TANGENT-SPACE PRECONDITIONING OF WASSERSTEIN GRADIENT FLOWS FOR PARTICLE SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a simple geometric perspective on Wasserstein gradient flows for designing efficient particle-based sampling algorithms in high-dimensional settings. The central idea is to precondition the Wasserstein gradient flow in its tangent space, compensating for slowly relaxing directions that can impede sampling. Remarkably, we show that using the 2-Wasserstein Hessian of the chi-squared flow at the target as a preconditioner recovers the idealized population dynamics underlying Laplacian-adjusted Wasserstein gradient descent (LAWGD), providing a geometric interpretation of LAWGD as Wasserstein preconditioning. This perspective further identifies a scalar-to-vector transport operator as the key object to approximate, leading to two principled finite-sample realizations based on kernels and neural networks. Experiments demonstrate improved sampling efficiency and reveal complementary regimes for the two realizations, with kernels performing well in low dimensions and neural approximations becoming increasingly advantageous as dimension grows.

1 INTRODUCTION

Sampling from a target probability distribution is a central problem in Bayesian inference, statistical computation, and generative modeling. A broad class of sampling methods can be understood through dynamics that evolve an initial distribution toward the target, with Langevin diffusion providing a canonical example (Liu & Wang, 2016; Liu, 2017). Notably, Jordan et al. showed that the marginal law of Langevin diffusion follows the Wasserstein gradient flow of the Kullback–Leibler divergence (Jordan et al., 1998), establishing a connection between sampling, optimal transport, and optimization that has since motivated a large body of work (Ambrosio et al., 2008; Villani, 2009; Wibisono, 2018; Santambrogio, 2017; Dalalyan, 2017). As in ordinary gradient descent, however, the resulting dynamics can relax at very different rates along different directions, with the slowest spectral modes governing the long-time convergence to equilibrium (Pavliotis, 2014).

A natural way to compensate for disparate relaxation rates is to exploit the spectral structure of the target dynamics. Laplacian-adjusted Wasserstein gradient descent (LAWGD) (Chewi et al., 2020) modifies the chi-squared Wasserstein flow using inverse spectral information from the associated Langevin operator, thereby accelerating slowly relaxing modes. Chewi et al. (2020) proved substantially improved convergence behavior for the resulting sampling dynamics. Subsequent work has developed data-driven implementations of this idea by approximating the target dynamics using kernel-based spectral methods (Li & Marzouk, 2025; Xu et al., 2025). These developments raise two related questions. First, is there a geometric principle that explains why this particular inverse operator is the appropriate correction to the Wasserstein dynamics? Second, how should this correction be represented and learned in practice, especially in higher-dimensional settings where kernel spectral approximations can become increasingly costly?

We answer the first question using the second-order geometry of Wasserstein space. In Wasserstein calculus, infinitesimal changes of a probability distribution are represented by tangent velocity fields, and the Wasserstein gradient of a functional specifies the velocity field of its steepest descent (Otto, 2001; Ambrosio et al., 2008). The Wasserstein Hessian is the corresponding second-order object: it describes how this gradient changes along tangent directions and therefore captures the local curvature of the objective (Villani, 2009). At the target distribution, the Wasserstein Hessian of the chi-squared objective is the Witten Laplacian (Helffer & Nier, 2005), the vector-field counterpart of

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.

the Langevin operator; this equilibrium curvature is shared, up to a scalar factor, by a broad class of f-divergences. Freezing this Hessian at the target and applying its inverse to the chi-squared gradient yields a Newton-type preconditioned flow that recovers the idealized population dynamics underlying LAWGD (Chewi et al., 2020). The inverse spectral correction in LAWGD can therefore be understood as compensating for the local Wasserstein curvature at equilibrium. Because the chi-squared gradient is linear in the density ratio, the resulting preconditioned velocity is also linear in the scalar density discrepancy from the target and can therefore be represented by a fixed target-dependent scalar-to-vector operator.

This observation also suggests taking this scalar-to-vector mapping itself as the computational object. Rather than first approximating an inverse scalar operator and subsequently differentiating its output to obtain the particle velocity as in (Chewi et al., 2020; Xu et al., 2025), we directly learn the mapping from the scalar density discrepancy to the resulting particle velocity field. For practical particle sampling, we develop kernel and neural low-rank implementations of this direct transport formulation. Kernel methods provide a direct nonparametric approximation and can be especially effective in lower-dimensional settings, while neural low-rank models offer a more flexible alternative whose approximation behavior can be more favorable as the ambient dimension grows, at the cost of nonconvex optimization. Our experiments examine this tradeoff and show that the direct transport formulation can be implemented effectively with both approaches.

2 PRELIMINARIES

We first introduce the Wasserstein geometric framework needed to describe sampling dynamics as optimization over probability distributions. We then define Wasserstein gradients and Hessians, which provide the first- and second-order objects underlying the Newton-type viewpoint developed in Section 3. Finally, we specialize these notions to the chi-squared objective and its Wasserstein gradient flow, which will serve as the starting point for our preconditioned dynamics.

2.1 WASSERSTEIN GRADIENT FLOWS

Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$ with finite second moment, equipped with the 2-Wasserstein distance. Wasserstein calculus provides a notion of differentiation for functionals $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and thereby allows optimization over probability distributions to be described in a form analogous to ordinary gradient descent; see Otto (2001), Ambrosio et al. (2008), and Villani (2009) for detailed treatments.

The first variation $\frac{\delta \mathcal{F}}{\delta \rho}(\rho) : \mathbb{R}^d \to \mathbb{R}$ describes the first-order change of $\mathcal{F}$ under a perturbation of the density ρ. The corresponding Wasserstein gradient is the vector field $\operatorname{grad}_{W_2} \mathcal{F}(\rho) = \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho)$. Unlike the gradient in finite-dimensional optimization, which is a single vector, the Wasserstein gradient is a vector field $\mathbb{R}^d \to \mathbb{R}^d$: it assigns a velocity to every point $\mathbf{x}$ in the sample space and therefore specifies how probability mass should move in order to decrease the objective.

More precisely, if a time-dependent density ρ_t is transported by a velocity field $\mathbf{v}_t : \mathbb{R}^d \to \mathbb{R}^d$, conservation of probability mass is expressed by the continuity equation $\partial_t \rho_t + \nabla \cdot (\rho_t \mathbf{v}_t) = 0$. The Wasserstein gradient flow of $\mathcal{F}$ chooses the steepest-descent velocity $\mathbf{v}_t = -\operatorname{grad}_{W_2} \mathcal{F}(\rho_t)$, and hence evolves according to $\partial_t \rho_t = \nabla \cdot (\rho_t \operatorname{grad}_{W_2} \mathcal{F}(\rho_t))$. Thus, Wasserstein gradient descent can be viewed as continuously transporting probability mass according to a velocity field determined by the objective. This first-order viewpoint will be the starting point for defining Wasserstein Hessians and Newton-type directions in the next subsection.

2.2 WASSERSTEIN HESSIANS AND NEWTON DIRECTIONS

The Wasserstein gradient introduced above is a vector field that depends on the current distribution ρ. To define a second-order analogue of Euclidean optimization, we therefore need to describe the space in which these vector fields live and how the gradient changes as ρ moves. At a distribution ρ, the Wasserstein tangent space $T_\rho \mathcal{P}_2$ consists, informally, of velocity fields that describe infinitesimal motion of ρ in Wasserstein space. In the 2-Wasserstein geometry, tangent directions are represented by gradient vector fields $\boldsymbol{\xi} = \nabla \varphi$, where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a scalar potential. More precisely, $T_\rho \mathcal{P}_2$ is the $L^2(\rho)$-closure of smooth gradient fields. For vector fields $\boldsymbol{\xi}, \boldsymbol{\zeta} \in T_\rho \mathcal{P}_2$, the tangent-space inner

product is $\langle \boldsymbol{\xi}, \boldsymbol{\zeta} \rangle_{T_\rho \mathcal{P}_2} := \int_{\mathbb{R}^d} \boldsymbol{\xi}(\mathbf{x})^\top \boldsymbol{\zeta}(\mathbf{x}) \rho(\mathbf{x})\, d\mathbf{x}$. Thus, both the Wasserstein gradient $\operatorname{grad}_{W_2} \mathcal{F}(\rho)$ and the velocity field generating a Wasserstein flow are elements of $T_\rho \mathcal{P}_2$. We refer the reader to Otto (2001); Ambrosio et al. (2008); Villani (2009) for the precise construction.

The Wasserstein Hessian describes how the Wasserstein gradient changes when the distribution is perturbed along one of these tangent directions. Formally, this change is defined through the covariant derivative of the Wasserstein gradient; we suppress the differential-geometric construction here and refer to Ambrosio et al. (2008); Villani (2009) for details. For our purposes, it is useful to view the Hessian as a linear operator $\mathsf{H}_\rho : T_\rho \mathcal{P}_2 \to T_\rho \mathcal{P}_2$. Given a tangent vector field $\boldsymbol{\xi} \in T_\rho \mathcal{P}_2$, the vector field $\mathsf{H}_\rho \boldsymbol{\xi}$ represents the first-order change of the Wasserstein gradient when ρ moves infinitesimally in the direction $\boldsymbol{\xi}$. Equivalently, the associated quadratic form $\langle \boldsymbol{\xi}, \mathsf{H}_\rho \boldsymbol{\xi} \rangle_{T_\rho \mathcal{P}_2}$ measures the second-order change of $\mathcal{F}$ along that direction. In this sense, H_ρ plays the same role as the Hessian matrix in finite-dimensional optimization.

This analogy also gives a Wasserstein version of the Newton direction. In Euclidean optimization, the Newton step applies the inverse Hessian to the gradient; similarly, when H_ρ is invertible on the relevant tangent space, the Wasserstein Newton velocity field $\boldsymbol{\xi}_{\mathrm{N}} \in T_\rho \mathcal{P}_2$ is defined by

$$\boldsymbol{\xi}_{\mathrm{N}} = -\mathsf{H}_\rho^{-1} \operatorname{grad}_{W_2} \mathcal{F}(\rho). \tag{1}$$

2.3 The chi-squared Wasserstein flow

The preceding framework applies to many objectives over probability distributions. A canonical example is the Kullback–Leibler divergence, whose Wasserstein gradient flow gives the Fokker–Planck dynamics of Langevin diffusion (Jordan et al., 1998) and which also underlies Stein Variational Gradient Descent (SVGD) (Liu & Wang, 2016). Here, following Chewi et al. (2020), we instead consider the chi-squared divergence. The key reason is that its Wasserstein gradient depends linearly on the density-ratio discrepancy $\rho/\pi - 1$. As we will see, this linear structure allows the preconditioned dynamics to be represented by a fixed linear operator and later evaluated from samples without estimating the density ratio pointwise. Concretely, we take the objective to be $\mathcal{F}(\rho) = \frac{1}{2}\chi^2(\rho\|\pi) = \frac{1}{2}\int_{\mathbb{R}^d} (\frac{\rho(\mathbf{x})}{\pi(\mathbf{x})} - 1)^2 \pi(\mathbf{x})\, d\mathbf{x}$. Defining the density-ratio discrepancy $u_\rho(\mathbf{x}) := \frac{\rho(\mathbf{x})}{\pi(\mathbf{x})} - 1$, its first variation is $\frac{\delta \mathcal{F}}{\delta \rho}(\rho) = u_\rho$ up to an additive constant, and hence its Wasserstein gradient is $\operatorname{grad}_{W_2} \mathcal{F}(\rho) = \nabla u_\rho$.

The corresponding Wasserstein gradient flow therefore transports probability mass with velocity $\mathbf{v}_\rho = -\nabla u_\rho$, and the density evolves according to $\partial_t \rho_t = \nabla \cdot (\rho_t \nabla u_{\rho_t})$, which we call the *chi-squared flow* (Chewi et al., 2020).

3 Tangent-Space Preconditioned Wasserstein Gradient Flow

In this section, we derive a Newton-type preconditioned Wasserstein gradient flow, identify the scalar-to-vector transport operator governing this flow (Section 3.1), and characterize its spectral structure (Section 3.2) We then develop a representation-independent low-rank approximation and analyze the convergence behavior of the resulting population dynamics, together with a corresponding particle sampling algorithm (Section 3.3).

3.1 Newton-Type Preconditioning with Frozen Hessian

The Wasserstein Newton direction in eq. (1) uses the Hessian at the current distribution. Because this operator varies with ρ, a literal Newton method would require recomputing and inverting a new Hessian throughout the flow. We instead freeze the curvature at the target distribution π. For the chi-squared objective, this target Hessian has a particularly simple operator form, which we now characterize. In what follows, we write $\mathbb{E}_\rho[f] := \int f\, d\rho$, and use $\langle \cdot, \cdot \rangle_\rho$ and $\| \cdot \|_\rho$ for the corresponding inner product and norm.

Let $\pi(\mathbf{x}) \propto e^{-V(\mathbf{x})}$ be the target distribution, and write $\mathcal{D} := \nabla$. Given π, there exists an associated positive Langevin operator on scalar functions defined as $\mathcal{L}_0 := \mathcal{D}_\pi^* D = -\Delta + \nabla V \cdot \nabla$, where $\mathcal{D}_\pi^*$ denotes the adjoint of $\mathcal{D}$ in $L^2(\pi)$. Since the Hessian acts on tangent vector fields rather than scalar functions, we also need the corresponding operator on vector fields. Its vector-field counterpart is

the *degree-one Witten Laplacian* $\mathcal{L}_1$ (Helffer & Nier, 2005) defined as, $\mathcal{L}_1\boldsymbol{\xi} := -\Delta\boldsymbol{\xi} + (\nabla V \cdot \nabla)\boldsymbol{\xi} + \nabla^2 V\,\boldsymbol{\xi}$. On sufficiently regular scalar functions, these operators satisfy the *intertwining identity*,

$$\mathcal{L}_1\mathcal{D} = \mathcal{D}\mathcal{L}_0. \tag{2}$$

We work under regularity and decay conditions that justify the operator identities and integrations by parts below. Whenever an inverse is used, we assume that π satisfies a Poincaré inequality, so that $\mathcal{L}_0$ is invertible on $L^2_0(\pi) := \{f \in L^2(\pi) : \mathbb{E}_\pi[f] = 0\}$. Throughout, $\mathcal{L}_0^{-1}$ acts on this mean-zero space, and $\mathcal{L}_1^{-1}$ denotes the inverse of $\mathcal{L}_1|_{T_\pi\mathcal{P}_2}$, the restriction of $\mathcal{L}_1$ to $T_\pi\mathcal{P}_2$.

Theorem 1 (Equilibrium Wasserstein Hessian)**.** *Let $\epsilon \mapsto \rho_\epsilon$, defined for $|\epsilon| < \epsilon_0$, be a smooth Wasserstein geodesic with $\rho_0 = \pi$ and initial tangent velocity $\mathcal{D}\phi$, where ϕ is sufficiently regular for the following expressions to be finite. Then $\frac{d^2}{d\epsilon^2}|_{\epsilon=0}\mathcal{F}(\rho_\epsilon) = \|\mathcal{L}_0\phi\|^2_\pi = \langle \mathcal{D}\phi, \mathcal{L}_1\mathcal{D}\phi\rangle_\pi$. Consequently, as an operator on the Wasserstein tangent space at π, $\mathsf{H}_\pi = \mathrm{Hess}_{W_2}\mathcal{F}(\pi) = \mathcal{L}_1|_{T_\pi\mathcal{P}_2}$.*

We use this equilibrium curvature to define the *frozen-Newton* velocity,

$$\mathbf{v}_{\mathrm{FN}}(\rho) := -\mathsf{H}_\pi^{-1}\,\mathrm{grad}_{W_2}\,\mathcal{F}(\rho) = -\mathcal{L}_1^{-1}Du_\rho. \tag{3}$$

Remarkably, this frozen-Newton velocity is exactly the *idealized* population velocity underlying LAWGD. In the kernel formulation of Chewi et al. (2020), choosing the scalar kernel operator as $\mathcal{K}_\pi = \mathcal{L}_0^{-1}$ yields the same velocity $-\mathcal{D}\mathcal{L}_0^{-1}u_\rho$, which is equivalent to eq. (3) via eq. (2). Thus, an inverse-generator correction introduced through the LAWGD kernel construction emerges here directly from inverse Wasserstein curvature at equilibrium. This identifies the ideal LAWGD dynamics as a Newton-type preconditioning of the chi-squared Wasserstein gradient flow.

Remark 1 (Equilibrium curvature is shared by f-divergences)**.** *The equilibrium Hessian in Theorem 1 is not specific to the chi-squared objective. For a sufficiently smooth f-divergence $\mathcal{D}_f(\rho\|\pi) := \int f(\rho/\pi)\,d\pi$, the expansion $u_{\rho_\epsilon} = \epsilon\mathcal{L}_0\phi + O(\epsilon^2)$ gives the same second variation $f''(1)\|\mathcal{L}_0\phi\|^2_\pi$ at π up to a multiplicative factor. In particular, for KL this agrees with the classical formula $\frac{d^2}{d\epsilon^2}|_{\epsilon=0}D_{\mathrm{KL}}(\rho_\epsilon\|\pi) = \int(\|\nabla^2\phi\|^2_F + \nabla\phi^\top\nabla^2 V\,\nabla\phi)\,d\pi$, which equals $\|\mathcal{L}_0\phi\|^2_\pi$ by the integrated Bochner identity (Otto & Villani, 2000; Bakry et al., 2014). What distinguishes the chi-squared objective here is its gradient: $\mathcal{D}u_\rho$ is linear in the density-ratio discrepancy. For KL, the analogous frozen velocity $-\mathcal{L}_1^{-1}\mathcal{D}\log(\rho/\pi)$ is nonlinear in this discrepancy and does not reduce to LAWGD.*

For the chi-squared flow, the frozen construction singles out a fixed scalar-to-vector operator,

$$\mathcal{T} := \mathcal{D}\mathcal{L}_0^{-1} = \mathcal{L}_1^{-1}\mathcal{D}, \tag{4}$$

which is a bounded map from $L^2_0(\pi)$ to $T_\pi\mathcal{P}_2$. We can then write the frozen-Newton velocity succinctly as $\mathbf{v}_{\mathrm{FN}}(\rho) = -\mathcal{T}u_\rho$. This formulation suggests a different approximation target. Note that the preconditioned dynamics depend on $\mathcal{L}_0^{-1}$ and $\mathcal{L}_1^{-1}$ only through the resulting scalar-to-vector action $\mathcal{T}$. It is therefore natural to approximate $\mathcal{T}$ directly, rather than first approximating the scalar inverse operator $\mathcal{L}_0^{-1}$ and then differentiating its output, or approximating the Hessian operator $\mathcal{L}_1$ and subsequently inverting it.

3.2 Spectral Structure of the Transport Operator

The transport operator admits a particularly simple spectral structure. Assume additionally that $\mathcal{L}_0^{-1}$ is compact on $L^2_0(\pi)$, and let $\mathcal{H}_{\mathrm{ex}} := T_\pi\mathcal{P}_2 = \overline{\mathrm{Range}(\mathcal{D})} \subset L^2(\pi;\mathbb{R}^d)$. Let (λ_k, ϕ_k) denote the nonconstant eigenpairs of $\mathcal{L}_0$, with $0 < \lambda_1 \le \lambda_2 \le \cdots$, $\langle\phi_i,\phi_j\rangle_\pi = \delta_{ij}$, where $\{\phi_k\}_{k\ge1}$ is a complete orthonormal basis of $L^2_0(\pi)$. Since $\mathcal{T}\phi_k = \lambda_k^{-1}\mathcal{D}\phi_k$, and $\langle\mathcal{D}\phi_i,\mathcal{D}\phi_j\rangle_\pi = \lambda_j\delta_{ij}$, the normalized fields $\mathbf{e}_k := \frac{D\phi_k}{\sqrt{\lambda_k}}$ are orthonormal in $\mathcal{H}_{\mathrm{ex}}$. The transport operator therefore admits the singular value decomposition, having $1/\sqrt{\lambda_k}$ being the k-th singular value, $\mathcal{T} = \sum_{k\ge1}\lambda_k^{-1/2}\mathbf{e}_k \otimes \phi_k$. Here $\mathbf{b}\otimes g$ denotes the rank-one operator $f \mapsto \mathbf{b}\langle g, f\rangle_\pi$. Thus, the slow Langevin modes with small λ_k are precisely the dominant singular components of the transport operator. More formally:

Proposition 2 (Singular structure of the transport operator)**.** *The transport operator satisfies $\mathcal{T}^*\mathcal{T} = \mathcal{L}_0^{-1}$ on $L^2_0(\pi)$ and $(\mathcal{T}\mathcal{T}^*)|_{\mathcal{H}_{\mathrm{ex}}} = (\mathcal{L}_1|_{\mathcal{H}_{\mathrm{ex}}})^{-1}$. If the codomain of $\mathcal{T}$ is taken to be all of $L^2(\pi;\mathbb{R}^d)$, then $\mathcal{T}\mathcal{T}^*$ vanishes on $\mathcal{H}_{\mathrm{ex}}^\perp$. Consequently, its singular values are $\sigma_k(\mathcal{T}) = \lambda_k^{-1/2}$, with right singular functions ϕ_k and left singular fields $\mathbf{e}_k$.*

3.3 Dynamics with Low-Rank Truncated Transport Operator

This structure motivates retaining only the leading r singular components of the transport operator. That is, rather than assuming access to the exact spectral components, we consider factorized approximations of the truncation $\widehat{\mathcal{T}}_r = \sum_{k=1}^{r} \mathbf{b}_k \otimes g_k$, where $g_k \in L_0^2(\pi)$ and $\mathbf{b}_k \in L^2(\pi; \mathbb{R}^d)$.

Writing $\mathbf{g}(\mathbf{x}) := (g_1(\mathbf{x}), \ldots, g_r(\mathbf{x}))^\top$, letting $\mathbf{B}(\mathbf{x}) \in \mathbb{R}^{d\times r}$ have columns $\mathbf{b}_1(\mathbf{x}), \ldots, \mathbf{b}_r(\mathbf{x})$, and defining $\mathbf{c} := \mathbb{E}_\pi[\mathbf{G}]$, the approximate transport operator generates the population velocity

$$\widehat{\mathbf{v}}_\rho = -\widehat{\mathcal{T}}_r u_\rho = -\sum_{k=1}^{r} \mathbf{b}_k \langle g_k, u_\rho\rangle_\pi = -\mathbf{B}(\mathbf{x})\left(\mathbb{E}_\rho[\boldsymbol{g}] - \mathbb{E}_\pi[\boldsymbol{g}]\right), \tag{5}$$

where the last equality invokes the exact identity $\langle g_k, u_\rho\rangle_\pi = \mathbb{E}_\rho[g_k] - \mathbb{E}_\pi[g_k]$. This identity allows finite-sample evaluation without pointwise estimation of the density ratio.

3.3.1 Convergence of the Truncated Population Dynamics

We can characterize when such an approximation preserves the contraction behavior of the ideal truncated dynamics. To describe convergence, assume that the factors are sufficiently regular that $\mathcal{S} := \mathcal{D}_\pi^* \widehat{\mathcal{T}}_r$ is a bounded operator on $L_0^2(\pi)$, and write $\mathrm{sym}(\mathcal{S}) := (\mathcal{S} + \mathcal{S}^*)/2$. For example, this holds when every $\mathbf{b}_k$ belongs to the domain of $\mathcal{D}_\pi^*$ with $\mathcal{D}_\pi^* \mathbf{b}_k \in L_0^2(\pi)$. Let Π_r be the orthogonal projector onto $\mathrm{span}\{\phi_1, \ldots, \phi_r\}$. For the exact truncation, $\mathcal{D}_\pi^* \mathcal{T}_r = \Pi_r$.

Along a sufficiently regular population solution of the continuity equation with velocity $-\widehat{\mathcal{T}}_r u_t$, $\frac{d}{dt} D_{\mathrm{KL}}(\rho_t\|\pi) = -\langle \mathcal{D} u_t, \widehat{\mathcal{T}}_r u_t\rangle_\pi = -\langle u_t, \mathrm{sym}(\mathcal{S}) u_t\rangle_\pi$. In particular, the full ideal operator gives $d\, D_{\mathrm{KL}}/dt = -\|u_t\|_\pi^2$, whereas the exact rank-r truncation gives $d\, D_{\mathrm{KL}}/dt = -\|\Pi_r u_t\|_\pi^2$.

Theorem 3 (Approximate KL contraction)**.** *Consider a sufficiently regular population flow generated by $\widehat{\mathcal{T}}_r$, and write $u_t := \rho_t/\pi - 1 \in L_0^2(\pi)$. Suppose that $\|\mathrm{sym}(\mathcal{S}) - \Pi_r\|_{\mathrm{op}} \le \epsilon < 1$, where the operator norm is taken on $L_0^2(\pi)$, and that, throughout the time interval under consideration, $\|(I - \Pi_r) u_t\|_\pi \le \eta_r$. Then, $D_{\mathrm{KL}}(\rho_t\|\pi) \le e^{-(1-\epsilon)t} D_{\mathrm{KL}}(\rho_0\|\pi) + \frac{\eta_r^2}{1-\epsilon}\left(1 - e^{-(1-\epsilon)t}\right)$.*

The bound separates dissipation error from unresolved density discrepancy. The quantity ϵ controls the deviation of the symmetric dissipation operator from Π_r on the full mean-zero space, including coupling to unresolved directions. The quantity η_r bounds the discrepancy outside the retained subspace along the flow. Thus, convergence is not a consequence of finite rank by itself: it requires both quantities to remain controlled. In particular, a small approximation objective $\mathcal{J}_r$ does not by itself imply the dissipation bound required by the theorem.

3.3.2 Particle Algorithm

In Section 4, we will propose kernel- and neural- methods that produce a low approximation $\widehat{\mathcal{T}}_r$ of this form with finite samples. Once either method provides factors $\{(g_k, \mathbf{b}_k)\}_{k=1}^r$, the resulting particle sampler is method-agnostic and simple to implement as we describe here. Given target samples $\{\mathbf{Y}_j\}_{j=1}^N$ and current particles $\{\mathbf{X}_n^{(i)}\}_{i=1}^M$, estimate the target feature mean and current feature mismatch by $\widehat{\mathbf{m}}_n := \frac{1}{M}\sum_{i=1}^M \boldsymbol{g}(\mathbf{X}_n^{(i)}) - \widehat{\boldsymbol{\mu}}_g$, where the target feature mean $\widehat{\boldsymbol{\mu}}_g := \frac{1}{N}\sum_{j=1}^N \boldsymbol{g}(\mathbf{Y}_j)$ can be precomputed given the learned factors. An explicit Euler step with eq. (5) therefore gives

$$\mathbf{X}_{n+1}^{(i)} = \mathbf{X}_n^{(i)} - \eta_n \mathbf{B}(\mathbf{X}_n^{(i)}) \widehat{\mathbf{m}}_n, \qquad i = 1, \ldots, M.$$

Thus, sampling requires factor evaluations and an r-dimensional feature average, but no density-ratio estimate. Any auxiliary objects used to construct the operator, including the neural Poisson-response networks to be introduced below, are not needed at sampling time.

4 Implementing with Finite Samples

We now describe how to learn the factors from target samples using a kernel-based method (Section 4.1) and a neural network (Section 4.2).

4.1 KERNEL-BASED IMPLEMENTATION

Proposition 2 suggests a natural kernel construction for approximating $\mathcal{T}$. Its right singular functions are the scalar Langevin eigenfunctions ϕ_k, while the corresponding left singular fields are proportional to their gradients, $\mathbf{e}_k = \mathcal{D}\phi_k/\sqrt{\lambda_k}$. We therefore couple the scalar and vector approximation spaces through differentiation: starting from a differentiable scalar kernel K, each scalar feature $K(\cdot, \mathbf{c})$ is paired with the vector feature $\mathcal{D}_1 K(\cdot, \mathbf{c})$, where $\mathcal{D}_1$ denotes differentiation in the first argument.

Let $\mathbf{c}_1, \ldots, \mathbf{c}_F$ be feature centers and define $\psi_j(\mathbf{x}) := K(\mathbf{x}, \mathbf{c}_j)$. Collect the scalar features into $\boldsymbol{\psi}(\mathbf{x}) := (\psi_1(\mathbf{x}), \ldots, \psi_F(\mathbf{x}))^\top$, center them under the target as $\bar{\boldsymbol{\psi}}(\mathbf{x}) := \boldsymbol{\psi}(\mathbf{x}) - \mathbb{E}_\pi[\boldsymbol{\psi}]$, and collect the corresponding gradient features into $\mathbf{J}(\mathbf{x}) := \mathcal{D}\boldsymbol{\psi}(\mathbf{x}) \in \mathbb{R}^{d\times F}$. A scalar mode and its associated transport direction are then represented using the same coefficient vector, $g(\mathbf{x}) = \bar{\boldsymbol{\psi}}(\mathbf{x})^\top \mathbf{q}$ and $\mathbf{b}(\mathbf{x}) = \mathbf{J}(\mathbf{x})\mathbf{q} = \mathcal{D}g(\mathbf{x})$. Thus, the finite-dimensional construction preserves the scalar-to-gradient relation of the exact singular system.

Projecting the scalar Langevin eigenproblem onto the span of $\bar{\boldsymbol{\psi}}$ gives the feature covariance matrix $\mathbf{G} := \mathbb{E}_\pi[\bar{\boldsymbol{\psi}}\bar{\boldsymbol{\psi}}^\top]$ and Dirichlet matrix $\mathbf{S} := \mathbb{E}_\pi[\mathbf{J}^\top\mathbf{J}]$. The resulting generalized eigenproblem is $\mathbf{G}\mathbf{q}_k = \sigma_k^2\mathbf{S}\mathbf{q}_k$, with normalization $\mathbf{q}_i^\top\mathbf{S}\mathbf{q}_j = \delta_{ij}$, where σ_k^2 approximates the inverse Langevin eigenvalue λ_k^{-1}. Under this normalization, $\mathbf{b}_k(\mathbf{x}) := \mathbf{J}(\mathbf{x})\mathbf{q}_k$ has unit $L^2(\pi)$ norm, while $g_k(\mathbf{x}) := \bar{\boldsymbol{\psi}}(\mathbf{x})^\top\mathbf{q}_k$ has norm σ_k. Hence g_k/σ_k and $\mathbf{b}_k$ approximate the right and left singular functions of $\mathcal{T}$, respectively, and the corresponding rank-one term is simply $\mathbf{b}_k \otimes g_k$.

Let $\mathbf{Q}_r = [\mathbf{q}_1, \ldots, \mathbf{q}_r]$ collect the r generalized eigenvectors with largest σ_k. The resulting rank-r transport approximation is $\widehat{\mathcal{T}}_r(\mathbf{x}, \mathbf{y}) = \mathbf{J}(\mathbf{x})\mathbf{Q}_r\mathbf{Q}_r^\top\bar{\boldsymbol{\psi}}(\mathbf{y})$. The kernel method therefore approximates $\mathcal{T}$ in a paired scalar/vector feature space generated by K and $\mathcal{D}_1 K$, with the pairing chosen specifically to reproduce the spectral structure of the transport operator.

With finite samples, the expectations defining $\mathbf{G}$ and $\mathbf{S}$ are replaced by empirical averages over target samples; the resulting procedure is summarized in Algorithm 1 in Appendix D.

4.2 NEURAL IMPLEMENTATION

The kernel construction mirrors the exact singular structure of $\mathcal{T}$ by pairing scalar kernel features with their gradients. Its limitation instead lies in the fixed scalar feature space: the relevant eigenfunctions must be represented within the span of a prescribed kernel dictionary. This can become costly in high dimensions; in Appendix B, we demonstrate this phenomenon by showing that even when the slow eigenspace lies in a fixed low-dimensional subspace, Gaussian kernel features can require a number of features that grows with the ambient dimension. This echoes known separations between fixed kernel representations and adaptive feature learning in high-dimensional problems with low-dimensional structure (Ghorbani et al., 2019; 2020; Mei et al., 2022).

Motivated by this, we replace the fixed kernel dictionary with an adaptive neural representation and parameterize a rank-r transport operator using scalar networks $g_k : \mathbb{R}^d \to \mathbb{R}$ and vector networks $\mathbf{b}_k : \mathbb{R}^d \to \mathbb{R}^d$, without imposing $\mathbf{b}_k = \mathcal{D}g_k$. Although the exact singular pairs satisfy this gradient relation, a low-rank operator admits nonunique factorizations, so recovering $\mathcal{T}_r$ does not require enforcing it factorwise. As shown below, the unrestricted population objective still recovers the same rank-r transport truncation.

Unlike the kernel realization, this adaptive parameterization no longer reduces to a finite-dimensional generalized eigendecomposition. We instead learn the dominant transport operator through a low-rank variational characterization following (Ryu et al., 2024; Jeong et al., 2025). For a rank-ℓ candidate $\widehat{\mathcal{T}}_\ell = \sum_{k=1}^{\ell} \mathbf{b}_k \otimes g_k$, we use the low-rank approximation objective,

$$\mathcal{J}_\ell(\widehat{\mathcal{T}}_\ell) := \|\widehat{\mathcal{T}}_\ell\|_{\mathrm{HS}}^2 - 2\,\mathrm{Tr}(\widehat{\mathcal{T}}_\ell^*\mathcal{T}) = \sum_{i,j\le\ell} \langle \mathbf{b}_i, \mathbf{b}_j\rangle_\pi \langle g_i, g_j\rangle_\pi - 2\sum_{i\le\ell} \langle \mathbf{b}_i, \mathcal{T} g_i\rangle_\pi.$$

Theorem 4 (Recovery of the transport truncation)**.** *Assume $\mathcal{T}$ is compact. Over operators of rank at most ℓ, the minimum of $\mathcal{J}_\ell$ is $-\sum_{k=1}^{\ell}\lambda_k^{-1}$, and a global optimizer is $\mathcal{T}_\ell$, the rank-ℓ truncation of $\mathcal{T}$. If $\lambda_\ell < \lambda_{\ell+1}$, the optimal rank-$\ell$ operator is unique.*

Thus, although the neural factors are not constrained individually to follow the singular structure of Proposition 2, the population objective still selects the same rank-r transport operator. We optimize $\mathcal{J}_r$ jointly over the neural factor pairs $\{(g_k, \mathbf{b}_k)\}_{k=1}^r$.

The remaining difficulty is that $\mathcal{J}_r$ contains the cross terms $\langle \mathbf{b}_k, \mathcal{T} g_k \rangle_\pi$, while $\mathcal{T} = \mathcal{D}\mathcal{L}_0^{-1}$ is precisely the operator we seek to approximate. We obtain the required forward and adjoint actions through *weak Poisson* problems. For a scalar factor g and a vector factor $\mathbf{b}$, define

$$h_g^* := \arg\min_h \left\{ \frac{1}{2}\|\mathcal{D}h\|_\pi^2 - \langle g, h \rangle_\pi \right\} \quad \text{and} \quad a_{\mathbf{b}}^* := \arg\min_a \left\{ \frac{1}{2}\|\mathcal{D}a\|_\pi^2 - \langle \mathbf{b}, \mathcal{D}a \rangle_\pi \right\}.$$

The following proposition shows that these variational problems recover the forward action $\mathcal{T}g$, the adjoint action $\mathcal{T}^*\mathbf{b}$, and equivalent sample-based representations of the cross term.

Proposition 5 (Weak Poisson responses)**.** *For $g \in L_0^2(\pi)$ and $\mathbf{b} \in L^2(\pi; \mathbb{R}^d)$, let h_g^* and $a_{\mathbf{b}}^*$ solve the variational problems above. Then $\mathcal{D}h_g^* = \mathcal{T}g$, $a_{\mathbf{b}}^* = \mathcal{T}^*\mathbf{b}$. Moreover, the population cross term admits the equivalent representations $\langle \mathbf{b}, \mathcal{T}g \rangle_\pi = \langle \mathbf{b}, \mathcal{D}h_g^* \rangle_\pi = \langle g, a_{\mathbf{b}}^* \rangle_\pi = \langle \mathcal{D}a_{\mathbf{b}}^*, \mathcal{D}h_g^* \rangle_\pi$.*

We parameterize the forward and adjoint response functions with auxiliary scalar networks and alternate between approximately solving these response problems and updating the deployed factors $(g_k, \mathbf{b}_k)$. This yields a sample-only optimization procedure involving expectations under π and derivatives of scalar functions. In particular, training does not require the target score, pointwise application of $\mathcal{D}_\pi^*$ to the learned vector fields, or direct inversion of $\mathcal{L}_0$.

Because the response problems are themselves solved approximately, the factor updates use a stopped-response construction rather than differentiating through the response optimization. Appendix C gives the precise surrogate, its value and gradient consistency, the alternating optimization procedure, and the centering and factor-balancing steps used to remove additive and reciprocal-scaling ambiguities. Algorithm 2 in Appendix D summarizes the resulting procedure: the auxiliary response networks are updated through their weak Poisson objectives, followed by a joint update of the transport factors $(g_k, \mathbf{b}_k)$ using the low-rank objective.

5 Related Work and Discussion

Our work connects to Wasserstein optimization, particle-based variational inference, spectral approximation of Langevin and transfer operators, and neural spectral and operator approximation. We summarize the closest connections here.

Wasserstein geometry, preconditioning, and particle flows. Wasserstein gradient flows originate from the Jordan–Kinderlehrer–Otto formulation and the geometric theory of optimal transport (Jordan et al., 1998; Otto, 2001; Ambrosio et al., 2008; Villani, 2009). Newton and preconditioning ideas from Euclidean optimization (Nocedal & Wright, 2006) have been extended to probability space through Wasserstein Newton methods, preconditioned Wasserstein descent, and second-order particle methods (Wang & Li, 2020; Detommaso et al., 2018; Bonet et al., 2024; Dong et al., 2023). SVGD transports interacting particles by following the gradient flow of the KL divergence under a kernel-induced Stein geometry (Liu & Wang, 2016; Liu, 2017). Chewi et al. (2020) developed an alternative viewpoint in which SVGD is interpreted as a kernelized Wasserstein gradient flow of the chi-squared divergence, and proposed LAWGD by using the inverse Langevin operator to precondition this flow. Related kernelized Wasserstein flows include MMD gradient flow (Arbel et al., 2019) and (de)-regularized MMD flow (Chen et al., 2025), which connects MMD and chi-squared geometries while retaining a sample-based implementation.

Kernel and spectral approximations. Recent particle samplers based on diffusion-map approximations of the Langevin generator (Li & Marzouk, 2025) and Koopman spectral approximations (Xu et al., 2025) build upon nonparametric approximations of differential generators and their eigenspaces using diffusion maps and related graph constructions (Coifman & Lafon, 2006; Singer, 2006; Berry & Harlim, 2016). More broadly, Galerkin, Nyström, and kernel methods are widely used to approximate integral, transfer, and Koopman operators from data (Williams et al., 2015; Klus et al., 2018; Mollenhauer et al., 2020). Their behavior in high dimension has motivated comparisons with learned representations, with several works identifying regimes in which fixed kernel methods incur dimension-dependent costs while neural networks exploit low-dimensional structure (Ghorbani et al., 2019; 2020; Mei et al., 2022).

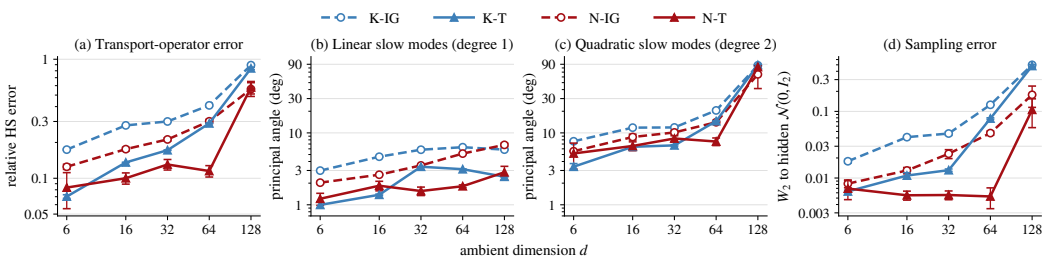


Figure 1: Spiked-OU dimension sweep with fixed intrinsic slow structure and increasing ambient dimension. We compare kernel and neural inverse-generator methods (K-IG, N-IG) with kernel and neural direct-transport methods (K-T, N-T). Panels report (a) relative transport-operator error, (b) principal-angle error for the linear slow modes, (c) principal-angle error for the quadratic slow modes, and (d) sampling error measured by W_2 on the hidden two-dimensional coordinates.

Neural spectral decomposition. Neural networks have been used to approximate eigenfunctions and singular functions of high-dimensional operators without relying on a prespecified basis; see, e.g., (Pfau et al., 2019; Deng et al., 2022; Ryu et al., 2024; 2025; Jeong et al., 2025). Recently, KSWGD (Xu et al., 2025) considered neural-network dictionaries for data-driven spectral approximation of the Langevin generator, alongside kernel and other finite-dimensional dictionaries. In this approach, neural networks serve as a flexible dictionary within a generator eigendecomposition pipeline. In contrast, our approach uses neural spectral learning to directly approximate the low-rank transport operator underlying the preconditioned Wasserstein flow, and systematically compares this construction with its kernel-based counterparts.

6 Experiments

We evaluate two aspects of the proposed approach: whether directly learning the transport operator is advantageous over estimating the inverse generator first, and also whether adaptive neural representations overcome the ambient-dimensional limitations of fixed kernels. We compare kernel and neural realizations of the inverse-generator approach (*K-IG*, *N-IG*) and of the direct transport formulation (*K-T*, *N-T*). Here, *K-IG* corresponds to the kernel-based KSWGD construction of Xu et al. (2025). We use the standard Gaussian kernel $K_\epsilon(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/(2\epsilon))$ for all kernel-based methods. The experimental details are deferred to Appendix E.

Experiment 1. Spiked Ornstein–Uhlenbeck Dynamics. To demonstrate the scaling behavior of the kernel and neural methods in terms of sampling performance as well as spectral estimation quality, we first use a spiked covariance Gaussian model, which embeds a fixed two-dimensional slow subspace in an ambient space of dimension d; see Appendix B. It allows us to increase ambient dimension without changing the intrinsic slow structure.

The results demonstrates the limitation of the kernel methods in high-dimensions. K-T is slightly more accurate at $d = 6$, but N-T becomes more accurate from $d = 16$ onward, with transport error 0.115 versus 0.289 for K-T at $d = 64$. The same separation appears in the particle dynamics, where hidden-coordinate Wasserstein error at $d = 64$ is 0.005 for N-T versus 0.078 for K-T. The degradation is strongest in the higher-order slow modes, consistent with the quadratic representation barrier of Theorem 6 in Appendix B. Thus, when the intrinsic structure is fixed but ambient dimension grows, the adaptive neural representation increasingly outperforms the fixed kernel representation.

Experiment 2. Stochastic Allen–Cahn equation. We then consider the stochastic Allen–Cahn benchmark of Xu et al. (2025), discretized on a 128×128 grid and compressed with an autoencoder before spectral learning. This experiment probes a different failure mode: estimating generator eigenvalues from very short-lag dynamics, where Koopman eigenvalues lie close to one.

K-IG becomes unstable because small errors in near-unit Koopman eigenvalues are amplified when converted into inverse generator eigenvalues. The direct transport methods avoid this intermediate spectral inversion and remain stable. N-T attains the best distributional metrics among the stable methods and fits roughly $6.2\times$ faster than N-IG. The absolute sampling improvement is modest

Table 1: Allen–Cahn sampling performance for seed 0. The Gaussian initialization serves as a pre-update upper reference, while the finite-sample floor serves as a lower empirical reference obtained from the latent W_2 between two independent target samples. Lower is better.

Method	Latent W_2	MMD^2	Energy	Fit (s)	Sample (s)
K-IG	diverged	0.6070	3.1×10^9	0.6	1.0
N-IG	1.916	0.0099	0.0578	187.5	4.1
K-T	1.911	0.0110	0.0642	**0.5**	0.5
N-T	**1.904**	**0.0091**	**0.0533**	30.2	**0.2**
Gaussian init.	1.925	0.0120	0.0709	–	–
Finite-sample floor	1.789	–	–	–	–

Table 2: Held-out VAMP-2 scores on pentapeptide molecular dynamics. The metric evaluates how well the learned rank-5 representation captures slow dynamical structure on unseen trajectory data; higher is better.

Method	VAMP-2
K-IG	1.495
N-IG	1.600
K-T	**1.730**
N-T	1.567

because the Gaussian initialization is already close to the target; the main takeaway is therefore improved spectral stability rather than a large reduction in sampling error.

Experiment 3. Pentapeptide molecular dynamics. Finally, we test whether the learned representations capture slow molecular dynamics on the pentapeptide benchmark distributed with PyEMMA (Scherer et al., 2015), consisting of 25 independent $500\,\mathrm{ns}$ molecular-dynamics trajectories. We use a 24-dimensional torsional representation and evaluate the learned rank-5 scalar feature space using held-out VAMP-2, which measures the time-predictive content retained by the representation.

All four methods recover nontrivial slow dynamical structure. K-T attains the highest held-out VAMP-2 (Wu & Noé, 2020) score, while N-T remains competitive with the inverse-generator methods. Unlike the controlled spiked experiment, there is no clear kernel–neural separation at this moderate dimension. Rather, this experiment shows that the transport formulation remains effective on realistic molecular dynamics, while the dimensional advantage of adaptive neural representations is more clearly isolated in the controlled ambient-dimension experiment.

7 Concluding Remarks

Viewed through Wasserstein geometry, the spectral correction underlying LAWGD emerges naturally as tangent-space preconditioning rather than as an ad hoc modification of the particle flow. For the chi-squared objective, the frozen-Newton step reduces to a single scalar-to-vector transport operator, separating the geometric question of what should be approximated from the representational question of how to approximate it. More broadly, the chi-squared divergence is particularly convenient for designing preconditioned Wasserstein dynamics: in terms of the density-ratio discrepancy $u_\rho = \rho/\pi - 1$, the objective is quadratic, $\mathcal{F}(\rho) = \frac{1}{2}\|u_\rho\|_\pi^2$, and its Wasserstein gradient is linear in u_ρ. As in Euclidean optimization of a quadratic objective, freezing the Hessian at the target therefore yields linear preconditioned dynamics. Although KL shares the same equilibrium curvature (Remark 1), its nonlinear gradient does not admit the same reduction. KL nevertheless remains natural in settings such as Langevin and score-based sampling, where only an unnormalized density may be available. Combining this advantage with the preconditioning structure exposed here is an interesting direction.

Several limitations remain. Our learning procedures assume access to samples from π and are therefore most directly suited to generative modeling rather than sampling from an unnormalized density. The theoretical guarantees are population, continuous-time results and do not account for finite-sample, neural approximation, or time-discretization errors. Moreover, the low-rank convergence guarantee contains an error floor controlled by unresolved modes, whose behavior depends on the quality of the learned transport approximation. A broader empirical comparison with external samplers and generative models also remains for future work.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing text, organizing parts of the paper, and supporting code development and debugging. AI tools were also used to help check intermediate mathematical derivations and suggest alternative formulations. All mathematical claims, proofs, experiments, and code used in the paper were independently reviewed and verified by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We implemented all methods in PyTorch and ran the experiments on a machine equipped with four NVIDIA H100 NVL GPUs. The appendices provide complete algorithmic descriptions, proofs, hyperparameters, data-generation procedures, and evaluation protocols. All reported multi-run experiments use explicitly specified random seeds. We will open-source the codebase and configuration files needed to reproduce the experiments upon acceptance.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2008.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2014.

Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 40(1):68–96, 2016. doi: 10.1016/j.acha.2015.01.001.

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, 2024.

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*, 26:1–77, 2025.

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, 2020.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. doi: 10.1016/j.acha.2006.04.006.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 678–689. PMLR, 2017.

Zhijie Deng, Jiaxin Shi, and Jun Zhu. NeuralEF: Deconstructing kernels by deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4976–4992, 2022.

Gianluca Detommaso, Tiangang Cui, Alessio Spantini, Youssef Marzouk, and Robert Scheichl. A stein variational newton method. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Bernard Helffer and Francis Nier. *Hypoelliptic Estimates and Spectral Theory for Fokker–Planck Operators and Witten Laplacians*, volume 1862 of *Lecture Notes in Mathematics*. Springer, 2005. doi: 10.1007/b104762.

Minchan Jeong, J. Jon Ryu, Se-Young Yun, and Gregory W. Wornell. Efficient parametric SVD of koopman operator for stochastic dynamical systems. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359.

Stefan Klus, Feliks Nüske, Péter Koltai, Hao Wu, Ioannis G. Kevrekidis, Christof Schütte, and Frank Noé. Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28:985–1010, 2018. doi: 10.1007/s00332-017-9437-7.

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.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. doi: 10.1016/j.acha.2021.12.003.

Mattes Mollenhauer, Stefan Klus, Christof Schütte, and Péter Koltai. Kernel autocovariance operators of stationary processes: Estimation and convergence. *arXiv preprint arXiv:2004.00891*, 2020.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1–2):101–174, 2001.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, 2014.

David Pfau, Stig Petersen, Ashish Agarwal, David Barrett, and Kimberly Stachenfeld. Spectral inference networks: Unifying deep and spectral learning. In *International Conference on Learning Representations*, 2019.

Jongha Jon Ryu, Xiangxiang Xu, Hasan Sabri Melihcan Erol, Yuheng Bu, Lizhong Zheng, and Gregory W. Wornell. Operator SVD with neural networks via nested low-rank approximation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42870–42905. PMLR, 2024.

Jongha Jon Ryu, Samuel Zhou, and Gregory Wornell. Revisiting orbital minimization method for neural operator decomposition. In *Advances in Neural Information Processing Systems*, volume 38, pp. 114671–114702, 2025.

Filippo Santambrogio. Euclidean, metric, and wasserstein gradient flows: An overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017. doi: 10.1007/s13373-017-0101-1.

Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. PyEMMA 2: A software package for estimation, validation, and analysis of markov models. *Journal of Chemical Theory and Computation*, 11(11):5525–5542, 2015. doi: 10.1021/acs.jctc.5b00743.

Amit Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006. doi: 10.1016/j.acha.2006.03.004.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Yifei Wang and Wuchen Li. Information newton's flow: Second-order optimization method in probability space. *arXiv preprint arXiv:2001.04341*, 2020.

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

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015. doi: 10.1007/s00332-015-9258-5.

Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *Journal of Nonlinear Science*, 30(1):23–66, 2020. doi: 10.1007/s00332-019-09567-y.

Yuanchao Xu, Fengyi Li, Masahiro Fujisawa, Xiaoyuan Cheng, Youssef Marzouk, and Isao Ishikawa. Generative modeling through koopman spectral analysis: An operator-theoretic perspective. *arXiv preprint arXiv:2512.18837*, 2025.

APPENDIX

A Proofs of Main Results **13**
A.1 Proof of Theorem 1 . . . 13
A.2 Proof of Proposition 2 . . . 13
A.3 Proof of Theorem 3 . . . 14
A.4 Proof of Theorem 4 . . . 14
A.5 Proof of Proposition 5 . . . 15

B Additional Theory for the Ambient-Dimensional Representation Barrier **15**
B.1 Spiked Ornstein–Uhlenbeck model and Gaussian feature spaces . . . 15
B.2 Hermite coefficients of Gaussian kernel sections . . . 16
B.3 Dirichlet projection onto the feature space . . . 16
B.4 Linear signal modes . . . 17
B.5 Quadratic signal modes . . . 18
B.6 Implication for the truncated inverse generator . . . 19

C Sample-Only Optimization Details **20**
C.1 Stopped-response surrogate . . . 20
C.2 Value and gradient consistency . . . 20
C.3 Alternating optimization . . . 20
C.4 Centering and factor balancing . . . 21

D Algorithms **21**

E Experimental details **21**
E.1 Spiked Ornstein–Uhlenbeck . . . 22
E.2 Stochastic Allen–Cahn . . . 22
E.3 Pentapeptide . . . 22

A PROOFS OF MAIN RESULTS

Throughout this section, we assume that π is smooth and strictly positive, and that the densities and test functions considered below have sufficient regularity and decay for the stated integrations by parts and differential identities to hold. We work on the mean-zero scalar subspace $L^2_0(\pi)$ and assume a Poincaré inequality whenever $\mathcal{L}_0^{-1}$ is invoked. Compactness or discreteness of the spectrum is assumed only in results that explicitly require it. Wasserstein Hessian statements are understood on the Wasserstein tangent space of gradient vector fields.

A.1 PROOF OF THEOREM 1

Proof. Let (ρ_ϵ) be the Wasserstein geodesic in the statement, with $\rho_0 = \pi$ and initial velocity $\mathcal{D}\phi$. Evaluating the continuity equation at $\epsilon = 0$ gives $\partial_\epsilon \rho_\epsilon|_{\epsilon=0} = -\nabla \cdot (\pi \mathcal{D}\phi)$. Since $\mathcal{D}_\pi^* \mathbf{v} = -\pi^{-1}\nabla \cdot (\pi \mathbf{v})$, it follows that $\partial_\epsilon(\rho_\epsilon/\pi - 1)|_{\epsilon=0} = \mathcal{D}_\pi^* \mathcal{D}\phi = \mathcal{L}_0\phi$, and hence $u_{\rho_\epsilon} = \epsilon \mathcal{L}_0\phi + O(\epsilon^2)$. Using $\mathcal{F}(\rho) = \frac{1}{2}\|u_\rho\|_\pi^2$, we obtain $\mathcal{F}(\rho_\epsilon) = \frac{\epsilon^2}{2}\|\mathcal{L}_0\phi\|_\pi^2 + O(\epsilon^3)$, so $\frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \mathcal{F}(\rho_\epsilon) = \|\mathcal{L}_0\phi\|_\pi^2$. Finally, using $\mathcal{L}_0 = \mathcal{D}_\pi^* \mathcal{D}$ and the intertwining identity $\mathcal{L}_1 \mathcal{D} = \mathcal{D}\mathcal{L}_0$,

$$\|\mathcal{L}_0\phi\|_\pi^2 = \langle \mathcal{D}\phi, \mathcal{D}\mathcal{L}_0\phi \rangle_\pi = \langle \mathcal{D}\phi, \mathcal{L}_1 \mathcal{D}\phi \rangle_\pi.$$

Thus the quadratic form of the Wasserstein Hessian at π agrees with that of $\mathcal{L}_1$ on tangent gradient fields, and polarization gives $\mathsf{H}_\pi = \mathrm{Hess}_{W_2} \mathcal{F}(\pi) = \mathcal{L}_1$ on $T_\pi \mathcal{P}_2$. □

A.2 PROOF OF PROPOSITION 2

Proof. Recall that $\mathcal{T} = \mathcal{D}\mathcal{L}_0^{-1}$ on $L^2_0(\pi)$. Since $\mathcal{D}^* = \mathcal{D}_\pi^*$ and $\mathcal{L}_0^{-1}$ is self-adjoint, $\mathcal{T}^* = \mathcal{L}_0^{-1}\mathcal{D}_\pi^*$, and therefore

$$\mathcal{T}^*\mathcal{T} = \mathcal{L}_0^{-1}\mathcal{D}_\pi^*\mathcal{D}\mathcal{L}_0^{-1} = \mathcal{L}_0^{-1}.$$

Hence the right singular functions of $\mathcal{T}$ are the eigenfunctions ϕ_k of $\mathcal{L}_0$, with singular values $\sigma_k(\mathcal{T}) = \lambda_k^{-1/2}$. Defining $\mathbf{e}_k := \mathcal{D}\phi_k/\sqrt{\lambda_k}$, the Dirichlet identity gives $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_\pi = \delta_{ij}$, while $\mathcal{T}\phi_k = \mathcal{D}\phi_k/\lambda_k = \lambda_k^{-1/2}\mathbf{e}_k$. Thus $\mathcal{T} = \sum_{k\geq 1} \lambda_k^{-1/2} \mathbf{e}_k \otimes \phi_k$.

The intertwining identity further gives $\mathcal{L}_1 \mathbf{e}_k = \lambda_k \mathbf{e}_k$. Since $\{\mathbf{e}_k\}$ spans $\mathcal{H}_{\mathrm{ex}} = \overline{\mathrm{Range}(\mathcal{D})}$, it follows that

$$\mathcal{T}\mathcal{T}^* = \sum_{k\geq 1} \lambda_k^{-1} \mathbf{e}_k \otimes \mathbf{e}_k = P_{\mathrm{ex}} (\mathcal{L}_1|_{\mathcal{H}_{\mathrm{ex}}})^{-1} P_{\mathrm{ex}}.$$

If the codomain is enlarged to all of $L^2(\pi; \mathbb{R}^d)$, $\mathcal{T}\mathcal{T}^*$ vanishes on $\mathcal{H}_{\mathrm{ex}}^\perp$, completing the proof. □

A.3 Proof of Theorem 3

Proof. Let $u_t = \rho_t/\pi - 1$ and $\mathbf{v}_t = -\widehat{\mathcal{T}}_r u_t$. Differentiating the KL divergence along the continuity equation gives $\frac{d}{dt}\mathrm{KL}(\rho_t\|\pi) = \int \rho_t \nabla \log(\rho_t/\pi)^\top \mathbf{v}_t \, d\mathbf{x}$. Since $\rho_t = \pi(1+u_t)$, we have $\rho_t \nabla \log(\rho_t/\pi) = \pi \mathcal{D} u_t$, and therefore

$$\frac{d}{dt}\mathrm{KL}(\rho_t\|\pi) = -\langle \mathcal{D}u_t, \widehat{\mathcal{T}}_r u_t \rangle_\pi = -\langle u_t, \mathrm{sym}(\mathcal{S}) u_t \rangle_\pi,$$

where $\mathcal{S} = \mathcal{D}_\pi^* \widehat{\mathcal{T}}_r$ and the skew-adjoint part does not contribute to the quadratic form.

Using $\|\mathrm{sym}(\mathcal{S}) - \Pi_r\|_{\mathrm{op}} \leq \epsilon$ gives $\frac{d}{dt}\mathrm{KL}(\rho_t\|\pi) \leq -\|\Pi_r u_t\|_\pi^2 + \epsilon \|u_t\|_\pi^2$. Since $u_t = \Pi_r u_t + (I - \Pi_r)u_t$ orthogonally, the right-hand side equals $-(1-\epsilon)\|u_t\|_\pi^2 + \|(I-\Pi_r)u_t\|_\pi^2$, and hence is at most $-(1-\epsilon)\|u_t\|_\pi^2 + \eta_r^2$. Finally, $\mathrm{KL}(\rho_t\|\pi) \leq \chi^2(\rho_t\|\pi) = \|u_t\|_\pi^2$, so

$$\frac{d}{dt}\mathrm{KL}(\rho_t\|\pi) \leq -(1-\epsilon)\mathrm{KL}(\rho_t\|\pi) + \eta_r^2.$$

Integrating this differential inequality yields $\mathrm{KL}(\rho_t\|\pi) \leq e^{-(1-\epsilon)t}\mathrm{KL}(\rho_0\|\pi) + \frac{\eta_r^2}{1-\epsilon}(1 - e^{-(1-\epsilon)t})$, as claimed. □

A.4 Proof of Theorem 4

Proof. By Proposition 2, the singular values of $\mathcal{T}$ are

$$\sigma_k = \lambda_k^{-1/2}, \qquad \sigma_1 \geq \sigma_2 \geq \cdots.$$

Let $\mathcal{S}$ be any operator of rank at most ℓ, with nonzero singular values $s_1, \ldots, s_\ell$, padding with zeros if necessary. The singular-value trace inequality gives

$$\mathrm{Tr}(\mathcal{S}^*\mathcal{T}) \leq \sum_{k=1}^{\ell} s_k \sigma_k.$$

Therefore

$$\mathcal{J}_\ell(\mathcal{S}) = \|\mathcal{S}\|_{\mathrm{HS}}^2 - 2\,\mathrm{Tr}(\mathcal{S}^*\mathcal{T}) \geq \sum_{k=1}^{\ell} s_k^2 - 2\sum_{k=1}^{\ell} s_k \sigma_k.$$

Completing the square yields

$$\mathcal{J}_\ell(\mathcal{S}) \geq \sum_{k=1}^{\ell} (s_k - \sigma_k)^2 - \sum_{k=1}^{\ell} \sigma_k^2 \geq -\sum_{k=1}^{\ell} \lambda_k^{-1}.$$

Equality is attained by taking the rank-ℓ truncated singular-value decomposition of $\mathcal{T}$, namely

$$\mathcal{T}_\ell = \sum_{k=1}^{\ell} \sigma_k \, \mathbf{e}_k \otimes \phi_k = \sum_{k=1}^{\ell} \frac{\mathcal{D}\phi_k}{\lambda_k} \otimes \phi_k.$$

Hence the minimum value is $-\sum_{k=1}^{\ell} \lambda_k^{-1}$.

If $\lambda_\ell < \lambda_{\ell+1}$, equivalently $\sigma_\ell > \sigma_{\ell+1}$, the leading ℓ-dimensional singular subspaces are uniquely determined. The best rank-ℓ operator is therefore unique and equals $\mathcal{T}_\ell$. □

A.5 Proof of Proposition 5

Proof. For $g \in L^2_0(\pi)$, let $\mathcal{E}_g(h) = \frac{1}{2}\|\mathcal{D}h\|^2_\pi - \langle g, h\rangle_\pi$. The first-order condition at h^*_g is $\langle \mathcal{D}h^*_g, \mathcal{D}\psi\rangle_\pi = \langle g, \psi\rangle_\pi$ for every admissible perturbation ψ, so $\mathcal{L}_0 h^*_g = g$ weakly and therefore $\mathcal{D}h^*_g = \mathcal{D}\mathcal{L}_0^{-1} g = \mathcal{T} g$. Similarly, for $\mathcal{E}_{\mathbf{b}}(a) = \frac{1}{2}\|\mathcal{D}a\|^2_\pi - \langle \mathbf{b}, \mathcal{D}a\rangle_\pi$, the first-order condition at $a^*_{\mathbf{b}}$ gives $\mathcal{L}_0 a^*_{\mathbf{b}} = \mathcal{D}^*_\pi \mathbf{b}$, and hence $a^*_{\mathbf{b}} = \mathcal{L}_0^{-1}\mathcal{D}^*_\pi \mathbf{b} = \mathcal{T}^* \mathbf{b}$.

The cross-term identities now follow directly from these two weak equations and adjointness:

$$\langle \mathbf{b}, \mathcal{T} g\rangle_\pi = \langle \mathbf{b}, \mathcal{D}h^*_g\rangle_\pi = \langle g, a^*_{\mathbf{b}}\rangle_\pi = \langle \mathcal{D}a^*_{\mathbf{b}}, \mathcal{D}h^*_g\rangle_\pi.$$

This proves the claim. □

B Additional Theory for the Ambient-Dimensional Representation Barrier

We provide the details underlying the ambient-dimensional kernel representation barrier used in Section 6. The argument proceeds by expanding Gaussian kernel sections in the Hermite eigenbasis of a spiked Ornstein–Uhlenbeck model and bounding how much Dirichlet energy a finite collection of such features can capture in a fixed low-dimensional signal mode.

B.1 Spiked Ornstein–Uhlenbeck model and Gaussian feature spaces

Let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2] \in \mathbb{R}^{d\times 2}$ satisfy $\mathbf{U}^\top \mathbf{U} = I_2$, fix $\tau > 1$, and consider the Gaussian target

$$\pi_d = \mathcal{N}(0, \mathbf{\Sigma}_d), \qquad \mathbf{\Sigma}_d = \tau^2 \mathbf{U}\mathbf{U}^\top + (I - \mathbf{U}\mathbf{U}^\top).$$

Extend $\mathbf{u}_1, \mathbf{u}_2$ to an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_d$ of $\mathbb{R}^d$, let $\sigma_1 = \sigma_2 = \tau$ and $\sigma_q = 1$ for $q \geq 3$, and define the standardized coordinates $z_q(\mathbf{x}) := \mathbf{u}_q^\top \mathbf{x}/\sigma_q$. Then $z_1, \dots, z_d$ are independent standard Gaussian random variables under π_d, and the positive Langevin operator is

$$\mathcal{L}_0 = \sum_{q=1}^{d} \frac{1}{\sigma_q^2}\left(-\partial^2_{z_q} + z_q \partial_{z_q}\right).$$

Let $h_n = \mathrm{He}_n/\sqrt{n!}$ denote the normalized probabilists' Hermite polynomials, and for a multi-index $\alpha \in \mathbb{N}_0^d$ write $h_\alpha(\mathbf{z}) := \prod_{q=1}^d h_{\alpha_q}(z_q)$. The family $\{h_\alpha\}$ is an orthonormal eigenbasis of $L^2(\pi_d)$ with

$$\mathcal{L}_0 h_\alpha = \lambda_\alpha h_\alpha, \qquad \lambda_\alpha = \sum_{q=1}^{d} \frac{\alpha_q}{\sigma_q^2}.$$

In particular, the linear signal modes z_1, z_2 have eigenvalue τ^{-2}, while the degree-two signal modes have eigenvalue $2\tau^{-2}$.

We measure approximation in the Dirichlet seminorm $a(f,g) := \int \langle \nabla f(\mathbf{x}), \nabla g(\mathbf{x})\rangle \, \pi_d(d\mathbf{x})$, with $\|f\|^2_a := a(f,f)$. In the Hermite basis, this form diagonalizes as

$$a(f,g) = \sum_\alpha \lambda_\alpha \widehat{f}_\alpha \widehat{g}_\alpha, \qquad \widehat{f}_\alpha := \langle f, h_\alpha\rangle_{L^2(\pi_d)}.$$

For bandwidth $\epsilon > 0$, define the Gaussian kernel $K_\epsilon(\mathbf{x}, \mathbf{z}) := \exp(-\|\mathbf{x} - \mathbf{z}\|^2/(2\epsilon))$, draw centers $\mathbf{Z}_1, \dots, \mathbf{Z}_N \overset{\text{iid}}{\sim} \pi_d$, and set $V_{N,\epsilon} := \mathrm{span}\{K_\epsilon(\cdot, \mathbf{Z}_i) : 1 \leq i \leq N\}$. For a nonconstant $\phi \in H^1(\pi_d)$, define the relative best-approximation error and captured Dirichlet-energy fraction by

$$R_{N,\epsilon}(\phi) := \frac{\mathrm{dist}_a(\phi, V_{N,\epsilon})^2}{\|\phi\|^2_a}, \qquad C_{N,\epsilon}(\phi) := 1 - R_{N,\epsilon}(\phi).$$

Thus, $C_{N,\epsilon}(\phi)$ measures the fraction of the Dirichlet energy of ϕ captured by the Gaussian feature space.

Theorem 6 (Ambient-dimensional representation barrier)**.** *Fix $\tau > 1$. For the spiked Ornstein–Uhlenbeck family above, as $d \to \infty$: (i) for the linear signal mode $\phi_{\mathrm{lin}} = z_1$, if $N = o(d)$, then $\sup_{\epsilon>0} C_{N,\epsilon}(\phi_{\mathrm{lin}}) \xrightarrow{\mathbb{P}} 0$; and (ii) for the quadratic signal mode $\phi_{\mathrm{quad}} = z_1 z_2$, if $N = o(d^2/\log^2 d)$, then $\sup_{\epsilon>0} C_{N,\epsilon}(\phi_{\mathrm{quad}}) \xrightarrow{\mathbb{P}} 0$. In particular, both conclusions remain valid when ϵ is chosen as a measurable function of the sampled centers.*

B.2 Hermite coefficients of Gaussian kernel sections

We first compute the Hermite coefficients of a Gaussian kernel section.

Lemma 7 (Hermite coefficients of a Gaussian section)**.** *Let γ be standard Gaussian measure on $\mathbb{R}^m$, let $a_1, \dots, a_m > 0$, and define*

$$k_{\mathbf{w}}(\mathbf{z}) := \exp\left(-\frac{1}{2}\sum_{q=1}^{m} a_q (z_q - w_q)^2\right), \qquad c_q := \frac{a_q}{1+a_q} \in (0,1).$$

Then, for every multi-index α,

$$\langle k_{\mathbf{w}}, h_\alpha \rangle_{L^2(\gamma)} = \left(\prod_{q=1}^{m} \frac{1}{\sqrt{1+a_q}} \exp\left(-\frac{a_q w_q^2}{2(1+a_q)}\right)\right) \prod_{q=1}^{m} \frac{c_q^{\alpha_q/2}\, \mathrm{He}_{\alpha_q}(\sqrt{c_q} w_q)}{\sqrt{\alpha_q!}}.$$

In particular, $\langle k_{\mathbf{w}}, z_q\rangle / \langle k_{\mathbf{w}}, 1\rangle = c_q w_q$, and for $p \neq q$, $\langle k_{\mathbf{w}}, z_p z_q\rangle / \langle k_{\mathbf{w}}, 1\rangle = c_p c_q w_p w_q$.

Proof. The integral factorizes over coordinates, so it suffices to consider the one-dimensional case. Dropping subscripts, the Hermite generating function gives

$$\sum_{n\geq 0} \frac{t^n}{n!} \mathrm{He}_n(z) = e^{tz - t^2/2}.$$

Therefore,

$$\sum_{n\geq 0} \frac{t^n}{n!} \langle k_w, \mathrm{He}_n\rangle = \int \exp\left(-\frac{a}{2}(z-w)^2 + tz - \frac{t^2}{2}\right) \gamma(dz).$$

Completing the square yields

$$\sum_{n\geq 0} \frac{t^n}{n!} \langle k_w, \mathrm{He}_n\rangle = \frac{1}{\sqrt{1+a}} \exp\left(-\frac{aw^2}{2(1+a)}\right) \exp\left(cwt - \frac{c}{2}t^2\right),$$

where $c = a/(1+a)$. Using $\exp(cwt - \frac{c}{2}t^2) = \sum_{n\geq 0} \frac{(\sqrt{c}\,t)^n}{n!} \mathrm{He}_n(\sqrt{c}\,w)$ and matching powers of t gives the claim. □

Write each center as $\mathbf{Z}_i = \sum_{q=1}^{d} \sigma_q W_{iq} \mathbf{u}_q$, where $W_{iq} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$. In standardized coordinates, $\|\mathbf{x} - \mathbf{Z}_i\|^2 = \sum_{q=1}^{d} \sigma_q^2 (z_q - W_{iq})^2$, so Lemma 7 applies with $c_q = \sigma_q^2/(\epsilon + \sigma_q^2)$. We write $c_s := \tau^2/(\epsilon+\tau^2)$ and $c_n := 1/(\epsilon+1)$ for the signal and noise coordinates, respectively, and define $\kappa := c_s/c_n = \tau^2(\epsilon+1)/(\epsilon+\tau^2)$. For every $\epsilon > 0$, $1 \leq \kappa < \tau^2$.

Normalize the kernel sections by their positive means,

$$\widetilde{K}_i := \frac{K_\epsilon(\cdot, \mathbf{Z}_i)}{\langle K_\epsilon(\cdot, \mathbf{Z}_i), 1\rangle_{L^2(\pi_d)}}.$$

Lemma 7 then gives $\langle \widetilde{K}_i, z_q\rangle_{L^2(\pi_d)} = c_q W_{iq}$ and, for $p \neq q$, $\langle \widetilde{K}_i, z_p z_q\rangle_{L^2(\pi_d)} = c_p c_q W_{ip} W_{iq}$.

B.3 Dirichlet projection onto the feature space

The next lemma expresses the best approximation error through the Dirichlet Gram matrix.

Lemma 8 (Captured Dirichlet energy). *Let $e \in H^1(\pi_d)$ satisfy $\|e\|_a = 1$, let $f_1, \dots, f_N \in H^1(\pi_d)$ span a space V, and define $G_{ij} := a(f_i, f_j)$ and $b_i := a(e, f_i)$. Then*

$$\mathrm{dist}_a(e, V)^2 = 1 - \mathbf{b}^\top \mathbf{G}^\dagger \mathbf{b}, \qquad C = \mathbf{b}^\top \mathbf{G}^\dagger \mathbf{b}.$$

This quantity depends only on V modulo constants and is unchanged if each f_i is multiplied by a nonzero scalar or shifted by a constant. Moreover, if $\mathbf{H} \preceq \mathbf{G}$ with $\mathbf{H} \succ 0$, then $C \le \mathbf{b}^\top \mathbf{H}^{-1} \mathbf{b}$.

Proof. If $\mathbf{G}\mathbf{v} = 0$, then $\|\sum_i v_i f_i\|_a^2 = \mathbf{v}^\top \mathbf{G}\mathbf{v} = 0$, and therefore $\mathbf{b}^\top \mathbf{v} = a(e, \sum_i v_i f_i) = 0$. Thus $\mathbf{b} \perp \ker(\mathbf{G})$. For any coefficient vector $\boldsymbol{\alpha}$,

$$\left\| e - \sum_i \alpha_i f_i \right\|_a^2 = 1 - 2\boldsymbol{\alpha}^\top \mathbf{b} + \boldsymbol{\alpha}^\top \mathbf{G} \boldsymbol{\alpha}.$$

Minimizing gives $\boldsymbol{\alpha} = \mathbf{G}^\dagger \mathbf{b}$ and hence $\mathrm{dist}_a(e, V)^2 = 1 - \mathbf{b}^\top \mathbf{G}^\dagger \mathbf{b}$. The invariance under rescaling and shifts follows because constants are null directions of the Dirichlet seminorm. Finally, if $\mathbf{H} \preceq \mathbf{G}$ with $\mathbf{H} \succ 0$, monotonicity of the inverse gives $C \le \mathbf{b}^\top \mathbf{H}^{-1} \mathbf{b}$. □

Since the Dirichlet form diagonalizes in the Hermite basis,

$$\mathbf{G} = \sum_\alpha \lambda_\alpha \mathbf{v}_\alpha \mathbf{v}_\alpha^\top, \qquad (\mathbf{v}_\alpha)_i := \langle f_i, h_\alpha \rangle_{L^2(\pi_d)}.$$

Every eigenmode therefore contributes a positive semidefinite rank-one term to $\mathbf{G}$. The proof below exploits this by lower-bounding $\mathbf{G}$ using only the large collection of irrelevant ambient modes.

B.4 Linear signal modes

We first prove the linear part of Theorem 6. Since $\lambda_{\mathrm{lin}} = \tau^{-2}$ and $\|z_1\|_{L^2(\pi_d)} = 1$, the energy-normalized target is $e_{\mathrm{lin}} := \tau z_1$. Using the normalized kernel sections $\widetilde{K}_i$, its Dirichlet correlation vector satisfies $b_i = a(e_{\mathrm{lin}}, \widetilde{K}_i) = \frac{c_s}{\tau} W_{i1}$.

Let $\mathbf{w} := (W_{11}, \dots, W_{N1})^\top$ and $\mathbf{W}_{\mathrm{noise}} := (W_{iq})_{\substack{1 \le i \le N \\ 3 \le q \le d}} \in \mathbb{R}^{N \times (d-2)}$. Each of the $d-2$ linear noise modes has eigenvalue 1 and coefficient $c_n W_{iq}$, so retaining only these modes in the Dirichlet Gram matrix gives

$$\mathbf{G} \succeq c_n^2 \mathbf{W}_{\mathrm{noise}} \mathbf{W}_{\mathrm{noise}}^\top.$$

For $N < d-2$, Lemma 8 yields

$$C_{N,\epsilon}(\phi_{\mathrm{lin}}) \le \frac{\kappa^2}{\tau^2} \mathbf{w}^\top (\mathbf{W}_{\mathrm{noise}} \mathbf{W}_{\mathrm{noise}}^\top)^{-1} \mathbf{w} \le \tau^2 \mathbf{w}^\top (\mathbf{W}_{\mathrm{noise}} \mathbf{W}_{\mathrm{noise}}^\top)^{-1} \mathbf{w} := \Xi_N,$$

where the second inequality uses $\kappa < \tau^2$. Crucially, Ξ_N does not depend on ϵ.

The vector $\mathbf{w}$ is independent of $\mathbf{W}_{\mathrm{noise}}$, with $\mathbb{E}[\mathbf{w}\mathbf{w}^\top] = I_N$, while $\mathbf{W}_{\mathrm{noise}} \mathbf{W}_{\mathrm{noise}}^\top$ is Wishart with $d-2$ degrees of freedom. Hence, for $N < d-3$,

$$\mathbb{E}\left[(\mathbf{W}_{\mathrm{noise}} \mathbf{W}_{\mathrm{noise}}^\top)^{-1} \right] = \frac{1}{d-N-3} I_N,$$

and therefore $\mathbb{E}[\Xi_N] = \tau^2 N/(d-N-3)$. If $N = o(d)$, this tends to zero. Since $\Xi_N \ge 0$, Markov's inequality gives

$$\sup_{\epsilon > 0} C_{N,\epsilon}(\phi_{\mathrm{lin}}) \le \Xi_N \xrightarrow{\mathbb{P}} 0,$$

proving the first part of Theorem 6.

B.5 QUADRATIC SIGNAL MODES

The quadratic case requires a lower bound on the Gram matrix generated by the many degree-two noise modes.

Lemma 9 (Gram lower bound for Gaussian tensor features)**.** *Let* $\mathbf{a}_i, \mathbf{d}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_s)$ *for* $i = 1, \dots, N$*, with all vectors mutually independent. Define* $\mathbf{q}_i := \mathbf{a}_i \otimes \mathbf{d}_i \in \mathbb{R}^D$ *with* $D := s^2$*, and let* $\mathbf{Q} \in \mathbb{R}^{N \times D}$ *have rows* $\mathbf{q}_i^\top$*. If* $N \log^2 s = o(s^2)$*, then*

$$\mathbb{P}\left(\mathbf{Q}\mathbf{Q}^\top \succeq \frac{1}{8} D I_N\right) \to 1.$$

Proof. Let $\mathbf{A} \in \mathbb{R}^{N \times s}$ have rows $\mathbf{a}_i^\top$, and for $q = 1, \dots, s$ define $\mathbf{v}_q := (d_{1q}, \dots, d_{Nq})^\top$. Conditional on $\mathbf{A}$, the vectors $\mathbf{v}_1, \dots, \mathbf{v}_s$ are independent standard Gaussian vectors in $\mathbb{R}^N$, and

$$\mathbf{Q}\mathbf{Q}^\top = \sum_{q=1}^{s} \operatorname{diag}(\mathbf{v}_q) \mathbf{A}\mathbf{A}^\top \operatorname{diag}(\mathbf{v}_q) := \sum_{q=1}^{s} \mathbf{Y}_q.$$

Consider the event $\mathcal{A} := \{\min_i \|\mathbf{a}_i\|_2^2 \geq s/2\} \cap \{\|\mathbf{A}\|_{\text{op}}^2 \leq 8(N+s)\}$. Standard Gaussian concentration gives $\mathbb{P}(\mathcal{A}^c) \leq N e^{-s/16} + e^{-(N+s)/2} = o(1)$ since $N = o(s^2)$. Set $T := 2\sqrt{\log(Ns)}$ and $\widehat{v}_{qi} := v_{qi} \mathbf{1}\{|v_{qi}| \leq T\}$, and define $\widehat{\mathbf{Y}}_q := \operatorname{diag}(\widehat{\mathbf{v}}_q) \mathbf{A}\mathbf{A}^\top \operatorname{diag}(\widehat{\mathbf{v}}_q)$. On the event $\mathcal{E} := \{\max_{q,i} |v_{qi}| \leq T\}$, we have $\widehat{\mathbf{Y}}_q = \mathbf{Y}_q$ for all q, while a union bound gives $\mathbb{P}(\mathcal{E}^c) \leq 2Nse^{-T^2/2} = 2/(Ns) = o(1)$.

Condition now on $\mathbf{A} \in \mathcal{A}$. The matrices $\widehat{\mathbf{Y}}_q$ are independent and positive semidefinite, with $\lambda_{\max}(\widehat{\mathbf{Y}}_q) \leq T^2 \|\mathbf{A}\|_{\text{op}}^2 \leq 32(N+s)\log(Ns)$. By symmetry, $\mathbb{E}[\widehat{v}_{qi}\widehat{v}_{qj}] = \beta \delta_{ij}$, where $\beta := \mathbb{E}[v^2 \mathbf{1}\{|v| \leq T\}] \geq 1/2$ for sufficiently large s. Hence $\mathbb{E}[\widehat{\mathbf{Y}}_q \mid \mathbf{A}] = \beta \operatorname{diag}(\|\mathbf{a}_1\|_2^2, \dots, \|\mathbf{a}_N\|_2^2)$, and therefore

$$\lambda_{\min}\left(\sum_{q=1}^{s} \mathbb{E}[\widehat{\mathbf{Y}}_q \mid \mathbf{A}]\right) = \beta s \min_i \|\mathbf{a}_i\|_2^2 \geq \frac{1}{4} s^2 = \frac{1}{4} D.$$

Applying the matrix Chernoff inequality with parameter $1/2$ gives

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{q=1}^{s} \widehat{\mathbf{Y}}_q\right) \leq \frac{1}{8} D \,\middle|\, \mathbf{A}\right) \leq N \exp\left(-\frac{s^2}{1024(N+s)\log(Ns)}\right).$$

Under $N \log^2 s = o(s^2)$, the right-hand side tends to zero. Combining this with $\mathbb{P}(\mathcal{A}^c) = o(1)$ and $\mathbb{P}(\mathcal{E}^c) = o(1)$ proves the claim. □

We now return to the quadratic signal mode $\phi_{\text{quad}} = z_1 z_2$. Its eigenvalue is $2\tau^{-2}$ and $\|\phi_{\text{quad}}\|_{L^2(\pi_d)} = 1$, so the energy-normalized target is $e_{\text{quad}} := (\tau/\sqrt{2}) z_1 z_2$. Its Dirichlet correlations with the normalized kernel sections satisfy $b_i = \frac{\sqrt{2}\, c_s^2}{\tau} W_{i1} W_{i2}$, and hence

$$\|\mathbf{b}\|_2^2 = \frac{2c_s^4}{\tau^2} \sum_{i=1}^{N} W_{i1}^2 W_{i2}^2.$$

Partition the noise coordinates $\{3, \dots, d\}$ into disjoint sets A and B of cardinality $s := \lfloor (d-2)/2 \rfloor$, discarding at most one coordinate, and let $D := s^2 = \Theta(d^2)$. For every $p \in A$ and $q \in B$, the function $z_p z_q$ is a quadratic noise eigenfunction with eigenvalue 2. Define $\mathbf{a}_i := (W_{ip})_{p \in A}$, $\mathbf{d}_i := (W_{iq})_{q \in B}$, and $\mathbf{q}_i := \mathbf{a}_i \otimes \mathbf{d}_i$, and let $\mathbf{Q} \in \mathbb{R}^{N \times D}$ have rows $\mathbf{q}_i^\top$. The contribution of these quadratic noise modes to the Dirichlet Gram matrix is $2c_n^4 \mathbf{Q}\mathbf{Q}^\top$, so

$$\mathbf{G} \succeq 2c_n^4 \mathbf{Q}\mathbf{Q}^\top.$$

If $N = o(d^2/\log^2 d)$, then $N \log^2 s = o(s^2)$, so Lemma 9 implies that with probability tending to one, $\mathbf{Q}\mathbf{Q}^\top \succeq \frac{1}{8} D I_N$. On this event, $\mathbf{G} \succeq \frac{1}{4} c_n^4 D I_N$, and Lemma 8 gives, simultaneously for all $\epsilon > 0$,

$$C_{N,\epsilon}(\phi_{\mathrm{quad}}) \le \frac{4\|\mathbf{b}\|_2^2}{c_n^4 D} = \frac{8\kappa^4}{\tau^2 D} \sum_{i=1}^N W_{i1}^2 W_{i2}^2 \le \frac{8\tau^6}{D} \sum_{i=1}^N W_{i1}^2 W_{i2}^2 := \Xi'_N.$$

Again, the final upper bound is independent of ϵ.

Because $\mathbb{E}[W_{i1}^2 W_{i2}^2] = 1$, we have $\mathbb{E}[\Xi'_N] = 8\tau^6 N/D \to 0$. Thus, for every $\delta > 0$,

$$\mathbb{P}\left(\sup_{\epsilon > 0} C_{N,\epsilon}(\phi_{\mathrm{quad}}) > \delta \right) \le o(1) + \frac{8\tau^6}{\delta} \frac{N}{D} \to 0,$$

proving the second part of Theorem 6.

B.6 Implication for the truncated inverse generator

The preceding theorem yields a direct consequence for the spectral preconditioner used in LAWGD. Assume $2 < \tau^2 < 3$. A nonconstant eigenvalue $\lambda_\alpha = \sum_{q=1}^d \alpha_q/\sigma_q^2$ is strictly below 1 only if α is supported entirely on the two signal coordinates, in which case the condition becomes $\alpha_1 + \alpha_2 < \tau^2$. Since $2 < \tau^2 < 3$, the complete nonconstant spectrum below 1 consists of

$$\phi_1 = z_1, \qquad \phi_2 = z_2, \qquad \phi_3 = \frac{z_1^2 - 1}{\sqrt{2}}, \qquad \phi_4 = z_1 z_2, \qquad \phi_5 = \frac{z_2^2 - 1}{\sqrt{2}},$$

with $\lambda_1 = \lambda_2 = \tau^{-2}$ and $\lambda_3 = \lambda_4 = \lambda_5 = 2\tau^{-2}$.

Define the oracle representation error of the rank-five inverse generator by

$$B_5(N,\epsilon)^2 := \sum_{k=1}^5 \frac{1}{\lambda_k^2} \operatorname{dist}_a(\phi_k, V_{N,\epsilon})^2.$$

For reference, the error of the zero approximation is

$$B_{5,\mathrm{triv}}^2 := \sum_{k=1}^5 \frac{\|\phi_k\|_a^2}{\lambda_k^2} = \sum_{k=1}^5 \frac{1}{\lambda_k} = \frac{7\tau^2}{2}.$$

Corollary 10 (Nonvanishing bias for the slow inverse generator)**.** *Assume $2 < \tau^2 < 3$ and $N = o(d^2/\log^2 d)$. Then*

$$\inf_{\epsilon > 0} B_5(N,\epsilon)^2 \ge \frac{\tau^2}{2} \left(1 - o_{\mathbb{P}}(1)\right).$$

Equivalently, $\inf_{\epsilon>0} B_5(N,\epsilon)^2 / B_{5,\mathrm{triv}}^2 \ge \frac{1}{7}(1 - o_{\mathbb{P}}(1))$.

Proof. Every term in $B_5(N,\epsilon)^2$ is nonnegative, so $B_5(N,\epsilon)^2 \ge \lambda_4^{-2} \operatorname{dist}_a(\phi_4, V_{N,\epsilon})^2$. Since $\phi_4 = z_1 z_2$ is $L^2(\pi_d)$-normalized, $\|\phi_4\|_a^2 = \lambda_4$, and hence $\operatorname{dist}_a(\phi_4, V_{N,\epsilon})^2 = \lambda_4 R_{N,\epsilon}(\phi_4)$. Therefore

$$B_5(N,\epsilon)^2 \ge \frac{R_{N,\epsilon}(\phi_4)}{\lambda_4} = \frac{\tau^2}{2} R_{N,\epsilon}(\phi_4).$$

By Theorem 6, $\inf_{\epsilon>0} R_{N,\epsilon}(\phi_4) = 1 - o_{\mathbb{P}}(1)$, proving the first claim. Dividing by $B_{5,\mathrm{triv}}^2 = 7\tau^2/2$ gives the relative form. □

The corollary is a statement about representation error rather than a convergence-rate lower bound for the resulting particle dynamics. It shows that, even when the slow eigenspace is confined to a fixed two-dimensional subspace, a Gaussian feature family that does not adapt to that subspace can retain nonvanishing error in its approximation of the slow inverse generator. This obstruction appears at the representation stage, before finite-particle or optimization error is introduced.

Remark 2 (Relation to random-feature lower bounds)**.** *The mechanism behind Theorem 6 is closely related to known high-dimensional limitations of random-feature and kernel methods, in which representing degree-ℓ structure can require on the order of d^ℓ features even when the target depends on only a bounded number of directions (Ghorbani et al., 2019; 2020; Mei et al., 2022). Here, the same phenomenon is established in the Dirichlet seminorm associated with the Langevin generator and translated into a representation barrier for the slow inverse generator relevant to spectral preconditioning.*

C Sample-Only Optimization Details

This section gives the implementation details of the neural transport objective introduced in Section 4.2. Proposition 5 shows that the exact forward and adjoint responses satisfy $\mathcal{D}h_g^* = \mathcal{T}g$ and $a_{\mathbf{b}}^* = \mathcal{T}^*\mathbf{b}$. In practice, these responses are represented by auxiliary scalar networks and solved only approximately. We therefore use a stopped-response surrogate that allows the rank-r transport factors to be updated without differentiating through the response optimization.

C.1 Stopped-response surrogate

The population low-rank objective contains the cross term $\mathcal{C}(\mathbf{b}, g) := \langle \mathbf{b}, \mathcal{T}g\rangle_\pi$. Although Proposition 5 gives several equivalent representations of its value at the exact responses, updating the scalar and vector factors independently requires the correct derivative with respect to both arguments. We therefore introduce the stopped-response surrogate

$$\widetilde{\mathcal{C}}(\mathbf{b}, g; h, a) := \langle \mathbf{b}, \mathcal{D}h\rangle_\pi + \langle g, a\rangle_\pi - \langle \mathcal{D}a, \mathcal{D}h\rangle_\pi.$$

During the factor update, h, a, and their input gradients are treated as fixed. The first term provides the response needed to update $\mathbf{b}$, the second provides the adjoint response needed to update g, and the final term corrects the surrogate value without contributing to the factor gradients.

C.2 Value and gradient consistency

Proposition 11 (Value and gradient consistency)**.** *Let $h = h_g^*$ and $a = a_{\mathbf{b}}^*$ be the exact forward and adjoint responses. Then*

$$\widetilde{\mathcal{C}}(\mathbf{b}, g; h_g^*, a_{\mathbf{b}}^*) = \langle \mathbf{b}, \mathcal{T}g\rangle_\pi.$$

Moreover, when the responses are held fixed during differentiation, the functional derivatives of $\widetilde{\mathcal{C}}$ with respect to $\mathbf{b}$ and g coincide with those of $\mathcal{C}$.

Proof. By the weak Poisson identities,

$$\langle \mathbf{b}, \mathcal{D}h_g^*\rangle_\pi = \langle g, a_{\mathbf{b}}^*\rangle_\pi = \langle \mathcal{D}a_{\mathbf{b}}^*, \mathcal{D}h_g^*\rangle_\pi = \langle \mathbf{b}, \mathcal{T}g\rangle_\pi.$$

Therefore the three terms in $\widetilde{\mathcal{C}}$ combine as $1 + 1 - 1$, giving the correct cross-term value. Now hold h_g^* and $a_{\mathbf{b}}^*$ fixed. For a perturbation $\delta\mathbf{b}$,

$$\mathcal{D}_{\mathbf{b}}\widetilde{\mathcal{C}}[\delta\mathbf{b}] = \langle \delta\mathbf{b}, \mathcal{D}h_g^*\rangle_\pi = \langle \delta\mathbf{b}, \mathcal{T}g\rangle_\pi = \mathcal{D}_{\mathbf{b}}\mathcal{C}[\delta\mathbf{b}].$$

Similarly, for a perturbation δg,

$$\mathcal{D}_g\widetilde{\mathcal{C}}[\delta g] = \langle \delta g, a_{\mathbf{b}}^*\rangle_\pi = \langle \delta g, \mathcal{T}^*\mathbf{b}\rangle_\pi = \mathcal{D}_g\mathcal{C}[\delta g].$$

Hence the stopped-response surrogate has both the correct value and the correct functional gradients with respect to the transport factors. □

C.3 Alternating optimization

At finite sample size, the response functions are only approximated by auxiliary networks. We therefore alternate between updating the response networks and updating the transport factors. With the current factors $(g_k, \mathbf{b}_k)$ fixed, we take K_H optimization steps on the forward responses and K_A steps on the adjoint responses, using fresh target minibatches for these response updates. The corresponding population objectives are

$$\mathcal{L}_H = \sum_{k=1}^{r}\left(\frac{1}{2}\|\mathcal{D}h_k\|_\pi^2 - \langle g_k, h_k\rangle_\pi\right), \qquad \mathcal{L}_A = \sum_{k=1}^{r}\left(\frac{1}{2}\|\mathcal{D}a_k\|_\pi^2 - \langle \mathbf{b}_k, \mathcal{D}a_k\rangle_\pi\right).$$

In our experiments, we use $K_H = K_A = 5$.

After the response updates, we hold the response networks and their input gradients fixed and draw a fresh minibatch for the transport-factor update. On this minibatch, we form the stopped-response

cross terms $\widetilde{c}_k := \langle \mathbf{b}_k, \mathrm{sg}(\mathcal{D}h_k)\rangle + \langle g_k, \mathrm{sg}(a_k)\rangle - \langle \mathrm{sg}(\mathcal{D}a_k), \mathrm{sg}(\mathcal{D}h_k)\rangle$, where sg denotes stop-gradient with respect to the transport-factor parameters. Then we jointly update all transport factors using

$$\widetilde{\mathcal{J}}_r = \sum_{i,j=1}^{r} \langle \mathbf{b}_i, \mathbf{b}_j\rangle\langle g_i, g_j\rangle - 2\sum_{k=1}^{r} \widetilde{c}_k,$$

where all expectations in this factor update are empirical averages over the fresh minibatch. This alternation is repeated throughout training, with each response update initialized from the current response-network parameters rather than reinitializing the networks.

Fresh sampling is used both across the response updates and for the subsequent factor update. This avoids repeatedly fitting the auxiliary Poisson problems to a fixed set of collocation points and reduces coupling between response-fitting error and the estimated factor gradient. The resulting procedure requires neither the target score $\nabla \log \pi$, the pointwise weighted divergence $\mathcal{D}_\pi^* \mathbf{b}$, nor explicit application of $\mathcal{L}_0^{-1}$.

C.4 Centering and factor balancing

The scalar factors are identifiable only up to additive constants in their action on the density-ratio discrepancy, since $\langle 1, u_\rho\rangle_\pi = 0$. We therefore center each factor as $g_k \leftarrow g_k - \mathbb{E}_\pi[g_k]$. The response potentials are likewise defined only up to additive constants, which do not affect their gradients.

The factorization $\mathbf{b}_k \otimes g_k$ also has a reciprocal scaling ambiguity: for any nonzero c, replacing g_k by $c\, g_k$ and $\mathbf{b}_k$ by $c^{-1}\mathbf{b}_k$ leaves the represented operator unchanged. To prevent severe imbalance during optimization, we periodically rescale each pair while preserving its rank-one product. Specifically, with $s_k := \sqrt{\|g_k\|_\pi / \|\mathbf{b}_k\|_\pi}$, we replace $g_k \leftarrow g_k / s_k$ and $\mathbf{b}_k \leftarrow s_k \mathbf{b}_k$. After rescaling, the corresponding response networks are refreshed because their defining Poisson problems depend on the current factors.

D Algorithms

For completeness, we collect here the finite-sample procedures corresponding to the kernel and neural transport approximations described in Section 4. Algorithm 1 summarizes the kernel construction, which replaces the population covariance and Dirichlet matrices by empirical averages and solves a generalized eigenvalue problem. Algorithm 2 summarizes the neural procedure, which alternates between weak Poisson-response updates and joint updates of the transport factors.

Algorithm 1 Kernel transport approximation

Require: Target samples $\{\mathbf{x}_n\}_{n=1}^N \sim \pi$, rank r

1: Choose a differentiable scalar feature family $\boldsymbol{\psi}$
2: Center the features: $\bar{\boldsymbol{\psi}}(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x}) - \frac{1}{N}\sum_{n=1}^N \boldsymbol{\psi}(\mathbf{x}_n)$
3: Compute feature gradients $\mathbf{J}(\mathbf{x}) = \nabla \bar{\boldsymbol{\psi}}(\mathbf{x})$
4: Form $\widehat{\mathbf{G}} = \frac{1}{N}\sum_{n=1}^N \bar{\boldsymbol{\psi}}(\mathbf{x}_n)\bar{\boldsymbol{\psi}}(\mathbf{x}_n)^\top$
5: Form $\widehat{\mathbf{S}} = \frac{1}{N}\sum_{n=1}^N \mathbf{J}(\mathbf{x}_n)^\top \mathbf{J}(\mathbf{x}_n)$
6: Solve $\widehat{\mathbf{G}}\mathbf{q}_k = \sigma_k^2 \widehat{\mathbf{S}}\mathbf{q}_k$, with $\mathbf{q}_i^\top \widehat{\mathbf{S}}\mathbf{q}_j = \delta_{ij}$
7: Let $\mathbf{q}_1, \dots, \mathbf{q}_r$ denote the generalized eigenvectors with largest σ_k
8: Set $g_k(\mathbf{x}) = \bar{\boldsymbol{\psi}}(\mathbf{x})^\top \mathbf{q}_k$ and $\mathbf{b}_k(\mathbf{x}) = \mathbf{J}(\mathbf{x})\mathbf{q}_k$
9: **return** $\widehat{\mathcal{T}}_r = \sum_{k=1}^r \mathbf{b}_k \otimes g_k$

E Experimental details

All methods were implemented in PyTorch on four NVIDIA H100 NVL GPUs. Unless stated otherwise, we use rank $r = 5$, report results over independent random seeds, and use $F = 1000$ kernel features. Hyperparameters are selected without access to test metrics.

Algorithm 2 Neural transport approximation

Require: Target samples $\{\mathbf{x}_n\}_{n=1}^N \sim \pi$, rank r, response steps K_H, K_A
Initialize neural factor networks $\{g_k, \mathbf{b}_k\}_{k=1}^r$ and auxiliary neural response networks $\{h_k, a_k\}_{k=1}^r$
repeat
  for $s = 1, \dots, K_H$ **do**
    Draw a fresh minibatch and let $\hat{\pi}_{B_s^H}$ denote its empirical distribution
    Update $\{h_k\}_{k=1}^r$ by minimizing $\sum_{k=1}^r \left(\frac{1}{2} \|\nabla h_k\|^2_{\hat{\pi}_{B_s^H}} - \langle g_k, h_k \rangle_{\hat{\pi}_{B_s^H}} \right)$
  end for
  for $s = 1, \dots, K_A$ **do**
    Draw a fresh minibatch and let $\hat{\pi}_{B_s^A}$ denote its empirical distribution
    Update $\{a_k\}_{k=1}^r$ by minimizing $\sum_{k=1}^r \left(\frac{1}{2} \|\nabla a_k\|^2_{\hat{\pi}_{B_s^A}} - \langle \mathbf{b}_k, \nabla a_k \rangle_{\hat{\pi}_{B_s^A}} \right)$
  end for
  Draw a fresh minibatch and let $\hat{\pi}_{B'}$ denote its empirical distribution
  for $k = 1, \dots, r$ **do**
    Form $\widetilde{c}_k = \langle \mathbf{b}_k, \mathrm{sg}(\nabla h_k) \rangle_{\hat{\pi}_{B'}} + \langle g_k, \mathrm{sg}(a_k) \rangle_{\hat{\pi}_{B'}} - \langle \mathrm{sg}(\nabla a_k), \mathrm{sg}(\nabla h_k) \rangle_{\hat{\pi}_{B'}}$
  end for
  Form $\widehat{\mathcal{J}}_r = \sum_{i,j=1}^r \langle \mathbf{b}_i, \mathbf{b}_j \rangle_{\hat{\pi}_{B'}} \langle g_i, g_j \rangle_{\hat{\pi}_{B'}} - 2 \sum_{k=1}^r \widetilde{c}_k$
  Update all factor pairs $(g_k, \mathbf{b}_k)_{k=1}^r$ jointly using $\widehat{\mathcal{J}}_r$
  Center the scalar factors g_k and balance the reciprocal scales of $(g_k, \mathbf{b}_k)$
until convergence
return $\widehat{\mathcal{T}}_r = \sum_{k=1}^r \mathbf{b}_k \otimes g_k$

E.1 Spiked Ornstein–Uhlenbeck

We use the spiked Gaussian model in Appendix B with ambient dimensions $d \in \{6, 16, 32, 64, 128\}$ and $N = 200{,}000$ target samples per dimension. The signal subspace remains two-dimensional as d varies, with two linear and three quadratic slow modes known analytically. To evaluate recovery, we form the scalar subspace spanned by the five learned factors and compute principal-angle errors against the known linear and quadratic eigenspaces, avoiding any assignment of individual factors to eigenfunctions. This metric is invariant to rotations or mixing within the learned subspace. We also report relative transport-operator error, singular-value error, and Wasserstein error of the resulting particle sampler.

E.2 Stochastic Allen–Cahn

Following Xu et al. (2025), we simulate

$$\mathrm{d}u = \left[D\nabla^2 u - \epsilon^{-2} u(u^2 - 1) \right] \mathrm{d}t + \sigma \, \mathrm{d}W_t$$

on a 128×128 grid with $D = 0.001$, $\epsilon = 0.01$, $\sigma = \sqrt{2}$, and $\Delta t = 10^{-5}$. We generate 300 independent training realizations and use the $t = 0$ and $t = 10^{-4}$ snapshots for the trajectory-based inverse-generator methods. Fields are compressed with a fully connected autoencoder $16384 \to 512 \to 128 \to 8$ with a symmetric decoder, and all methods operate in the resulting eight-dimensional latent space. Sampling uses 150 particles, 800 iterations, and step size 0.03. We report latent W_2, MMD^2, and energy distance over four independent pipelines, counting divergent runs as failures.

E.3 Pentapeptide

We use a 24-dimensional torsional representation, with trajectories 0–18 for training, 19–21 for validation, and 22–24 for testing. A reference reversible Markov state model is constructed from six TICA coordinates followed by clustering into 300 states at lag 2.5 ns. We evaluate the learned rank-5 scalar feature space using held-out VAMP-2 on the test trajectories.