RANDOM SUBSPACE DESCENT FOR PROBABILITY FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wasserstein gradient flows connect sampling and generative modeling with optimization over probability measures. Kernel-based flows such as MMD descent, Sobolev descent, and Laplacian Adjusted Wasserstein Gradient Descent (LAWGD) yield interacting particle algorithms, but their descent directions can be expensive to compute. Random features and slicing reduce this cost, but many slicing algorithms ignore the geometry of the flow and need not decrease the original objective. We approximate the flows by solving their defining variational problem over a random span of ridge potentials. This preserves the original objective and geometry, reduces the update to a Gram solve, and yields an exact decomposition of captured and lost dissipation. Particle experiments demonstrate the advantage of projection over naive slicing.

1 INTRODUCTION

Kernel-based Wasserstein gradient flows give attractive dynamics for sampling and distribution matching, but their descent directions are expensive to compute. MMD descent (Arbel et al., 2019) sums a kernel over all particles and target samples, Sobolev descent (Mroueh et al., 2019) requires a dense linear solve in feature space, and LAWGD (Chewi et al., 2020a) requires inverting the Langevin generator. Each flow is determined by two choices (Jordan et al., 1998; Otto, 2001): an objective, which specifies the discrepancy to reduce, and a geometry, which specifies which descent directions to emphasize. Together, they define the descent direction as the solution of a variational problem over all potentials, and solving this problem exactly is what makes these flows expensive.

Random features (Rahimi & Recht, 2007) and slicing (Liutkus et al., 2019; Bonet et al., 2022) are the standard ways to reduce this cost, but both change the problem being solved. Random features replace the objective with a finite-feature approximation, and slicing typically defines a new objective (Deshpande et al., 2018) or a new geometry (Bonet et al., 2022). The resulting flow need not decrease the original objective, and properties such as preconditioning need not survive the approximation.

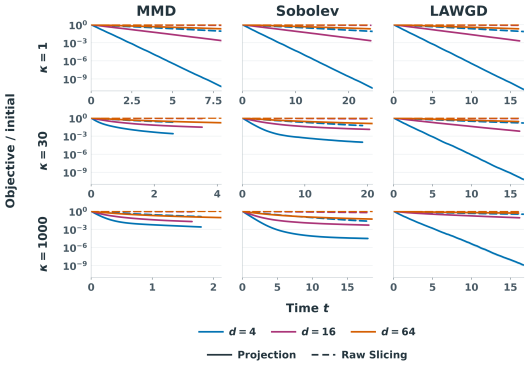

Figure 1: Additional dissipation captured by projection over slicing in an ill-conditioned Gaussian target example.

Our key idea is to treat random features and slices as trial directions, not as new objectives or geometries. We sample a small span of ridge potentials and solve the original flow's variational problem exactly within that span. Thus every population update is the native projection of the full descent direction, guaranteeing descent and giving an exact decomposition of captured versus lost dissipation. This draws inspiration from the works on random subspace descent (Kozak et al., 2021; Gower et al., 2019) in Euclidean space. Figure 1 displays the advantage of projection over slicing in a simple Gaussian example. Full details of this plot are given in Section 5.1.

Because the projections live in L^2, no finite random subspace captures a fixed fraction of every direction in expectation. We therefore express expected progress through the sampling law's average

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

projector and bound it using a random-feature norm. This yields quantitative rates under conditions verifiable in synthetic examples.

Contributions. Our main contribution is a random subspace formulation that preserves the variational structure of MMD, Sobolev, and LAWGD flows. This yields a projected dissipation identity that quantifies the quality of a random-subspace flow. We then analyze random sampling through the average projector, and show that we can recover variants of randomized subspace descent exactly on translation families.

Related work. Many methods change the geometry of a WGF to improve its conditioning, including Stein variational gradient descent (Liu & Wang, 2016), mirror and preconditioned descent (Hsieh et al., 2018; Zhang et al., 2020; Chewi et al., 2020b; Bonet et al., 2024), Sobolev descent (Mroueh et al., 2019), and LAWGD (Chewi et al., 2020a), but they compute the full direction. Others fit the direction in a function class (Dong et al., 2023; Wang et al., 2024) or restrict the flow to a parametric family (Chen & Li, 2020; Jin et al., 2025). Our classes are small random linear spaces, so the fit is a Gram solve and the effect of the sampling law on dissipation is explicit. Sliced methods (Liutkus et al., 2019; Bonet et al., 2022; Hertrich et al., 2024) use slices to define the objective, the geometry, or the evaluation scheme, whereas here slices only propose trial potentials. Randomized methods have been used for kernel approximation (Rahimi & Recht, 2007), numerical linear algebra Martinsson & Tropp (2020), as well as approximation and smoothing in optimization (Gower & Richtárik, 2015; Nesterov & Spokoiny, 2017; Kozak et al., 2021). A more complete discussion of related work is given in Appendix A.

2 Background

We provide some background for our method. Section 2.1 reviews subspace descent, and Section 2.2 formulates the Wasserstein gradient flows used below.

2.1 Subspace descent

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, and let $H \succ 0$ define the metric $\|u\|_H^2 = u^\top H u$. The preconditioned gradient flow $\dot{x}_t = -H^{-1}\nabla f(x_t)$ moves along

$$u_\star = H^{-1}\nabla f(x) = \operatorname*{argmin}_{u\in\mathbb{R}^d}\Big\{\tfrac{1}{2}\|u\|_H^2 - \langle \nabla f(x), u\rangle\Big\}.$$

Along a flow $\dot{x}_t = -u_t$, the chain rule gives $-\frac{d}{dt}f(x_t) = \langle \nabla f(x_t), u_t\rangle$, the *dissipation* of u_t. For the full direction, $\mathcal{D}_\star(x) = \langle \nabla f(x), u_\star\rangle = \|u_\star\|_H^2$. Subspace descent restricts the same variational problem to the span of a matrix $U \in \mathbb{R}^{d\times r}$ of trial directions,

$$u_U = \operatorname*{argmin}_{u\in\mathrm{span}(U)}\Big\{\tfrac{1}{2}\|u\|_H^2 - \langle \nabla f(x), u\rangle\Big\} = U\big(U^\top H U\big)^\dagger U^\top \nabla f(x). \tag{1}$$

The step uses only the directional derivatives $U^\top\nabla f(x)$ and the Gram matrix $U^\top HU$. With $H = I_d$ it is random subspace descent (Kozak et al., 2021), and with H a Hessian it is randomized subspace Newton (Gower et al., 2019). Restricted stationarity makes u_U the H-orthogonal projection of $u_\star$ onto $\mathrm{span}(U)$, so its dissipation $\mathcal{D}_U = \|u_U\|_H^2$ is nonnegative and the Pythagorean identity gives the exact cost of the restriction,

$$\underbrace{\|u_\star\|_H^2}_{\text{full}} = \underbrace{\|u_U\|_H^2}_{\text{captured}} + \underbrace{\|u_\star - u_U\|_H^2}_{\text{lost}}. \tag{2}$$

The metric of the projection matters. Projecting the full direction in Euclidean coordinates, $\dot{x} = -UU^\top H^{-1}\nabla f(x)$, gives dissipation $\langle \nabla f(x), UU^\top H^{-1}\nabla f(x)\rangle$, which can be negative when $H \neq I_d$. Fitting in the flow's own metric guarantees descent.

The decomposition also separates a rate into a constant of the full flow and the fraction captured by the trial spaces (Appendix B).

2.2 PROBABILITY FLOWS

A potential u_t moves a law μ_t on $\mathbb{R}^d$ with velocity $-\nabla u_t$, so that $\partial_t \mu_t = \nabla \cdot (\mu_t \nabla u_t)$. Write $\langle f, g \rangle_\rho = \mathbb{E}_\rho[f \cdot g]$ and $\|f\|_\rho^2 = \langle f, f \rangle_\rho$. If an objective F admits a first variation and the flow is regular enough for the chain rule and integration by parts, then $\frac{d}{dt} F(\mu_t) = -\ell_{\mu_t}(u_t)$, where

$$\ell_\mu(u) = \mathbb{E}_\mu \Big[\nabla \frac{\delta F}{\delta \mu}(\mu) \cdot \nabla u \Big]$$

is the dissipation of u at μ, the analogue of $\langle \nabla f(x), u \rangle$. A symmetric bilinear form B_μ fixes the geometry, and the full direction solves

$$u_\star = \operatorname*{argmin}_u \left\{ \tfrac{1}{2} B_\mu(u, u) - \ell_\mu(u) \right\}. \tag{3}$$

The kinetic form $B_\mu^{\mathrm{W}}(u, h) = \mathbb{E}_\mu[\nabla u \cdot \nabla h]$ gives the Wasserstein gradient flow, whose full potential is the first variation $u_\star = \delta F / \delta \mu$. Other forms precondition the flow. The objective specifies what decreases, B_μ specifies how the direction is chosen, and the trial space of Section 3 specifies how that direction is approximated. Appendix C treats general convex penalties in place of $\frac{1}{2} B_\mu$, and Appendix **??** treats optimal transport objectives.

MMD and Sobolev descent. Let k be a symmetric positive-definite kernel with continuous and bounded derivatives through order two, and let $(\mathcal{K}\rho)(x) = \int k(x, y) d\rho(y)$ for finite signed measures ρ. MMD descent (Arbel et al., 2019) minimizes $F_k(\mu) = \frac{1}{2} \|m_\mu - m_\pi\|_{\mathcal{H}_k}^2$, with mean embeddings $m_\mu = \mathbb{E}_\mu[k(X, \cdot)]$, in the kinetic geometry. Its full potential is the kernel witness $u_\star = \mathcal{K}(\mu - \pi)$. Sobolev descent (Mroueh et al., 2019) keeps a finite-feature version of this objective and changes the geometry. For $k(x, y) = \psi(x)^\top \psi(y)$ with $\psi : \mathbb{R}^d \to \mathbb{R}^{d_\psi}$, write $\delta_\mu = \mathbb{E}_\mu \psi - \mathbb{E}_\pi \psi$, so that $F_\psi(\mu) = \frac{1}{2} \|\delta_\mu\|^2$, and $(D_\mu)_{ij} = \mathbb{E}_\mu[\nabla \psi_i \cdot \nabla \psi_j]$. The vector $(C_\mu h)_i = \langle \nabla \psi_i, \nabla h \rangle_\mu$ records how a potential h moves the feature means, and $\ell_\mu(h) = \delta_\mu^\top C_\mu h$. For $\lambda > 0$, the Sobolev geometry and full potential are

$$B_\mu(u, h) = \langle C_\mu u, C_\mu h \rangle + \lambda \langle \nabla u, \nabla h \rangle_\mu, \quad u_\star = \psi^\top (D_\mu + \lambda I_{d_\psi})^{-1} \delta_\mu, \tag{4}$$

so Sobolev descent inserts the regularized inverse $(D_\mu + \lambda I_{d_\psi})^{-1}$ into the MMD dynamics without changing the objective (Appendix C).

LAWGD. Let $\pi \propto e^{-V}$, let $\mathcal{L} = -\Delta + \nabla V \cdot \nabla$ be the negative Langevin generator, and let $f_\mu = d\mu / d\pi - 1$ be the density error. Assume that $\mathcal{L}$ has a self-adjoint realization in $L^2(\pi)$ whose kernel consists of the constants and whose spectral gap is $\gamma > 0$, and that $f_\mu \in L_0^2(\pi)$, the mean-zero functions on which $\mathcal{L}^{-1}$ is defined. For $F(\mu) = \mathsf{KL}(\mu \| \pi)$, integration by parts gives $\ell_\mu(h) = \mathbb{E}_\mu[\mathcal{L} h] = \langle f_\mu, \mathcal{L} h \rangle_\pi$. LAWGD (Chewi et al., 2020a) uses the target-dependent geometry $B_\mu(u, h) = \langle \mathcal{L} u, \mathcal{L} h \rangle_\pi$. Its full potential $u_\star = \mathcal{L}^{-1} f_\mu$ satisfies $B_\mu(u_\star, h) = \ell_\mu(h)$, so the flow and its dissipation are

$$\partial_t \mu_t = \nabla \cdot \big(\mu_t \nabla \mathcal{L}^{-1} f_{\mu_t} \big), \quad \mathcal{D}_\star(\mu) = \|f_\mu\|_\pi^2 = \chi^2(\mu \| \pi). \tag{5}$$

In this geometry the variational objective is a Poisson residual,

$$\tfrac{1}{2} B_\mu(u, u) - \ell_\mu(u) = \tfrac{1}{2} \|\mathcal{L} u - f_\mu\|_\pi^2 - \tfrac{1}{2} \|f_\mu\|_\pi^2.$$

Our method minimizes this residual over an inexpensive space of potentials instead of inverting $\mathcal{L}$. More recent methods for approximating LAWGD use diffusion maps (Li & Marzouk, 2025) or Koopman operators (Xu et al., 2025).

3 RANDOM SUBSPACE PROJECTIONS OF PROBABILITY FLOWS

Each flow in Section 2 minimizes $\frac{1}{2} B_\mu(u, u) - \ell_\mu(u)$ over all potentials. Our proposed method keeps the objective and the geometry fixed and uses a random subspace to project the descent direction at each iteration. More specifically, at iteration n, we draw m ridge potentials from a sampling law ν_n, project the full potential onto their span in the geometry B_μ by a Gram solve, and move the particles with the projected potential (Algorithm 1).

We proceed with the description of our method. Section 3.1 describes the sampled trial spaces, Section 3.2 gives the variational projection, and Section 3.3 specializes the construction to MMD, Sobolev, and LAWGD flows

3.1 SAMPLING THE SUBSPACE

We use ridge functions $\varphi_z(x) = h_z(\theta_z^\top x)$ because they are inexpensive to evaluate while supporting rich trial spaces (Pinkus, 2015). Here $\theta_z \in \mathbb{S}^{d-1}$ and h_z is a twice differentiable profile. At iteration n, Z_j^n are drawn independently from ν_n given the current state and the trial space is $\mathcal{U}_n = \operatorname{span}\{\varphi_{Z_1^n}, \dots, \varphi_{Z_m^n}\}$. The support of ν_n determines which descent directions can be captured, and its weighting determines which components it captures efficiently.

Examples of ridge function classes that can be used in practice are:

1. Random Fourier ridges $h(s) = \cos(\omega s + \beta)$, with $\theta = \xi/\|\xi\|$ for $\xi \sim \mathcal{N}(0, I_d)$, ω uniform on a finite set of positive frequencies, and $\beta \sim \mathrm{Unif}[0, 2\pi)$ (Rahimi & Recht, 2007).
2. Sliced proposals that solve a one-dimensional problem on the projected measures $\mu_\theta = (x \mapsto \theta^\top x)_{\#}\mu$ and π_θ. For example, one may solve the one-dimensional LAWGD problem on the projected measures and lift the solution to the ambient space.
3. Curvature-informed directions draw ξ from a Gaussian whose covariance is the inverse of a particle estimate of the objective curvature, which turns out to be optimal in the translation model of Section 4.3.

Further implementation details, including concrete sampling laws, are given in Appendix D.

3.2 PROJECTION ONTO THE SAMPLED SUBSPACE

Let $\mathcal{U} = \operatorname{span}\{\varphi_1, \dots, \varphi_m\}$ and $u_a = \sum_{j=1}^m a_j \varphi_j$. The restricted problem

$$u_{\mathcal{U}} = \operatorname*{argmin}_{u \in \mathcal{U}} \left\{ \tfrac{1}{2} B_\mu(u, u) - \ell_\mu(u) \right\} \tag{6}$$

is the quadratic program $\min_a \{\frac{1}{2} a^\top G a - b^\top a\}$ with

$$G_{ij} = B_\mu(\varphi_i, \varphi_j), \quad b_i = \ell_\mu(\varphi_i). \tag{7}$$

The entries of b are the directional derivatives of the objective along the sampled potentials, which is the analogue of $U^\top \nabla f(x)$ from subspace descent, and G is the analogue of $U^\top H U$. With exact moments, the coefficients are $a = G^\dagger b$, which is well defined because continuity of ℓ_μ puts b in the range of G.

Along the population flow at $\mu = \mu_t$, the captured dissipation is $-\frac{d}{dt} F(\mu_t) = \ell_\mu(u_{\mathcal{U}}) = b^\top G^\dagger b \geq 0$, and with $\|u\|_{B_\mu}^2 = B_\mu(u, u)$, we see that the dissipation can be decomposed as

$$\underbrace{\|u_\star\|_{B_\mu}^2}_{\text{full decrease}} = \underbrace{\|u_{\mathcal{U}}\|_{B_\mu}^2}_{\text{captured decrease}} + \underbrace{\|u_\star - u_{\mathcal{U}}\|_{B_\mu}^2}_{\text{lost decrease}}. \tag{8}$$

This is the Euclidean identity (2) but for the specific flows being considered.

In the particle implementation, we must implement the moments used in this projection. In particular, $\widehat{G}_n \succeq 0$ and $\widehat{b}_n$, which are evaluated using the particle distributions, replace the moments, and the coefficients solve $(\widehat{G}_n + \varepsilon_n I_m)\widehat{a}_n = \widehat{b}_n$. The regularization $\varepsilon_n > 0$ is needed because an estimated $\widehat{b}_n$ need not lie in the range of a singular $\widehat{G}_n$. With exact moments, the regularized fit still descends, since $b^\top a = a^\top G a + \varepsilon \|a\|^2 \geq 0$. The particles then move along $-\nabla \widehat{u}_n$ with $\widehat{u}_n = \sum_j \widehat{a}_{n,j} \varphi_j^n$. A full description of the resulting subspace descent algorithm is given in Algorithm 1.

3.3 IMPLEMENTATIONS OF THE FLOWS OF INTEREST

For LAWGD, each trial potential enters through its image under the generator,

$$q_j(x) = \mathcal{L}\varphi_j(x) = -h_j''(\theta_j^\top x) + h_j'(\theta_j^\top x)\theta_j^\top \nabla V(x), \tag{9}$$

which costs one directional derivative $\theta_j^\top \nabla V$ per point. The system has $G_{ij} = \mathbb{E}_\pi[q_i q_j]$ and $b_i = \mathbb{E}_\mu[q_i]$. With $(Q_X)_{rj} = q_j(X_n^r)$ and $(Q_Y)_{sj} = q_j(Y_s)$ for a target reference sample $Y_1, \dots, Y_M$,

Algorithm 1 Random subspace descent for probability flows

Require: Initial particles $\{X_0^r\}_{r=1}^N$, number of updates K, number of trial potentials m, refresh period $k_{\mathrm{ref}} \geq 1$, sampling laws ν_n, step sizes $\eta_n > 0$, regularizers $\varepsilon_n > 0$, moment estimators for the chosen objective and geometry.

1: **for** $n = 0, \dots, K-1$ **do**
2: **if** $n \bmod k_{\mathrm{ref}} = 0$ **then**
3: Draw $Z_1^n, \dots, Z_m^n \overset{\text{i.i.d.}}{\sim} \nu_n$ and set $\varphi_j^n = \varphi_{Z_j^n}$.
4: **else**
5: Keep $\varphi_j^n = \varphi_j^{n-1}$.
6: **end if**
7: Assemble $\widehat{G}_{n,ij} = \widehat{B}_n(\varphi_i^n, \varphi_j^n)$ and $\widehat{b}_{n,i} = \widehat{\ell}_n(\varphi_i^n)$, reusing cached quantities when the trial potentials are unchanged.
8: Solve $(\widehat{G}_n + \varepsilon_n I_m)\widehat{a}_n = \widehat{b}_n$.
9: Update all particles, $X_{n+1}^r = X_n^r - \eta_n \sum_{j=1}^m \widehat{a}_{n,j} \nabla\varphi_j^n(X_n^r)$.
10: **end for**

Ensure: Final particles $\{X_K^r\}_{r=1}^N$.

the estimates are $\widehat{G}_n = Q_Y^\top Q_Y/M$ and $\widehat{b}_n = Q_X^\top \mathbf{1}_N/N$. With $\mathcal{W} = \mathcal{L}\mathcal{U}$, the projection satisfies $\mathcal{L}u_{\mathcal{U}} = P_{\mathcal{W}} f_\mu$, so the captured dissipation is $\|P_{\mathcal{W}} f_\mu\|_\pi^2$ and the part of the density error outside $\mathcal{W}$ is exactly the lost KL dissipation. While b must be recomputed at each iteration, if the subspace is not updated then there is no need to recompute G: its factorization can be reused while the trial potentials are held fixed. A new draw costs $O((N+M)dm + Mm^2 + m^3)$ to recompute $\widehat{G}_n$.

For MMD and Sobolev descent with the fixed-feature objective F_ψ, each trial potential enters through its response in the features,

$$J_{pi} = \mathbb{E}_\mu[\nabla\psi_p \cdot \nabla\varphi_i] = \mathbb{E}_\mu\big[h_i'(\theta_i^\top X)\, \theta_i^\top \nabla\psi_p(X)\big], \tag{10}$$

which costs one directional derivative $\theta_i^\top \nabla\psi_p$ per point. Because the two flows share the objective, they share the directional derivatives $b = J^\top \delta_\mu$. They differ only in the Gram matrix, which is $G_{ij}^{\mathrm{W}} = \mathbb{E}_\mu[\nabla\varphi_i \cdot \nabla\varphi_j]$ for MMD descent and $J^\top J + \lambda G^{\mathrm{W}}$ for Sobolev descent. If we write $(P_X)_{ri} = h_i'(\theta_i^\top X_n^r)$ and $\Theta = [\theta_1, \dots, \theta_m]$, then these moments can be estimated as $\widehat{\delta}_n = N^{-1}\sum_r \psi(X_n^r) - \mathbb{E}_\pi\psi$, $(\widehat{J}_n)_\pi = N^{-1}\sum_r h_i'(\theta_i^\top X_n^r)\, \theta_i^\top \nabla\psi_p(X_n^r)$, and $\widehat{G}_n^{\mathrm{W}} = (P_X^\top P_X/N) \odot (\Theta^\top \Theta)$. Target samples are used to compute the feature mean $\mathbb{E}_\pi\psi$. For MMD descent, the projection is the best kinetic-norm approximation of the witness $\psi^\top \delta_\mu$ in $\mathcal{U}$. For Sobolev descent, the projection solves a regularized least-squares problem. Since $\frac{d}{dt}\delta_{\mu_t} = -J_t a_t$, only the component of the feature error in the range of J can be reduced. A key difference from LAWGD is that the Gram matrix depends on the current law, so both $\widehat{G}_n$ and $\widehat{b}_n$ must be recomputed at every iteration. Each iteration costs $O(Nd_\psi dm + Nm^2 + m^3)$, compared with $O(Nd_\psi^2 d + d_\psi^3)$ for full Sobolev descent.

4 Theory

We characterize how the sampling law in Algorithm 1 controls captured dissipation, derive expected convergence rates, and analyze a translation model that separates the roles of objective and geometry. Supporting results on the dissipation captured by sampled subspaces, conditions for asymptotic coverage, mode-wise sampling rates, and lifted one-dimensional Poisson proposals are collected in Appendix E.

4.1 Setting

Throughout this section, we suppress the dependence on μ when it is obvious: we write $B = B_\mu$, $\ell = \ell_\mu$, and $\|u\|_B^2 = B(u,u)$.

The following assumption ensures that the full potential $u_\star$ exists and has finite dissipation, that projections onto sampled subspaces are well defined, and that the random subspaces are measurable, so that expectations such as $\mathbb{E}\mathcal{D}_{\mathcal{U}_m}$ make sense.

Assumption 4.1. *The potentials modulo B-null potentials form a separable Hilbert space $(\mathcal{V}, B)$. The functional ℓ is continuous on $\mathcal{V}$ with Riesz representer $u_\star$, so that $\mathcal{D}_\star = \|u_\star\|_B^2$. The map $z \mapsto \varphi_z \in \mathcal{V}$ is measurable, and ν is a probability measure on its index set $\mathcal{Z}$.*

Draw $Z_1, \dots, Z_m$ i.i.d. from ν and let $\mathcal{U}_m = \operatorname{span}\{\varphi_{Z_1}, \dots, \varphi_{Z_m}\}$. By (8), a subspace $\mathcal{U}$ captures $\mathcal{D}_\mathcal{U} = \|P_\mathcal{U} u_\star\|_B^2$, so the expected capture is a quadratic form in the *average projector* $S_{\nu,m} = \mathbb{E}[P_{\mathcal{U}_m}]$,

$$\mathbb{E}\mathcal{D}_{\mathcal{U}_m} = B\big(u_\star, S_{\nu,m} u_\star\big). \tag{11}$$

In $\mathbb{R}^d$, uniformly random subspaces give $S_{\nu,m} = (m/d)I$, and so dissipation in all directions is uniformly decreased by a multiplicative factor. In infinite dimensions, $S_{\nu,m}$ has trace at most m, and so no uniform fraction exists, and the capture depends on how well ν represents $u_\star$.

One way to quantify this is to use the random-feature norm (Bach, 2017). Letting $\widetilde{\varphi}_z = \varphi_z / \|\varphi_z\|_B$, and $\widetilde{\varphi}_z = 0$ when $\varphi_z = 0$, we define

$$R_\nu(u)^2 = \inf\Big\{ \|c\|_{L^2(\nu)}^2 : u = \int c(z)\widetilde{\varphi}_z \nu(dz) \Big\}, \tag{12}$$

with $R_\nu(u) = \infty$ when u has no such representation. If $R_\nu(u_\star) < \infty$, Maurey's empirical method gives (see Theorem E.3)

$$\mathbb{E}\mathcal{D}_{\mathcal{U}_m} \geq \frac{m\mathcal{D}_\star^2}{(m-1)\mathcal{D}_\star + R_\nu(u_\star)^2} \geq \frac{\mathcal{D}_\star^2}{R_\nu(u_\star)^2}. \tag{13}$$

In the next section, Theorem 4.3 turns these two bounds into linear and sublinear rates based on conditions on the sampling distribution ν.

4.2 Expected convergence

We analyze convergence of the population scheme without particle approximations. Let the law and the sampling law evolve according to $\mu_{n+1} = (\mathrm{id} - \eta \nabla u_{\mathcal{U}_n})_\# \mu_n$ from a deterministic μ_0, and the potentials in $\mathcal{U}_n$ are sampled from an $\mathcal{F}_n$-measurable law ν_n with $\mathcal{F}_n = \sigma(\mu_0, \dots, \mu_n)$. We assume that Assumption 4.1 holds at every μ_n and that $F \geq 0$. Write $B_n = B_{\mu_n}$, $u_{\star,n} = u_\star(\mu_n)$, $\mathcal{D}_{\star,n} = \|u_{\star,n}\|_{B_n}^2$, $F_n = F(\mu_n)$, and S_n for the conditional average of $P_{\mathcal{U}_n}$ given $\mathcal{F}_n$.

We also assume the following descent bound, which is analogous to a descent lemma in smooth Euclidean optimization.

Assumption 4.2 (Descent Lemma)**.** *For some $c > 0$ and all $0 < \eta \leq \eta_0$, $F_{n+1} \leq F_n - c\eta \mathcal{D}_{\mathcal{U}_n}$ almost surely.*

While this condition holds under bounded-derivative conditions on MMD and Sobolev descent, it remains an open question when it holds for LAWGD in general. We verify that this condition holds in some simple examples in Appendix F.2.

Theorem 4.3 (Expected rates)**.** *Let Assumption 4.2 hold, and suppose that the full flow satisfies $\mathcal{D}_{\star,n} \geq \kappa F_n$ almost surely.*

- *(a) If $B_n(u_{\star,n}, S_n u_{\star,n}) \geq \alpha \mathcal{D}_{\star,n}$ almost surely and $c\eta\alpha\kappa \leq 1$, then $\mathbb{E}F_n \leq (1 - c\eta\alpha\kappa)^n F_0$.*
- *(b) If $R_{\nu_n}(u_{\star,n}) \leq R_0$ almost surely and $\beta = c\eta\kappa^2/R_0^2$ satisfies $\beta F_0 < 1$, then $\mathbb{E}F_n \leq F_0/(1 + \beta F_0 n)$.*

The proof of this theorem is in Appendix F. The constant κ belongs to the full flow, and α or R_0 belongs to the sampling law. In the appendix, Theorem E.3 gives $\alpha = m/(m + C_R - 1)$. Since the relative bound in (a) may not hold along the flow, the absolute norm does not increase, which yields (b). For LAWGD, $\kappa = 1$, since $\mathcal{D}_\star = \chi^2(\mu\|\pi) \geq \mathsf{KL}(\mu\|\pi)$. For the fixed-feature objective with $D_\mu \succeq d_0 I_{d_\psi}$, $\kappa = 2d_0$ for MMD descent and $\kappa = 2d_0/(d_0 + \lambda)$ for Sobolev descent.

The assumptions of Theorem 4.3 can be verified in specific examples, which we now discuss. In the appendix, Remark F.2 gives a case in which part (a) holds but part (b) fails. For ease of reading, we present the following example here. This toy construction relies on a known orthogonal factorization and a single active factor. Here, α is the worst-case probability that the active direction is sampled,

while the LAWGD full-flow constant is $\kappa = 1$. The proof hinges on each discrete update preserving the density bound $1 - \delta \le d\mu_n/d\pi \le 1 + \delta$, which gives both the one-step descent inequality and the uniform bound on the random-feature norm.

Example 4.4 (Product targets with one active factor). Let $e_1, \dots, e_d$ be an orthonormal basis, write $z_i = e_i^\top x$, and let

$$\pi(x) = \prod_{i=1}^{d} p_i(z_i), \qquad p_i \propto e^{-v_i}, \qquad \lambda_i \le v_i'' \le \Lambda_i, \qquad v_i'(0) = 0,$$

with $K = \max_i \Lambda_i/\lambda_i$. Suppose the initial law differs from the target in one unknown factor j,

$$\frac{d\mu_0}{d\pi}(x) = 1 + g_0(z_j), \qquad \int g_0\, p_j = 0, \qquad \|g_0\|_\infty \le \delta \le \tfrac{1}{2},$$

with g_0 continuous. We run LAWGD with marginal Poisson proposals along the basis directions. At iteration n, draw m indices i.i.d. with probabilities $\varpi_i > 0$, and for each drawn index i use the ridge potential $\varphi_{i,n}(x) = h_{i,n}(z_i)$, where $h_{i,n}$ solves the one-dimensional Poisson equation

$$-h_{i,n}'' + v_i' h_{i,n}' = \frac{d\mu_{n,i}}{dp_i} - 1,$$

with $\mu_{n,i}$ the marginal of μ_n along e_i and C_ρ the distribution function of ρ. Then, for every $0 < \eta \le 1/(4K)$, Assumption 4.2 holds with $c = \frac{1}{2}$, the full flow satisfies $\mathcal{D}_{\star,n} \ge F_n$, and the hypotheses of both parts of Theorem 4.3 hold with

$$\alpha = 1 - (1 - \omega_{\min})^m, \qquad R_0^2 = \frac{\delta^2}{\omega_{\min}}, \qquad \omega_{\min} = \min_i \omega_i.$$

Consequently,

$$\mathbb{E}F_n \le \Big(1 - \frac{\eta\alpha}{2}\Big)^n F_0 \qquad \text{and} \qquad \mathbb{E}F_n \le \frac{F_0}{1 + \eta\omega_{\min} F_0 n/(2\delta^2)}.$$

For uniform sampling, $\alpha = 1 - (1 - 1/d)^m$, which equals $1/d$ when $m = 1$, and $R_0^2 = d\delta^2$. The rigorous proof is in Appendix F.2.

4.3 TRANSLATIONS OF THE TARGET

To explicitly see the roles of the geometry and the sampling law in our construction, consider translations $\mu_\tau = \pi(\cdot - \tau)$ and suppose we restrict to linear ridge potentials $\theta^\top x$, which we collect in $\Theta \in \mathbb{R}^{d\times m}$. Linear potentials have constant gradients, so their flows stay in this family. Under Assumption F.3, with Fisher information $I_F = \mathbb{E}_\pi[\nabla V \nabla V^\top] \succ 0$, the KL objective on the family is $g(\tau) = \mathbb{E}_\pi[V(X + \tau) - V(X)]$, with $\nabla^2 g(0) = I_F$.

Proposition 4.5 (Reduction to Euclidean subspace descent)**.** *Let $\mu_0 = \mu_{\tau_0}$ and let Θ_t be piecewise constant. The fitted flow with trial potentials $\Theta_t^\top x$ satisfies $\mu_t = \mu_{\tau_t}$ with*

$$\dot\tau_t = -\Theta_t(\Theta_t^\top H \Theta_t)^\dagger \Theta_t^\top \nabla \bar F(\tau_t), \tag{14}$$

where, with features $\psi = Wx$ for MMD and Sobolev descent,

Flow	$\bar F(\tau) = F(\mu_\tau)$	H
Wasserstein gradient flow of KL	$g(\tau)$	I_d
LAWGD	$g(\tau)$	$I_F = \nabla^2 g(0)$
MMD	$\frac{1}{2}\|W\tau\|^2$	I_d
Sobolev	$\frac{1}{2}\|W\tau\|^2$	$W^\top W + \lambda I_d$

With $\Theta = I_d$, LAWGD is the Newton flow with Hessian frozen at the optimum, and for a Gaussian target it gives $\dot\tau = -\tau$ regardless of the covariance.

In (14), Θ chooses the trial subspace and H the metric of the projection. With $H = I_d$ the method is random subspace descent (Kozak et al., 2021), and with $H = I_F$, it is a random subspace Fisher

Table 1: High-dimensional ripple capture (mean and SD; three paired seeds).

flow	sliced	projected	gain (pp)
LAWGD	0.943 ± 0.003	0.987 ± 0.001	$+4.4\%$
Sobolev	0.943 ± 0.002	0.992 ± 0.001	$+4.9\%$
MMD	0.859 ± 0.007	0.968 ± 0.001	$+10.9\%$

natural gradient method (or equivalently, it is randomized subspace Newton (Gower et al., 2019) with the Hessian frozen at the optimum).

On quadratic objectives $\bar{F}(\tau) = \frac{1}{2}\tau^\top Q\tau$ with $Q \succ 0$, which include Gaussian KL and linear-feature MMD and Sobolev descent, we can use this reduction to compare geometries exactly. For a law P of Θ with rank at most m, let $A_\mathsf{P} = \mathbb{E}_{\Theta\sim\mathsf{P}}[\Theta(\Theta^\top H\Theta)^\dagger\Theta^\top]$, $r(\mathsf{P}) = \lambda_{\min}(Q^{1/2}A_\mathsf{P}Q^{1/2})$, and $L_H = \lambda_{\max}(H^{-1/2}QH^{-1/2})$. The discrete scheme with independent draws $\Theta_n \sim \mathsf{P}$ and step $1/L_H$ contracts $\mathbb{E}\bar{F}$ by the factor $1 - r(\mathsf{P})/L_H$ (Gower & Richtárik, 2015). Theorem F.4 in Appendix F shows that $r(\mathsf{P})/L_H \le m/(L_H \operatorname{tr}(HQ^{-1})) \le m/d$ for every law of rank at most m, with equality in the second bound only when H is a multiple of Q, and that a Gaussian sketch attains m/d when $H = Q$.

For a Gaussian target $\mathcal{N}(0, \Sigma)$ and $m = 1$, uniform directions give $r(\mathsf{P}) = \lambda_{\min}(\Sigma^{-1})/d$ in the Wasserstein geometry, and the best law gives $1/\operatorname{tr}\Sigma$. In the LAWGD geometry, if we let $\theta \sim \mathcal{N}(0, \Sigma)$, then we find $r(\mathsf{P}) = 1/d$ independent of the conditioning of Σ. The sampling law thus carries the covariance information that LAWGD stores in its Gram matrix. This motivates the choice of the curvature-informed directions mentioned in Section 3.1.

5 Experiments

This section discusses the empirical results related to our method. In Section 5.1 we discuss the Gaussian experiment that isolates the effect of projection on flow dissipation. Sections 5.2 and 5.3 test whether projection preserves fine distributional structure that uniform slicing can smooth away. Section 5.4 then asks whether the same comparison survives in a setting where the moments are estimated and the objective is not the reported quantity, by using the flow as the coupling of a generative model on CIFAR-100.

Appendix G gives supplementary experiments in more regimes.

5.1 Gaussian Dissipation Experiment

In Figure 1, we compare full Gaussian flow trajectories under MMD, Sobolev, and LAWGD dynamics, varying the dimension $d \in \{4, 16, 64\}$ and target covariance condition number $\kappa \in \{1, 30, 1000\}$. Using exact Gaussian moments, we evolve the covariance while recomputing each method's velocity at every integration stage. All methods use shared, periodically refreshed dictionaries of $2d$ random quadratic ridge potentials. We compare the subspace flows with uniform slicing. We plot the remaining objective relative to its initial value against time. We see that uniformly in this set of examples, projection preserves more dissipation than slicing.

5.2 High-dimensional ripple

We test a 16-dimensional standard-normal product modulated in each coordinate by $1 + 0.25\cos(16\pi\Phi(x_j))$. Gaussian initialization matches its envelope but lacks this high-frequency ripple. We run 300 steps with 2,048 particles, 4,096 target particles, 64 trial slices, and 256 probe slices, where we refresh the probe at each step. Capture is $c_8(\mu_T)/c_8(\pi)$, where $c_8(\mu) = 16^{-1}\sum_j \mathbb{E}_\mu[\cos(16\pi\Phi(X_j))]$. We see that projection uniformly improves capture of the fine structure in this example.

Table 2: Multiscale-ring global W_2 (three paired seeds; lower is better).

flow	sliced global W_2	projected global W_2	ΔW_2
LAWGD	0.327	0.317	-0.009
Sobolev	0.235	0.234	-0.002
MMD	0.502	0.491	-0.011

5.3 Multiscale rings

The third experiment compares rings centered at $-4, -2, 0, 2, 4$ with radii $0.75, 0.20, 0.75, 0.20, 0.75$, uniform angles, radial noise 0.035, and equal total mass at each scale. Using the same sampling budgets, we run 200 steps of size 0.5. Global W_2 measures the distance between generated and target samples. Again, we see that the projected flows are uniformly better at capturing finer structure in the generated data.

Projection improves all nine ripple runs (up to 10.9 points) and lowers ring W_2 for every flow and 34 of 36 family–horizon–seed comparisons (Tables 1–2), consistent with recovering descent components lost by uniform slicing.

5.4 Generative Modeling on CIFAR-100

We compare flow-derived couplings for flow matching on four CIFAR-100 superclasses to test whether projected flows can yield better couplings than sliced flows. A convolutional autoencoder maps each five-class group ($N = 2500$) to a whitened 256-dimensional latent, and a fixed-budget velocity network learns the transported pairs. We redraw 512 trial directions on 4×4 latent patches each step and report GlobFID in a 128-dimensional CNN feature space, using each method's best step size found on a grid. We see that again projected methods typically recover better fine structure, exhibited through their ability to better fit these specific classes. Projection improves 10 of 12 cells, including every flow on large carnivores, aquatic mammals, and vehicles. In one example, trees, LAWGD and Sobolev do slightly worse.

Table 3: CIFAR-100 GlobFID. Cells are sliced/projected/Δ. Δ is projected minus sliced, and lower is better.

superclass	MMD	LAWGD	Sobolev
large carnivores	$9.327/9.058/-0.269$	$9.976/9.229/-0.747$	$10.034/9.192/-0.842$
aquatic mammals	$3.708/3.573/-0.135$	$3.905/3.593/-0.312$	$3.935/3.660/-0.275$
trees	$11.667/11.566/-0.101$	$12.303/12.442/+0.139$	$11.735/11.965/+0.230$
vehicles	$8.128/7.431/-0.697$	$8.320/7.833/-0.487$	$7.703/7.587/-0.116$

6 Conclusion

Randomized approximations of Wasserstein gradient flows can be used to develop efficient, implementable algorithms, but there are many design choices that need to be made. Proper projections that respect the objective and geometry in our specific examples of MMD descent, Sobolev descent, and LAWGD reduce to Gram solves, which capture the full dissipation. We are able to analyze the effect of this projection through the average projector, which acts as a preconditioner on the infinite-dimensional space. Experiments demonstrate the advantage of this framework in practice over naive random slicing methods.

Several questions remain. We use particle approximations, but we currently do not have theoretical control of their error. LAWGD also requires target samples to form the Gram matrix, and its theoretical guarantees change if we use alternative matrices, which should be explored. Generative-modeling experiments could also test whether tailored preconditioned flows emphasize fine distributional features.

7 AI DISCLOSURE

In this work, we used generative AI tools for the full scope of the research: to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology and experiments, and implement methods. Additionally, we used generative AI tools for creating figures, suggesting experimental parameters, creating and editing software code, drafting parts of the research paper, summarizing existing literature, brainstorming, searching for information, and formatting references. All LLM-generated text has been revised by human authors for readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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, 2nd edition, 2008. doi: 10.1007/978-3-7643-8722-8. URL https://link.springer.com/book/10.1007/978-3-7643-8722-8.

Michael Arbel, Danica J. Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for MMD GANs. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.neurips.cc/paper_files/paper/2018/hash/07f75d9144912970de5a09f5a305e10c-Abstract.html.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/944a5ae3483ed5c1e10bbccb7942a279-Abstract.html.

Michael Arbel, Arthur Gretton, Wuchen Li, and Guido Montúfar. Kernelized Wasserstein natural gradient. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hklz71rYvS.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017. URL https://www.jmlr.org/papers/v18/15-178.html.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Clément Bonet, Nicolas Courty, François Septier, and Lucas Drumetz. Efficient gradient flows in sliced-Wasserstein space. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2110.10972.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and preconditioned gradient descent in Wasserstein space. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25311–25374, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/2cf153951b5e9b39564fc4a0ef6adc1a-Abstract-Conference.html.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi: 10.1007/s10851-014-0506-3. URL https://doi.org/10.1007/s10851-014-0506-3.

Siwan Boufadène and François-Xavier Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *SIAM Journal on Mathematical Analysis*, 57(4):4556–4587, 2025. doi: 10.1137/24M1647631. URL https://epubs.siam.org/doi/10.1137/24M1647631.

Paul Caucheteux, Clément Bonet, and Anna Korba. A unifying view of variational generative Wasserstein flows. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.31369.

Peng Chen and Omar Ghattas. Projected Stein variational gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1947–1958, 2020. URL https://papers.nips.cc/paper/2020/hash/14faf969228fc18fcd4fcf59437b0c97-Abstract.html.

Yifan Chen and Wuchen Li. Optimal transport natural gradient for statistical manifolds with continuous sample space. *Information Geometry*, 3(1):1–32, 2020. doi: 10.1007/s41884-020-00028-0. URL https://arxiv.org/abs/1805.08380.

Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Anna Korba, Arthur Gretton, and Bharath K. Sriperumbudur. (De)-regularized Maximum Mean Discrepancy Gradient Flow. *Journal of Machine Learning Research*, 26(235):1–77, 2025. URL https://jmlr.org/papers/v26/24-1574.html.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2098–2109, 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/16f8e136ee5693823268874e58795216-Abstract.html.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. Exponential ergodicity of mirror-Langevin diffusions. *Advances in Neural Information Processing Systems*, 33:19573–19585, 2020b.

Lénaïc Chizat, Maria Colombo, Roberto Colombo, and Xavier Fernández-Real. Quantitative convergence of Wasserstein gradient flows of kernel mean discrepancies, 2026. URL https://arxiv.org/abs/2603.01977. Preprint.

Nicolas Chopin, Francesca Crucinio, and Anna Korba. A connection between tempering and entropic mirror descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8782–8800. PMLR, 2024. URL https://proceedings.mlr.press/v235/chopin24a.html.

Giacomo Cozzi and Filippo Santambrogio. Long-time asymptotics of the sliced-Wasserstein flow. *SIAM Journal on Imaging Sciences*, 18(1):1–19, 2025. doi: 10.1137/24M1656414. URL https://arxiv.org/abs/2405.06313.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Zhiyan Ding, Qin Li, Jianfeng Lu, and Stephen J Wright. Random coordinate langevin monte carlo. In *Conference on learning theory*, pp. 1683–1710. PMLR, 2021.

Masao Doi. Onsager principle as a tool for approximation. *Chinese Physics B*, 24(2):020505, 2015. doi: 10.1088/1674-1056/24/2/020505. URL https://cpb.iphy.ac.cn/CN/10.1088/1674-1056/24/2/020505.

Hanze Dong, Xi Wang, Yong Lin, and Tong Zhang. Particle-based variational inference with preconditioned functional gradient flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6OphWWAE3cS.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023. URL https://www.jmlr.org/papers/v24/20-602.html.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

Alexandre Galashov, Valentin de Bortoli, and Arthur Gretton. Deep MMD gradient flow without adversarial training. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/717b9fd2ede6b8a9971a296d5179df89-Abstract-Conference.html.

Weiye Gan, Tangjun Wang, and Zuoqiang Shi. Weighted laplacian flow: A deterministic particle flow with exponential convergence, 2026. URL https://arxiv.org/abs/2608.21831v2. Version 2, September 14, 2026.

Pierre Glaser, Michael Arbel, and Arthur Gretton. KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/433a6ea5429d6d75f0be9bf9da26e24c-Abstract.html.

Isabella Carla Gonnella, Olga Mula, Federico Pichi, and Gianluigi Rozza. A dynamical approximation scheme on the Stiefel manifold for Wasserstein gradient flows, 2026. URL https://arxiv.org/abs/2607.17235.

Robert M. Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

Robert M. Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. RSN: Randomized subspace Newton. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/bc6dc48b743dc5d013b1abaebd2faed2-Abstract.html.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2): 217–288, 2011. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

Ye He, Krishnakumar Balasubramanian, Bharath K. Sriperumbudur, and Jianfeng Lu. Regularized Stein variational gradient flow. *Foundations of Computational Mathematics*, 25:1199–1257, 2025. doi: 10.1007/s10208-024-09663-w. URL https://doi.org/10.1007/s10208-024-09663-w.

Johannes Hertrich. Fast kernel summation in high dimensions via slicing and Fourier transforms. *SIAM Journal on Mathematics of Data Science*, 6(4):1109–1137, 2024. doi: 10.1137/24M1632085. URL https://doi.org/10.1137/24M1632085.

Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. Generative sliced MMD flows with Riesz kernels. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5b288823575bb29654b0953a251e933b-Abstract-Conference.html.

Johannes Hertrich, Tim Jahn, and Michael Quellmalz. Fast summation of radial kernels via QMC slicing. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iNmVX9lx9l.

Elias Hess-Childs, Dejan Slepčev, and Lantian Xu. Radon–Wasserstein gradient flows for interacting-particle sampling in high dimensions, 2026. URL https://arxiv.org/abs/2602.05227.

Ya-Ping Hsieh, Ali Kavis, Paul Rolland, and Volkan Cevher. Mirrored Langevin dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Yijie Jin, Shu Liu, Hao Wu, Xiaojing Ye, and Haomin Zhou. Parameterized Wasserstein gradient flow. *Journal of Computational Physics*, 524:113660, 2025. doi: 10.1016/j.jcp.2024.113660. URL https://arxiv.org/abs/2404.19133.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359. URL https://epubs.siam.org/doi/10.1137/S0036141096303359.

Sophia Seulkee Kang, Louis Sharrock, Xiaoyuan Cheng, François-Xavier Briol, and Zonghao Chen. A gradient flow perspective on minimum MMD estimation, 2026. URL https://arxiv.org/abs/2607.03871.

Jun Kitagawa and Asuka Takatsu. Sliced optimal transport: Is it a suitable replacement? *Indiana University Mathematics Journal*, 75(2):487–524, 2026. doi: 10.1512/iumj.2026.75.60788. URL https://arxiv.org/abs/2311.15874.

Othmar Koch and Christian Lubich. Dynamical low-rank approximation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):434–454, 2007. doi: 10.1137/050639703. URL https://epubs.siam.org/doi/10.1137/050639703.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5719–5730, 2021. URL https://proceedings.mlr.press/v139/korba21a.html.

David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Computational Optimization and Applications*, 79(2):339–368, 2021. doi: 10.1007/s10589-021-00271-w. URL https://arxiv.org/abs/2003.02684.

Lauro Langosco di Langosco, Vincent Fortuin, and Heiko Strathmann. Neural variational gradient descent, 2021. URL https://arxiv.org/abs/2107.10731. Preprint.

Frédéric Legoll and Tony Lelièvre. Effective dynamics using conditional expectations. *Nonlinearity*, 23(9):2131–2163, 2010. doi: 10.1088/0951-7715/23/9/006. URL https://arxiv.org/abs/0906.4865.

Fengyi Li and Youssef Marzouk. Diffusion map particle systems for generative modeling. *Foundations of Data Science*, 7(3):814–837, 2025. doi: 10.3934/fods.2024054. URL https://www.aimsciences.org/article/doi/10.3934/fods.2024054.

V. Ya. Lin and Allan Pinkus. Fundamentality of ridge functions. *Journal of Approximation Theory*, 75(3):295–311, 1993.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/17ed8abedc255908be746d245e50263a-Abstract.html.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.neurips.cc/paper_files/paper/2016/hash/b3ba8f1bee1238a2f37603d90b58898d-Abstract.html.

Qiang Liu and Dilin Wang. Stein variational gradient descent as moment matching. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/125b93c9b50703fe9dac43ec231f5f83-Abstract.html.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4104–4113. PMLR, 2019. URL https://proceedings.mlr.press/v97/liutkus19a.html.

Per-Gunnar Martinsson and Joel A Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

Tyler Maunu and Jiayi Yao. Subspace Langevin Monte Carlo. *SIAM Journal on Mathematics of Data Science*, 8(3):850–886, 2026. doi: 10.1137/25M1739327. URL https://doi.org/10.1137/25M1739327.

Youssef Mroueh. Regularized finite dimensional kernel sobolev discrepancy, 2018. URL https://arxiv.org/abs/1805.06441.

Youssef Mroueh and Truyen Nguyen. On the convergence of gradient descent in GANs: MMD GAN as a gradient flow. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1720–1728. PMLR, 2021. URL https://proceedings.mlr.press/v130/mroueh21a.html.

Youssef Mroueh and Mattia Rigotti. Unbalanced Sobolev descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17034–17043, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c5f5c23be1b71adb51ea9dc8e9d444a8-Abstract.html.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJA7xfb0b.

Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2976–2985. PMLR, 2019. URL https://proceedings.mlr.press/v89/mroueh19a.html.

Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001. doi: 10.1081/PDE-100002243. URL https://www.mis.mpg.de/publications/preprint-repository/article/1999/issue-8.

Sangmin Park and Dejan Slepčev. Geometry and analytic properties of the sliced Wasserstein space. *Journal of Functional Analysis*, 289(7):110975, 2025. doi: 10.1016/j.jfa.2025.110975. URL https://arxiv.org/abs/2311.05134.

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*, pp. 5072–5081. PMLR, 2019. URL https://proceedings.mlr.press/v97/paty19a.html.

Allan Pinkus. *Ridge functions*, volume 205. Cambridge University Press, 2015.

Gilles Pisier. Remarques sur un résultat non publié de B. Maurey. In *Séminaire d'Analyse Fonctionnelle 1980–1981*. École Polytechnique, 1981. Exposé V.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://papers.nips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, volume 21, 2008. URL https://papers.nips.cc/paper/2008/hash/0efe32849d230d7f53049ddc4a4b0c60-Abstract.html.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/61b1fb3f59e28c67f3925f3c79be81a1-Abstract.html.

Louis Sharrock and Christopher Nemeth. Coin sampling: Gradient-based Bayesian inference without learning rates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30850–30882. PMLR, 2023. URL https://proceedings.mlr.press/v202/sharrock23a.html.

Zheyang Shen, Markus Heinonen, and Samuel Kaski. De-randomizing MCMC dynamics with the diffusion Stein operator. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17507–17517, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9271905e840548b8cada6d60c0cfd93b-Abstract.html.

Gauthier Thurin, Claire Boyer, and Kimia Nadjahi. Convergence rates for distribution matching with sliced optimal transport. In *Proceedings of Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6156–6196. PMLR, 2026. URL https://proceedings.mlr.press/v336/thurin26a.html.

Chenyang Tian, Bharath K. Sriperumbudur, Arthur Gretton, and Zonghao Chen. Sobolev regularized MMD gradient flow, 2026. URL https://arxiv.org/abs/2605.11884. Preprint.

Christophe Vauthier, Anna Korba, and Quentin Mérigot. Towards understanding gradient dynamics of the sliced-Wasserstein distance via critical point analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61071–61107. PMLR, 2025. URL https://proceedings.mlr.press/v267/vauthier25a.html.

Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/5dcd0ddd3d918c70d380d32bce4e733a-Abstract.html.

Yifei Wang, Peng Chen, and Wuchen Li. Projected Wasserstein gradient descent for high-dimensional Bayesian inference. *SIAM/ASA Journal on Uncertainty Quantification*, 10(4):1513–1532, 2022. doi: 10.1137/21M1454018. URL https://arxiv.org/abs/2102.06350.

Yifei Wang, Peng Chen, Mert Pilanci, and Wuchen Li. Optimal neural network approximation of Wasserstein gradient direction via convex optimization. *SIAM Journal on Mathematics of Data Science*, 6(4):978–999, 2024. doi: 10.1137/23M1573173. URL https://arxiv.org/abs/2205.13098.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi: 10.1561/0400000060. URL https://arxiv.org/abs/1411.4357.

Jinchao Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34(4): 581–613, 1992. doi: 10.1137/1034116. URL https://doi.org/10.1137/1034116.

Yewei Xu and Qin Li. Random coordinate descent on the Wasserstein space of probability measures. *arXiv preprint arXiv:2604.01606*, 2026.

Yuanchao Xu, Fengyi Li, Masahiro Fujisawa, Xiaoyuan Cheng, Youssef Marzouk, and Isao Ishikawa. Generative modeling through Koopman spectral analysis: An operator-theoretic perspective, 2025. URL https://arxiv.org/abs/2512.18837. Preprint.

Zhuoran Yang, Yufeng Zhang, Yongxin Chen, and Zhaoran Wang. Variational transport: A convergent particle-based algorithm for distributional optimization, 2020. URL https://arxiv.org/abs/2012.11554. Preprint, revised 2024.

Rentian Yao, Xiaohui Chen, and Yun Yang. Wasserstein proximal coordinate gradient algorithms. *Journal of Machine Learning Research*, 25(269):1–66, 2024. URL https://www.jmlr.org/papers/v25/23-0889.html.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In *Conference on learning theory*, pp. 3814–3841. PMLR, 2020.

Jia-Jie Zhu and Alexander Mielke. Kernel approximation of Fisher–Rao gradient flows, 2024. URL https://arxiv.org/abs/2410.20622.

Xinzhe Zuo, Jiaxi Zhao, Shu Liu, Stanley Osher, and Wuchen Li. Numerical analysis on neural network projected schemes for approximating one dimensional Wasserstein gradient flows. *Journal of Computational Physics*, 546:114501, 2026. doi: 10.1016/j.jcp.2025.114501. URL https://arxiv.org/abs/2402.16821.

A Extended related work

A.1 Wasserstein geometry

The Wasserstein gradient-flow view of sampling begins with the variational formulation of the Fokker–Planck equation (Jordan et al., 1998). Otto (2001) develops its formal Riemannian structure and Ambrosio et al. (2008) its metric theory. Many methods keep the objective and change the geometry, including mirror and preconditioned descent in Wasserstein space (Bonet et al., 2024) and tempering viewed as entropic mirror descent (Chopin et al., 2024). Stein variational gradient descent (Liu & Wang, 2016) descends KL in a kernel-induced geometry (Liu, 2017; Duncan et al., 2023). Matrix-valued kernels extend that geometry (Wang et al., 2019), regularization interpolates it toward the Wasserstein geometry (He et al., 2025), and Zhu & Mielke (2024) relate its kernel approximations to regression and dissipation. Our restriction preserves whichever geometry is chosen. At the population level, every fitted direction therefore descends, and the sampling law acts as an additional preconditioner whose effect we quantify.

Several works fit a descent direction within a function class (Yang et al., 2020; Langosco di Langosco et al., 2021; Wang et al., 2024). Closest to our approach, Dong et al. (2023) regularize the KL directional derivative over a function class and interpret quadratic regularization as weighted approximation of a preconditioned Wasserstein gradient. Our function classes are small random linear spaces of scalar potentials, so the fit reduces to a Gram solve. We treat MMD, Sobolev, and LAWGD in a common framework and analyze how the sampling law governs coverage, captured dissipation, and expected rates.

A related line of work restricts the flow to a parametric family. Arbel et al. (2020) restrict a dual formulation of the Wasserstein natural gradient to an RKHS. Other works pull the Wasserstein metric back to model or map parameters (Chen & Li, 2020; Jin et al., 2025; Zuo et al., 2026), and Caucheteux et al. (2026) show that the induced Gram geometry need not equal the Wasserstein pullback. On location families the Wasserstein natural gradient is Euclidean (Chen & Li, 2020), and our translation model extends this reduction to the LAWGD and Sobolev geometries. Kang et al. (2026) project an MMD descent field onto the directions realizable by a generator, and Gonnella et al. (2026) evolve transport maps in adaptive subspaces. Unlike these works, we restrict the instantaneous transport potential and assume no parametric family of distributions or maps. Other methods restrict the state variables instead (Chen & Ghattas, 2020; Wang et al., 2022) or use coordinate methods on products of Wasserstein spaces (Yao et al., 2024).

Arbel et al. (2019) study the Wasserstein gradient flow of the squared MMD, our unpreconditioned kernel example. Sobolev GAN introduces a gradient-constrained critic (Mroueh et al., 2018), related to a weighted negative Sobolev norm with explicit finite-feature approximation (Mroueh, 2018). Arbel et al. (2018) study gradient regularization of MMD critics. Sobolev descent (Mroueh et al., 2019) keeps the feature-MMD objective and preconditions transport with a regularized inverse derivative Gram matrix. Tian et al. (2026) propose a new Sobolev-regularized MMD flow, and their convergence analysis rests on a regularity condition on the difference between kernel mean embeddings. This extends to parametric generators (Mroueh & Nguyen, 2021) and transport-reaction dynamics (Mroueh & Rigotti, 2020). Quantitative convergence of full kernel-discrepancy flows is known only under specific structural conditions (Boufadène & Vialard, 2025; Chizat et al., 2026), which is why our rates assume an explicit dissipation inequality for the full flow. Other works modify the objective itself (Glaser et al., 2021; Chen et al., 2025; Korba et al., 2021; Galashov et al., 2025). These are complementary to ours: once an objective and a geometry are fixed, our analysis describes the randomized approximation of the resulting flow. Our mode-rate comparison shows how the spectrum of the kernel determines what sampling can recover of LAWGD's conditioning in these flows.

Chewi et al. (2020a) introduce LAWGD, which preconditions KL descent with the inverse Langevin generator and dissipates at the rate of the χ^2 divergence. Implementing it requires this inverse. Li & Marzouk (2025) estimate it with diffusion maps. Xu et al. (2025) learn the generator through Koopman spectral analysis and bound the dissipation lost by spectral truncation, a close precedent for our captured-dissipation identities. We instead solve a least-squares problem over the span of the functions $\mathcal{L}\varphi_j$ for random trial potentials, which needs neither the density ratio nor a solution of the full Poisson equation. Related work inverts general diffusion generators (Shen et al., 2021), removes step-size tuning (Sharrock & Nemeth, 2023), or uses a different Poisson forcing (Gan et al., 2026).

The functions $\mathcal{L}\varphi_j$ depend on the target only through its score, so the directional derivatives of KL are finite-feature Stein moments comparing the current law with the target (Liu & Wang, 2016; 2018). This gives two score-only fits: the restricted Wasserstein gradient flow of KL and a proxy for the LAWGD geometry. At the population level, the proxy Gram matrix approaches the target Gram matrix as the current law converges in χ^2 divergence. Kernel Stein discrepancy descent (Korba et al., 2021) instead uses a Stein discrepancy as the objective. To bound how much a lifted one-dimensional Poisson solution captures, we use the conditional-expectation identities of effective dynamics (Legoll & Lelièvre, 2010).

A.2 Random subspaces, sketching, and random features

Randomized subspace descent replaces the full gradient with directional derivatives along a few random directions (Kozak et al., 2021). Sketch-and-project methods analyze such updates through the expected sketch (Gower & Richtárik, 2015). Randomized subspace Newton keeps the Newton geometry inside the restricted step (Gower et al., 2019). Coordinate descent adapts the sampling probabilities to directional smoothness (Nesterov, 2012). Random coordinate Langevin (Ding et al., 2021) and coordinate descent in Wasserstein space (Xu & Li, 2026) extend these ideas to the space of measures. Maunu & Yao (2026) extend randomized subspace descent to Langevin by updating random eigenblocks of a preconditioner. Proposing a small trial space and then solving a reduced problem exactly also underlies subspace correction methods (Xu, 1992), sketching (Woodruff, 2014), and randomized low-rank approximation (Halko et al., 2011). Dynamical low-rank approximation (Koch & Lubich, 2007) and Onsager-principle approximations (Doi, 2015) similarly restrict the variational principle that defines an evolution.

Random features approximate kernels and functions by sampled nonlinear features (Rahimi & Recht, 2007; 2008). Maurey's empirical method (Pisier, 1981; Barron, 1993) gives $O(1/m)$ approximation error for functions with finite norm in the space generated by the feature law, while subsequent work characterizes how the feature distribution and number of sampled features affect approximation and statistical efficiency (Bach, 2017; Rudi & Rosasco, 2017). Random Fourier features approximate shift-invariant kernels (Rahimi & Recht, 2007). Here we instead use random features to construct transport functions: adding profiles enlarges the trial space while each profile remains univariate, and the joint Gram system accounts for interactions between these profiles in the ambient space.

A common randomized approximation in Wasserstein space uses one-dimensional slices. Examples include sliced Wasserstein barycenters (Bonneel et al., 2015), sliced-Wasserstein flows (Liutkus et al., 2019), and JKO schemes in sliced-Wasserstein space (Bonet et al., 2022). Sliced-Wasserstein space is not a length space (Park & Slepčev, 2025), and sliced metrics are generally not bi-Lipschitz equivalent to their classical counterparts (Kitagawa & Takatsu, 2026). Monotone descent of a sliced objective does not settle long-time behavior (Cozzi & Santambrogio, 2025; Vauthier et al., 2025; Thurin et al., 2026). Radon–Wasserstein geometries build a new ambient geometry whose velocities are computable from projections (Hess-Childs et al., 2026). For kernel discrepancies, slicing gives exact identities for Riesz-kernel MMDs (Hertrich et al., 2024) and fast kernel summation (Hertrich, 2024; Hertrich et al., 2025). Low-dimensional projections also define robust transport distances (Paty & Cuturi, 2019). In each of these works, slicing defines the objective, the geometry, or the evaluation scheme. In ours, it only proposes trial potentials.

B Remarks on Euclidean subspace descent

Table 4 records how each Euclidean quantity from Section 2.1 carries over to probability flows in Sections 2.2 and 3.

The decomposition (2) separates the ingredients of a convergence rate. Suppose that along $\dot{x}_t = -u_{U_t}$ the trial spaces capture a fraction $\alpha \in (0, 1]$ of the full dissipation, $\mathcal{D}_{U_t}(x_t) \geq \alpha \mathcal{D}_\star(x_t)$, and that the full flow satisfies the Polyak–Łojasiewicz-type inequality $\mathcal{D}_\star(x) \geq \kappa\big(f(x) - f_\star\big)$ with $f_\star = \inf f > -\infty$. Grönwall's inequality then gives $f(x_t) - f_\star \leq e^{-\alpha\kappa t}\big(f(x_0) - f_\star\big)$. The constant κ belongs to the full flow, and α belongs to the trial spaces.

If ∇f is L-Lipschitz, the update $x^+ = x - \eta u_U$ satisfies

$$f(x^+) \leq f(x) - \eta \mathcal{D}_U(x) + \tfrac{L\eta^2}{2}\|u_U\|^2 \leq f(x) - \eta\Big(1 - \tfrac{L\eta}{2\lambda_{\min}(H)}\Big)\mathcal{D}_U(x),$$

Table 4: Euclidean subspace descent and its probability-flow analogue.

	Euclidean	Probability flow
State	$x \in \mathbb{R}^d$	μ
Trial space	$\mathrm{span}(U)$	$\mathrm{span}\{\varphi_1, \dots, \varphi_m\}$
Directional derivatives	$U^\top \nabla f(x)$	$b_i = \ell_\mu(\varphi_i)$
Metric	$u^\top H h$	$B_\mu(u, h)$
Gram matrix	$U^\top H U$	$G_{ij} = B_\mu(\varphi_i, \varphi_j)$
Full direction	$H^{-1}\nabla f(x)$	$u_\star$ in (6)
Dissipation identity	(2)	(8)
Full-flow inequality	$\mathcal{D}_\star \geq \kappa(f - f_\star)$	$\mathcal{D}_\star \geq \kappa F$

using $\|u_U\|^2 \leq \|u_U\|_H^2/\lambda_{\min}(H) = \mathcal{D}_U(x)/\lambda_{\min}(H)$. For $\eta \leq \lambda_{\min}(H)/L$, this gives $f(x^+) \leq f(x) - \frac{\eta}{2}\mathcal{D}_U(x)$, the Euclidean instance of Assumption 4.2 with $c = 1/2$.

Another way to retain descent is to project the transformed gradient $H^{-1/2}\nabla f(x)$ onto $\mathrm{span}(U)$ in Euclidean coordinates and map back, giving $\widetilde{u}_U = H^{-1/2} P H^{-1/2} \nabla f(x)$ with $P = UU^\top$ for orthonormal U. This is not a separate construction. It is (1) with trial matrix $U' = H^{-1/2}U$, since $U'^\top H U' = I_r$. The restricted variational problem therefore covers both options, and the only choice is the trial space. When trial directions are sampled, as in our method, the natural option is to keep the sampled span and let the metric enter only through the Gram matrix.

C Details on the flows

General convex penalties. The quadratic penalty $\frac{1}{2}B_\mu$ in (3) is one choice of a differentiable convex penalty $\mathcal{R}_\mu$ minimized at zero. For such a penalty, the full direction and its restriction to a linear space $\mathcal{U}_\mu$ solve

$$u_\star \in \operatorname*{argmin}_u \{\mathcal{R}_\mu(u) - \ell_\mu(u)\} \text{ and } u_{\mathcal{U}} \in \operatorname*{argmin}_{u \in \mathcal{U}_\mu} \{\mathcal{R}_\mu(u) - \ell_\mu(u)\}. \tag{15}$$

The first-order condition for the constrained minimizer is $D\mathcal{R}_\mu(u_{\mathcal{U}})[h] = \ell_\mu(h)$ for $h \in \mathcal{U}_\mu$. Taking $h = u_{\mathcal{U}}$ and using convexity gives

$$\frac{d}{dt}F(\mu_t) = -\ell_{\mu_t}(u_{\mathcal{U}}) = -D\mathcal{R}_{\mu_t}(u_{\mathcal{U}})[u_{\mathcal{U}}] \leq -\big(\mathcal{R}_{\mu_t}(u_{\mathcal{U}}) - \mathcal{R}_{\mu_t}(0)\big) \leq 0,$$

so restricting the variational problem preserves objective descent for every such penalty. Appendix **??** gives two nonquadratic examples.

MMD descent. With the kinetic geometry, the MMD potential $u_\star = \mathcal{K}(\mu - \pi)$ gives the flow

$$\partial_t \mu_t = \nabla \cdot \big(\mu_t \nabla \mathcal{K}(\mu_t - \pi)\big), \tag{16}$$

with full dissipation $\mathcal{D}_\star(\mu) = \|\nabla \mathcal{K}(\mu - \pi)\|_\mu^2$.

Sobolev descent. Retain the finite-moment and gradient assumptions of MMD descent, with an admissible potential space containing the ψ_i modulo null potentials. The feature assumptions make C_μ bounded, and $\lambda > 0$ ensures that the Sobolev geometry controls the kinetic norm. To compute the full direction, set $c_\mu = (D_\mu + \lambda I_{d_\psi})^{-1}\delta_\mu$. Since $C_\mu(\psi^\top c) = D_\mu c$,

$$B_\mu(\psi^\top c_\mu, h) = c_\mu^\top D_\mu C_\mu h + \lambda c_\mu^\top C_\mu h = c_\mu^\top (D_\mu + \lambda I_{d_\psi}) C_\mu h = \delta_\mu^\top C_\mu h = \ell_\mu(h),$$

which gives the variational representation (4) and the flow $\partial_t \mu_t = \nabla \cdot \big(\mu_t \nabla [\psi^\top (D_{\mu_t} + \lambda I_{d_\psi})^{-1} \delta_{\mu_t}]\big)$. The operator measures the feature error, applies a regularized inverse derivative Gram matrix, and lifts the result to a potential. Along the flow, $\frac{d}{dt}\mathbb{E}_{\mu_t}\psi_i = -(C_{\mu_t}u_\star)_i$, so

$$\frac{d}{dt}\delta_{\mu_t} = -D_{\mu_t}(D_{\mu_t} + \lambda I_{d_\psi})^{-1}\delta_{\mu_t}$$

and $\mathcal{D}_\star(\mu) = \ell_\mu(u_\star) = \delta_\mu^\top D_\mu (D_\mu + \lambda I_{d_\psi})^{-1}\delta_\mu$. Regularization allows D_μ to be singular.

Transport admissibility for LAWGD. Transport admissibility requires additional control under the current law. For example, if $d\mu/d\pi \le C$, then for mean-zero u,

$$\|\nabla u\|_\mu^2 \le C\|\nabla u\|_\pi^2 = C\langle u, \mathcal{L}u\rangle_\pi \le \frac{C}{\gamma}\|\mathcal{L}u\|_\pi^2.$$

This is a sufficient condition for the generator norm to control kinetic energy. The KL chain rule and integration by parts remain required along the flow.

D Implementation details

D.1 Particle update, regularization, and refreshing

Algorithm 1 summarizes the method. The projected potential $\widehat{u}_n = \sum_j \widehat{a}_{n,j}\varphi_j^n$ defines a transport field on the ambient space, and all particles move simultaneously under it,

$$X_{n+1}^r = X_n^r - \eta_n \sum_{j=1}^m \widehat{a}_{n,j}(h_j^n)'\big((\theta_j^n)^\top X_n^r\big)\theta_j^n, \tag{17}$$

for $r = 1, \dots, N$. The field lies in the span of the m sampled directions. Storing the maps $T_n = \mathrm{id} - \eta_n \nabla\widehat{u}_n$ also transports new samples from the initial distribution through the composition $T_{K-1} \circ \cdots \circ T_0$. In continuous time, the same construction gives the particle system $\dot{X}_t^i = -\sum_j \widehat{a}_j(t)\nabla\varphi_j(X_t^i)$ for the empirical measure $\widehat{\mu}_t^N = N^{-1}\sum_i \delta_{X_t^i}$, and the population dissipation identities describe the underlying subspace flow.

The implementation fixes the scaling of the trial functions before fitting and solves $(\widehat{G}_n + \varepsilon_n I_m)\widehat{a}_n = \widehat{b}_n$ by a matrix factorization. The regularizer is a numerical choice, distinct from any regularization already included in B_μ. With estimated moments, $\widehat{b}_n^\top \widehat{a}_n$ is a fitting diagnostic rather than a certificate of objective decrease, since moment error and the finite step η_n must be controlled separately.

The trial potentials are redrawn every k_{ref} iterations and held fixed in between, while the coefficients are refit at every iteration because the directional derivatives, and for MMD and Sobolev descent also the geometry, change with μ_n. Redrawing at every iteration matches the independent draws analyzed in Section 4. Longer holds save computation because cached quantities are reused, and for LAWGD with a fixed reference sample the Gram matrix and its factorization do not change. Without refreshing, the flow stalls when the current descent direction is nearly orthogonal to a single finite-dimensional span.

D.2 Sampling laws

Example D.1 (Random Fourier ridges). The default law draws independently for each function $\xi_j \sim \mathcal{N}(0, I_d)$, $\theta_j = \xi_j/\|\xi_j\|$, $\omega_j \sim \mathrm{Unif}(\Omega)$, and $\beta_j \sim \mathrm{Unif}[0, 2\pi)$, and sets

$$h_j(s) = \cos(\omega_j s + \beta_j), \tag{18}$$

with a finite set of positive frequencies Ω. More generally, for a nonsingular matrix $A \in \mathbb{R}^{d\times d}$, the direction law may be anisotropic, $\theta_j = A\xi_j/\|A\xi_j\|$, and the frequency law may be any law on $(0, \infty)$. For a continuous frequency law, full support transfers coverage from a dictionary whose span is already dense in the flow norm, but full support alone does not establish that density (Proposition E.6). The tails of the law determine how cheaply smooth versus oscillatory potentials are represented, which is why we also use mixtures of bandwidths.

Example D.2 (Slice-based proposals). Given a sampled direction θ, a one-dimensional problem on the marginals μ_θ and π_θ can propose a profile h_θ, which defines the ridge potential $\varphi_\theta(x) = h_\theta(\theta^\top x)$ with $\nabla\varphi_\theta(x) = h_\theta'(\theta^\top x)\theta$. Options include a marginal Poisson solution for LAWGD, a marginal transport potential for transport objectives (Appendix **??**), and a marginal critic for Sobolev descent. For Sobolev descent, we group the fixed objective features into ridge blocks $\Psi_i(\theta_i^\top x)$, including their fixed weights, where $\Psi_i : \mathbb{R} \to \mathbb{R}^{d_{\psi,i}}$. With μ_{θ_i} and π_{θ_i} the corresponding marginals, we set $D_i = \mathbb{E}_{\mu_{\theta_i}}[\Psi_i'\Psi_i'^\top]$, $\delta_i = \mathbb{E}_{\mu_{\theta_i}}\Psi_i - \mathbb{E}_{\pi_{\theta_i}}\Psi_i$, and

$$\varphi_i(x) = \Psi_i(\theta_i^\top x)^\top (D_i + \lambda I_{d_{\psi,i}})^{-1}\delta_i. \tag{19}$$

Algorithm 2 LAWGD moment assembly for Algorithm 1

Require: Current particles $\{X_r\}_{r=1}^N$, target reference points $\{Y_s\}_{s=1}^M$, target gradient ∇V, sampled ridge potentials $\varphi_j(x) = h_j(\theta_j^\top x)$ with $\|\theta_j\| = 1$.

1: Evaluate ∇V at all current and target points (reuse cached target gradients when available).
2: **for** $j = 1, \dots, m$ **do**
3: Set $(Q_X)_{rj} = -h_j''(\theta_j^\top X_r) + h_j'(\theta_j^\top X_r)\theta_j^\top \nabla V(X_r)$ for all r.
4: Set $(Q_Y)_{sj} = -h_j''(\theta_j^\top Y_s) + h_j'(\theta_j^\top Y_s)\theta_j^\top \nabla V(Y_s)$ for all s.
5: **end for**
6: Set $\widehat{G} = Q_Y^\top Q_Y/M$ and $\widehat{b} = Q_X^\top \mathbf{1}_N/N$.

Ensure: $\widehat{G} \succeq 0$ and $\widehat{b}$.

Such proposals depend on the current marginals, so they change as μ_n evolves while the objective features ψ stay fixed. Let $\widetilde{\nu}_n$ denote this proposal law, and let ν_0 be a base law that covers the potential space. The mixture $\nu_n = (1-\varepsilon_{\mathrm{mix}})\widetilde{\nu}_n + \varepsilon_{\mathrm{mix}}\nu_0$ with $\varepsilon_{\mathrm{mix}} \in (0,1]$ preserves coverage and satisfies $R_{\nu_n}(u)^2 \le \varepsilon_{\mathrm{mix}}^{-1} R_{\nu_0}(u)^2$ whenever the right-hand side is finite. Indeed, $S_{\nu_n,1} \succeq \varepsilon_{\mathrm{mix}} S_{\nu_0,1}$, and the bound follows from the variational formula for R_ν in Appendix E.1. The direction law can itself be adapted. For sliced objectives, drawing directions in proportion to their marginal discrepancy $W_2(\mu_\theta, \pi_\theta)$ minimizes the bound on $R_\nu(u_\star)^2$ in Proposition **??**. Marginal Poisson proposals for LAWGD require the marginal of the target and hence target samples. Marginal transport proposals require samples from the target marginals.

Example D.3 (Curvature-informed directions). The direction law can carry preconditioning that would otherwise come from the geometry. In the quadratic translation model with matched geometry, the optimal Gaussian sketch has covariance proportional to the inverse objective Hessian (Theorem F.4). At a refresh, we form a particle-based curvature estimate $\widehat{Q}_n \succeq 0$ from the current particles and scores, draw $\xi \sim \mathcal{N}(0, (\widehat{Q}_n + \delta_Q I_d)^{-1})$ with $\delta_Q > 0$, and set $\theta = \xi/\|\xi\|$. We hold $\widehat{Q}_n$ fixed with the trial potentials and store it through a few leading eigenblocks, as in Maunu & Yao (2026).

D.3 LAWGD MOMENT ASSEMBLY

Algorithm 2 assembles the LAWGD moments for Algorithm 1. We assume access to ∇V and to a target reference sample $Y_1, \dots, Y_M \sim \pi$, and target quadrature can replace this sample when available. For a newly drawn set of trial potentials, evaluating the q_j, Gram assembly, the coefficient solve, and particle transport cost $O((N+M)dm + Mm^2 + m^3)$, excluding score evaluations and any marginal proposal solves. While the trial potentials are held fixed, $\widehat{G}_n$ and its regularized factorization can be cached, only $\widehat{b}_n$ must be recomputed, and the cost drops to $O(Ndm + m^2)$ per iteration. Replacing the target Gram matrix by $Q_X^\top Q_X/N$ changes the geometry, and Appendix D.5 analyzes this replacement.

D.4 MMD AND SOBOLEV IMPLEMENTATION

Both flows minimize the fixed-feature objective $F_\psi(\mu) = \frac{1}{2}\|\delta_\mu\|^2$. To run Algorithm 1, we need the feature discrepancy δ_μ, the matrix $J_{pi} = \mathbb{E}_\mu[\nabla\psi_p \cdot \nabla\varphi_i]$, and the kinetic Gram matrix $G^{\mathrm{W}}_{ij} = \mathbb{E}_\mu[\nabla\varphi_i \cdot \nabla\varphi_j]$, for $p = 1, \dots, d_\psi$ and $i, j = 1, \dots, m$. The feature discrepancy compares the current law with the target through the objective features, J records how each sampled potential moves those features, and G^{W} is the transport cost of the sampled potentials. The two flows share $b = J^\top \delta_\mu$ and differ only in the Gram matrix,

$$G_{\mathrm{MMD}} = G^{\mathrm{W}}, \quad G_{\mathrm{Sobolev}} = J^\top J + \lambda G^{\mathrm{W}}. \tag{20}$$

Changing G is what turns MMD descent into Sobolev descent. Along the flow, the fitted coefficients satisfy $\frac{d}{dt}\delta_{\mu_t} = -J_t a_t$, where J_t and a_t are evaluated at μ_t. Thus the sampled subspace can only reduce the components of δ_{μ_t} in the range of J_t.

With particles $X_1, \dots, X_N$ and target samples $Y_1, \dots, Y_M$, the estimates are

$$\widehat{\delta} = \frac{1}{N}\sum_r \psi(X_r) - \frac{1}{M}\sum_s \psi(Y_s), \quad \widehat{J}_{pi} = \frac{1}{N}\sum_r h_i'(\theta_i^\top X_r)\theta_i^\top \nabla\psi_p(X_r),$$

and $\widehat{G}^{\mathrm{W}} = \left(\frac{1}{N} P_X^\top P_X\right) \odot \left(\Theta^\top \Theta\right)$ with $(P_X)_{ri} = h_i'(\theta_i^\top X_r)$, where $\Theta = [\theta_1, \dots, \theta_m]$ and $\odot$ is the entrywise product. Target samples enter only through the feature means $\mathbb{E}_\pi \psi$, which are computed once and reused. For MMD with a general kernel, the directional derivatives $b_i = \mathbb{E}_\mu[\nabla \varphi_i \cdot \nabla \mathcal{K}(\mu - \pi)]$ are estimated by replacing μ and π in the witness with the particle and target samples, at cost $O(N(N+M)d)$. The sampled potentials then approximate the gradient of this fixed witness. Because the objective kernel and features remain fixed when the trial potentials are redrawn, sampling changes only the subspace, never the objective.

D.5 Sampling from scores

In score-based generative modeling, the target may be available only through its score $s(x) = \nabla \log \pi(x) = -\nabla V(x)$. The potential V may be known only up to an additive constant, or not at all when s is a learned score model, and samples from π may be unavailable. MMD and OT flows require additional target information, but KL directional derivatives can be evaluated from particles and scores alone. For a ridge potential,

$$q_j(x) = \mathcal{L}\varphi_j(x) = -h_j''(\theta_j^\top x) - h_j'(\theta_j^\top x)\theta_j^\top s(x),$$

so the directional derivatives $b_j = \mathbb{E}_\mu[q_j]$ can be estimated by $\widehat{b} = Q_X^\top \mathbf{1}_N / N$. Neither V, its normalizing constant, nor the density of μ enters. A coefficient fit also requires a Gram matrix, whose evaluation depends on the chosen geometry.

The Gram matrix

$$G_{ij}^{\mathrm{W}} = \mathbb{E}_\mu\left[h_i'(\theta_i^\top X) h_j'(\theta_j^\top X)\right] \theta_i^\top \theta_j$$

requires only particles. Fitting b against G^{W} restricts the Wasserstein gradient flow of KL, which is the deterministic Fokker–Planck flow. The unrestricted flow has velocity $-\nabla \log(\mu/\pi)$ and so needs the score of the current law, which particles do not provide. For the restricted fit, integration by parts moves the derivative of $\log \mu$ onto the trial potential, $\mathbb{E}_\mu[\nabla \log(\mu/\pi) \cdot \nabla \varphi] = \mathbb{E}_\mu[\mathcal{L}\varphi]$, which avoids estimating the current score. This is the variational principle of Dong et al. (2023), applied to a finite linear space of potentials.

We now give a particle proxy for the LAWGD geometry. The LAWGD Gram matrix $G_{ij}^\pi = \mathbb{E}_\pi[q_i q_j]$ is a target expectation. Replacing it by $G_{ij}^\mu = \mathbb{E}_\mu[q_i q_j]$, estimated by $Q_X^\top Q_X / N$, changes the geometry, but in a controlled way.

Proposition D.4 (Score-only fits)**.** *Let $q = (q_1, \dots, q_m)^\top$ with $q_j \in L^4(\pi)$, let $b = \mathbb{E}_\mu[q]$, and let $f_\mu = d\mu/d\pi - 1 \in L^2(\pi)$.*

- *(i) For every $G \succeq 0$ and $\varepsilon > 0$, the solution of $(G + \varepsilon I_m)a = b$ satisfies $b^\top a \geq \varepsilon \|a\|^2$, with $a = 0$ only if $b = 0$. Both score-only fits are therefore descent directions for KL at the population level, and they stall only when every directional derivative vanishes.*
- *(ii) $\left|G_{ij}^\mu - G_{ij}^\pi\right| \leq \chi^2(\mu\|\pi)^{1/2} \|q_i q_j\|_{L^2(\pi)}$. The proxy geometry therefore converges to the LAWGD geometry as μ approaches π.*
- *(iii) The proxy fit $a = (G^\mu)^\dagger b$ minimizes $\mathbb{E}_\mu\left[\left(\sum_j a_j q_j - 1\right)^2\right]$.*

Proof. (i) $b^\top a = a^\top (G + \varepsilon I_m) a \geq \varepsilon \|a\|^2$. (ii) $G_{ij}^\mu - G_{ij}^\pi = \int q_i q_j (d\mu - d\pi) = \langle f_\mu, q_i q_j \rangle_\pi$. Apply Cauchy–Schwarz and $\|f_\mu\|_\pi^2 = \chi^2(\mu\|\pi)$. (iii) Since

$$\tfrac{1}{2} a^\top G^\mu a - b^\top a = \tfrac{1}{2}\mathbb{E}_\mu\left[(q^\top a)^2\right] - \mathbb{E}_\mu[q^\top a] = \tfrac{1}{2}\mathbb{E}_\mu\left[(q^\top a - 1)^2\right] - \tfrac{1}{2},$$

the two objectives have the same minimizers. □

Part (iii) gives a direct interpretation. Exact LAWGD fits the density error f_μ by Stein features in $L^2(\pi)$, whereas the proxy fits the constant 1 by the same features in $L^2(\mu)$. Both use functions with zero π-mean to test for $\mu \neq \pi$. Part (ii) describes how the proxy approaches the LAWGD geometry as the current law approaches the target. Away from the target, it defines a different, law-dependent geometry that still guarantees population descent by (i).

Algorithm 3 Score-only moment assembly for Algorithm 1

Require: Particles $\{X_r\}_{r=1}^N$, score s, ridge potentials $\varphi_j(x) = h_j(\theta_j^\top x)$ with $\|\theta_j\| = 1$, $\Theta = [\theta_1, \dots, \theta_m]$, geometry choice (kinetic or proxy).

1: Evaluate $s(X_r)$ for all r.
2: **for** $j = 1, \dots, m$ **do**
3: $(P_X)_{rj} = h_j'(\theta_j^\top X_r)$ and $(Q_X)_{rj} = -h_j''(\theta_j^\top X_r) - h_j'(\theta_j^\top X_r)\theta_j^\top s(X_r)$ for all r.
4: **end for**
5: $\widehat{b} = Q_X^\top \mathbf{1}_N / N$.
6: **if** geometry is kinetic **then**
7: $\widehat{G} = (P_X^\top P_X / N) \odot (\Theta^\top \Theta)$ {$\odot$: entrywise product}
8: **else**
9: $\widehat{G} = Q_X^\top Q_X / N$
10: **end if**

Ensure: $\widehat{G} \succeq 0$ and $\widehat{b}$.

Marginal Poisson proposals need the negative target marginal score $\mathbb{E}_\pi[\theta^\top \nabla V(X) \mid \theta^\top X]$ for $X \sim \pi$, which is a target expectation. In the score-only setting, directly evaluable profiles such as the random Fourier ridges (18) replace them. With a learned score $\widehat{s}$, the error in each directional derivative is bounded by $\|h_j'(\theta_j^\top \cdot)\|_{L^2(\mu)} \|s - \widehat{s}\|_{L^2(\mu)}$. This error is measured under the current law, which may differ from the distribution on which $\widehat{s}$ was trained, so it can be large early in the flow. Finally, KL itself cannot be evaluated without V and the density of μ. The fitted dissipation $\widehat{b}^\top \widehat{a}$ and the Stein discrepancy $\|\widehat{b}\|$ over the current trial potentials provide score-only diagnostics for monitoring convergence. Using Algorithm 3 within Algorithm 1 costs $O(Ndm + Nm^2 + m^3)$ per iteration, plus N score evaluations. Compared with Algorithm 2, it requires no reference sample and has no M-dependent term.

E Supporting theory

The results in this section supplement the theory in Section 4.

E.1 Tools from sketching and random features

For a realized subspace $\mathcal{U}$, $P_{\mathcal{U}} v$ is a measurable function of $(Z_1, \dots, Z_m)$ through the Gram system and a Borel pseudoinverse, so all expectations below are well defined. The average projector turns the random span $\mathcal{U}_m$ into a deterministic operator. Despite its name, this average is generally not itself a projector, and Lemma E.2 shows instead that it is a positive contraction. It plays the role of the expected sketch matrix in sketch-and-project methods (Gower & Richtárik, 2015), and of $\mathbb{E}P = cI$ in random subspace descent (Kozak et al., 2021; Maunu & Yao, 2026).

Definition E.1 (Average projector)**.** $S_{\nu,m} = \mathbb{E}[P_{\mathcal{U}_m}]$, *defined weakly by* $B(v, S_{\nu,m} w) = \mathbb{E} B(v, P_{\mathcal{U}_m} w)$.

Lemma E.2 (Properties of the average projector)**.** *(i)* $0 \preceq S_{\nu,m} \preceq I$, *and* $\operatorname{tr} S_{\nu,m} = \mathbb{E} \dim \mathcal{U}_m \le m$.

(ii) $\mathbb{E}\mathcal{D}_{\mathcal{U}_m} = B(u_\star, S_{\nu,m} u_\star)$.

(iii) $S_{\nu,1} = \int \widetilde{\varphi}_z \otimes \widetilde{\varphi}_z \nu(dz)$, *so that* $\mathbb{E}\mathcal{D}_{\mathcal{U}_1} = \int \ell(\varphi_z)^2 \|\varphi_z\|_B^{-2} \nu(dz)$, *with the integral taken over* $\{\varphi_z \neq 0\}$.

(iv) $S_{\nu,1} \preceq S_{\nu,m} \preceq S_{\nu,m+1}$, *and* $\nu \mapsto S_{\nu,1}$ *is linear.*

(v) If $\dim \mathcal{V} = \infty$, *then* $\inf_{\|u\|_B = 1} B(u, S_{\nu,m} u) = 0$ *for every* m.

Proof. Items (i)–(iv) follow from linearity of expectation, $\operatorname{tr} P_{\mathcal{U}} = \dim \mathcal{U}$, the projection identity $\mathcal{D}_{\mathcal{U}} = B(u_\star, P_{\mathcal{U}} u_\star)$, and the nesting of the subspaces. For (v), item (i) shows that $S_{\nu,m}$ is trace class, so it cannot be bounded below on an infinite-dimensional space. □

Items (ii) and (iv) connect this operator directly to the convergence analysis: the quadratic form of $S_{\nu,m}$ is the expected progress, and enlarging the subspace can only increase it. Item (v), however, rules out the Euclidean condition $\mathbb{E}P \succeq cI$ on an infinite-dimensional space of transport potentials.

The random-feature norm (12) is finite precisely when u admits an exact integral representation with square-integrable coefficients under the sampling law. Equivalently,

$$R_\nu(u)^2 = \sup_{v \in \mathcal{V}} \{2B(u,v) - B(v, S_{\nu,1} v)\},$$

which equals $B(u, S_{\nu,1}^{-1} u)$ when $S_{\nu,1}$ is invertible, and the infimum in (12) is attained when finite. We also have $R_\nu(u) \geq \|u\|_B$. Thus $R_\nu(u)$ records both whether the sampling law can represent u and the smallest coefficient norm of such a representation.

Theorem E.3 (Bounds for sampled subspaces). *If $R = R_\nu(u_\star) < \infty$, then for every $m \geq 1$,*

$$\mathbb{E}\mathcal{D}_{\mathcal{U}_m} \geq \frac{m\mathcal{D}_\star^2}{(m-1)\mathcal{D}_\star + R^2} \text{ and } \mathbb{E}\|u_\star - P_{\mathcal{U}_m} u_\star\|_B^2 \leq \frac{R^2 - \mathcal{D}_\star}{m}.$$

In particular, if $R^2 \leq C_R \mathcal{D}_\star$, then $\mathbb{E}\mathcal{D}_{\mathcal{U}_m} \geq \frac{m}{m + C_R - 1} \mathcal{D}_\star$.

Proof. The second bound is Maurey's empirical method (Pisier, 1981; Barron, 1993), as used for random features (Rahimi & Recht, 2008). Take a minimizer c in (12) and use the competitor $v_m = m^{-1} \sum_j c(Z_j) \widetilde{\varphi}_{Z_j} \in \mathcal{U}_m$. Then $\mathbb{E}v_m = u_\star$ and $\mathbb{E}\|v_m\|_B^2 \leq \frac{m-1}{m}\mathcal{D}_\star + \frac{R^2}{m}$. The best-approximation property of $P_{\mathcal{U}_m}$ gives the second bound. For the first, $v_m \in \mathcal{U}_m$ implies $\mathcal{D}_\star = \mathbb{E}B(P_{\mathcal{U}_m} u_\star, v_m)$, and Cauchy–Schwarz yields $\mathcal{D}_\star^2 \leq \mathbb{E}\mathcal{D}_{\mathcal{U}_m} \cdot \mathbb{E}\|v_m\|_B^2$. □

The projection error decays as $1/m$, while the expected captured dissipation approaches the full dissipation as m grows. A finite random-feature norm is a quantitative requirement. The weaker coverage condition below is enough to rule out directions that are invisible to the sampling law, although it does not give a uniform rate. Here the subspaces are nested by taking $\mathcal{U}_m$ from the first m draws of a single i.i.d. sequence.

Definition E.4 (Coverage). *The law ν* covers *$\mathcal{V}$ if $\ker S_{\nu,1} = \{0\}$. Equivalently, $B(v, \varphi_z) = 0$ for ν-a.e. z implies $v = 0$.*

Theorem E.5 (Coverage). *Suppose ν covers $\mathcal{V}$. Then:*

- *(i) $\mathbb{E}\mathcal{D}_{\mathcal{U}_m} > 0$ whenever $u_\star \neq 0$, so a fresh draw captures positive dissipation in expectation at every nonstationary fixed law.*
- *(ii) Almost surely $\bigcup_m \mathcal{U}_m$ is dense in $\mathcal{V}$, and $B(u, S_{\nu,m} u) \uparrow \|u\|_B^2$ for every u.*
- *(iii) By Lemma E.2(v), the convergence in (ii) is not uniform when $\dim \mathcal{V} = \infty$.*

Proof. Part (i) follows from Lemma E.2(ii)–(iv), since $B(u_\star, S_{\nu,1} u_\star) = \int B(u_\star, \widetilde{\varphi}_z)^2 \nu(dz) > 0$. For part (ii), i.i.d. samples from the law of $\widetilde{\varphi}_Z$ on the separable space $\mathcal{V}$ are almost surely dense in its support. Any v orthogonal to all samples is therefore orthogonal to $\widetilde{\varphi}_z$ for ν-a.e. z, and coverage gives $v = 0$. Monotone convergence of the nested projections completes the proof. □

Thus coverage gives pointwise but not uniform consistency. Every fixed transport direction can be approximated as $m \to \infty$, but coverage alone cannot replace the quantitative assumptions in Theorem 4.3.

Proposition E.6 (Sufficient conditions for coverage). *Suppose $\mathcal{Z}$ is a topological space, $z \mapsto \varphi_z$ is continuous, ν has full topological support, and $\operatorname{span}\{\varphi_z : z \in \mathcal{Z}\}$ is dense in $\mathcal{V}$. Then ν covers $\mathcal{V}$.*

Proof. The set $\{z : B(v, \varphi_z) \neq 0\}$ is open, so if it is ν-null, full support implies that it is empty. Density of the span then gives $v = 0$. □

This proposition reduces coverage to familiar conditions on the feature family and its sampling law. Density of ridge functions is classical in L^2 and in spaces of continuous functions (Lin & Pinkus,

1993), but the flow norms used here are stronger. For LAWGD, the map $u \mapsto \mathcal{L}u$ is an isometry from $\mathcal{V}$ onto $L^2_0(\pi)$, and density of a ridge dictionary in the LAWGD geometry is therefore equivalent to density of the functions $\mathcal{L}\varphi_z$ in $L^2_0(\pi)$. For trigonometric ridges $\cos(\omega^\top x + \beta)$, these functions are

$$\|\omega\|^2 \cos(\omega^\top x + \beta) - \sin(\omega^\top x + \beta)\omega^\top \nabla V(x).$$

They are continuous in (ω, β) in $L^2(\pi)$ whenever $\mathbb{E}_\pi\|\nabla V\|^2 < \infty$, by dominated convergence.

E.2 Absolute and relative norm bounds in a fixed-operator model

The two parts of Theorem 4.3 impose different assumptions. Part (a) requires a fixed fraction of the full dissipation at every step, which Theorem E.3 obtains from a uniform bound on the ratio $R_\nu^2/\mathcal{D}_\star$. Part (b) instead uses a uniform bound on R_ν^2. The distinction matters because these two quantities evolve in opposite directions even in averaged linear dynamics.

To see this, freeze a bounded average operator S on $(\mathcal{V}, B)$ and consider the linear model $f_t = e^{-tS} f_0$. Denote the quadratic form associated with S^{-1} by $R_S(f_t)^2$. For $S = S_{\nu,1}$, wherever the representation formula above applies, $R_S(f_t) = R_\nu(f_t)$. If f_t is the full descent direction, then its dissipation is $\|f_t\|_B^2$, so define $\varrho(t) = R_S(f_t)^2/\|f_t\|_B^2$. The proposition concerns this fixed-operator model, rather than the nonlinear flow in which the geometry and sampling law may change with time.

Proposition E.7 (Evolution under a fixed operator)**.** *Let S be bounded, self-adjoint, injective, and nonnegative on $(\mathcal{V}, B)$. Let $f_0 \neq 0$, set $f_t = e^{-tS} f_0$ for $t \geq 0$, and let σ_0 be the spectral measure of f_0. Assume $R_S(f_0)^2 = \int s^{-1}\sigma_0(ds) < \infty$ and write $R_S(f_t)^2 = \int s^{-1} e^{-2ts}\sigma_0(ds)$. Then:*

(i) $R_S(f_t)^2$ is nonincreasing, and $\|f_t\|_B^2 \leq R_S(f_0)^2/(2et)$ for $t > 0$.

(ii) The ratio $\varrho(t) = R_S(f_t)^2/\|f_t\|_B^2$ is nondecreasing.

Proof. For part (i), monotonicity is immediate. The bound is the classical rate under a source condition (Engl et al., 1996):

$$\|f_t\|_B^2 = \int (se^{-2ts}) s^{-1}\sigma_0(ds) \leq R_S(f_0)^2 \sup_s se^{-2ts}.$$

For part (ii), let $p_t \propto e^{-2ts}\sigma_0(ds)$, so that $\varrho(t) = \int s^{-1} dp_t$. Differentiating gives $\varrho'(t) = -2\mathrm{Cov}_{p_t}(s^{-1}, s)$, which is nonnegative because s^{-1} and s are oppositely monotone, by Chebyshev's association inequality. □

The spectral picture explains both conclusions. Components associated with larger eigenvalues of S decay first, leaving more of the error in directions where the sampling operator is small. The source norm $R_S(f_t)^2$ therefore decreases, but its ratio to the remaining error increases, with $\varrho(t) \to 1/\inf\operatorname{supp}\sigma_0$. In infinite dimensions the limiting ratio can be arbitrarily large. Thus, even in the fixed-operator model, an initial relative bound need not persist, whereas the absolute norm does not increase. This distinction motivates the assumption in Theorem 4.3(b). Exponential decay remains available on spectrally resolved components: if σ_0 is supported in $[s_0, \infty)$ with $s_0 > 0$, then $\|f_t\|_B \leq e^{-s_0 t}\|f_0\|_B$.

E.3 Sampling as preconditioning: mode rates

One way to reproduce a target preconditioner in expectation is to sample its blocks and rescale each selected block by the inverse of its sampling probability. This is the identity $\mathbb{E}[h_k P_k] = hA$ used in subspace Langevin Monte Carlo (Maunu & Yao, 2026, Assumption 4.10), and the same idea underlies importance-sampled coordinate descent (Nesterov, 2012). Without inverse-probability rescaling, sampling in proportion to the block eigenvalues gives the same expected direction multiplied by the reciprocal of their sum. The desired direction is then retained, but its averaged rate is reduced by this normalization. The following proposition makes this comparison mode by mode for the four flows.

Proposition E.8 (Mode rates)**.** *Let $e_1, \ldots, e_J$ be orthonormal in $L^2_0(\pi)$ with $\mathcal{L}e_j = \gamma_j e_j$, so that $\gamma_j \geq \gamma$. For the fixed-feature kernel $k(x,y) = \psi(x)^\top\psi(y)$, define $K_\pi f = \mathcal{K}(f\pi) = \int k(\cdot, y) f(y)\pi(dy)$ and suppose $K_\pi e_j = k_j e_j$. Write $f = f_\mu$ and $f_j = \langle f, e_j\rangle_\pi$. Freeze the*

geometry at π and linearize the directional derivatives and dynamics about π, so that $\partial_t f = -\mathcal{L}u$. Fitting the single potential e_j changes only mode j, at the rate

$$r_j^{\mathrm{WGF}} = \gamma_j, \quad r_j^{\mathrm{MMD}} = \gamma_j k_j, \quad r_j^{\mathrm{Sob}} = \frac{\gamma_j k_j}{\gamma_j k_j + \lambda}, \quad r_j^{\mathrm{LAWGD}} = 1.$$

For any one of the four flows, write r_j for its corresponding rate and assume $r_j > 0$ for every j. If modes are drawn with probabilities p in the simplex Δ_J, the best worst-case averaged rate is $\max_{p\in\Delta_J} \min_j p_j r_j = (\sum_j r_j^{-1})^{-1}$, attained at $p_j \propto r_j^{-1}$. Equivalently, the optimum is $1/J$ times the harmonic mean of (r_j).

Proof. For a mode with $f_j \neq 0$, the fitted potential is $u = (\ell(e_j)/\|e_j\|_B^2)e_j$, so mode j decays at the rate $r_j = \gamma_j \ell(e_j)/(f_j\|e_j\|_B^2)$. Using $\langle \nabla u, \nabla h\rangle_\pi = \langle u, \mathcal{L}h\rangle_\pi$ and

$$\|C_\pi e_j\|^2 = \sum_{i=1}^{d_\psi} \langle \nabla\psi_i, \nabla e_j\rangle_\pi^2 = \gamma_j^2 \langle e_j, K_\pi e_j\rangle_\pi = \gamma_j^2 k_j,$$

the pairs $(\|e_j\|_B^2, \ell(e_j)/f_j)$ are (γ_j, γ_j) for the Wasserstein gradient flow of KL, from $\ell(h) = \langle f, \mathcal{L}h\rangle_\pi$, then $(\gamma_j, \gamma_j k_j)$ for MMD, from $\ell(h) = \langle K_\pi f, \mathcal{L}h\rangle_\pi$, then $(\gamma_j^2 k_j + \lambda\gamma_j, \gamma_j k_j)$ for Sobolev descent, and (γ_j^2, γ_j) for LAWGD. Substitution gives the four stated rates. The max–min problem is solved by equalizing $p_j r_j$ across modes. □

The proposition gives a mode-wise version of the conditioning comparison in Section 4.3. The calculation evaluates the operators at π and therefore describes the linearized dynamics around the target. The first display gives the decay rate r_j conditional on selecting mode j, and under randomized selection its averaged rate is $p_j r_j$. For LAWGD, $r_j = 1$ and uniform sampling gives the common rate $1/J$. For another flow, the optimized common rate is the harmonic mean of its conditional rates divided by J, and no universal ordering with $1/J$ follows without a normalization of those rates. When the kernel spectrum decays rapidly, some MMD or Sobolev rates can be small, which lowers their harmonic mean. Importance sampling redistributes updates toward these slow modes, whereas the LAWGD geometry removes the mode dependence before sampling.

E.4 Dissipation captured by slices

Conditional-expectation identities from effective dynamics (Legoll & Lelièvre, 2010) quantify how much LAWGD dissipation a lifted one-dimensional Poisson solution captures.

Proposition E.9 (Dissipation captured by a lifted one-dimensional Poisson solution)**.** *In the LAWGD setting, fix a unit vector θ, let $Z = \theta^\top X$ under π have density $p_\theta > 0$, let $V'_\theta(z) = \mathbb{E}_\pi[\theta^\top \nabla V \mid Z = z]$, and let $\mathcal{L}_\theta = -\partial_z^2 + V'_\theta \partial_z$. Suppose that $f_\mu = f_\theta(Z)$ with $f_\theta = \mathcal{L}_\theta h$ and $\mathcal{D}_\theta = \|f_\theta\|^2_{L^2(p_\theta)} > 0$, and that $\varphi = h(\theta^\top x) \in \mathcal{V}$ with $C(h) = \mathbb{E}_\pi[h'(Z)^2\zeta_\theta^2] < \infty$, where $\zeta_\theta = \theta^\top \nabla V - V'_\theta(Z)$.*

- *(i) The subspace $\mathcal{U} = \mathrm{span}\{\varphi\}$ captures the fraction $\mathcal{D}_\mathcal{U}/\mathcal{D}_\star = \mathcal{D}_\theta/(\mathcal{D}_\theta + C(h))$.*
- *(ii) If $\mathrm{Var}_\pi(\theta^\top \nabla V \mid Z) \le \sigma^2$ almost surely and $\mathcal{L}_\theta$ has spectral gap γ_θ, then $\mathcal{D}_\mathcal{U}/\mathcal{D}_\star \ge (1 + \sigma^2/\gamma_\theta)^{-1}$.*

The residual ζ_θ is the part of the directional target score left after conditioning on Z, and $C(h)$ is its cost in the LAWGD norm. A small conditional score variance relative to the marginal spectral gap therefore guarantees that one lifted solution captures a substantial fraction of the full dissipation.

Proof of Proposition E.9. Expanding $\mathcal{L}\varphi(x) = -h''(Z) + h'(Z)\theta^\top \nabla V(x)$ and adding and subtracting $h'(Z)V'_\theta(Z)$ gives $\mathcal{L}\varphi = (\mathcal{L}_\theta h)(Z) + h'(Z)\zeta_\theta$ pointwise. The second term has conditional mean zero given Z, because $\mathbb{E}_\pi[\zeta_\theta \mid Z] = 0$. It is therefore orthogonal in $L^2(\pi)$ to the first term, which is a function of Z, and $C(h) = \|h'(Z)\zeta_\theta\|_\pi^2$.

For (i), $f_\mu = f_\theta(Z)$ sees only the marginal term in the decomposition. Consequently, $\ell(\varphi) = \langle f_\mu, \mathcal{L}\varphi\rangle_\pi = \mathcal{D}_\theta$ and $\|\varphi\|_B^2 = \|\mathcal{L}\varphi\|_\pi^2 = \mathcal{D}_\theta + C(h)$. For a single potential, $\mathcal{D}_\mathcal{U} = \ell(\varphi)^2/\|\varphi\|_B^2$, while $\mathcal{D}_\star = \|f_\mu\|_\pi^2 = \mathcal{D}_\theta$, which gives the ratio in (i).

For (ii), conditioning on Z gives

$$C(h) \le \sigma^2 \int h'(z)^2 p_\theta(z) dz = \sigma^2 \langle f_\theta, \mathcal{L}_\theta^{-1} f_\theta \rangle_{L^2(p_\theta)}.$$

The spectral gap bounds the final inner product by $\mathcal{D}_\theta/\gamma_\theta$, and substitution into part (i) gives the result. □

F Proofs for Section 4

F.1 Proof of Theorem 4.3

Proof. Condition on $\mathcal{F}_n$. Assumption 4.2 and (11) give $\mathbb{E}[F_{n+1} \mid \mathcal{F}_n] \le F_n - c\eta B_n(u_{\star,n}, S_n u_{\star,n})$. In case (a), the capture assumption and $\mathcal{D}_{\star,n} \ge \kappa F_n$ give $\mathbb{E}[F_{n+1} \mid \mathcal{F}_n] \le (1 - c\eta\alpha\kappa)F_n$. Taking expectations and iterating proves the geometric bound. In case (b), (13) gives $B_n(u_{\star,n}, S_n u_{\star,n}) \ge \mathcal{D}_{\star,n}^2/R_0^2 \ge \kappa^2 F_n^2/R_0^2$. Jensen's inequality then gives $\mathbb{E}F_{n+1} \le \mathbb{E}F_n - \beta(\mathbb{E}F_n)^2$. The condition $\beta F_0 < 1$ keeps this scalar recursion in its decreasing regime, and the reciprocal argument $1/\mathbb{E}F_{n+1} \ge 1/\mathbb{E}F_n + \beta$ gives the stated $O(1/n)$ bound. □

F.2 Proof of Example 4.4

The proof reduces to a one-dimensional statement about a single LAWGD step with the exact Poisson velocity.

Lemma F.1 (One-dimensional LAWGD step)**.** *Let $p = Z^{-1}e^{-v}$ on $\mathbb{R}$ with $v \in C^2$, $v'(0) = 0$, and $\lambda \le v'' \le \Lambda$, and write $k = \Lambda/\lambda$. Let μ be a probability measure with $d\mu/dp = r = 1 + f$, where f is continuous and $\|f\|_\infty \le \delta \le \frac{1}{2}$. Let h solve $-h'' + v'h' = f$, with*

$$h'(s) = -\frac{1}{p(s)} \int_{-\infty}^{s} f\,p = \frac{1}{p(s)} \int_{s}^{\infty} f\,p.$$

For $0 < \eta \le 1/(4k)$, set $T = \mathrm{id} - \eta h'$ and $\mu^+ = T_\# \mu$. Then:

(i) $d\mu^+/dp = 1 + f^+$ with f^+ continuous and $\|f^+\|_\infty \le \delta$;

(ii) $\mathrm{KL}(\mu^+\|p) \le \mathrm{KL}(\mu\|p) - \frac{\eta}{2}\|f\|^2_{L^2(p)}$.

Proof. Step 1: bounds on the velocity. Since $p' = -v'p$, we have $-h'' + v'h' = -p^{-1}(ph')'$, so the stated h' solves the Poisson equation. The two expressions agree because $\int fp = 0$. For $s > 0$, strong convexity and $v'(0) = 0$ give $v'(s) > 0$, and convexity gives $v(t) \ge v(s) + v'(s)(t - s)$, so $\int_s^\infty p \le p(s)/v'(s)$. Hence $|v'(s)h'(s)| \le \delta$. The case $s < 0$ is symmetric, using the first expression for h'. Since $h'' = v'h' - f$,

$$\|h''\|_\infty \le 2\delta.$$

For $s \ge 0$ and $t \ge s$, we also have $v(t) - v(s) \ge v'(s)(t - s) + \frac{\lambda}{2}(t - s)^2 \ge \frac{\lambda}{2}(t - s)^2$, so

$$|h'(s)| \le \delta \int_0^\infty e^{-\lambda u^2/2}\, du \le \delta\sqrt{2/\lambda},$$

and the same bound holds for $s < 0$.

Step 2: the transported density. Since $T' = 1 - \eta h'' \ge 1 - 2\eta\delta \ge \frac{3}{4}$ and $T - \mathrm{id}$ is bounded, T is a C^1 diffeomorphism of $\mathbb{R}$. The change of variables formula gives

$$r^+(T(s)) = r(s)\,\frac{\exp\big(v(T(s)) - v(s)\big)}{1 - \eta h''(s)}, \qquad r^+ = \frac{d\mu^+}{dp}.$$

By Taylor's theorem, $v(T(s)) - v(s) = -\eta v'(s)h'(s) + \frac{\eta^2}{2}v''(\xi_s)h'(s)^2$ for some ξ_s. For $|a| \le \frac{1}{2}$, we have $-\log(1 - a) = a + \rho(a)$ with $0 \le \rho(a) \le a^2$, and $a = \eta h''(s)$ satisfies $|a| \le \frac{1}{4}$. Since $-\eta v'h' + \eta h'' = -\eta f$,

$$\log r^+(T(s)) = \log r(s) - \eta f(s) + E(s), \qquad 0 \le E(s) \le \eta^2\Big(\tfrac{\Lambda}{2}h'(s)^2 + h''(s)^2\Big). \tag{21}$$

By Step 1 and $k \geq 1$, $E(s) \leq \eta^2\delta^2(k+4) \leq 5k\eta^2\delta^2$.

Step 3: the density bound is preserved. Write $r^+(T(s)) = \psi(r(s))e^{E(s)}$ with $\psi(a) = a\, e^{-\eta(a-1)}$. Since $\psi'(a) = (1-\eta a)e^{-\eta(a-1)} > 0$ for $a \leq 1+\delta$ (because $\eta(1+\delta) \leq \frac{3}{8}$), ψ is increasing on $[1-\delta, 1+\delta]$. Hence

$$r^+(T(s)) \geq \psi(1-\delta) = (1-\delta)e^{\eta\delta} \geq 1-\delta$$

and

$$r^+(T(s)) \leq \psi(1+\delta)e^{5k\eta^2\delta^2} = (1+\delta)\exp\big(-\eta\delta(1-5k\eta\delta)\big) \leq 1+\delta,$$

because $5k\eta\delta \leq \frac{5}{8}$. Since T is onto and all terms are continuous in s, this proves (i).

Step 4: decrease of KL. Since $\mathrm{KL}(\mu^+\|p) = \int \log r^+ \, d\mu^+ = \int \log r^+(T(s))\, \mu(ds)$, equation (21) gives

$$\mathrm{KL}(\mu^+\|p) - \mathrm{KL}(\mu\|p) = -\eta \int f\, r\, p + \int E\, r\, p.$$

Because $\int fp = 0$, $\int f\, r\, p = \int f(1+f)p = \|f\|^2_{L^2(p)} =: D$. Expanding $f^2 = (-h'' + v'h')^2$ and integrating the cross term by parts,

$$-2\int v'h'h''p = -\int v'p\,(h'^2)' = \int h'^2(v'p)' = \int h'^2(v'' - v'^2)p,$$

where the boundary terms vanish because h' is bounded and $v'p \to 0$ at $\pm\infty$. Hence

$$D = \int \big(h''^2 + v''h'^2\big)p.$$

Since $\frac{\Lambda}{2}h'^2 \leq \frac{k}{2}v''h'^2$ and $k \geq 1$, the bound on E gives $\int E\, r\, p \leq (1+\delta)\eta^2 k D \leq \frac{3}{8}\eta D$. Therefore $\mathrm{KL}(\mu^+\|p) \leq \mathrm{KL}(\mu\|p) - \frac{5}{8}\eta D$, which implies (ii). □

Proof of Example 4.4. We prove by induction that, for every n,

$$\frac{d\mu_n}{d\pi}(x) = 1 + g_n(z_j), \qquad \int g_n\, p_j = 0, \qquad \|g_n\|_\infty \leq \delta, \tag{22}$$

with g_n continuous. The case $n = 0$ is the hypothesis. Suppose (22) holds at step n.

The full potential is one ridge. In the coordinates z, the law μ_n is the product of $\mu_{n,j}$, which has density $(1+g_n)p_j$, with the factors p_i for $i \neq j$. Hence $\mu_{n,i} = p_i$ for $i \neq j$, so $h_{i,n}$ is constant and $\varphi_{i,n} = 0$ in $\mathcal{V}$. Since $V(x) = \sum_i v_i(z_i)$ up to a constant and the e_i are orthonormal, the generator separates as $\mathcal{L} = \sum_i \mathcal{L}_i$, where $\mathcal{L}_i = -\partial^2_{z_i} + v_i'(z_i)\partial_{z_i}$. Therefore

$$\mathcal{L}\big[h_{j,n}(z_j)\big] = (\mathcal{L}_j h_{j,n})(z_j) = g_n(z_j) = f_{\mu_n}(x),$$

so $u_{\star,n} = \varphi_{j,n}$ and $\mathcal{D}_{\star,n} = \chi^2(\mu_n\|\pi) = \int g_n^2 p_j$. The target satisfies a Poincaré inequality with constant $\min_i \lambda_i$ by the Bakry–Émery criterion, and f_{μ_n} is bounded, so Assumption 4.1 holds at μ_n. Consequently, $u_{\mathcal{U}_n} = u_{\star,n}$ if index j is drawn at least once, and $u_{\mathcal{U}_n} = 0$ otherwise.

Invariance and one-step descent. If j is not drawn, then $\mu_{n+1} = \mu_n$ and $\mathcal{D}_{\mathcal{U}_n} = 0$. If j is drawn, then $\nabla u_{\mathcal{U}_n}(x) = h'_{j,n}(z_j)e_j$, so the update moves only the coordinate z_j, by the map $T_j = \mathrm{id} - \eta h'_{j,n}$. Thus μ_{n+1} is the product of $(T_j)_\#\mu_{n,j}$ with the factors p_i, $i \neq j$. Lemma F.1, applied with $(p, \mu, f) = (p_j, \mu_{n,j}, g_n)$ and $k = \Lambda_j/\lambda_j \leq K$, gives (22) at step $n+1$. Since the KL divergence between two products that differ in one factor equals the KL divergence of that factor, it also gives

$$F_{n+1} \leq F_n - \tfrac{\eta}{2}\,\mathcal{D}_{\star,n} = F_n - \tfrac{\eta}{2}\,\mathcal{D}_{\mathcal{U}_n}.$$

This completes the induction and proves Assumption 4.2 with $c = \frac{1}{2}$ and $\eta_0 = 1/(4K)$.

Full-flow inequality. Since $\log r \leq r-1$, we have $F_n = \int r \log r\, d\pi \leq \int r(r-1)\, d\pi = \chi^2(\mu_n\|\pi) = \mathcal{D}_{\star,n}$, so $\kappa = 1$.

Part (a). The proposal law is $\mathcal{F}_n$-measurable because the profiles are functions of μ_n. By (22) and the dichotomy above,

$$B_n(u_{\star,n}, S_n u_{\star,n}) = \mathbb{E}[\mathcal{D}_{\mathcal{U}_n} \mid \mathcal{F}_n] = \big(1 - (1-\varpi_j)^m\big)\mathcal{D}_{\star,n} \geq \alpha\, \mathcal{D}_{\star,n}.$$

Moreover, $c\eta\alpha\kappa = \eta\alpha/2 \le 1$.

Part (b). At iteration n, every atom except $\varphi_{j,n} = u_{\star,n}$ vanishes. The normalized atoms are therefore $\widetilde{\varphi}_{i,n} = 0$ for $i \neq j$ and $\widetilde{\varphi}_{j,n} = u_{\star,n}/\|u_{\star,n}\|_{B_n}$. A representation $u_{\star,n} = \sum_i \varpi_i c_i \widetilde{\varphi}_{i,n}$ requires $\varpi_j c_j = \mathcal{D}_{\star,n}^{1/2}$, and the minimal choice sets $c_i = 0$ for $i \neq j$. Hence

$$R_{\nu_n}(u_{\star,n})^2 = \varpi_j c_j^2 = \frac{\mathcal{D}_{\star,n}}{\varpi_j} \le \frac{\delta^2}{\varpi_{\min}},$$

using $\mathcal{D}_{\star,n} = \int g_n^2 p_j \le \delta^2$. This bound holds trivially when $\mathcal{D}_{\star,n} = 0$. With $\beta = c\eta\kappa^2/R_0^2 = \eta\varpi_{\min}/(2\delta^2)$, we have $\beta F_0 \le \beta \mathcal{D}_{\star,0} \le \eta\varpi_{\min}/2 < 1$. The two rates now follow from Theorem 4.3.

Instances. For $\pi = \mathcal{N}(0,\Sigma)$, $v_i(s) = s^2/(2\sigma_i^2)$, so $\lambda_i = \Lambda_i$ and $K = 1$. The function $\sin(z_j/\sigma_j)$ is odd and p_j is even, so g_0 is centered. For $v_i(s) = s^2/2 + a\log\cosh s$, we have $v_i'(s) = s + a\tanh s$ and $v_i''(s) = 1 + a\,\mathrm{sech}^2 s \in [1, 1+a]$, so $K = 1+a$, and $\sin(z_j)$ is centered because p_i is even. □

Remark F.2 (Translations: part (a) without part (b)). Consider the translation model of Section 4.3 with $\pi = \mathcal{N}(0,\Sigma)$, $\mu_n = \mathcal{N}(\tau_n,\Sigma)$, and linear potentials $\theta^\top x$ with $\theta \sim \mathcal{N}(0,\Sigma)$. Write $y_n = \Sigma^{-1/2}\tau_n$. The discrete version of Proposition 4.5 gives $y_{n+1} = (I - \eta P_n)y_n$, where P_n projects onto the span of $m \le d$ independent standard Gaussian vectors. Since $F_n = \frac{1}{2}\|y_n\|^2$ and $\mathcal{D}_{\mathcal{U}_n} = \|P_n y_n\|^2$,

$$F_{n+1} = F_n - \eta(1-\eta/2)\,\mathcal{D}_{\mathcal{U}_n},$$

so Assumption 4.2 holds with $c = \frac{1}{2}$ for $\eta \le 1$. By rotation invariance, $\mathbb{E}[\mathcal{D}_{\mathcal{U}_n} \mid \mathcal{F}_n] = \frac{m}{d}\|y_n\|^2$. The full LAWGD dissipation is $\mathcal{D}_{\star,n} = \chi^2(\mu_n\|\pi) = e^{\|y_n\|^2} - 1$. Since $\|y_n\|$ is nonincreasing and $s \mapsto s/(e^s - 1)$ is decreasing, part (a) holds with $\kappa = 1$ and $\alpha = \frac{m}{d}\, s_0/(e^{s_0} - 1)$, where $s_0 = \|y_0\|^2$. Part (b) fails, however. The density error $f_{\mu_n}(x) = \exp\left(y_n^\top \Sigma^{-1/2} x - \frac{1}{2}\|y_n\|^2\right) - 1$ is not linear, while $\mathcal{L}$ maps linear functions to linear functions. Hence $u_{\star,n}$ does not lie in the span of linear potentials, and $R_{\nu_n}(u_{\star,n}) = \infty$ whenever $\tau_n \neq 0$.

F.3 Proof of Proposition 4.5

Assumption F.3 (Translation model)**.** *The potential V is C^2, π has finite second moments, and $V(\cdot+\tau)$, $\|\nabla V(\cdot+\tau)\|^2$, and $\|\nabla^2 V(\cdot+\tau)\|$ are dominated by a π-integrable function locally uniformly in τ. Moreover, $I_F = \mathbb{E}_\pi[\nabla V \nabla V^\top] = \mathbb{E}_\pi \nabla^2 V \succ 0$, and centered linear functions lie in $D(\mathcal{L})$ with $\mathcal{L}(\theta^\top x) = \theta^\top \nabla V$.*

Proof of Proposition 4.5. Fix t and write $\Theta = \Theta_t$ and $\tau = \tau_t$. For a linear potential, the spatial gradient is constant. The continuity equation therefore translates the law and gives $\dot\tau_t = -\Theta a$, where $a = G^\dagger b$ is evaluated at μ_τ. It remains to identify G and b for each flow. For the KL directional derivative $\ell(h) = \mathbb{E}_\mu[\mathcal{L}h]$,

$$b_j = \mathbb{E}_{\mu_\tau}[\theta_j^\top \nabla V] = \theta_j^\top \nabla g(\tau).$$

The kinetic Gram matrix is $\Theta^\top\Theta$, whereas the LAWGD Gram matrix is $\mathbb{E}_\pi[(\theta_i^\top\nabla V)(\theta_j^\top \nabla V)] = (\Theta^\top I_F \Theta)_{ij}$. For $\psi = Wx$, we have $\delta_{\mu_\tau} = W\tau$ and $J = W\Theta$. Hence $b = J^\top \delta_{\mu_\tau} = \Theta^\top W^\top W\tau$, the MMD Gram matrix is $\Theta^\top\Theta$, and the Sobolev Gram matrix is $\Theta^\top(W^\top W + \lambda I_d)\Theta$. Thus, in every case, $G = \Theta^\top H \Theta$ and $b = \Theta^\top \nabla \bar F \in \operatorname{range} G$, and substitution into $\dot\tau_t = -\Theta G^\dagger b$ proves (14). With $\Theta = I_d$, the LAWGD flow is $\dot\tau = -\nabla^2 g(0)^{-1}\nabla g(\tau)$ and the Sobolev flow is $\dot\tau = -(W^\top W + \lambda I_d)^{-1} W^\top W\tau$. For a Gaussian target, $g(\tau) = \frac{1}{2}\tau^\top\Sigma^{-1}\tau$ and $I_F = \Sigma^{-1}$, so LAWGD gives $\dot\tau = -\tau$. □

F.4 Contraction and optimal geometry in the quadratic translation model

Contraction of the discrete scheme. The discrete scheme is $\tau_{n+1} = \tau_n - \eta\Theta_n(\Theta_n^\top H\Theta_n)^\dagger\Theta_n^\top Q\tau_n$ with independent $\Theta_n \sim \mathsf{P}$. Write $\Pi = \Theta(\Theta^\top H\Theta)^\dagger\Theta^\top$ and $y = Q\tau$. Since $\Pi H \Pi = \Pi$ and $Q \preceq L_H H$,

$$\bar F(\tau - \eta\Pi y) = \bar F(\tau) - \eta y^\top \Pi y + \tfrac{\eta^2}{2} y^\top \Pi Q \Pi y \le \bar F(\tau) - \left(\eta - \tfrac{\eta^2 L_H}{2}\right) y^\top \Pi y.$$

With $\eta = 1/L_H$ and $\mathbb{E} y^\top \Pi y = \tau^\top Q A_{\mathsf{P}} Q\tau \ge r(\mathsf{P})\tau^\top Q\tau = 2r(\mathsf{P})\bar F(\tau)$, this gives $\mathbb{E}\bar F(\tau_{n+1}) \le (1 - r(\mathsf{P})/L_H)\bar F(\tau_n)$. The trace identity follows from $\operatorname{tr}(H\Pi) = \operatorname{rank}\Theta$.

The comparison of geometries in Section 4.3 rests on the following result.

Theorem F.4 (Optimal geometry in the quadratic translation model)**.** *For every $H \succ 0$ and every law P supported on matrices of rank at most m,*

$$\frac{r(\mathsf{P})}{L_H} \le \frac{m}{L_H \operatorname{tr}(HQ^{-1})} \le \frac{m}{d}.$$

The second inequality is an equality if and only if H is a multiple of Q. For $H = Q$ and $1 \le m \le d$, the Gaussian sketch $\Theta = Q^{-1/2}Z$, with $Z \in \mathbb{R}^{d\times m}$ having i.i.d. $\mathcal{N}(0,1)$ entries, attains $r(\mathsf{P}) = m/d$. For general H and $m = 1$, the bound $m/\operatorname{tr}(HQ^{-1})$ is attained by $\theta = H^{-1/2}v_i$, where v_i are the eigenvectors of $H^{1/2}Q^{-1}H^{1/2}$, each chosen with probability proportional to its eigenvalue.

Proof. If $Q^{1/2}A_{\mathsf{P}}Q^{1/2} \succeq r(\mathsf{P})I_d$, then $A_{\mathsf{P}} \succeq r(\mathsf{P})Q^{-1}$. Taking the H-weighted trace gives $m \ge \operatorname{tr}(HA_{\mathsf{P}}) \ge r(\mathsf{P})\operatorname{tr}(HQ^{-1})$. Next, $Q \preceq L_H H$ gives $H^{1/2}Q^{-1}H^{1/2} \succeq L_H^{-1}I_d$, so $\operatorname{tr}(HQ^{-1}) \ge d/L_H$, with equality exactly when $H = L_H^{-1}Q$. For the Gaussian sketch with $m \le d$, $Q^{1/2}A_{\mathsf{P}}Q^{1/2} = \mathbb{E}[P_{\operatorname{range} Z}] = (m/d)I_d$ by rotation invariance, and $L_Q = 1$. For $m = 1$, write $\Gamma = H^{1/2}Q^{-1}H^{1/2}$ with eigenpairs (ς_i, v_i), and draw $\theta = H^{-1/2}v_i$ with probability $\varsigma_i/\operatorname{tr}\Gamma$. Since $\theta^\top H\theta = 1$ for $\theta = H^{-1/2}v_i$,

$$A_{\mathsf{P}} = \sum_i \frac{\varsigma_i}{\operatorname{tr}\Gamma} H^{-1/2}v_i v_i^\top H^{-1/2} = \frac{Q^{-1}}{\operatorname{tr}\Gamma},$$

so $r(\mathsf{P}) = 1/\operatorname{tr}\Gamma = 1/\operatorname{tr}(HQ^{-1})$, which attains the first bound of the theorem. □

The Gaussian example. Let $\pi = \mathcal{N}(0,\Sigma)$, so that $Q = \Sigma^{-1}$, and take $m = 1$. In the kinetic geometry $H = I_d$, Theorem F.4 gives the best value $r(\mathsf{P}) = 1/\operatorname{tr}\Sigma$, attained by eigenvectors of Σ drawn with probability proportional to their eigenvalues. Uniform directions give $A_{\mathsf{P}} = I_d/d$ and hence $r(\mathsf{P}) = \lambda_{\min}(\Sigma^{-1})/d$. In the LAWGD geometry $H = \Sigma^{-1}$, the Gaussian sketch $\theta \sim \mathcal{N}(0,\Sigma)$ gives $r(\mathsf{P}) = 1/d$ and $L_H = 1$.

G Supplementary Experiments

G.1 Particle experiments on non-Gaussian targets

We test Algorithm 1 on two-dimensional Gaussian mixture targets, whose densities, scores, and marginals are available in closed form. Throughout, u and v denote orthonormal vectors in $\mathbb{R}^2$.

For MMD and Sobolev descent, we minimize the fixed-feature objective $F_\psi(\mu) = \frac{1}{2}\|\mathbb{E}_\mu\psi - \mathbb{E}_\pi\psi\|^2$, where ψ stacks random Fourier ridge features $\cos(\omega\,\theta_r^\top x)$ and $\sin(\omega\,\theta_r^\top x)$ at three frequencies along a fixed pool of R random directions θ_r, for $6R$ features in total. Sobolev descent uses $\lambda = 0.02$. LAWGD targets $\mathrm{KL}(\mu\|\pi)$. At each refresh, we draw m directions from the pool with replacement. Each draw θ_r supplies one trial potential $h_r(\theta_r^\top x)$, whose profile is a combination of the six features along θ_r obtained by solving the flow's problem on the one-dimensional marginals: the marginal MMD witness for MMD, the marginal Sobolev potential for Sobolev, and a regularized Galerkin approximation of the marginal Poisson solution for LAWGD.

We compare three ways of turning these potentials into a velocity field. Projection in Algorithm 1 normalizes the m trial potentials and refits their coefficients at every update by solving $(\widehat{G}+\varepsilon I_m)a = \widehat{b}$ in the flow's own geometry. Uniform slicing averages the same marginal potentials with weights $1/m$. And, for each flow, we give a Reference that solves the full problem over all $6R$ features, using the exact finite-feature direction for MMD and Sobolev and a dense Galerkin solve for LAWGD.

All methods use explicit Euler steps with a backtracking step-size rule that enforces descent of the objective. We report the unnormalized objective F_ψ (denoted F in the plots) or KL, evaluated on held-out test particles that are used neither for fitting nor for step selection. KL is computed by propagating the source log-density through the exact Jacobians of the transport maps. All horizontal axes show wall-clock time.

G.1.1 CONCENTRATED VERSUS UNIFORM DIRECTION PROPOSALS

The target is $\pi = \frac{1}{2}\mathcal{N}(-2u, I_2) + \frac{1}{2}\mathcal{N}(2u, I_2)$, and the source is π translated by $1.5v$. We use $R = 64$ directions (384 features), $m = 24$ trial potentials refreshed every 10 updates, and $\varepsilon = 0.01$. We compare two proposal laws on the same pool of directions. The first is uniform. The second is concentrated, placing most of its mass on a narrow cone of directions about 15° from the mixture axis u. The concentrated law produces many nearly redundant potentials, which is where the two ways of combining them should differ most.

Figure 2 shows the results. Under concentrated proposals, uniform slicing makes little progress for all three flows, because averaging many nearly collinear potentials yields a direction dominated by a few correlated slices. Projection fits the potentials jointly and continues to decrease the objective. It plateaus above the Reference because the concentrated law covers only a limited set of directions. Under uniform proposals, both methods improve, but Projection still decreases the objective faster than uniform slicing and approaches the Reference.

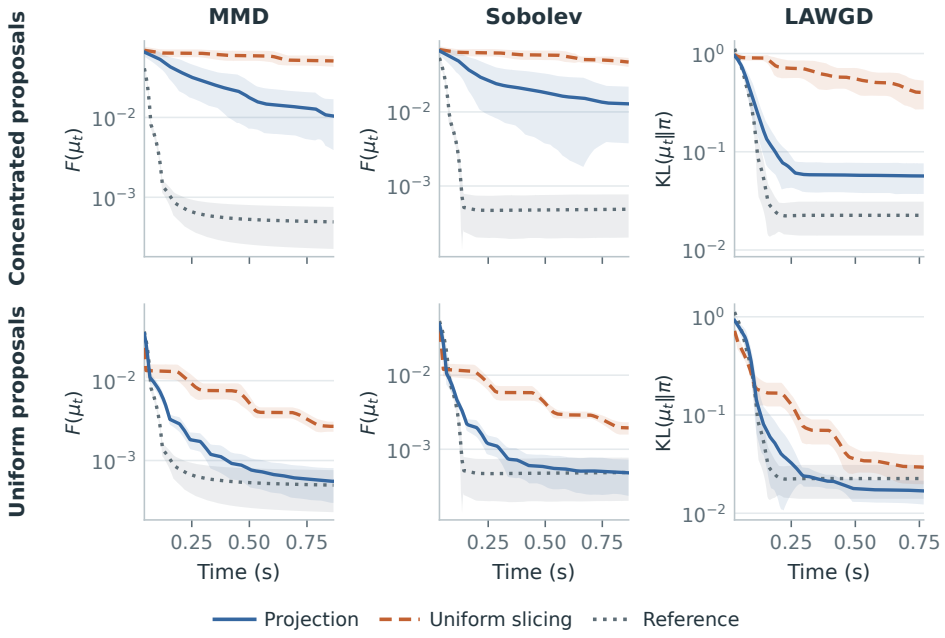

Figure 2: Projection versus uniform slicing on a Gaussian mixture. Columns show MMD, Sobolev, and LAWGD; rows show concentrated and uniform direction proposals. Vertical axes show F_ψ or KL on logarithmic scales against wall-clock time. Lines and bands show the mean and one standard deviation over eight runs. Reference uses all 384 features and is shared between rows.

G.1.2 LAWGD UNDER COVARIANCE ANISOTROPY

We next vary the conditioning of the target. The target is $\pi_\kappa = \frac{1}{2}\mathcal{N}(-2u, \Sigma_\kappa) + \frac{1}{2}\mathcal{N}(2u, \Sigma_\kappa)$ with $\Sigma_\kappa = uu^\top + \kappa^{-1}vv^\top$ and $\kappa \in \{1, 25, 100\}$. The Gaussian source is chosen so that the initial KL and χ^2 divergences are identical across κ, so differences between panels reflect conditioning alone. We draw $m = 16$ potentials uniformly from $R = 32$ directions (192 features), refresh them at every update, and use $\varepsilon = 10^{-3}$. For each κ, all methods are shown over a common wall-clock window.

Figure 3 shows that at $\kappa = 1$ all three methods converge at comparable speed, with uniform slicing slightly faster early on. As κ grows, uniform slicing slows markedly, while Projection remains close to the Reference. This is consistent with Section 4: the projected fit retains the LAWGD geometry through its Gram matrix, whereas averaging marginal potentials ignores the interactions between directions that the geometry accounts for.

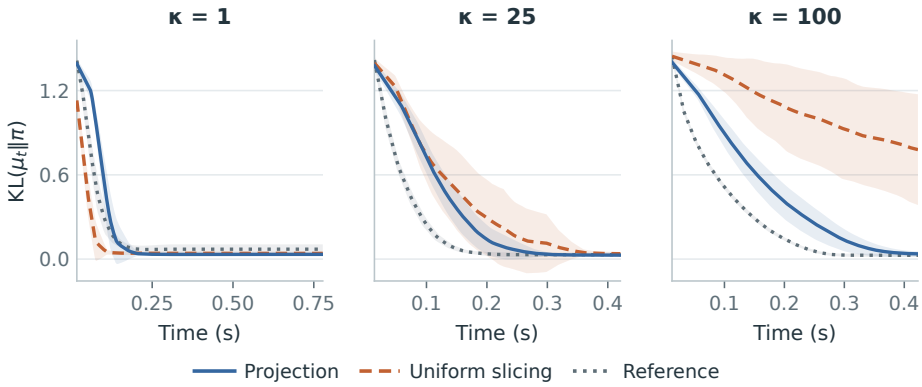

Figure 3: LAWGD at equal computational cost under anisotropy. Panels vary the component-covariance condition number κ with fixed initial KL and χ^2 divergence. Curves show test KL against wall-clock time, with mean and one standard deviation over eight runs. Reference is dense regularized Galerkin LAWGD in the 192-feature bank.

G.1.3 Number of trial potentials at fixed runtime

Finally, we examine what happens when we vary the number of trial potentials. We return to the concentrated-proposal problem of Appendix G.1.1 and vary $m \in \{8, 16, 24, 48\}$, keeping all other settings fixed. Each flow is run for a fixed wall-clock budget shared by both methods and all values of m, so larger m is charged its additional fitting and transport cost.

Figure 4 shows that Projection improves steadily with m for all three flows, even at equal runtime. Uniform slicing is essentially insensitive to m, since additional draws from the concentrated proposal only average in more redundant potentials.

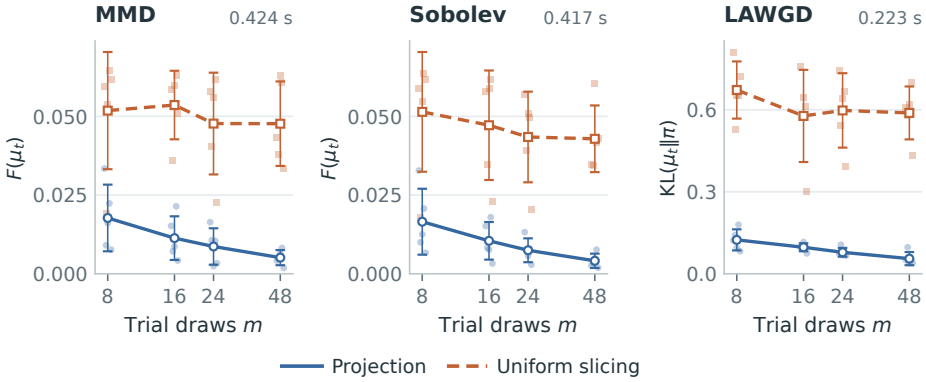

Figure 4: **Additional trial potentials at a fixed wall-time budget.** Concentrated-proposal mixture experiment with varying number of trial draws m. All points in a panel share the runtime printed above it. Vertical axes show F_ψ for MMD/Sobolev and KL for LAWGD. Large markers and error bars show the mean and one standard deviation over five runs; faint points show individual runs.