Muon Dynamics as a Spectral Wasserstein Flow

Gabriel Peyré
CNRS and ENS, PSL Université
gabriel.peyre@ens.fr

Abstract

Gradient normalization can stabilize deep-learning optimization, and spectral normalizations are especially natural for matrix-shaped parameter blocks; Muon is the motivating example. We study an idealized deterministic, continuous-time, vanishing-momentum version of this idea in the mean-field regime, where wide models are represented by probability measures on parameter space. Starting from normalized matrix flows, we introduce Spectral Wasserstein distances indexed by norms γ on positive semidefinite matrices: the trace norm recovers the classical W_2 distance, the operator norm gives the Muon geometry, and Schatten norms interpolate between them. We develop a static Kantorovich formulation, a max–min robust-cost representation, and Gaussian reductions extending the Bures formula. For monotone norms, we prove equivalence with a Benamou–Brenier formulation and obtain a gradient-flow interpretation of normalized mean-field training dynamics. Numerical experiments illustrate the resulting flows for MMD objectives, Gaussian reductions, two-layer ReLU models, and shallow attention.

1 Introduction

Spectrally normalized optimizers such as Muon replace the Euclidean geometry of gradient descent by a matrix-aware geometry adapted to block-shaped parameters. This paper develops a mathematical framework for such optimizers in very wide, possibly infinite-width, models. The central object is a Spectral Wasserstein geometry in which their idealized dynamics become mean-field gradient flows.

Normalized gradient descent and mean field dynamics. On the optimization side, a growing literature studies normalized first-order methods. The recent framework of Pethick et al. [2025] is particularly relevant because it treats norm-constrained linear minimization oracles as a general language for normalized gradient methods and includes spectral normalizations as special cases. Earlier works such as Cutkosky and Mehta [2020] and Murray et al. [2019] show how normalized gradient methods change the optimization dynamics even in nonconvex settings. For deep architectures, matrix-aware normalizations are especially natural: Shampoo is an early influential example of tensor/matrix-aware preconditioning [Gupta et al., 2018], and Muon has become a leading example of spectral normalization in large-scale training [Jordan et al., 2024, Liu et al., 2025]. This connects naturally with the mean-field description of wide neural networks through probability measures on parameter space, which underlies the landscape analysis of two-layer networks by Mei et al. [2018], the optimal-transport convergence analysis of over-parameterized models by Chizat and Bach [2018], and the metric-gradient-flow viewpoint developed in Ambrosio et al. [2008]. Our work keeps this mean-field perspective but changes the underlying metric from Euclidean Wasserstein geometry to matrix-aware Spectral Wasserstein geometries.

Generalized optimal distances. On the transport side, our work is closest in spirit to generalized coupling costs and weak transport, as developed for instance by Gozlan et al. [2017], Backhoff-Veraguas et al. [2019], and Backhoff-Veraguas and Pammer [2022]. It is also related to covariance-dependent transport costs [Burger et al., 2025], to matrix-valued optimal transport [Ning et al., 2015,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Chen et al., 2018], and to the Bures–Wasserstein geometry of covariance matrices [Bhatia et al., 2019]. Another useful perspective is to robustify the Wasserstein distance by optimizing over the cost itself. The closest antecedent is the subspace robust Wasserstein distance of Paty and Cuturi [2019], which corresponds to the special case of our spectral Wasserstein distance W_γ obtained when γ is a Ky Fan norm [Fan, 1951]. Maximizing over costs is also connected to metric learning in Wasserstein discriminant analysis [Flamary et al., 2018], ground metric learning [Cuturi and Avis, 2014], and congestion models for transportation networks [Carlier et al., 2008]. The opposite direction, minimizing over costs, appears, for instance, in Sebbouh et al. [2024]; this leads to a concave minimization problem useful for Gromov–Wasserstein-type structure. By contrast, our robustification is a concave maximization over admissible quadratic costs, leading to a convex static problem.

Contributions. The construction has three steps: a matrix normalization defines a descent direction; its force-covariance formula extends that direction to arbitrary measures; and a transport distance supplies the corresponding geometry. We first derive spectrally normalized matrix flows and their mean-field continuity equations in Section 2. We then introduce the static Spectral Wasserstein cost in Section 3, compare it with W_2 in Proposition 3.2, establish its max–min robust-cost representation in Theorem 3.4, and derive its Gaussian reduction in Proposition 3.6. For monotone norms, Section 4 proves the equivalence between the static and Benamou–Brenier formulations in Theorem 4.2, yielding the metric property in Corollary 4.3 and explicit geodesics in Corollary 4.4. Section 5 identifies the normalized PDE as the corresponding metric gradient flow and discusses MLP, shallow-attention, MMD, and linear-network examples. A Python computational toolbox, illustrative notebooks, and code to reproduce the numerical experiments are available at `https://github.com/gpeyre/muon-dynamics`.

Scope. This paper does not attempt to explain why Muon-type algorithms outperform vanilla gradient descent for particular architectures, especially attention layers. Its goal is to provide the mathematical formalism needed to investigate such questions in the regime of very wide networks. The resulting transport geometries also raise theoretical questions that extend beyond neural-network training. The finite-particle identity is exact, while the gradient-flow characterization is conditional on stated regularity assumptions. We do not prove convergence of particle systems to the limiting PDE, global PDE well-posedness, or convergence of practical stochastic, momentum-based Muon implementations.

2 Spectral Gradient Flow and Mean Field Dynamics

Spectral gradient flows on matrices. This section introduces the normalized dynamics studied in the paper, whose geometry is governed by a matrix gauge γ. The "Muon-type" descent corresponds to the case where γ is the operator norm. We impose monotonicity throughout the main text. It is used in the static–dynamic equivalence of Theorem 4.2 and is crucial for the Spectral Wasserstein gradient-flow interpretation in Section 5. Appendix C explains what can fail without monotonicity, while Appendix E.4 gives a different extension beyond convex gauges.

Definition 2.1 (Monotone gauge γ and norm $\mathrm{N}(V)$)**.** *Unless explicitly stated otherwise, γ denotes the restriction to $\mathbb{S}^d_+$ of a norm on $\mathbb{S}^d$. We assume that this restriction is monotone: whenever $0 \preceq S \preceq S'$, one has $\gamma(S) \le \gamma(S')$. The matrix norm induced by γ is $\mathrm{N}(V) \coloneqq \gamma(V^\top V)^{1/2}$; for $\gamma = \gamma_p$, one has $\mathrm{N}(V) = \|V\|_{S_{2p}}$. As shown in Proposition A.3, $\mathrm{N}(V)$ is indeed a norm when γ is monotone.*

The canonical examples used throughout this article are the Schatten gauges: if $\lambda(S)$ denotes the vector of eigenvalues of $S \succeq 0$, then $\gamma_p(S) \coloneqq \|S\|_{S_p} = \|\lambda(S)\|_{\ell_p}$, $p \in [1, \infty]$. Other examples include the Ky Fan norms [Fan, 1951], $\gamma(S) \coloneqq \sum_{i=1}^k \lambda_i(S)$, obtained by summing the k largest eigenvalues. They interpolate between the operator norm ($k = 1$) and the trace norm ($k = d$) and recover the subspace-robust Wasserstein setting of Paty and Cuturi [2019]. Although the main text assumes that γ is convex, many metric and Benamou–Brenier results extend to the larger class of possibly nonconvex root-norm gauges introduced in Appendix E.4.

Definition 2.2 (Linear Minimization Oracle (LMO))**.** *For $G \in \mathbb{R}^{n\times d}$, the matrix LMO associated with γ is any minimizer $\mathrm{J}_\gamma(G) \in \operatorname{argmin}_V \{\langle G, V\rangle_F + \frac{1}{2}\mathrm{N}(V)^2\}$.*

The name is a slight abuse of the usual terminology: a classical LMO minimizes a linear functional on a fixed ball. Here it is followed by an optimal scalar step. Indeed, if N^* is the dual norm and $G \neq 0$, then

$$U_G \in \operatorname*{argmin}_{\mathrm{N}(U)\leq 1} \langle G, U\rangle_F \quad \Longrightarrow \quad \mathrm{J}_\gamma(G) = \mathrm{N}^*(G)U_G. \tag{1}$$

This follows by minimizing $-r\mathrm{N}^*(G) + r^2/2$ over $r \geq 0$; $\mathrm{J}_\gamma(0) = 0$. We use the same convention for the Measure-LMO below.

Let $\mathrm{F} : \mathbb{R}^{n\times d} \to \mathbb{R}$ be a smooth objective. Following Pethick et al. [2025], it yields both the discrete step and the ODE flow

$$X_{k+1} = X_k + \tau \mathrm{J}_\gamma(\nabla \mathrm{F}(X_k)) \quad \text{and} \quad \dot{X}_t = \mathrm{J}_\gamma(\nabla \mathrm{F}(X_t)). \tag{2}$$

For $p = 1$, $\mathrm{J}_{\gamma_1}(G) = -G$, so one recovers the usual gradient flow; for $p = \infty$, $\mathrm{J}_{\gamma_\infty}$ is the trace-scaled partial polar factor of G, giving a simplified "Muon" ODE. Explicit Schatten formulas are given in Appendix A. The scalar factor matters: the unscaled polar update and this squared-norm flow follow the same deterministic paths, away from stationary points, but with different time parametrizations. Our metric fixes the latter clock.

Mean-field lift and normalized continuity equation. Given particles $x_1, \ldots, x_n \in \mathbb{R}^d$, write $X \coloneqq [x_1^\top; \ldots; x_n^\top] \in \mathbb{R}^{n\times d}$ and $\mu_X \coloneqq n^{-1}\sum_i \delta_{x_i}$. We consider permutation-invariant objectives lifted from measures, $\mathrm{F}(X) \coloneqq n\mathcal{F}(\mu_X)$. This scaling applies, in particular, to the two-layer MLP and attention models discussed in Section 5.

Definition 2.3 (Measure-LMO)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. For $v \in L^2(\mu;\mathbb{R}^d)$, define $\mathcal{N}_\mu(v)^2 \coloneqq \gamma\big(\int v(x)v(x)^\top d\mu(x)\big)$. For $g \in L^2(\mu;\mathbb{R}^d)$, the measure-LMO is any member of $\mathcal{J}_\gamma^\mu(g) \in \operatorname{argmin}_{v\in L^2(\mu;\mathbb{R}^d)} \big\{\int g(x)\cdot v(x)\, d\mu(x) + \frac{1}{2}\mathcal{N}_\mu(v)^2\big\}$.*

For an empirical measure $\mu_X = n^{-1}\sum_i \delta_{x_i}$ transported by particle velocities stacked into a matrix V, one has $\mathcal{N}_{\mu_X}(v)^2 = n^{-1}\mathrm{N}(V)^2$. If G denotes the matrix with rows $g(x_i)^\top$, then the rows of $\mathcal{J}_\gamma^{\mu_X}(g)$ coincide with those of $\mathrm{J}_\gamma(G)$. For a more general measure μ, however, it is not a priori clear how to compute $\mathcal{J}_\gamma^\mu(g)$. The following proposition reduces this infinite-dimensional LMO to a finite-dimensional matrix problem.

Proposition 2.4 (Finite-dimensional Measure-LMO)**.** *Let $S_g \coloneqq \mathcal{S}_\mu(g) = \int g(x)g(x)^\top d\mu(x)$ and let P_g be the orthogonal projector onto* $\operatorname{range}(S_g)$*. If $S_g = 0$, then $\mathcal{J}_\gamma^\mu(g) = 0$. Otherwise, a valid Measure-LMO is given by*

$$A_g \in \operatorname*{argmin}_{A=P_gA} \left\{ \operatorname{tr}(S_g^{1/2}A) + \frac{1}{2}\gamma(A^\top A)\right\}, \qquad \mathcal{J}_\gamma^\mu(g)(x) = A_g^\top S_g^{\dagger/2} g(x).$$

Proof. Associate to each field $h \in L^2(\mu;\mathbb{R}^d)$ the operator $T_h : \mathbb{R}^d \to L^2(\mu)$, $(T_h a)(x) = a\cdot h(x)$. Then $T_h^* T_h = \int hh^\top d\mu$ and $\int g\cdot h\, d\mu = \langle T_g, T_h\rangle_{\mathrm{HS}}$. Writing the polar decomposition $T_g = U_g S_g^{1/2}$, set $A = U_g^* T_h$ and $B = (I - U_g U_g^*)T_h$. Then $P_g A = A$, $T_h^*T_h = A^\top A + B^*B$, and $\langle T_g, T_h\rangle_{\mathrm{HS}} = \operatorname{tr}(S_g^{1/2}A)$. Monotonicity of γ shows that deleting B cannot increase the objective. Conversely, every $A = P_gA$ is realized by $T_h = U_gA$, or $h(x) = A^\top S_g^{\dagger/2} g(x)$. The finite-dimensional objective is continuous and coercive, so a minimizer exists. If $S_g = 0$, then $g = 0$ μ-almost everywhere and the unique optimal field is zero. □

For Schatten gauges, this finite-dimensional LMO gives the explicit formula in Corollary A.6; in particular, it covers the Muon endpoint. For force and displacement fields, we use "covariance" to mean the uncentered second-moment matrix. Covariances of Gaussian laws retain their usual centered meaning.

Following Ambrosio et al. [2008], define the Wasserstein force and the first variation by

$$g_\mu(x) \coloneqq \nabla_x \frac{\delta \mathcal{F}}{\delta \mu}(x), \qquad \frac{d}{d\varepsilon}\mathcal{F}(\mu + \varepsilon\sigma)\Big|_{\varepsilon=0} = \int \frac{\delta\mathcal{F}}{\delta\mu}\, d\sigma$$

for signed perturbations σ of zero total mass for which the perturbed measures are admissible, with one-sided derivatives when necessary. The force g_μ is a cotangent quantity, not the spectral velocity: the latter is $\mathcal{J}_\gamma^\mu(g_\mu)$. The normalized mean-field dynamics is then the PDE

$$\partial_t \mu_t + \operatorname{div}\big(\mu_t \mathcal{J}_\gamma^{\mu_t}(g_{\mu_t})\big) = 0. \tag{3}$$

When $\gamma(S) = \operatorname{tr}(S)$, $\mathcal{J}_\gamma^\mu(g) = -g$, and this reduces to the usual formal W_2 gradient-flow PDE. The particle interpretation requires the following transport chain rule, not just a first variation along mixtures.

Proposition 2.5 (Finite-width interpretation)**.** *Assume $\mathrm{F}(X) = n\mathcal{F}(\mu_X)$ is C^1 and, at every empirical state μ, $\frac{d}{dh}\mathcal{F}((\mathrm{Id} + hw)_\#\mu)\big|_{h=0} = \int g_\mu \cdot w \, d\mu$ for every field w on its support. Let X_t be an absolutely continuous particle curve and set $\mu_t = \mu_{X_t}$, using matching matrix and measure LMO selections. As long as the atoms remain pairwise distinct, X_t solves ODE* (2) *if and only if μ_{X_t} solves* (3) *weakly. The same statement holds through collisions when coincident atoms are assigned the same selected velocity. The proof is in Appendix A.*

Extending the global-in-time PDE well-posedness theory of Chizat and Bach [2018], which covers the trace case $p = 1$ under smoothness and growth assumptions on the feature map, to general Schatten geometries remains a difficult open problem. The Measure-LMO nevertheless defines the same evolution rule for empirical and continuous distributions, without a prescribed number of particles. For suitable continuous families it can be substantially simpler: affine velocities preserve Gaussian distributions and close the dynamics on their moments (Appendix F). The major question left open by the construction above is whether (3) is the gradient flow of an actual distance on probability measures. The next two sections define this distance and prove a Benamou–Brenier formula with velocity action $\mathcal{N}_\mu^2$. Different velocities may generate the same continuity equation, so metric speed need not equal the norm of an arbitrary representative. Theorem 5.1 establishes equality for the selected descent velocity under its stated hypotheses.

Remark 2.6 (Connection with Muon)**.** *Appendix B relates the normalized-flow viewpoint to practical ingredients of Muon. Finitely many Newton–Schulz iterations replace the exact LMO by a polynomial spectral map generated by an explicit potential. Momentum can be modeled by an exponentially averaged force variable, leading in the mean-field limit to a phase-space Liouville equation; this is no longer a gradient flow on the position marginal alone, and thus lies partly outside the Spectral Wasserstein geometry developed below. Blockwise normalization is covered by block gauges, with different gauges for independently scaled LMOs and for raw polar updates. The same appendix distinguishes these deterministic extensions from noisy minibatch updates, for which normalization and expectation do not commute.*

3 Static Spectral Wasserstein Geometry

Kantorovich cost. This section introduces the Spectral Wasserstein transport cost. The key point is that the correct object is coupling-based and depends on a norm γ acting on the global displacement covariance.

Definition 3.1 (Static Spectral Wasserstein)**.** *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\Pi(\mu, \nu)$ be the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν, and define*

$$\mathsf{W}_\gamma(\mu, \nu)^2 \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \gamma\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} (y - x)(y - x)^\top \, d\pi(x, y) \right).$$

The Monge restriction, the possible strictness of the coupling relaxation, and the obstruction to the triangle inequality for non-monotone norms are discussed in Appendix C. In the monotone case, the full metric property is proved later through the dynamic formulation in Corollary 4.3.

Comparison with W_2 and cost-robust formulation. We now show that W_γ has the same topology as W_2. The proof, explicit Schatten constants, and sharpness examples are given in Appendix A.2.

Proposition 3.2 (Norm comparison)**.** *Let $0 < c_\gamma \le C_\gamma < \infty$ satisfy $c_\gamma \operatorname{tr}(S) \le \gamma(S) \le C_\gamma \operatorname{tr}(S)$ for every $S \succeq 0$; such constants always exist. Then $\sqrt{c_\gamma}\, \mathsf{W}_2(\mu, \nu) \le \mathsf{W}_\gamma(\mu, \nu) \le \sqrt{C_\gamma}\, \mathsf{W}_2(\mu, \nu)$.*

The comparison bounds of Proposition 3.2 already imply separation and topological equivalence with W_2. What is not immediate from the static formulation alone is the triangle inequality. To address this, and to clarify the geometry of W_γ, Theorem 3.4 below recasts W_γ as a cost-robust Wasserstein distance. This reformulation has two immediate consequences: a conditional Brenier theorem in the Monge setting (Appendix C) and the fact that W_γ is a distance (Appendix A.3). The latter can also be obtained from the Benamou–Brenier formulation; see Corollary 4.3.

Definition 3.3 (Representing set $\mathcal{K}_\gamma$)**.** *A representing set for γ is a convex compact set $\mathcal{K}_\gamma \subset \mathbb{S}^d_+$ such that, for every $S \succeq 0$,*

$$\gamma(S) = \max_{Q \in \mathcal{K}_\gamma} \operatorname{tr}(QS).$$

As recalled in Appendix A, monotonicity of γ on $\mathbb{S}^d_+$ allows the representing set to be chosen inside $\mathbb{S}^d_+$. Remark A.4 gives examples of such sets $\mathcal{K}_\gamma$ for the Schatten norms $\gamma = \gamma_p$.

Theorem 3.4 (Max-min and cost-robust representation)**.** *For every positive semidefinite matrix $Q \in \mathbb{S}^d_+$, let W^Q denote the quadratic transport functional associated with the cost $\langle Q(x-y), x-y\rangle$, namely*

$$\mathsf{W}^Q(\mu,\nu)^2 \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} (y-x)^\top Q(y-x)\, d\pi(x,y).$$

In particular, $\mathsf{W}^{\mathrm{Id}} = \mathsf{W}_2$. Then, for any representing set $\mathcal{K}_\gamma$ for γ as in Definition 3.3, for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \max_{Q \in \mathcal{K}_\gamma} \inf_{\pi \in \Pi(\mu,\nu)} \int (y-x)^\top Q(y-x)\, d\pi(x,y) = \max_{Q \in \mathcal{K}_\gamma} \mathsf{W}^Q(\mu,\nu)^2.$$

Denoting $\Sigma_\pi \coloneqq \int (y-x)(y-x)^\top \, d\pi(x,y)$, any saddle point $(Q^\star, \pi^\star)$ satisfies $Q^\star \in \partial\gamma(\Sigma_{\pi^\star})$, where the subdifferential is understood on the cone $\mathbb{S}^d_+$.

Proof sketch. Details are given in Appendix A.2. The support representation of γ gives $\gamma(\Sigma_\pi) = \max_{Q \in \mathcal{K}_\gamma} \operatorname{tr}(Q\Sigma_\pi)$, with $\operatorname{tr}(Q\Sigma_\pi) = \int (y-x)^\top Q(y-x)\, d\pi(x,y)$. Hence $\mathsf{W}_\gamma(\mu,\nu)^2 = \inf_{\pi \in \Pi(\mu,\nu)} \max_{Q \in \mathcal{K}_\gamma} F(\pi, Q)$, where $F(\pi,Q) \coloneqq \int (y-x)^\top Q(y-x)\, d\pi(x,y)$. The conclusion follows by exchanging the infimum and the maximum. Indeed, $\Pi(\mu,\nu)$ is convex, $\mathcal{K}_\gamma$ is compact and convex, and F is affine in both variables, continuous in Q, and lower semicontinuous in π. Sion's minimax theorem therefore gives $\inf_\pi \max_Q F(\pi,Q) = \max_Q \inf_\pi F(\pi,Q)$. The inner infimum is precisely $\mathsf{W}^Q(\mu,\nu)^2$, yielding $\mathsf{W}_\gamma(\mu,\nu)^2 = \max_{Q \in \mathcal{K}_\gamma} \mathsf{W}^Q(\mu,\nu)^2$. □

Corollary 3.5 (Kantorovich potentials)**.** *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \max_{Q \in \mathcal{K}_\gamma} \sup_{\varphi(x)+\psi(y) \le (x-y)^\top Q(x-y)} \left(\int \varphi\, d\mu + \int \psi\, d\nu \right),$$

where the supremum ranges over integrable admissible potentials.

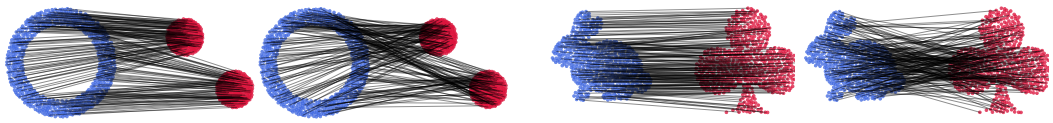
The proof is in Appendix A.

Discrete setting. This subsection spells out the finite-dimensional version used in computations and shows how the abstract coupling problem becomes a concrete convex program. We consider two discrete measures $\mu \coloneqq \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu \coloneqq \sum_{j=1}^m b_j \delta_{y_j}$, with weights $a_i, b_j \ge 0$ summing to one. As in classical Kantorovich transport, a coupling is represented by a transport matrix $P \in \mathbb{R}_+^{n \times m}$ satisfying $P\mathbf{1}_m = a$ and $P^\top \mathbf{1}_n = b$. The discrete static Spectral Wasserstein problem can then be written as the convex optimization problem

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \min_{P \ge 0,\ P\mathbf{1}_m = a,\ P^\top \mathbf{1}_n = b} \gamma\Big(\sum_{i=1}^n \sum_{j=1}^m P_{ij} (y_j - x_i)(y_j - x_i)^\top \Big).$$

For $\gamma = \gamma_p$ and $p = 1$, this reduces to the usual linear optimal-transport problem with quadratic cost, hence to a classical assignment problem in the equal-weight case. For $p = \infty$, it admits a semidefinite programming formulation through an epigraph constraint on the largest eigenvalue of the displacement covariance. Figure 1 illustrates couplings obtained for the trace and operator

objectives. The trace assignments are exact; the operator panels use a rank-one alternating heuristic, not a certified SDP optimizer (Appendix G). Proposition A.10 explains this effect in the regime where the two measures have well-separated means: while the trace case remains isotropic, every Schatten geometry with $p > 1$ asymptotically selects a quadratic cost aligned with the mean displacement direction, making the matching primarily sensitive to displacements parallel to the vector joining the two means. As explained in Appendix C, for $p = \infty$ an active matrix $Q^\star$ is often rank one, so the conditional Brenier argument neither ensures uniqueness nor forces every optimizer to be induced by a map. Figure 1 illustrates the resulting freedom in directions orthogonal to the active one; see also Remark 5.2. Additional image-defined shape pairs are shown in Appendix G.



(a) $p = 1$ (b) $p = \infty$ (c) $p = 1$ (d) $p = \infty$

Figure 1: Static coupling illustrations for two source-target pairs: exact trace assignments ($p = 1$) and heuristic operator couplings ($p = \infty$). Only a subsample of matched segments is displayed.

Gaussian setting and Spectral Bures distance. Gaussian marginals compress the transport problem to a finite-dimensional optimization over admissible covariance blocks. The proof of Proposition 3.6 and the closed commuting formula for Schatten norms are collected in Appendix D.

Proposition 3.6 (Gaussian reduction)**.** *Let $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$. Then*

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \inf_{K:\left(\begin{smallmatrix}\Sigma_0 & K\\ K^\top & \Sigma_1\end{smallmatrix}\right)\succeq 0} \gamma\big((m_1-m_0)(m_1-m_0)^\top + \Sigma_0 + \Sigma_1 - K - K^\top\big)\,.$$

4 Dynamic Formulation and Geodesics

We now turn to the Benamou–Brenier formulation [Benamou and Brenier, 2000]. Throughout this section, γ is monotone. Appendix A then allows us to choose $\mathcal{K}_\gamma \subset \mathbb{S}^d_+$; this positive-semidefinite representation is the key ingredient in the static–dynamic equivalence below.

Definition 4.1 (Dynamic Spectral Wasserstein)**.** *For $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, define*

$$\mathsf{W}^{\mathrm{BB}}_\gamma(\mu_0,\mu_1)^2 := \inf_{(\mu_t, v_t)} \int_0^1 \gamma\left(\int v_t(x) v_t(x)^\top \, d\mu_t(x)\right) dt, \tag{4}$$

where the infimum is over narrowly continuous curves $t \mapsto \mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ and Borel velocity fields satisfying $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$ and the endpoint conditions $\mu_{t=0} = \mu_0$, $\mu_{t=1} = \mu_1$.

The equivalent convex momentum formulation is given in Appendix E. The following theorem is the main result of the paper. Its proof, together with proofs of the subsequent corollaries, is given in Appendix E.

Theorem 4.2 (Static-dynamic equivalence)**.** *If γ is monotone, then $\mathsf{W}^{\mathrm{BB}}_\gamma(\mu_0,\mu_1) = \mathsf{W}_\gamma(\mu_0,\mu_1)$.*

Proof sketch. The detailed argument appears in Appendix E. For the dynamic-to-static inequality, let (μ_t, v_t) solve the continuity equation. By the superposition principle, there is a probability measure η on absolutely continuous paths ω such that $(e_t)_\#\eta = \mu_t$ and $\dot\omega_t = v_t(\omega_t)$ for a.e. t, where $e_t(\omega) = \omega_t$. The endpoint coupling $\pi = (e_0, e_1)_\#\eta$ then satisfies $\omega_1 - \omega_0 = \int_0^1 v_t(\omega_t)\, dt$ for η-a.e. ω. By monotonicity and Appendix A, the dual support set $\mathcal{K}_\gamma$ may be chosen in $\mathbb{S}^d_+$. Hence, for every $Q \in \mathcal{K}_\gamma$, Jensen's inequality gives $\langle Q, (\omega_1-\omega_0)(\omega_1-\omega_0)^\top\rangle \le \int_0^1 \langle Q, v_t(\omega_t) v_t(\omega_t)^\top\rangle\, dt$. Integrating with respect to η and taking the supremum over $Q \in \mathcal{K}_\gamma$ yields the static cost of π bounded by the dynamic action. Taking the infimum over admissible (μ_t, v_t) gives one inequality. Conversely, let π be an optimal static coupling, set $z_t(x,y) = (1-t)x + ty$, and define $\mu_t = (z_t)_\#\pi$. An Eulerian velocity field for this curve is $v_t(z) = \mathbb{E}_\pi[y - x \mid z_t = z]$. Conditional Jensen's inequality yields $\int v_t v_t^\top\, d\mu_t \preceq \int (y-x)(y-x)^\top\, d\pi$ for a.e. t. Monotonicity of γ therefore bounds the dynamic action by the static cost of π, proving the reverse inequality. □

Corollary 4.3 (Metric properties). *W_γ is a distance on $\mathcal{P}_2(\mathbb{R}^d)$ inducing the same topology as W_2.*

Corollary 4.4 (Geodesics). *If π_* is any optimal coupling for $\mathsf{W}_\gamma(\mu_0, \mu_1)$, then $\mu_t = ((1-t)x + ty)_\# \pi_*$ is a constant-speed geodesic and $\mathsf{W}_\gamma(\mu_s, \mu_t) = |t-s|\,\mathsf{W}_\gamma(\mu_0, \mu_1)$ for every $0 \le s \le t \le 1$.*

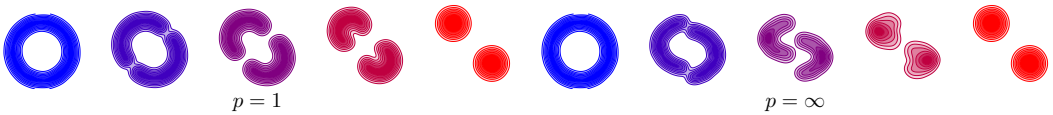


Figure 2: Displacement-interpolation density snapshots induced by the computed static couplings (exact for $p = 1$, heuristic for $p = \infty$), for $p = 1$ (first five panels) and for $p = \infty$ (last five panels). We use $n = 1500$ particles. As in Figure 1, the first and last snapshots correspond to the same source and target distributions ($t = 0$ and $t = 1$).

Figure 2 complements Figure 1: once a coupling is fixed, its displacement interpolation is linear in time at the particle level, but the macroscopic density evolution differs markedly between $p = 1$ and $p = \infty$. Appendix G shows additional examples. For Schatten gauges these effects do not privilege the coordinate axes: for every orthogonal matrix R, simultaneous rotation of the two measures conjugates each displacement covariance by R, leaving W_γ unchanged. The exact transport problem is rotation invariant; the displayed heuristic couplings may additionally reflect the solver initialization.

5 Spectral Wasserstein Gradient Flows

This section is written directly for arbitrary measures. Proposition 2.5 already identifies its empirical version with the normalized matrix ODE, hence with the idealized Muon dynamics at $p = \infty$. The different role of Theorem 5.1 below is to identify this evolution as a curve of maximal slope for a transport metric, including when the distribution is not empirical.

Metric interpretation of the normalized PDE. We now show that (3) is the gradient flow of $\mathcal{F}$ for the metric W_γ, in the metric-gradient-flow sense of Ambrosio et al. [2008]. Recall that the metric derivative of an absolutely continuous curve $t \mapsto \mu_t$ is $|\mu'|_\gamma(t) := \lim_{h\to 0} \mathsf{W}_\gamma(\mu_{t+h}, \mu_t)/|h|$, whenever the limit exists, and the local slope of $\mathcal{F}$ is $|\partial\mathcal{F}|_\gamma(\mu) := \limsup_{\nu\to\mu}(\mathcal{F}(\mu) - \mathcal{F}(\nu))_+ / \mathsf{W}_\gamma(\mu, \nu)$. We use the energy-identity formulation of a curve of maximal slope: the curve and its energy are locally absolutely continuous, and, for every $s \le t$,

$$\mathcal{F}(\mu_s) - \mathcal{F}(\mu_t) = \frac{1}{2}\int_s^t \left(|\mu'|_\gamma(r)^2 + |\partial\mathcal{F}|_\gamma(\mu_r)^2\right) dr.$$

The velocity action is minimized over representatives of the same measure derivative, formally $\inf\{\mathcal{N}_\mu(v) : -\operatorname{div}(\mu v) = \dot\mu\}$. We will use the speed bound supplied by BB and prove equality for the LMO flow, without assuming equality for every velocity field.

Theorem 5.1 (W_γ gradient flow). *Fix $T < \infty$ and $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Let $v_t(x)$ be continuous in (t, x), uniformly globally Lipschitz in x, with uniform linear growth on $[0, T]$. Let X_t be its characteristic flow and $\mu_t = (X_t)_\sharp \mu_0$. Assume that $\mathcal{F}$ is finite on a neighborhood of this curve and admits forces $g_t = g_{\mu_t} \in L^2(\mu_t; \mathbb{R}^d)$ such that, for constants $C, \delta > 0$, every $t \in [0, T]$, and every $\pi \in \Pi(\mu_t, \nu)$ with $\int |y - x|^2 d\pi \le \delta$,*

$$\left|\mathcal{F}(\nu) - \mathcal{F}(\mu_t) - \int g_t(x)\cdot(y - x)\, d\pi(x, y)\right| \le C \int |y - x|^2\, d\pi(x, y). \tag{5}$$

If $v_t \in \mathcal{J}_\gamma^{\mu_t}(g_t)$, then μ_t is a curve of maximal slope for $\mathcal{F}$ with respect to W_γ. More precisely, almost everywhere,

$$|\mu'|_\gamma(t) = |\partial\mathcal{F}|_\gamma(\mu_t) = \mathcal{N}^*_{\mu_t}(g_t) = \mathcal{N}_{\mu_t}(v_t), \qquad \frac{d}{dt}\mathcal{F}(\mu_t) = -\mathcal{N}^*_{\mu_t}(g_t)^2,$$

where $\mathcal{N}^*_\mu$ is the dual norm under the $L^2(\mu)$ pairing.

Proof sketch. The detailed proof is in Appendix E.5. The transport expansion (5) and barycentric projection of an optimal coupling give $|\partial\mathcal{F}|_\gamma(\mu_t) \le \mathcal{N}^*_{\mu_t}(g_t)$. Perturbing μ_t by $(\mathrm{Id} + hw)_\sharp \mu_t$, for

arbitrary $w \in L^2(\mu_t)$, proves the opposite inequality by duality. The characteristic flow gives the chain rule $\frac{d}{dt}\mathcal{F}(\mu_t) = \int g_t \cdot v_t \, d\mu_t = -\mathcal{N}^*_{\mu_t}(g_t)^2$. The dynamic formulation bounds the metric speed by $\mathcal{N}_{\mu_t}(v_t) = \mathcal{N}^*_{\mu_t}(g_t)$; the slope inequality and the chain rule give the reverse bound. These equalities imply the energy dissipation identity. □

These are verification assumptions on an existing curve, not an existence or uniqueness theorem. They require neither a density nor invertibility of the force covariance. A concrete smooth class satisfying (5) is $\mathcal{F}(\mu) = \Phi(\int \varphi_1 d\mu, \dots, \int \varphi_m d\mu)$, with C^2 features of bounded Hessian and $\Phi \in C^2$ on a neighborhood of the relevant moments; see Appendix E.5. Regularity of the selected velocity must still be checked. In particular, the theorem does not automatically cover ReLU kinks or changes of rank at the operator endpoint.

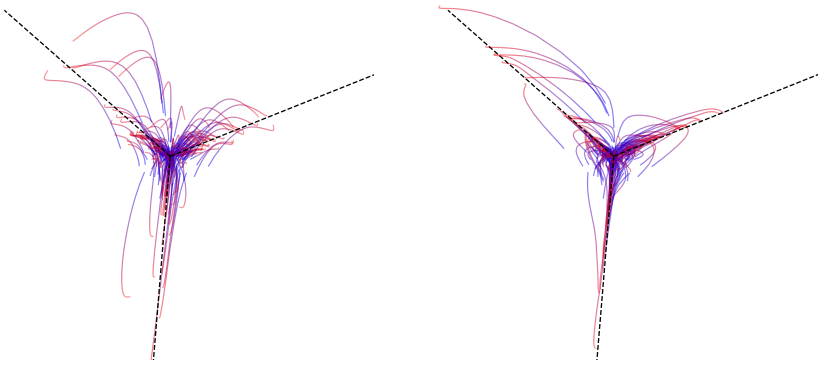
Remark 5.2 (Static and infinitesimal views of Muon)**.** *For $\gamma_\infty(S) = \lambda_{\max}(S)$, the optimal cost certificate $Q^\star \in \partial\gamma_\infty(\Sigma_{\pi^\star})$ of Theorem 3.4 is supported on the top eigenspace of the displacement covariance. If this eigenspace is spanned by a unit vector $u^\star$, then $Q^\star = u^\star(u^\star)^\top$ and $\pi^\star$ minimizes the projected cost $\langle y - x, u^\star\rangle^2$. Infinitesimally, write $S_g = \int gg^\top d\mu$ and let P_g project onto its range. Corollary A.6 gives the selector and its velocity covariance:*

$$v = \mathcal{J}^\mu_{\gamma_\infty}(g) = -\operatorname{tr}(S_g^{1/2})S_g^{\dagger/2}g, \qquad \int vv^\top d\mu = \operatorname{tr}(S_g^{1/2})^2 P_g.$$

Thus the flow equalizes the velocity covariance on the active force subspace, rather than restricting motion to one direction. This is the mean-field counterpart of Muon's matrix orthogonalization.

Mean-field models for two-layer MLPs. We now specialize the abstract gradient-flow PDE to the standard mean-field parameterization of two-layer MLPs. The mean-field predictor is $H_\mu(z) \coloneqq \int_{\mathbb{R}^d} \phi(z, x)\, d\mu(x)$, and the training objective is $\mathcal{F}(\mu) \coloneqq R(H_\mu)$ for some loss R. For a two-layer network with scalar output and activation σ, one may take parameter $x = (u, v) \in \mathbb{R}^d$ and feature $\phi(z, x) \coloneqq u\,\sigma(v \cdot z)$, where u is the outer weight and v is the inner weight. This is the standard mean-field representation of a two-layer multilayer perceptron [Mei et al., 2018, Chizat and Bach, 2018]. The notation is a deliberate departure from usual machine-learning conventions: the network input is denoted by z, while the trainable variable is denoted by x so that it matches the transport notation used throughout the paper. Here d is the sum of the block dimensions of u and v. The experiment normalizes the single $n \times d$ particle matrix. Separate normalization of inner and outer weight blocks is also covered, using the sum or maximum block gauges of Appendix B, depending on the scalar normalization of each update. A basic example is the quadratic risk $R(H) = \frac{1}{2}\int |H(z) - y^\star(z)|^2 \, d\rho(z)$, or its empirical version on a dataset. Because H_μ depends linearly on μ, this makes $\mathcal{F}(\mu)$ a quadratic interaction functional. These energies are of MMD-type [Gretton et al., 2012], which is why the MMD experiment below is a natural test case for the Spectral Wasserstein flow. In Figure 3, we train a two-dimensional to one-dimensional ReLU network, with $\sigma(s) = \max(s, 0)$, particles $x = (u, v_1, v_2) \in \mathbb{R}^3$, and covariates $z \in \mathbb{R}^2$. The target $y^\star$ is generated by a three-neuron teacher on Gaussian covariates $z \sim \mathcal{N}(0, \mathrm{Id}_2)$. We display both runs on their respective unit-normalized nominal horizons, chosen by hand. This is a qualitative comparison, not a calibration of the intrinsic gradient-flow clocks: the implementation also clips large Euler steps. Appendix G.1 records the exact normalization, safeguards, and an initialization-only dissipation check. We display the particle evolution in the reduced two-dimensional plane $|u|v$. The rays associated with the teacher neurons are shown as dashed half-lines in Figure 3; additional loss and circular histogram plots are in Appendix G.2. Changing p visibly changes the paths and the loss profile on these normalized plots. We do not infer an intrinsic convergence-rate or wall-clock advantage from this comparison. The positively two-homogeneous structure of ReLU motivates a spherical reduction of the energy and homogeneous velocity action. The corresponding regularity and selector assumptions are stated in Appendix J; the nonsmooth ReLU case requires its own chain-rule justification.

Mean-field shallow attention. Using the same spectral-flow construction, we also consider a shallow mean-field multi-head attention model. This is a useful stress test because attention uses softmax nonlinearities, precisely the regime where Muon-type spectral normalization is empirically attractive. A datum is a sequence $z = (z_1, \dots, z_m)$ of tokens in $\mathbb{R}^{d_T}$. A head is a particle $x = (Q, K, V, O)$, with $Q, K, V, O \in \mathbb{R}^{d_H \times d_T}$, so $x \in \mathbb{R}^d$ with $d = 4d_H d_T$. Its elementary feature is



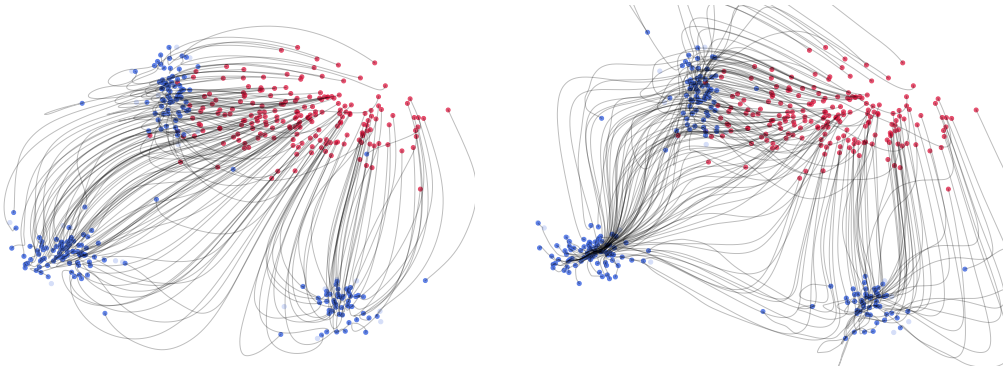
(a) $p = 1$ (b) $p = \infty$

Figure 3: Two-layer ReLU MLP trajectories in the reduced parameter plane $(|u|v_1, |u|v_2)$. Dashed rays indicate the teacher neurons. The archived runs use separately normalized nominal horizons; neither intrinsic flow times nor wall-clock runtimes are matched.

the softmax attention map $\phi(z, x)_j := O^\top \sum_{i=1}^m \frac{\exp(\langle Qz_j, Kz_i\rangle/\sqrt{d_H})}{\sum_{\ell=1}^m \exp(\langle Qz_j, Kz_\ell\rangle/\sqrt{d_H})} Vz_i$, and the mean-field layer is again $H_\mu(z) = \int \phi(z, x)d\mu(x)$. For an empirical measure μ_X, this is a multi-head attention layer averaged over an arbitrary number n of heads, which is the natural mean-field analogue of increasing the number of heads without fixing a finite architecture. The measure-level oracle requires $g_\mu \in L^2(\mu)$, which is not automatic from $\mu \in \mathcal{P}_2$. For bounded covariates and targets, the output has at most quadratic parameter growth, but derivatives in Q and K can have cubic growth. A finite sixth parameter moment is a sufficient integrability condition for the least-squares force. Preservation of this moment bound and regularity of the selected velocity remain additional requirements for a rigorous flow application.

Figure 9 in Appendix G.2 reports an experiment with $m = 16$ tokens, $d_T = 4$, and $d_H = 3$: we fit a random five-head teacher by least squares using a student with $n = 200$ heads, over five paired trials with new teacher and student initializations. The unfolding is a single 200×48 particle matrix. The query/key scale controls the teacher's attention sparsity, measured by the effective number of attended tokens $N_{\text{eff}} := (\sum_i w_i^2)^{-1}$ averaged over attention rows. The dense high-temperature regime has $N_{\text{eff}} \approx 16$, while the sparse low-temperature regime has $N_{\text{eff}} \approx 2.1$. The plots use separately normalized nominal horizons, with the safeguards specified in Appendix G.1; they are not calibrated intrinsic flow times. The different loss profiles illustrate the dependence on the geometry without establishing a convergence-rate ordering.

MMD gradient flows. As a simple model for generative fitting and sampling, in the spirit of MMD-based generative models [Li et al., 2015, Bińkowski et al., 2018], we also minimize $\mathcal{F}(\mu) = \text{MMD}(\mu, \nu)^2$ for the energy-distance kernel $k(x, y) = -\|x - y\|_2$. For empirical measures $\mu = n^{-1} \sum_i \delta_{x_i}$ and $\nu = m^{-1} \sum_j \delta_{y_j}$, with $n = m = 200$ particles in $\mathbb{R}^2$, this objective is $\text{MMD}(\mu, \nu)^2 = n^{-2} \sum_{i,i'} k(x_i, x_{i'}) + m^{-2} \sum_{j,j'} k(y_j, y_{j'}) - 2(nm)^{-1} \sum_{i,j} k(x_i, y_j)$, with a small smoothing of $\|x - y\|_2$ in the implementation. The same initial cloud evolves with $p = 1$ and $p = \infty$. Strictly speaking, $-\|x - y\|_2$ is conditionally positive definite on zero-mass signed measures, not a positive-definite kernel on all measures. Adding the usual anchoring terms gives a positive-definite kernel with the same quadratic energy on $\mu - \nu$ [Sejdinovic et al., 2013]. Figure 4 isolates the effect of the spectral tangent norm without the additional architectural structure of a neural network. The trace-norm flow reacts locally to the MMD force field, while the operator-norm flow produces more globally coordinated trajectories; the scalar loss curves remain close on the separately normalized horizons. A plausible explanation for the early outward motion in the operator case is amplification of small transverse force components: the canonical selector multiplies an active covariance mode of eigenvalue λ_i by $\text{tr}(S_g^{1/2})/\sqrt{\lambda_i}$. Its velocity covariance is isotropic on the active range, as in Remark 5.2. This is a mechanism consistent with the figure, not a separate proof of its transient behavior.



(a) $p = 1$ (b) $p = \infty$

Figure 4: MMD particle trajectories on separately normalized nominal horizons. The corresponding loss curves appear in Figure 11, Appendix G.2.

Gaussian-preserving linear networks. A particularly tractable example is a linear two-layer network, $H_\mu(z) = \int u(v^\top z)\, d\mu(u, v)$, trained with quadratic loss. Its force is linear in the parameter $x = (u, v)$. The canonical Measure-LMO preserves this property, because it multiplies the force by a matrix depending only on its covariance. An initially Gaussian distribution therefore remains Gaussian along any sufficiently regular solution, and the PDE closes on its mean and covariance. This illustrates a concrete advantage of the measure formulation: infinitely many neurons can be described by a finite-dimensional moment system. Appendix F derives this system, its modal reductions, and the p-dependent implicit covariance bias illustrated in Figures 5 and 6.

Geodesic convexity. Geodesic convexity is a useful structural benchmark for metric gradient flows [Ambrosio et al., 2008], but is generally too strong for the MLP and MMD objectives above. Appendix H studies convex potentials and interactions along static displacement geodesics, and relative entropy along a selected geodesic; Proposition H.1 collects the basic examples. Convexity along every metric geodesic is stronger and can fail at the Muon endpoint, where geodesics need not be unique.

A weaker route to convergence is a Polyak–Lojasiewicz (P-L) inequality, which requires $|\partial\mathcal{F}|_\gamma^2 \geq 2\kappa(\mathcal{F} - \mathcal{F}_\star)$, with $\mathcal{F}_\star = \inf \mathcal{F} > -\infty$, without geodesic convexity. If $\gamma(S) \leq C_\gamma \operatorname{tr}(S)$, then $|\partial\mathcal{F}|_\gamma \geq C_\gamma^{-1/2} |\partial\mathcal{F}|_2$; thus a W_2 P-L constant κ transfers as κ/C_γ (Proposition I.3). For every Schatten $p \geq 1$, including Muon, one can take $C_\gamma = 1$. Writing $E(t) = \mathcal{F}(\mu_t) - \mathcal{F}_\star$, a global κ-P-L flow with $E(0) < \infty$ satisfying $E' = -|\mu'|_\gamma^2 = -|\partial\mathcal{F}|_\gamma^2$ obeys $E(t) \leq E(0) e^{-2\kappa t}$. If $\mathcal{F}$ is also lower semicontinuous, then $\mathsf{W}_\gamma(\mu_t, \mu_\infty) \leq \sqrt{2E(0)/\kappa}\, e^{-\kappa t}$ for some minimizer μ_∞ (Proposition I.2). Appendix I gives the proofs and the local Poincaré analysis for relative entropy. These are conditional guarantees, not a convergence theorem or a trajectory-wise speed comparison for the networks above.

6 Conclusion

The main message of this paper is that matrix-normalized optimizers for mean-field neural models are naturally encoded by Spectral Wasserstein distances. In this view, matrix normalization is not merely a preconditioning device, but the local steepest-descent geometry generated by a spectral transport cost. This clarifies the transport side, through coupling-based costs, geodesics, and Gaussian reductions, and the optimization side, through a continuum interpretation of normalized matrix updates. The resulting picture is that the choice of spectral gauge shapes which directions of motion are emphasized or suppressed in the mean-field dynamics, thereby providing a geometric language for comparing matrix-normalized optimizers. A major open question is the rigorous study of the Spectral Wasserstein gradient-flow PDE: global-in-time existence, stability, and the avoidance of local minima for shallow neural-network and MMD losses.

Acknowledgments and Disclosure of Funding

The author thanks Tony Silveti-Falls, from whom he learned about the intricate details of Muon and with whom he had insightful discussions about Newton–Schulz approximation. This work was supported by the European Research Council (ERC project WOLF) and the French government under the management of Agence Nationale de la Recherche as part of the "France 2030" program, reference ANR-23-IACL-0008 (PRAIRIE-PSAI). The author has no competing financial interests to disclose.

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Basel, 2 edition, 2008. doi: 10.1007/978-3-7643-8722-8. URL https://link.springer.com/book/10.1007/978-3-7643-8722-8.

Julio Daniel Backhoff-Veraguas and Gudmund Pammer. Applications of weak transport theory. *Bernoulli*, 28(1):370–394, 2022. doi: 10.3150/21-BEJ1346. URL https://arxiv.org/abs/2003.05338.

Julio Daniel Backhoff-Veraguas, Mathias Beiglböck, and Gudmund Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6), 2019. doi: 10.1007/s00526-019-1624-y. URL https://link.springer.com/article/10.1007/s00526-019-1624-y. Article 203.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, Cham, 2014. ISBN 978-3-319-00226-2. doi: 10.1007/978-3-319-00227-9. URL https://link.springer.com/book/10.1007/978-3-319-00227-9.

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. doi: 10.1007/s002110050002. URL https://link.springer.com/article/10.1007/s002110050002.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. doi: 10.1016/j.exmath.2018.01.002. URL https://www.sciencedirect.com/science/article/pii/S0723086918300021.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Martin Burger, Matthias Erbar, Franca Hoffmann, Daniel Matthes, and André Schlichting. Covariance-modulated optimal transport and gradient flows. *Archive for Rational Mechanics and Analysis*, 249 (1), 2025. doi: 10.1007/s00205-024-02065-w. URL https://link.springer.com/article/10.1007/s00205-024-02065-w. Article 7.

Guillaume Carlier, Carlos Jimenez, and Filippo Santambrogio. Optimal transportation with traffic congestion and Wardrop equilibria. *SIAM Journal on Control and Optimization*, 47(3):1330–1350, 2008. doi: 10.1137/060672832. URL https://epubs.siam.org/doi/10.1137/060672832.

Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Matrix optimal mass transport: A quantum mechanical approach. *IEEE Transactions on Automatic Control*, 63(8):2612–2619, 2018. doi: 10.1109/TAC.2017.2767707. URL https://par.nsf.gov/servlets/purl/10114182.

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, pages 3036–3046, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a1afc58c6ca9540d057299ec3016d726-Abstract.html.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018. doi: 10.1016/j.jfa.2018.03.008. URL https://www.sciencedirect.com/science/article/pii/S0022123618301058.

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. URL https://proceedings.mlr.press/v119/cutkosky20b.html.

Marco Cuturi and David Avis. Ground metric learning. *Journal of Machine Learning Research*, 15 (17):533–564, 2014. URL https://jmlr.org/papers/v15/cuturi14a.html.

Ky Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences of the United States of America*, 37(11):760–766, 1951. doi: 10.1073/pnas.37.11.760. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC1063464/.

Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018. doi: 10.1007/s10994-018-5717-1. URL https://link.springer.com/article/10.1007/s10994-018-5717-1.

Nathael Gozlan, Cyril Roberto, Paul-Marie Samson, and Prasad Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017. doi: 10.1016/j.jfa.2017.08.015. URL https://www.sciencedirect.com/science/article/am/pii/S0022123617303294.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

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. URL https://proceedings.mlr.press/v80/gupta18a.html.

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, 8 December 2024.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3252–3261. PMLR, 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1718–1727, 2015. URL https://proceedings.mlr.press/v37/li15.html.

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. doi: 10.1007/s00222-017-0759-8. URL https://link.springer.com/article/10.1007/s00222-017-0759-8.

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. arXiv:2502.16982.

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. doi: 10.1073/pnas.1806579115. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6099898/.

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. URL https://arxiv.org/abs/1711.05224.

Lipeng Ning, Tryphon T. Georgiou, and Allen Tannenbaum. On matrix-valued Monge–Kantorovich optimal mass transport. *IEEE Transactions on Automatic Control*, 60(2):373–382, 2015. doi: 10.1109/TAC.2014.2350171. URL https://pubmed.ncbi.nlm.nih.gov/26997667/.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5072–5081, 2019. URL https://proceedings.mlr.press/v97/paty19a.html.

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. URL https://proceedings.mlr.press/v267/pethick25a.html.

Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Structured transforms across spaces with cost-regularized optimal transport. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 586–594, 2024. URL https://proceedings.mlr.press/v238/sebbouh24a.html.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140. URL https://arxiv.org/abs/1207.6076.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. URL https://msp.org/pjm/1958/8-1/pjm-v8-n1-p14-p.pdf.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

A Matrix Recap and Explicit LMOs

This appendix collects the matrix facts used in Section 2. The canonical polar representing set is $\mathcal{P}_\gamma := \{Q \in \mathbb{S}^d : \operatorname{tr}(QS) \le \gamma(S) \text{ for every } S \succeq 0\}$; finite-dimensional convex duality gives $\gamma(S) = \max_{Q\in\mathcal{P}_\gamma} \operatorname{tr}(QS)$ for every $S \succeq 0$.

We first show that the use of positive gauges on the cone $\mathbb{S}^d_+$ is not essentially more restrictive than working with norms on the full space of symmetric matrices: under monotonicity, every such gauge admits a canonical norm extension.

Proposition A.1 (Norm extension of a monotone positive gauge)**.** *Let $\gamma : \mathbb{S}^d_+ \to \mathbb{R}_+$ be a monotone positive gauge, i.e., γ is convex, positively one-homogeneous, $\gamma(S) > 0$ for every nonzero $S \succeq 0$, and $0 \preceq S \preceq S'$ implies $\gamma(S) \le \gamma(S')$. Define, for $A \in \mathbb{S}^d$,*

$$\|A\|_\gamma := \inf\Big\{\gamma(P) + \gamma(N) : P, N \succeq 0,\ A = P - N\Big\}.$$

Then $\|\cdot\|_\gamma$ is a norm on $\mathbb{S}^d$, and its restriction to $\mathbb{S}^d_+$ coincides with γ.

Proof. We first note that the admissible set is nonempty for every $A \in \mathbb{S}^d$: for instance, the spectral decomposition $A = A_+ - A_-$ gives $A_+, A_- \succeq 0$. Hence $\|A\|_\gamma$ is well-defined and finite.

Positive homogeneity follows directly from the definition. If $a \geq 0$, then

$$\|aA\|_\gamma = \inf_{A=P-N}\{\gamma(aP) + \gamma(aN)\} = a\|A\|_\gamma.$$

If $a < 0$, then $aA = (-a)N - (-a)P$ whenever $A = P - N$, so the same argument gives $\|aA\|_\gamma = |a|\|A\|_\gamma$. In particular, $\| - A\|_\gamma = \|A\|_\gamma$.

We next prove the triangle inequality. Let $A = P_1 - N_1$ and $B = P_2 - N_2$ with $P_i, N_i \succeq 0$. Then

$$A + B = (P_1 + P_2) - (N_1 + N_2),$$

and $P_1+P_2, N_1+N_2 \succeq 0$. By subadditivity of γ, which follows from convexity and one-homogeneity,

$$\gamma(P_1 + P_2) + \gamma(N_1 + N_2) \leq \gamma(P_1) + \gamma(P_2) + \gamma(N_1) + \gamma(N_2).$$

Taking the infimum over all decompositions of A and B yields $\|A + B\|_\gamma \leq \|A\|_\gamma + \|B\|_\gamma$.

It remains to prove nondegeneracy. Monotonicity and subadditivity imply $|\gamma(S)-\gamma(T)| \leq \gamma(I)\|S-T\|_{\mathrm{op}}$ on the PSD cone, so γ is continuous. Since it is strictly positive on the compact section $\{S \succeq 0 : \mathrm{tr}(S) = 1\}$, there exists $c > 0$ such that $\gamma(S) \geq c\ \mathrm{tr}(S)$ for every $S \succeq 0$. Therefore, for every decomposition $A = P - N$,

$$\gamma(P) + \gamma(N) \geq c\ \mathrm{tr}(P + N).$$

Moreover, from $A = P - N$ we have $P + N \succeq A$ and $P + N \succeq -A$. Let E_+ and E_- denote the spectral projectors of A associated with its positive and negative eigenvalues. Taking traces after compression to these two subspaces gives

$$\mathrm{tr}(E_+(P + N)) \geq \mathrm{tr}(A_+), \qquad \mathrm{tr}(E_-(P + N)) \geq \mathrm{tr}(A_-).$$

Since $P + N \succeq 0$, its trace on the kernel of A is nonnegative, and hence $\mathrm{tr}(P + N) \geq \mathrm{tr}\,|A|$. Therefore

$$\gamma(P) + \gamma(N) \geq c\ \mathrm{tr}\,|A|.$$

Taking the infimum gives $\|A\|_\gamma \geq c\ \mathrm{tr}\,|A|$. Thus $\|A\|_\gamma = 0$ implies $A = 0$.

Finally, let $S \succeq 0$. The decomposition $S = S - 0$ gives $\|S\|_\gamma \leq \gamma(S)$. Conversely, if $S = P - N$ with $P, N \succeq 0$, then $P = S + N \succeq S$, and by monotonicity, $\gamma(P) \geq \gamma(S)$. Hence

$$\gamma(P) + \gamma(N) \geq \gamma(P) \geq \gamma(S).$$

Taking the infimum over all such decompositions gives $\|S\|_\gamma \geq \gamma(S)$. Therefore $\|S\|_\gamma = \gamma(S)$ for every $S \succeq 0$. □

The next proposition characterizes the monotone norms for which representing sets may be chosen inside the positive-semidefinite cone.

Proposition A.2 (PSD representation and monotonicity)**.** *The following are equivalent:*

- γ *is monotone on* $\mathbb{S}^d_+$*, namely*

$$0 \preceq S \preceq T \Longrightarrow \gamma(S) \leq \gamma(T);$$

- *there exists a convex compact representing set* $\mathcal{K}_\gamma \subset \mathbb{S}^d_+$ *such that*

$$\gamma(S) = \max_{Q\in\mathcal{K}_\gamma} \mathrm{tr}(QS) \qquad \textit{for every } S \succeq 0.$$

Moreover, when these properties hold, the canonical choice

$$\mathcal{K}_\gamma = \mathcal{P}_\gamma \cap \mathbb{S}^d_+$$

is admissible.

Proof. If $\mathcal{K}_\gamma \subset \mathbb{S}^d_+$, then for $0 \preceq S \preceq T$ one has $\mathrm{tr}(QS) \leq \mathrm{tr}(QT)$ for every $Q \in \mathcal{K}_\gamma$, hence $\gamma(S) \leq \gamma(T)$ by taking suprema.

Conversely, assume γ is monotone and use Proposition A.1 to extend it to a norm on $\mathbb{S}^d$. Define

$$\mathcal{K}^+_\gamma := \{Q \in \mathbb{S}^d_+ :\ \mathrm{tr}(QS) \leq \gamma(S) \text{ for every } S \succeq 0\}.$$

It is immediate that $\mathcal{K}_\gamma^+$ is convex, closed, bounded, and contained in $\mathcal{P}_\gamma \cap \mathbb{S}_+^d$. We claim that it represents γ. Fix $S \succeq 0$. If $S \succ 0$, choose a subgradient $Q \in \partial\gamma(S)$ in the ambient space of symmetric matrices. Since γ is convex and one-homogeneous, $\operatorname{tr}(QS) = \gamma(S)$ and $\operatorname{tr}(QT) \le \gamma(T)$ for every $T \succeq 0$. Moreover $Q \succeq 0$: indeed, for every $R \succeq 0$ and every sufficiently small $\varepsilon > 0$ such that $S - \varepsilon R \succeq 0$, the subgradient inequality gives

$$\gamma(S - \varepsilon R) \ge \gamma(S) - \varepsilon \operatorname{tr}(QR),$$

while monotonicity gives $\gamma(S - \varepsilon R) \le \gamma(S)$. Hence $\operatorname{tr}(QR) \ge 0$ for all $R \succeq 0$, i.e., $Q \succeq 0$. Thus $Q \in \mathcal{K}_\gamma^+$ and $\gamma(S) \le \sup_{P \in \mathcal{K}_\gamma^+} \operatorname{tr}(PS)$. The reverse inequality follows from the definition of $\mathcal{K}_\gamma^+$. For singular $S \succeq 0$, apply the previous argument to $S + \varepsilon I$. Compactness gives a convergent subsequence of supporting matrices; passing to the limit and using continuity of γ yields equality at S. Therefore $\gamma(S) = \max_{Q \in \mathcal{K}_\gamma^+} \operatorname{tr}(QS)$ for every $S \succeq 0$. □

Proposition A.3 (Norm property of N)**.** *Assume that γ is monotone and admits a positive semidefinite representing set $\mathcal{K}_\gamma \subset \mathbb{S}_+^d$, so that $\gamma(S) = \max_{Q \in \mathcal{K}_\gamma} \operatorname{tr}(QS)$ for every $S \succeq 0$. Define $\mathrm{N}(V) := \gamma(V^\top V)^{1/2}$. Then N is a seminorm. If, in addition, $\gamma(S) > 0$ for every nonzero $S \succeq 0$, then N is a norm.*

Proof. For every matrix V, one has

$$\mathrm{N}(V)^2 = \gamma(V^\top V) = \max_{Q \in \mathcal{K}_\gamma} \operatorname{tr}(QV^\top V) = \max_{Q \in \mathcal{K}_\gamma} \operatorname{tr}(VQV^\top).$$

Since $Q \succeq 0$, we may write $\operatorname{tr}(VQV^\top) = \|VQ^{1/2}\|_F^2$. Hence

$$\mathrm{N}(V) = \max_{Q \in \mathcal{K}_\gamma} \|VQ^{1/2}\|_F.$$

For each fixed $Q \in \mathcal{K}_\gamma$, the map $V \mapsto \|VQ^{1/2}\|_F$ is a seminorm. Therefore, for any matrices V, W and scalar a,

$$\mathrm{N}(aV) = \max_{Q \in \mathcal{K}_\gamma} \|aVQ^{1/2}\|_F = |a|\mathrm{N}(V),$$

and

$$\mathrm{N}(V+W) = \max_{Q \in \mathcal{K}_\gamma} \|(V+W)Q^{1/2}\|_F \le \max_{Q \in \mathcal{K}_\gamma} \big(\|VQ^{1/2}\|_F + \|WQ^{1/2}\|_F\big) \le \mathrm{N}(V) + \mathrm{N}(W).$$

Thus N is a seminorm. If moreover $\gamma(S) > 0$ for every nonzero $S \succeq 0$, then $\mathrm{N}(V) = 0$ implies $\gamma(V^\top V) = 0$, hence $V^\top V = 0$, and therefore $V = 0$. Thus N is a norm. □

Remark A.4 ($\mathcal{K}_\gamma$ for the Schatten p-norms)**.** *For $\gamma = \gamma_p$, a convenient supporting set is obtained from the dual Schatten unit ball. Namely, if $p \in [1, \infty]$ and $p' \in [1, \infty]$ is the conjugate exponent, $1/p + 1/p' = 1$, one may take*

$$\mathcal{K}_{\gamma_p} = \mathcal{P}_{\gamma_p} \cap \mathbb{S}_+^d := \Big\{ G \in \mathbb{S}_+^d : \|G\|_{S_{p'}} \le 1 \Big\}.$$

One can, however, choose a smaller supporting set in the endpoint cases. For $p = 1$, since $\|A\|_{S_1} = \operatorname{tr}(A)$ on $\mathbb{S}_+^d$, it is enough to take $\mathcal{K}_{\gamma_1} = \{I\}$. For $p = \infty$, one may take the set of density matrices,

$$\mathcal{K}_{\gamma_\infty} = \Big\{ G \in \mathbb{S}_+^d : \operatorname{tr}(G) = 1 \Big\}.$$

The next proposition gives the explicit LMO associated with each Schatten geometry.

Proposition A.5 (Explicit Schatten selectors)**.** *Let $\gamma = \gamma_p = \|\cdot\|_{S_p}$, let $r = 2p$, and let $q = r' = 2p/(2p-1)$ be the dual exponent, with $q = 1$ for $p = \infty$. If*

$$G = U \operatorname{diag}(\sigma_i) W^\top$$

is a compact singular value decomposition over the positive singular values, then

$$\mathrm{J}_p(G) = -\|G\|_{S_q}^{2-q}\, U \operatorname{diag}(\sigma_i^{q-1}) W^\top.$$

For $G = 0$, set $\mathrm{J}_p(G) = 0$. In particular,

- $p = 1$: $\mathrm{J}_1(G) = -G$, *which recovers ordinary gradient flow;*
- $p = 2$: $\mathrm{J}_2(G) = -\|G\|_{S_{4/3}}^{2/3} U \operatorname{diag}(\sigma_i^{1/3}) W^\top$, *which we call the Frobenius-intermediate normalization;*
- $p = \infty$:
$$\mathrm{J}_\infty(G) = -\|G\|_{S_1} UW^\top = -\operatorname{tr}((G^\top G)^{1/2})\, G\, (G^\top G)^{\dagger/2}, \tag{6}$$
which is the Muon normalization.

Proof. The displayed formula is the negative gradient of $\frac{1}{2}\|G\|_{S_q}^2$ when $1 < q \le 2$. At the endpoint $q = 1$, the compact factor $UW^\top$ is a canonical partial-polar subgradient; it also fixes the value on the null singular subspaces. □

A.1 Proofs and dual formulas for measure LMOs

Proof of Proposition 2.4.

Proof. Associate to every vector field $h \in L^2(\mu; \mathbb{R}^d)$ the Hilbert–Schmidt operator $T_h : \mathbb{R}^d \to L^2(\mu)$ defined by $(T_h a)(x) = a \cdot h(x)$. Then

$$T_h^* T_h = \int h h^\top \, d\mu, \qquad \int g \cdot h \, d\mu = \langle T_g, T_h \rangle_{\mathrm{HS}}.$$

Let $S_g = T_g^* T_g$ and write the polar decomposition

$$T_g = U_g S_g^{1/2}, \qquad U_g^* U_g = P_g,$$

where P_g is the orthogonal projector onto $\operatorname{range}(S_g)$. Let $\Pi_g = U_g U_g^*$ be the orthogonal projector onto the range of T_g in $L^2(\mu)$. For any h, replacing T_h by $\Pi_g T_h$ leaves the linear term unchanged and decreases the covariance:

$$\langle T_g, \Pi_g T_h \rangle_{\mathrm{HS}} = \langle T_g, T_h \rangle_{\mathrm{HS}}, \qquad (\Pi_g T_h)^* (\Pi_g T_h) = T_h^* \Pi_g T_h \preceq T_h^* T_h.$$

By monotonicity of γ, this projection cannot increase the LMO objective. Hence it is enough to minimize over operators of the form $T_h = U_g A$ with $A = P_g A$. For such fields,

$$\int g \cdot h \, d\mu = \operatorname{tr}(S_g^{1/2} A), \qquad \int h h^\top \, d\mu = A^\top A.$$

Finally, the operator $T_h = U_g A$ corresponds to the vector field $h(x) = A^\top S_g^{\dagger/2} g(x)$. This proves the finite-dimensional reduction and the lifted selector. □

Corollary A.6 (Schatten measure selectors)**.** *Let $\gamma = \gamma_p = \|\cdot\|_{S_p}$ and set $q = (2p)'$, with $q = 1$ for $p = \infty$. Then*

$$\mathcal{J}_{\gamma_p}^{\mu}(g)(x) = -\|S_g^{1/2}\|_{S_q}^{2-q} S_g^{(q-2)/2} g(x), \qquad S_g = \int g g^\top \, d\mu,$$

where the power is taken on $\operatorname{range}(S_g)$ and is zero on its kernel; when $S_g = 0$, the selector is zero. In particular, $p = 1$ gives $\mathcal{J}_{\gamma_1}^{\mu}(g) = -g$, while $p = \infty$ gives

$$\mathcal{J}_{\gamma_\infty}^{\mu}(g)(x) = -\operatorname{tr}(S_g^{1/2})\, S_g^{\dagger/2} g(x).$$

Proof. By Proposition 2.4, one only needs to solve

$$\min_A \left\{ \operatorname{tr}(S_g^{1/2} A) + \frac{1}{2}\|A\|_{S_{2p}}^2 \right\}$$

on the active range. The minimizer is the Schatten duality map

$$A_* = -\|S_g^{1/2}\|_{S_q}^{2-q} S_g^{(q-1)/2}.$$

Lifting by $h(x) = A_*^\top S_g^{\dagger/2} g(x)$ gives the formula. □

The displayed Schatten selector is Borel measurable, including at singular covariances, by its spectral-calculus definition. Measurability does not imply continuity: for example, the canonical matrix selector sends $G_\varepsilon = \operatorname{diag}(1, \varepsilon)$ to $-(1+\varepsilon)I$ for $\varepsilon > 0$, but sends $G_0 = \operatorname{diag}(1, 0)$ to $-\operatorname{diag}(1, 0)$. Thus rank changes at the Muon endpoint require a separate analysis for ODE or PDE well-posedness. The full argmin is also not necessarily a singleton: for $G = \operatorname{diag}(1, 0)$ and the operator tangent norm, every $V = -\operatorname{diag}(1, a)$ with $|a| \le 1$ attains the minimum $-1/2$. The displayed canonical selector chooses $a = 0$.

The support representation of γ gives another useful construction. It is less intrinsic than the finite-dimensional LMO above, but when it applies it produces the explicit formula $v = -Q^{-1}g$ and identifies an active quadratic cost. Its limitation is the required attainment at an invertible matrix, which may fail for rank-deficient force covariances, including at $p = \infty$. This is not an inherent rank-one restriction at the operator endpoint: for $S_g \succ 0$ and $\mathcal{K}_\gamma = \{Q \succeq 0 : \operatorname{tr} Q \le 1\}$, the minimizer is $Q^* = S_g^{1/2} / \operatorname{tr}(S_g^{1/2}) \succ 0$. The rank-one certificates that can appear in the static transport problem play a different role.

Proposition A.7 (Inverse-trace dual certificate)**.** *Assume that $\mathcal{K}_\gamma \subset \mathbb{S}^d_+$ is a compact convex representing set for γ and that $\mathcal{K}_\gamma \cap \mathbb{S}^d_{++} \neq \emptyset$. Then*

$$\inf_v \left\{ \int g \cdot v \, d\mu + \frac{1}{2}\gamma\left(\int vv^\top \, d\mu \right) \right\} = -\frac{1}{2} \inf_{Q \in \mathcal{K}_\gamma \cap \mathbb{S}^d_{++}} \operatorname{tr}\big(Q^{-1} S_g\big), \qquad S_g = \mathcal{S}_\mu(g).$$

If the matrix infimum is attained at

$$Q^*_\mu \in \operatorname*{argmin}_{Q \in \mathcal{K}_\gamma \cap \mathbb{S}^d_{++}} \operatorname{tr}\big(Q^{-1} S_g\big),$$

then $v_*(x) = -(Q^*_\mu)^{-1} g(x)$ is a Measure-LMO.

Proof. For $Q \in \mathcal{K}_\gamma$, define

$$\Phi_Q(v) \coloneqq \int g \cdot v \, d\mu + \frac{1}{2} \int v^\top Q v \, d\mu.$$

Since $\gamma(R) = \max_{Q \in \mathcal{K}_\gamma} \operatorname{tr}(QR)$ for $R \succeq 0$, the Measure-LMO objective is $\inf_v \max_{Q \in \mathcal{K}_\gamma} \Phi_Q(v)$. The convex-concave minimax theorem gives

$$\inf_v \max_{Q \in \mathcal{K}_\gamma} \Phi_Q(v) = \max_{Q \in \mathcal{K}_\gamma} \inf_v \Phi_Q(v).$$

Indeed, the matrix set is compact and convex, while any $Q_0 \in \mathcal{K}_\gamma \cap \mathbb{S}^d_{++}$ provides coercivity in v. For fixed positive definite Q, the inner minimization gives

$$v_Q(x) = -Q^{-1} g(x), \qquad \inf_v \Phi_Q(v) = -\frac{1}{2} \operatorname{tr}\big(Q^{-1} \mathcal{S}_\mu(g)\big).$$

Convexity of $\mathcal{K}_\gamma$ allows singular matrices to be approximated by $(1-\varepsilon)Q + \varepsilon Q_0 \succ 0$, which proves the displayed inverse-trace identity. Now let Q^*_μ attain the matrix infimum and set $v_* = -(Q^*_\mu)^{-1} g$. The first-order optimality condition for the convex function $Q \mapsto \operatorname{tr}(Q^{-1} S_g)$ over $\mathcal{K}_\gamma$ gives, for every $Q \in \mathcal{K}_\gamma$,

$$\operatorname{tr}\big(Q\,(Q^*_\mu)^{-1} S_g (Q^*_\mu)^{-1}\big) \le \operatorname{tr}\big(Q^*_\mu (Q^*_\mu)^{-1} S_g (Q^*_\mu)^{-1}\big)\,.$$

Since

$$\int v_* v_*^\top \, d\mu = (Q^*_\mu)^{-1} S_g (Q^*_\mu)^{-1},$$

this says precisely that Q^*_μ is active in the support representation of γ at the covariance of v_*. Therefore

$$\int g \cdot v_* \, d\mu + \frac{1}{2}\gamma\left(\int v_* v_*^\top \, d\mu \right) = -\frac{1}{2} \operatorname{tr}\big((Q^*_\mu)^{-1} \mathcal{S}_\mu(g)\big),$$

which equals the common minimax value. Thus, v_* minimizes the Measure-LMO objective. □

Proof of Proposition 2.5.

Proof. At distinct atoms, prescribe $w(x_i) = H_i$ for an arbitrary matrix H. The transport chain rule gives

$$D\mathrm{F}(X)[H] = n \left.\frac{d}{dh}\mathcal{F}((\mathrm{Id} + hw)_{\#}\mu_X)\right|_{h=0} = \sum_i g_{\mu_X}(x_i) \cdot H_i.$$

Thus $\nabla \mathrm{F}(X) = G$, where $G_i = g_{\mu_X}(x_i)^\top$. At coincident atoms the corresponding rows of $\nabla \mathrm{F}$ are equal by permutation symmetry; moving the whole coincident group in the chain rule identifies each row with the same force. The transport expansion in Theorem 5.1 is a sufficient condition for this chain rule.

For distinct atoms, testing the weak continuity equation near each atom identifies its velocity almost everywhere. The same identification applies to coincident groups assigned a common velocity. The identity $\mathcal{N}_{\mu_X}(v)^2 = n^{-1}\mathrm{N}(V)^2$ and pairing $n^{-1}\operatorname{tr}(G^\top V)$ reduce Definition 2.3 to

$$\operatorname*{argmin}_V \left\{ \operatorname{tr}(G^\top V) + \frac{1}{2}\mathrm{N}(V)^2 \right\} = \mathrm{J}_\gamma(G).$$

With matching selections, stacking the velocities therefore gives $\dot{X}_t = \mathrm{J}_\gamma(\nabla \mathrm{F}(X_t))$, exactly ODE (2). Conversely, this ODE yields the weak continuity equation by differentiating test functions along the particles. For the unscaled objective $\widetilde{\mathrm{F}}(X) = \mathcal{F}(\mu_X)$, one instead has $\nabla\widetilde{\mathrm{F}}(X) = n^{-1}G_t$ and $\dot{X}_t = n^{-1}\mathrm{J}_\gamma(G_t)$ by one-homogeneity. This is the same weak measure equation after rescaling time by n. □

A.2 Comparison between Classical and Spectral Wasserstein Distances

Proof of Proposition 3.2.

Proof. Since γ is a norm on the finite-dimensional space $\mathbb{S}^d$, its restriction to the compact set $\{S \succeq 0 : \operatorname{tr}(S) = 1\}$ is continuous and strictly positive. Hence there exist constants $0 < c_\gamma \le C_\gamma < \infty$ such that $c_\gamma \operatorname{tr}(S) \le \gamma(S) \le C_\gamma \operatorname{tr}(S)$ for every $S \succeq 0$.

Let $\pi \in \Pi(\mu, \nu)$ be any coupling, and define its displacement covariance by

$$\Sigma_\pi := \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - y)(x - y)^\top \, d\pi(x, y).$$

Then $\Sigma_\pi \succeq 0$, and therefore

$$c_\gamma \operatorname{tr}(\Sigma_\pi) \le \gamma(\Sigma_\pi) \le C_\gamma \operatorname{tr}(\Sigma_\pi).$$

Moreover,

$$\operatorname{tr}(\Sigma_\pi) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \operatorname{tr}\big((x - y)(x - y)^\top\big) \, d\pi(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, d\pi(x, y).$$

Thus, for every coupling $\pi \in \Pi(\mu, \nu)$,

$$c_\gamma \int |x - y|^2 \, d\pi(x, y) \le \gamma(\Sigma_\pi) \le C_\gamma \int |x - y|^2 \, d\pi(x, y).$$

Taking infima over $\pi \in \Pi(\mu, \nu)$ gives

$$c_\gamma \mathsf{W}_2(\mu, \nu)^2 \le \mathsf{W}_\gamma(\mu, \nu)^2 \le C_\gamma \mathsf{W}_2(\mu, \nu)^2,$$

where we used the definitions of W_2 and W_γ. Taking square roots yields

$$\sqrt{c_\gamma}\, \mathsf{W}_2(\mu, \nu) \le \mathsf{W}_\gamma(\mu, \nu) \le \sqrt{C_\gamma}\, \mathsf{W}_2(\mu, \nu).$$

□

Schatten constants and sharpness. For Schatten norms, $d^{1/p-1}\,\mathsf{W}_2(\mu,\nu)^2 \le \mathsf{W}_{\gamma_p}(\mu,\nu)^2 \le \mathsf{W}_2(\mu,\nu)^2$, equivalently $c_{\gamma_p} = d^{1/p-1}$ and $C_{\gamma_p} = 1$. In particular, $(c_{\gamma_2}, C_{\gamma_2}) = (d^{-1/2}, 1)$ and $(c_{\gamma_\infty}, C_{\gamma_\infty}) = (d^{-1}, 1)$. These constants follow from $\|\lambda\|_{\ell_p} \le \|\lambda\|_{\ell_1} \le d^{1-1/p}\,\|\lambda\|_{\ell_p}$.

The lower Schatten bound is sharp. For $\gamma = \gamma_p$, equality in the lower bound is attained by isotropic Gaussian pairs $\mu = \mathcal{N}(0, \alpha I_d)$ and $\nu = \mathcal{N}(0, \beta I_d)$, $\alpha, \beta > 0$. By the Gaussian computation of Appendix D, the displacement covariance is $(\sqrt{\alpha} - \sqrt{\beta})^2 I_d$, hence $\mathsf{W}_{\gamma_p}(\mu,\nu)^2 = d^{1/p-1}\mathsf{W}_2(\mu,\nu)^2$.

Proof of Theorem 3.4.

Proof. Let $\pi \in \Pi(\mu,\nu)$ and set

$$\Sigma_\pi \coloneqq \int_{\mathbb{R}^d\times\mathbb{R}^d} (y-x)(y-x)^\top \, d\pi(x,y).$$

Since $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the matrix Σ_π is well-defined and belongs to $\mathbb{S}^d_+$. By the support representation of γ,

$$\gamma(\Sigma_\pi) = \max_{Q\in\mathcal{K}_\gamma} \operatorname{tr}(Q\Sigma_\pi).$$

Moreover, for every $Q \in \mathcal{K}_\gamma$,

$$\operatorname{tr}(Q\Sigma_\pi) = \int_{\mathbb{R}^d\times\mathbb{R}^d} (y-x)^\top Q(y-x)\, d\pi(x,y).$$

Therefore

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \max_{Q\in\mathcal{K}_\gamma} \int (y-x)^\top Q(y-x)\, d\pi(x,y).$$

It remains to justify the exchange of $\inf_\pi$ and $\max_Q$. Define

$$F(\pi,Q) \coloneqq \int_{\mathbb{R}^d\times\mathbb{R}^d} (y-x)^\top Q(y-x)\, d\pi(x,y).$$

The coupling set $\Pi(\mu,\nu)$ is convex and weakly compact, since its marginals are fixed. The set $\mathcal{K}_\gamma$ is convex and compact by assumption. For fixed π, the map $Q \mapsto F(\pi,Q)$ is affine, hence both concave and convex, and is continuous on $\mathcal{K}_\gamma$. For fixed $Q \succeq 0$, the cost $(x,y) \mapsto (y-x)^\top Q(y-x)$ is nonnegative and lower semicontinuous. Hence $\pi \mapsto F(\pi,Q)$ is lower semicontinuous for weak convergence of probability measures, by the Portmanteau theorem. It is also affine in π.

Thus the lower-semicontinuous form of Sion's minimax theorem Sion [1958] applies and gives

$$\inf_{\pi\in\Pi(\mu,\nu)} \max_{Q\in\mathcal{K}_\gamma} F(\pi,Q) = \max_{Q\in\mathcal{K}_\gamma} \inf_{\pi\in\Pi(\mu,\nu)} F(\pi,Q).$$

The only point requiring care is that the quadratic cost is not bounded on $\mathbb{R}^d \times \mathbb{R}^d$. This is precisely why one uses the lower-semicontinuous version of the minimax theorem rather than the bounded continuous one. Consequently,

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \max_{Q\in\mathcal{K}_\gamma} \inf_{\pi\in\Pi(\mu,\nu)} \int (y-x)^\top Q(y-x)\, d\pi(x,y).$$

Finally, by definition of the quadratic optimal transport cost associated with Q,

$$\mathsf{W}^Q(\mu,\nu)^2 = \inf_{\pi\in\Pi(\mu,\nu)} \int (y-x)^\top Q(y-x)\, d\pi(x,y).$$

Therefore

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \max_{Q\in\mathcal{K}_\gamma} \mathsf{W}^Q(\mu,\nu)^2,$$

which proves the value formula. Finally, if $(Q^\star, \pi^\star)$ is a saddle point, then $Q^\star$ maximizes $Q \mapsto \operatorname{tr}(Q\Sigma_{\pi^\star})$ over $\mathcal{K}_\gamma$. Since γ is the support function of $\mathcal{K}_\gamma$ on $\mathbb{S}^d_+$, this is equivalent to the cone subgradient relation $Q^\star \in \partial\gamma(\Sigma_{\pi^\star})$, proving the stated optimality relation. □

Proof of Corollary 3.5. For each fixed $Q \succeq 0$, Kantorovich duality for the continuous, nonnegative quadratic cost identifies the inner supremum with $\mathsf{W}^Q(\mu,\nu)^2$. This remains valid for singular Q: equivalently, one may transport the pushforwards by $Q^{1/2}$. Apply Theorem 3.4. The outer maximum is attained by compactness of $\mathcal{K}_\gamma$ and upper semicontinuity of the infimum of linear functions of Q. □

A.3 Metric property from the robust-cost formulation

The max–min representation of Theorem 3.4 gives a direct proof of the triangle inequality when γ is monotone. This avoids using the Benamou–Brenier formulation for this specific purpose.

Proposition A.8 (Static proof of the metric property)**.** *Assume that γ is monotone on $\mathbb{S}^d_+$. Then W_γ is a distance on $\mathcal{P}_2(\mathbb{R}^d)$ and induces the same topology as W_2.*

Proof. By Proposition A.2, monotonicity allows us to choose a representing set $\mathcal{K}_\gamma \subset \mathbb{S}^d_+$ such that

$$\gamma(S) = \max_{Q \in \mathcal{K}_\gamma} \operatorname{tr}(QS), \qquad S \succeq 0.$$

Hence Theorem 3.4 gives

$$\mathsf{W}_\gamma(\mu, \nu) = \max_{Q \in \mathcal{K}_\gamma} \mathsf{W}^Q(\mu, \nu).$$

For a fixed $Q \in \mathcal{K}_\gamma$, write

$$d_Q(x, y) \coloneqq \left\| Q^{1/2}(x - y) \right\|_2.$$

Since $Q \succeq 0$, d_Q is a seminorm distance on $\mathbb{R}^d$, and therefore W^Q is the corresponding quadratic Wasserstein pseudodistance. In particular, for all $\mu_0, \mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathsf{W}^Q(\mu_0, \mu_2) \le \mathsf{W}^Q(\mu_0, \mu_1) + \mathsf{W}^Q(\mu_1, \mu_2).$$

Taking the maximum over $Q \in \mathcal{K}_\gamma$ yields

$$\begin{aligned}
\mathsf{W}_\gamma(\mu_0, \mu_2) &= \max_{Q \in \mathcal{K}_\gamma} \mathsf{W}^Q(\mu_0, \mu_2) \\
&\le \max_{Q \in \mathcal{K}_\gamma} \left\{ \mathsf{W}^Q(\mu_0, \mu_1) + \mathsf{W}^Q(\mu_1, \mu_2) \right\} \\
&\le \max_{Q \in \mathcal{K}_\gamma} \mathsf{W}^Q(\mu_0, \mu_1) + \max_{Q \in \mathcal{K}_\gamma} \mathsf{W}^Q(\mu_1, \mu_2) \\
&= \mathsf{W}_\gamma(\mu_0, \mu_1) + \mathsf{W}_\gamma(\mu_1, \mu_2).
\end{aligned}$$

Symmetry is immediate from the definition. Moreover $\mathsf{W}_\gamma(\mu, \mu) = 0$, and separation follows from Proposition 3.2: if $\mathsf{W}_\gamma(\mu, \nu) = 0$, then $\mathsf{W}_2(\mu, \nu) = 0$, hence $\mu = \nu$. The same comparison bounds also show that W_γ and W_2 induce the same topology on $\mathcal{P}_2(\mathbb{R}^d)$. □

Remark A.9 (Relation with subspace robust Wasserstein distances)**.** *The proof is the same robust-cost argument used by Paty and Cuturi [2019] for Ky Fan gauges. The only point needed in their argument is that the maximizing cost matrices are positive semidefinite. By Proposition A.2, this positivity holds for any monotone gauge γ, not only for Ky Fan gauges. Thus the Paty–Cuturi triangle-inequality proof extends verbatim to the present monotone-gauge setting.*

A.4 Large-mean asymptotics for Schatten geometries

We now describe the behavior of W_γ when the displacement between the means is much larger than the centered fluctuations of the measures. This regime is useful to interpret the optimizer in the max-min formula: for Schatten geometries with $p > 1$, the optimal certificate becomes increasingly aligned with the direction of the mean displacement, and the residual transport problem is effectively measured only along this direction.

Proposition A.10 (Large-mean asymptotics for Schatten geometries)**.** *Let $\gamma = \gamma_p$ be the Schatten-p norm, with $p \in [1, \infty]$. Fix centered measures $\bar{\mu}, \bar{\nu} \in \mathcal{P}_2(\mathbb{R}^d)$ and a unit vector $u \in \mathbb{R}^d$. For each $R > 0$, let μ_R, ν_R have centered laws $\bar{\mu}, \bar{\nu}$ and means whose difference is Ru. Then, as $R \to \infty$, the following asymptotics hold. For $p = 1$,*

$$\mathsf{W}_\gamma(\mu_R, \nu_R)^2 = R^2 + \mathsf{W}_2(\bar{\mu}, \bar{\nu})^2.$$

For $1 < p \le \infty$,

$$\mathsf{W}_\gamma(\mu_R, \nu_R)^2 = R^2 + \mathsf{W}^{uu^\top}(\bar{\mu}, \bar{\nu})^2 + o(1).$$

Proof. We first separate the mean displacement from the centered transport. Fix $R > 0$, abbreviate $\mu = \mu_R$, $\nu = \nu_R$, set $m_\mu = \int x \, d\mu(x)$, $m_\nu = \int y \, d\nu(y)$, and let $\Delta = m_\nu - m_\mu = Ru$. Fix $Q \succeq 0$. If $\pi \in \Pi(\mu, \nu)$ and if $\bar{\pi}$ denotes the corresponding coupling of $\bar{\mu}$ and $\bar{\nu}$, then writing $x = m_\mu + \bar{x}$ and $y = m_\nu + \bar{y}$ gives

$$y - x = \Delta + (\bar{y} - \bar{x}).$$

Therefore

$$\int (y-x)^\top Q(y-x)\, d\pi(x,y) = \Delta^\top Q \Delta + 2\Delta^\top Q \int (\bar{y} - \bar{x})\, d\bar{\pi}(\bar{x}, \bar{y}) + \int (\bar{y} - \bar{x})^\top Q (\bar{y} - \bar{x})\, d\bar{\pi}(\bar{x}, \bar{y}).$$

The middle term vanishes because $\bar{\mu}$ and $\bar{\nu}$ have zero mean. Taking the infimum over couplings yields the exact decomposition

$$\mathsf{W}^Q(\mu, \nu)^2 = \Delta^\top Q \Delta + \mathsf{W}^Q(\bar{\mu}, \bar{\nu})^2.$$

By Theorem 3.4,

$$\mathsf{W}_\gamma(\mu, \nu)^2 = \max_{Q \in \mathcal{K}_{\gamma_p}} \left\{ \Delta^\top Q \Delta + \mathsf{W}^Q(\bar{\mu}, \bar{\nu})^2 \right\}.$$

For $p = 1$, we use the reduced representing set $\mathcal{K}_{\gamma_1} = \{\mathrm{Id}\}$. Then

$$\mathsf{W}_\gamma(\mu, \nu)^2 = \Delta^\top \Delta + \mathsf{W}^{\mathrm{Id}}(\bar{\mu}, \bar{\nu})^2 = R^2 + \mathsf{W}_2(\bar{\mu}, \bar{\nu})^2,$$

as claimed.

Assume now that $1 < p \leq \infty$, and let p' be the conjugate exponent, with $p' = 1$ when $p = \infty$. We use the positive dual Schatten representing set

$$\mathcal{K}_{\gamma_p} = \{Q \succeq 0 : \|Q\|_{S_{p'}} \leq 1\}.$$

Since $\Delta = Ru$, the max-min formula becomes

$$\mathsf{W}_\gamma(\mu, \nu)^2 = \max_{Q \in \mathcal{K}_{\gamma_p}} \left\{ R^2 u^\top Q u + \mathsf{W}^Q(\bar{\mu}, \bar{\nu})^2 \right\}.$$

By Schatten duality, applied to the rank-one matrix $uu^\top$,

$$\max_{Q \in \mathcal{K}_{\gamma_p}} u^\top Q u = \max_{\substack{Q \succeq 0 \\ \|Q\|_{S_{p'}} \leq 1}} \mathrm{tr}(Q\, uu^\top) = \left\| uu^\top \right\|_{S_p} = 1.$$

For $1 < p < \infty$, the equality case in Hölder's inequality is unique and gives $Q = uu^\top$. For $p = \infty$, the constraint is $\mathrm{tr}(Q) \leq 1$, and $u^\top Q u \leq \mathrm{tr}(Q) \leq 1$; equality again forces $Q = uu^\top$. Thus the exposed face of $\mathcal{K}_{\gamma_p}$ in the direction $uu^\top$ is the singleton $\{uu^\top\}$ for every $1 < p \leq \infty$.

Let Q_R be an optimal certificate for the previous maximization problem. Since $\mathcal{K}_{\gamma_p}$ is compact, any sequence $R_n \to \infty$ admits a subsequence such that $Q_{R_n} \to Q_\infty \in \mathcal{K}_{\gamma_p}$. Dividing the optimality relation by R_n^2 and passing to the limit shows that Q_∞ maximizes $Q \mapsto u^\top Q u$ over $\mathcal{K}_{\gamma_p}$. By the uniqueness of the exposed maximizer, $Q_\infty = uu^\top$. Hence every maximizing certificate satisfies

$$Q_R \longrightarrow uu^\top \qquad \text{as } R \to \infty.$$

Set $b(Q) = \mathsf{W}^Q(\bar{\mu}, \bar{\nu})^2$. This map is Lipschitz on $\mathbb{S}_+^d$: indeed, comparing the two infima and using $|(\bar{y} - \bar{x})^\top (Q - Q')(\bar{y} - \bar{x})| \leq \|Q - Q'\|_{\mathrm{op}} |\bar{y} - \bar{x}|^2$ gives

$$|b(Q) - b(Q')| \leq 2\|Q - Q'\|_{\mathrm{op}} \left(\int |\bar{x}|^2 \, d\bar{\mu} + \int |\bar{y}|^2 \, d\bar{\nu} \right).$$

Consequently,

$$b(Q_R) \longrightarrow b(uu^\top).$$

Convergence of the certificates alone would not yet control the R^2-weighted term. Optimality of Q_R, compared with the admissible choice $uu^\top$, supplies the missing estimate:

$$0 \leq R^2(1 - u^\top Q_R u) \leq b(Q_R) - b(uu^\top) = o(1).$$

Substituting both estimates into the max–min formula gives

$$\mathsf{W}_\gamma(\mu,\nu)^2 = R^2 + b(uu^\top) + o(1).$$

Taking square roots yields

$$\mathsf{W}_\gamma(\mu,\nu) = R + \frac{1}{2R}\mathsf{W}^{uu^\top}(\bar\mu,\bar\nu)^2 + o(R^{-1}),$$

which completes the proof. □

The proposition shows that all Schatten geometries have the same leading large-mean behavior, namely $\mathsf{W}_\gamma(\mu_R,\nu_R) \sim R$. The difference appears at the next order and in the optimizer selected by the max-min formula. For the reduced trace geometry $p = 1$, the certificate is fixed and isotropic, $Q = \mathrm{Id}$, so the centered correction is the usual Wasserstein cost $\mathsf{W}_2(\bar\mu,\bar\nu)^2$. By contrast, for every $p > 1$, the optimal certificate aligns with the normalized mean displacement u, and the centered correction becomes $\mathsf{W}^{uu^\top}(\bar\mu,\bar\nu)^2$. This is exactly the one-dimensional W_2^2 cost between the projected marginals $(u^\top\cdot)_\#\bar\mu$ and $(u^\top\cdot)_\#\bar\nu$. Thus the first centered correction depends only on fluctuations along u; transverse fluctuations do not enter at this order.

B Connection with Muon, Orthogonalization, and Momentum

This appendix complements Section 2 by linking the matrix ODE for $p = \infty$ to practical Muon variants used at scale, and by showing how Newton–Schulz orthogonalization and momentum fit in the same normalized-flow language.

Taking into account Newton–Schulz orthogonalization. Setting $p = \infty$ in Proposition A.5 yields the LMO (6), so the spectral matrix flow

$$\dot X_t = \mathrm{J}_\infty(\nabla \mathrm{F}(X_t))$$

is exactly the vanishing-step-size, vanishing-momentum, full-batch limit of idealized Muon in the normalization convention used throughout this paper.

An important practical feature of Muon is that one does not compute the exact polar factor $UW^\top$ by an SVD at every step. To make the method scalable and GPU-friendly, practical implementations replace $UW^\top = G(G^\top G)^{\dagger/2}$ by a few Newton–Schulz iterations [Jordan et al., 2024, Liu et al., 2025]. After rescaling G by a scalar $\alpha(G) > 0$ chosen so that the spectrum of $\alpha(G)^{-2}G^\top G$ lies in $[0,1]$, the classical Newton–Schulz iteration reads

$$Z_{k+1} = \frac{1}{2}Z_k(3I - Z_k^\top Z_k), \qquad Z_0 = \alpha(G)^{-1}G.$$

For the discussion below, let us freeze the scale and regard $\alpha > 0$ as fixed a priori. Then after one iteration one obtains

$$\mathcal{U}_{\alpha,1}(G) \coloneqq -Z_1 = -\frac{1}{2\alpha}\, G\left(3I - \alpha^{-2}G^\top G\right),$$

and after k iterations one obtains more generally a polynomial update of the form

$$\mathcal{U}_{\alpha,k}(G) = -G\, p_{\alpha,k}\left(G^\top G\right),$$

where $p_{\alpha,k}$ is a matrix polynomial approximating $s \mapsto s^{-1/2}$ on the spectral interval of interest. In this fixed-α setting, the polynomial orthogonalization part does admit a scalar generating function: if

$$p_{\alpha,k}(s) = \sum_{\ell=0}^{m} a_\ell s^\ell,$$

then

$$\Psi_{\alpha,k}(S) \coloneqq \sum_{\ell=0}^{m} \frac{a_\ell}{\ell+1}\,\mathrm{tr}(S^{\ell+1})$$

satisfies

$$\nabla\Psi_{\alpha,k}(S) = p_{\alpha,k}(S) \qquad \text{and} \qquad \mathcal{U}_{\alpha,k}(G) = -\nabla_G\left(\frac{1}{2}\Psi_{\alpha,k}(G^\top G)\right).$$

For one Newton–Schulz step this gives the explicit formula

$$\Psi_{\alpha,1}(S) = \frac{3}{2\alpha}\operatorname{tr}(S) - \frac{1}{4\alpha^3}\operatorname{tr}(S^2).$$

This function is not a norm, and it is not globally monotone on $\mathbb{S}^d_+$, because

$$\nabla\Psi_{\alpha,1}(S) = \frac{3}{2\alpha}I - \frac{1}{2\alpha^3}S$$

ceases to be positive semidefinite once S has eigenvalues larger than $3\alpha^2$. It is nevertheless monotone on the truncated spectral region

$$0 \preceq S \preceq 3\alpha^2 I,$$

and in particular on the standard Newton–Schulz regime $0 \preceq S \preceq \alpha^2 I$.

Thus a fixed-α Newton–Schulz update is generated by the spectral potential $\Psi_{\alpha,k}$. It must be interpreted on the *force* side, not substituted for the primal velocity gauge γ. Although this polynomial is generally neither convex in S nor homogeneous, there is a precise convex-dissipation interpretation on the stable region.

Proposition B.1 (Fixed-scale Newton–Schulz dissipation)**.** *For the classical iteration above with fixed α and $k \geq 1$, set $D^*_{\alpha,k}(G) = \frac{1}{2}\Psi_{\alpha,k}(G^\top G)$. On $\{G : \|G\|_{\mathrm{op}} \leq \alpha\}$ this is convex, and $\Psi_{\alpha,k}$ is Loewner-monotone on $0 \preceq S \preceq \alpha^2 I$. There is a global convex extension $\widehat{D}^*_{\alpha,k}$ agreeing with $D^*_{\alpha,k}$ on this region. With $D_{\alpha,k} = (\widehat{D}^*_{\alpha,k})^*$, for $\|G\|_{\mathrm{op}} \leq \alpha$ one has*

$$\mathcal{U}_{\alpha,k}(G) = -\nabla\widehat{D}^*_{\alpha,k}(G) \in \operatorname*{argmin}_V\{\langle G, V\rangle_F + D_{\alpha,k}(V)\}.$$

Proof. Let $h_0(s) = s/\alpha$ and $h_{j+1}(s) = (3h_j(s) - h_j(s)^3)/2$ for $0 \leq s \leq \alpha$. Induction gives $0 \leq h_j \leq 1$ and $h_j' \geq 0$ there, with $h_j(\alpha) = 1$. The singular-value response is h_k, so $D^*_{\alpha,k}(G) = \sum_i \int_0^{\sigma_i(G)} h_k(s)\,ds$. Extend h_k by 1 for $s \geq \alpha$. Its primitive is convex and nondecreasing; denote the resulting even convex spectral function by $\widehat{D}^*_{\alpha,k}$. Its gradient agrees with $-\mathcal{U}_{\alpha,k}$ in the stable region. Moreover, $p_{\alpha,k}(s) = h_k(\sqrt{s})/\sqrt{s} \geq 0$ there (with the continuous value at $s = 0$), which proves Loewner monotonicity by integration along ordered matrix segments. For $\|G\|_{\mathrm{op}} \leq \alpha$, Fenchel duality gives $\mathcal{U}_{\alpha,k}(G) \in \arg\min_V\{\langle G, V\rangle_F + D_{\alpha,k}(V)\}$. □

For one step, the singular-value derivative of the dual dissipation is $h_1(s) = 3s/(2\alpha) - s^3/(2\alpha^3)$, with $h_1'(s) = 3(1 - s^2/\alpha^2)/(2\alpha) \geq 0$ on the stable region. This illustrates why convexity of the dissipation, rather than convexity of Ψ as a function of the covariance, is the relevant condition. The same spectral construction works for fields $g \in L^2(\mu; \mathbb{R}^d)$, using $S_g = \int gg^\top d\mu$ and the Hilbert-space operator T_g from Proposition 2.4. It gives a convex dissipation and an energy-decreasing generalized gradient system while $S_{g_{\mu_t}} \preceq \alpha^2 I$. Here α bounds the square root of the *normalized* force covariance. For $\mu = \mu_X$ and $M_i = g_\mu(x_i)^\top$, the stacked field velocity is

$$-Mp_{\alpha,k}(M^\top M/n) = \sqrt{n}\,\mathcal{U}_{\sqrt{n}\alpha,k}(M), \qquad M^\top M/n \preceq \alpha^2 I. \tag{7}$$

Indeed, rescaling $M/\sqrt{n}$ by α and M by $\sqrt{n}\alpha$ gives the same initial Newton–Schulz iterate. The field convention therefore differs from applying $\mathcal{U}_{\alpha,k}$ directly to M; the homogeneous clock conversion of Proposition 2.5 cannot be reused at fixed α.

What does not follow is a gradient flow for a fixed W_γ metric: finite-step fixed-scale maps are not one-homogeneous in the force, whereas the duality map of any squared norm is. A dynamic transport *action* can be defined using the conjugate dissipation, but the squared-distance BB theorem in this paper does not apply to it. Monotonicity alone does not fix this homogeneity obstruction. A priori control keeping the evolving force covariance in the stable region is an additional requirement. Adaptive rescaling and the nonclassical polynomial coefficients used in practical implementations require their own analysis.

Taking into account momentum. A natural momentum extension, in the spirit of Muon with exponential moving average, is obtained by introducing an auxiliary matrix variable A_t and considering

$$\tau \dot{A}_t + A_t = \nabla \mathrm{F}(X_t), \qquad \dot{X}_t = \mathrm{J}_\gamma(A_t), \tag{8}$$

where $\tau \geq 0$ is the memory time scale. The variable A_t plays the role of a continuous-time exponentially averaged gradient. When $\tau = 0$, one recovers exactly the first-order spectral flow studied above.

For finite Schatten geometries ($1 \leq p < \infty$), this system is equivalent to the second-order inertial equation

$$\tau \frac{d}{dt}\big(\mathrm{I}_p(\dot{X}_t)\big) + \mathrm{I}_p(\dot{X}_t) + \nabla \mathrm{F}(X_t) = 0,$$

where

$$\mathrm{I}_p(V) \coloneqq \|V\|_{S_r}^{2-r}\, U \operatorname{diag}(\sigma_i^{r-1}) W^\top, \qquad r = 2p,$$

for $V = U \operatorname{diag}(\sigma_i) W^\top$. In particular, for $p = 1$ one has $\mathrm{I}_1(V) = V$ and one recovers the heavy-ball equation

$$\tau \ddot{X}_t + \dot{X}_t + \nabla \mathrm{F}(X_t) = 0.$$

We set $\mathrm{I}_p(0) = 0$. At the endpoint $p = \infty$, (8) reads

$$\tau \dot{A}_t + A_t = \nabla \mathrm{F}(X_t), \qquad \dot{X}_t = -\,\|A_t\|_{S_1}\, U_t W_t^\top,$$

which is the exact-polar continuous-time Muon dynamics with momentum.

The corresponding mean-field description naturally lives on phase space. Let

$$\eta_t \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$$

be the law of pairs (x, a), where x is the particle position and a is the exponentially averaged force. Its spatial marginal is

$$\mu_t = (\pi_x)_{\#} \eta_t.$$

For a functional $\mathcal{F}$ on measures, write its Wasserstein force as

$$g_\mu(x) = \nabla_x \frac{\delta \mathcal{F}}{\delta \mu}(x).$$

Given a phase-space law η, define its normalized spatial velocity by the phase-space Measure-LMO

$$u_\eta \in \operatorname*{argmin}_{u \in L^2(\eta; \mathbb{R}^d)} \left\{ \int a \cdot u(x,a)\, d\eta(x,a) + \frac{1}{2}\gamma\left(\int uu^\top \, d\eta \right) \right\}.$$

The operator proof of Proposition 2.4 applies verbatim with phase space as the underlying probability space and computes u_η from the momentum covariance $S(\eta) = \int aa^\top \, d\eta$. In particular, this definition remains valid for singular selectors such as $p = \infty$. The phase-space characteristics are

$$\dot{x}_t = u_{\eta_t}(x_t, a_t), \qquad \tau \dot{a}_t = g_{\mu_t}(x_t) - a_t. \tag{9}$$

Consequently, the law η_t solves the Liouville-type equation

$$\partial_t \eta_t + \operatorname{div}_x(\eta_t u_t) + \frac{1}{\tau} \operatorname{div}_a\big(\eta_t (g_{\mu_t}(x) - a)\big) = 0, \tag{10}$$

where

$$u_t(x,a) = u_{\eta_t}(x,a), \qquad \mu_t = (\pi_x)_{\#} \eta_t.$$

When the inverse-trace problem in Proposition A.7 is attained at an invertible Q_η^*, this constitutive relation reduces to the closed form $u_\eta(x,a) = -(Q_\eta^*)^{-1} a$. When $\tau = 0$, the averaged force instantaneously relaxes to $a = g_{\mu_t}(x)$, and the spatial marginal formally recovers the first-order normalized PDE (3). For $\tau > 0$, the dynamics is no longer a metric gradient flow for W_γ on the position marginal alone; it is a damped kinetic system in which the spectral geometry enters through the constitutive relation between the phase-space force covariance and the spatial velocity.

Stochastic minibatches. An unbiased force estimate need not give an unbiased spectral velocity: $\mathbb{E}[\mathcal{J}_\gamma^\mu(g_\mu + \xi)]$ generally differs from $\mathcal{J}_\gamma^\mu(g_\mu)$ even if $\mathbb{E}\xi = 0$. Bias after nonlinear normalization can prevent convergence; sign-based counterexamples are given by Karimireddy et al. [2019]. These examples do not establish failure for every stochastic Muon variant, but rule out transferring unbiased-gradient arguments without further analysis. Under a fast-resampling closure with fixed-variance random force fields, the candidate transport equation is

$$\partial_t \mu_t + \operatorname{div}\big(\mu_t \, \mathbb{E}_\xi[\mathcal{J}_\gamma^{\mu_t}(g_{\mu_t} + \xi)]\big) = 0.$$

The covariance inside each LMO is that of the entire random force field; shared minibatch noise is not independent across particles. There is no automatic diffusion term in this scaling. Its gradient-flow structure is open in general.

A different toy model adds independent Brownian noise *after* selecting the deterministic velocity. Its equation is $\partial_t \mu + \operatorname{div}(\mu \mathcal{J}_\gamma^\mu(g_\mu)) = \varepsilon \Delta \mu$. For $\gamma = \operatorname{tr}$ this is the W_2 flow of $\mathcal{F} + \varepsilon \operatorname{Ent}$, where $\operatorname{Ent}(\mu) = \int \rho \log \rho \, dx$ for $d\mu = \rho \, dx$. For a general spectral gauge, however, $\mathcal{J}_\gamma^\mu(g_\mu + \varepsilon \nabla \log \rho)$ is not in general $\mathcal{J}_\gamma^\mu(g_\mu) - \varepsilon \nabla \log \rho$. Thus this same entropy-penalization interpretation does not carry over.

Block-wise normalization. For parameters $x = (u, v)$, separate normalization of the outer and inner weight blocks corresponds to a block gauge. The choice depends on whether each block uses its own squared-norm LMO or an unscaled polar update. For $S = \left(\begin{smallmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{smallmatrix}\right) \succeq 0$, define

$$\gamma_{\mathrm{blk}}(S) \coloneqq \gamma_1(S_{11}) + \gamma_2(S_{22}),$$

where γ_1 and γ_2 are monotone gauges on the two positive semidefinite block cones. This gauge is convex, positively homogeneous, and positive on every nonzero PSD matrix. It is monotone because $0 \preceq S \preceq T$ implies $S_{ii} \preceq T_{ii}$ for each principal block. Proposition A.1 therefore supplies a norm extension to the whole symmetric space. Furthermore, writing $v = (v_1, v_2)$ gives $\mathcal{N}_\mu(v)^2 = \gamma_1(\int v_1 v_1^\top d\mu) + \gamma_2(\int v_2 v_2^\top d\mu)$; the LMO objective separates into two independent minimizations. For operator block gauges and a gradient matrix partitioned as $G = (G_1, \dots, G_B)$, the resulting update in block b is $-\|G_b\|_* P_b$, where P_b is its compact polar factor. These individual speed factors cannot generally be absorbed into a single time change.

To model raw blockwise polar updates with one common clock, instead use

$$\gamma_{\max}(S) = \max_{1 \le b \le B} \lambda_{\max}(S_{bb}).$$

This is again a positive monotone convex gauge. Its tangent norm is $\mathrm{N}(V) = \max_b \|V_b\|_{\mathrm{op}}$, with dual $\mathrm{N}^*(G) = \sum_b \|G_b\|_*$. The unit-ball formula (1) therefore gives the canonical selection

$$\mathrm{J}_{\max}(G)_b = -\left(\sum_a \|G_a\|_*\right) P_b,$$

where $P_b = 0$ for a zero-gradient block. Thus it follows the raw blockwise polar trajectories up to a common time change away from stationarity. Fixed positive block learning-rate factors η_b are obtained by replacing the maximum gauge by $\max_b \lambda_{\max}(S_{bb})/\eta_b^2$. The static and dynamic theory and conditional gradient-flow theorem apply to both constructions, for any finite number of blocks. Neither construction replaces the joint measure by independent marginals: the static problem still uses a common coupling, and the chosen block layout must match the modeled parameterization.

B.1 Time Scales and Discretization

The choice of gauge fixes a time scale. Replacing γ by $a\gamma$ multiplies W_γ by $\sqrt{a}$ and divides the selected velocity by a. Thus an arbitrary acceleration of a continuous-time curve is not itself an optimization gain. Practical iteration counts additionally depend on smoothness in the chosen geometry and on the discretization.

For a smooth matrix objective on a convex region Ω, a precise Lipschitz-gradient constant is

$$L_\gamma \coloneqq \sup_{X \neq Y \in \Omega} \frac{\mathrm{N}^*(\nabla \mathrm{F}(X) - \nabla \mathrm{F}(Y))}{\mathrm{N}(X - Y)}. \tag{11}$$

If it is finite, integration along a segment gives the descent lemma $\mathrm{F}(X+V) \le \mathrm{F}(X) + \langle \nabla \mathrm{F}(X), V\rangle_F + \frac{1}{2} L_\gamma \mathrm{N}(V)^2$. For $V = \tau \mathrm{J}_\gamma(\nabla \mathrm{F}(X))$, this yields

$$\mathrm{F}(X+V) \le \mathrm{F}(X) - \tau\left(1 - \frac{L_\gamma \tau}{2}\right) \mathrm{N}^*(\nabla \mathrm{F}(X))^2,$$

provided the segment remains in Ω. Hence $0 < \tau < 2/L_\gamma$ is a sufficient descent condition, not a universal numerical accuracy or optimal stability bound. A one-sided quadratic upper model is enough for this conclusion; it need not imply the two-sided gradient bound (11).

For the unscaled trace and operator gauges on $\mathbb{R}^{n\times d}$, let $r_n = \min(n, d)$. Since $\|V\|_{\mathrm{op}} \le \|V\|_F \le \sqrt{r_n}\|V\|_{\mathrm{op}}$ and $\|G\|_F \le \|G\|_* \le \sqrt{r_n}\|G\|_F$, one obtains on the same region

$$L_1 \le L_\infty \le r_n L_1.$$

The operator smoothness constant cannot be smaller under this common normalization. This worst-case comparison does not order convergence in iterations: the operator update also has the larger dissipation $\|G\|_*^2$, and the directions and states visited differ. For fixed particle dimension d, the factor remains bounded by d as $n \to \infty$; there is no dimension-independent bound when d also grows. Existence of a finite smoothness constant, for either geometry, is a separate property of the objective. Neither ReLU losses nor arbitrary smooth neural-network losses have a global such bound automatically.

A transport analogue, avoiding comparison of covectors at different measures, is the upper estimate

$$\mathcal{F}(\nu) \le \mathcal{F}(\mu) + \int g_\mu(x)\cdot(y-x)\, d\pi + \frac{L_\gamma}{2}\gamma\left(\int (y-x)(y-x)^\top\, d\pi\right)$$

for the couplings and states under consideration. Applied to $\nu = (\mathrm{Id} + \tau v)_\sharp \mu$ with $v \in \mathcal{J}_\gamma^\mu(g_\mu)$, it gives the same descent bound with $\mathcal{N}_\mu^*(g_\mu)^2$ in place of the matrix dual norm. Establishing this estimate along a neural-network flow, and controlling discretization or particle approximation errors, are beyond the present analysis. In particular, a Lipschitz bound on the energy differential must not be confused with a Lipschitz bound on the Muon velocity map, which can be discontinuous at rank changes.

C Monge Maps and Conditional Brenier Theorem

This appendix collects the map-based statements associated with the static formulation of Section 3.

The next definition records the Monge restriction, which is useful for comparison but is not the correct notion in general.

Definition C.1 (Monge restriction)**.** *Define*

$$\mathsf{W}_\gamma^{\mathrm{M}}(\mu,\nu)^2 \coloneqq \inf_{T_\#\mu=\nu} \gamma\left(\int_{\mathbb{R}^d} (T(x)-x)(T(x)-x)^\top\, d\mu(x)\right),$$

with value $+\infty$ if no transport map exists.

The next proposition simply records that the map-based problem is a restriction of the coupling-based one.

Proposition C.2 (Monge is a restriction)**.** *For every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathsf{W}_\gamma(\mu,\nu) \le \mathsf{W}_\gamma^{\mathrm{M}}(\mu,\nu).$$

Proof. Every transport map T induces the coupling $(\mathrm{Id}, T)_\#\mu$, and the static cost evaluated on this coupling is exactly the Monge cost associated with T. □

The following remark shows on a two-Dirac example that the relaxation can be genuinely strict.

Remark C.3 (Strictness of the Monge restriction)**.** *Consider*

$$\mu = \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)}), \qquad \nu = \frac{1}{2}(\delta_{(0,-1)} + \delta_{(0,1)}).$$

Any transport map from μ to ν is a bijection between the two atoms. The two possible displacement covariance matrices are

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

so for the operator norm and Frobenius norm one gets

$$\mathsf{W}_\gamma^{\mathrm{M}}(\mu,\nu)^2 = 2.$$

By contrast, the split coupling assigning mass $1/4$ to each source-target pair has displacement covariance equal to the identity matrix, hence

$$\mathsf{W}_\gamma(\mu,\nu)^2 = 1 \quad \text{for } \gamma = \|\cdot\|_{S_\infty}, \qquad \mathsf{W}_\gamma(\mu,\nu)^2 = \sqrt{2} \quad \text{for } \gamma = \|\cdot\|_{S_2}.$$

Therefore the inequality in Proposition C.2 is strict already for two-point measures.

The next remark explains that, for the restriction of an arbitrary matrix norm to $\mathbb{S}_+^d$, the static cost need not be a bona fide metric.

Remark C.4 (Triangle inequality may fail for non-monotone norms)**.** *For Dirac masses, the unique coupling gives*

$$\mathsf{W}_\gamma(\delta_x,\delta_y)^2 = \gamma\big((y-x)(y-x)^\top\big).$$

Hence if W_γ were always a distance, then the pointwise cost

$$d_\gamma(x,y) \coloneqq \sqrt{\gamma((y-x)(y-x)^\top)}$$

would in particular have to define a metric on $\mathbb{R}^d$.

This fails for general norms on $\mathbb{S}_+^d$. Fix $d \geq 2$ and a parameter $M > 2$, and define

$$\gamma_{\mathrm{ns}}(S) \coloneqq \operatorname{tr}(S) + M \sum_{1\leq i<j\leq d} |S_{ij}|, \qquad S \succeq 0.$$

This is the restriction to $\mathbb{S}_+^d$ of the weighted entrywise norm $A \mapsto \sum_i |A_{ii}| + M\sum_{i<j}|A_{ij}|$ on $\mathbb{S}^d$. It is neither spectral nor monotone. Indeed, in dimension two,

$$S = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \preceq T = 2I, \qquad \gamma_{\mathrm{ns}}(S) = 2 + M > 4 = \gamma_{\mathrm{ns}}(T).$$

For the displacement

$$\Delta = y - x = (\Delta_1, \ldots, \Delta_d)$$

one has

$$d_{\gamma_{\mathrm{ns}}}(x,y)^2 = \|\Delta\|_2^2 + M \sum_{1\leq i<j\leq d} |\Delta_i \Delta_j|.$$

Taking

$$x_0 = 0, \qquad x_1 = e_1, \qquad x_2 = e_1 + e_2$$

gives

$$d_{\gamma_{\mathrm{ns}}}(x_0,x_1) = 1, \qquad d_{\gamma_{\mathrm{ns}}}(x_1,x_2) = 1, \qquad d_{\gamma_{\mathrm{ns}}}(x_0,x_2) = \sqrt{2+M} > 2,$$

so the triangle inequality fails.

By contrast, if γ is monotone on $\mathbb{S}_+^d$, then Proposition A.2 allows us to choose $\mathcal{K}_\gamma \subset \mathbb{S}_+^d$, and therefore

$$d_\gamma(x,y) = \sup_{Q\in\mathcal{K}_\gamma} \left\| Q^{1/2}(x-y) \right\|_2,$$

which is a supremum of seminorms and, by positivity of γ, a norm of $x - y$. For general measures, however, this pointwise argument does not control the cross terms created by gluing couplings. The full metric property of W_γ follows either from the robust static argument in Proposition A.8 or from the dynamic formulation in Corollary 4.3.

The following proposition explains when an optimal coupling is induced by a Monge map of Brenier type.

Proposition C.5 (Conditional Brenier theorem)**.** *Assume μ is absolutely continuous. If a dual maximizing matrix $Q_* \in \mathcal{K}_\gamma$ in Theorem 3.4 is positive definite, then every optimal coupling for $\mathsf{W}_\gamma(\mu,\nu)$ is induced by a map. More precisely, there exists a convex function u such that*

$$T_*(x) = Q_*^{-1/2}\nabla u(Q_*^{1/2}x)$$

is optimal.

Without monotonicity, a polar certificate need not be positive semidefinite and may even be indefinite. Under the standing monotonicity assumption, however, Proposition A.2 allows one to choose the representing set $\mathcal{K}_\gamma$ inside $\mathbb{S}^d_+$; the active matrix is then automatically positive semidefinite, and the remaining nondegeneracy assumption is simply that it be invertible. For the trace gauge $p = 1$, one may use the reduced set $\mathcal{K}_\gamma = \{I\}$, so the condition is automatic. For $1 < p < \infty$, the canonical active matrix is $\Sigma_{\pi^\star}^{p-1}/\|\Sigma_{\pi^\star}\|_{S_p}^{p-1}$ and is therefore positive definite whenever the optimal displacement covariance is. The endpoint $p = \infty$ is sharply different: in that case $\gamma(\Sigma) = \lambda_{\max}(\Sigma)$ and Q_* is supported on the top eigenspace of $\Sigma_{\pi^\star}$. When the leading eigenvalue is simple, Q_* is rank one. The associated degenerate quadratic cost does not control directions orthogonal to the active eigenspace, so Brenier's uniqueness argument no longer applies. Monge optimizers may still exist, but the dual certificate alone neither selects one nor implies that every optimizer is map-induced.

Proof. Fix such a positive definite dual maximizer Q_* and let π_* be any primal optimizer. With $F(\pi, Q) = \int (y-x)^\top Q(y-x)\, d\pi$, minimax equality gives

$$F(\pi_*, Q_*) \le \max_{Q\in\mathcal{K}_\gamma} F(\pi_*, Q) = \mathsf{W}_\gamma(\mu,\nu)^2 = \min_{\pi\in\Pi(\mu,\nu)} F(\pi, Q_*) \le F(\pi_*, Q_*).$$

Thus π_* is optimal for the quadratic cost induced by Q_*. To identify this optimizer, let

$$\widetilde{\pi}_* = (Q_*^{1/2}, Q_*^{1/2})_\# \pi_*,$$

and define

$$\widetilde{\mu} = (Q_*^{1/2})_\#\mu, \qquad \widetilde{\nu} = (Q_*^{1/2})_\#\nu.$$

Then

$$\int (y-x)^\top Q_*(y-x)\, d\pi_*(x,y) = \int |\tilde{y}-\tilde{x}|^2 \, d\widetilde{\pi}_*(\tilde{x},\tilde{y}).$$

Because μ is absolutely continuous and $Q_*^{1/2}$ is invertible, $\widetilde{\mu}$ is also absolutely continuous. Brenier's theorem therefore yields a convex potential u such that the unique optimal coupling between $\widetilde{\mu}$ and $\widetilde{\nu}$ is induced by the map

$$\widetilde{T}(\tilde{x}) = \nabla u(\tilde{x}).$$

Pulling this map back to the original variables gives

$$T_*(x) = Q_*^{-1/2}\widetilde{T}(Q_*^{1/2}x) = Q_*^{-1/2}\nabla u(Q_*^{1/2}x).$$

Since π_* is Q_*-optimal and the latter coupling is unique, π_* must be induced by T_*. The argument applies to every primal optimizer. □

D Gaussian Reductions

This appendix gathers the Gaussian reductions used in the static geometry and in the linear-network flow. Gaussian marginals compress the transport problem to a finite-dimensional optimization over admissible covariance blocks. The next proposition shows that, for Gaussian marginals, the infinite-dimensional transport problem collapses to an optimization over the cross-covariance matrix.

Proposition D.1 (Gaussian reduction)**.** *Let $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$. Then*

$$\mathsf{W}_\gamma(\mu,\nu)^2 = \inf_K \gamma\big((m_1-m_0)(m_1-m_0)^\top + \Sigma_0 + \Sigma_1 - K - K^\top\big),$$

where the infimum runs over matrices K such that

$$\begin{pmatrix} \Sigma_0 & K \\ K^\top & \Sigma_1 \end{pmatrix} \succeq 0.$$

In particular, for centered Gaussians the covariance cost

$$\mathsf{B}_\gamma(\Sigma_0,\Sigma_1)^2 \coloneqq \inf_{K:\left[\begin{smallmatrix}\Sigma_0 & K\\ K^\top & \Sigma_1\end{smallmatrix}\right]\succeq 0} \gamma(\Sigma_0+\Sigma_1-K-K^\top)$$

is well defined on the cone of covariance matrices. In the monotone regime of Section 4, it is the restriction of the metric W_γ to centered Gaussian laws and therefore defines a metric on covariance matrices.

Proof. Let (X,Y) be any coupling of the two Gaussian marginals and set $K=\mathbb{E}[(X-m_0)(Y-m_1)^\top]$. Its block covariance is

$$\begin{pmatrix}\Sigma_0 & K\\ K^\top & \Sigma_1\end{pmatrix}.$$

It must be positive semidefinite. Conversely, every feasible block matrix is the covariance of a jointly Gaussian vector with marginals μ and ν. For any coupling with cross-covariance K, the displacement covariance is

$$(m_1-m_0)(m_1-m_0)^\top+\Sigma_0+\Sigma_1-K-K^\top.$$

The cost therefore depends only on K, and every feasible K is realized by a Gaussian coupling. Optimizing over all couplings is thus equivalent to the stated finite-dimensional problem. □

The next corollary shows that commuting covariances lead to a closed form for the Schatten family, exactly as in the classical Bures case.

Corollary D.2 (Commuting covariances)**.** *If Σ_0 and Σ_1 commute, then for Schatten norms γ_p, $1\le p\le\infty$, one has*

$$\mathsf{B}_{\gamma_p}(\Sigma_0,\Sigma_1)^2=\gamma_p\Big((\Sigma_0^{1/2}-\Sigma_1^{1/2})^2\Big).$$

Equivalently, if a_i and b_i are the diagonal entries of Σ_0 and Σ_1 in a common eigenbasis, then, for $p<\infty$,

$$\mathsf{B}_{\gamma_p}(\Sigma_0,\Sigma_1)^2=\left(\sum_{i=1}^d\left|\sqrt{a_i}-\sqrt{b_i}\right|^{2p}\right)^{1/p}.$$

For $p=\infty$, the right-hand side is $\max_i|\sqrt{a_i}-\sqrt{b_i}|^2$. When $p=1$, this is the usual Bures–Wasserstein formula.

Proof. Simultaneously diagonalize $\Sigma_0=\operatorname{diag}(a_i)$ and $\Sigma_1=\operatorname{diag}(b_i)$. For any feasible K, each 2×2 principal minor involving the coordinates X_i and Y_i gives $|K_{ii}|\le\sqrt{a_ib_i}$. Hence the displacement covariance $D=\Sigma_0+\Sigma_1-K-K^\top\succeq0$ satisfies

$$D_{ii}=a_i+b_i-2K_{ii}\ge(\sqrt{a_i}-\sqrt{b_i})^2.$$

Pinching onto the common eigenbasis is contractive for every unitarily invariant norm, so

$$\|D\|_{S_p}\ge\|\operatorname{diag}(D_{ii})\|_{S_p}\ge\Big\|\operatorname{diag}\big((\sqrt{a_i}-\sqrt{b_i})^2\big)\Big\|_{S_p}.$$

The diagonal choice $K_0=\operatorname{diag}(\sqrt{a_ib_i})$ is feasible and attains both inequalities, proving the claim. □

The following remark clarifies the scope of Corollary D.2.

Remark D.3. *The commuting formula is stated for Schatten norms because their value depends only on eigenvalues. A general gauge may retain basis-dependent information. The restriction $p\ge1$ is also essential: at $p=1/2$, take $\Sigma_0=\operatorname{diag}(4,1)$ and $\Sigma_1=\operatorname{diag}(1,4)$. In the basis $(e_1+e_2)/\sqrt{2},(e_1-e_2)/\sqrt{2}$, the map $Y_1=X_1$, $Y_2=X_2-\frac{6}{5}X_1$ has the required target covariance and displacement covariance $\operatorname{diag}(0,18/5)$. Its cost is $18/5$, strictly below the value $\gamma_{1/2}(I)=4$ predicted by the commuting expression. Thus the root-norm extension below does not extend this particular formula.*

E Dynamic Formulation Details

This appendix gathers the convex momentum formulation and the proofs deferred from Section 4.

E.1 Convex momentum reformulation

Passing to momenta linearizes the constraint. If $\mu = \rho\lambda$ and $m = w\lambda$ for some reference measure λ, define

$$\mathcal{A}_\gamma(\mu, m) \coloneqq \sup_{Q \in \mathcal{K}_\gamma} \int_{\mathbb{R}^d} \frac{w(x)^\top Q\, w(x)}{\rho(x)}\, d\lambda(x),$$

with the usual perspective convention on $\{\rho = 0\}$.

The following proposition shows that the momentum formulation is convex and exactly matches the velocity action.

Proposition E.1 (Convex momentum action)**.** *The action $\mathcal{A}_\gamma(\mu, m)$ is intrinsic, convex, and weak-* lower semicontinuous. If $m = \mu v$, then*

$$\mathcal{A}_\gamma(\mu, m) = \gamma\left(\int v(x)v(x)^\top\, d\mu(x)\right).$$

Hence

$$\mathsf{W}_\gamma^{\mathrm{BB}}(\mu_0, \mu_1)^2 = \inf_{(\mu_t, m_t)} \int_0^1 \mathcal{A}_\gamma(\mu_t, m_t)\, dt$$

under the linear constraint

$$\partial_t \mu_t + \operatorname{div}(m_t) = 0.$$

Proof. Let λ be any positive measure dominating both μ and m, and write $\mu = \rho\lambda$ and $m = w\lambda$, with $\rho \geq 0$ and $w : \mathbb{R}^d \to \mathbb{R}^d$. For $Q \succeq 0$, define the perspective quadratic integrand

$$\phi_Q(\rho, w) \coloneqq \begin{cases} \dfrac{w^\top Q w}{\rho}, & \rho > 0, \\ 0, & \rho = 0 \text{ and } Q^{1/2} w = 0, \\ +\infty, & \rho = 0 \text{ and } Q^{1/2} w \neq 0. \end{cases}$$

Then ϕ_Q is convex and lower semicontinuous on $\mathbb{R}_+ \times \mathbb{R}^d$. Indeed, for $\rho > 0$ it is the perspective of the convex quadratic form $w \mapsto w^\top Q w$, and the above convention is its closed extension at $\rho = 0$. Notice that kernel directions have zero recession cost when Q is singular.

The momentum action may be written as

$$\mathcal{A}_\gamma(\mu, m) = \sup_{Q \in \mathcal{K}_\gamma} \int \phi_Q(\rho(x), w(x))\, d\lambda(x).$$

This formula is intrinsic, i.e., independent of the dominating measure λ. To see this, choose another dominating measure λ' and write $\mu = \rho'\lambda'$ and $m = w'\lambda'$. Passing to a common dominating measure, for instance $\eta = \lambda + \lambda'$, and using the one-homogeneity of ϕ_Q, namely $\phi_Q(a\rho, aw) = a\phi_Q(\rho, w)$ for $a \geq 0$, gives the same value of $\int \phi_Q\, d\lambda$ and $\int \phi_Q\, d\lambda'$. Thus $\mathcal{A}_\gamma(\mu, m)$ depends only on the pair (μ, m).

We next prove the absolutely continuous representation. If $m = \mu v$, then one may take $\lambda = \mu$, so that $\rho = 1$ and $w = v$. Hence, for every $Q \in \mathcal{K}_\gamma$,

$$\int \phi_Q(1, v(x))\, d\mu(x) = \int v(x)^\top Q v(x)\, d\mu(x) = \operatorname{tr}\left(Q \int v(x)v(x)^\top\, d\mu(x)\right).$$

Taking the supremum over $Q \in \mathcal{K}_\gamma$ and using the support representation of γ yields

$$\mathcal{A}_\gamma(\mu, \mu v) = \sup_{Q \in \mathcal{K}_\gamma} \operatorname{tr}\left(Q \int vv^\top\, d\mu\right) = \gamma\left(\int vv^\top\, d\mu\right).$$

Conversely, strict positivity of γ implies that $\mathcal{K}_\gamma$ contains a positive definite matrix. Indeed, for every unit vector a there is a $Q_a \in \mathcal{K}_\gamma$ with $a^\top Q_a a > 0$; compactness of the sphere gives finitely many such matrices whose average is positive definite and belongs to $\mathcal{K}_\gamma$ by convexity. Fix one such matrix Q_0. If $\mathcal{A}_\gamma(\mu, m) < +\infty$, then $\int \phi_{Q_0}(\rho, w)\, d\lambda$ is finite. Since $Q_0 \succ 0$, the singular case $\rho = 0$ and $w \neq 0$ cannot occur on a set of positive λ-measure. Thus m is absolutely continuous with respect to μ, and $m = \mu v$ with $v = w/\rho$ on $\{\rho > 0\}$.

Convexity follows from the same perspective formulation. For fixed $Q \succeq 0$, the map $(\mu, m) \mapsto \int \phi_Q(\rho, w)\, d\lambda$ is convex, because ϕ_Q is convex and one-homogeneous. Taking the supremum over $Q \in \mathcal{K}_\gamma$ preserves convexity, hence $(\mu, m) \mapsto \mathcal{A}_\gamma(\mu, m)$ is convex.

Finally, we prove weak-* lower semicontinuity. For each fixed $Q \succeq 0$, the functional

$$(\mu, m) \longmapsto \int \phi_Q\left(\frac{d\mu}{d\lambda}, \frac{dm}{d\lambda}\right) d\lambda$$

is lower semicontinuous under weak-* convergence of measures, by the standard lower semicontinuity theorem for convex, lower semicontinuous, one-homogeneous integral functionals. Since $\mathcal{A}_\gamma$ is the supremum over $Q \in \mathcal{K}_\gamma$ of these lower semicontinuous functionals, $\mathcal{A}_\gamma$ is also weak-* lower semicontinuous.

The dynamic formula now follows by rewriting the velocity formulation in momentum variables. If (μ_t, v_t) satisfies $\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0$ and $m_t = \mu_t v_t$, then the constraint becomes the linear continuity equation $\partial_t \mu_t + \operatorname{div}(m_t) = 0$, and the previous identity gives

$$\mathcal{N}_{\mu_t}(v_t)^2 = \gamma\left(\int v_t v_t^\top \, d\mu_t\right) = \mathcal{A}_\gamma(\mu_t, m_t).$$

Therefore the dynamic action $\int_0^1 \mathcal{N}_{\mu_t}(v_t)^2\, dt$ is exactly $\int_0^1 \mathcal{A}_\gamma(\mu_t, m_t)\, dt$ in momentum variables. Taking the infimum over all admissible curves gives

$$\mathsf{W}_\gamma^{\mathrm{BB}}(\mu_0, \mu_1)^2 = \inf_{(\mu_t, m_t)} \int_0^1 \mathcal{A}_\gamma(\mu_t, m_t)\, dt,$$

under the linear constraint $\partial_t \mu_t + \operatorname{div}(m_t) = 0$. □

E.2 Proof of Theorem 4.2 (static/dynamic equivalence)

Proof. Let $\pi \in \Pi(\mu_0, \mu_1)$ and consider the displacement interpolation

$$z_t(x, y) = (1 - t)x + ty, \qquad \mu_t = (z_t)_\# \pi.$$

Its Eulerian velocity is the conditional mean $v_t(z) = \mathbb{E}_\pi[y - x \mid z_t = z]$. The pair (μ_t, v_t) solves the continuity equation, and conditional Jensen's inequality gives

$$\int v_t v_t^\top \, d\mu_t \preceq \int (y - x)(y - x)^\top \, d\pi$$

for a.e. t. By monotonicity of γ, its dynamic action is no larger than the static cost of π. Taking the infimum over π gives

$$\mathsf{W}_\gamma^{\mathrm{BB}}(\mu_0, \mu_1)^2 \leq \mathsf{W}_\gamma(\mu_0, \mu_1)^2.$$

Conversely, take any admissible dynamic plan (μ_t, v_t). By the superposition principle, there exists a probability measure η on absolutely continuous paths ω such that

$$\mu_t = (e_t)_\# \eta, \qquad \dot{\omega}_t = v_t(\omega_t)$$

for η-a.e. path and a.e. t. Let $\pi = (e_0, e_1)_\# \eta$. For every $Q \in \mathcal{K}_\gamma$,

$$\int_0^1 \int v_t^\top Q v_t \, d\mu_t \, dt = \int \int_0^1 \dot{\omega}_t^\top Q \dot{\omega}_t \, dt \, d\eta(\omega).$$

This is the key point where monotonicity is used: by Proposition A.2 we have chosen $\mathcal{K}_\gamma \subset \mathbb{S}_+^d$, so every test matrix satisfies $Q \succeq 0$. Without this reduction one would have to work with possibly

indefinite matrices, and the quadratic Jensen inequality below would no longer apply. Applying Jensen to $t \mapsto Q^{1/2}\dot{\omega}_t$ gives

$$\int_0^1 \dot{\omega}_t^\top Q \dot{\omega}_t \, dt \geq (\omega_1 - \omega_0)^\top Q (\omega_1 - \omega_0).$$

Hence

$$\int_0^1 \gamma\Big(\int v_t v_t^\top \, d\mu_t\Big) \, dt \geq \int (y-x)^\top Q (y-x) \, d\pi(x,y)$$

for every $Q \in \mathcal{K}_\gamma$. Taking the supremum over Q yields

$$\int_0^1 \gamma\Big(\int v_t v_t^\top \, d\mu_t\Big) \, dt \geq \gamma\Big(\int (y-x)(y-x)^\top \, d\pi\Big).$$

Finally, take the infimum over admissible dynamic plans. □

E.3 Proof of Corollary 4.4 (Spectral Geodesics)

Proof. The upper bound follows by restricting the same displacement interpolation to the time interval $[s,t]$. The coupling $((1-s)x + sy, (1-t)x + ty)_\# \pi_*$ has displacement $(t-s)(y-x)$, so the static definition gives

$$\mathsf{W}_\gamma(\mu_s,\mu_t)^2 \leq (t-s)^2 \gamma\Big(\int (y-x)(y-x)^\top \, d\pi_*(x,y)\Big) = (t-s)^2 \mathsf{W}_\gamma(\mu_0,\mu_1)^2.$$

Applying the same argument to the segments $[0,s]$ and $[t,1]$ yields

$$\mathsf{W}_\gamma(\mu_0,\mu_s) \leq s\mathsf{W}_\gamma(\mu_0,\mu_1), \qquad \mathsf{W}_\gamma(\mu_t,\mu_1) \leq (1-t)\mathsf{W}_\gamma(\mu_0,\mu_1).$$

By the triangle inequality from Corollary 4.3,

$$\mathsf{W}_\gamma(\mu_0,\mu_1) \leq \mathsf{W}_\gamma(\mu_0,\mu_s) + \mathsf{W}_\gamma(\mu_s,\mu_t) + \mathsf{W}_\gamma(\mu_t,\mu_1) \leq \mathsf{W}_\gamma(\mu_s,\mu_t) + (1-(t-s))\mathsf{W}_\gamma(\mu_0,\mu_1).$$

Rearranging gives

$$\mathsf{W}_\gamma(\mu_s,\mu_t) \geq (t-s)\mathsf{W}_\gamma(\mu_0,\mu_1),$$

hence equality holds. □

E.4 Root-norm gauges beyond convexity

The main text assumes that γ is convex, monotone, and positively one-homogeneous. Convexity is essential for the support-function representation $\gamma(S) = \max_Q \operatorname{tr}(QS)$ and hence for the cost-robust formulation. Some parts of the theory, however, only use that the square-root action is a norm.

Definition E.2 (Root-norm gauge)**.** *A positive one-homogeneous function $\gamma : \mathbb{S}_+^d \to \mathbb{R}_+$ is a root-norm gauge if, for every Hilbert space $\mathcal{H}$, the map*

$$T : \mathbb{R}^d \to \mathcal{H} \quad \longmapsto \quad \mathfrak{N}_\gamma(T) \coloneqq \gamma(T^*T)^{1/2}$$

is a norm on finite-rank operators.

For a vector field $v \in L^2(\mu;\mathbb{R}^d)$, this gives exactly the tangent norm used in the paper:

$$\mathcal{N}_\mu(v) = \mathfrak{N}_\gamma(T_v), \qquad (T_v a)(x) = a \cdot v(x),$$

since $T_v^* T_v = \int v v^\top \, d\mu$.

Proposition E.3 (Characterization)**.** *A positive one-homogeneous γ is a root-norm gauge if and only if $V \mapsto \gamma(V^\top V)^{1/2}$ is a norm on $\mathbb{R}^{d\times d}$. Equivalently, $\gamma(S) = \mathfrak{N}(S^{1/2})^2$ for a left-orthogonally invariant norm $\mathfrak{N}$ on square matrices. In the spectral case, this is equivalent to*

$$\gamma(S) = \varphi\Big(\lambda(S)^{1/2}\Big)^2$$

for some symmetric gauge norm φ on $\mathbb{R}^d$. Every root-norm gauge is monotone on $\mathbb{S}_+^d$, but it need not be convex.

Proof. Necessity follows by choosing the target $\mathbb{R}^d$. Conversely, suppose the square-matrix map $\mathfrak{N}$ is a norm. It is left-orthogonally invariant. Every square contraction is a convex combination of orthogonal matrices, as follows by writing its singular values as convex combinations of ± 1. Hence $\mathfrak{N}(CV) \le \mathfrak{N}(V)$ for $\|C\|_{\mathrm{op}} \le 1$. If $0 \preceq S \preceq T$, then $C = S^{1/2}T^{\dagger/2}$ is a contraction and $S^{1/2} = CT^{1/2}$, proving $\gamma(S) \le \gamma(T)$.

For operators $T_1, T_2 : \mathbb{R}^d \to \mathcal{H}$, choose a contraction $R : \mathcal{H} \to \mathbb{R}^d$ that is isometric on the range of $T_1 + T_2$, whose dimension is at most d. Then

$$\mathfrak{N}_\gamma(T_1 + T_2) = \mathfrak{N}(RT_1 + RT_2) \le \mathfrak{N}(RT_1) + \mathfrak{N}(RT_2) \le \mathfrak{N}_\gamma(T_1) + \mathfrak{N}_\gamma(T_2).$$

The last inequality uses $(RT_i)^* RT_i \preceq T_i^* T_i$ and the monotonicity just proved. Positivity and homogeneity follow from the square-matrix norm. This proves the norm property for every target. Polar decomposition gives the equivalent description by left-invariant norms; the spectral case is the usual symmetric-gauge description of unitarily invariant norms. Finally, for any target contraction B, the inequality $T^*B^*BT \preceq T^*T$ proves $\mathfrak{N}_\gamma(BT) \le \mathfrak{N}_\gamma(T)$. □

Corollary E.4 (Schatten threshold)**.** *For $d \ge 2$ and $p \in (0, \infty]$, the function $\gamma_p(S) = \|S\|_{S_p}$ is a root-norm gauge if and only if $p \ge 1/2$. It is convex if and only if $p \ge 1$. In dimension one, all these functions equal $\gamma_p(s) = s$, so both properties hold for every $p > 0$.*

Proof. One has

$$\gamma_p(V^\top V)^{1/2} = \|V\|_{S_{2p}}.$$

When $d \ge 2$, this is a norm exactly when $2p \ge 1$. Convexity of the Schatten p-functional on the positive cone holds exactly for $p \ge 1$. When $d = 1$, there is only one singular value, so neither threshold applies. □

For nonconvex root-norm gauges, the static Kantorovich formula still makes sense,

$$\mathsf{W}_\gamma(\mu, \nu)^2 = \inf_{\pi \in \Pi(\mu,\nu)} \gamma\left(\int (y - x)(y - x)^\top \, d\pi(x, y) \right),$$

but it is no longer a convex optimization problem in the coupling. Moreover, there is no exact cost-robust formula of the form $\gamma(S) = \max_{Q \in \mathcal{K}_\gamma} \mathrm{tr}(QS)$, since a maximum of linear forms is convex. Thus the direct robust-cost proof of the triangle inequality from Appendix A.3 is no longer available.

Theorem E.5 (Benamou–Brenier formula for root-norm gauges)**.** *Let γ be a root-norm gauge. Then, for every $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathsf{W}_\gamma(\mu_0, \mu_1)^2 = \inf_{\partial_t \mu_t + \mathrm{div}(\mu_t v_t) = 0} \int_0^1 \gamma\left(\int v_t v_t^\top \, d\mu_t \right) dt,$$

where the infimum is over narrowly continuous curves with endpoints μ_0, μ_1 and square-integrable velocity fields.

Proof. Let (μ_t, v_t) be admissible. By the superposition principle, there exists a probability measure η on absolutely continuous paths ω such that $\mu_t = (e_t)_\# \eta$ and $\dot\omega_t = v_t(\omega_t)$ for a.e. t. Set $\pi = (e_0, e_1)_\# \eta$ and $\Delta(\omega) = \omega_1 - \omega_0$. Then

$$\Delta = \int_0^1 v_t(\omega_t) \, dt.$$

At the operator level, $T_\Delta = \int_0^1 T_{v_t \circ e_t} \, dt$ as maps $\mathbb{R}^d \to L^2(\eta)$. Hence, by the triangle inequality for $\mathfrak{N}_\gamma$ and Cauchy–Schwarz,

$$\begin{aligned} \gamma\left(\int \Delta\Delta^\top \, d\eta \right)^{1/2} &\le \int_0^1 \gamma\left(\int v_t v_t^\top \, d\mu_t \right)^{1/2} dt \\ &\le \left(\int_0^1 \gamma\left(\int v_t v_t^\top \, d\mu_t \right) dt \right)^{1/2}. \end{aligned}$$

Since Δ depends only on the endpoints, this gives $\mathsf{W}_\gamma(\mu_0,\mu_1)^2$ bounded above by the dynamic action.

Conversely, fix $\pi \in \Pi(\mu_0,\mu_1)$ and set $z_t = (1-t)x + ty$, $\mu_t = (z_t)_\#\pi$. Let $v_t(z) = \mathbb{E}_\pi[y - x \mid z_t = z]$. Then (μ_t, v_t) solves the continuity equation. The conditional-expectation operator $C_t : L^2(\pi) \to L^2(\mu_t)$ is a contraction and satisfies $T_{v_t} = C_t T_{y-x}$. Since $\mathfrak{N}_\gamma$ is contractive under target contractions,

$$\mathcal{N}_{\mu_t}(v_t) \le \gamma\left(\int (y-x)(y-x)^\top \, d\pi\right)^{1/2}.$$

Integrating over t gives a dynamic action no larger than the static cost of π. Taking the infimum over π proves the reverse inequality. □

Corollary E.6 (Metric property for root-norm gauges)**.** *If γ is a root-norm gauge, then W_γ is a distance on $\mathcal{P}_2(\mathbb{R}^d)$ and induces the same topology as W_2.*

Proof. Nonnegativity and symmetry are immediate. Apply finite-dimensional norm equivalence on $\mathbb{R}^{d\times d}$ to the square root $S^{1/2}$. This gives constants $0 < c \le C < \infty$, independent of the Hilbert target, such that

$$c\,\operatorname{tr}(S) \le \gamma(S) \le C\,\operatorname{tr}(S), \qquad S \succeq 0.$$

Thus $\sqrt{c}\,\mathsf{W}_2 \le \mathsf{W}_\gamma \le \sqrt{C}\,\mathsf{W}_2$, which gives separation and topological equivalence.

For the triangle inequality, one may either use Theorem E.5 and concatenate dynamic curves, or argue directly. Given couplings π_{01} and π_{12}, glue them into a measure on triples (x_0, x_1, x_2). With $\Delta_{ij} = x_j - x_i$, one has $\Delta_{02} = \Delta_{01} + \Delta_{12}$, and the norm property gives

$$\mathfrak{N}_\gamma(T_{\Delta_{02}}) \le \mathfrak{N}_\gamma(T_{\Delta_{01}}) + \mathfrak{N}_\gamma(T_{\Delta_{12}}).$$

Taking infima over the two input couplings proves $\mathsf{W}_\gamma(\mu_0,\mu_2) \le \mathsf{W}_\gamma(\mu_0,\mu_1) + \mathsf{W}_\gamma(\mu_1,\mu_2)$. □

E.5 Proof of Theorem 5.1 (Spectral Gradient Flow)

Proof. Write $N = \mathcal{N}_\mu$ and $N^* = \mathcal{N}_\mu^*$ at a fixed state on the curve. First, norm duality gives, for every $v \in \mathcal{J}_\gamma^\mu(g)$,

$$N(v) = N^*(g), \qquad \int g \cdot v \, d\mu = -N^*(g)^2. \tag{12}$$

Indeed, optimizing the radius along a unit-ball minimizing direction reduces the objective to $-rN^*(g) + r^2/2$.

We next prove the slope identity without assuming that μ has a density. Let π be an optimal spectral coupling from μ to ν, and let $w(x) = \int (y-x)\, d\pi(y \mid x)$ be its conditional mean displacement. Conditional Jensen and monotonicity give $N(w) \le \mathsf{W}_\gamma(\mu,\nu)$. Also, $\int |y-x|^2 d\pi \le c_\gamma^{-1}\mathsf{W}_\gamma(\mu,\nu)^2$ by Proposition 3.2. The transport expansion therefore implies

$$\mathcal{F}(\mu) - \mathcal{F}(\nu) \le N^*(g)\mathsf{W}_\gamma(\mu,\nu) + Cc_\gamma^{-1}\mathsf{W}_\gamma(\mu,\nu)^2,$$

and hence $|\partial\mathcal{F}|_\gamma(\mu) \le N^*(g)$. Conversely, for $w \in L^2(\mu;\mathbb{R}^d)$ and $h > 0$, use the deterministic coupling to $\nu_h = (\mathrm{Id} + hw)_\sharp\mu$:

$$\mathsf{W}_\gamma(\mu,\nu_h) \le hN(w), \qquad \mathcal{F}(\nu_h) - \mathcal{F}(\mu) = h\int g \cdot w \, d\mu + O(h^2\|w\|^2_{L^2(\mu)}).$$

For $\int g \cdot w\, d\mu < 0$, division and passage to the limit give $|\partial\mathcal{F}|_\gamma(\mu) \ge -\int g \cdot w\, d\mu / N(w)$. Taking the supremum proves $|\partial\mathcal{F}|_\gamma(\mu) = N^*(g)$; the case $g = 0$ follows from the upper bound.

The assumed spatial Lipschitz and growth bounds provide a characteristic flow, a uniform second-moment bound, and

$$\frac{X_{t+h} - X_t}{h} \longrightarrow v_t(X_t) \quad \text{in } L^2(\mu_0;\mathbb{R}^d).$$

Apply (5) to the coupling $(X_t, X_{t+h})_\sharp\mu_0$. Its squared displacement is $O(h^2)$, so (12) yields

$$\frac{d}{dt}\mathcal{F}(\mu_t) = \int g_t \cdot v_t \, d\mu_t = -\mathcal{N}_{\mu_t}^*(g_t)^2.$$

The energy is locally Lipschitz: $N^*(g_t) = N(v_t)$ is uniformly bounded by the growth and moment bounds, as is $\|g_t\|_{L^2(\mu_t)}$ by norm equivalence.

The BB formula, rescaled from $[0,1]$ to $[s,t]$, gives $\mathsf{W}_\gamma(\mu_s,\mu_t)^2 \le (t-s)\int_s^t N_{\mu_r}(v_r)^2\,dr$. Thus the curve is absolutely continuous and its metric derivative exists almost everywhere, with $|\mu'|_\gamma(t) \le N_{\mu_t}(v_t)$. At such a time, the definition of descending slope and the energy derivative give $N^*_{\mu_t}(g_t)^2 \le |\partial\mathcal{F}|_\gamma(\mu_t)|\mu'|_\gamma(t)$. Together with the slope identity this proves equality of speed, slope, and dual norm whenever $g_t \neq 0$. If $g_t = 0$, the speed bound already gives zero. The stated energy dissipation identity follows. □

A concrete smooth class. Let $\varphi_k \in C^2(\mathbb{R}^d)$ have bounded Hessian, and let $\mathcal{F}(\mu) = \Phi(m(\mu))$, where $m(\mu) = (\int \varphi_k\,d\mu)_{k=1}^m$ and Φ is C^2 near the moment vectors of the curve. The forces are $g_\mu = \sum_k \partial_k\Phi(m(\mu))\nabla\varphi_k$. On a uniformly second-moment-bounded curve,

$$|m(\nu) - m(\mu)| \le C_1\left[\left(\int |y-x|^2 d\pi\right)^{1/2} + \int |y-x|^2 d\pi\right].$$

Taylor expansion of each feature and then of Φ proves (5), uniformly for small displacements. Potential energies are included by taking a single feature. Smooth symmetric interaction energies $\frac{1}{2}\iint K(x,y)\,d\mu(x)d\mu(y)$ satisfy the same estimate when K has bounded Hessian, by applying its Taylor formula to $\pi \otimes \pi$. These examples concern the differential of the energy; they do not by themselves establish regularity of the selected spectral velocity. At empirical measures, their transport expansion gives the chain rule of Proposition 2.5; their particle lifts are C^1 by ordinary finite-dimensional differentiation.

F Gaussian-Preserving Flows and Linear Networks

This section isolates a tractable regime in which the spectral flow reduces from an infinite-dimensional PDE to finite-dimensional covariance dynamics. The general Gaussian-preservation principle comes first, and the rest of the section specializes it to centered Gaussian training of linear two-layer networks.

F.1 Gaussian-Preserving Gradient Flows

This subsection isolates a simple class of spectral gradient flows that can still be analyzed rather explicitly. In general, understanding the long-time behavior of the flow is hard, but when Gaussian measures are preserved the infinite-dimensional evolution reduces to a finite-dimensional ODE on means and covariances.

The next corollary explains why Gaussian preservation transfers from the classical W_2 flow to the spectral flow. The finite-dimensional Measure-LMO reduction is important here because it also covers singular endpoints such as $p = \infty$.

Corollary F.1 (From W_2 Gaussian preservation to spectral Gaussian preservation)**.** *Fix a class of Gaussian states in the domain of $\mathcal{F}$ and assume that, at each such state, the classical Wasserstein gradient*

$$g_\mu(x) = \nabla_x \frac{\delta\mathcal{F}}{\delta\mu}(x)$$

is affine in x. Then the Measure-LMO admits an affine selection at each state in this class, and the spectral flow using this selection,

$$\partial_t\mu_t + \operatorname{div}\big(\mu_t\mathcal{J}_\gamma^{\mu_t}(g_{\mu_t})\big) = 0$$

admits a closed Gaussian ansatz: every solution of the induced mean and covariance ODE that stays in this class and has locally integrable affine coefficients defines a Gaussian solution of the PDE. This does not assert uniqueness among all PDE solutions.

Proof. Fix a state μ in this class and write

$$g_\mu(x) = b_\mu + B_\mu x$$

for the affine W_2 gradient provided by the assumption. Set $S_g = \mathcal{S}_\mu(g_\mu)$. By Proposition 2.4, there is a matrix A_g such that

$$\mathcal{J}^\mu_\gamma(g_\mu)(x) = A_g^\top S_g^{\dagger/2}(b_\mu + B_\mu x).$$

This is again affine in x (and is zero when $S_g = 0$). Therefore, the spectral continuity equation is of the form

$$v_t(x) = a_t + A_t x$$

with coefficients depending on the current Gaussian state. A continuity equation with affine drift preserves the Gaussian class, and the associated mean m_t and covariance Σ_t solve the closed ODE system

$$\dot{m}_t = a_t + A_t m_t, \qquad \dot{\Sigma}_t = A_t\Sigma_t + \Sigma_t A_t^\top.$$

Conversely, a solution of these moment equations with locally integrable a_t, A_t is transported by their affine characteristic flow, so its Gaussian law solves the PDE. Existence and uniqueness of this state-dependent moment system, particularly at rank changes, are separate questions. □

Two important examples fit into this mechanism. First, relative entropy toward a nondegenerate Gaussian target has an affine Wasserstein force at every nondegenerate Gaussian state; singular Gaussian states have infinite relative entropy and are excluded. Second, if $\mathcal{F}$ depends smoothly on the mean and covariance, its first variation is quadratic and its Wasserstein force is affine. This includes the linear two-layer model with quadratic risk below, whose loss is a quadratic function of its uncentered cross-moment.

F.2 Centered Gaussian Linear Two-Layer Model

This subsection develops a concrete Gaussian-preserving example where the predictor is a linear map and the spectral flow closes on covariance matrices. Write m for the input and output dimension; the particle dimension in the general theory is then $d = 2m$. Consider particles

$$x = (u, v) \in \mathbb{R}^m \times \mathbb{R}^m,$$

and the linear two-layer feature map

$$\phi(z, x) = u(v^\top z), \qquad z \in \mathbb{R}^m.$$

Here u is the outer weight and v is the inner weight. For a probability measure μ on (u, v), the predictor is

$$H_\mu(z) = \int u(v^\top z)\, d\mu(u, v) = M_\mu z, \qquad M_\mu \coloneqq \int uv^\top\, d\mu(u, v).$$

With a linear teacher $y^\star(z) = A_\star z$ and data law ρ, the population least-squares loss is

$$\mathcal{F}(\mu) = \frac{1}{2}\int |H_\mu(z) - A_\star z|^2\, d\rho(z).$$

Writing

$$C_\rho \coloneqq \int zz^\top\, d\rho(z), \qquad B_\rho \coloneqq \int y^\star(z)z^\top\, d\rho(z) = A_\star C_\rho,$$

one obtains

$$\mathcal{F}(\mu) = \frac{1}{2}\operatorname{tr}(M_\mu C_\rho M_\mu^\top) - \operatorname{tr}(B_\rho M_\mu^\top) + \text{const}.$$

For centered Gaussian states

$$\mu = \mathcal{N}(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \Sigma_U & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_V \end{pmatrix},$$

the predictor is simply $M_\mu = \Sigma_{UV}$. Thus the cross-covariance block carries the learned linear map, while the diagonal covariance blocks influence the normalized metric direction.

The next proposition gives the closed covariance equation for Schatten spectral flows.

Proposition F.2 (Covariance ODE for centered Gaussian states)**.** *Let $p \in [1, \infty]$, let $\gamma = \gamma_p$ be the Schatten-p norm, and let*

$$q = \frac{2p}{2p-1}$$

be the dual exponent of $2p$, with $q = 1$ when $p = \infty$. On any interval where a centered Gaussian solution $\mu_t = \mathcal{N}(0, \Sigma_t)$ uses the canonical Schatten selector and Σ_t is positive definite, one has

$$\dot{\Sigma}_t = A_t \Sigma_t + \Sigma_t A_t^\top, \qquad A_t = B_t \Sigma_t^{-1/2},$$

where

$$B_t = -\|K_t\|_{S_q}^{2-q} U_t \operatorname{diag}(\sigma_i(t)^{q-1}) W_t^\top, \qquad K_t = L_t \Sigma_t^{1/2},$$

$K_t = U_t \operatorname{diag}(\sigma_i(t)) W_t^\top$ is a compact singular value decomposition over its positive singular values, and

$$L_t = \begin{pmatrix} 0 & E_t \\ E_t^\top & 0 \end{pmatrix}, \qquad E_t = \Sigma_{UV,t} C_\rho - B_\rho.$$

Equivalently,

$$\dot{\Sigma}_t = B_t \Sigma_t^{1/2} + \Sigma_t^{1/2} B_t^\top.$$

When $K_t = 0$, set $B_t = 0$. This inverse-covariance formula is only asserted up to the first loss of positive definiteness.

Proof. For centered Gaussian $\mu = \mathcal{N}(0, \Sigma)$, the loss depends on μ only through $M_\mu = \Sigma_{UV}$, so differentiating with respect to this block gives $E = \Sigma_{UV} C_\rho - B_\rho$. Therefore

$$g_\mu(u, v) = \nabla_x \frac{\delta \mathcal{F}}{\delta \mu}(u, v) = \begin{pmatrix} Ev \\ E^\top u \end{pmatrix} = Lx.$$

The Wasserstein gradient is affine, hence the Gaussian class is preserved by Corollary F.1. For an affine velocity $v(x) = Ax$, the selected direction solves

$$\min_A \left\{ \operatorname{tr}(L \Sigma A^\top) + \frac{1}{2} \left\| A \Sigma^{1/2} \right\|_{S_{2p}}^2 \right\}.$$

Setting $B = A\Sigma^{1/2}$ and $K = L\Sigma^{1/2}$ reduces this to the finite-dimensional Schatten duality-map problem

$$\min_B \left\{ \operatorname{tr}(K B^\top) + \frac{1}{2} \|B\|_{S_{2p}}^2 \right\},$$

whose minimizer is the displayed formula for B_t. Finally, if X_t follows $\dot{X}_t = A_t X_t$ and $X_t \sim \mathcal{N}(0, \Sigma_t)$, then $\dot{\Sigma}_t = A_t \Sigma_t + \Sigma_t A_t^\top$. □

F.3 Modal Reduction Under Simultaneous Diagonalization

This subsection shows that, under a compatible data basis and balanced initialization, the covariance ODE decomposes into interpretable feature modes. Assume that after orthogonal changes of coordinates one has

$$C_\rho = \mathrm{Id}, \qquad B_\rho = \operatorname{diag}(\beta_1, \ldots, \beta_m),$$

and use a balanced diagonal initialization

$$\Sigma_{U,0} = \operatorname{diag}(\alpha_i), \qquad \Sigma_{V,0} = \operatorname{diag}(\alpha_i), \qquad \Sigma_{UV,0} = 0, \qquad \alpha_i > 0.$$

Then the covariance remains diagonal mode by mode. We write

$$r_i(t) = (\Sigma_{UV,t})_{ii}, \qquad s_i(t) = (\Sigma_{U,t})_{ii} = (\Sigma_{V,t})_{ii},$$

so that r_i is the predictor coefficient and s_i measures the corresponding layer magnitude.

Proposition F.3 (Modal ODE)**.** *Let $p \in [1, \infty]$, let $q = (2p)'$ (in particular $q = 1$ for $p = \infty$), and define $\psi_q(a) = a|a|^{q-2}$ for $a \neq 0$ and $\psi_q(0) = 0$. Under the assumptions above, set*

$$a_i = s_i + r_i, \qquad b_i = s_i - r_i, \qquad \mathcal{N}_t^q = \sum_{j=1}^{m} |r_j - \beta_j|^q \left(a_j^{q/2} + b_j^{q/2}\right).$$

Then each mode solves

$$\dot{r}_i = -\mathcal{N}_t^{2-q}\, \psi_q(r_i - \beta_i) \left(a_i^{q/2} + b_i^{q/2}\right),$$

$$\dot{s}_i = -\mathcal{N}_t^{2-q}\, \psi_q(r_i - \beta_i) \left(a_i^{q/2} - b_i^{q/2}\right).$$

Equivalently,

$$\dot{a}_i = -2\mathcal{N}_t^{2-q}\, \psi_q(r_i - \beta_i) a_i^{q/2},$$

$$\dot{b}_i = \quad 2\mathcal{N}_t^{2-q}\, \psi_q(r_i - \beta_i) b_i^{q/2}.$$

If $\mathcal{N}_t = 0$, all right-hand sides are understood to be zero.

Proof. When Σ_{UV} is diagonal, the error $E_t = \Sigma_{UV,t} - B_\rho$ is diagonal, hence the affine gradient matrix

$$L_t = \begin{pmatrix} 0 & E_t \\ E_t & 0 \end{pmatrix}$$

preserves each coordinate plane. The balanced diagonal covariance structure is therefore invariant. In the i-th plane, the covariance block is

$$\Sigma_t^{(i)} = \begin{pmatrix} s_i & r_i \\ r_i & s_i \end{pmatrix},$$

with eigenvalues $a_i = s_i + r_i$ and $b_i = s_i - r_i$, while the corresponding block of L_t has eigenvalues $\pm(r_i - \beta_i)$. The singular values of $K_t = L_t \Sigma_t^{1/2}$ are therefore $|r_i - \beta_i|\, a_i^{1/2}$ and $|r_i - \beta_i|\, b_i^{1/2}$, with no additional factor of two. Applying the Schatten LMO of Proposition F.2 blockwise and collecting the common normalization gives the stated equations. □

F.4 Rank-One and Rank-Two Targets

This subsection records the two simplest modal systems, which are useful for interpreting how the spectral normalization changes learning profiles. If only one target coefficient is nonzero,

$$B_\rho = \mathrm{diag}(\beta, 0, \ldots, 0), \qquad \beta \neq 0,$$

then all inactive modes remain frozen and the dynamics reduces to the single pair (r, s) governed by Proposition F.3. In the positive-definite interior, the variables $a = s + r$ and $b = s - r$ are positive and encode the two layer variances in the active symmetric and antisymmetric directions. This one-mode system is the basic phase portrait behind the Gaussian experiments: the predictor coefficient r_t moves toward β, while the terminal value of s_t records the implicit covariance bias selected by p. The admissible covariance cone is

$$\{(r, s) \in \mathbb{R}^2 : \; s \geq |r|\},$$

with positive-definite interior $s > |r|$, equivalently $a > 0$ and $b > 0$. In Figure 5(a), we display interior trajectories from two initial conditions. The next proposition gives the interior conserved leaves and explains why they need not describe an entire trajectory.

Proposition F.4 (Interior invariants and finite-time rank loss)**.** *For each mode, on every interval where $a_i, b_i > 0$, the trace geometry $p = 1$ preserves $a_i b_i = s_i^2 - r_i^2$. For $p > 1$, set $\eta = 1 - q/2$; then $a_i^\eta + b_i^\eta$ is constant. In particular, $p = \infty$ preserves $\sqrt{s_i + r_i} + \sqrt{s_i - r_i}$. From the balanced initialization $a_i = b_i = \alpha_i$, an interior trajectory with $p > 1$ satisfies*

$$|r_i| < R_{p,i} := \alpha_i 2^{p/(p-1)}, \qquad R_{\infty,i} = 2\alpha_i.$$

If $|\beta_i| > R_{p,i}$, any global nonnegative modal solution must lose rank in this mode in finite time.

Proof. Write $\Lambda_i = \mathcal{N}_t^{2-q}\psi_q(r_i - \beta_i)$. The equations $\dot{a}_i = -2\Lambda_i a_i^{q/2}$ and $\dot{b}_i = 2\Lambda_i b_i^{q/2}$ give the invariants by differentiation. For $p > 1$, the balanced invariant is $a_i^\eta + b_i^\eta = 2\alpha_i^\eta$, so $a_i, b_i < (2\alpha_i^\eta)^{1/\eta}$ in the interior, giving the displayed bound on $r_i = (a_i - b_i)/2$. The sign of $\dot{r}_i$ points toward β_i. If $|\beta_i| > R_{p,i}$, the residual stays bounded away from zero until rank loss, and the growing eigenvalue is at least α_i. Consequently $|\Lambda_i| \geq c > 0$. The shrinking eigenvalue z then satisfies $\dot{z} \leq -2cz^{q/2}$. Since $q/2 < 1$, it reaches zero in finite time. This argument allows other modes to have already lost rank. □

After rank loss the preceding invariant need no longer be conserved. To specify the subsequent dynamics, we use the modal equations with every extinct eigenvalue kept at zero. This is a choice of continuation, not a claim of uniqueness for the full measure-valued PDE.

Proposition F.5 (Selected continuation on the covariance boundary)**.** *Let (a_i, b_i) be a locally absolutely continuous, nonnegative solution of the modal equations, keeping zero eigenvalues at zero. Then its centered Gaussian law solves the continuity equation with the canonical Schatten selector, including after rank loss, and is a curve of maximal slope for the linear-network loss. For almost every time, its energy satisfies*

$$\frac{d}{dt}\mathcal{F}(\mu_t) = -\mathcal{N}_t^2, \qquad \mathcal{F}(\mu_t) = \frac{1}{2}\sum_i (r_i - \beta_i)^2.$$

This statement is conditional on the specified modal solution; it does not assert well-posedness for arbitrary initial measures.

Proof. Use the fixed orthonormal symmetric/antisymmetric coordinates in each mode and realize the law by independent coordinates $\sqrt{a_i(t)}\xi_i^+$ and $\sqrt{b_i(t)}\xi_i^-$, with standard normal $\xi_i^\pm$. Before extinction,

$$\frac{d}{dt}\sqrt{a_i} = -\Lambda_i a_i^{(q-1)/2}, \qquad \frac{d}{dt}\sqrt{b_i} = \Lambda_i b_i^{(q-1)/2}.$$

These derivatives are locally bounded, since $q \geq 1$ and the modal solution is locally bounded; after extinction they are zero. Thus this realization is locally absolutely continuous in L^2 and its derivative is precisely the canonical Schatten velocity on the support. Its action satisfies $\mathcal{N}_{\mu_t}(v_t)^2 = \mathcal{N}_t^2$ by (12). This proves the continuity equation with locally finite action, even if its affine coefficient is not integrable through extinction; no inverse of Σ_t is needed. Differentiating the loss gives $\mathcal{F}' = -\mathcal{N}_t^2$ by the modal ODE. The transport–Taylor estimate of Theorem 5.1 holds for this quadratic moment functional, including at singular Gaussian measures: for each coupling $\pi \in \Pi(\mu_t, \nu)$ with small squared displacement, expanding the moments gives a remainder $O(\int |y - x|^2 \, d\pi)$ at a fixed state. This is the displacement cost of the chosen coupling, not necessarily the optimal $\mathsf{W}_2(\mu_t, \nu)^2$. Its slope is therefore $\mathcal{N}_t$, by the slope argument in that theorem, which does not require regularity of the selected velocity. The dynamic formula gives $|\mu'|_{\mathsf{W}_\gamma} \leq \mathcal{N}_t$; the slope inequality and the energy identity give the reverse bound almost everywhere. Integrating yields the metric energy-dissipation identity across rank changes. □

For the first mode in Figure 6(b), $\alpha_1 = 0.1$, so $R_{2,1} = 0.4$, $R_{4,1} \simeq 0.252$, and $R_{\infty,1} = 0.2$. All four plotted targets exceed the latter two bounds in absolute value, and three exceed the first. These trajectories therefore require the boundary continuation, not just the positive-definite covariance formula.

Figures 5 and 6 illustrate these rank-one and rank-two reductions using the archived trajectories.

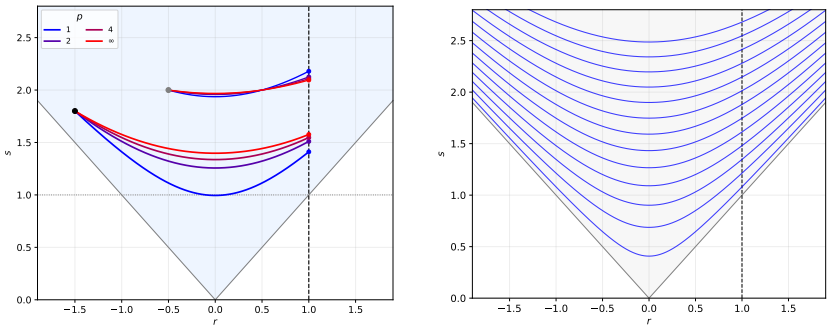
For a rank-two target

$$B_\rho = \mathrm{diag}(\beta_1, \beta_2, 0, \ldots, 0),$$

the first two modes solve the same equations but are coupled through

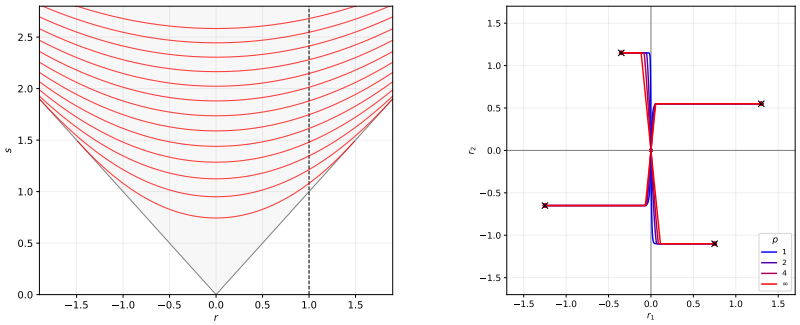
$$\mathcal{N}_t^q = \sum_{j=1}^{2} |r_j - \beta_j|^q \left(a_j^{q/2} + b_j^{q/2}\right).$$

Thus the trace geometry evolves the two learned directions independently, whereas every $p > 1$ geometry shares a global normalization across them. This is the simplest Gaussian setting where the matrix nature of the spectral flow is visible at the predictor level rather than only through a time rescaling of a single scalar mode. In Figure 6(b), this coupling appears as distinct planar paths toward different targets (β_1, β_2) under the same initialization.



(a) One-mode trajectories (b) Conserved leaves, $p = 1$

Figure 5: Gaussian one-mode geometry. (a) Trajectories in (r, s) for $p \in \{1, 2, 4, \infty\}$. (b) Interior conserved foliation for $p = 1$. After rank loss, trajectories use the continuation of Proposition F.5; interior invariants no longer apply.



(a) Conserved leaves, $p = \infty$ (b) Two-mode dynamics

Figure 6: Gaussian operator leaves and coupled modes. (a) Interior conserved foliation for $p = \infty$. (b) Rank-two trajectories in (r_1, r_2), initially $r_1 = r_2 = 0$ and $(s_1, s_2) = (0.1, 10)$, toward the indicated targets (β_1, β_2). The same rank-boundary continuation is used as in Figure 5.

G Numerical Protocols and Additional Examples

G.1 Time Normalization and Numerical Safeguards

The experiments are qualitative illustrations of the chosen geometry, not wall-clock or generalization benchmarks. Their notebooks differentiate the empirical loss $\mathcal{F}(\mu_X)$, whereas Proposition 2.5 uses $\mathrm{F}(X) = n\mathcal{F}(\mu_X)$. By homogeneity of the LMO, an unclipped notebook step of size Δt therefore corresponds to a mean-field Euler step of size $\Delta t/n$. Table 1 lists the nominal step sizes and counts. Each loss plot independently maps its run's nominal horizon to $[0, 1]$; these display coordinates are not common intrinsic flow times.

Example	n	steps $(1, \infty)$	Δt_1	Δt_∞	step cap
ReLU MLP	1920	$(2000, 2000)$	75	50	0.20
Attention	200	$(960, 480)$	100	180	0.45
MMD	200	$(6000, 6000)$	5	3.2	none

Table 1: Nominal notebook clocks. Caps apply to the Frobenius norm of the entire particle-matrix increment, not separately to particles.

The MLP and attention implementations multiply an oversized increment by a scalar to enforce the stated cap, and additionally guard coordinates at ± 30 and ± 10, respectively. The cap makes effective step sizes state-dependent; an active coordinate guard can also change the direction. The MMD experiment uses the smoothed distance $\sqrt{\|x - y\|^2 + 0.01^2}$ and unclipped Euler steps. The

archived plots do not record cumulative effective times or all safeguard activations, so no corrected continuum-clock comparison is inferred from them. The slight difference between pre-update loss samples and post-update snapshot labels is also a discretization convention, not a theoretical time rescaling.

An initialization-only check of the MLP notebook (seed 7, double precision, 2000 training covariates) makes this distinction concrete. Writing $G = \nabla_X \mathcal{F}(\mu_X)$, the ideal mean-field dissipations are $D_1 = n\|G\|_F^2$ and $D_\infty = n\|G\|_*^2$. Their initial ratio is $D_\infty/D_1 \simeq 2.55765$, whereas the nominal horizon ratio is 1.5. Both initial updates are already clipped, with factors approximately 0.74469 and 0.40326; their first-order predicted loss-decrease ratio per displayed step is consequently 0.92334, not 2.55765. These initialization checks do not require running training. An initial dissipation ratio alone cannot calibrate the entire trajectories. We therefore retain the archived curves without relabeling them by this ratio; their interpretation is qualitative.

For attention, the five paired trials redraw both teacher and student. The effective-token statistics use all attention weights, whereas the displayed weight histograms restrict to $[0, 0.1]$ and renormalize there. The Gaussian modal figures use the ODE of Proposition F.3, integrated in double precision by RK4 with $T = 18$ and $\Delta t = 2 \times 10^{-4}$. Negative numerical covariance eigenvalues are projected to zero; extinct modes are not artificially revived. We check the vector field and interior conserved leaves independently. A separate event-resolving integration uses transformed variances to locate rank-loss and zero-error events, with the continuation of Proposition F.5. Across the 24 parameter cases, it detects 11 rank-loss events. Reference refinement changes their times by at most 4.9×10^{-8}. In local windows around all detected events, refining the notebook step from 2×10^{-4} to 5×10^{-5} reduces the maximum tested covariance-eigenvalue error from 3.5×10^{-4} to 8.6×10^{-5}. These are local consistency checks, not a certified global error bound; in particular, classical fourth-order accuracy is not assumed at nonsmooth events. This path interpretation is separate from the empirical neural-network clocks.

The public reproduction runner records software versions, source and input hashes, elapsed time, and host information for new runs. Complete hardware, peak-memory, and elapsed-time records for the archived full experiments are unavailable; recent smoke-test timings are not substitutes for them.

G.2 Archived Diagnostics

These panels show the same archived runs discussed in Section 5. The nominal-time and safeguard qualifications of Section G.1 apply throughout; the curves have not been recalibrated or replaced by new simulations.

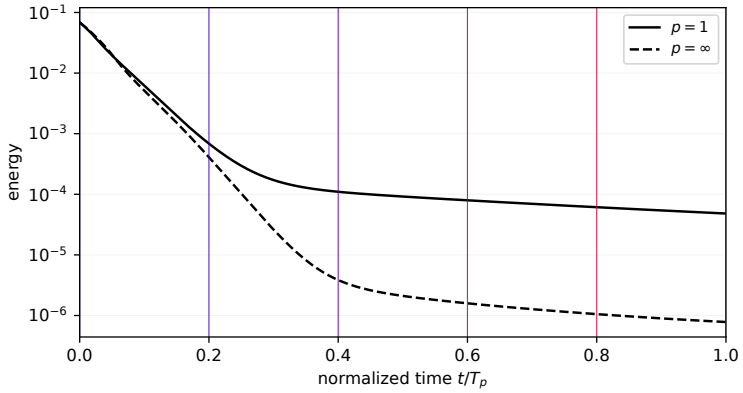


Figure 7: ReLU MLP energy on separately normalized nominal horizons: solid $p = 1$, dashed $p = \infty$. Colored vertical lines identify the snapshots in Figure 8; equal horizontal coordinates do not represent equal intrinsic flow times. Particle trajectories appear in Figure 3.

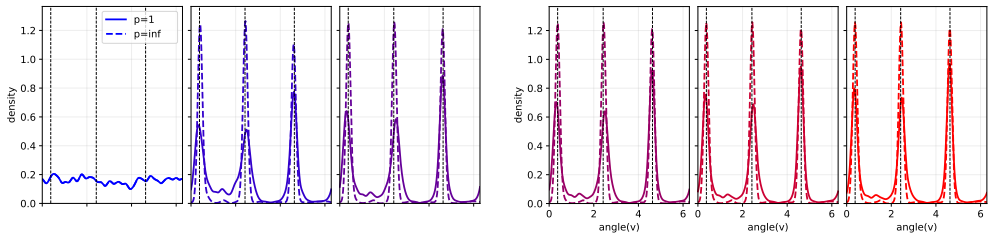
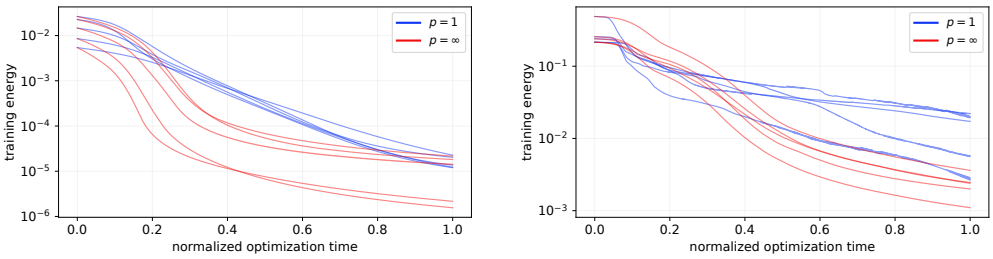


Figure 8: Circular histogram snapshots for the ReLU MLP, in chronological order from left to right, with the same angular axis in all panels. Solid curves correspond to $p = 1$ and dashed curves to $p = \infty$ (Muon); colors match the snapshot markers in Figure 7. Dashed vertical lines mark teacher directions. Only the legacy legend text has been corrected.



(a) Dense attention

(b) Sparse attention

Figure 9: Shallow-attention loss curves for five paired trials per regime, with $p = 1$ in blue and $p = \infty$ in red. Each trial redraws both teacher and student. Effective support is approximately 16 tokens in the dense regime and 2.1 in the sparse regime. The horizontal axes use separately normalized nominal horizons, not matched intrinsic flow times.

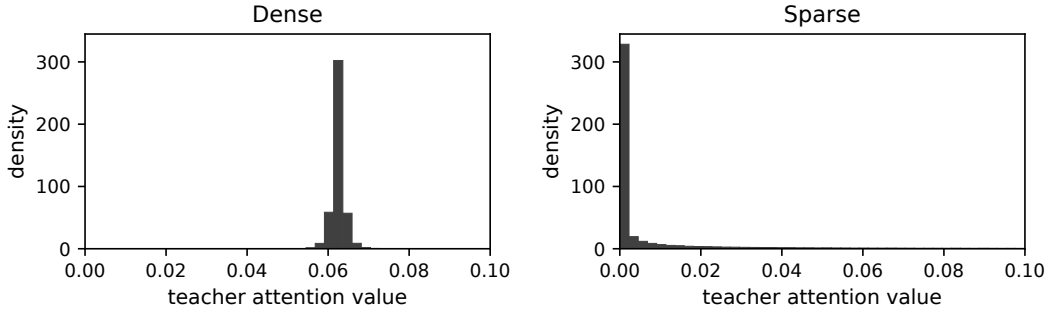


Figure 10: Teacher-attention histograms for the dense and sparse regimes. Weights are truncated to $[0, 0.1]$ and renormalized on that interval; sparse weights larger than 0.1 are not displayed. Effective-token statistics in the text instead use all attention weights. Both panels use the same horizontal scale.

G.3 Couplings and Interpolations

This subsection complements Figures 1 and 2 with additional planar shape pairs. Each example uses an image-defined source and a translated image-defined target. Trace couplings are exact assignment solutions. Operator couplings are obtained by alternating a projected assignment with a rank-one cost update, for at most 40 iterations from a horizontal-axis initialization. This heuristic supplies no primal–dual gap certificate and is not asserted to solve the full convex SDP. Accordingly, its interpolations are illustrative displacement paths, not certified W_γ-geodesics. Rotation invariance holds for the underlying Schatten transport problem, not necessarily for this finitely iterated, fixed-initialization solver.

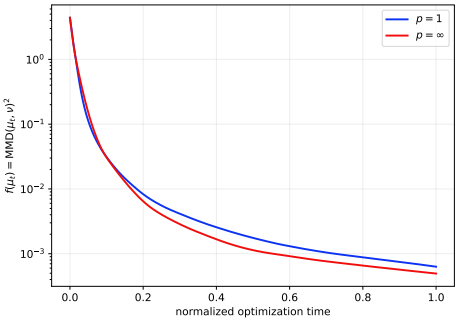
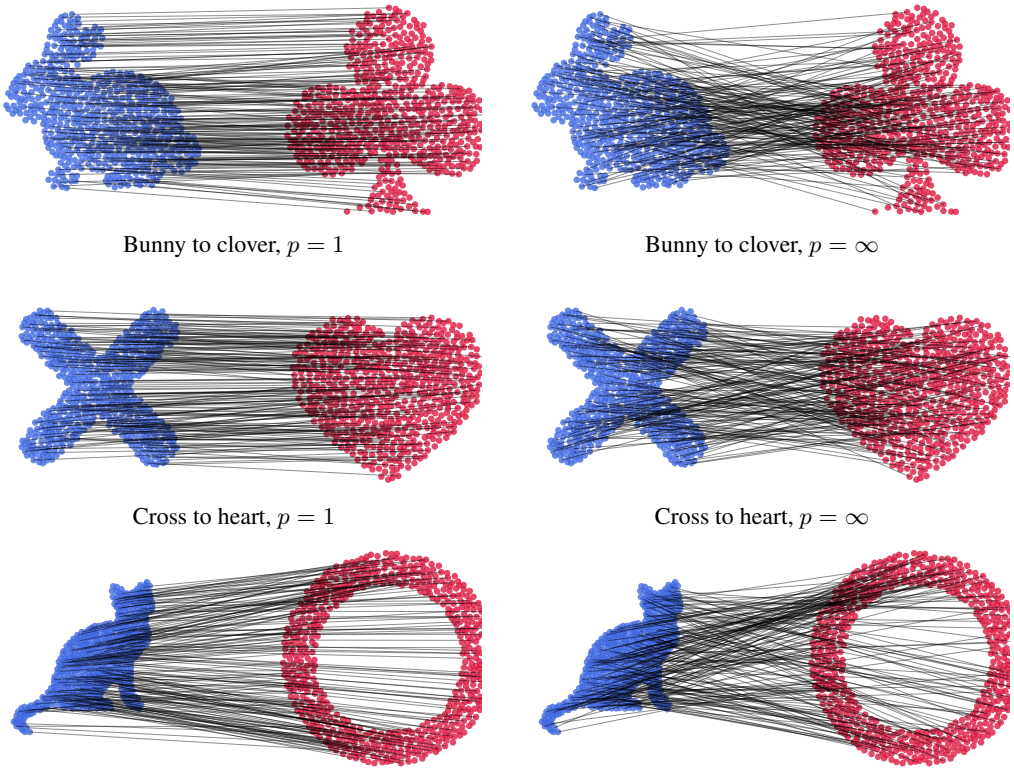


Figure 11: Archived MMD objective values for the trajectories in Figure 4. Nominal horizons are normalized separately for $p = 1$ and $p = \infty$. These curves do not establish an intrinsic convergence-rate or wall-clock ordering.

Figure 12 shows additional static couplings for three image-defined source and target pairs, complementing the annulus-to-disks example in the main text.



Bunny to clover, $p = 1$ | Bunny to clover, $p = \infty$

Cross to heart, $p = 1$ | Cross to heart, $p = \infty$

Cat to annulus, $p = 1$ | Cat to annulus, $p = \infty$

Figure 12: Additional static Spectral Wasserstein couplings. Only a subsample of matching segments is displayed.

Figure 13 displays the corresponding displacement-interpolation density snapshots, using the same computed couplings as in Figure 12.

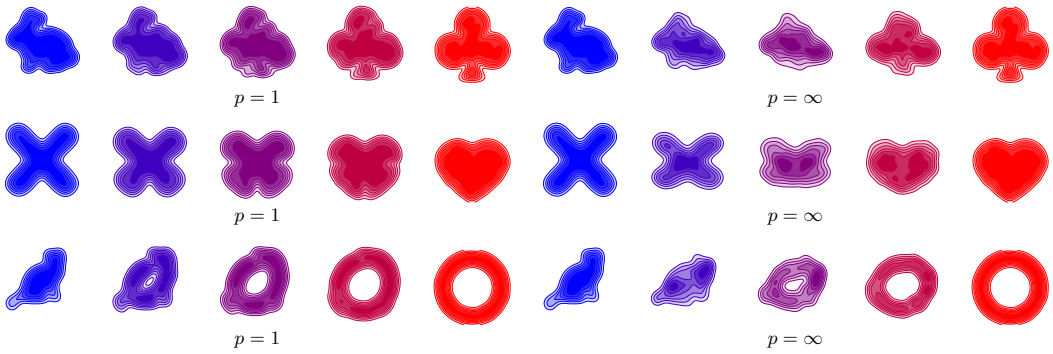


Figure 13: Additional displacement-interpolation density snapshots for the same shape pairs as Figure 12.

H Geodesic Convexity

This section studies one of the basic structural notions behind metric gradient flows, namely geodesic convexity, for the Spectral Wasserstein geometries introduced above. Geodesic convexity is desirable because it supports stability and global convergence statements [Ambrosio et al., 2008], but for many machine-learning objectives it is arguably too strong, so here it is used as a structural benchmark rather than as a blanket modeling assumption.

Proposition H.1 (Basic displacement-convexity examples)**.** *Assume that γ is monotone. A linear functional $\mathcal{F}(\mu) = \int h\, d\mu$, with $h : \mathbb{R}^d \to \mathbb{R}$, is convex along every static displacement geodesic if and only if h is convex. A quadratic interaction functional $\mathcal{F}(\mu) = \int h(x, y)\, d\mu(x) d\mu(y)$, with $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, is convex along every static displacement geodesic if h is convex. Finally, let $d\nu = Z^{-1} e^{-V} dx$, with $V \in C^2(\mathbb{R}^d)$ convex, and define the relative entropy by $\mathcal{F}(\mu) = \int \log(d\mu/d\nu)\, d\mu$ if $\mu \ll \nu$, and $\mathcal{F}(\mu) = +\infty$ otherwise. If the representing set $\mathcal{K}_\gamma$ contains at least one positive definite matrix, then for every $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ there exists a constant-speed W_γ-geodesic $(\mu_t)_{t\in[0,1]}$ joining them along which $\mathcal{F}$ is convex.*

H.1 Definition and General Setup

This subsection recalls the relevant notion of convexity along Spectral Wasserstein geodesics and explains why it matters for the variational analysis of the flow. When one studies gradient flows in metric spaces, geodesic convexity plays the same role as ordinary convexity in Euclidean optimization: it governs uniqueness, stability, and the variational structure of the dynamics. Corollary 4.4 provides a distinguished class of geodesics from optimal couplings, for which convexity inequalities can be tested directly along Euclidean segments.

The next definition fixes the notion used in the remainder of this section.

Definition H.2 (Geodesic convexity)**.** *A functional $F : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is called κ-geodesically convex for W_γ if along every constant-speed W_γ-geodesic $(\mu_t)_{t\in[0,1]}$ one has*

$$F(\mu_t) \le (1-t)F(\mu_0) + tF(\mu_1) - \frac{\kappa}{2} t(1-t)\mathsf{W}_\gamma(\mu_0, \mu_1)^2.$$

We call a geodesic a *static displacement geodesic* if it has the form $\mu_t = ((1-t)x + ty)_\# \pi_*$ for a static optimal coupling π_*, as in Corollary 4.4. Full geodesic convexity concerns every constant-speed geodesic and is stronger when the metric admits branching.

Remark H.3 (Branching at the operator endpoint)**.** *This distinction is genuine for the Muon geometry. Let $\gamma(S) = \lambda_{\max}(S)$ on $\mathbb{S}^2_+$, choose a nonzero 1-Lipschitz function $f : [0, 1] \to \mathbb{R}$ with $f(0) = f(1) = 0$, and set*

$$\mu_t = \tfrac{1}{2}\delta_{(t,f(t))} + \tfrac{1}{2}\delta_{(t,-f(t))}.$$

For $s < t$, matching equal signs gives displacement covariance $\mathrm{diag}((t-s)^2, (f(t)-f(s))^2)$ and therefore cost $(t-s)^2$. Conversely, the change of the first-coordinate mean gives the same lower bound for every coupling. Hence (μ_t) is a constant-speed W_γ-geodesic from $\delta_{(0,0)}$ to $\delta_{(1,0)}$, but

it is not a static displacement geodesic when $f \neq 0$. For the convex function $h(x) = x_2^2$, one has $\int h\, d\mu_0 = \int h\, d\mu_1 = 0$ but $\int h\, d\mu_t = f(t)^2 > 0$ whenever $f(t) \neq 0$. Thus convexity of h ensures convexity along every static displacement geodesic, but not full geodesic convexity at this endpoint.

H.2 Linear Functionals

This subsection shows that for linear functionals the qualitative and quantitative convexity criteria reduce to explicit pointwise conditions on the integrand. For a Borel function $h : \mathbb{R}^d \to \mathbb{R}$, define the linear functional

$$F_h(\mu) \coloneqq \int_{\mathbb{R}^d} h(x)\, d\mu(x).$$

Because static displacement geodesics are given by affine interpolation in space, the convexity question for F_h along this class is especially transparent.

The next theorem gives the exact criterion.

Theorem H.4 (Linear functionals along displacement geodesics)**.** *Assume γ is monotone and let $\mathcal{K}_\gamma$ be a convex compact PSD representing set for γ as in Proposition A.2.*

1. *The functional F_h is convex along every static displacement geodesic if and only if h is convex on $\mathbb{R}^d$.*
2. *If in addition $h \in C^2(\mathbb{R}^d)$ and $\kappa \geq 0$, then F_h satisfies the κ-convexity inequality along every static displacement geodesic if and only if*

$$\nabla^2 h(z) \succeq \kappa Q \qquad \text{for every } z \in \mathbb{R}^d,\ Q \in \mathcal{K}_\gamma.$$

Equivalently,

$$\xi^\top \nabla^2 h(z)\, \xi \geq \kappa \max_{Q \in \mathcal{K}_\gamma} \xi^\top Q\, \xi \qquad \text{for every } z, \xi \in \mathbb{R}^d.$$

For $\kappa < 0$, this displayed scalar reformulation with a maximum is not equivalent to the matrix inequality; multiplication by the negative scalar reverses the extremum, so the corresponding pointwise scalar condition would involve $\min_{Q \in \mathcal{K}_\gamma} \xi^\top Q \xi$. No equivalence for negative κ is asserted here.

Proof. Assume first that h is convex and let

$$\mu_t = ((1-t)x + ty)_{\#}\pi_*$$

be any static displacement geodesic. Then

$$F_h(\mu_t) = \int h((1-t)x + ty)\, d\pi_*(x, y).$$

By the convexity of h,

$$h((1-t)x + ty) \leq (1-t)h(x) + th(y),$$

and integrating yields geodesic convexity of F_h.

Conversely, apply the displacement-convexity assumption to Dirac masses $\mu_0 = \delta_x$ and $\mu_1 = \delta_y$. Their static displacement geodesic is

$$\mu_t = \delta_{(1-t)x+ty},$$

so

$$h((1-t)x + ty) = F_h(\mu_t) \leq (1-t)h(x) + th(y),$$

which proves convexity of h.

Assume now that $h \in C^2$ and that

$$\nabla^2 h(z) \succeq \kappa Q \qquad \text{for every } z \in \mathbb{R}^d,\ Q \in \mathcal{K}_\gamma.$$

For each fixed $Q \in \mathcal{K}_\gamma$, the function $z \mapsto h(z) - \frac{\kappa}{2} z^\top Q z$ is convex. Hence

$$h((1-t)x + ty) \leq (1-t)h(x) + th(y) - \frac{\kappa}{2} t(1-t)(y-x)^\top Q (y-x).$$

Integrate against any static optimal coupling π_* and then maximize $\mathrm{tr}(Q\Sigma_{\pi_*})$ over $Q \in \mathcal{K}_\gamma$. Since $\kappa \geq 0$, this gives

$$F_h(\mu_t) \leq (1-t)F_h(\mu_0) + tF_h(\mu_1) - \frac{\kappa}{2}t(1-t)\mathsf{W}_\gamma(\mu_0,\mu_1)^2.$$

No differentiation under the integral is needed. Convexity of h gives an affine lower bound, so these integrals are well-defined on $\mathcal{P}_2$; the argument applies to finite endpoint energies, with the other cases understood in the extended-value sense.

Conversely, test the midpoint inequality between $\delta_{z-\varepsilon\xi}$ and $\delta_{z+\varepsilon\xi}$. It gives

$$h(z) \leq \frac{h(z-\varepsilon\xi)+h(z+\varepsilon\xi)}{2} - \frac{\kappa}{2}\varepsilon^2\gamma(\xi\xi^\top).$$

Subtracting $h(z)$, dividing by $\varepsilon^2/2$, and sending $\varepsilon \to 0$ yields

$$\xi^\top \nabla^2 h(z)\,\xi \geq \kappa\gamma(\xi\xi^\top) = \kappa \max_{Q\in\mathcal{K}_\gamma} \xi^\top Q\,\xi.$$

Because $\kappa \geq 0$, this scalar inequality is equivalent to the matrix inequality

$$\nabla^2 h(z) \succeq \kappa Q \qquad \text{for every } Q \in \mathcal{K}_\gamma.$$

□

The next remark interprets the condition of Theorem H.4 for the Schatten family.

Remark H.5 (Schatten norms)**.** *For the Schatten geometries used throughout the paper and for $\kappa \geq 0$, the condition*

$$\nabla^2 h(z) \succeq \kappa Q \qquad \textit{for every } Q \in \mathcal{K}_\gamma$$

is equivalent to

$$\nabla^2 h(z) \succeq \kappa \mathrm{Id}.$$

Indeed, for $p = 1$ we may choose $\mathcal{K}_\gamma = \{\mathrm{Id}\}$. For $1 < p \leq \infty$, a convenient choice is

$$\mathcal{K}_\gamma = \{Q \succeq 0 :\ \|Q\|_{S_{p'}} \leq 1\}, \qquad p' = \frac{p}{p-1},$$

with the usual interpretation $p' = 1$ when $p = \infty$. Under the standing condition $\kappa \geq 0$ in Theorem H.4, every such Q satisfies $Q \preceq \mathrm{Id}$, and every rank-one projector belongs to $\mathcal{K}_\gamma$; hence the condition is equivalent to $\nabla^2 h(z) \succeq \kappa\mathrm{Id}$, exactly as in the classical W_2 geometry. For negative κ, this simplification is not asserted.

H.3 Interaction Functionals

The same segmentwise argument applies to convex interactions.

Proposition H.6 (Convex interactions)**.** *Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be jointly convex and define*

$$F_h^{\mathrm{int}}(\mu) = \int h(x,x')\,d\mu(x)d\mu(x').$$

Then F_h^{int} is convex along every static displacement geodesic.

Proof. Let $\mu_t = ((1-t)x + ty)_\#\pi_*$ and take two independent copies $(x,y),(x',y')$ with law π_*. Joint convexity gives

$$h((1-t)x+ty,(1-t)x'+ty') \leq (1-t)h(x,x') + th(y,y').$$

Integrating with respect to $\pi_* \otimes \pi_*$ proves the claim. □

H.4 Relative Entropy

This subsection examines the geodesic convexity of relative entropy, which is a fundamental nonlinear example underlying diffusion-type gradient flows. Section H.5 treats in detail the case of a Gaussian target ν. For the canonical selector, the Gaussian ansatz closes as an ODE for the mean and covariance. Let

$$d\nu(x) = Z^{-1}e^{-V(x)}\,dx, \qquad V \in C^2(\mathbb{R}^d),$$

and define the relative entropy

$$\mathrm{Ent}_\nu(\mu) \coloneqq \begin{cases} \displaystyle\int \log\left(\frac{d\mu}{d\nu}\right)\,d\mu, & \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases}$$

The trace case is the classical displacement-convexity theory of entropy. For general spectral geometries the same argument works only when the active quadratic costs remain uniformly elliptic, and otherwise one only gets a weaker statement by approximation.

The next theorem records the all-geodesics result in the uniformly elliptic regime.

Theorem H.7 (Full entropy convexity under uniform ellipticity)**.** *Assume γ is monotone and that $\mathcal{K}_\gamma$ can be chosen so that*

$$\mathcal{K}_\gamma \subset \mathbb{S}^d_{++}.$$

Assume moreover that

$$\nabla^2 V(x) \succeq \kappa Q \qquad \textit{for every } x \in \mathbb{R}^d,\ Q \in \mathcal{K}_\gamma.$$

Then Ent_ν is κ-geodesically convex on $(\mathcal{P}_2(\mathbb{R}^d), \mathsf{W}_\gamma)$.

Proof. Let (μ_t) be any constant-speed W_γ-geodesic and set $D = \mathsf{W}_\gamma(\mu_0, \mu_1)$. Choose a dual maximizer Q_* for its endpoints. Then $Q_* \succ 0$ and $\mathsf{W}^{Q_*}(\mu_0, \mu_1) = D$. For $0 \le s \le t \le 1$, the triangle inequality for W^{Q_*} and the bound $\mathsf{W}^{Q_*} \le \mathsf{W}_\gamma$ give

$$\begin{aligned} D &\le \mathsf{W}^{Q_*}(\mu_0, \mu_s) + \mathsf{W}^{Q_*}(\mu_s, \mu_t) + \mathsf{W}^{Q_*}(\mu_t, \mu_1) \\ &\le sD + (t-s)D + (1-t)D = D. \end{aligned}$$

Equality throughout implies $\mathsf{W}^{Q_*}(\mu_s, \mu_t) = (t-s)D$. Thus the given curve is also a constant-speed W^{Q_*}-geodesic. With $L = Q_*^{1/2}$, set $\widetilde{\mu}_t = L_\sharp \mu_t$ and $\widetilde{\nu} = L_\sharp \nu$. The transformed curve is a W_2-geodesic and

$$\mathsf{W}_2(\widetilde{\mu}_0, \widetilde{\mu}_1)^2 = \mathsf{W}_\gamma(\mu_0, \mu_1)^2.$$

The reference measure $\widetilde{\nu}$ has density

$$d\widetilde{\nu}(z) = \widetilde{Z}^{-1}e^{-\widetilde{V}(z)}\,dz, \qquad \widetilde{V}(z) = V(L^{-1}z) + \text{constant},$$

so

$$\nabla^2 \widetilde{V}(z) = L^{-1}\nabla^2 V(L^{-1}z)L^{-1}.$$

Because $Q_* \in \mathcal{K}_\gamma$, the curvature assumption gives

$$\nabla^2 V(x) \succeq \kappa Q_* = \kappa L^2,$$

hence

$$\nabla^2 \widetilde{V}(z) \succeq \kappa \mathrm{Id}.$$

The classical entropy convexity theorem for W_2 therefore yields

$$\mathrm{Ent}_{\widetilde{\nu}}(\widetilde{\mu}_t) \le (1-t)\,\mathrm{Ent}_{\widetilde{\nu}}(\widetilde{\mu}_0) + t\,\mathrm{Ent}_{\widetilde{\nu}}(\widetilde{\mu}_1) - \frac{\kappa}{2}t(1-t)\mathsf{W}_2(\widetilde{\mu}_0, \widetilde{\mu}_1)^2.$$

Finally, relative entropy is invariant under the invertible change of variables L, so

$$\mathrm{Ent}_{\widetilde{\nu}}(\widetilde{\mu}_t) = \mathrm{Ent}_\nu(\mu_t),$$

and similarly at the endpoints. This gives the result. □

The next theorem explains what remains true in the more relevant but singular regime where the condition $\mathcal{K}_\gamma \subset \mathbb{S}^d_{++}$ fails. This lack of ellipticity is precisely what happens for Schatten-p geometries with $p > 1$, because the corresponding representing sets contain singular matrices, but one still retains a weaker geodesic-convexity statement by approximation.

Theorem H.8 (Weak entropy convexity)**.** *Assume γ is monotone, that $\mathcal{K}_\gamma \subset \mathbb{S}^d_+$ is convex compact, and that*

$$\mathcal{K}_\gamma \cap \mathbb{S}^d_{++} \neq \varnothing.$$

Assume moreover that

$$\nabla^2 V(x) \succeq \kappa Q \qquad \text{for every } x \in \mathbb{R}^d,\ Q \in \mathcal{K}_\gamma.$$

Then for every $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ there exists at least one constant-speed W_γ-geodesic $(\mu_t)_{t\in[0,1]}$ from μ_0 to μ_1 such that

$$\mathrm{Ent}_\nu(\mu_t) \le (1-t)\,\mathrm{Ent}_\nu(\mu_0) + t\,\mathrm{Ent}_\nu(\mu_1) - \frac{\kappa}{2} t(1-t)\mathsf{W}_\gamma(\mu_0,\mu_1)^2.$$

Proof. Choose $Q_0 \in \mathcal{K}_\gamma \cap \mathbb{S}^d_{++}$ and define

$$\mathcal{K}_{\gamma,\varepsilon} := (1-\varepsilon)\mathcal{K}_\gamma + \varepsilon Q_0, \qquad \varepsilon \in (0,1).$$

Then every element of $\mathcal{K}_{\gamma,\varepsilon}$ is positive definite. Let γ_ε be the support function of $\mathcal{K}_{\gamma,\varepsilon}$, namely

$$\gamma_\varepsilon(S) = \max_{Q\in\mathcal{K}_{\gamma,\varepsilon}} \mathrm{tr}(QS) = (1-\varepsilon)\gamma(S) + \varepsilon\,\mathrm{tr}(Q_0 S).$$

Hence

$$(1-\varepsilon)\gamma(S) \le \gamma_\varepsilon(S) \le \gamma(S)$$

for all $S \succeq 0$, and therefore

$$(1-\varepsilon)\mathsf{W}_\gamma(\mu_0,\mu_1)^2 \le \mathsf{W}_{\gamma_\varepsilon}(\mu_0,\mu_1)^2 \le \mathsf{W}_\gamma(\mu_0,\mu_1)^2.$$

Moreover, convexity of $\mathcal{K}_\gamma$ gives $\mathcal{K}_{\gamma,\varepsilon} \subset \mathcal{K}_\gamma$, so the curvature assumption is unchanged.

By Theorem H.7, each regularized geometry has full entropy convexity. Choose an optimal coupling π_ε for $\mathsf{W}_{\gamma_\varepsilon}$ and set

$$\mu_t^\varepsilon = ((1-t)x + ty)_\# \pi_\varepsilon.$$

Then

$$\mathrm{Ent}_\nu(\mu_t^\varepsilon) \le (1-t)\,\mathrm{Ent}_\nu(\mu_0) + t\,\mathrm{Ent}_\nu(\mu_1) - \frac{\kappa}{2} t(1-t)\mathsf{W}_{\gamma_\varepsilon}(\mu_0,\mu_1)^2.$$

By compactness of couplings, one may extract a subsequence $\pi_{\varepsilon_n} \rightharpoonup \pi_*$. Setting

$$\mu_t = ((1-t)x + ty)_\# \pi_*,$$

the lower semicontinuity of relative entropy and the convergence

$$\mathsf{W}_{\gamma_{\varepsilon_n}}(\mu_0,\mu_1)^2 \to \mathsf{W}_\gamma(\mu_0,\mu_1)^2$$

yield the claimed inequality along the limit geodesic. The limit coupling π_* is W_γ-optimal by continuity of the displacement covariance and of the support functions on trace-bounded sets. □

The next proposition shows that, in the operator case, full geodesic convexity of relative entropy fails in a very concrete way.

Proposition H.9 (Operator-norm counterexample)**.** *Let $\gamma(S) = \lambda_{\max}(S)$ on $\mathbb{S}^2_+$, and let*

$$d\nu(x) = \frac{1}{2\pi} e^{-|x|^2/2}\,dx.$$

Then Ent_ν is not geodesically convex on $(\mathcal{P}_2(\mathbb{R}^2), \mathsf{W}_\gamma)$.

Proof. Fix parameters $L > 0$ and $a > 0$ such that

$$\frac{4a^2}{3} \le L^2.$$

Set

$$R_0 = [0,1] \times [-a,a], \qquad R_1 = [L, L+1] \times [-a,a],$$

and let

$$\mu_i = \frac{1}{2a} \mathbf{1}_{R_i}\, dx, \qquad i = 0, 1.$$

Both endpoints are absolutely continuous with bounded support, hence satisfy

$$\mathrm{Ent}_\nu(\mu_0) < +\infty, \qquad \mathrm{Ent}_\nu(\mu_1) < +\infty.$$

For any coupling π between μ_0 and μ_1, define

$$S_\pi = \int (y-x)(y-x)^\top \, d\pi(x,y).$$

Then

$$\lambda_{\max}(S_\pi) \ge e_1^\top S_\pi e_1 = \int (y_1 - x_1)^2 \, d\pi(x,y) \ge \left(\int (y_1 - x_1)\, d\pi(x,y) \right)^2 = L^2.$$

Hence every admissible coupling has operator cost at least L^2.

The translation map

$$T(x_1, x_2) = (x_1 + L, x_2)$$

pushes μ_0 to μ_1 and has covariance

$$S_T = \mathrm{diag}(L^2, 0),$$

so $\mathsf{W}_\gamma(\mu_0, \mu_1)^2 = L^2$.

Now consider the reflected transport

$$R(x_1, x_2) = (x_1 + L, -x_2).$$

It also pushes μ_0 to μ_1 because the second marginal of μ_0 is symmetric on $[-a, a]$. Its displacement is

$$R(x) - x = (L, -2x_2),$$

hence

$$S_R = \mathrm{diag}\left(L^2, \frac{4a^2}{3} \right).$$

By the choice of a,

$$\lambda_{\max}(S_R) = L^2,$$

so the coupling induced by R is again optimal for W_γ.

Consider the corresponding constant-speed geodesic

$$\mu_t = ((1-t)\mathrm{Id} + tR)_{\#}\mu_0.$$

At time $t = \frac{1}{2}$ one gets

$$\mu_{1/2} = \left(x_1 + \frac{L}{2}, 0 \right)_{\#} \mu_0,$$

which is the uniform measure on the horizontal segment

$$\left[\frac{L}{2}, \frac{L}{2} + 1 \right] \times \{0\}.$$

Therefore $\mu_{1/2}$ is singular with respect to Lebesgue measure, hence also singular with respect to ν, and thus

$$\mathrm{Ent}_\nu(\mu_{1/2}) = +\infty.$$

So along this optimal W_γ-geodesic the endpoints have finite entropy while the midpoint has infinite entropy, which rules out geodesic convexity. □

The next remark explains why Proposition H.9 does not contradict Theorem H.8.

Remark H.10 (Why there is no contradiction)**.** *Theorem H.8 is an existence statement: for each pair of endpoints, it guarantees at least one W_γ-geodesic along which the entropy inequality holds. Proposition H.9 proves that in the operator geometry there also exist optimal geodesics along which entropy blows up at an interior time. Thus the counterexample rules out convexity along every geodesic, but not the existence of a selected geodesic satisfying the estimate.*

The following remark explains the scope of the entropy results for the Schatten family.

Remark H.11 (Schatten norms)**.** *The full theorem applies to the trace case $p = 1$, where one may choose $\mathcal{K}_\gamma = \{\mathrm{Id}\}$ and recover the classical W_2 displacement-convexity theory. For every Schatten geometry with $1 < p \le \infty$, the natural representing sets contain singular matrices, so the uniformly elliptic argument above does not apply directly. Theorem H.8 nevertheless yields a weak geodesic-convexity statement as soon as $\mathcal{K}_\gamma$ contains one positive definite matrix, which is the case for the standard Schatten choices.*

Metric slope of relative entropy. Theorem 5.1 does not directly apply to entropy: atomic measures have infinite entropy and are dense in $\mathcal{P}_2$, so its finite-neighborhood hypothesis fails. The following one-sided argument provides the slope identity needed for Gaussian states and smooth local perturbations, also for root-norm gauges.

Proposition H.12 (Entropy slope at a regular relative density)**.** *Let γ be a root-norm gauge and let $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ have strictly positive C^1 densities. Assume that $\ell = \log(d\alpha/d\beta) \in C^2$ has bounded Hessian. Then*

$$|\partial \mathrm{KL}(\cdot \mid \beta)|_\gamma(\alpha) = \mathcal{N}_\alpha^*(\nabla \ell).$$

Proof. The Hessian bound implies quadratic growth of ℓ and linear growth of $g = \nabla \ell$, so the entropy is finite and $g \in L^2(\alpha)$. The supporting inequality for relative entropy gives

$$\mathrm{KL}(\nu \mid \beta) - \mathrm{KL}(\alpha \mid \beta) \ge \int \ell \, d(\nu - \alpha).$$

It follows either by convexity or from the nonnegativity of $\mathrm{KL}(\nu \mid \alpha)$; it is trivial when the left side is infinite. For any coupling π of α, ν, Taylor's lower bound therefore gives

$$\mathrm{KL}(\nu \mid \beta) - \mathrm{KL}(\alpha \mid \beta) \ge \int g(x) \cdot (y - x) \, d\pi - \frac{M}{2} \int |y - x|^2 \, d\pi, \qquad M = \sup_x \|D^2 \ell(x)\|_{\mathrm{op}}.$$

Choose $c_\gamma > 0$ with $c_\gamma \operatorname{tr}(S) \le \gamma(S)$, as provided by Corollary E.6. Apply the inequality to an optimal spectral coupling. Conditional Jensen and monotonicity of a root-norm gauge give

$$\mathrm{KL}(\alpha \mid \beta) - \mathrm{KL}(\nu \mid \beta) \le \mathcal{N}_\alpha^*(g) \mathsf{W}_\gamma(\alpha, \nu) + \frac{M}{2c_\gamma} \mathsf{W}_\gamma(\alpha, \nu)^2.$$

This proves the slope upper bound.

For $w \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$, the map $\mathrm{Id} + hw$ is a diffeomorphism for small $|h|$. Change of variables in the entropy and integration by parts on the compact support give

$$\frac{d}{dh} \mathrm{KL}((\mathrm{Id} + hw)_\sharp \alpha \mid \beta)\Big|_{h=0} = -\int (\operatorname{div} w + \nabla \log \beta \cdot w) \, d\alpha = \int g \cdot w \, d\alpha,$$

where β inside the logarithm denotes its density. Together with $\mathsf{W}_\gamma(\alpha, (\mathrm{Id} + hw)_\sharp \alpha) \le |h| \mathcal{N}_\alpha(w)$, this proves the opposite slope bound by duality and density of C_c^∞ in $L^2(\alpha)$. □

H.5 Relative-entropy flows over Gaussians

The W_2-gradient flow of the relative entropy $\mathrm{Ent}_\nu(\mu) = \mathrm{KL}(\mu \,|\, \nu)$ is the Langevin/Fokker–Planck equation, whose stochastic single-particle realization is the overdamped Langevin diffusion. This makes it a basic tool for sampling. We now describe what happens for the spectral Schatten geometries. For $p > 1$, the resulting evolution is still explicit on Gaussian measures, but the PDE is generally nonlocal: its velocity depends on global statistics of the current law. In general, a stochastic representation would therefore need law-dependent coefficients, rather than a fixed autonomous single-particle diffusion. In dimension one all these geometries reduce to W_2.

Proposition H.13 (Gaussian KL flow for Schatten geometries)**.** *Let $\nu = \mathcal{N}(\bar{m}, \bar{\Sigma})$ with $\bar{\Sigma} \succ 0$, and consider $\mathcal{F}(\mu) = \mathrm{KL}(\mu \,|\, \nu)$ with the canonical Schatten selector of Corollary A.6. Given $\mu_0 = \mathcal{N}(m_0, \Sigma_0)$ with $\Sigma_0 \succ 0$, any locally absolutely continuous solution (m_t, Σ_t) of the equations below, with these initial data and $\Sigma_t \succ 0$, defines a Gaussian solution $\mu_t = \mathcal{N}(m_t, \Sigma_t)$ of the normalized continuity equation. On every compact time interval on which $\Sigma_t \succ 0$, this solution is also a curve of maximal slope for $\mathrm{KL}(\cdot \mid \nu)$ in the ambient W_γ metric. Conversely, every sufficiently regular Gaussian solution with this selector satisfies these equations. This is a conditional reduction, not an assertion of existence or uniqueness through changes in the rank of the force covariance. Define the classical $p = 1$ Gaussian velocity by $\dot{m}_t^{(1)} := -\bar{\Sigma}^{-1}(m_t - \bar{m})$ and $\dot{\Sigma}_t^{(1)} := 2I - \bar{\Sigma}^{-1}\Sigma_t - \Sigma_t\bar{\Sigma}^{-1}$. Then the trace geometry $p = 1$ satisfies $\dot{m}_t = \dot{m}_t^{(1)}$ and $\dot{\Sigma}_t = \dot{\Sigma}_t^{(1)}$.*

For a general Schatten geometry $p \in [1, \infty]$, set $q := 2p/(2p-1)$, with the convention $q = 1$ when $p = \infty$, and define

$$F_t := (\bar{\Sigma}^{-1} - \Sigma_t^{-1})\Sigma_t(\bar{\Sigma}^{-1} - \Sigma_t^{-1}) + \bar{\Sigma}^{-1}(m_t - \bar{m})(m_t - \bar{m})^\top\bar{\Sigma}^{-1}.$$

If $F_t \neq 0$, define

$$\mathcal{N}_t^q := \mathrm{tr}(F_t^{q/2}), \qquad P_{p,t} := \mathcal{N}_t^{2-q}F_t^{q/2-1}.$$

For $1 < p \leq \infty$, the negative power is taken on $\mathrm{range}(F_t)$ and set to zero on its kernel; for $p = 1$, take $P_{1,t} = I$. If $F_t = 0$, the velocity and both right-hand sides below are set to zero. Then the Schatten-p Gaussian KL flow is

$$\dot{m}_t = P_{p,t}\dot{m}_t^{(1)}, \qquad \dot{\Sigma}_t = P_{p,t}(I - \bar{\Sigma}^{-1}\Sigma_t) + (I - \Sigma_t\bar{\Sigma}^{-1})P_{p,t}.$$

If $\bar{\Sigma}^{-1}$ and Σ_t commute, the covariance equation reduces to

$$\dot{\Sigma}_t = \frac{1}{2}\left(P_{p,t}\dot{\Sigma}_t^{(1)} + \dot{\Sigma}_t^{(1)}P_{p,t}\right).$$

In particular, for $p = 1$, one has $q = 2$, $P_{1,t} = I$, and the classical Langevin covariance equation is recovered. For $p = \infty$, one has

$$P_{\infty,t} = \mathrm{tr}(F_t^{1/2})F_t^{\dagger/2}.$$

Proof. For $\mu_t = \mathcal{N}(m_t, \Sigma_t)$ and $\nu = \mathcal{N}(\bar{m}, \bar{\Sigma})$, the first variation of $\mathrm{KL}(\mu \,|\, \nu)$ has Wasserstein gradient

$$g_t(x) = \nabla_x \log \frac{d\mu_t}{d\nu}(x) = (\bar{\Sigma}^{-1} - \Sigma_t^{-1})(x - m_t) + \bar{\Sigma}^{-1}(m_t - \bar{m}).$$

Thus g_t is affine. Set $H_t := \bar{\Sigma}^{-1} - \Sigma_t^{-1}$ and $b_t := \bar{\Sigma}^{-1}(m_t - \bar{m})$, so that $g_t(x) = H_t(x - m_t) + b_t$. Since $\int (x - m_t)\, d\mu_t(x) = 0$ and $\int (x - m_t)(x - m_t)^\top\, d\mu_t(x) = \Sigma_t$, the covariance matrix of g_t is exactly

$$\int g_t(x)g_t(x)^\top\, d\mu_t(x) = H_t\Sigma_t H_t + b_t b_t^\top = F_t.$$

For the Schatten geometry associated with p, the Measure-LMO gives the velocity

$$v_t(x) = -P_{p,t}g_t(x), \qquad P_{p,t} = \mathcal{N}_t^{2-q}F_t^{q/2-1}, \qquad \mathcal{N}_t^q = \mathrm{tr}(F_t^{q/2}).$$

This formula is the Schatten duality map applied to the vector field g_t. When $F_t = 0$, one has $g_t = 0$ μ_t-a.e., and the flow is stationary. Because g_t is affine and $P_{p,t}$ is independent of x, the selected velocity v_t is affine in x. The continuity equation for this prescribed affine velocity transports Gaussian laws to Gaussian laws. The moment equations computed below therefore verify the claimed Gaussian solution; they do not establish uniqueness among all solutions of the nonlinear PDE.

We now compute the induced ODE on mean and covariance. Since $v_t(x) = -P_{p,t}H_t(x - m_t) - P_{p,t}b_t$, the mean satisfies

$$\dot{m}_t = \int v_t(x)\, d\mu_t(x) = -P_{p,t}b_t.$$

But $b_t = \bar{\Sigma}^{-1}(m_t - \bar{m})$, hence $-b_t = \dot{m}_t^{(1)}$. Therefore $\dot{m}_t = P_{p,t}\dot{m}_t^{(1)}$.

For the covariance, using $\dot{\Sigma}_t = \int \big(v_t(x)(x - m_t)^\top + (x - m_t)v_t(x)^\top\big)\, d\mu_t(x)$, we obtain

$$\dot{\Sigma}_t = -P_{p,t}H_t\Sigma_t - \Sigma_t H_t P_{p,t}.$$

On the other hand,

$$\dot{\Sigma}_t^{(1)} = 2I - \bar{\Sigma}^{-1}\Sigma_t - \Sigma_t\bar{\Sigma}^{-1}.$$

Since $H_t = \bar{\Sigma}^{-1} - \Sigma_t^{-1}$ and both $\bar{\Sigma}^{-1}$ and Σ_t^{-1} are symmetric,

$$-H_t\Sigma_t = I - \bar{\Sigma}^{-1}\Sigma_t, \qquad -\Sigma_t H_t = I - \Sigma_t\bar{\Sigma}^{-1}.$$

Therefore

$$-P_{p,t}H_t\Sigma_t - \Sigma_t H_t P_{p,t} = P_{p,t}(I - \bar{\Sigma}^{-1}\Sigma_t) + (I - \Sigma_t\bar{\Sigma}^{-1})P_{p,t}.$$

This is the intrinsic covariance equation stated above. Whenever $\bar{\Sigma}^{-1}\Sigma_t = \Sigma_t\bar{\Sigma}^{-1}$, this reduces to the displayed symmetrized formula because $\dot{\Sigma}_t^{(1)} = 2(I - \bar{\Sigma}^{-1}\Sigma_t)$.

Finally, for $p = 1$ one has $q = 2$, hence $\mathcal{N}_t^{2-q} = 1$ and $F_t^{q/2-1} = I$, so $P_{1,t} = I$. The equations therefore reduce to $\dot{m}_t = -\bar{\Sigma}^{-1}(m_t - \bar{m})$ and $\dot{\Sigma}_t = 2I - \bar{\Sigma}^{-1}\Sigma_t - \Sigma_t\bar{\Sigma}^{-1}$, which is the Gaussian restriction of the classical Langevin/Fokker–Planck flow. For $p = \infty$, $q = 1$, giving $P_{\infty,t} = \mathrm{tr}(F_t^{1/2})F_t^{\dagger/2}$.

To verify the metric interpretation, work on a compact time interval with $\Sigma_t \succ 0$. The moments and Σ_t^{-1} are bounded there, so the force norm and the selected velocity norm are bounded. For an affine field $v_t(x) = A_t(x - m_t) + a_t$, its $L^2(\mu_t)$ norm controls A_t, a_t on such an interval. Thus the affine coefficients are locally integrable, even if they jump at a force-rank change, and the Gaussian continuity equation is valid. The Gaussian entropy is a smooth function of (m, Σ); its derivative is $\int g_t \cdot v_t \, d\mu_t = -\mathcal{N}_{\mu_t}^*(g_t)^2$ almost everywhere. Proposition H.12 identifies the ambient slope because the log density ratio is quadratic. BB bounds the metric speed by $\mathcal{N}_{\mu_t}(v_t) = \mathcal{N}_{\mu_t}^*(g_t)$. The chain rule and slope inequality give the reverse speed bound, exactly as in the proof of Theorem 5.1. The energy is absolutely continuous, hence the integrated energy identity follows. No existence or uniqueness claim for the moment ODE is needed for this verification. □

I P-L and Poincaré Inequalities

The preceding section shows that geodesic convexity is useful but sometimes too rigid for Spectral Wasserstein geometries. A more direct route to convergence rates is a Polyak–Łojasiewicz (P-L) inequality: a first-order estimate requiring the metric slope to control the energy gap.

I.1 P-L inequalities

Let $\mathcal{F}_\star := \inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu) > -\infty$. We use the descending metric slope defined in Section 5. Whenever its identification with the dual force norm holds, as under Theorem 5.1, it can be computed from $g_\mu = \nabla\delta\mathcal{F}/\delta\mu$ by

$$|\partial\mathcal{F}|_\gamma(\mu)^2 = \sup_{\mathcal{N}_\mu(v)\le 1} \left(\int g_\mu(x)\cdot v(x)\, d\mu(x)\right)^2.$$

Definition I.1 (W_γ–P-L inequality)**.** *We say that $\mathcal{F}$ satisfies the W_γ–P-L inequality with constant $\kappa > 0$ if*

$$|\partial\mathcal{F}|_\gamma(\mu)^2 \ge 2\kappa\big(\mathcal{F}(\mu) - \mathcal{F}_\star\big)$$

for every μ in the domain of $\mathcal{F}$.

Proposition I.2 (W_γ–P-L convergence)**.** *Assume that $\mathcal{F}$ satisfies Definition I.1 and is lower semicontinuous for W_γ convergence. Let $(\mu_t)_{t\ge0}$ be a global, locally absolutely continuous W_γ gradient flow with $\mathcal{F}(\mu_0) < +\infty$. Assume that $t \mapsto \mathcal{F}(\mu_t)$ is locally absolutely continuous and satisfies the energy dissipation identity*

$$\frac{d}{dt}\mathcal{F}(\mu_t) = -|\dot{\mu}_t|_\gamma^2 = -|\partial\mathcal{F}|_\gamma(\mu_t)^2 \qquad \textit{for a.e. } t > 0.$$

Set $E(t) = \mathcal{F}(\mu_t) - \mathcal{F}_\star$. Then, for $0 \le s \le t$,

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

and

$$\mathrm{Length}_{\mathsf{W}_\gamma}\big((\mu_r)_{r\in[s,t]}\big) \le \sqrt{\frac{2}{\kappa}}\,\big(\sqrt{E(s)} - \sqrt{E(t)}\big).$$

Consequently, μ_t converges to some $\mu_\infty \in \operatorname{argmin}\mathcal{F}$, and

$$\mathsf{W}_\gamma(\mu_t,\mu_\infty) \le \sqrt{\frac{2}{\kappa}}\sqrt{E(t)} \le \sqrt{\frac{2E(0)}{\kappa}}\,e^{-\kappa t}.$$

In particular,

$$\mathrm{dist}_{\mathsf{W}_\gamma}\big(\mu_t, \operatorname{argmin}\mathcal{F}\big) \le \sqrt{\frac{2E(0)}{\kappa}}\,e^{-\kappa t}.$$

Proof. The energy identity and the P-L inequality give $E'(t) \le -2\kappa E(t)$, hence the exponential bound by Gronwall's lemma. On the set where $E(t) > 0$, the same two identities give

$$|\dot\mu_t|_\gamma = |\partial\mathcal{F}|_\gamma(\mu_t) \le \frac{|\partial\mathcal{F}|_\gamma(\mu_t)^2}{\sqrt{2\kappa E(t)}} = \frac{-E'(t)}{\sqrt{2\kappa E(t)}}.$$

Integrating from s to t gives the length estimate; intervals on which $E = 0$ are stationary by the energy identity. The tail length is finite, so (μ_t) is Cauchy. Proposition 3.2 and completeness of W_2 imply completeness of W_γ, hence $\mu_t \to \mu_\infty$. Lower semicontinuity gives $\mathcal{F}(\mu_\infty) = \mathcal{F}_\star$, and the last bound follows by sending the final time to $+\infty$ and using that distance is bounded by metric length. □

The following comparison explains why P-L estimates transfer more readily than geodesic convexity: only the tangent norms and their dual slopes enter.

Proposition I.3 (Transfer from W_2 to W_γ)**.** *Assume that*

$$\gamma(S) \le C_\gamma \operatorname{tr}(S) \qquad \textit{for every } S \succeq 0.$$

If $\mathcal{F}$ satisfies the W_2–P-L inequality with constant κ, then it satisfies the W_γ–P-L inequality with constant κ/C_γ.

Proof. The static comparison gives $\mathsf{W}_\gamma \le \sqrt{C_\gamma}\mathsf{W}_2$. Both distances induce the same topology, so the definitions of descending slope yield $|\partial\mathcal{F}|_\gamma \ge C_\gamma^{-1/2}|\partial\mathcal{F}|_2$. Combining this with the W_2–P-L inequality proves the claim, without a differentiability assumption on $\mathcal{F}$. □

The normalization of the gauge matters. Replacing γ by $a\gamma$, $a > 0$, multiplies squared distances and squared tangent norms by a, while dividing squared slopes, P-L constants, and the Poincaré constants below by a. All comparisons in this section use the standard Schatten normalization $\gamma_p(aa^\top) = |a|^2$.

For the Schatten gauges $\gamma_p(S) = \|S\|_{S_p}$ with $p \ge 1$, including the Muon endpoint $p = \infty$, one has $\gamma_p(S) \le \operatorname{tr}(S)$ on $\mathbb{S}^d_+$, so $C_\gamma = 1$. Thus every W_2–P-L inequality transfers to these spectral geometries with no loss of constant. A basic instance is the relative entropy $\mathcal{F}(\alpha) = \mathrm{KL}(\alpha\,|\,\beta)$: if $d\beta = Z^{-1}e^{-V}dx$ and $\nabla^2 V \succeq \lambda I$ for some $\lambda > 0$, then the classical log-Sobolev inequality gives the λ–P-L inequality in W_2 [Bakry et al., 2014, Villani, 2009]. Hence the same functional is also λ–P-L for the Muon geometry. Thus any global Muon entropy flow satisfying the energy identity above inherits at least the same guaranteed exponential rate in function value as the W_2 flow. This is a comparison of certified rates, not a pointwise ordering of the two trajectories.

I.2 Spectral Poincaré inequalities

Throughout this subsection, assume that $\beta \in \mathcal{P}_2(\mathbb{R}^d)$ has a strictly positive C^1 density. Test functions are taken in the weighted Sobolev space $H^1(\beta)$, the completion of smooth compactly supported functions for the norm $\big(\int (u^2 + |\nabla u|^2)\,d\beta\big)^{1/2}$. These assumptions make the weak gradient well defined and the smooth and Sobolev infima below agree.

The comparison above gives a simple way to transfer a W_2–P-L estimate, but it does not identify the sharp W_γ–P-L constant in general. Deciding whether a geometry such as Muon yields a strictly

better global constant is difficult because its nonlinear slope must be controlled over the entire energy landscape. We therefore study a small-amplitude expansion at a minimizer, measuring a variational local dissipation gap in the spectral geometry. The leading evolution need not be linear: a spectral duality map is generally nonlinear even though it is one-homogeneous. For concreteness, consider the relative entropy

$$\mathcal{F}(\alpha) = \mathrm{KL}(\alpha \,|\, \beta).$$

Its W_2 gradient flow is the Fokker–Planck, or drift–diffusion, process with invariant law β. Its variational local gap in the W_γ geometry is a γ-Poincaré constant, generalizing the classical Poincaré constant of β.

Let $u = \varphi - \int \varphi \, d\beta$ with $\varphi \in C_c^\infty(\mathbb{R}^d)$ and set $\alpha_\varepsilon = (1 + \varepsilon u)\beta$ for sufficiently small $|\varepsilon|$. Then

$$\mathrm{KL}(\alpha_\varepsilon \,|\, \beta) = \frac{\varepsilon^2}{2} \operatorname{Var}_\beta(u) + O(\varepsilon^3), \qquad \nabla \log \frac{d\alpha_\varepsilon}{d\beta} = \frac{\varepsilon \nabla u}{1 + \varepsilon u} = \varepsilon \nabla u + O_{L^2(\beta)}(\varepsilon^2).$$

Proposition H.12 applies because $\log(1 + \varepsilon u)$ has bounded Hessian. These expansions therefore give

$$|\partial \mathcal{F}|_\gamma(\alpha_\varepsilon)^2 = \mathcal{N}_{\alpha_\varepsilon}^* \left(\nabla \log \frac{d\alpha_\varepsilon}{d\beta} \right)^2 = \varepsilon^2 \mathcal{E}_\gamma^\beta(u) + o(\varepsilon^2),$$

where the spectral Dirichlet energy is

$$\mathcal{E}_\gamma^\beta(u) \coloneqq \sup_{\mathcal{N}_\beta(v) \le 1} \left(\int \nabla u \cdot v \, d\beta \right)^2.$$

Indeed, the pairing is exactly $\int \nabla \log(d\alpha_\varepsilon / d\beta) \cdot v \, d\alpha_\varepsilon = \varepsilon \int \nabla u \cdot v \, d\beta$. Writing $\delta = |\varepsilon| \|u\|_\infty < 1$, monotonicity and homogeneity give $(1 - \delta)\mathcal{N}_\beta(v)^2 \le \mathcal{N}_{\alpha_\varepsilon}(v)^2 \le (1 + \delta)\mathcal{N}_\beta(v)^2$. Duality then bounds the squared slope between $\varepsilon^2 \mathcal{E}_\gamma^\beta(u)/(1 + \delta)$ and $\varepsilon^2 \mathcal{E}_\gamma^\beta(u)/(1 - \delta)$, justifying the expansion despite the changing reference measure. Thus a κ–P-L inequality implies $\kappa \operatorname{Var}_\beta(u) \le \mathcal{E}_\gamma^\beta(u)$, first for these tests and then on $H^1(\beta)$ by density and norm comparison. The converse is not a global P-L theorem, nor a uniform local P-L statement without specifying a neighborhood and uniform remainder bounds. The energy $\mathcal{E}_\gamma^\beta$ is convex and two-homogeneous but is generally not quadratic. We therefore use "Dirichlet energy" rather than a bilinear Dirichlet form. For the canonical Schatten selection ($p \ge 1$), the formal leading perturbation equation is

$$\partial_t u = -\beta^{-1} \operatorname{div}\big(\beta \mathcal{J}_\gamma^\beta(\nabla u)\big), \qquad \frac{1}{2} \frac{d}{dt} \int u^2 \, d\beta = -\mathcal{E}_\gamma^\beta(u),$$

where β in the PDE denotes its density and the identity holds for sufficiently regular solutions. This explains the dissipation meaning of the local constant without asserting a linear generator or existence of the nonlinear entropy flow.

Definition I.4 (γ-Poincaré constant)**.** *The γ-Poincaré constant of β is*

$$\lambda_{\mathrm{P}}^\gamma(\beta) \coloneqq \inf_{\operatorname{Var}_\beta(u) > 0} \frac{\mathcal{E}_\gamma^\beta(u)}{\operatorname{Var}_\beta(u)}.$$

Equivalently,

$$\lambda_{\mathrm{P}}^\gamma(\beta) \operatorname{Var}_\beta(u) \le \mathcal{E}_\gamma^\beta(u)$$

for all smooth u.

Equivalently, one may take the infimum over the corresponding Sobolev form domain; the finite-dimensional norm comparisons make the two formulations identical by density.

Proposition I.5 (Robust Poincaré representation)**.** *Let $\mathcal{K}_\gamma \subset \mathbb{S}_+^d$ be a compact convex representing set for γ, and assume that $\mathcal{K}_\gamma \cap \mathbb{S}_{++}^d \neq \emptyset$. Then, for every smooth u,*

$$\mathcal{E}_\gamma^\beta(u) = \inf_{Q \in \mathcal{K}_\gamma \cap \mathbb{S}_{++}^d} \int \nabla u(x)^\top Q^{-1} \nabla u(x) \, d\beta(x).$$

The infimum need not be attained; a singular optimizer is understood through the lower-semicontinuous limit from $\mathcal{K}_\gamma \cap \mathbb{S}_{++}^d$. Consequently,

$$\lambda_{\mathrm{P}}^\gamma(\beta) = \inf_{Q \in \mathcal{K}_\gamma \cap \mathbb{S}_{++}^d} \lambda_{\mathrm{P}}^Q(\beta), \qquad \lambda_{\mathrm{P}}^Q(\beta) \coloneqq \inf_{\operatorname{Var}_\beta(u) > 0} \frac{\int \nabla u^\top Q^{-1} \nabla u \, d\beta}{\operatorname{Var}_\beta(u)}.$$

Proof. Let $\ell_u(v) = \int \nabla u \cdot v \, d\beta$ and $\mathcal{N}_\beta(v)^2 = \gamma(\int vv^\top d\beta)$. Fenchel duality for a squared norm gives

$$\inf_v \left\{ \ell_u(v) + \frac{1}{2}\mathcal{N}_\beta(v)^2 \right\} = -\frac{1}{2} \sup_{\mathcal{N}_\beta(v)\le 1} \ell_u(v)^2 = -\frac{1}{2}\mathcal{E}_\gamma^\beta(u).$$

On the other hand, set $\Phi_Q(v) = \ell_u(v) + \frac{1}{2}\int v^\top Q v \, d\beta$. The support representation of γ and the inverse-trace minimax argument of Proposition A.7 give

$$\begin{aligned}\inf_v \left\{ \ell_u(v) + \frac{1}{2}\mathcal{N}_\beta(v)^2 \right\} &= \inf_v \max_{Q\in\mathcal{K}_\gamma} \Phi_Q(v) = \max_{Q\in\mathcal{K}_\gamma} \inf_v \Phi_Q(v) \\ &= -\frac{1}{2} \inf_{Q\in\mathcal{K}_\gamma\cap\mathbb{S}^d_{++}} \operatorname{tr}\left(Q^{-1} \int \nabla u \nabla u^\top \, d\beta \right).\end{aligned}$$

Here any $Q_0 \in \mathcal{K}_\gamma \cap \mathbb{S}^d_{++}$ provides the coercivity needed for the minimax exchange, and convexity of $\mathcal{K}_\gamma$ lets one approximate every singular Q by $(1-\varepsilon)Q + \varepsilon Q_0 \succ 0$. Comparing the two displays proves the formula for $\mathcal{E}_\gamma^\beta(u)$. Taking the infimum over all u with positive variance gives the formula for $\lambda_{\mathrm{P}}^\gamma(\beta)$; the two infima commute because this is an infimum over the product of admissible functions and matrices. □

Equivalently, $\lambda_{\mathrm{P}}^Q(\beta)$ is the usual Poincaré constant after the linear change of variables $(Q^{1/2})_\#\beta$. Hence the spectral Poincaré constant is the worst anisotropic Poincaré constant over the dual ball $\mathcal{K}_\gamma$. At singular endpoints, this worst case may be approached by positive definite matrices without being attained in $\mathbb{S}^d_{++}$.

For the Schatten family γ_p, $p \ge 1$, this becomes especially explicit. With $p' = p/(p-1)$, using the conventions $p' = \infty$ for $p = 1$ and $p' = 1$ for $p = \infty$, one may take

$$\mathcal{K}_p = \{Q \succeq 0 : \ \|Q\|_{S_{p'}} \le 1\}.$$

Hence

$$\lambda_{\mathrm{P}}^p(\beta) = \inf_{\substack{Q\succ 0 \\ \|Q\|_{S_{p'}}\le 1}} \lambda_{\mathrm{P}}\big((Q^{1/2})_\#\beta\big).$$

For $p = \infty$, this is the Muon formula

$$\lambda_{\mathrm{P}}^\infty(\beta) = \inf_{\substack{Q\succ 0 \\ \operatorname{tr} Q\le 1}} \lambda_{\mathrm{P}}\big((Q^{1/2})_\#\beta\big).$$

Since every admissible Q satisfies $Q \preceq I$, one has $Q^{-1} \succeq I$ and therefore

$$\lambda_{\mathrm{P}}^p(\beta) \ge \lambda_{\mathrm{P}}^1(\beta) \qquad (p \ge 1).$$

Thus the variational local dissipation constant is no smaller than the classical W_2 constant for all Schatten geometries from the trace case to the Muon endpoint. Along solutions for which the preceding perturbation identity is justified, it controls exponential decay of the perturbation's squared $L^2(\beta)$ norm.

The same local computation also gives a simple way to compare examples across the full root-norm Schatten range $p \ge 1/2$ discussed in Appendix E.4. For a smooth test function u, set

$$G_u := \int \nabla u \, \nabla u^\top \, d\beta.$$

For $\gamma_p(S) = \|S\|_{S_p}$, $p \ge 1/2$, the tangent norm is the Schatten S_{2p} norm of the operator $T_v : a \mapsto a \cdot v$. By Schatten duality, the spectral Dirichlet energy is

$$\mathcal{E}_p^\beta(u) = \left\| G_u^{1/2} \right\|_{S_{\theta_p}}^2, \qquad \theta_p = (2p)',$$

with the conventions $\theta_{1/2} = +\infty$ and $\theta_\infty = 1$. Relative to the trace case,

$$\mathcal{E}_p^\beta(u) \ge \mathcal{E}_1^\beta(u) \quad (p > 1), \qquad \mathcal{E}_p^\beta(u) \le \mathcal{E}_1^\beta(u) \quad (1/2 \le p < 1),$$

and, away from $p = 1$, equality holds precisely when G_u has rank at most one. Thus the difference between the geometries is visible exactly through perturbations with several active gradient directions.

Proposition I.6 (Gaussian targets)**.** *Let $\beta = \mathcal{N}(m, \Sigma)$ be a nondegenerate Gaussian measure on $\mathbb{R}^d$, $d \geq 2$. Then, for every $p \geq 1$,*

$$\lambda_{\mathrm{P}}^p(\beta) = \lambda_{\mathrm{P}}^1(\beta) = \frac{1}{\lambda_{\max}(\Sigma)}.$$

For $1/2 \leq p < 1$, the constant is strictly smaller:

$$d^{1-1/p}\lambda_{\mathrm{P}}^1(\beta) \leq \lambda_{\mathrm{P}}^p(\beta) < \lambda_{\mathrm{P}}^1(\beta).$$

Proof. For $p \geq 1$, $\theta_p \leq 2$, so $\left\|G^{1/2}\right\|_{S_{\theta_p}}^2 \geq \operatorname{tr} G$. The classical Gaussian Poincaré inequality gives $\lambda_{\mathrm{P}}^p(\beta) \geq 1/\lambda_{\max}(\Sigma)$. Equality follows by testing with a linear function $u(x) = a^\top x$, where a is an eigenvector of Σ associated with $\lambda_{\max}(\Sigma)$: then $G_u = aa^\top$ is rank one, so all Schatten norms coincide on $G_u^{1/2}$.

Let now $1/2 \leq p < 1$, so $\theta_p > 2$. The Schatten comparison gives

$$\left\|G^{1/2}\right\|_{S_{\theta_p}}^2 \geq d^{1-1/p} \operatorname{tr} G,$$

which proves the lower bound. For the strict upper bound, translate the measure so that $m = 0$, diagonalize Σ, and write $\sigma_1 = \lambda_{\max}(\Sigma)$ and $\sigma_2 > 0$ for another eigenvalue. In these coordinates, take

$$u_\varepsilon(x) = x_1 + \varepsilon(x_2^2 - \sigma_2).$$

Then

$$\operatorname{Var}_\beta(u_\varepsilon) = \sigma_1 + 2\varepsilon^2\sigma_2^2, \qquad G_{u_\varepsilon} = e_1e_1^\top + 4\varepsilon^2\sigma_2\, e_2e_2^\top.$$

Thus $\mathcal{E}_p^\beta(u_\varepsilon) = 1 + o(\varepsilon^2)$ as $\varepsilon \to 0$ for every $\theta_p > 2$ (and it is exactly 1 for small ε when $\theta_p = \infty$). Therefore

$$\frac{\mathcal{E}_p^\beta(u_\varepsilon)}{\operatorname{Var}_\beta(u_\varepsilon)} = \frac{1}{\sigma_1} - \frac{2\sigma_2^2}{\sigma_1^2}\varepsilon^2 + o(\varepsilon^2),$$

which is strictly smaller than $1/\sigma_1$ for all sufficiently small $\varepsilon > 0$. □

Proposition I.7 (Generalized Gaussian targets)**.** *Let $d \geq 2$, let $r > 1$, $r \neq 2$, and let*

$$d\beta_r(x) = Z_{r,d}^{-1} e^{-|x|^r}\, dx.$$

Then the Schatten Poincaré constant is strictly improved for $p > 1$ and strictly decreased for $1/2 \leq p < 1$:

$$\lambda_{\mathrm{P}}^p(\beta_r) > \lambda_{\mathrm{P}}^1(\beta_r) \quad \textit{for } p > 1, \qquad \lambda_{\mathrm{P}}^p(\beta_r) < \lambda_{\mathrm{P}}^1(\beta_r) \quad \textit{for } 1/2 \leq p < 1.$$

Thus, for the family $e^{-|x|^r}$ in dimension $d > 1$, the Gaussian exponent $r = 2$ is the only case with no strict improvement of the local Poincaré rate for $p > 1$.

Proof. The reversible generator is

$$L_r = \Delta - r|x|^{r-2}x \cdot \nabla.$$

Writing $V_r(x) = |x|^r$, the ground-state transform of $-L_r$ has effective potential

$$\frac{1}{4}|\nabla V_r|^2 - \frac{1}{2}\Delta V_r = \frac{r^2}{4}|x|^{2r-2} - \frac{r(r+d-2)}{2}|x|^{r-2},$$

which tends to $+\infty$ as $|x| \to \infty$. The transform is understood at the level of the closed quadratic form. For $1 < r < 2$, its negative singularity at zero is a multiple of $|x|^{-(2-r)}$, which belongs locally to L^a for some $a > d/2$ and $a > 1$. Sobolev estimates make this term infinitesimally form-bounded relative to the Laplacian. For $r \geq 2$ the potential is locally bounded. In both cases, the confining part controls tails and local Rellich compactness gives compact embedding of the form domain. Thus the classical Poincaré problem has compact resolvent and its spectral gap is attained.

We first prove that no first eigenfunction has rank-one gradient covariance. Indeed, if G_u has rank one, then the positive density of β_r implies that ∇u is a.e. parallel to some unit vector a. Hence $u(x) = \psi(a \cdot x)$, and, with $t = a \cdot x$, the eigenvalue equation $-L_r u = \lambda u$ becomes

$$-\psi''(t) + r|x|^{r-2}t\psi'(t) = \lambda\psi(t).$$

Elliptic regularity away from the origin justifies this identity pointwise for $t \neq 0$. Since u is nonconstant, one can choose such a t with $\psi'(t) \neq 0$. Setting $x = ta + z$ and varying $z \perp a$ forces $|ta + z|^{r-2}$ to be independent of $|z|$, which is possible only for $r = 2$. This contradiction proves the claim.

Let $\lambda_1 = \lambda_{\mathrm{P}}^1(\beta_r)$. If $p > 1$, then $\theta_p < 2$ and $\mathcal{E}_p^{\beta_r} \geq \mathcal{E}_1^{\beta_r}$, so $\lambda_{\mathrm{P}}^p(\beta_r) \geq \lambda_1$. Suppose equality holds and take a minimizing sequence (u_n) with zero mean and unit variance. Since

$$\lambda_1 \leq \mathcal{E}_1^{\beta_r}(u_n) \leq \mathcal{E}_p^{\beta_r}(u_n) \longrightarrow \lambda_1,$$

its classical Dirichlet energy also converges to λ_1. Compactness of the resolvent and the spectral decomposition then give, after extraction, convergence in the form norm to a first classical eigenfunction $u_\star$. Moreover, $\mathcal{E}_p^{\beta_r}(u_n) - \mathcal{E}_1^{\beta_r}(u_n) \to 0$, and the continuity of $u \mapsto G_u$ along this convergence gives

$$\left\| G_{u_\star}^{1/2} \right\|_{S_{\theta_p}} = \left\| G_{u_\star}^{1/2} \right\|_{S_2}.$$

For $\theta_p < 2$, equality in the Schatten comparison is possible only when $G_{u_\star}$ has rank one, impossible by the rigidity fact above. Hence the inequality is strict for $p > 1$.

If $1/2 \leq p < 1$, take any first classical eigenfunction u. The previous paragraph shows that G_u has rank at least two. Since now $\theta_p > 2$,

$$\mathcal{E}_p^{\beta_r}(u) = \left\| G_u^{1/2} \right\|_{S_{\theta_p}}^2 < \left\| G_u^{1/2} \right\|_{S_2}^2 = \mathcal{E}_1^{\beta_r}(u) = \lambda_1 \operatorname{Var}_{\beta_r}(u),$$

which gives $\lambda_{\mathrm{P}}^p(\beta_r) < \lambda_{\mathrm{P}}^1(\beta_r)$. □

The restriction $d \geq 2$ is essential: in dimension one, every gradient covariance is scalar, so all the Schatten Poincaré constants coincide.

J Positively 2-Homogeneous Case and a Generalized Unbalanced Geometry

This section continues the two-layer mean-field discussion of Section 5 in the positively two-homogeneous case. The energy factors through a weighted spherical projection, and the action of a homogeneous velocity has a spherical continuity–reaction form. These are energy and action identities, not an isometry between fixed ambient lifts and spherical measures. Let

$$\mathcal{F}(\mu) = R\left(\int_{\mathbb{R}^d} \Phi(x)\, d\mu(x) \right), \qquad \Phi(\lambda x) = \lambda^2 \Phi(x).$$

Here $\Phi : \mathbb{R}^d \to \mathbb{R}^k$ is continuous and R is defined on the relevant moment vectors. The energy identity below also applies to nonsmooth features. For the subsequent force construction we assume $\Phi \in C^1$ and $R \in C^1$; applying it to ReLU requires a separate justification at its kinks. Writing $x = r\omega$ with $\omega \in \mathbb{S}^{d-1}$, define the weighted spherical projection

$$\int_{\mathbb{S}^{d-1}} \psi(\omega)\, d\Pi_2(\mu)(\omega) \coloneqq \int_{\mathbb{R}^d} |x|^2\, \psi\left(\frac{x}{|x|} \right) d\mu(x).$$

The integrand on the right is defined to be zero at $x = 0$. The first point is that two-homogeneous models only depend on this weighted spherical projection.

Proposition J.1 (Energy factorization)**.** *There exists a functional $\overline{f}$ on $\mathcal{M}_+(\mathbb{S}^{d-1})$ such that*

$$\mathcal{F}(\mu) = \overline{f}(\Pi_2(\mu)).$$

Proof. The homogeneity identity $\Phi(r\omega) = r^2 \Phi(\omega)$ immediately gives

$$\int_{\mathbb{R}^d} \Phi(x)\, d\mu(x) = \int_{\mathbb{S}^{d-1}} \Phi(\omega)\, d\Pi_2(\mu)(\omega),$$

which defines $\overline{f}$. □

The second point is that the ambient normalized velocity splits into radial and tangential parts on the sphere. If a vector field v is 1-homogeneous, namely $v(\lambda x) = \lambda v(x)$ for every $\lambda > 0$, then for every $x = r\omega \neq 0$ it can be written uniquely as

$$v(r\omega) = r\big(b(\omega)\omega + \tau(\omega)\big), \qquad \tau(\omega) \in T_\omega \mathbb{S}^{d-1}.$$

Indeed, one defines

$$b(\omega) \coloneqq v(\omega) \cdot \omega, \qquad \tau(\omega) \coloneqq v(\omega) - \big(v(\omega) \cdot \omega\big)\omega,$$

so that $\tau(\omega) \cdot \omega = 0$, and then 1-homogeneity gives

$$v(r\omega) = r\, v(\omega) = r\big(b(\omega)\omega + \tau(\omega)\big).$$

The splitting is orthogonal and unique. In the smooth two-homogeneous model, the first variation is two-homogeneous, so its force is one-homogeneous. Choose the projected Measure-LMO of Proposition 2.4, using a fixed measurable matrix selection $A_g = A(S_g)$, and set the velocity to zero at the origin. This selected velocity is one-homogeneous because it is a spatially constant matrix times the force. The canonical Schatten selector is one such choice. Other minimizing velocities need not be homogeneous when the oracle is nonunique, so this selection requirement cannot be omitted.

The coefficients close on $\nu = \Pi_2(\mu)$: the force depends on μ only through $\int \Phi \, d\nu$, and homogeneity gives

$$S_g = \int_{\mathbb{S}^{d-1}} g_\mu(\omega) g_\mu(\omega)^\top \, d\nu(\omega).$$

Consequently the fixed selection, and hence b and τ, depend only on ν, not on its radial lift.

Proposition J.2 (Projected continuity-reaction equation)**.** *Let (μ_t, v_t) solve the continuity equation on a finite interval, with μ_t continuous in W_2, $\int_0^T \int |v_t|^2 d\mu_t dt < \infty$, and v_t a one-homogeneous Borel field with $v_t(0) = 0$. If $\nu_t = \Pi_2(\mu_t)$, then*

$$\partial_t \nu_t + \mathrm{div}_{\mathbb{S}^{d-1}}(\nu_t \tau_t) = 2 b_t \nu_t.$$

Proof. Test the ambient continuity equation against $\widetilde{\psi}(r\omega) = r^2 \psi(\omega)$. For $r > 0$,

$$\nabla \widetilde{\psi}(r\omega) = r\big(2\psi(\omega)\omega + \nabla_{\mathbb{S}^{d-1}} \psi(\omega)\big).$$

Using $v(r\omega) = r(b(\omega)\omega + \tau(\omega))$ and $\omega \cdot \tau(\omega) = 0$, we obtain

$$\nabla \widetilde{\psi}(r\omega) \cdot v(r\omega) = r^2\big(2b(\omega)\psi(\omega) + \nabla_{\mathbb{S}^{d-1}} \psi(\omega) \cdot \tau(\omega)\big).$$

The weak form of the ambient continuity equation and the definition of Π_2 therefore give

$$\frac{d}{dt} \int_{\mathbb{S}^{d-1}} \psi \, d\nu_t = \int_{\mathbb{S}^{d-1}} \big(2 b_t \psi + \nabla_{\mathbb{S}^{d-1}} \psi \cdot \tau_t\big) \, d\nu_t,$$

which is precisely the weak form of the stated continuity–reaction equation. The test function is C^1 at the origin with zero gradient. Spatial cutoffs justify its quadratic growth: the error is controlled by $\int_0^T \int |x||v_t(x)| \, d\mu_t dt < \infty$, using the uniform second-moment bound and finite kinetic action. The W_2 continuity also implies narrow continuity of ν_t, by uniform integrability of the weighted projection. □

Motivated by Proposition J.2, define for finite nonnegative measures ν_0, ν_1 on $\mathbb{S}^{d-1}$

$$\mathsf{UW}_\gamma(\nu_0, \nu_1)^2 \coloneqq \inf_{(\nu_t, b_t, \tau_t)} \int_0^1 \gamma\left(\int_{\mathbb{S}^{d-1}} (b_t(\omega)\omega + \tau_t(\omega))(b_t(\omega)\omega + \tau_t(\omega))^\top \, d\nu_t(\omega) \right) dt,$$

over narrowly continuous finite-measure curves with the prescribed endpoints and Borel fields b_t, τ_t, with $\tau_t(\omega) \perp \omega$ almost everywhere and $\int_0^1 \int (b_t^2 + |\tau_t|^2) d\nu_t dt < \infty$, under the distributional continuity–reaction constraint

$$\partial_t \nu_t + \mathrm{div}_{\mathbb{S}^{d-1}}(\nu_t \tau_t) = 2 b_t \nu_t.$$

The next proposition identifies the ambient homogeneous action with this spherical unbalanced action.

Proposition J.3 (Ambient action equals spherical action)**.** *For each ambient curve and homogeneous velocity satisfying Proposition J.2 on $[0,T]$, $T>0$, its ambient action equals the action of its spherical projection on the same interval. In particular,*

$$\mathsf{UW}_\gamma(\nu_0,\nu_T)^2 \le T\int_0^T \gamma\Big(\int v_t v_t^\top\, d\mu_t\Big)\,dt.$$

Proof. Write $x=r\omega$ and $v(r\omega)=r(b(\omega)\omega+\tau(\omega))$. By definition of $\Pi_2(\mu)$,

$$\int_{\mathbb{R}^d} v(x)v(x)^\top\, d\mu(x) = \int_{\mathbb{S}^{d-1}} (b(\omega)\omega+\tau(\omega))(b(\omega)\omega+\tau(\omega))^\top\, d\Pi_2(\mu)(\omega).$$

Integrating yields equality of the two actions. Rescaling $[0,T]$ to $[0,1]$ multiplies velocities by T, hence multiplies the total action by T; the definition of UW_γ then gives the inequality. □

This identity does not equate the distances of fixed lifts. For example, $\mu_0=\delta_{e_1}$ and $\mu_1=\frac{1}{2}\delta_0+\frac{1}{2}\delta_{\sqrt{2}e_1}$ both project to δ_{e_1}, while $\mathsf{W}_\gamma(\mu_0,\mu_1)^2=(2-\sqrt{2})\gamma(e_1e_1^\top)>0$. A metric-quotient theorem would require an additional minimization over lifts and a converse lifting result, neither of which is asserted here.

The following remark explains how the trace-norm case collapses to the classical Wasserstein–Fisher–Rao geometry.

Remark J.4. *When $\gamma(S)=\mathrm{tr}(S)$, the integrand becomes*

$$\int_{\mathbb{S}^{d-1}} \big(|\tau_t(\omega)|^2+b_t(\omega)^2\big)\,d\nu_t(\omega).$$

Since the reaction rate is $\alpha_t=2b_t$, the reaction term is $\alpha_t^2/4$. Thus the trace case is the Wasserstein–Fisher–Rao action with this transport/reaction normalization, whose static formulations are developed in Chizat et al. [2018], Liero et al. [2018]. Writing WFR *for this normalization, trace comparison gives*

$$\sqrt{c_\gamma}\,\mathsf{WFR} \le \mathsf{UW}_\gamma \le \sqrt{C_\gamma}\,\mathsf{WFR}.$$

Thus UW_γ is finite and separates measures. Symmetry follows by reversing time, and concatenating curves with durations proportional to the square roots of their actions proves the triangle inequality. Hence it is a metric on finite spherical measures. Whether it admits an analogous static formulation is an interesting open direction.

K Notation

This section collects the main notation used throughout the paper. The first table lists standard background notation, while the second table lists the notation introduced for the Spectral Wasserstein geometry.

Table 2: Classical notation.

Notation	Name	First use
$\mathbb{R}^d$	Euclidean space	Sec. 2
$\mathbb{S}^d$	Symmetric matrices	Sec. 2
$\mathbb{S}^d_+$	PSD matrices	Sec. 2
$\mathbb{S}^d_{++}$	Positive definite matrices	Prop. A.7
$S \preceq T$	Loewner order	Sec. 2
I, Id	Identity matrix/map	Sec. 2
$\mathrm{tr}(S)$	Trace	Def. 2.1
$\mathrm{diag}(a_i)$	Diagonal matrix	Sec. 2
$\lambda(S)$	Eigenvalues of S	Sec. 2
$\|\cdot\|_F$	Frobenius norm	Def. 2.2
$\langle\cdot,\cdot\rangle_F$	Frobenius product	Def. 2.2

Notation	Name	First use
$\|A\|_{S_p}$	Schatten p-norm	Sec. 2
$A^\dagger$	Pseudoinverse	Eq. (6)
$\mathcal{P}_2(\mathbb{R}^d)$	Measures with finite second moment	Def. 2.3
δ_x	Dirac mass	Sec. 2
$T_\#\mu$	Pushforward	Sec. 4
$\Pi(\mu,\nu)$	Couplings	Def. 3.1
$L^2(\mu;\mathbb{R}^d)$	Square-integrable vector fields	Def. 2.3
∇, ∇_x	Euclidean gradient	Sec. 2
div	Divergence	Eq. (3)
$\delta\mathcal{F}/\delta\mu$	First variation	Sec. 2
W_2	Quadratic Wasserstein distance	Sec. 1
$\mathcal{N}(m,\Sigma)$	Gaussian law	Prop. 3.6
Σ	Covariance matrix	Prop. 3.6
$\mathbb{S}^{d-1}$	Unit sphere	App. J
$\nabla_{\mathbb{S}^{d-1}}$, $\mathrm{div}_{\mathbb{S}^{d-1}}$	Spherical gradient/divergence	Prop. J.2

Table 3: Main notation introduced in the paper.

Notation	Name	Introduction
γ	Monotone matrix gauge on $\mathbb{S}^d_+$	Sec. 2
γ_p	Schatten-p gauge	Sec. 2
p'	Conjugate exponent of p	Rem. A.4
$q=(2p)'$	Tangent-space dual exponent	Prop. A.5
$\mathcal{K}_\gamma$	Representing set	Def. 3.3
$\mathcal{P}_\gamma$	Polar set	App. A
$\mathcal{K}_{\gamma_p}$	Schatten representing set	Rem. A.4
$\mathrm{N}(V)$	Matrix tangent norm	Sec. 2
$\mathrm{J}_\gamma(G)$	Matrix LMO	Def. 2.2
$\mathrm{J}_p(G)$	Schatten matrix LMO	Prop. A.5
X	Particle matrix	Sec. 2
μ_X	Empirical particle measure	Sec. 2
$\mathrm{F}(X)$	Particle objective	Sec. 2
$\mathcal{F}(\mu)$	Mean-field objective	Sec. 2
$\mathcal{N}_\mu(v)$	Measure tangent norm	Def. 2.3
$\mathcal{J}^\mu_\gamma(g)$	Measure-LMO	Def. 2.3
$\mathcal{S}_\mu(g)$	Uncentered force second moment	Prop. 2.4
$\mathfrak{N}_\gamma(T)$	Root norm of an operator	Def. E.2
Q^*_μ	Inverse-trace minimizer	Prop. A.7
g_μ	Wasserstein force (before normalization)	Sec. 2
W_γ	Static Spectral Wasserstein distance	Def. 3.1
W^Q	Fixed-metric transport cost	Thm. 3.4
c_γ, C_γ	Trace comparison constants	Prop. 3.2
$\mathsf{W}^{\mathrm{BB}}_\gamma$	Dynamic Spectral Wasserstein distance	Def. 4.1
$\mathcal{A}_\gamma(\mu,m)$	Momentum action	Prop. E.1
$\mathsf{W}^{\mathrm{M}}_\gamma$	Monge-restricted cost	Def. C.1
$\phi(z,x)$	Feature map	Sec. 5
u, v	Outer and inner weights (MLP context)	Sec. 5
H_μ	Mean-field predictor	Sec. 5
$R(H)$	Risk functional	Sec. 5
$\mathrm{MMD}(\mu,\nu)^2$	Squared MMD	Sec. 5
(Q,K,V,O)	Attention weights	Sec. 5
N_{eff}	Effective tokens	Sec. 5
C_ρ	Data covariance	App. F

Notation	Name	Introduction
B_ρ	Teacher cross-moment	App. F
M_μ	Learned linear predictor	App. F
$A_\star$	Teacher matrix	App. F
E_t	Linear prediction error	Prop. F.2
$r_i(t)$	Modal predictor coefficient	Prop. F.3
$s_i(t)$	Modal variance	Prop. F.3
β_i	Teacher mode	Prop. F.3
$\mathcal{N}_t$	Modal normalization	Prop. F.3
γ_{blk}	Sum block gauge (separate LMOs)	App. B
$\gamma_{\max}$	Maximum block gauge (common polar clock)	App. B
L_γ	Matrix smoothness constant	Eq. (11)
$r_n = \min(n, d)$	Matrix rank bound	App. B.1
$\mathrm{I}_p(V)$	Momentum constitutive map	Eq. (8)
$F_h(\mu)$	Linear measure functional	Thm. H.4
κ	Convexity or P-L constant (by context)	Def. H.2
$\mathrm{Ent}_\nu(\mu)$	Relative entropy	App. H
$\mathrm{KL}(\mu \,\|\, \nu)$	Relative entropy / KL divergence	Sec. H.5
$\lvert\partial\mathcal{F}\rvert_\gamma$	Metric slope of $\mathcal{F}$	Sec. 5
$\mathcal{E}_\gamma^\beta$	Spectral Dirichlet energy	App. I
$\lambda_{\mathrm{P}}^\gamma(\beta)$	Spectral Poincaré constant	App. I
$\theta_p = (2p)'$	Dirichlet-energy Schatten exponent	App. I
Φ	Two-homogeneous feature map	App. J
$\Pi_2(\mu)$	Spherical projection	App. J
UW_γ	Unbalanced spherical action	App. J

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sections 1–4 state the geometric contributions, and Section 5 makes the conditional nature of the gradient-flow theorem explicit. The experiments illustrate geometry rather than establish a runtime or convergence-rate advantage.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The scope paragraph in Section 1, the discussion after Theorem 5.1, and Appendices B and G.1 state the regularity, practical-optimizer, and numerical limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The statements specify their hypotheses, with proofs in the main text or appendices. Proposition 2.5 explicitly assumes a transport chain rule; Theorem H.4 uses integrated pointwise convexity. Theorem 5.1 identifies regular solutions under a strong transport expansion and does not claim global existence or uniqueness.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix G.1 records nominal clocks, clipping, smoothing, and validation checks. The public repository linked in Section 1 supplies figure notebooks, generators, and reproduction instructions; archived plots do not reconstruct unrecorded effective clocks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository `https://github.com/gpeyre/muon-dynamics` contains the computational toolbox, illustrative notebooks, synthetic-data generators, and scripts for reproducing paper figures. Its documentation distinguishes reduced smoke tests from full experimental runs.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 specifies the models and interpretation of the plots; Appendix G.1 and the figure notebooks provide the numerical settings. These are synthetic demonstrations, not held-out generalization benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Attention results use five paired trials with new teacher and student initializations, but no general statistical-significance claim is made. The MLP and MMD trajectories are individual illustrative runs, so uncertainty is not quantified for every figure.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The repository specifies software dependencies and execution commands, but complete hardware, peak-memory, and wall-clock records for all archived full experiments are not available. The paper does not make computational-efficiency claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The author has reviewed the NeurIPS Code of Ethics and confirmed compliance. The reported study is mathematical and uses synthetic examples, without human-subject experiments or personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: This is foundational work on transport geometry and idealized optimization, without a deployed application. A dedicated assessment of societal benefits and risks is not included; no empirical claim of safer or more resource-efficient deployment is made.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The work does not release pretrained language or image-generation models, scraped datasets, or other high-risk data or models. Its released artifacts are mathematical software and synthetic examples.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The original toolbox code is released under the MIT license, and the repository identifies software dependencies and inventories illustration assets. Provenance and reuse permissions for some legacy illustration assets remain unverified; no unsupported attribution or blanket license is claimed for them.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The repository documents the numerical toolbox, its API, illustrative notebooks, and paper-figure reproduction workflow. The licensing limitation is disclosed separately above.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research contains no crowdsourcing or human-subject experiments; the examples use synthetic data and teacher models.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: There are no human participants or personal data in the reported experiments, so no human-subject approval is applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM assistance was used in developing and checking mathematical arguments and code, as well as in editing and formatting. The author retains responsibility for all claims, proofs, references, and experimental results; no LLM is a component of the proposed optimizer or numerical models.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.