t^{-1}z,$$

which proves (4.11). To solve the characteristic equation, whiten the source by setting

$$C = \Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}, \qquad \widetilde{Z}_t = \Sigma_0^{-1/2}Z_t.$$

In these coordinates the source covariance is Id and

$$\widetilde{\Sigma}_t = (1-t)^2\mathrm{Id} + t^2C.$$

Because Id and C commute, the affine flow map in whitened coordinates is simply $\widetilde{T}_t = \widetilde{\Sigma}_t^{1/2}$. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\Sigma}_t^{1/2} = \big(tC - (1-t)\mathrm{Id}\big)\widetilde{\Sigma}_t^{-1/2},$$

which is exactly the equation $\dot{\widetilde{T}}_t = \widetilde{A}_t\widetilde{T}_t$ with $\widetilde{T}_0 = \mathrm{Id}$. Returning to the original coordinates gives (4.12), and $t = 1$ gives (4.13).

Both T_1^{FM} and T^{OT} push $\mathcal{N}(0,\Sigma_0)$ to $\mathcal{N}(0,\Sigma_1)$. The Brenier map between nondegenerate Gaussians is the unique symmetric positive definite linear map with this property. Hence $T_1^{\mathrm{FM}} = T^{\mathrm{OT}}$ if and only if T_1^{FM} is symmetric positive definite. The map T_1^{FM} is similar to $C^{1/2}$, so if it is symmetric then it is automatically positive definite. It remains to characterize symmetry. Since $C^{1/2}$ is symmetric positive definite,

$$(T_1^{\mathrm{FM}})^\top = \Sigma_0^{-1/2}C^{1/2}\Sigma_0^{1/2}.$$

Thus symmetry of T_1^{FM} is equivalent to $\Sigma_0C^{1/2} = C^{1/2}\Sigma_0$, hence to $\Sigma_0C = C\Sigma_0$ by functional calculus. Multiplying this identity on the left and right by $\Sigma_0^{1/2}$ gives $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$. Conversely, if Σ_0 and Σ_1 commute, they are orthogonally co-diagonalizable, and both (4.13) and (4.14) reduce in that basis to the diagonal map with entries $\sqrt{\lambda_{1,k}/\lambda_{0,k}}$. This proves the equivalence. □

The proposition gives a compact warning about a common overinterpretation of flow matching.

> **Remark 4.8 (Changing the bridge speed does not restore optimality).** The same terminal map (4.13) is obtained for every continuously differentiable scalar schedule $Z_t = a_tX_0 + b_tX_1$ satisfying $(a_0, b_0) = (1, 0)$, $(a_1, b_1) = (0, 1)$, and $a_t^2\mathrm{Id} + b_t^2C \succ 0$ for all t. Indeed, after whitening, the covariance path is $a_t^2\mathrm{Id} + b_t^2C$, and the flow map is its positive square root. Thus changing the speed of a nondegenerate scalar Gaussian bridge, for instance by using an OU schedule, cannot repair the non-optimality created by non-commuting covariances. Commuting covariances reduce the terminal map to independent one-dimensional scalings, whereas non-

commuting covariances create a non-symmetric affine map, hence a transport with a rotational or shearing component. More generally, mixture-like paths can create the same obstruction even when every instantaneous velocity looks natural. This distinction is closely related to counterexamples showing that flow maps associated with Fokker–Planck or diffusion-type evolutions do not in general provide optimal transport maps Lavenant and Santambrogio (2022). In particular, starting from an isotropic Gaussian does not by itself guarantee optimality once the target distribution is non-Gaussian; additional structural assumptions on the path or on the coupling are needed.

Variations on the interpolant. The geometry of the generated trajectories depends on the chosen interpolant, not only on the two endpoint laws. There is first a harmless ambiguity: a monotone reparametrization $Z_t = (1-\lambda(t))X + \lambda(t)Y$ of the linear bridge only changes the speed of the flow,

$$v_t(z) = \lambda'(t)\, v^{\mathrm{lin}}_{\lambda(t)}(z), \qquad v^{\mathrm{lin}}_r(z) = \mathbb{E}[Y - X \mid (1-r)X + rY = z].$$

It therefore leaves the spatial integral curves unchanged. Diffusion models use a genuinely different family of noising paths. If

$$Z_t = a_t X + b_t Y, \qquad Y \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}),$$

then both the mixture centers and the component variances are changed. Writing p_t for the density of Z_t and $s_t = \nabla \log p_t$, Tweedie's formula gives, away from times where $a_t = 0$,

$$v_t(z) = a_t' \,\mathbb{E}[X \mid Z_t = z] + b_t' \,\mathbb{E}[Y \mid Z_t = z] = \frac{a_t'}{a_t} z + \left(\frac{a_t' b_t^2}{a_t} - b_t' b_t \right) \sigma^2 s_t(z).$$

For the linear bridge, $a_t = 1 - t$ and $b_t = t$, this recovers the formula above. For the variance-preserving Ornstein–Uhlenbeck noising used in diffusion models,

$$a_\tau = e^{-\tau}, \qquad b_\tau = \sqrt{1 - e^{-2\tau}},$$

one obtains the forward probability-flow velocity $v_\tau(z) = -z - \sigma^2 \nabla \log p_\tau(z)$. Sampling integrates this ODE backward in τ. Equivalently, after introducing reverse time s, the velocity changes sign and becomes $z + \sigma^2 \nabla \log p_\tau(z)$. This is the noising law used in the left panel of Figure 4.4; the trajectories are more curved than for the linear bridge because the centers and variances evolve according to the OU/Fokker–Planck scaling rather than by affine interpolation. Numerically, the integration is stopped at a small positive noising time before the Dirac endpoint, where the score becomes singular.

The finite-time coefficients $a_t = \cos(\pi t/2)$ and $b_t = \sin(\pi t/2)$ are not a new spatial interpolant: they are exactly the OU coefficients after the time change $\tau = -\log \cos(\pi t/2)$. Figure 4.5 therefore compares OU with a genuinely different scalar bridge,

$$a_t = (1-t)(1-2t), \qquad b_t = t,$$

whose data coefficient changes sign before vanishing. This overshooting bridge is mainly a diagnostic example: it keeps the same endpoints, but its intermediate mixture reflects through the origin and produces visibly different reverse trajectories.

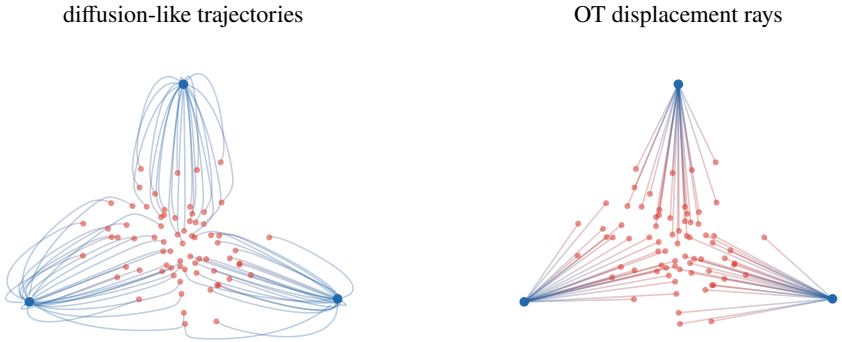


Figure 4.4: Diffusion-style sampling trajectories compared with OT rays in the three-Dirac setting of Figure 4.3. Red particles are sampled from the centered Gaussian endpoint and transported toward the three blue atoms. The left panel integrates the reverse probability-flow ODE for the variance-preserving OU noising $a_\tau = e^{-\tau}$, $b_\tau = \sqrt{1 - e^{-2\tau}}$, using the closed-form Gaussian-mixture score and stopping just before the singular Dirac endpoint. The right panel uses the straight displacement rays selected by a quadratic OT matching to the same atoms.

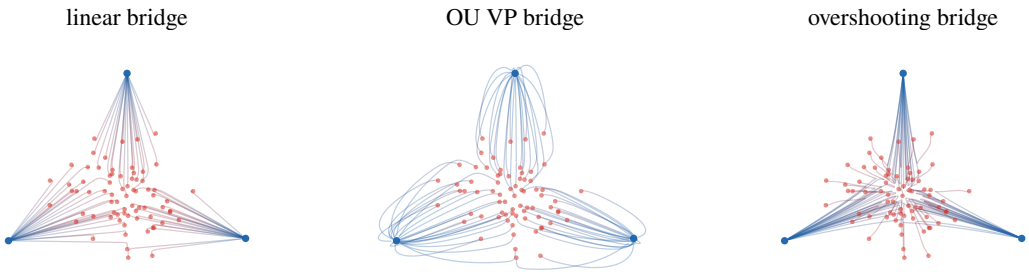


Figure 4.5: Effect of the interpolant on exact reverse-flow trajectories. All panels use the same three-Dirac target and the same Gaussian endpoint. The linear bridge $a_t = 1 - t$, $b_t = t$ produces almost radial curves. The variance-preserving OU bridge $a_\tau = e^{-\tau}$, $b_\tau = \sqrt{1 - e^{-2\tau}}$ changes the relative speed of contraction and noising. The overshooting bridge $a_t = (1 - t)(1 - 2t)$, $b_t = t$ is not a time reparameterization of either one and produces a more pronounced bending of the reverse trajectories.

4.2. One-Step Generative Models

One-step generative models try to keep the geometric training principle of flows while removing the expensive multi-step integration at sampling time. The idea is to evolve the model distribution during training, but to store the final evolution in a single generator evaluation.

One-step models via parameter-domain discrepancy flows. The most direct one-step strategy is to train the generator parameters themselves by descending a discrepancy between generated and data distributions. Let ζ be a simple latent distribution, let $f_\theta : \mathcal{Z} \to \mathcal{X}$ be a neural generator, and let β be the data distribution. The objective is to find θ such that

$$(f_\theta)_\sharp \zeta = \beta,$$

or, in practice, such that the generated law is close to β. Given a discrepancy $\mathcal{D}$ on probability measures, define

$$\mathcal{E}_\beta(\gamma) \coloneqq \mathcal{D}(\gamma, \beta), \qquad H_\beta(\theta) \coloneqq \mathcal{E}_\beta\big((f_\theta)_\sharp \zeta\big) = \mathcal{D}\big((f_\theta)_\sharp \zeta, \beta\big). \tag{4.15}$$

This viewpoint includes Wasserstein GANs, MMD-GANs and Sinkhorn generative models Arjovsky et al. (2017); Dziugaite et al. (2015); Genevay et al. (2018); when $\mathcal{D}$ is written through a dual potential or discriminator, it connects directly with adversarial training. The resulting training dynamics is the ordinary Euclidean gradient flow in parameter space,

$$\dot\theta_t = -\nabla_\theta H_\beta(\theta_t). \tag{4.16}$$

This parameter flow induces a flow of generated measures, but this induced flow is not intrinsic in general. Set

$$\alpha_t \coloneqq (f_{\theta_t})_\sharp \zeta, \qquad X_t = f_{\theta_t}(Z), \qquad Z \sim \zeta.$$

If $t \mapsto \theta_t$ is smooth and f_θ is differentiable with respect to θ, then

$$\dot X_t = \partial_\theta f_{\theta_t}(Z)\dot\theta_t.$$

For α_t-a.e. x, let $\eta_{t,x}$ be the disintegration of the latent law ζ with respect to the map f_{θ_t}. Thus $\eta_{t,x}$ is supported on the fiber $\{z\ ;\ f_{\theta_t}(z) = x\}$, and for every bounded measurable ψ,

$$\int \psi(z)\,\mathrm{d}\zeta(z) = \int_{\mathcal{X}} \left(\int \psi(z)\,\mathrm{d}\eta_{t,x}(z) \right) \mathrm{d}\alpha_t(x).$$

The Eulerian velocity of the generated measure is therefore the conditional average

$$v_t(x) = \int \partial_\theta f_{\theta_t}(z)\dot\theta_t\,\mathrm{d}\eta_{t,x}(z) = -\int \partial_\theta f_{\theta_t}(z)\nabla_\theta H_\beta(\theta_t)\,\mathrm{d}\eta_{t,x}(z). \tag{4.17}$$

Indeed, for every smooth test function φ,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \varphi(x)\,\mathrm{d}\alpha_t(x) = \int \langle \nabla\varphi(x),\ v_t(x)\rangle\,\mathrm{d}\alpha_t(x),$$

which is the weak form of the continuity equation

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0. \tag{4.18}$$

The velocity (4.17) depends on the parameter value θ_t and on the latent disintegration, not only on the measure α_t. If the parametrization is non-identifiable, two different parameters can generate the same measure while inducing different velocities. Thus (4.16) is a Euclidean gradient flow on parameter space, and only after push-forward a model-dependent flow on measures. It coincides with an intrinsic Wasserstein gradient flow only in special cases where v_t agrees with the Wasserstein velocity

$$-\nabla \delta_\gamma \mathcal{E}_\beta(\gamma)\big|_{\gamma=\alpha_t}.$$

More generally, a projected Wasserstein flow on the model manifold would require projecting this intrinsic velocity onto the infinitesimal velocities attainable by variations of θ, using the $L^2(\alpha_t)$ metric. The ordinary Euclidean parameter gradient flow (4.16) does not perform this projection in general. In machine learning, however, β is accessed through minibatches. If

$$\hat{\beta}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}, \qquad Y_i \overset{\text{i.i.d.}}{\sim} \beta,$$

then the implemented update often replaces H_β by the random objective $H_{\hat{\beta}_n}$. For nonlinear discrepancies this is not, in general, an unbiased gradient estimator:

$$\mathbb{E}\left[\nabla_\theta H_{\hat{\beta}_n}(\theta)\right] \neq \nabla_\theta H_\beta(\theta).$$

The expectation is over the data minibatch. This bias is a finite-sample effect of inserting the empirical measure inside a nonlinear transport or adversarial optimization before differentiating; it is distinct from the optimization noise of stochastic gradient descent. MMD losses admit unbiased U-statistic variants, whereas OT and entropic OT objectives generally exhibit a nontrivial statistical bias–variance tradeoff.

Example 4.9 (Application to perturbation-response prediction). In perturbation biology, one observes an unperturbed control law α and a perturbed law β, but not paired cells. The goal is to learn how a new control cell would respond. Neural OT parameterizes a Monge map, a conditional Monge map or a stochastic semi-coupling and trains it from unpaired samples so that the pushed or transported control population matches β. The distinction between a coupling, a deterministic map and an out-of-sample extrapolator is essential here: the learned object should act on cells not seen during training, not merely explain one empirical coupling Bunne et al. (2022); Lübeck et al. (2022); Chen et al. (2024); Klein et al. (2023). Conditional Wasserstein flows in Section 3.8 give a related geometric language when the conditioning variable is depth, context or treatment.

One-step model using Wasserstein flow of discrepancy. This construction keeps the discrepancy-minimization viewpoint of the previous paragraph but changes the dynamics. Instead of following the Euclidean gradient of the discrepancy in parameter space, one prescribes the Wasserstein gradient flow of the generated law and then learns a generator update realizing this motion. This distribution-space, natural-gradient-type dynamics is less tied to a particular parametrization and can have better convergence behavior, in particular by reducing parameter-space traps when the discrepancy has a favorable geometry.

Let ζ be a simple latent distribution and let $\alpha_\theta = (G_\theta)_\sharp \zeta$ be the model distribution. Assume that the target data distribution is β. The Wasserstein-flow construction chooses a discrepancy

$$\mathcal{E}_\beta(\alpha),$$

for instance a smoothed $\mathrm{KL}(\alpha|\beta)$, an MMD/IPM loss, or the debiased Sinkhorn divergence $\bar{\mathcal{L}}_c^\varepsilon(\alpha,\beta)$. The associated formal descent is

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t w_t) = 0, \qquad w_t(x) = -\nabla \delta_\alpha \mathcal{E}_\beta(\alpha_t)(x). \tag{4.19}$$

Instead of integrating (4.19) at inference time, one fits, at each training time t, a parametric residual field U_{η_t} along the current model distribution:

$$\min_\eta \int \|U_\eta(x) - w_t(x)\|^2 \, \mathrm{d}\alpha_t(x). \tag{4.20}$$

In a particle or generator implementation, the learned residual is then used to update the current generator by

$$\alpha_\theta^+ = (\mathrm{Id} + \tau U_{\eta_t})_\sharp \alpha_\theta, \qquad \text{or equivalently} \qquad G_\theta^+(z) = G_\theta(z) + \tau U_{\eta_t}(G_\theta(z)).$$

This is an ideal function-space update. If the residuals were stored as a growing composition, evaluation would still require one layer per training update and the model would not be one-step. A genuine one-step implementation therefore fits the updated outputs back into a fixed generator architecture, or distills the accumulated transport into one network, so that the final sampler remains a single evaluation of G_θ.

After many training updates, the fixed-architecture or distilled generator is evaluated once at test time. This is the organizing principle behind recent one-step methods based on Wasserstein gradient flows: W-Flow uses such a construction with the Sinkhorn divergence as a tractable global discrepancy Han et al. (2026), while drifting methods evolve the generated distribution during training through a fitted vector field and also admit one-step inference Deng et al. (2026). The gradient-flow interpretation of drifting models, and its relation to KL, MMD, sliced-Wasserstein and Sinkhorn-type discrepancies, is analyzed in Gretton et al. (2026); He et al. (2026). These ideas are also connected to the Sinkhorn-type normalization dynamics used to model attention in Sinkformers Sander et al. (2022).

Sliced-Wasserstein flow. A particularly transparent instance uses the sliced objective introduced for imaging and barycenters by Rabin, Peyré, Delon and Bernot Rabin et al. (2011). Take

$$\mathcal{E}_\beta(\alpha) = \frac{1}{2}\,\mathrm{SW}_2^2(\alpha, \beta),$$

where SW_2 averages the one-dimensional quadratic Wasserstein distances of all linear projections. With the continuity-equation convention $\partial_t \alpha_t + \mathrm{div}(\alpha_t w_t) = 0$, the descent velocity points from the current projected law toward the target projected law. More precisely, assume that α has a density and sufficient moments so that the projected monotone maps are uniquely defined for σ-almost every direction. If T_θ denotes the one-dimensional monotone transport from $(P_\theta)_\sharp \alpha$ to $(P_\theta)_\sharp \beta$, then the formal $\mathcal{W}_2$-gradient-flow velocity is

$$w_{\mathrm{SW}}[\alpha, \beta](x) = \int_{\mathbb{S}^{d-1}} \big(T_\theta(P_\theta x) - P_\theta x\big)\, \theta \, \mathrm{d}\sigma(\theta). \tag{4.21}$$

The sign follows from the one-dimensional identity: the first variation of $\frac{1}{2}\mathcal{W}_2^2(\rho, \nu)$ has spatial derivative $s - T(s)$, so the descent velocity is $T(s) - s$, and composing with P_θ lifts it in the direction θ. Thus empirical implementations only require one-dimensional sorting along sampled directions, followed by averaging the lifted projected displacements. This is the sliced-Wasserstein flow studied by Cozzi and Santambrogio Cozzi and Santambrogio (2024), who prove long-time convergence under their hypotheses when the target is Gaussian and also show that the limiting characteristic map is not, in general, the optimal transport map. When both the evolving law and the target are Gaussian, the averaged velocity is affine and the flow closes on means and covariances; this finite-dimensional closure is revisited in Section 4.5, where the sliced objective appears in the Gaussian closure catalogue of Proposition 4.20.

The difference between the full $\mathcal{W}_2$ descent and its sliced analogue is already visible on a simple shape example. Figure 4.6 compares the exact empirical flow of $\frac{1}{2}\mathcal{W}_2^2(\cdot, \beta)$, where one fixed optimal assignment gives straight relaxation curves, with the $\mathcal{W}_2$-gradient flow of $\frac{1}{2}\mathrm{SW}_2^2(\cdot, \beta)$, where the velocity is recomputed from projected monotone rearrangements.

Stein variational gradient descent. Stein variational gradient descent (SVGD) is another deterministic particle flow that fits naturally in this one-step viewpoint Liu and Wang (2016). Its original motivation is Bayesian sampling: given a target probability $\beta = \rho_\beta \, \mathrm{d}x$ known through its score $\nabla \log \rho_\beta = -\nabla V$, but not necessarily through its normalizing constant, drive a particle cloud toward β without estimating the score of the current empirical law. Geometrically, this replaces the Wasserstein gradient flow of $\mathrm{KL}(\alpha|\beta)$, whose tangent norm is $L^2(\alpha)$, by the kernelized Benamou–Brenier geometry of Section 2.3, whose velocities live in a vector-valued RKHS.

For the formal density-level calculation, assume $\alpha = \rho_\alpha \, \mathrm{d}x$ and $\beta = \rho_\beta \, \mathrm{d}x$ have smooth positive densities, and let v be a smooth compactly supported vector field. For the perturbation $\alpha_\varepsilon = (\mathrm{Id} + \varepsilon v)_\sharp \alpha$, integration by parts gives the Stein first variation

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\,\mathrm{KL}(\alpha_\varepsilon|\beta)\Big|_{\varepsilon=0} = -\int \big(\langle \nabla \log \rho_\beta(x),\, v(x)\rangle + \mathrm{div}\, v(x)\big)\mathrm{d}\alpha(x).$$

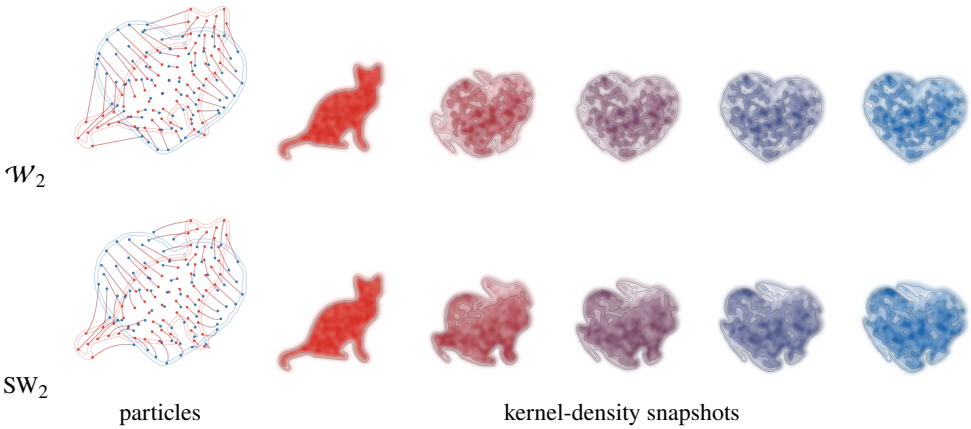


Figure 4.6: Full Wasserstein and sliced-Wasserstein flows from a cat-shaped empirical law to a heart-shaped target. The left panel in each row shows representative particle trajectories colored from red to blue. The five panels on the right render kernel-density estimates of all particles at common normalized times. The $\mathcal{W}_2$ flow follows the fixed optimal assignment and therefore straight relaxation rays, whereas the sliced flow averages one-dimensional sorted rearrangements over projection directions, producing curved trajectories and a different transient density evolution.

The bracket is the Langevin–Stein operator applied to v and averaged under α; because it only evaluates v, $\operatorname{div} v$, and the target score at sample locations, it remains meaningful when α is empirical. Optimizing this linear functional over the unit ball of $\mathcal{H}_k^d$ and using the reproducing property gives the RKHS steepest-descent direction,

$$v_\alpha^{\mathrm{SVGD}}(x) = \int \Big(k(y,x)\nabla \log \rho_\beta(y) + \nabla_y k(y,x)\Big)\mathrm{d}\alpha(y). \tag{4.22}$$

The associated mean-field equation is the transport equation

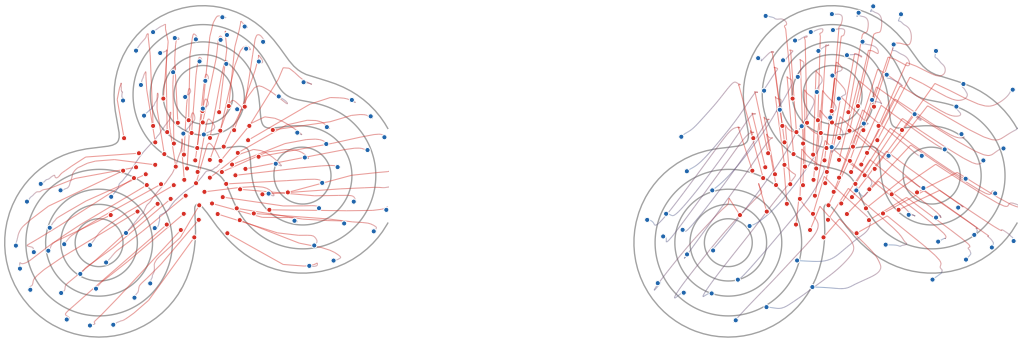
$$\partial_t \alpha_t + \operatorname{div}\big(\alpha_t v_{\alpha_t}^{\mathrm{SVGD}}\big) = 0. \tag{4.23}$$

For particles $\alpha_n^\ell = n^{-1}\sum_i \delta_{X_i^\ell}$, this becomes the explicit update

$$X_i^{\ell+1} = X_i^\ell + \tau \frac{1}{n}\sum_{j=1}^n \Big(k(X_j^\ell, X_i^\ell)\nabla \log \rho_\beta(X_j^\ell) + \nabla_{X_j} k(X_j^\ell, X_i^\ell)\Big). \tag{4.24}$$

The first term attracts particles toward increasing target log-density; the second term is a kernel repulsion that prevents immediate collapse. Liu's gradient-flow interpretation Liu (2017), the geometric analysis of Duncan, Nuesken and Szpruch Duncan et al. (2023), and the many-particle/long-time analysis of Nuesken and Renger Nüsken and Renger (2023) make precise in which sense this is a gradient flow in a Stein or kernelized transport geometry. In generative modeling terms, SVGD transports a simple empirical latent law by repeated smooth particle updates. It is therefore close in spirit to the drifting fields above, but its velocity is not learned by regression: it is the closed-form RKHS steepest-descent direction of the KL functional.

Figure 4.7 contrasts this RKHS flow with a particle closure of the Wasserstein gradient flow of $\mathrm{KL}(\alpha|\beta)$. The latter evaluates the current score $\nabla \log \rho_\alpha$ by a Gaussian KDE, while SVGD avoids this density estimate and uses the target score together with a kernel repulsion.



$\mathcal{W}_2$ KDE closure SVGD

Figure 4.7: Particle trajectories for two deterministic descents of relative entropy. Both descents target the same three-Gaussian density, shown by gray contours. The left panel approximates the $\mathcal{W}_2$ gradient flow of $\mathrm{KL}(\alpha|\beta)$ by the KDE velocity $\nabla \log \rho_\beta - \nabla \log \rho_{\alpha,h}$. The right panel uses the RBF-kernel SVGD velocity (4.22), where target-score attraction is coupled with RKHS repulsion. Both panels use the same initial particles and color trajectory time from red to blue.

Self-corrected drifting fields. Drifting methods need not start from an exact Wasserstein gradient. They often prescribe an attraction-minus-repulsion field and then regress this field in $L^2(\alpha_t)$. A simple continuous version uses a strictly positive kernel $K_\varepsilon(x,y)$ for which the following integrals are finite, and defines, for any probability measure ν,

$$B_\varepsilon[\nu](x) := \frac{\int (y-x) K_\varepsilon(x,y)\,\mathrm{d}\nu(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\nu(y)}. \tag{4.25}$$

For the Gaussian kernel $K_\varepsilon(x,y) = \exp(-\|x-y\|^2/(2\varepsilon))$, this normalized field is a score of a smoothed density:

$$B_\varepsilon[\nu](x) = \varepsilon \nabla \log\left(\int K_\varepsilon(x,y)\,\mathrm{d}\nu(y) \right). \tag{4.26}$$

The drifting velocity is then

$$u_t(x) = B_\varepsilon[\beta](x) - B_\varepsilon[\alpha_t](x) = \varepsilon \nabla \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\alpha_t(y)}. \tag{4.27}$$

The first term pulls samples toward data, while the second term corrects self-attraction and prevents all particles from collapsing onto the same high-density region. For a fixed reference measure, $B_\varepsilon[\nu]$ is precisely the Gaussian mean-shift displacement in (4.34): it moves x toward the local kernel barycenter of ν. Hence self-corrected drifting can be read as the difference between a target mean-shift field and the current model's own mean-shift field. Sinkhorn drifting replaces these one-sided kernel normalizations by two-sided entropic OT couplings, so that the cross and self terms are normalized by Sinkhorn scaling rather than by a single denominator He et al. (2026).

Figure 4.8 illustrates why the self term matters: it corrects the collapse created by target attraction alone and improves coverage of separated modes.

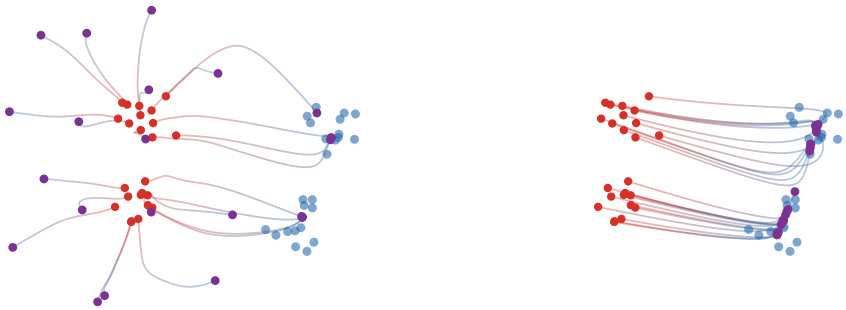
Proposition 4.10 (Instantaneous gradient representation of drifting)**.** *Let $\alpha_t = \rho_t\,\mathrm{d}x$ be a smooth curve of probability measures with positive densities, and let $u_t = \nabla\varphi_t$ be a smooth time-dependent gradient field. Define the semi-relaxed functional*

$$\mathcal{R}_t(\alpha|\alpha_t) := -\int \varphi_t(x)\,\mathrm{d}\alpha(x) + \int \varphi_t(x)\,\mathrm{d}\alpha_t(x). \tag{4.28}$$

Here α_t and φ_t are frozen when taking the first variation with respect to the first argument α. Then the continuity equation

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t u_t) = 0$$

is the formal Wasserstein gradient descent of the frozen time-dependent functional $\alpha \mapsto \mathcal{R}_t(\alpha|\alpha_t)$.



raw kernel drift self-corrected drift

Figure 4.8: Drifting trajectories for a small particle generator. The raw Laplacian-kernel drift has weak long-range attraction and can leave particles away from the data modes. The self-corrected field uses the difference $B_\varepsilon[\beta] - B_\varepsilon[\alpha_t]$, so a longer integration brings particles to the blue modes while repelling them from their own current concentration.

Proof. Since α_t and φ_t are fixed in the variation with respect to α, the first variation is

$$\delta_\alpha \mathcal{R}_t(\alpha|\alpha_t)(x) = -\varphi_t(x).$$

By Proposition 3.2,

$$\nabla_W \mathcal{R}_t(\alpha|\alpha_t) = \nabla \delta_\alpha \mathcal{R}_t(\alpha|\alpha_t) = -\nabla\varphi_t = -u_t.$$

The Wasserstein gradient-descent velocity is the negative of this gradient, namely u_t. Substituting this velocity in the continuity equation gives the claimed flow. □

The proposition is an instantaneous representation, not a claim that $(\alpha_t)_t$ is the gradient flow of one fixed energy: both φ_t and the reference argument α_t are frozen before differentiating with respect to α.

Example 4.11 (Kernel drifting as a frozen surrogate). For the Gaussian-kernel drift (4.27), set

$$\varphi_t(x) = \varepsilon \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\alpha_t(y)}.$$

Then $u_t = \nabla\varphi_t$, so Proposition 4.10 shows that kernel drifting is the Wasserstein gradient descent of

$$\mathcal{R}_t^{\mathrm{drift}}(\alpha|\alpha_t) = \varepsilon \int \log \frac{\int K_\varepsilon(x,y)\,\mathrm{d}\alpha_t(y)}{\int K_\varepsilon(x,y)\,\mathrm{d}\beta(y)}\,\mathrm{d}\alpha(x) + \text{constant}.$$

The surrogate vanishes at $\alpha = \alpha_t$, but it need not be nonnegative and is therefore not a divergence. It is "semi-relaxed" because the current model α_t is used to build the potential, but it is not varied inside the denominator when computing the first variation in α.

Remark 4.12 (General fields and projection onto gradients). A general regressed field b_t is not necessarily the minimal Wasserstein tangent representative. Such representatives belong to the $L^2(\alpha_t)$ closure of gradient fields, and fields producing the same continuity-equation variation can differ by a weighted divergence-free component. The gradient component is obtained by the weighted projection

$$\nabla\varphi_t = \operatorname*{argmin}_{\nabla\varphi} \int \|\nabla\varphi(x) - b_t(x)\|^2\,\mathrm{d}\alpha_t(x).$$

One may first normalize b_t pointwise, for instance by $b_t/(\|b_t\| + \eta)$, or globally by $\|b_t\|_{L^2(\alpha_t)}$, before this projection. Proposition 4.10 then applies to the projected field. Non-gradient components can still be useful in a parametric model, but they are not descent directions of a scalar functional for the $\mathcal{W}_2$ Riemannian metric.

4.3. Moment Measures

Moment measures give another way to make a whole distribution from one convex potential. Instead of first fixing a simple source law and then learning a transport map, one asks for a convex function whose own

log-concave density is pushed forward by its gradient. This couples sampling and mapping in a rigid way: the same potential defines both the source density and the Brenier map. The reward is a hidden convex structure: after a suitable optimal-transport reformulation, a nonlinear equation on convex functions becomes a convex minimization problem for probability densities. This is one of the cleanest places where optimal transport, Prékopa-type inequalities and convex geometry meet.

Definition 4.13 (Moment measure of a convex function)**.** Let $u : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous convex function such that

$$Z_u := \int_{\mathbb{R}^d} e^{-u(x)}\,\mathrm{d}x < +\infty.$$

The normalized log-concave measure associated with u is

$$\eta_u := Z_u^{-1} e^{-u(x)}\,\mathrm{d}x.$$

Whenever ∇u is defined η_u-almost everywhere, the moment measure of u is

$$\mathfrak{M}(u) := (\nabla u)_\sharp \eta_u.$$

The normalization removes additive constants in u. Translations of the argument are another invariance: if $u_a(x) = u(x-a)$, then η_{u_a} is the translate of η_u, while $\nabla u_a(x) = \nabla u(x-a)$, hence $\mathfrak{M}(u_a) = \mathfrak{M}(u)$. A first obstruction is immediate. Formally, if u is smooth and e^{-u} decays fast enough for the boundary term to vanish, then

$$\int y\,\mathrm{d}\mathfrak{M}(u)(y) = Z_u^{-1}\int \nabla u(x) e^{-u(x)}\,\mathrm{d}x = -Z_u^{-1}\int \nabla(e^{-u(x)})\,\mathrm{d}x = 0.$$

Thus moment measures are necessarily centered. The nonsmooth theory uses essentially continuous convex functions: these are lower-semicontinuous convex functions whose set of discontinuity points has zero $\mathcal{H}^{d-1}$ measure. Since a convex function is continuous in the interior of its effective domain, this condition controls only its boundary behavior.

Figure 4.9 shows the forward construction in one dimension. The map u' is implicit in the push-forward, but the display focuses on the two visible measures: the log-concave source $\eta_u = Z_u^{-1}e^{-u}\mathrm{d}x$ and the resulting moment measure $\mathfrak{M}(u)$.

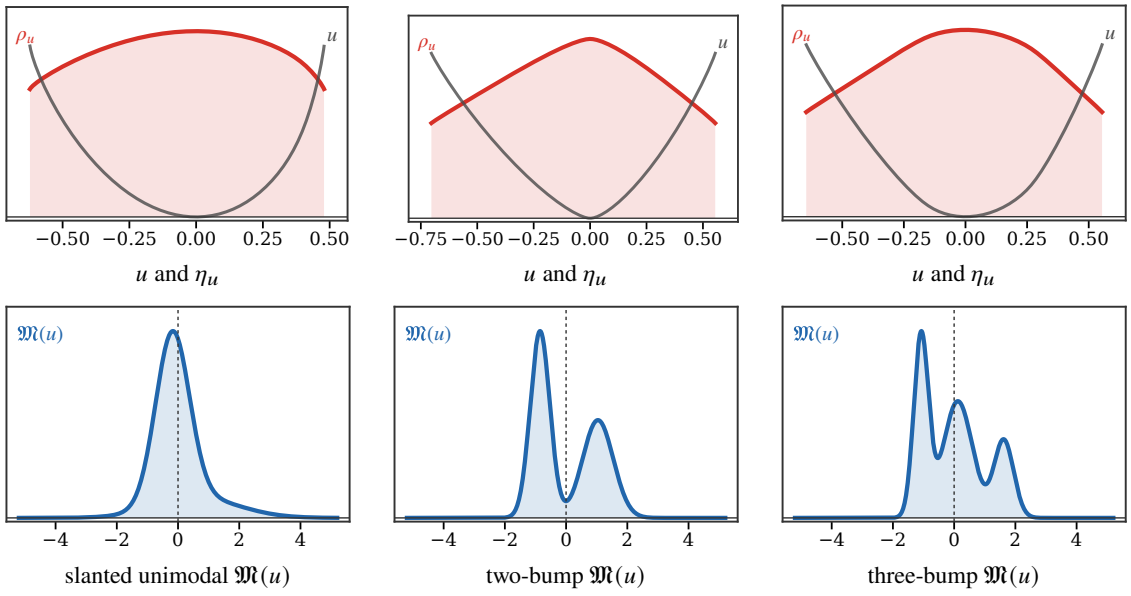


Figure 4.9: Forward moment-measure construction in one dimension. Each column shows a convex potential u chosen so that the moment measure has a prescribed shape: a skewed unimodal density, two bumps, and three bumps with different widths and heights. The top row overlays u (gray, vertically rescaled) with the density of the log-concave source $\eta_u = Z_u^{-1}e^{-u}\mathrm{d}x$ (red), while the bottom row shows $\mathfrak{M}(u) = (u')_\sharp\eta_u$ (blue); the dashed vertical line marks the zero barycenter.

Theorem 4.14 (Cordero–Erausquin–Klartag)**.** *Let $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$ be a probability measure. There exists an essentially continuous convex function u such that $\alpha = \mathfrak{M}(u)$ if and only if*

$$\int y \, \mathrm{d}\alpha(y) = 0 \qquad \text{and} \qquad \operatorname{supp}(\alpha) \text{ is not contained in a hyperplane.}$$

Under these conditions, u is unique up to translations of the argument and additive constants.

This theorem is due to Cordero–Erausquin and Klartag Cordero-Erausquin and Klartag (2015). It is a functional analogue of a Minkowski-type problem: the target measure prescribes how the gradient image of a log-concave density should be distributed. The hyperplane condition is the natural non-degeneracy assumption; otherwise the prescribed gradient image lives in a lower-dimensional affine direction and no coercive full-dimensional convex potential can be recovered.

Example 4.15 (Quadratic potentials). Let $A \in \mathbb{R}^{d\times d}$ be symmetric positive definite and

$$u(x) = \frac{1}{2} x^\top A x + c.$$

Then $\eta_u = \mathcal{N}(0, A^{-1})$ and $\nabla u(x) = Ax$, so

$$\mathfrak{M}(u) = \mathcal{N}(0, A).$$

Thus centered non-degenerate Gaussians are moment measures. This example also shows the self-dual flavor of the construction: the covariance of the log-concave source is inverted by the linear gradient map.

Optimal-transport variational formulation. Santambrogio Santambrogio (2016) reformulates the moment-measure problem as a minimization over absolutely continuous probability measures. For a centered $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$, define the maximal-correlation transport functional, with values in $\mathbb{R} \cup \{+\infty\}$,

$$C_\alpha(\eta) := \sup_{\pi \in \mathcal{U}(\eta,\alpha)} \int_{\mathbb{R}^d \times \mathbb{R}^d} x \cdot y \, \mathrm{d}\pi(x, y). \tag{4.29}$$

By Kantorovich duality for the scalar-product cost,

$$C_\alpha(\eta) = \inf_{v \text{ convex}} \left\{ \int v(x) \, \mathrm{d}\eta(x) + \int v^*(y) \, \mathrm{d}\alpha(y) \right\}, \tag{4.30}$$

where the infimum is over convex functions for which both integrals are well defined and v^* is the Legendre transform. If $\eta, \alpha \in \mathcal{P}_2(\mathbb{R}^d)$, then

$$C_\alpha(\eta) = \frac{1}{2} \int \|x\|^2 \, \mathrm{d}\eta(x) + \frac{1}{2} \int \|y\|^2 \, \mathrm{d}\alpha(y) - \frac{1}{2} \mathcal{W}_2^2(\eta, \alpha). \tag{4.31}$$

The variational problem attached to a centered target α is

$$\min_{\eta \in \mathcal{P}_1(\mathbb{R}^d)} \{\mathcal{H}(\eta) + C_\alpha(\eta)\}, \qquad \mathcal{H}(r \, \mathrm{d}x) := \int r(x) \log r(x) \, \mathrm{d}x, \tag{4.32}$$

with $\mathcal{H}(\eta) = +\infty$ when η is not absolutely continuous. The centering of α makes this functional invariant under translations of η, since translating η by a vector a changes $\int x \cdot y \, \mathrm{d}\pi$ by $a \cdot \int y \, \mathrm{d}\alpha(y) = 0$.

Proposition 4.16 (Variational characterization of moment measures)**.** *Let $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$ be centered and not supported on a hyperplane. Then the minimization problem (4.32) admits a solution, unique up to translations. Every minimizer is a probability measure with log-concave density of the form*

$$\eta = Z_u^{-1} e^{-u} \, \mathrm{d}x,$$

where u is convex, essentially continuous, and satisfies the moment-measure equation

$$\alpha = (\nabla u)_\sharp \eta.$$

Conversely, any essentially continuous convex u satisfying this equation yields a global minimizer. If $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, the objective $\eta \mapsto \mathcal{H}(\eta) + C_\alpha(\eta)$ is displacement convex along $\mathcal{W}_2$ geodesics of absolutely continuous measures in $\mathcal{P}_2(\mathbb{R}^d)$ whenever the terms are finite. When merely $\alpha \in \mathcal{P}_1(\mathbb{R}^d)$, the maximal-correlation term remains convex along such finite-second-moment geodesics by approximation; existence and uniqueness on the full $\mathcal{P}_1$ domain follow from the lower-semicontinuity argument of Santambrogio.

Proof. The existence proof has two ingredients. First, since α is centered, translating η does not change either $\mathcal{H}(\eta)$ or $C_\alpha(\eta)$, so one can center a minimizing sequence. Second, the assumption that α is not supported on a hyperplane gives a coercive lower bound of the form $C_\alpha(\eta) \geqslant c_\alpha \int \|x\| \, \mathrm{d}\eta(x)$ for centered absolutely continuous η, with $c_\alpha > 0$. Together with the lower-semicontinuity estimates for entropy and maximal correlation, this yields a minimizer Santambrogio (2016).

Let $\eta = r \, \mathrm{d}x$ be a minimizer and let u be a convex optimizer in the dual formula (4.30). Keeping u fixed and varying η in (4.32) gives the Euler equation

$$\log r(x) + 1 + u(x) = \text{constant} \qquad \text{on } \{r > 0\},$$

so $\eta = Z_u^{-1} e^{-u} \mathrm{d}x$. The optimality condition for the scalar-product transport problem says that an optimal coupling is supported on $\{(x, y) : y \in \partial u(x)\}$. Since η is absolutely continuous and u is convex, $\partial u(x) = \{\nabla u(x)\}$ for η-almost every x, hence $\alpha = (\nabla u)_\sharp \eta$.

Conversely, assume $\eta = Z_u^{-1} e^{-u} \mathrm{d}x$ and $\alpha = (\nabla u)_\sharp \eta$. Let ν be a smooth compactly supported competitor, and let T be the Brenier map from η to ν. Along the geodesic $\eta_t = ((1-t)\mathrm{Id} + tT)_\sharp \eta$, the right derivative of the entropy at $t = 0$ is

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{H}(\eta_t)\Big|_{t=0^+} = -\int \langle T(x) - x, \, \nabla u(x) \rangle \, \mathrm{d}\eta(x),$$

where the identity follows by differentiating the Jacobian formula and integrating by parts. The dual optimizer u gives the upper directional bound

$$\frac{\mathrm{d}}{\mathrm{d}t} C_\alpha(\eta_t)\Big|_{t=0^+} \leqslant \int \langle T(x) - x, \, \nabla u(x) \rangle \, \mathrm{d}\eta(x).$$

Conversely, Santambrogio's derivative estimate gives

$$\frac{\mathrm{d}}{\mathrm{d}t} C_\alpha(\eta_t)\Big|_{t=0^+} \geqslant \int \langle T(x) - x, \, \nabla u(x) \rangle \, \mathrm{d}\eta(x),$$

because $(\mathrm{Id}, \nabla u)_\sharp \eta$ is optimal for the scalar-product problem. The two bounds coincide, so the first-order terms cancel. Hence the one-sided derivative of $\mathcal{H} + C_\alpha$ at η in every such direction is zero. Displacement convexity implies global minimality, and approximation removes the smooth compact-support restriction. Strict displacement convexity of entropy gives uniqueness, except for translations; translations do not change C_α because α is centered.

It remains to justify the convexity assertion. The entropy term is displacement convex by McCann's theorem, recalled in Theorem 3.11. If α has finite second moment, identity (4.31) writes C_α as the sum of the 1-convex moment term $\eta \mapsto \frac{1}{2} \int \|x\|^2 \, \mathrm{d}\eta$ and the (-1)-convex term $\eta \mapsto -\frac{1}{2} \mathcal{W}_2^2(\eta, \alpha)$, hence C_α is displacement convex. For a target with only a finite first moment, Santambrogio obtains the same convexity along $\mathcal{P}_2$ geodesics by approximation and proves the full variational characterization by lower semicontinuity. □

Remark 4.17 (Where the convexity is hidden). If one eliminates η first, the problem becomes the convex-potential functional

$$u \mapsto \int u^*(y) \, \mathrm{d}\alpha(y) - \log\left(\int e^{-u(x)} \, \mathrm{d}x \right).$$

This is a Toland-type duality: the functional is not visibly convex as a function of u, because the first term is convex while the logarithmic partition term is concave in this parametrization. Cordero–Erausquin and Klartag make the hidden convexity visible by changing variables to the dual potential $\varphi = u^*$. Since $u = \varphi^*$ for closed convex potentials, the same functional becomes

$$\varphi \mapsto \int \varphi(y) \, \mathrm{d}\alpha(y) - \log\left(\int e^{-\varphi^*(x)} \, \mathrm{d}x \right).$$

The first term is now affine in φ. The core Prékopa–Leindler input is that $\varphi \mapsto \log \int e^{-\varphi^*}$ is concave along convex combinations of convex functions; equivalently, the negative log-partition in the display is convex. Santambrogio's formulation reveals the same mechanism in transport language: the difficult convexity becomes the displacement convexity of an entropy-plus-maximal-correlation functional in the measure variable η.

Conjugate moment measures for generation. The moment-measure factorization suggests a generative recipe: sample X from the log-concave law $Z_u^{-1} e^{-u}$ and output $\nabla u(X)$. This ties sampling and mapping through the same

convex potential. Vesseron, Béthune and Cuturi Vesseron et al. (2025) argue that this direct factorization can be poorly adapted to practical generative modeling, and propose instead the conjugate factorization

$$\beta = (\nabla w^*)_\sharp \left(Z_w^{-1} e^{-w(z)} \, \mathrm{d}z \right). \tag{4.33}$$

Here ∇w^* is the Brenier map from the learned log-concave source to the target distribution β. This keeps the single-convex-potential philosophy, but places the transport map on the conjugate side; it can be parameterized by input-convex neural networks and trained using OT solvers. From the viewpoint of this chapter, moment measures are therefore a rigorous convex-analytic prototype for one-step generators based on gradients of convex potentials.

4.4. Evolution in Depth of Transformers

Deep residual architectures can be read as time discretizations of ODEs or PDEs. For transformers, the transported objects are token measures and the velocity is induced by attention.

Transformers were introduced as sequence-to-sequence architectures driven by self-attention Vaswani et al. (2017) and have since become a central architecture for language and vision models Brown et al. (2020); Dosovitskiy et al. (2021). Their distinctive feature is that each token is updated by a data-dependent average of all other tokens. This makes an attention layer permutation-equivariant before positional encoding, context dependent after conditioning on the input sequence, and naturally compatible with a measure viewpoint in which a prompt is regarded as an empirical distribution of tokens.

The mathematical limit used below concerns depth rather than model scale: one lets the number of residual attention layers grow while each layer makes a small update, as in continuous-depth neural networks Chen et al. (2018). For attention, the resulting velocity is nonlinear in the current token law because it is normalized by the whole context. This measure-theoretic view appears in the analysis of attention as a Lipschitz or interacting-particle operator Vuckovic et al. (2020); Geshkovski et al. (2025), in the Sinkhorn-normalized dynamics of Sinkformers Sander et al. (2022), and in recent well-posedness and mean-field-limit results for several attention mechanisms Castin et al. (2025). It also separates the infinite-depth limit studied here from the token-limit question, where one controls how a finite empirical context approximates its limiting attention operator Bohbot et al. (2025).

We now consider very deep transformers, focusing on a single-head attention mechanism for simplicity while ignoring MLP layers, layer normalization, causality, and masking. This stripped-down framework is best suited to modeling encoders and vision transformers; the references above indicate which parts of this simplified picture extend to richer attention mechanisms.

Attention as a context-dependent velocity. After tokenization, embedding, and positional encoding, each input is represented as a point cloud $(x_i)_{i=1}^n$ of n vectorized tokens. An attention layer with a skip connection and residual scale $1/T$, where T is the depth, transforms the tokens according to

$$x_i \mapsto x_i + \frac{1}{T} \sum_j \frac{e^{\langle Qx_i, Kx_j \rangle} V x_j}{\sum_\ell e^{\langle Qx_i, Kx_\ell \rangle}},$$

where $\theta = (K, Q, V)$ denotes the three parameter matrices. The conventional factor $1/\sqrt{r}$, with r the query/key dimension, can be absorbed into Q or K and is omitted here.

Token measure evolution. To handle an arbitrary number of tokens, we define $\alpha = \frac{1}{n} \sum_i \delta_{x_i}$ as the empirical measure of tokens and rewrite the transformer mapping as:

$$x_i \mapsto x_i + \frac{1}{T} \Gamma_\theta[\alpha](x_i),$$

where

$$\Gamma_\theta[\alpha](x) := \frac{\int e^{\langle Qx, Ky \rangle} V y \, \mathrm{d}\alpha(y)}{\int e^{\langle Qx, Kz \rangle} \, \mathrm{d}\alpha(z)}.$$

At the level of the token distribution, the layer uses the context-dependent velocity $\Gamma_{\theta_t}[\alpha]$ and pushes α forward by $\mathrm{Id} + \tau \Gamma_{\theta_t}[\alpha]$. This map depends on the whole context α as well as on the depth-dependent parameters θ_t. Denoting normalized depth by $t \in [0, 1]$ and setting $\tau = 1/T$ gives

$$\alpha_{t+\tau} = (\mathrm{Id} + \tau \Gamma_{\theta_t}[\alpha_t])_\sharp \alpha_t.$$

As $\tau \to 0$, this converges formally to the conservation equation

$$\partial_t \alpha_t + \mathrm{div}(\alpha_t \Gamma_{\theta_t}[\alpha_t]) = 0.$$

L^2 attention and mean shift. A particularly geometric variant replaces the dot-product score $\langle Qx, Ky\rangle$ by a negative squared Euclidean score $s_\varepsilon(x,y) = -\|x-y\|^2/(2\varepsilon)$. Take, for simplicity, the same token space for queries, keys and values, and set

$$K_\varepsilon(x,y) = \exp(-\|x-y\|^2/(2\varepsilon)), \qquad \rho_\varepsilon[\alpha](x) = \int K_\varepsilon(x,y)\mathrm{d}\alpha(y), \qquad m_\varepsilon[\alpha](x) = \frac{\int yK_\varepsilon(x,y)\mathrm{d}\alpha(y)}{\rho_\varepsilon[\alpha](x)}.$$

The map $x \mapsto m_\varepsilon[\alpha](x)$ is exactly Gaussian-kernel attention, i.e. normalized kernel regression over tokens; such L^2 or Gaussian-kernel scores are used explicitly in transformer variants motivated by Lipschitz control and projection-free attention Kim et al. (2020); Kundu et al. (2026). Classical mean shift, however, uses the displacement from the current point to this local barycenter. This gives

$$M_\varepsilon[\alpha](x) := m_\varepsilon[\alpha](x) - x = \frac{\int (y-x)K_\varepsilon(x,y)\mathrm{d}\alpha(y)}{\rho_\varepsilon[\alpha](x)} = \varepsilon\nabla\log\rho_\varepsilon[\alpha](x) \tag{4.34}$$

and, when α is empirical, $\rho_\varepsilon[\alpha]$ is a Gaussian kernel density estimate up to normalization. Thus $M_\varepsilon[\alpha]$ is the classical Gaussian mean-shift vector Fukunaga and Hostetler (1975); Cheng (1995); Comaniciu and Meer (2002). If the data measure is frozen, $x \leftarrow m_\varepsilon[\alpha](x)$ is the classical mode-seeking iteration. If every support point moves and the empirical measure is recomputed, one obtains the self-consistent, or blurring, mean-shift process Chen (2015). Its damped update

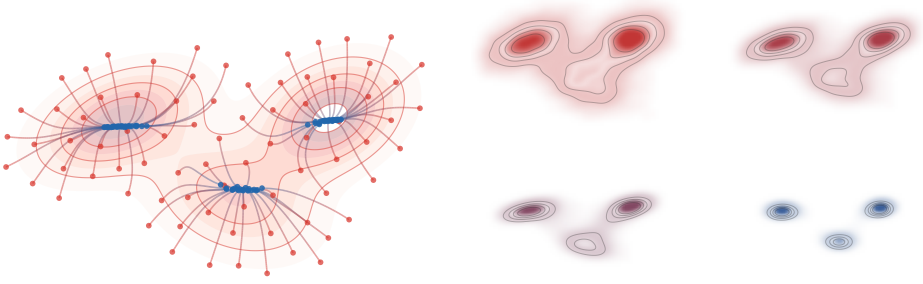
$$x_i^{k+1} = (1-\tau)x_i^k + \tau m_\varepsilon[\alpha_k](x_i^k) = x_i^k + \tau M_\varepsilon[\alpha_k](x_i^k)$$

is an explicit Euler step of the continuous-time mean-shift equation

$$\partial_t\alpha_t + \mathrm{div}\big(\alpha_t M_\varepsilon[\alpha_t]\big) = 0.$$

For $\tau = 1$ this is discrete blurring mean shift; for small τ it approximates a transport PDE that moves each token uphill along the log of the smoothed token density. This distinction between the frozen and self-consistent processes, and between the raw barycentric output m_ε and the velocity $M_\varepsilon = m_\varepsilon - \mathrm{Id}$, is important: adding m_ε directly as a residual would produce a different drift. The mean-shift form isolates a purely metric attention mechanism from the learned bilinear geometry of $\langle Qx, Ky\rangle$.

Figure 4.10 visualizes this positive-feedback mechanism before complete modal collapse: particles climb the current smoothed density while the density itself sharpens.



trajectories on the initial density later density snapshots

Figure 4.10: Continuous-time mean shift for a densely sampled three-Gaussian mixture. Left: initial density level sets, in red, and representative particle paths of $\dot{x} = M_\varepsilon[\alpha_t](x)$, colored from red to blue. Right: four later kernel-density renderings of α_t at increasing times, with the same red-to-blue time palette; the initial density is omitted because it is shown on the left. The snapshots are chosen before complete mode collapse, so that the flow visibly advects mass uphill along $\log\rho_\varepsilon[\alpha_t]$ and sharpens the overlapping modes.

Consensus contraction. When the averaging law evolves with the particles, blurring mean shift is also a consensus model, closely related to Hegselmann–Krause and Motsch–Tadmor opinion dynamics Hegselmann and Krause (2002); Blondel et al. (2009); Motsch and Tadmor (2014). This viewpoint gives a direct convergence mechanism. For a positive kernel K and an empirical law $\alpha = \sum_i a_i\delta_{x_i}$, define

$$(\mathsf{P}_X)_{ij} = \frac{a_j K(x_i,x_j)}{\sum_r a_r K(x_i,x_r)}. \tag{4.35}$$

Then P_X is row-stochastic and the damped update is $X^+ = ((1-\tau)I + \tau\mathsf{P}_X)X$. Its contraction is measured by the Dobrushin coefficient

$$\delta(\mathsf{P}) = \frac{1}{2}\max_{i,\ell}\sum_j |\mathsf{P}_{ij} - \mathsf{P}_{\ell j}| = 1 - \min_{i,\ell}\sum_j \min\{\mathsf{P}_{ij}, \mathsf{P}_{\ell j}\}. \tag{4.36}$$

This is the variation-seminorm analogue of projective contraction for positive operators: it measures how much two normalized averaging rows can still disagree Gaubert and Qu (2015).

Theorem 4.18 (Dobrushin consensus for positive blurring mean shift)**.** *Let $\alpha_0 = \sum_i a_i \delta_{x_i^0}$ have positive weights, and let C_0 be the convex hull of its support. Assume that $K : C_0 \times C_0 \to (0, +\infty)$ is continuous. Set*

$$\bar{\delta} = \sup_{X \in C_0^n} \delta(\mathsf{P}_X) < 1, \qquad q_\tau = 1 - \tau(1-\bar{\delta}), \qquad 0 < \tau \leqslant 1.$$

For the iteration $X^{k+1} = ((1-\tau)I + \tau\mathsf{P}_{X^k})X^k$,

$$\max_{i,j} \|x_i^k - x_j^k\| \leqslant q_\tau^k \max_{i,j} \|x_i^0 - x_j^0\|. \tag{4.37}$$

Consequently all particles converge to a common point in C_0. In particular, strict positivity on the compact invariant hull rules out persistent multimodal clustering; several limiting clusters require a kernel with vanishing or asymptotically negligible interactions.

Proof. Every updated point is a convex combination of the current points, hence the convex hulls are nested. For a scalar vector $z \in \mathbb{R}^n$, the common-mass decomposition of two rows of a stochastic matrix gives

$$\max_i (\mathsf{P}z)_i - \min_i (\mathsf{P}z)_i \leqslant \delta(\mathsf{P})(\max_i z_i - \min_i z_i).$$

Apply this inequality to every projection $z_i = \langle e, x_i^k\rangle$, $\|e\| = 1$, and to $(1-\tau)I + \tau\mathsf{P}_{X^k}$, whose Dobrushin coefficient is at most $1 - \tau(1 - \delta(\mathsf{P}_{X^k})) \leqslant q_\tau$. Taking the supremum over e yields (4.37). The increments are bounded by τ times the current diameter, so the geometric bound makes every particle trajectory Cauchy; the vanishing diameter forces a common limit. Finally, continuity and strict positivity of K on the compact set $C_0 \times C_0$ imply $\bar{\delta} < 1$. □

Gradient structure and limitations. When the token space has dimension d and the query/key space has dimension r, take $Q, K \in \mathbb{R}^{r\times d}$ and $V \in \mathbb{R}^{d\times d}$. If $V = Q^\top K$, the field $\Gamma_\theta[\alpha]$ is a gradient vector field in the token variable. Indeed, define the log-partition potential

$$\Phi_\alpha(x) = \int \exp(\langle Qx, Ky\rangle)\mathrm{d}\alpha(y), \qquad U_\alpha(x) = \log \Phi_\alpha(x).$$

Then

$$\nabla_x U_\alpha(x) = \frac{\int Q^\top K y \exp(\langle Qx, Ky\rangle)\mathrm{d}\alpha(y)}{\int \exp(\langle Qx, Kz\rangle)\mathrm{d}\alpha(z)} = \Gamma_\theta[\alpha](x).$$

This proves only that the velocity is an instantaneous gradient in x; it does not by itself identify a Wasserstein energy. Indeed, the natural scalar candidate

$$f_{\mathrm{att}}(\alpha) = \int U_\alpha(x)\,\mathrm{d}\alpha(x)$$

has first variation

$$\delta f_{\mathrm{att}}(\alpha)(z) = U_\alpha(z) + \int \frac{\exp(\langle Qx, Kz\rangle)}{\Phi_\alpha(x)}\,\mathrm{d}\alpha(x),$$

up to an additive constant. The second term is the response of every query normalization to a perturbation of the key distribution, and its spatial gradient is absent from $\Gamma_\theta[\alpha]$. Thus, without additional symmetry or integrability conditions, the displayed attention PDE is a transportation dynamics rather than the Wasserstein gradient flow of this fixed scalar functional. Special variants recover additional structure: Sinkhorn attention can be interpreted through doubly stochastic normalization and Wasserstein-type gradient flows Sander et al. (2022); Castin et al. (2025), while layer normalization leads naturally to dynamics on the sphere and to modified metrics. The key open difficulty for the present viewpoint is training: after the architecture has been rewritten as a controlled transport equation, learning corresponds to optimizing the time-dependent parameters $(\theta_t)_t$ rather than merely analyzing the forward PDE for fixed parameters.

4.5. Flows over the Gaussian Manifold

Gaussian measures provide a useful testing ground for the preceding dynamics. They are not invariant under a general Wasserstein gradient flow: a nonlinear velocity usually creates non-Gaussian densities immediately. The useful substitute is to either identify affine velocities, which exactly preserve Gaussianity, or to project the dynamics onto the Gaussian manifold. In both cases the measure PDE reduces to matrix ODEs for the mean and covariance. This viewpoint is emphasized in the survey Peyré (2025) and is useful for comparing diffusion paths, Wasserstein gradient flows, drifting fields and transformer-type dynamics.

For constrained gradient flows on this family, the covariance equation is the finite-dimensional Bures–Wasserstein gradient flow on positive definite matrices. Thus Gaussian closure is not just a computational shortcut: it is the restriction of Wasserstein geometry to the Gaussian submanifold, where affine gradient fields encode tangent vectors. Figure 4.11 first compares three bridge-type Gaussian closures from a source α_0 to a target γ; the exact gradient-flow closures for specified energies $f(\alpha)$ are catalogued afterwards in Proposition 4.20.

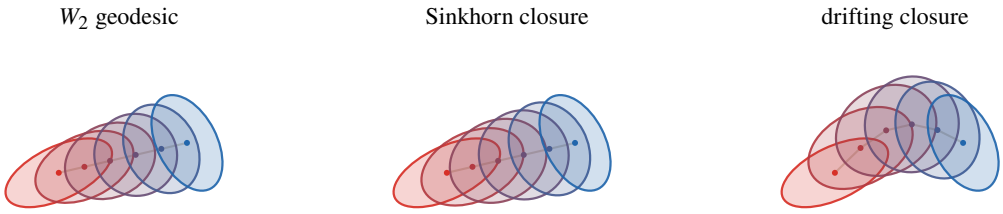


Figure 4.11: Gaussian closures from a red source α_0 to a blue target $\gamma = \mathcal{N}(\bar{m}, \bar{\Sigma})$. The left panel is the constant-speed W_2 geodesic, equivalently the displacement interpolation minimizing the Benamou–Brenier action between α_0 and γ. The middle panel is an entropic Sinkhorn/Schrödinger bridge-style closure for the quadratic cost $|x-y|^2$ and regularization strength $\varepsilon > 0$; it is a bridge toward γ, not the gradient flow of a fixed energy $f(\alpha)$, and the entropic noise inflates intermediate covariances. The right panel is a prescribed non-variational drifting flow, governed by a continuity equation with an affine Gaussian-preserving velocity, chosen so that the mean follows a curved path while the covariance is moment-matched to the same endpoint γ.

Gaussianity preservation. The first question is invariance: one wants a simple criterion ensuring that the continuity equation does not leave the finite-dimensional Gaussian family.

Proposition 4.19 (Affine velocities preserve Gaussianity)**.** *Let $\alpha_0 = \mathcal{N}(m_0, \Sigma_0)$, with $\Sigma_0 \succ 0$. Let $b_t \in \mathbb{R}^d$ and $A_t \in \mathbb{R}^{d\times d}$ be locally integrable on a time interval, and let (m_t, Σ_t) solve*

$$\dot{m}_t = b_t, \qquad \dot{\Sigma}_t = A_t\Sigma_t + \Sigma_t A_t^\top, \qquad (m_{t=0}, \Sigma_{t=0}) = (m_0, \Sigma_0).$$

Then, as long as this matrix ODE is defined, $\Sigma_t \succ 0$ and

$$\alpha_t = \mathcal{N}(m_t, \Sigma_t)$$

is the solution of the continuity equation

$$\partial_t \alpha_t + \operatorname{div}(\alpha_t v_t) = 0, \qquad v_t(x) = b_t + A_t(x - m_t).$$

In particular, if a smooth functional f has a Wasserstein gradient on Gaussian measures of the affine form

$$\nabla_W f(\mathcal{N}(m,\Sigma))(x) = b_f(m,\Sigma) + A_f(m,\Sigma)(x-m),$$

with $A_f(m,\Sigma)$ symmetric, then the Wasserstein gradient flow of f, initialized from any non-degenerate Gaussian, stays Gaussian and satisfies

$$\dot{m}_t = h_f(m_t,\Sigma_t), \qquad \dot{\Sigma}_t = H_f(m_t,\Sigma_t),$$

where

$$h_f(m,\Sigma) = -b_f(m,\Sigma), \qquad H_f(m,\Sigma) = -\big(A_f(m,\Sigma)\Sigma + \Sigma A_f(m,\Sigma)\big).$$

Conversely, fix a non-degenerate Gaussian $\mathcal{N}(m,\Sigma)$. Suppose that a finite-energy Wasserstein tangent field $V_{m,\Sigma}$ is represented by a gradient in $L^2(\mathcal{N}(m,\Sigma))$ and is tangent to the Gaussian manifold at this Gaussian. Then there exist $b(m,\Sigma)$ and a symmetric matrix $A(m,\Sigma)$ such that

$$V_{m,\Sigma}(x) = b(m,\Sigma) + A(m,\Sigma)(x-m) \qquad \mathcal{N}(m,\Sigma)\text{-a.e.}$$

Consequently, under the same smoothness assumptions, if the Wasserstein gradient flow of a functional preserves the Gaussian family for all non-degenerate Gaussian initial data, then $\nabla_W f(\mathcal{N}(m,\Sigma))$ is affine on each Gaussian. Without the Wasserstein gradient representative, this converse is false because one may add velocity fields with zero $\mathcal{N}(m,\Sigma)$-weighted divergence.

Finally, any smooth Gaussian curve with positive definite covariance can be generated by an affine velocity. If one wants the velocity to be a Wasserstein tangent gradient, one chooses the unique symmetric solution of the Lyapunov equation

$$A_t\Sigma_t + \Sigma_t A_t = \dot\Sigma_t.$$

Proof. Let X_t follow the characteristic ODE $\dot X_t = b_t + A_t(X_t - m_t)$ with $X_0 \sim \mathcal{N}(m_0,\Sigma_0)$. Since $\dot m_t = b_t$, the centered variable $\tilde X_t = X_t - m_t$ solves the homogeneous linear ODE $\dot{\tilde X}_t = A_t\tilde X_t$. If Φ_t is the fundamental matrix $\dot\Phi_t = A_t\Phi_t$, $\Phi_{t=0} = \mathrm{Id}$, then

$$X_t = m_t + \Phi_t(X_0 - m_0), \qquad \Sigma_t = \Phi_t\Sigma_0\Phi_t^\top.$$

Hence X_t is Gaussian and $\Sigma_t \succ 0$, and

$$\dot\Sigma_t = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\big(\tilde X_t\tilde X_t^\top\big) = A_t\Sigma_t + \Sigma_t A_t^\top.$$

This proves Gaussian preservation and the moment ODE. The Wasserstein gradient-flow statement follows by inserting the descent velocity $v_t = -\nabla_W f(\alpha_t)$.

For the converse, fix $\alpha = \mathcal{N}(m,\Sigma)$ and denote its density by ρ. Tangency to the Gaussian manifold means that the density variation $-\operatorname{div}(\rho V)$ is generated by some moment variation $(\dot m, \dot\Sigma)$, with $\dot\Sigma$ symmetric. Set $b = \dot m$, and let $A = A^\top$ be the unique solution of

$$A\Sigma + \Sigma A = \dot\Sigma.$$

By the first part of the proposition, the affine gradient field $V_0(x) = b + A(x-m)$ generates exactly the same infinitesimal Gaussian variation. Hence

$$\operatorname{div}\big(\rho(V - V_0)\big) = 0$$

in the distributional sense. Both V and V_0 belong to the $L^2(\alpha)$ closure of gradient fields. The weighted Helmholtz decomposition therefore makes $V - V_0$ simultaneously a tangent gradient and orthogonal to every tangent gradient, so $V = V_0$ in $L^2(\alpha)$. Equivalently, for a smooth representative $V - V_0 = \nabla\psi$, integration by parts gives $\int \|\nabla\psi\|^2 \mathrm{d}\alpha = 0$. This proves that the Wasserstein tangent representative is affine. The qualification is essential: without selecting the gradient representative, one can add a nonzero field with zero α-weighted divergence without changing the Gaussian curve.

For a prescribed smooth Gaussian curve, set $b_t = \dot m_t$ and choose any matrix A_t satisfying $A_t\Sigma_t + \Sigma_t A_t^\top = \dot\Sigma_t$. Since Σ_t is positive definite, the Lyapunov map $A \mapsto A\Sigma_t + \Sigma_t A$ is invertible on symmetric matrices, which gives the unique symmetric choice when a gradient velocity is required. In that case v_t is the gradient of the quadratic potential $x \mapsto \langle b_t, x\rangle + \langle A_t(x-m_t), x - m_t\rangle/2$. □

Gaussian-preserving gradient flows. We now instantiate the affine-gradient viewpoint by tracking functionals whose full Wasserstein gradient is affine on Gaussian inputs. The catalogue below contains exact Gaussian-preserving Wasserstein flows, not projected or constrained flows; the separate constrained construction is discussed only afterwards.

Proposition 4.20 (Gaussian closure catalogue)**.** *Let $\gamma = \mathcal{N}(\bar m, \bar\Sigma)$ on $\mathbb{R}^d$, with $\bar\Sigma \succ 0$, and let the initial condition be $\alpha_0 = \mathcal{N}(m_0,\Sigma_0)$, with $\Sigma_0 \succ 0$. For each functional displayed below, let α_t be the usual Wasserstein gradient flow on $\mathcal{P}_2(\mathbb{R}^d)$, initialized at α_0. Then the Gaussian family is invariant for these dynamics: as long as the solution exists and $\Sigma_t \succ 0$,*

$$\alpha_t = \mathcal{N}(m_t,\Sigma_t),$$

and the mean and covariance satisfy

$$\dot m_t = h(m_t,\Sigma_t), \qquad \dot\Sigma_t = H(m_t,\Sigma_t).$$

Write $\delta_m = m - \bar m$, $A = \bar\Sigma^{-1}$, and

$$M_{\Sigma,\bar\Sigma} := \Sigma^{-1/2}\big(\Sigma^{1/2}\bar\Sigma\Sigma^{1/2}\big)^{1/2}\Sigma^{-1/2}.$$

With the normalizations displayed in the first column, the mean vector field h and covariance vector field H are listed in the following table. Gradients with respect to Σ are symmetric Frobenius gradients on the cone of covariance matrices.

Gaussian-preserving Wasserstein gradient flows.

Functional $f(\alpha)$	$h(m,\Sigma)$	$H(m,\Sigma)$
$g(m_\alpha,\Sigma_\alpha)$	$-\nabla_m g$	$-2(\Sigma\nabla_\Sigma g+\nabla_\Sigma g\,\Sigma)$
$\int\Big(\frac{1}{2}x^\top Bx+\langle \ell,\,x\rangle\Big)\mathrm{d}\alpha(x),\ B=B^\top$	$-(Bm+\ell)$	$-(\Sigma B+B\Sigma)$
$\frac{1}{4}\iint (x-y)^\top G(x-y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y),\ G=G^\top$	0	$-(\Sigma G+G\Sigma)$
$\mathrm{KL}(\alpha\|\gamma)$	$-A\delta_m$	$2\mathrm{Id}-\Sigma A-A\Sigma$
$\mathcal{I}(\alpha\|\gamma)$	$-2A^2\delta_m$	$4\Sigma^{-1}-2\Sigma A^2-2A^2\Sigma$
$\mathcal{W}_2^2(\alpha,\gamma)$	$-2\delta_m$	$2(M_{\Sigma,\bar\Sigma}\Sigma+\Sigma M_{\Sigma,\bar\Sigma}-2\Sigma)$
$\mathrm{MMD}_k^2(\alpha,\gamma),\ k(x,y)=\langle x,\,y\rangle^2$	$-4Rm$	$-4(\Sigma R+R\Sigma)$
$\bar{\mathcal{L}}^\varepsilon_{\|\cdot-\cdot\|^2}(\alpha,\gamma)$	$-2\delta_m$	$-2(\Sigma G_\varepsilon+G_\varepsilon\Sigma)$
$\mathrm{SW}_2^2(\alpha,\gamma)$	$-\frac{2}{d}\delta_m$	$-2(\Sigma G_{\mathrm{sw}}+G_{\mathrm{sw}}\Sigma)$

Here, in the MMD row,

$$R=\Sigma+m\,m^\top-\bar\Sigma-\bar m\,\bar m^\top.$$

For Gaussian inputs, the debiased Sinkhorn divergence has the closed form

$$\bar{\mathcal{L}}^\varepsilon_{\|\cdot-\cdot\|^2}(\alpha,\gamma)=\|\delta_m\|^2+\mathcal{B}_\varepsilon(\Sigma,\bar\Sigma)^2,$$

with the closed-form covariance gradient

$$G_\varepsilon(\Sigma,\bar\Sigma)=\tau_\varepsilon(\Sigma)-\bar\Sigma^{1/2}\tau_\varepsilon\big(B^{1/2}_{\Sigma,\bar\Sigma}\big)B^{-1/2}_{\Sigma,\bar\Sigma}\bar\Sigma^{1/2},\qquad B_{\Sigma,\bar\Sigma}=\bar\Sigma^{1/2}\Sigma\bar\Sigma^{1/2}.$$

Here τ_ε is the scalar function

$$\tau_\varepsilon(r)\coloneqq\frac{\sqrt{\varepsilon^2+16r^2}-\varepsilon}{4r},\qquad r>0,$$

applied to positive matrices by spectral calculus. Equivalently, for $M\succ 0$,

$$\tau_\varepsilon(M)=\big(\sqrt{\varepsilon^2 I+16M^2}-\varepsilon I\big)(4M)^{-1}.$$

With this convention, $G_\varepsilon=\nabla_\Sigma\mathcal{B}_\varepsilon(\Sigma,\bar\Sigma)^2$. In the sliced row, σ is the normalized spherical measure on $\mathbb{S}^{d-1}$, and

$$G_{\mathrm{sw}}(\Sigma,\bar\Sigma)=\int_{\mathbb{S}^{d-1}}\left(1-\sqrt{\frac{\theta^\top\bar\Sigma\theta}{\theta^\top\Sigma\theta}}\right)\theta\theta^\top\,\mathrm{d}\sigma(\theta).$$

Here

$$\mathcal{I}(\alpha|\gamma)=\int\left|\nabla\log\frac{\rho(x)}{\rho_\gamma(x)}\right|^2\rho(x)\,\mathrm{d}x=\int|\nabla\log\rho(x)+A(x-\bar m)|^2\,\rho(x)\,\mathrm{d}x\qquad(\alpha=\rho\,\mathrm{d}x),$$

where ρ_γ is the density of γ.

Proof. For the moment-functional row, the formula is exact on the full Wasserstein space. Indeed, write

$$m_\alpha=\int x\,\mathrm{d}\alpha(x),\qquad\Sigma_\alpha=\int(x-m_\alpha)(x-m_\alpha)^\top\mathrm{d}\alpha(x).$$

If η is a signed perturbation with zero total mass, then

$$\mathrm{d}m_\alpha[\eta]=\int x\,\mathrm{d}\eta(x),\qquad\mathrm{d}\Sigma_\alpha[\eta]=\int(x-m_\alpha)(x-m_\alpha)^\top\mathrm{d}\eta(x).$$

Thus, up to an irrelevant additive constant, the first variation is

$$\delta f(\alpha)(x)=\langle\nabla_m g,\,x\rangle+\langle\nabla_\Sigma g,\,(x-m_\alpha)(x-m_\alpha)^\top\rangle,$$

and its spatial gradient is the affine field

$$\nabla_x \delta f(\alpha)(x) = \nabla_m g + 2\nabla_\Sigma g\,(x - m_\alpha).$$

Consequently the full Wasserstein descent velocity is affine, and Proposition 4.19 gives the displayed moment equations.

For the quadratic potential row, the ambient first variation is

$$\delta f(\alpha)(x) = \frac{1}{2}x^\top Bx + \langle \ell,\, x\rangle, \qquad \nabla_x \delta f(\alpha)(x) = Bm + \ell + B(x - m),$$

which gives the displayed affine velocity. For the quadratic interaction row, the ambient first variation is

$$\delta f(\alpha)(x) = \frac{1}{2}\int (x - y)^\top G(x - y)\mathrm{d}\alpha(y), \qquad \nabla_x \delta f(\alpha)(x) = G(x - m_\alpha).$$

These first variations are quadratic in x, hence the affine Gaussian flow coincides with the full Wasserstein gradient flow.

For the KL row, write $\alpha = \rho\,\mathrm{d}x$ and let ρ_γ denote the density of γ. The ambient first variation is $\log(\rho/\rho_\gamma)$ up to an additive constant. If $\alpha = \mathcal{N}(m, \Sigma)$, then

$$\nabla \log(\rho/\rho_\gamma)(x) = -\Sigma^{-1}(x - m) + A(x - \bar{m}),$$

so the descent velocity is

$$v(x) = -A\delta_m + (\Sigma^{-1} - A)(x - m).$$

Proposition 4.19 gives $h(m, \Sigma) = -A\delta_m$ and $H(m, \Sigma) = 2\mathrm{Id} - \Sigma A - A\Sigma$. For the Fisher row, writing $u = \log(\rho/\rho_\gamma)$, the ambient first variation of $\frac{1}{2}\mathcal{I}(\alpha|\gamma)$ is, up to constants,

$$-\Delta u - \frac{1}{2}|\nabla u|^2 - \langle \nabla u,\, \nabla \log \rho_\gamma\rangle.$$

For Gaussian ρ and Gaussian γ, this is quadratic in x. Multiplying by two for $\mathcal{I}(\alpha|\gamma)$ gives the descent velocity

$$v(x) = -2A^2\delta_m + 2(\Sigma^{-2} - A^2)(x - m).$$

Hence the Wasserstein velocity is affine and the Gaussian family is closed, with the displayed h and H. Although the ambient Fisher flow is a fourth-order PDE for a generic density, it is therefore an exact finite-dimensional closure in this quadratic-reference case.

For $\mathcal{W}_2^2(\alpha, \gamma)$, the Brenier map from $\mathcal{N}(m, \Sigma)$ to γ is

$$T(x) = \bar{m} + M_{\Sigma,\bar{\Sigma}}(x - m).$$

The descent velocity for the unhalved squared distance is $2(T - \mathrm{Id})$, which gives the mean and covariance equations through Proposition 4.19. For the MMD row, $k(x, y) = \langle x,\, y\rangle^2$ identifies the kernel mean embedding with the raw second moment. Thus

$$\mathrm{MMD}_k^2(\alpha, \gamma) = \|\Sigma + m\,m^\top - \bar{\Sigma} - \bar{m}\,\bar{m}^\top\|_{\mathrm{F}}^2 = \|R\|_{\mathrm{F}}^2,$$

and the ambient first variation is $2x^\top Rx$, whose descent velocity is $-4Rx$. This gives $h(m, \Sigma) = -4Rm$ and $H(m, \Sigma) = -4(\Sigma R + R\Sigma)$.

For the Sinkhorn row, the closed Gaussian formula is a squared mean displacement plus the smoothed Bures term $\mathcal{B}_\varepsilon^2$. This is not only a formula on the Gaussian submanifold. The first variation of the entropic value with respect to the first marginal is a Sinkhorn dual potential, and for Gaussian marginals the cross and self dual potentials are quadratic. Their debiased combination is therefore quadratic, so its spatial gradient is affine. Since the restriction of the functional to Gaussians is

$$(m, \Sigma) \mapsto \|m - \bar{m}\|^2 + \mathcal{B}_\varepsilon(\Sigma, \bar{\Sigma})^2,$$

the quadratic coefficient is $G_\varepsilon = \nabla_\Sigma \mathcal{B}_\varepsilon(\Sigma, \bar{\Sigma})^2$, and the moment-functional computation gives the displayed h and H. To compute this gradient, set $B_{\Sigma,\bar{\Sigma}} = \bar{\Sigma}^{1/2}\Sigma\bar{\Sigma}^{1/2}$. Since $\psi_\varepsilon'(r) = -2\tau_\varepsilon(r)$, differentiating $\mathrm{tr}\,\psi_\varepsilon(B_{\Sigma,\bar{\Sigma}}^{1/2})$ gives

$$-\bar{\Sigma}^{1/2}\tau_\varepsilon\big(B_{\Sigma,\bar{\Sigma}}^{1/2}\big)B_{\Sigma,\bar{\Sigma}}^{-1/2}\bar{\Sigma}^{1/2},$$

whereas differentiating the self term $-\frac{1}{2}\operatorname{tr}\psi_\varepsilon(\Sigma)$ gives $\tau_\varepsilon(\Sigma)$. This proves the displayed closed form for G_ε. When $\bar\Sigma = \mathrm{Id}$, this covariance contribution is a spectral function of Σ. If λ is an eigenvalue of Σ, the corresponding scalar velocity is

$$4\sqrt{\lambda+\varepsilon^2/16} - 4\sqrt{\lambda^2+\varepsilon^2/16},$$

which gives the displayed covariance-eigenvalue equation by functional calculus.

For sliced Wasserstein, the exact ambient Wasserstein gradient is obtained by averaging the one-dimensional gradients. Each projection is Gaussian:

$$(P_\theta)_\sharp\alpha = \mathcal{N}(\langle\theta,\, m\rangle, \theta^\top\Sigma\theta), \qquad (P_\theta)_\sharp\gamma = \mathcal{N}(\langle\theta,\, \bar m\rangle, \theta^\top\bar\Sigma\theta).$$

Hence

$$\mathrm{SW}_2^2(\alpha,\gamma) = \int_{\mathbb{S}^{d-1}} \left[\langle\theta,\, \delta_m\rangle^2 + \left(\sqrt{\theta^\top\Sigma\theta} - \sqrt{\theta^\top\bar\Sigma\theta}\right)^2\right] \mathrm{d}\sigma(\theta).$$

If T_θ is the monotone transport from $(P_\theta)_\sharp\alpha$ to $(P_\theta)_\sharp\gamma$, then the descent velocity for the unhalved sliced objective is

$$2\int_{\mathbb{S}^{d-1}} \big(T_\theta(P_\theta x) - P_\theta x\big)\theta\,\mathrm{d}\sigma(\theta).$$

For Gaussian marginals, T_θ is affine. Thus this velocity is affine in x, and Proposition 4.19 applies. The spherical identity $\int \theta\theta^\top \mathrm{d}\sigma(\theta) = \mathrm{Id}/d$ gives the mean equation, and differentiating the covariance term gives G_{sw}. □

Not every PDE preserves Gaussianity exactly. Wasserstein flows of generic higher-order regularizers usually create higher moments immediately and require a Gaussian projection to close on (m,Σ). Such projected closures are still useful: they expose the finite-dimensional dynamics predicted by a variational model and make it easy to compare variational flows with non-variational affine dynamics such as drifting fields or the Gaussian transformer closure below.

Example 4.21 (Linear mean-field networks as cross-covariance flows). Consider the two-layer mean-field model of Section 3.4, and take the linear activation $\sigma(s) = s$, so that

$$\psi((u,v),z) = v\,\langle u,\, z\rangle.$$

We restrict this example to centered neuron laws,

$$\int (u,v)\mathrm{d}\alpha(u,v) = 0, \qquad \Sigma_\alpha = \begin{pmatrix}\Sigma_{uu}(\alpha) & \Sigma_{uv}(\alpha)\\ \Sigma_{vu}(\alpha) & \Sigma_{vv}(\alpha)\end{pmatrix},$$

and use the lower-left cross-covariance block

$$\Sigma_{vu}(\alpha) = \int vu^\top \mathrm{d}\alpha(u,v) \in \mathbb{R}^{d'\times d}.$$

The predictor is therefore the linear map

$$G_{\alpha_t}(z) = \Sigma_{vu}(\alpha_t)z.$$

For the squared Euclidean loss, set

$$S = \int zz^\top \mathrm{d}\rho(z,y), \qquad R = \int yz^\top \mathrm{d}\rho(z,y).$$

The learning energy of (3.28) is then the covariance functional

$$f(\alpha) = g(\Sigma_\alpha), \qquad g(\Sigma) = \frac{1}{2}\operatorname{tr}\big(\Sigma_{vu}S\Sigma_{uv}\big) - \operatorname{tr}\big(R\Sigma_{uv}\big) + \frac{1}{2}\int \|y\|^2\mathrm{d}\rho(z,y).$$

This puts the model exactly in the centered moment-functional row of Proposition 4.20. To see that it recovers the usual particle equation, write

$$E_\alpha \coloneqq \Sigma_{vu}(\alpha)S - R.$$

The first variation is

$$\delta f(\alpha)(u,v) = \langle E_\alpha,\, vu^\top\rangle = v^\top E_\alpha u.$$

Hence the particle velocity in parameter space is linear:

$$-\nabla_{(u,v)}\delta f(\alpha)(u,v) = -\begin{pmatrix} 0 & E_\alpha^\top \\ E_\alpha & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Equivalently, at the level of g,

$$\nabla_\Sigma g = \frac{1}{2}\begin{pmatrix} 0 & E_\alpha^\top \\ E_\alpha & 0 \end{pmatrix}.$$

The factor $1/2$ in the covariance gradient comes from the symmetry of Σ: the upper-right block is the transpose of the lower-left block. Substituting this gradient in Proposition 4.20 gives

$$\dot{m}_t = 0, \qquad \dot{\Sigma}_t = -(\Sigma_t L_t + L_t \Sigma_t), \qquad L_t = \begin{pmatrix} 0 & E_t^\top \\ E_t & 0 \end{pmatrix}, \qquad E_t = \Sigma_{vu}(\alpha_t)S - R.$$

Thus a centered Gaussian law of neurons remains centered Gaussian, and the dynamics is driven by the cross-covariance block alone. This exact closure is special to the linear activation; for nonlinear activations, Gaussian closures are usually projections rather than invariant families.

One-dimensional case. The one-dimensional Gaussian family makes the preceding catalogue especially transparent. Fix a target $\gamma = \mathcal{N}(\bar{m}, \bar{\sigma}^2)$, $\bar{\sigma} > 0$, and write $\alpha = \mathcal{N}(m, \sigma^2)$, $\sigma > 0$. By Proposition 1.12, $(m, \sigma) \mapsto \mathcal{N}(m, \sigma^2)$ identifies this family isometrically with the Euclidean upper half-plane. For a functional f, set $g_f(m, \sigma) = f(\mathcal{N}(m, \sigma^2))$ and $\delta = m - \bar{m}$. For the Sinkhorn row, define

$$\psi_\varepsilon(r) = -\frac{\sqrt{\varepsilon^2 + 16r^2} - \varepsilon}{2} - \frac{\varepsilon}{2}\log\left(\frac{2\varepsilon}{\sqrt{\varepsilon^2 + 16r^2} + \varepsilon}\right), \qquad r > 0.$$

The principal exact closures reduce to the following scalar landscapes.

Functional $f(\alpha)$	Restriction $g_f(m, \sigma)$
$\mathrm{KL}(\alpha\|\gamma)$	$\log\frac{\bar{\sigma}}{\sigma} + \frac{\sigma^2 + \delta^2}{2\bar{\sigma}^2} - \frac{1}{2}$
$\mathcal{W}_2^2(\alpha, \gamma)$	$\delta^2 + (\sigma - \bar{\sigma})^2$
$\bar{\mathcal{L}}^\varepsilon_{\|\cdot-\cdot\|^2}(\alpha, \gamma)$	$\delta^2 + \psi_\varepsilon(\sigma\bar{\sigma}) - \frac{1}{2}\psi_\varepsilon(\sigma^2) - \frac{1}{2}\psi_\varepsilon(\bar{\sigma}^2)$
$\mathcal{I}(\alpha\|\gamma)$	$\frac{\delta^2}{\bar{\sigma}^4} + \frac{(\sigma^2 - \bar{\sigma}^2)^2}{\sigma^2\bar{\sigma}^4}$

Here $\bar{\mathcal{L}}^\varepsilon$ denotes the debiased entropic transport cost used in the catalogue, and $\mathcal{I}$ is relative Fisher information. Since $\psi_\varepsilon(r) \to -2r$ as $\varepsilon \downarrow 0$, the Sinkhorn landscape converges to the squared-Wasserstein landscape. Figure 4.12 compares the four exact ambient Wasserstein flows; none is a projection onto the Gaussian family.

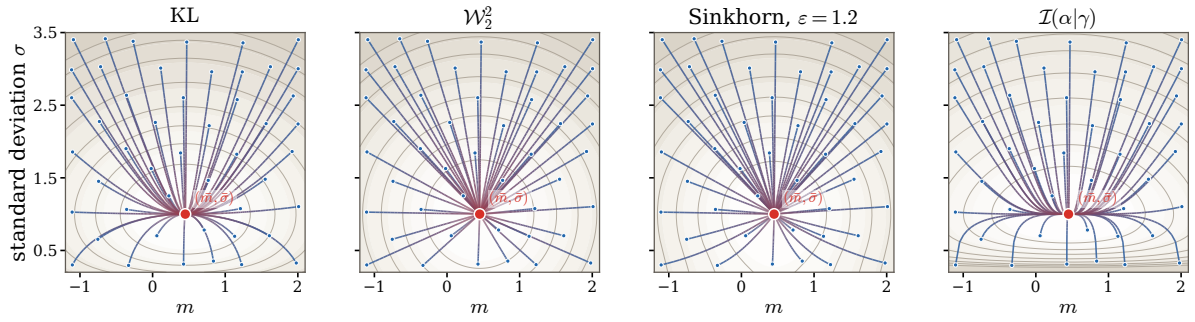


Figure 4.12: One-dimensional Gaussian Wasserstein gradient flows in the parameter half-plane (m, σ). From left to right: relative entropy, squared Wasserstein distance, debiased Sinkhorn divergence, and relative Fisher information. The trajectories start from the same well-spread initial parameters and converge to the common target $(\bar{m}, \bar{\sigma})$.

Constrained evolution on the Gaussian manifold. The preceding affine-gradient examples have a limited scope: most Wasserstein gradient flows of a functional $f(\alpha)$ are not closed on the Gaussian manifold. A nonlinear ambient velocity creates higher-order moments immediately, so the exact Gaussian closure usually fails. The

constrained viewpoint deliberately replaces the full evolution by its projection onto $\mathcal{G}$, forcing the curve to remain Gaussian while keeping the Wasserstein tangent geometry.

Let

$$\mathcal{G} = \{\mathcal{N}(m, \Sigma) : m \in \mathbb{R}^d,\ \Sigma > 0\}$$

be the Gaussian submanifold of $\mathcal{P}_2(\mathbb{R}^d)$. The Wasserstein gradient of a functional constrained to a smooth submanifold $\mathcal{M} \subset \mathcal{P}_2$ is defined as the Riesz representative of the differential restricted to tangent velocities of $\mathcal{M}$. Equivalently, it is the small-step limit of the constrained JKO scheme

$$\alpha^{k+1} \in \operatorname*{argmin}_{\alpha \in \mathcal{M}} \frac{1}{2\tau} \mathcal{W}_2^2(\alpha, \alpha^k) + f(\alpha).$$

For $\mathcal{M} = \mathcal{G}$, tangent velocities are affine gradient fields $v(x) = b + A(x - m)$ with $A = A^\top$. The constrained gradient is therefore the $L^2(\mathcal{N}(m, \Sigma))$ projection of the ambient Wasserstein gradient onto this finite-dimensional affine space, whenever the ambient gradient exists.

Proposition 4.22 (Gaussian-constrained Wasserstein gradients)**.** *Let f be a smooth functional and assume that its restriction to nondegenerate Gaussian measures can be written as*

$$f(\mathcal{N}(m, \Sigma)) = F(m, \Sigma).$$

Then the Wasserstein gradient constrained to the Gaussian family is the affine vector field

$$v_F(x) = \nabla_m F(m, \Sigma) + 2\nabla_\Sigma F(m, \Sigma)(x - m),$$

where $\nabla_\Sigma F$ denotes the symmetric matrix derivative. Equivalently, v_F is the $L^2(\mathcal{N}(m, \Sigma))$ projection of the ambient Wasserstein gradient onto affine gradient fields, whenever the ambient gradient exists. Hence the gradient descent flow constrained to Gaussian measures satisfies

$$\dot{m}_t = -\nabla_m F(m_t, \Sigma_t), \qquad \dot{\Sigma}_t = -2\big(\Sigma_t \nabla_\Sigma F(m_t, \Sigma_t) + \nabla_\Sigma F(m_t, \Sigma_t)\Sigma_t\big), \tag{4.38}$$

and the descent velocity is affine.

Proof. Test the functional along a Gaussian tangent vector, represented by an affine gradient field

$$v(x) = b + A(x - m)$$

with A symmetric. The induced first-order variations are $\dot{m} = b$ and $\dot{\Sigma} = A\Sigma + \Sigma A$. Therefore

$$\mathrm{d}F(m, \Sigma)[b, A\Sigma + \Sigma A] = \langle \nabla_m F,\ b\rangle + \operatorname{tr}(\nabla_\Sigma F(A\Sigma + \Sigma A))\,.$$

Since A, Σ and $\nabla_\Sigma F$ are symmetric, the second term equals

$$2\operatorname{tr}(\nabla_\Sigma F\, A\Sigma) = \int \langle 2\nabla_\Sigma F(x - m),\ A(x - m)\rangle \mathrm{d}\mathcal{N}(m, \Sigma)(x).$$

Together with the mean term, this gives

$$\mathrm{d}F(m, \Sigma)[\dot{m}, \dot{\Sigma}] = \int \langle v_F(x),\ v(x)\rangle \mathrm{d}\mathcal{N}(m, \Sigma)(x)$$

for all affine gradient fields v. This identifies the constrained Wasserstein gradient in the induced $L^2(\alpha)$ metric, or equivalently the projection of the ambient gradient when it exists. Substituting the descent velocity $-v_F$ in Proposition 4.19 gives (4.38). □

This proposition is the organizing rule for constrained Gaussian closures: once the scalar energy has been reduced to a function of (m, Σ), its constrained Wasserstein gradient is automatically affine and the covariance follows the Bures-type ODE (4.38). When the first variation of f is quadratic, this constrained gradient coincides with the full Wasserstein gradient.

Non-variational Gaussian-preserving flows. The last examples are not ordinary gradient flows of a fixed scalar energy on the full Wasserstein space. They preserve Gaussianity because the prescribed velocity field is affine when evaluated on Gaussian measures.

Example 4.23 (Flow matching and diffusion paths between Gaussians). Consider a prescribed Gaussian interpolation $\alpha_t = \mathcal{N}(m_t, \Sigma_t)$. Proposition 4.19 shows that an exact flow-matching velocity can be taken affine:

$$v_t(x) = \dot{m}_t + A_t(x - m_t), \qquad A_t\Sigma_t + \Sigma_t A_t = \dot{\Sigma}_t.$$

In the isotropic case $\Sigma_t = s_t^2\mathrm{Id}$, this reduces to the transparent formula

$$v_t(x) = \dot{m}_t + \frac{\dot{s}_t}{s_t}(x - m_t).$$

For instance, the diffusion noising path

$$X_t = a_t X_0 + \sigma_t Z, \qquad Z \sim \mathcal{N}(0, \mathrm{Id}),$$

has $m_t = a_t m_0$ and $\Sigma_t = a_t^2\Sigma_0 + \sigma_t^2\mathrm{Id}$. Thus, in the Gaussian case, diffusion paths and flow-matching paths reduce to the same mean-covariance bookkeeping, although the corresponding training objectives are different.

Example 4.24 (Gaussian kernel drifting). Let the target be $\gamma = \mathcal{N}(\bar{m}, \bar{\Sigma})$ and assume $\alpha_t = \mathcal{N}(m_t, \Sigma_t)$. For the Gaussian kernel

$$K_\varepsilon(x, y) = \exp(-\|x - y\|^2/(2\varepsilon)),$$

the normalized field (4.25) satisfies

$$B_\varepsilon[\alpha_t](x) = -\varepsilon(\Sigma_t + \varepsilon\mathrm{Id})^{-1}(x - m_t).$$

Indeed the smoothed density $x \mapsto \int K_\varepsilon(x, y)\mathrm{d}\alpha_t(y)$ is proportional to the Gaussian density with mean m_t and covariance $\Sigma_t + \varepsilon\mathrm{Id}$. Thus $B_\varepsilon[\alpha_t]$ is the mean-shift vector of a Gaussian density: it points linearly toward the Gaussian mode, with strength set by the bandwidth. The drifting velocity (4.27) is therefore the difference of two affine mean-shift fields; it is affine and preserves Gaussianity. With

$$A_t = (\Sigma_t + \varepsilon\mathrm{Id})^{-1}, \qquad \bar{A} = (\bar{\Sigma} + \varepsilon\mathrm{Id})^{-1},$$

the ODE is

$$\dot{m}_t = \varepsilon\bar{A}(\bar{m} - m_t), \qquad \dot{\Sigma}_t = \varepsilon\big((A_t - \bar{A})\Sigma_t + \Sigma_t(A_t - \bar{A})\big).$$

This finite-dimensional model explains the stabilizing role of the self-normalized repulsion term in drifting: without it, the covariance equation loses the $A_t\Sigma_t + \Sigma_t A_t$ contribution.

Example 4.25 (Gaussian closure of attention dynamics). For the transformer PDE, assume $\alpha = \mathcal{N}(m, \Sigma)$. Since exponential tilting preserves Gaussianity,

$$\frac{\int e^{\langle Qx, Ky\rangle}\, y\, \mathrm{d}\alpha(y)}{\int e^{\langle Qx, Kz\rangle}\, \mathrm{d}\alpha(z)} = m + \Sigma K^\top Q x.$$

Therefore

$$\Gamma_\theta[\alpha](x) = Vm + V\Sigma K^\top Q x$$

is affine. The Gaussian token law is preserved and satisfies

$$\dot{m}_t = (V_t + V_t\Sigma_t K_t^\top Q_t)m_t, \qquad \dot{\Sigma}_t = B_t\Sigma_t + \Sigma_t B_t^\top, \qquad B_t = V_t\Sigma_t K_t^\top Q_t.$$

When $V_t = Q_t^\top K_t$, the matrix $B_t = Q_t^\top K_t\Sigma_t K_t^\top Q_t$ is symmetric positive semidefinite, matching the gradient-field case mentioned above. This closure is not a convergence theorem for trained transformers. It is instead a tractable model of how attention can shear, amplify or contract a cloud of tokens through its covariance.

Contractive Gaussian projection. The preceding examples show when Gaussianity is preserved or imposed by projection. Gelbrich's inequality Gelbrich (1990) gives a useful variational explanation: replacing a measure by the Gaussian with the same first two moments cannot increase its Wasserstein distance to another similarly projected measure.

Theorem 4.26 (Gelbrich theorem)**.** *For $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, let*

$$\mathcal{R}\alpha := \mathcal{N}(m_\alpha, \Sigma_\alpha)$$

be the Gaussian with the same mean and covariance as α. Then

$$\mathcal{W}_2^2(\mathcal{R}\alpha, \mathcal{R}\nu) = \|m_\alpha - m_\nu\|^2 + \mathcal{B}^2(\Sigma_\alpha, \Sigma_\nu) \leqslant \mathcal{W}_2^2(\alpha, \nu).$$

Proof. Take any coupling (X, Y) of α and ν, center the variables, and write $C = \mathbb{E}[(X - m_\alpha)(Y - m_\nu)^\top]$. In the positive definite case, positivity of the block covariance matrix implies the factorization $C = \Sigma_\alpha^{1/2} K \Sigma_\nu^{1/2}$ with $\|K\|_{\mathrm{op}} \leqslant 1$, and therefore, by operator/nuclear norm duality,

$$\operatorname{tr} C \leqslant \operatorname{tr}\left((\Sigma_\alpha^{1/2}\Sigma_\nu\Sigma_\alpha^{1/2})^{1/2}\right).$$

The semidefinite case follows by adding $\eta\mathrm{Id}$ to both covariance matrices and letting $\eta \downarrow 0$. Expanding $\mathbb{E}\|X - Y\|^2$ gives the lower bound

$$\mathbb{E}\|X - Y\|^2 \geqslant \|m_\alpha - m_\nu\|^2 + \mathcal{B}^2(\Sigma_\alpha, \Sigma_\nu).$$

Taking the infimum over couplings proves the inequality, while equality for Gaussian laws is Proposition 1.12. □

The following preservation criterion is a direct consequence of Gelbrich's theorem and was explained to us by Hugo Lavenant. It says that a functional which does not increase under moment-matched Gaussian projection admits Gaussian minimizing movements from Gaussian initial data.

Theorem 4.27 (Hugo Lavenant Gaussian-preservation criterion)**.** *Let $f : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ satisfy*

$$f(\mathcal{R}\alpha) \leqslant f(\alpha) \qquad \forall \alpha \in \mathcal{P}_2(\mathbb{R}^d),$$

with $\mathcal{R}$ defined in Theorem 4.26. If γ is Gaussian and ν minimizes the JKO step

$$\eta \mapsto f(\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \eta),$$

then $\mathcal{R}\nu$ is also a minimizer. If this JKO minimizer is unique, it is Gaussian. Consequently, if every step from Gaussian data has a unique minimizer and the resulting minimizing movements converge in $\mathcal{W}_2$, each discrete iterate is Gaussian and every limit curve is Gaussian as well, possibly with a singular limiting covariance.

Proof. For the JKO claim, $\mathcal{R}\gamma = \gamma$ because γ is Gaussian. Hence, for any competitor η,

$$f(\mathcal{R}\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \mathcal{R}\eta) \leqslant f(\eta) + \frac{1}{2\tau}\mathcal{W}_2^2(\gamma, \eta).$$

Applying this to a minimizer $\eta = \nu$ shows that $\mathcal{R}\nu$ is again a minimizer. Uniqueness forces $\nu = \mathcal{R}\nu$. □

> **Remark 4.28 (Gaussian barycenters from contraction).** The same projection argument also explains why quadratic Wasserstein barycenters of Gaussian measures are Gaussian. If β_s are Gaussian and
>
> $$f_{\mathrm{bar}}(\alpha) = \sum_s \lambda_s \mathcal{W}_2^2(\alpha, \beta_s),$$
>
> then $\mathcal{R}\beta_s = \beta_s$, and Theorem 4.26 gives $f_{\mathrm{bar}}(\mathcal{R}\alpha) \leqslant f_{\mathrm{bar}}(\alpha)$. Thus the moment-matched Gaussian projection of any barycenter is again a barycenter; when the barycenter is unique, it must itself be Gaussian. This is the contraction mechanism behind Gaussian closure of quadratic Wasserstein barycenters.

Conclusion

The central message of this review is that many machine-learning methods can be read as PDEs on probability measures. Optimal transport supplies the geometry behind this statement: it turns a loss into a metric gradient flow, a coupling into a geodesic, a score into a Fokker–Planck drift, and a generative model into a transported law. This language is valuable because it separates modelling choices from representation choices. A finite particle algorithm, a neural parametrization and a continuum PDE may then be understood as different approximations of the same measure-valued evolution.

The same viewpoint also clarifies the limits of current theory. Empirical measures are singular, neural parametrizations are non-convex, high-dimensional transport is statistically expensive, and generative flows often use interpolants that are transportation dynamics without being Wasserstein geodesics or gradient flows. These tensions are not defects of the PDE viewpoint; they identify where mathematical analysis is still needed. For machine learning, PDEs are therefore both an explanatory tool and a source of new problems.

Competing interests

The author declares no competing interests.

Data availability statement

No new data were generated for this review. The manuscript sources and the code used to generate the numerical figures are available at https://github.com/gpeyre/ot4ml.

References

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, volume 2062 of *Lecture Notes in Mathematics*, pages 1–155. Springer, 2013.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2006.

David Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, 2 edition, 2009.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, May 2010.

Hédy Attouch, Juan Peypouquet, and Patrick Redont. Fast convex optimization via inertial dynamics with hessian driven damping, 2016. URL https://arxiv.org/abs/1601.07113.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX, 1983/84*, volume 1123 of *Lecture Notes in Mathematics*, pages 177–206. Springer, 1985. doi: 10.1007/BFb0075847.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014. doi: 10.1007/978-3-319-00227-9.

Raphaël Barboni, Gabriel Peyré, and Franccois-Xavier Vialard. Understanding the training of infinitely deep and wide ResNets with conditional optimal transport. *arXiv preprint arXiv:2403.12887*, 2024. doi: 10.48550/arXiv.2403.12887. URL https://arxiv.org/abs/2403.12887.

Martin Beckmann. A continuous model of transportation. *Econometrica*, 20:643–660, 1952.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the Monge–Ampère operator. *Numerische Mathematik*, 134(3):611–636, 2016.

Vincent D. Blondel, Julien M. Hendrickx, and John N. Tsitsiklis. On Krause's multi-agent consensus model with state-dependent connectivity. *IEEE Transactions on Automatic Control*, 54(11):2586–2597, 2009. doi: 10.1109/TAC.2009.2031211.

Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and Franccois-Xavier Vialard. Token sample complexity of attention. *arXiv preprint arXiv:2512.10656*, 2025.

Jérôme Bolte and Adrien Blanchet. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *arXiv preprint arXiv:1612.02619*, 2016. URL https://arxiv.org/abs/1612.02619.

Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010. doi: 10.1090/S0002-9947-10-05048-X.

Walter Braun and Klaus Hepp. The Vlasov dynamics and its fluctuations in the $1/N$ limit of interacting classical particles. *Communications in Mathematical Physics*, 56(2):101–113, 1977. doi: 10.1007/BF01611497.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional Monge maps. *arXiv preprint arXiv:2206.14262*, 2022. URL https://arxiv.org/abs/2206.14262.

Guillaume Carlier, Chloé Jimenez, and Filippo Santambrogio. Optimal transportation with traffic congestion and Wardrop equilibria. *SIAM Journal on Control and Optimization*, 47(3):1330–1350, 2008.

Guillaume Carlier, Lénaïc Chizat, and Maxime Laborde. Displacement smoothness of entropic optimal transport. *ESAIM: Control, Optimisation and Calculus of Variations*, 30:25, 2024. doi: 10.1051/cocv/2024013.

José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17:233–258, 1 2015.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

Carlo Cercignani, Reinhard Illner, and Mario Pulvirenti. *The Mathematical Theory of Dilute Gases*. Springer, New York, 1994. doi: 10.1007/978-1-4419-8524-8.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025. URL https://jmlr.org/papers/v26/24-0586.html.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ting-Li Chen. On the convergence and consistency of the blurring mean-shift process. *Annals of the Institute of Statistical Mathematics*, 67(1):157–176, 2015. doi: 10.1007/s10463-013-0443-8.

Yanshuo Chen, Zhengmian Hu, Wei Chen, and Heng Huang. Fast and scalable Wasserstein-1 neural optimal transport solver for single-cell perturbation prediction. *arXiv preprint arXiv:2411.00614*, 2024. URL https://arxiv.org/abs/2411.00614.

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995. doi: 10.1109/34.400568.

Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194:487–532, 2022a. doi: 10.1007/s10107-021-01636-z.

Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *Open Journal of Mathematical Optimization*, 3:8, 2022b. doi: 10.5802/ojmo.20.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and Franccois-Xavier Vialard. Unbalanced optimal transport: dynamic and Kantorovich formulation. *Journal of Functional Analysis*, 274(11):3090–3123, 2018a.

Lénaïc Chizat, Bernhard Schmitzer, Gabriel Peyré, and Franccois-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018b.

Shui-Nee Chow, Wen Huang, Yao Li, and Haomin Zhou. Fokker-Planck equations for a free energy functional or Markov process on a graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. doi: 10.1109/34.1000236.

Dario Cordero-Erausquin. Some applications of mass transport to Gaussian-type inequalities. *Archive for Rational Mechanics and Analysis*, 161(3):257–269, 2002. doi: 10.1007/s002050100185.

Dario Cordero-Erausquin and Bo'az Klartag. Moment measures. *Journal of Functional Analysis*, 268(12): 3834–3866, 2015. URL https://arxiv.org/abs/1304.0630.

Dario Cordero-Erausquin, Robert J. McCann, and Michael Schmuckenschläger. A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Inventiones Mathematicae*, 146(2):219–257, 2001.

Giacomo Cozzi and Filippo Santambrogio. Long-time asymptotics of the sliced-Wasserstein flow. *arXiv preprint arXiv:2405.06313*, 2024. doi: 10.48550/arXiv.2405.06313. URL https://arxiv.org/abs/2405.06313.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268, 2020.

Bernard Dacorogna and Jürgen Moser. On a partial differential equation involving the Jacobian determinant. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 7(1):1–26, 1990.

Lorenzo Dello Schiavo, Jan Maas, and Francesco Pedrotti. Local conditions for global convergence of gradient flows and proximal point sequences in metric spaces. *Transactions of the American Mathematical Society*, 377 (6):3779–3804, 2024. doi: 10.1090/tran/9156.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Andrew B. Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023. URL https://www.jmlr.org/papers/v24/20-602.html.

Richard Duong, Nicolaj Rux, Viktor Stein, and Gabriele Steidl. Wasserstein gradient flows of MMD functionals with distance kernels under Sobolev regularization. *arXiv preprint arXiv:2411.09848*, 2024. doi: 10.48550/arXiv.2411.09848. URL https://arxiv.org/abs/2411.09848.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence-Proceedings of the 31st Conference, UAI 2015*, pages 258–267, 2015.

Weinan E, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):10, 2019. doi: 10.1007/s40687-018-0172-y. URL https://arxiv.org/abs/1807.01083.

Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496): 1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Matthias Erbar. The heat equation on manifolds as a gradient flow in the Wasserstein space. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 46(1):1–23, 2010.

Matthias Erbar. Gradient flows of the entropy for jump processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(3):920–945, 2014. doi: 10.1214/12-AIHP537.

Matthias Erbar and Jan Maas. Gradient flow structures for discrete porous medium equations. *Discrete and Continuous Dynamical Systems*, 34(4):1355–1374, 2014.

Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon. Computation of optimal transport on discrete metric measure spaces. *Numerische Mathematik*, 144(1):157–200, 2020. doi: 10.1007/s00211-019-01077-z.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Keinosuke Fukunaga and Larry D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975. doi: 10.1109/TIT.1975.1055330.

Thomas O Gallouët and Leonard Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

Stéphane Gaubert and Zheng Qu. Dobrushin's ergodicity coefficient for Markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015. doi: 10.1007/s00020-014-2193-2.

Matthias Gelbrich. On a formula for the l^2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025. doi: 10.1090/bull/1863.

Ugo Gianazza, Giuseppe Savaré, and Giuseppe Toscani. The Wasserstein gradient flow of the Fisher information and the quantum drift-diffusion equation. *Archive for Rational Mechanics and Analysis*, 194(1):133–220, 2009.

Arthur Gretton, Kevin Wenliang Li, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975. doi: 10.2307/2373688.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850, 2018.

Mert Gürbüzbalaban, Umut cSimcsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017. doi: 10.1088/1361-6420/aa9a90. URL https://arxiv.org/abs/1705.03341.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J. Candès. One-step generative modeling via wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026.

Daniel Hauer and José Mazón. Kurdyka–łojasiewicz–Simon inequality for gradient flows in metric spaces. *Transactions of the American Mathematical Society*, 372(7):4917–4976, 2019. URL https://arxiv.org/abs/1707.03129.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://arxiv.org/abs/1512.03385.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002. URL https://www.jasss.org/5/3/2.html.

Johannes Hertrich, Antonin Chambolle, and Julie Delon. On the relation between rectified flows and optimal transport. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.19712.

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

Bamdad Hosseini, Alexander W. Hsu, and Amirhossein Taghvaei. Conditional optimal transport on function spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):304–338, 2025. doi: 10.1137/23M1618922. URL https://arxiv.org/abs/2311.05672.

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

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/. Blog post.

Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak–łojasiewicz condition, 2016. URL https://arxiv.org/abs/1608.04636.

Gavin Kerrigan, Giosuè Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2968. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/aa1d8cf866c4a684ef2e066dd200f8e4-Abstract-Conference.html.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. *arXiv preprint arXiv:2006.04710*, 2020. URL https://arxiv.org/abs/2006.04710.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. GENOT: Entropic (Gromov) Wasserstein flow matching with applications to single-cell genomics. *arXiv preprint arXiv:2310.09254*, 2023. URL https://arxiv.org/abs/2310.09254.

Debarshi Kundu, Archisman Ghosh, Swaroop Ghosh, and Vasant Honavar. Projection-free transformers via Gaussian kernel attention. *arXiv preprint arXiv:2605.02144*, 2026. URL https://arxiv.org/abs/2605.02144.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48 (3):769–783, 1998. doi: 10.5802/aif.1638.

Hugo Lavenant and Filippo Santambrogio. The flow map of the Fokker–Planck equation does not provide optimal transport. *Applied Mathematics Letters*, 133:108225, 2022.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021. URL https://arxiv.org/abs/2102.09204.

Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

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*, 48(4): 2869–2911, 2016.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training, 2025. URL https://arxiv.org/abs/2502.16982.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/17ed8abedc255908be746d245e50263a-Abstract.html.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1608.04471.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pages 87–89. CNRS, Paris, 1963.

John Lott and Cédric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169(3):903–991, 2009.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3276–3285, 2018. URL https://arxiv.org/abs/1710.10121.

Frederike Lübeck, Charlotte Bunne, Gabriele Gut, Jacobo Sarabia del Castillo, Lucas Pelkmans, and David Alvarez-Melis. Neural unbalanced optimal transport via cycle-consistent semi-couplings. *arXiv preprint arXiv:2209.15621*, 2022. URL https://arxiv.org/abs/2209.15621.

Jan Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8): 2250–2292, 2011.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. *Mathematical Models and Methods in Applied Sciences*, 20(10):1787–1821, 2010.

Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Alexander Mielke. Geodesic convexity of the relative entropy in reversible Markov chains. *Calculus of Variations and Partial Differential Equations*, 48(1-2):1–31, 2013.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM Review*, 56(4):577–621, 2014. doi: 10.1137/120901866.

Ryan Murray, Brian Swenson, and Soummya Kar. Revisiting normalized gradient descent: Fast evasion of saddle points. *IEEE Transactions on Automatic Control*, 64(11):4818–4824, 2019. doi: 10.1109/TAC.2019.2914998.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

Nikolas Nüsken and D. R. Michiel Renger. Stein variational gradient descent: Many-particle and long-time asymptotics. *Foundations of Data Science*, 5(3):286–320, 2023. doi: 10.3934/fods.2022023.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. doi: 10.1006/jfan.1999.3557.

Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

Jan Peszek and David Poyato. Heterogeneous gradient flows in the topology of fibered optimal transport. *Calculus of Variations and Partial Differential Equations*, 62(9):258, 2023. URL https://arxiv.org/abs/2203.08104.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 49069–49104, 2025.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4): 2323–2351, 2015.

Gabriel Peyré. Optimal and diffusion transports in machine learning. *arXiv preprint arXiv:2512.06797*, 2025. Proc. 2026 International Congress of Mathematicians.

Gabriel Peyré. Muon dynamics as a spectral Wasserstein flow. *arXiv preprint arXiv:2604.04891*, 2026.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Boris T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. doi: 10.1016/0041-5553(63)90382-3.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-5553(64)90137-5.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. doi: 10.1016/S0893-6080(98)00116-6.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*. Springer, 1998a.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume II: Applications*. Springer, 1998b.

Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022. doi: 10.1002/cpa.22074.

Grant M. Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5508–5517. PMLR, 2019. URL https://proceedings.mlr.press/v97/rotskoff19a.html.

Stefan G. Samko, Anatoly A. Kilbas, and Oleg I. Marichev. *Fractional Integrals and Derivatives: Theory and Applications*. Gordon and Breach, 1993.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3515–3530. PMLR, 2022.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, 2015.

Filippo Santambrogio. Dealing with moment measures via entropy and optimal transport. *Journal of Functional Analysis*, 271(2):418–436, 2016. doi: 10.1016/j.jfa.2016.04.009.

Filippo Santambrogio. Crowd motion and population dynamics under density constraints. *GMT preprint 3728*, 2018.

Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations, 2018. URL https://arxiv.org/abs/1810.08907.

Umut cSimcsekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5827–5837. PMLR, 2019.

Dejan Slepčev and Andrew Warren. Nonlocal wasserstein distance: Metric and asymptotic properties. *arXiv preprint arXiv:2209.08407*, 2022.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 2015.

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

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

Karl-Theodor Sturm. On the geometry of metric measure spaces. I. *Acta Mathematica*, 196(1):65–131, 2006a.

Karl-Theodor Sturm. On the geometry of metric measure spaces. II. *Acta Mathematica*, 196(1):133–177, 2006b.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016. URL https://jmlr.org/papers/v17/15-084.html.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, 6(3):587–600, 1996. doi: 10.1007/BF02249265.

Alexander Tong, Jessie Huang, Guy Wolf, David van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. *arXiv preprint arXiv:2002.04461*, 2020. URL https://arxiv.org/abs/2002.04461.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Nina Vesseron, Louis Béthune, and Marco Cuturi. Sample and map from a single convex potential: Generation using conjugate moment measures. *arXiv preprint arXiv:2503.10576*, 2025. URL https://arxiv.org/abs/2503.10576.

Cédric Villani. A review of mathematical topics in collisional kinetic theory. In *Handbook of Mathematical Fluid Dynamics*, volume 1, pages 71–305. North-Holland, 2002. doi: 10.1016/S1874-5792(02)80004-0.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics Series*. American Mathematical Society, 2003. ISBN 9780821833124.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

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

Max-K. von Renesse and Karl-Theodor Sturm. Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on Pure and Applied Mathematics*, 58(7):923–940, 2005.

James Vuckovic, Aristide Baratin, and Rémi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

Andrew Warren. Gradient flow structure for some nonlocal diffusion equations. *arXiv preprint arXiv:2412.20969*, 2024.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016. doi: 10.1073/pnas.1614734113.

Diyuan Wu, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis for heavy ball methods: Dropout-stability, connectivity, and global convergence, 2022. URL https://arxiv.org/abs/2210.06819.