Sobolev Regularized MMD Gradient Flow

Chenyang Tian[1,*] **Bharath K. Sriperumbudur**[2], **Arthur Gretton**[1,3], **Zonghao Chen**[1,*]
[1]University College London, [2]Pennsylvania State University, [3]Google Deepmind
{chenyang.tian.26, zonghao.chen.22}@ucl.ac.uk,
{bharathsv.ucsd, arthur.gretton}@gmail.com

Abstract

We analyze Sobolev descent [66], a regularized variant of maximum mean discrepancy (MMD) gradient flow based on a gradient penalty on the MMD witness function. This regularization mitigates the non-convexity of the MMD objective, and yields provable *global* convergence guarantees in MMD, in both continuous and discrete time. A surprising appeal is that our convergence analysis does not rely on isoperimetric assumptions on the target distribution. Instead, it is based on a regularity condition on the difference between kernel mean embeddings. We extend Sobolev descent such that it is applicable in both sampling (from an unnormalized target distribution)—using Stein kernels—and generative modeling settings, unlike previous works, where a gradient flow is suitable for only generative modeling or sampling but not both. We empirically validate the method on a broad range of tasks across both settings.

1 Introduction

Sampling and generative modelling, while often studied as distinct tasks, can be cast within a common framework of optimization over the space of probability measures $\mathcal{P}$ when only partial information about the target distribution π is available. In sampling, π is typically known only up to a normalizing constant, whereas in generative modelling it is accessible only through some i.i.d. samples. Indeed, algorithmic developments in one area have often inspired progress in the other. For instance, diffusion models, which have proved successful in generative modelling [76], have recently motivated the development of new diffusion-based sampling methods [51, 80]. Another example is provided by Wasserstein gradient flows, such as the maximum mean discrepancy (MMD) flow [4] and the kernel Stein discrepancy (KSD) flow [56]: the former is studied in generative modelling, while the latter is primarily used for sampling.

We address the case of a gradient flow based on a gradient-regularized integral probability metric (IPM), known as Sobolev descent [66]. This notion of gradient regularization was originally popularized in generative adversarial networks (GANs), a class of generative models since overshadowed by diffusion models: when an IPM is used as the critic, then *gradient regularization* of the critic witness function leads to higher-quality images [71, 41, 3, 66, 65]. Moving beyond the GAN setting, gradient regularization as in [66] also improves the convergence of both flow-based generative modelling via MMD flow and sampling via KSD flow [56]. As KSD is a special case of MMD with the Langevin Stein kernel [45], we refer to both resulting flows collectively as *Sobolev-regularized Maximum Mean Discrepancy gradient flow* (SrMMD flow)—in the literature, it is referred to as *Sobolev Descent* [66]—emphasizing the gradient regularization imposed on the witness function.

In this work, we establish that SrMMD flow enjoys provably fast convergence in both sampling and generative modelling. For sampling, its convergence rate is independent of isoperimetric characteristics of the target distribution π, denoted by β, such as log-Sobolev or Poincaré constants [5]. In

[*]Equal Contribution. This work was done when the first author was an undergraduate student at Tsinghua University.

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

particular, the Unadjusted Langevin Algorithm (ULA) requires $\mathcal{O}(\delta^{-1}\beta)$ iterations to achieve error $\delta > 0$ [23, 81], whereas SrMMD flow achieves the same accuracy in $\mathcal{O}(\delta^{-1/\alpha})$ iterations, where α characterizes the regularity of the trajectory. This suggests a clear advantage of SrMMD flow for multi-modal targets, where isoperimetric constants β may scale exponentially with the separation between modes [62, 21]. We admit, however, that our guarantee is stated in KSD, which is a weaker metric than the KL or χ^2 divergence commonly used to analyze ULA [52].

For generative modelling, although convergence in MMD is not the primary metric of practical interest, compared for instance with perceptual metrics such as the FID score [49], our analysis provides a *theoretical justification* that gradient regularization can accelerate convergence. Our analysis complements the original Sobolev Descent [66] in the following aspects: (i) The RKHS can be possibly infinite-dimensional, unlike [66], which handles only finite-dimensional RKHS; (ii) we establish a discrete-time convergence guarantee; and (iii) we explicitly characterize the residual term arising from the regularization. Together with the recent work on DrMMD flow [18], which provides a theoretical justification for L_2 regularization of the witness/critic function, these results help explain why combining L_2 and gradient regularization leads to improved empirical performance in MMD GANs [3, 65]. This addresses a longstanding open question in the literature. Although both of these regularization schemes are most prominent in the GAN literature, they have also appeared in flow-based generative models [35, 37, 40].

Contributions. We analyze Sobolev Descent [66], a.k.a. SrMMD flow, a regularized variant of MMD/KSD flow based on a gradient norm penalty on the witness function. The proposed framework applies both to generative modelling and to sampling from unnormalized targets, the latter through the use of Stein kernels. We begin in section 2 with an introduction to Wasserstein gradient flows on the MMD and KSD, and then describe the Sobolev-regularized MMD flow in section 3. We prove in section 4 that the SrMMD gradient regularization yields *global* convergence guarantees in both continuous- and discrete-time settings, under a source condition on the trajectory. In particular, the convergence rate measured in MMD/KSD is independent of the isoperimetric properties of the target distribution. In section 5, we show that SrMMD flow admits a tractable finite-sample implementation which, in contrast to diffusion-based approaches [75, 50, 51], does not require learning a score function at each iteration. Finally, in section 6, we validate the proposed method empirically on a broad range of tasks, including sampling, color transfer, and student-teacher network.

Notations. Let $\mathcal{P}_2(\mathbb{R}^d)$ be the space of probability distributions on $\mathbb{R}^d$ with finite second moment. For a continuous mapping $T : \mathbb{R}^d \to \mathbb{R}^d$, $T_{\#}\mu$ denotes the push-forward measure of μ. Given a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we denote by $\mathcal{H}$ its corresponding RKHS, endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\| \cdot \|_{\mathcal{H}}$. For any probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$, $(L_2(\mu), \| \cdot \|_{L_2(\mu)})$ is the space of all real valued μ-square integrable measurable functions on $\mathbb{R}^d$ with $\|f\|^2_{L_2(\mu)} := \int_{\mathbb{R}^d} |f(x)|^2 \mathrm{d}\mu(x)$. $(L^d_2(\mu), \| \cdot \|_{L^d_2(\mu)})$ is defined analogously with $\mathbb{R}^d$-valued functions. Given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, let $\partial_\ell f(x)$ denote the partial derivative with respect to the ℓ-th coordinate in x and denote $\nabla f(x) = [\partial_1 f(x), \dots, \partial_d f(x)]^\top \in \mathbb{R}^d$. Let $\partial_{1,\ell} k(x, x')$ denote the partial derivative with respect to the ℓ-th coordinate of x and let $\partial_{1,\ell}\partial_{2,\ell} k(x, x')$ denote the mixed partial derivative with respect to the ℓ-th coordinate in x and the ℓ-th coordinate in x'. Moreover, denote $\nabla_1 k(x, x') = [\partial_{1,1} k(x, x'), \dots, \partial_{1,d} k(x, x')]^\top \in \mathbb{R}^d$, and define $\nabla_2 k(x, x')$ analogously. Denote $[S] = \{0, 1, \dots, S\}$ for $S \in \mathbb{N}$. The notation $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ denotes explicitly the target distribution. We assume throughout that the kernel k and the target density π are sufficiently continuously differentiable for all displayed derivatives to be well defined.

2 Background & Related Work

In this section, we introduce the necessary background and present related work in the literature.

2.1 Wasserstein gradient flow

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional with a unique minimizer π. In many cases, $\mathcal{F}$ is of the form $\mathcal{F}(\cdot) = \mathcal{D}(\cdot \| \pi)$, where $\mathcal{D}$ is some probability divergence. The Wasserstein gradient flow (WGF) provides a principled approach to approximating π by evolving a curve of probability measures $(\mu_t)_{t \geq 0}$ along the steepest descent direction of $\mathcal{F}$ in the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, namely the

space of probability measures endowed with the Wasserstein-2 distance [2, 72]. Formally, the WGF is characterized by the continuity equation [2]: given an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\partial_t \mu_t - \nabla \cdot (\mu_t \boldsymbol{v}[\mu_t]) = 0, \tag{1}$$

where the vector field $\boldsymbol{v}[\mu_t] : \mathbb{R}^d \to \mathbb{R}^d$ is the Fréchet subdifferential of $\mathcal{F}$ evaluated at μ_t [2, Definition 11.1.1]. Under some mild growth conditions on $\boldsymbol{v}[\mu_t]$, the trajectory $(\mu_t)_{t\geq 0}$ is a unique absolutely continuous curve in $\mathcal{P}_2(\mathbb{R}^d)$ [2, Lemma 8.1.4]. Moreover, the trajectory $(\mu_t)_{t\geq 0}$ admits the following characteristic (particle) representation:

$$\mathrm{d}X_t = -\boldsymbol{v}[\mu_t](X_t)\mathrm{d}t, \quad \mathrm{Law}(X_t) = \mu_t. \tag{2}$$

Eq. (1) and Eq. (2) provide two equivalent yet complementary viewpoints. Eq. (1) gives the Eulerian perspective, which describes the evolution of the density, whereas Eq. (2) gives the Lagrangian perspective, which focuses on the trajectories of individual particles. In practice, implementations of WGF are typically based on the Lagrangian formulation, which amounts to simulating an interacting particle system. Specifically, given a fixed time horizon $S \in \mathbb{N}$ and an initial collection of N particles $\{x_0^{(i)}\}_{i=1}^N \sim \mu_0$, for $s \in [S]$,

$$x_{s+1}^{(i)} = x_s^{(i)} - \gamma \boldsymbol{v}[\hat{\mu}_s](x_s^{(i)}), \quad \hat{\mu}_s = \tfrac{1}{N}\textstyle\sum_{j=1}^N \delta_{x_s^{(j)}}. \tag{3}$$

Here, $\gamma > 0$ is the step size. Compared with Eq. (2), we see that the particle system in Eq. (3) involves both a forward Euler discretization in *time* and a finite-particle discretization in *space*.

2.2 MMD and KSD gradient flow

The maximum mean discrepancy (MMD) [39] between two probability measures is defined as the RKHS norm of the difference between their corresponding kernel mean embeddings [67], i.e., $\mathrm{MMD}(\mu, \pi) = \| \int k(x, \cdot)\, \mathrm{d}(\mu - \pi)(x) \|_{\mathcal{H}} =: \|m_\mu - m_\pi\|_{\mathcal{H}}$, where $m_\mu := \int k(x, \cdot)\, \mathrm{d}\mu(x)$ is called the kernel mean element w.r.t. μ. When the functional $\mathcal{F}$ takes the form of a squared MMD, i.e., $\mathcal{F}_{\mathrm{MMD}}(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \pi)$, the corresponding Fréchet subdifferential is $\boldsymbol{v}_{\mathrm{MMD}}[\mu](\cdot) = \int \nabla_2 k(x, \cdot)\, \mathrm{d}(\mu - \pi)(x)$ [4], which makes the MMD flow well-suited to generative modeling since $\boldsymbol{v}_{\mathrm{MMD}}[\mu](\cdot)$ can be estimated from samples from μ and π. Indeed, given M i.i.d. target samples $\{y^{(i)}\}_{i=1}^M$ from π and N particles $\{x_s^{(j)}\}_{j=1}^N$ at iteration s, $\boldsymbol{v}_{\mathrm{MMD}}[\hat{\mu}_s]$ admits a direct unbiased plug-in estimator $\boldsymbol{v}_{\mathrm{MMD}}[\hat{\mu}_s] \approx \frac{1}{N}\sum_{j=1}^N \nabla_2 k(x_s^{(j)}, \cdot) - \frac{1}{M}\sum_{j=1}^M \nabla_2 k(y^{(j)}, \cdot)$. The kernel Stein discrepancy (KSD) is a special instance of MMD under a specific choice of kernel k_π that satisfies the Stein identity $\int k_\pi(x, y)\mathrm{d}\pi(x) = 0, \forall y \in \mathbb{R}^d$ [61, 25]. Arguably, the most popular choice of k_π is the *Langevin Stein kernel*, defined through the *score*, $\mathfrak{s}(x) = \nabla \log \pi(x)$, and a base kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$k_\pi(x, y) := \mathfrak{s}(x)^\top \mathfrak{s}(y) k(x, y) + \mathfrak{s}(x)^\top \nabla_2 k(x, y) + \nabla_1 k(x, y)^\top \mathfrak{s}(y) + \mathrm{Tr}(\nabla_1 \nabla_2 k(x, y)), \tag{4}$$

where $\mathrm{Tr}(\nabla_1 \nabla_2 k(x, y)) = \sum_{\ell=1}^d \partial_{1,\ell}\partial_{2,\ell} k(x, y)$. Under mild tail conditions on the base kernel k, k_π satisfies the Stein's identity $\int k_\pi(x, y)\mathrm{d}\pi(x) = 0$ for all $y \in \mathbb{R}^d$ (Lemma B.5). As a result, the corresponding Fréchet subdifferential at iteration s equals $\boldsymbol{v}_{\mathrm{KSD}}[\hat{\mu}_s](\cdot) = \frac{1}{N}\sum_{j=1}^N \nabla_2 k_\pi(x_s^{(j)}, \cdot)$ [56]. This makes the KSD flow particularly suitable for sampling from the target π, which is known only up to a normalization constant.

To ensure the well-definedness of the MMD (resp. KSD) gradient flow, in the sense that the trajectory is absolutely continuous in $\mathcal{P}_2(\mathbb{R}^d)$, unique and does not explode in finite-time, we need to impose the following sufficient conditions on the growth and boundedness of the kernel k (resp. k_π).

Assumption 1 (Growth I)**.** There exists a constant $M > 0$ such that for any $x \in \mathbb{R}^d$, $|k(x, x)| \leq M(1 + \|x\|^4)$; and for any $i \in [d]$, $|\partial_{1,i}\partial_{2,i} k(x, x)| \leq M(1 + \|x\|^2)$.

Assumption 2 (Local Boundedness)**.** Let $\mathcal{K} = B(0, R)$ be a ball of radius R in $\mathbb{R}^d$. There exists a constant B_R such that $\sup_{x\in\mathcal{K}} \|\partial_{1,i}\partial_{1,j} k(x, \cdot)\|_{\mathcal{H}} \leq B_R$ for any $i, j \in [d]$.

Well-posedness. A proof of the well-posedness is provided in Proposition B.8. Compared with the conditions imposed in the original MMD flow paper [4], our assumptions constitute a strict relaxation. In particular, Assumptions (A)–(D) in [4] enforce global boundedness of $\sup_{x\in\mathbb{R}^d} \|\partial_{1,i} k(x, \cdot)\|_{\mathcal{H}}$ and

$\sup_{x\in\mathbb{R}^d} \|\partial_{1,i}\partial_{1,j}k(x,\cdot)\|_{\mathcal{H}}$, whereas here we only require *local* boundedness. As a consequence, beyond the globally bounded kernels commonly used in the Wasserstein gradient flow literature—such as Gaussian and Matérn kernels—our Assumptions 1 and 2 also accommodate certain unbounded kernels, including the polynomial kernel $k(x,y) = (x^\top y + 1)^2$, and the Stein kernel k_π under some conditions on π [56]; see the next section for details. Extension to kernels with even faster growth would be challenging; indeed, classical theory of differential equations shows that linear-growth conditions on the vector field are crucial in ruling out finite-time blow-up [79, Problem 2.18]. These conditions exclude, however, the negative distance kernel $k(x,y) = -\|x-y\|$ [48, 42], Riesz kernels $k(x,y) = -\|x-y\|^r$ for $r \in (0,2)$ [73, 1, 47] and the Coulomb kernels $k(x,y) = (d-2)\|x-y\|^{-(d-2)}$ [12] due to the singular behaviour at the origin. Although the MMD gradient flows of these kernels exhibit strong empirical performances in practice, establishing rigorous well-posedness remains an open problem; existing results are essentially restricted to the torus [12, 24] or the real line $\mathbb{R}^1$ [32]. For this reason, we do not consider such kernels in the present work.

Convergence. Beyond well-posedness, another central question for Wasserstein gradient flows is convergence, namely whether $\lim_{t\to\infty} \mu_t = \pi$ under some proper metric. A classical sufficient condition is that the functional $\mathcal{F}$ be convex along Wasserstein-2 geodesics, i.e., *geodesically* convex [82]. However, this property typically holds only for f-divergences and a restricted class of log-concave target distributions π [84, 21]. By contrast, the MMD and KSD functionals are *not* geodesically convex, even when π is log-concave [4, 56]. To address this limitation, a recent line of work has sought to improve convergence of these WGFs through spectral regularization [22, 18, 46].

3 Sobolev-regularized Maximum Mean Discrepancy (SrMMD) Flow

In this section, we review the Sobolev descent procedure defined in [66], aka SrMMD flow, which is a form of Sobolev regularization for MMD and KSD gradient flows. We then establish a new well-posedness result for this flow. In the next section, we study its convergence in both continuous and discrete time.

Definition 3.1 (SrMMD, Definition 2 of [66])**.** Let $\lambda > 0$ be fixed. For two fixed $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, the Sobolev regularized Maximum Mean Discrepancy (SrMMD) is defined as

$$\mathrm{SrMMD}(\mu,\pi) = \sup_{f\in\mathcal{H}:\, \|\nabla f\|^2_{L^d_2(\mu)} + \lambda\|f\|^2_{\mathcal{H}} \le 1} \int f \mathrm{d}(\mu - \pi). \tag{5}$$

Definition 3.2 (SrMMD flow, Section 3 of [66])**.** Let $\lambda > 0$ be fixed. For an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, the SrMMD flow is a trajectory of probability measures $(\mu_t)_{t\ge 0}$ satisfying the continuity equation

$$\partial_t \mu_t - \nabla\cdot(\mu_t \nabla f_{\mu_t,\pi}) = 0,$$

or equivalently, admitting a characteristic representation:

$$\mathrm{d}X_t = -\nabla f_{\mu_t,\pi}(X_t)\mathrm{d}t, \quad \mathrm{Law}(X_t) = \mu_t, \tag{6}$$

where $f_{\mu,\pi} \in \mathcal{H}$ is the witness function that attains the supremum in Definition 3.1.

It is immediate that SrMMD belongs to the family of integral probability metrics [68]. In contrast to $\mathrm{MMD}(\mu,\pi) = \|m_\mu - m_\pi\|_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}\le 1} \int f \mathrm{d}(\mu-\pi)$ whose hypothesis class is the unit ball of RKHS $\mathcal{H}$ [39], SrMMD enforces an additional *gradient penalty* on the witness function f. The resulting gradient field $\nabla f_{\mu,\pi}$ is then used in SrMMD flow to transport particles. Gradient penalization of the witness function has been widely used as a critical heuristic in generative modelling, both in classical generative adversarial networks [41, Eq. (3)][10][3, Eq.(6)] and in recent flow-based approaches [37, Eq. (9)].

The witness function $f_{\mu,\pi}$ actually admits a closed-form expression. To see this, we consider a linear operator $D_\mu : f \in \mathcal{H} \mapsto \nabla f$ which maps a function f to its first-order derivatives. When the kernel satisfies $\int \partial_{1,i}\partial_{2,i}k(x,x)\mathrm{d}\mu(x) < \infty$ for any $i \in [d]$, the operator D_μ is well defined, Hilbert-Schmidt, $\mathrm{ran}(D_\mu) \subseteq L^d_2(\mu)$, and it admits a well-defined adjoint operator $D^*_\mu : L^d_2(\mu) \to \mathcal{H}$ (see Lemma B.3). Let $S_\mu = D^*_\mu D_\mu : \mathcal{H} \to \mathcal{H}$ be the *gradient covariance* operator, which is bounded, linear, Hilbert-Schmidt (see Lemma B.4), satisfying $\langle f, S_\mu g\rangle_{\mathcal{H}} = \langle D_\mu f, D_\mu g\rangle_{L^d_2(\mu)} = \langle \nabla f, \nabla g\rangle_{L^d_2(\mu)}$ for all $f, g \in \mathcal{H}$.

Proposition 3.3 (Witness function $f_{\mu,\pi}$, Proposition 1 of [66]). *Suppose* $\int k(x,x)\mathrm{d}\mu(x) < \infty$, $\int k(x,x)\mathrm{d}\pi(x) < \infty$, $\int \partial_{1,i}\partial_{2,i}k(x,x)\mathrm{d}\mu(x) < \infty$ *for all* $i \in [d]$. *Then,* $f_{\mu,\pi}$ *attaining the supremum in the definition of SrMMD (Definition 3.1) admits a closed form expression:* $f_{\mu,\pi} = (S_\mu + \lambda \mathrm{Id})^{-1}(m_\mu - m_\pi)$.

The original Proposition 1 of [66] is proved under a bounded-kernel assumption. We extend this result to unbounded kernels, which is essential for the sampling setting. A proof is given in Appendix A.1 for completeness. Although the expression for $f_{\mu,\pi}$ appears involved, since it depends on an inverse operator, we show in Section 5 that its dual representation leads to a closed-form expression with a direct plug-in finite particle implementation. If the target π is known only up to a normalization constant, then under a Stein kernel k_π, we have $f_{\mu,\pi} = (S_\mu + \lambda \mathrm{Id})^{-1} m_\mu$ since $m_\pi = 0$, which admits a direct plug-in finite particle implementation.

The following result establishes the well-posedness of the SrMMD flow defined in Definition 3.2, and is proved in Appendix A.2.

Theorem 3.4 (Well-posedness of the SrMMD flow). *Suppose that Assumptions 1 and 2 hold, and let* $T \in [0,\infty)$ *be a finite time horizon. Let* $\mu_0, \pi \in \mathcal{P}_2(\mathbb{R}^d)$ *with* $\mathrm{MMD}(\mu_0,\pi) < C_{\mathrm{MMD}}$. *Suppose* μ_0 *admits a density and is supported on a compact domain. Then,* $(\mu_t)_{t\in[0,T]}$ *is a unique absolutely continuous curve in* $\mathcal{P}_2(\mathbb{R}^d)$, *each* μ_t *admits a density, and the flow does not blow up in finite time.*

In the context of sampling, for a Stein kernel k_π constructed from a smooth translation invariant base kernel k, Assumptions 1 and 2 hold when the score function satisfies $\|\mathfrak{s}(x)\| \le M_0(1+\|x\|)$ and $\|\mathbf{J}\mathfrak{s}(x)\|_F \le M_0(1+\|x\|)$ for some constant M_0 (Proposition B.11). Here, $\mathbf{J}\mathfrak{s}(x) \in \mathbb{R}^{d\times d}$ denotes the Jacobian matrix of the score function s. Compared with the original KSD flow of [56], which assumes $\|\mathfrak{s}(x)\| \le M_0(1+\sqrt{\|x\|})$ and $\|\mathbf{J}\mathfrak{s}(x)\|_F \le M_0(1+\sqrt{\|x\|})$, our Theorem 3.4 allows weaker growth conditions on s. Our growth conditions on s are satisfied by a broader class of targets, such as Gaussian-type and exponential-type distributions [60, 56]. A closely related work is [31], which studies MMD flows with positive and negative distance kernels on the real line, however, their Sobolev regularization is introduced in the quantile-space to ensure existence of the flow [32].

Remark 3.5. The SrMMD flow in Definition 3.2 is *not* the Wasserstein gradient flow of the functional $\mathcal{F}(\mu) := \mathrm{SrMMD}(\mu,\pi)$. Indeed, such a gradient flow would require differentiating the operator $(S_\mu + \lambda\mathrm{Id})^{-1}$ with respect to μ, which is analytically cumbersome. Instead, we introduce a Sobolev-regularized variant of the MMD flow, whose vector field is given by $\nabla f_{\mu,\pi} = \nabla(S_\mu + \lambda\mathrm{Id})^{-1}(m_\mu - m_\pi)$, in contrast to the standard MMD flow, whose vector field is $\nabla(m_\mu - m_\pi)$. However, the SrMMD flow can be identified as the gradient flow of MMD functional under a different metric; see Appendix A.6.

Remark 3.6 (Connection with DrMMD flow). A closely related flow is the (De)-regularized MMD flow, which is the Wasserstein flow of $\mathrm{DrMMD}(\mu,\pi) = \sup_{\|f\|^2_{L_2(\pi)}+\lambda\|f\|^2_{\mathcal{H}}\le 1} \int f \,\mathrm{d}(\mu-\pi)$ [18, 77]. Compared with the SrMMD flow, there are two key differences: (i) SrMMD employs gradient regularization, and (ii) the regularization in SrMMD is defined with respect to μ instead of π. We also study a hybrid regularization that combines both L_2 and gradient penalties, referred to as Hybrid-regularized MMD (HrMMD) flow, which matches the actual regularization schemes used in MMD GANs [3]. The HrMMD objective is the same as the Sobolev-Fisher discrepancy studied by [65], where it was subsequently used as the objective to be minimized in the Wasserstein-Fisher-Rao geometry. See Appendix D.6 for details.

4 Global Convergence of SrMMD Gradient Flow

In this section, we study convergence of the SrMMD flow $(\mu_t)_{t\ge 0}$ towards the target π.

Theorem 4.1 (Exponential decay in continuous time). *Let* $\lambda > 0$. *Let* $T > 0$ *be a fixed time horizon. Suppose all conditions in Theorem 3.4 hold. Additionally, suppose that* $m_{\mu_t} - m_\pi \in \overline{\mathrm{Ran}(S_{\mu_t})}$ *for all* $t \in [0,T]$. *Then, for the SrMMD flow* $(\mu_t)_{t\in[0,T]}$,

$$\mathrm{MMD}^2(\mu_T,\pi) \le e^{-2T}\,\mathrm{MMD}^2(\mu_0,\pi) + 2e^{-2T}\int_0^T e^{2t} R(\lambda,t)\,\mathrm{d}t,$$

where the residual $R(\lambda,t)$ *satisfies* $\lim_{\lambda\to 0+} R(\lambda,t) = 0$ *for all* $t \in [0,T]$. *If, in addition, there exists* $0 < \alpha \le 1$ *such that* $m_{\mu_t} - m_\pi \in \mathrm{Ran}\, S_{\mu_t}^{\alpha/2}$ *for all* $t \in [0,T]$; *then,* $R(\lambda,t) = \mathcal{O}(\lambda^\alpha)$.

Algorithm	Metric	Iteration complexity
LMC [23, 81, 20]	KL / χ^2	$\tilde{\mathcal{O}}(\delta^{-1}\beta)$
Averaged LMC [7]	Fisher	$\mathcal{O}(d^2\delta^{-2})$
SVGD [8]	KSD^2	$\mathcal{O}(\delta^{-1})$
R-SVGD [46]	Fisher	$\mathcal{O}(\delta^{-1})$
DrMMD [18]	χ^2	$\tilde{\mathcal{O}}(\delta^{-(\alpha+1)/\alpha}\beta)$
SrMMD	MMD^2 / KSD^2	$\tilde{\mathcal{O}}(\delta^{-1/\alpha})$

Table 1: Iteration complexity of several sampling methods. Here, β denotes the isoperimetric constant of π, and α denotes the regularity parameter of the flow. The guarantees are stated in different metrics: χ^2 divergence controls KL divergence and both typically control KSD [38, 52].

The proof, which is given in Appendix A.3, builds on the dissipation analysis in Appendix C.2 of [66]. However, our result is more general: it holds for all $\lambda > 0$, allows for possibly infinite-dimensional RKHSs, and accommodates unbounded kernels. Another technical distinction is that we do not impose a vanishing boundary condition; see Remark A.1. The condition $m_{\mu_t} - m_\pi \in \overline{\mathrm{Ran}(S_{\mu_t})}$ is mild: it holds whenever μ_t has full support on $\mathbb{R}^d$; see Lemma A.2. Theorem 4.1 shows that the SrMMD flow converges exponentially fast to the target distribution π, as measured either by the MMD (or KSD with a Stein kernel k_π), up to a residual barrier term. This barrier term is of explicit order $\mathcal{O}(\lambda^\alpha)$ under a stronger condition $m_{\mu_t} - m_\pi \in \mathrm{Ran}\, S_{\mu_t}^{\alpha/2}$, which quantifies the smoothness of $m_{\mu_t} - m_\pi$ relative to the operator S_{μ_t}. Such source conditions are standard in nonparametric regression [27, 36, 19, 18]. See Remark 4.2 for a detailed discussion. Our Theorem 4.1 shall be contrasted with the convergence of the original MMD flow (c.f. Theorem 6 of [3]) where the barrier term cannot be explicitly characterized.

In the context of sampling from an unnormalized target π, most of the existing convergence rates are established under various tail conditions on π, such as Poincaré (sub-exponential tail) or log-Sobolev (sub-Gaussian tail) [84, 81, 64, 23]. By contrast, Theorem 4.1 relies only on conditions on π that are implicitly encoded through the Stein kernel k_π in Assumptions 1 and 2. As a result, the SrMMD flow enjoys an explicit decay rate $\exp(-2T)$ independent of the tail conditions of π. A likely explanation is that our convergence is stated with the weaker MMD/KSD metric compared with KL or χ^2-divergence metrics as in ULA [23, 81] or SVGD [55]. We postpone a detailed iteration-complexity analysis in this context to the discrete-time setting, after establishing the corresponding discrete-time convergence.

Remark 4.2 (Source condition). To motivate the condition that $m_\mu - m_\pi \in \mathrm{Ran}\, S_\mu^{\alpha/2}$, consider a periodic Matérn kernel on a d-dimensional torus $\mathcal{X} = \mathbb{T}^d = [0, 2\pi)^d$ whose RKHS is norm equivalent to a Sobolev space $H^r(\mathcal{X})$ for $r \geq 1$ [83, Corollary 10.48]. Then, according to Lemma A.3,

$$\mathrm{Ran}(S_\mu^{\alpha/2}) \cong H^{r(\alpha+1)-\alpha}(\mathcal{X}) \cap \mathrm{span}\{1\}^{\perp_{\mathcal{H}}}.$$

Since $\langle m_\mu - m_\pi, 1\rangle_{\mathcal{H}} = 0$, so the $\mathrm{span}\{1\}^{\perp_{\mathcal{H}}}$ part is satisfied. Next, we consider the $H^{r(\alpha+1)-\alpha}(\mathcal{X})$ part. Since $r > 1$, this implies that $m_\mu - m_\pi$ belongs to a smaller Sobolev space $H^{r(\alpha+1)-\alpha}(\mathcal{X}) \subset H^r(\mathcal{X}) \cong \mathcal{H}$. A sufficient condition for this to hold is when $\frac{\mathrm{d}\pi}{\mathrm{d}\mu} - 1 \in H^{r\alpha-r-\alpha}(\mathcal{X})$ (Lemma A.4), which is weaker than the condition imposed in DrMMD flow that $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1 \in H^{r\alpha}(\mathcal{X})$.

Next, we consider SrMMD descent, the forward Euler discretization of the SrMMD flow, which is referred to as *Sobolev descent* in [66]. Specifically, given an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, a learning rate $\gamma > 0$, and a fixed time horizon $S \in \mathbb{N}$, SrMMD descent is a sequence of probability measures $(\mu_s)_{s\in[S]}$ defined by:

$$\mu_{s+1} = (\mathrm{Id} - \gamma \nabla f_{\mu_s,\pi})_{\#}\mu_s. \tag{7}$$

Similarly as in continuous time, the SrMMD descent scheme is well-posed under Assumptions 1 and 2. We now turn to the convergence analysis of the SrMMD descent.

Assumption 3 (Growth II)**.** There exists a constant $M > 0$ such that for any $x, y \in \mathbb{R}^d$, $\|\mathbf{H}_1 k(x,y)\|_{\mathrm{op}} \leq M(1 + \|y\|^2)$, where $[\mathbf{H}_1 k(x,y)]_{i,j} := \partial_{1,i}\partial_{1,j}k(x,y)$ denotes the Hessian of k with respect to its first argument and $\|\cdot\|_{\mathrm{op}}$ is the matrix operator norm.

The additional condition controls the growth of the Hessian of the kernel k. For a Langevin Stein kernel k_π, this condition is satisfied when the score function $\mathfrak{s}$ is Lipschitz and has a bounded Hessian (See Proposition B.11). Such additional conditions on $\mathfrak{s}$ are commonly used for *discrete-time* convergence analysis of sampling, e.g., [8, Assumption 2], [55, Assumption A2], [23], etc.

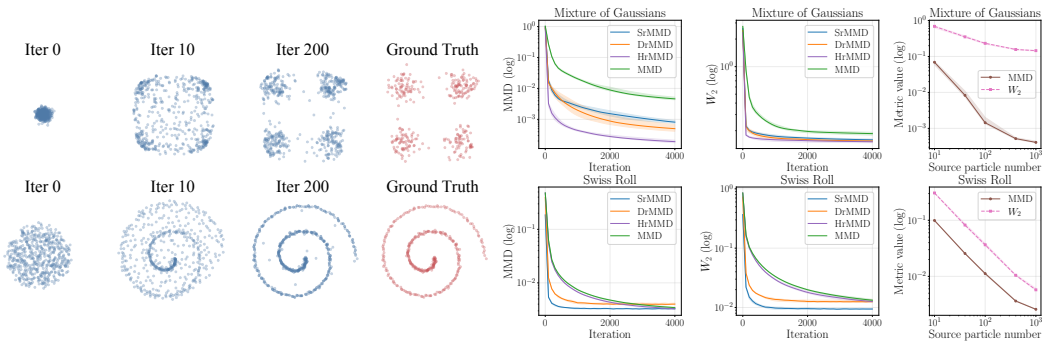

Figure 1: **Top row:** Mixture of Gaussians. **Bottom row:** Swiss roll. From left to right: particle evolution during training; MMD and W_2 versus iteration; and final MMD and W_2 versus the number of source particles N. Results are aggregated over 10 random seeds; solid curves show the median, and shaded regions show the 25%–75% percentile.

Theorem 4.3 (Exponential decay in discrete time)**.** *Suppose Assumptions 1 to 3 hold. Let $S \in \mathbb{N}$ be a fixed time horizon. Let $\gamma \in (0, \frac{1}{2})$ be a fixed step size. Let $(\mu_s)_{s\in[S]}$ be the SrMMD descent defined in Eq.* (7)*. Suppose that $(\mu_s)_{s\in[S]}$ has uniformly bounded second moment $C > 0$. Suppose there exists $0 < \alpha \leq 1$ such that $m_{\mu_s} - m_\pi = S_{\mu_s}^{\alpha/2} q_s$ with $q_s \in \mathcal{H}$ for all $s \in [S]$. Then,*

$$\mathrm{MMD}^2(\mu_S, \pi) \leq \exp(-2S\gamma)\,\mathrm{MMD}^2(\mu_0, \pi) + 4(dC_0)^\alpha \gamma^{1+\alpha} \textstyle\sum_{s=0}^{S-1} (1-2\gamma)^{S-1-s} \|q_s\|_{\mathcal{H}}^2.$$

Here, the constant C_0 only depends on C and M. If additionally there exists a constant Q independent of S such that $\|q_s\|_{\mathcal{H}}^2 \leq Q$ for all $s \in [S]$, then with $\lambda = dC_0\gamma$, we have

$$\mathrm{MMD}^2(\mu_S, \pi) \leq \exp(-2S\gamma)\,\mathrm{MMD}^2(\mu_0, \pi) + 4(dC_0)^\alpha \gamma^\alpha Q.$$

The proof is given in Appendix A.4. We are now ready to analyze the iteration complexity of SrMMD descent. To reach final error $\mathrm{MMD}^2(\mu_S, \pi) \leq \delta$ (or equivalently $\mathrm{KSD}^2(\mu_S, \pi) \leq \delta$), it suffices to pick a step size $\gamma = \mathcal{O}(\delta^{1/\alpha})$ and iteration number $S = \mathcal{O}(\delta^{-1/\alpha} \log(\delta^{-1}))$. Among existing results, the one closest in spirit to ours is Stein variational gradient descent (SVGD) [60], which has recently been shown to admit an iteration complexity of $\mathcal{O}(\delta^{-1})$ [8] also under the KSD metric. The iteration complexity of SrMMD is constrained by $0 < \alpha \leq 1$ and does not improve for further smoothness $\alpha > 1$. This limitation arises from essentially using Tikhonov regularization $(S_\mu + \lambda \mathrm{Id})^{-1}$ in the formulation of SrMMD—a well-known saturation phenomenon [69, 58, 63]. Such a saturation can be avoided by using other spectral regularization strategies, such as Showalter, Landweber, or cutoff regularization [33, 9, 43, 44]. Moreover, the KSD guarantee of SVGD in [8] is *ergodic*, in the sense that it controls the averaged quantity $\frac{1}{S}\sum_{s=1}^{S} \mathrm{KSD}^2(\mu_s, \pi)$ in contrast to our last-iterate guarantee.

Beyond this, most existing iteration-complexity results in the sampling literature rely on structural constants β of the target π, such as Poincaré or log-Sobolev constants [81, 22, 23]. These constants, however, are exponential in the separation between modes when the distribution π is multi-modal [62, 21], making convergence too slow to occur in a reasonable number of iterations. In contrast, our convergence result does not rely on such constants, and therefore does not suffer from the multi-modality of π. The price we pay here is the regularity condition $m_\mu - m_\pi \in \mathrm{Ran}(S_\mu^{\alpha/2})$ which we have discussed in Remark 4.2. We admit that this condition is hard to verify in practice. See Table 1 for a comparison on the iteration complexities of all methods. The original KSD flow is shown to converge under a Stein-Poincaré inequality [55]—a condition that is proved never to hold [30]. This highlights another advantage of Sobolev regularization. Recent theoretical advances have also established iteration complexity bounds for diffusion-based samplers that are independent of β; see, e.g., [57, 14, 58].

5 Particle Implementation of SrMMD Descent

In practice, the SrMMD descent scheme in Eq. (7) is implemented by a system of interacting particles [66, Section 5]. Let $\{x_0^{(i)}\}_{i=1}^N$ be N i.i.d samples from the initial distribution μ_0 on $\mathbb{R}^d$. For a learning rate $\gamma > 0$, and a fixed time horizon $S \in \mathbb{N}$, the SrMMD particle descent is defined as:

$$x_{s+1}^{(i)} = x_s^{(i)} - \gamma \nabla f_{\hat{\mu}_s, \pi}(x_s^{(i)}), \quad \forall s \in [S]. \tag{8}$$

where $\hat{\mu}_s = \frac{1}{N}\sum_{i=1}^N \delta_{x_s^{(i)}}$ denotes the empirical distribution of the particles at time step s. Proposition 5.1 below shows that the witness function $f_{\hat{\mu}_n,\pi}$ admits a tractable closed-form expression.

Proposition 5.1. *Let $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N \delta_{x^{(i)}}$ be an empirical distribution. Define $D_{XX} \in \mathbb{R}^{(Nd)\times N}$ with $[D_{XX}]_{(i,l),j} = \partial_{1,l}k(x^{(i)}, x^{(j)})$, and $H_{XX} \in \mathbb{R}^{(Nd)\times(Nd)}$ with $[H_{XX}]_{(i,l),(j,m)} = \partial_{1,l}\partial_{2,m}k(x^{(i)}, x^{(j)})$.[2] Also, denote $\mathbb{K}_X : \mathbb{R}^d \to \mathbb{R}^N$ and $\mathbb{D}_X : \mathbb{R}^d \to \mathbb{R}^{Nd}$:*

$$\mathbb{K}_X(z) = [k(x^{(1)}, z), \ldots, k(x^{(N)}, z)]^\top, \quad \mathbb{D}_X(z) = [\partial_{1,1}k(x^{(1)}, z), \ldots, \partial_{1,d}k(x^{(N)}, z)]^\top.$$

Let $\mathbf{1}_N \in \mathbb{R}^N$ be a vector of all 1s. Then, the witness function $f_{\hat{\mu},\pi} \in \mathcal{H}$ is given by

$$f_{\hat{\mu},\pi}(z) = \frac{\mathbf{1}_N^\top \mathbb{K}_X(z)}{N\lambda} - \frac{\mathbb{E}_\pi[k(Y,z)]}{\lambda} - \mathbb{D}_X(z)^\top (H_{XX} + N\lambda \mathrm{Id})^{-1}\left(\frac{D_{XX}\mathbf{1}_N}{N\lambda} - \frac{\mathbb{E}_\pi[\mathbb{D}_X(Y)]}{\lambda}\right).$$

Proof is provided in Appendix A.5. In the context of generative modeling when π is known up to M i.i.d samples $\{y^{(j)}\}_{j=1}^M$, the above expectation with respect to π is replaced by an empirical average of these M samples. In the context of sampling when π is known up to its normalization constant, the above expectation with respect to π vanishes under a Stein kernel k_π. The additional gradient $\nabla f_{\hat{\mu},\pi}$ required in Eq. (8) can be obtained with automatic differentiation libraries such as JAX [13].

Next, we analyze finite-particle convergence of SrMMD in Eq. (8), under the following assumption.

Assumption 4. We impose the following conditions.

1. *Global boundedness.* There exists a constant $B > 0$ such that, for all $i, j \in [d]$, $\sup_{x\in\mathbb{R}^d} \|\partial_{1,i}\partial_{1,j}k(x,\cdot)\|_{\mathcal{H}} \le B$, where the derivatives are taken in $\mathcal{H}$.
2. *Bounded fourth moments.* $\{\mu_s\}_{s\ge 0}$ defined in Eq. (7) satisfies $\sup_{s\ge 0}\int \|x\|^4 \,\mathrm{d}\mu_s(x) < \infty$.

Theorem 5.2 (Convergence of finite-particle implementation of SrMMD)**.** *Let $\mu_0, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, let $\{x_0^{(i)}\}_{i=1}^N$ be i.i.d. with μ_0, let $S \in \mathbb{N}$ be a fixed time horizon. Suppose there exists $0 < \alpha \le 1$ such that $m_{\mu_s} - m_\pi = S_{\mu_s}^{\alpha/2} q_s$ with $q_s \in \mathcal{H}$ for any $s \in [S]$, and there exists a constant $Q < \infty$ such that $\sup_{1\le s\le S}\|q_s\|_{\mathcal{H}}^2 \le Q$. Fix $\lambda > 0$ and $0 < \gamma \le \frac{1}{2}$ such that $4dB\gamma(\mathrm{MMD}(\mu_0,\pi) + 1) \le \lambda$. Then there exists a constant C depending only on $d, B, \gamma, \lambda, \mathrm{MMD}(\mu_0,\pi), \alpha, Q$, the kernel k, and the uniform fourth-moment bound in Assumption 4, such that for each $\delta \in (0,1)$ and $N \ge C/\delta$, with probability at least $1-\delta$,*

$$\mathrm{MMD}(\hat{\mu}_s, \pi) \le \mathrm{MMD}(\mu_s, \pi) + C[\log(N\delta)]^{-\alpha/(2-\alpha)}, \quad \forall\, 0 \le s \le S.$$

The proof is given in Appendix C. To obtain the high-probability bound on $\mathrm{MMD}(\hat{\mu}_s, \pi)$ holds simultaneously for all $0 \le s \le S$, we truncate particle and population updates up to a chosen iteration, then use the monotonicity of $\mathrm{MMD}(\hat{\mu}_s, \pi)$ and the decay of $\mathrm{MMD}(\mu_s, \pi)$ to control later iterations. Combining Theorem 5.2 and Theorem 4.3 provides an upper bound on $\mathrm{MMD}(\hat{\mu}_s, \pi)$ in terms of iteration S and particle number N. Compared with closely related prior work, the finite-particle bounds for both MMD [4] and DrMMD [18] flows are obtained via propagation-of-chaos arguments in terms of the expected W_2 distance. These bounds achieve an $\mathcal{O}(1/\sqrt{N})$ rate with respect to the particle number N, but grow exponentially with the number of iterations. [8] studied the finite-particle convergence of SVGD and obtained an $\mathcal{O}(1/\sqrt{N})$ convergence rate, improving upon the previous $\mathcal{O}(1/\sqrt{\log\log N})$ rate established in [74]. More recently, [6] studied the finite-particle convergence of the drifting model proposed in [28], obtaining a conditional root residual-velocity rate of $\mathcal{O}(N^{-1/(d+4)})$ under uniform regularity assumptions. In contrast, our result provides a uniform-in-iteration bound, albeit at the cost of a logarithmic convergence rate with respect to N.

6 Numerical Experiments

In this section, we conduct extensive experiments to evaluate the empirical performance of SrMMD in the context of both generative modelling (π is known from i.i.d. samples) and sampling (π is known up to a normalization constant). The code is available at `https://github.com/FerdTian/SrMMD`.

[2] Strictly speaking, D_{XX} and H_{XX} here are higher-order tensors. To simplify notation, we implicitly identify them with their vectorized matrix representations, so that all subsequent computations can be written using standard matrix operations.

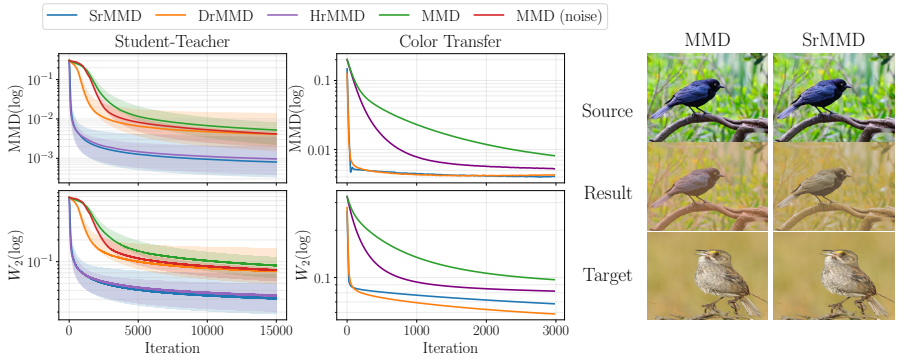

Figure 2: **Left:** Student-teacher. **Middle & Right:** Color transfer. Results are aggregated over 10 random seeds; solid curves show the median, and shaded regions show the 25%–75% percentile.

Toy examples. We consider two synthetic settings in which 1) the target distribution π is a mixture of two-dimensional Gaussian distributions, and 2) π has the shape of a Swiss roll. The Gaussian mixture provides a standard multimodal benchmark for assessing whether a sampling method can recover all modes of the target, whereas the Swiss roll is a simple example of a distribution supported on a manifold. We compare the SrMMD flow with the vanilla MMD flow [4], the DrMMD flow [18], and HrMMD flow [65]—[65] refers to HrMMD objective as kernel Sobolev-Fisher discrepancy—, a hybrid of DrMMD and SrMMD that penalizes both the L_2 norm and the gradient norm of the witness function; see Appendix D.6 for details. For the Gaussian mixture experiments, we use the Gaussian kernel $k(x, y) = \exp(-(2\sigma^2)^{-1}\|x - y\|^2)$ with length scale $\sigma > 0$, so that the kernel and the target π satisfy all Assumptions 1 to 3. For the Swiss roll, we use the Riesz kernel $k(x, x') = -\|x - x'\|$; this choice is intentional and is meant to test whether the SrMMD flow continues to perform well even when the conditions on kernels are not perfectly met.

For all flows considered here, we use regularization parameter $\lambda = 0.1$ and step size $\gamma = 0.1$, except for HrMMD where $\gamma = 1.0$. An ablation study on different values of λ and σ is given in Appendix D.1. All flows are simulated for $4{,}000$ iterations when convergence is empirically observed. From Figure 1, SrMMD flow substantially outperforms the vanilla MMD flow. This is consistent with the theory: MMD flow has only polynomial convergence under restrictive conditions [4], whereas SrMMD flow enjoys exponential convergence; see Theorems 4.1 and 4.3. The performance of SrMMD flow is comparable to that of DrMMD flow, suggesting that either L_2 or gradient regularization improves convergence. Finally, the hybrid HrMMD flow achieves the best performance on Gaussian mixtures, which is expected because it finds a best combination of both the regularization schemes.

Student-Teacher network. Next, we consider mean-field student-teacher networks, a more challenging and high-dimensional benchmark for MMD flows [4, 18]. The teacher network is $\Psi_\pi^{\mathsf{T}}(z) = \int \psi(z, \theta)\, \mathrm{d}\pi(\theta)$, where π is the target teacher distribution, and the student network is $\Psi_\mu^{\mathsf{S}}(z) = \int \psi(z, \theta)\, \mathrm{d}\mu(\theta)$, where μ is the student distribution. Training the student network amounts to the optimization problem: $\min_\mu \mathbb{E}_{z\sim\mathbb{P}_{\mathrm{data}}}[\|\Psi_\pi^{\mathsf{T}}(z) - \Psi_\mu^{\mathsf{S}}(z)\|^2]$, where the data distribution $\mathbb{P}_{\mathrm{data}}$ is a uniform spherical distribution on $\mathbb{R}^{50}$. This is equivalent to MMD flow under a new kernel, given by $k(\theta, \theta') = \mathbb{E}_{z\sim\mathbb{P}_{\mathrm{data}}}[\psi(z, \theta)^\top \psi(z, \theta')]$, which is bounded and hence satisfies all Assumptions 1 to 3. See Appendix D.3 for details.

We compare SrMMD flow with the vanilla MMD flow (both with and without noise injection), DrMMD flow, and the hybrid version, HrMMD flow. All of them are simulated for $15,000$ iterations on 1000 training samples and then evaluated on a held-out 1000 validation samples. We set the regularization parameter to $\lambda = 0.1$ and step size $\gamma = 0.1$ for all the flows. An ablation study on different values of λ is given in Appendix D.3. In the left panel of Figure 2, both SrMMD and HrMMD flows achieve faster convergence and better final performance.

Image color transfer. We consider an image color transfer task following [66], where the goal is to modify a source image A so that its color distribution matches that of a target image B. The task can be formulated as the alignment of the empirical color distributions of the source and target image in the RGB space $[0, 1]^3$. To this end, we sample N color particles from the source image and also N color particles from the target image. We then evolve the source particles toward the target

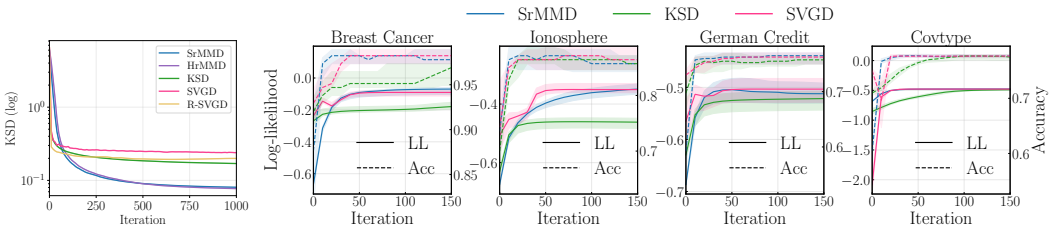

Figure 3: **Left:** Sampling from Gaussian mixtures. **Rest:** Bayesian logistic regression. Solid curves show the median, and shaded regions show the 25%–75% percentile.

particles using our SrMMD flow, and compare it against the vanilla MMD flow, DrMMD flow, and HrMMD flow. After obtaining the transported particles Y_T, we reconstruct the recolored image by nearest-neighbor assignment; see Appendix D.4 for additional details. As shown in the middle panel of Figure 2, SrMMD, DrMMD, and HrMMD all exhibit faster decay in both MMD and W_2 distance than the vanilla MMD flow, demonstrating the positive effect of regularization on convergence. This improvement is also visible in the final recolored images; see result figures in the right panel of Figure 2, where SrMMD more faithfully transfers the target color to the source image.

Sampling. While the previous experiments concern the ability of SrMMD flow as a generative model, here we evaluate the performance of SrMMD as a sampling method, under a Stein kernel k_π. The considered baselines are KSD flow [56] and Stein variational gradient descent (SVGD) [60]. DrMMD flow is not applicable here since its regularization requires access to the unknown $\|f\|_{L_2(\pi)}$.

First, we consider sampling from a two-dimensional mixture of 10 Gaussian distributions, a challenging multi-modal setting. The Stein kernel here is constructed from a base Gaussian kernel $k(x, x') = \exp(-(2\sigma^2)^{-1}\|x - x'\|^2)$ with bandwidth $\sigma = 0.3$. In the left panel of Figure 3, we perform SrMMD flow with $N = 500$ source particles. More experiments with other values of N can be found in Appendix D.5. Beyond KSD flow and SVGD, we also compare against regularized SVGD (R-SVGD) [46], a spectral regularized version of SVGD. For SrMMD, HrMMD, and KSD flows, we use the best tuned step sizes $\gamma = 0.01$ for KSD and $\gamma = 0.1$ for SrMMD and HrMMD. For SVGD and R-SVGD, we use the best-tuned step size $\gamma = 0.01$. The regularization parameters are $\lambda = 0.5$ for SrMMD and HrMMD, and $\lambda = 0.01$ for R-SVGD, each selected from a pre-specified grid according to final performance. The left panel of Figure 3 shows that SrMMD attains the fastest KSD decay and the lowest final KSD value.

Next, we consider the task of Bayesian logistic regression, following [60, 56]. Given a dataset $\mathscr{D}$ with a Bernoulli likelihood, the target distribution π is the posterior $\pi(x) \propto p(\mathscr{D} \mid x)\, p(x)$, which is available up to an intractable normalizing constant. This benchmark is considerably more challenging than the Gaussian mixture example above, since the resulting posterior distributions are higher-dimensional, with dimensions corresponding to BREAST CANCER ($d = 30$), IONOSPHERE ($d = 34$), GERMAN CREDIT ($d = 24$), and COVTYPE ($d = 54$). We simulate SrMMD flow, KSD flow, and SVGD, each with $N = 20$ particles, and report the evolution of test accuracy and test log-likelihood. From Figure 3, we see that SrMMD flow significantly improves upon the original KSD flow, while matching the performance of SVGD across all tasks. Since KSD flow is known to underperform relative to SVGD [56], the Sobolev regularization provides a promising alternative.

7 Discussion

We analyze SrMMD flow (a.k.a. Sobolev Descent), a flow with a gradient penalty to the MMD/KSD witness function for globally convergent continuous- and discrete-time schemes, applicable for both generative modelling and sampling. A primary limitation of SrMMD flow is its computational cost: each iteration requires inverting an $Nd \times Nd$ matrix, leading to a cost of $\mathcal{O}(N^3 d^3)$. Nyström approximations [85], random features [70], and caching of vector field [37] are standard approaches for reducing such computational burdens. Since the present paper is primarily theoretical, our empirical evaluation of SrMMD as a sampling method is restricted to small-scale benchmarks, where it achieves performance comparable to SVGD. A natural next step is to assess the performance of SrMMD in larger-scale tasks where SVGD has proved effective, e.g., earthquake modelling [26].

References

[1] F. Altekrüger, J. Hertrich, and G. Steidl. Neural Wasserstein gradient flows for discrepancies with Riesz kernels. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning*, pages 664–690. PMLR, PMLR, 2023.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

[3] M. Arbel, D. J. Sutherland, M. Bińkowski, and A. Gretton. On gradient regularizers for MMD GANs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[4] M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 103. Springer, 2014.

[6] K. Balasubramanian. Finite-particle convergence rates for conservative and non-conservative drifting models. *arXiv preprint arXiv:2605.22795*, 2026.

[7] K. Balasubramanian, S. Chewi, M. A. Erdogdu, A. Salim, and S. Zhang. Towards a theory of non-log-concave sampling: First-order stationarity guarantees for Langevin Monte Carlo. In P.-L. Loh and M. Raginsky, editors, *Conference on Learning Theory*, pages 2896–2923. PMLR, PMLR, 2022.

[8] S. Banerjee, K. Balasubramanian, and P. Ghosal. Improved finite-particle convergence rates for Stein variational gradient descent. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=sbG8qhMjkZ`.

[9] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

[10] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1lUOzWCW`.

[11] B. Bonnet and H. Frankowska. Differential inclusions in Wasserstein spaces: the Cauchy-Lipschitz framework. *Journal of Differential Equations*, 271:594–637, 2021.

[12] S. Boufadène and F.-X. Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. *SIAM Journal on Mathematical Analysis*, 57(4):4556–4587, 2025. doi: 10.1137/24M1647631. URL `https://doi.org/10.1137/24M1647631`.

[13] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

[14] H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, volume 202, pages 4735–4763. PMLR, 2023.

[15] Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2010.

[16] Z. Chen, M. Naslidnyk, A. Gretton, and F.-X. Briol. Conditional Bayesian quadrature. In N. Kiyavash and J. M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 648–684. PMLR, PMLR, 2024.

[17] Z. Chen, T. Karvonen, H. Kanagawa, F.-X. Briol, and C. Oates. Stationary MMD points for cubature. *https://arxiv.org/abs/2505.20754*, 2025.

[18] Z. Chen, A. Mustafi, P. Glaser, A. Korba, A. Gretton, and B. K. Sriperumbudur. (De)-regularized maximum mean discrepancy gradient flow. *Journal of Machine Learning Research*, 26(235): 1–77, 2025.

[19] Z. Chen, M. Naslidnyk, and F.-X. Briol. Nested expectations with kernel quadrature. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 8760–8793. PMLR, 2025.

[20] X. Cheng and P. Bartlett. Convergence of langevin mcmc in kl-divergence. In F. Janoos, M. Mohri, and K. Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 186–211. PMLR, 07–09 Apr 2018. URL `https://proceedings.mlr.press/v83/cheng18a.html`.

[21] S. Chewi. Log-concave sampling. `https://chewisinho.github.io/`, 2024. Book draft.

[22] S. Chewi, T. Le Gouic, C. Lu, T. Maunu, and P. Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2098–2109. Curran Associates, Inc., 2020.

[23] S. Chewi, M. A. Erdogdu, M. Li, R. Shen, and M. S. Zhang. Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev. *Foundations of Computational Mathematics*, pages 1–51, 2024.

[24] L. Chizat, M. Colombo, R. Colombo, and X. Fernández-Real. Quantitative convergence of Wasserstein gradient flows of kernel mean discrepancies. *arXiv preprint arXiv:2603.01977*, 2026.

[25] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In M. F. Balcan and K. Q. Weinberger, editors, *International Conference on Machine Learning*, pages 2606–2615. PMLR, PMLR, 2016.

[26] M. Corrales, S. Berti, B. Denel, P. Williamson, M. Aleardi, and M. Ravasi. Annealed Stein variational gradient descent for improved uncertainty estimation in full-waveform inversion. *Geophysical Journal International*, 241(2):1088–1113, 2025.

[27] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

[28] M. Deng, H. Li, T. Li, Y. Du, and K. He. Generative modeling via drifting, 2026. URL `https://arxiv.org/abs/2602.04770`.

[29] R. G. Douglas. On majorization, factorization, and range inclusion of operators on Hilbert space. *Proceedings of the American Mathematical Society*, 17(2):413–415, 1966.

[30] A. Duncan, N. Nüsken, and L. Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39, 2023.

[31] R. Duong, N. Rux, V. Stein, and G. Steidl. Wasserstein gradient flows of maximum mean discrepancy functionals with distance kernels under Sobolev regularization. *Philosophical Transactions of the Royal Society A*, 383(2298):20240243, 2025. doi: 10.1098/rsta.2024.0243.

[32] R. Duong, V. Stein, R. Beinert, J. Hertrich, and G. Steidl. Wasserstein gradient flows of MMD functionals with distance kernel and Cauchy problems on quantile functions. *ESAIM: Control, Optimisation and Calculus of Variations*, 32:10, 2026.

[33] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media, 1996.

[34] L. Evans and R. Gariepy. *Measure Theory and Fine Properties of Functions, Revised Edition*. CRC Press, 2015.

[35] C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization. In H. D. III and A. Singh, editors, *International Conference on Machine Learning*, pages 3154–3164. PMLR, PMLR, 2020.

[36] S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

[37] A. Galashov, V. D. Bortoli, and A. Gretton. Deep MMD gradient flow without adversarial training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pf85K2wtz8.

[38] J. Gorham and L. Mackey. Measuring sample quality with kernels. In D. Precup and Y. W. Teh, editors, *International Conference on Machine Learning*, pages 1292–1301. PMLR, PMLR, 2017.

[39] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[40] H. Gu, P. Birmpa, Y. Pantazis, L. Rey-Bellet, and M. A. Katsoulakis. Lipschitz-regularized gradient flows and generative particle algorithms for high-dimensional scarce data. *SIAM Journal on Mathematics of Data Science*, 6(4):1205–1235, 2024.

[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In U. von Luxburg, S. Bengio, R. Fergus, R. Garnett, I. Guyon, H. Wallach, and S. V. N. Vishwanathan, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[42] P. Hagemann, J. Hertrich, F. Altekrüger, R. Beinert, J. Chemseddine, and G. Steidl. Posterior sampling based on gradient flows of the MMD with negative distance kernel. In *The Twelfth International Conference on Learning Representations*, 2024.

[43] O. Hagrass, B. K. Sriperumbudur, and B. Li. Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101, 2024.

[44] O. Hagrass, B. K. Sriperumbudur, and B. Li. Spectral regularized kernel goodness-of-fit tests. *Journal of Machine Learning Research*, 25(309):1–52, 2024.

[45] O. Hagrass, B. Sriperumbudur, and K. Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *Bernoulli*, 32(1):299–324, 2026.

[46] Y. He, K. Balasubramanian, B. K. Sriperumbudur, and J. Lu. Regularized Stein variational gradient flow. *Foundations of Computational Mathematics*, pages 1–59, 2024.

[47] J. Hertrich, M. Gräf, R. Beinert, and G. Steidl. Wasserstein steepest descent flows of discrepancies with Riesz kernels. *Journal of Mathematical Analysis and Applications*, 531(1):127829, 2024.

[48] J. Hertrich, C. Wald, F. Altekrüger, and P. Hagemann. Generative sliced MMD flows with Riesz kernels. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VdkGRV1vcf.

[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In U. von Luxburg, S. Bengio, R. Fergus, R. Garnett, I. Guyon, H. Wallach, and S. V. N. Vishwanathan, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[50] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[51] X. Huang, H. Dong, Y. Hao, Y.-A. Ma, and T. Zhang. Reverse diffusion Monte Carlo. In *International Conference on Learning Representations*, 2024.

[52] H. Kanagawa, A. Barp, A. Gretton, and L. Mackey. Controlling moments with kernel Stein discrepancies. *The Annals of Applied Probability*, 35(6):3818–3843, 2025.

[53] M. Kelly, R. Longjohn, and K. Nottingham. The UCI Machine Learning Repository. https://archive.ics.uci.edu, 2024.

[54] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38: 3660–3695, 2010. URL `https://api.semanticscholar.org/CorpusID:17525518`.

[55] A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton. A non-asymptotic analysis for Stein variational gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4672–4682. Curran Associates, Inc., 2020.

[56] A. Korba, P.-C. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel Stein discrepancy descent. In M. Meila and T. Zhang, editors, *International Conference on Machine Learning*, pages 5719–5730. PMLR, PMLR, 2021.

[57] H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *Algorithmic Learning Theory*, volume 201, pages 946–985. PMLR, 2023.

[58] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *International Conference on Learning Representations*, 2024.

[59] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[60] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[61] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In M. F. Balcan and K. Q. Weinberger, editors, *International Conference on Machine Learning*, pages 276–284. PMLR, PMLR, 2016.

[62] G. Menz and A. Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884, 2014.

[63] D. Meunier, Z. Shen, M. Mollenhauer, A. Gretton, and Z. Li. Optimal rates for vector-valued spectral regularization learning algorithms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 82514–82559. Curran Associates, Inc., 2024.

[64] A. Mousavi-Hosseini, T. K. Farghly, Y. He, K. Balasubramanian, and M. A. Erdogdu. Towards a complete analysis of Langevin Monte Carlo: Beyond Poincaré inequality. In G. Neu and L. Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory*, pages 1–35. PMLR, PMLR, 2023.

[65] Y. Mroueh and M. Rigotti. Unbalanced Sobolev descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17034–17043. Curran Associates, Inc., 2020.

[66] Y. Mroueh, T. Sercu, and A. Raj. Sobolev descent. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2976–2985. PMLR, PMLR, 2019.

[67] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Scholköpf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2): 1–141, 2017.

[68] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[69] A. Neubauer. On converse and saturation results for Tikhonov regularization of linear ill-posed problems. *SIAM Journal on Numerical Analysis*, 34(2):517–527, 1997.

[70] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

[71] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. Stabilizing training of generative adversarial networks through regularization. In U. von Luxburg, S. Bengio, R. Fergus, R. Garnett, I. Guyon, H. Wallach, and S. V. N. Vishwanathan, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[72] F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: An overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

[73] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[74] J. Shi and L. Mackey. A finite-particle convergence rate for Stein variational gradient descent. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2024.

[75] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[76] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[77] V. Stein, S. Neumayer, N. Rux, and G. Steidl. Wasserstein gradient flows for Moreau envelopes of f-divergences in reproducing kernel Hilbert spaces. *Analysis and Applications*, pages 1–45, 2025.

[78] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[79] G. Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140. American Mathematical Soc., 2012.

[80] A. Vacher, O. Chehab, and A. Korba. Sampling from multi-modal distributions with polynomial query complexity in fixed dimension via reverse diffusion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

[81] S. Vempala and A. Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[82] C. Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

[83] H. Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

[84] A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Conference on Learning Theory*, pages 2093–3027. PMLR, PMLR, 2018.

[85] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 13, 2000.

Supplementary Material

A Proof of Main Results

Additional notations Denote a d-fold tensor product Hilbert space $\mathcal{H}^d = \mathcal{H} \otimes \ldots \otimes \mathcal{H}$ and we say $\mathbf{f} \in \mathcal{H}^d$ if $\mathbf{f} = [f_1, \cdots, f_d]^\top$ such that $f_i \in \mathcal{H}$ for all $i \in [d]$. Denote $(L_2^d(\mu), \|\cdot\|_{L_2^d(\mu)})$ as the space of all vector valued μ-square integrable measurable function with $\|f\|^2_{L_2^d(\mu)} := \int_{\mathbb{R}^d} \|f(x)\|^2 \mathrm{d}\mu(x)$. Define $\mathcal{L}^d$ as the Lebesgue measure on $\mathbb{R}^d$. $\mu \ll \mathcal{L}^d$ means that a measure μ is absolutely continuous with respect to the Lebesgue measure.

A.1 Proof of Proposition 3.3

By definition of S_μ, there is $\langle f, S_\mu f\rangle_\mathcal{H} = \|\nabla f\|^2_{L_2^d(\mu)}$. By a direct calculation,

$$\begin{aligned}
\mathrm{SrMMD}(\mu, \pi) &= \sup_{f\in\mathcal{H}:\ \langle f,(S_\mu+\lambda\mathrm{Id})f\rangle_\mathcal{H}\le 1} \langle f, m_\mu - m_\pi\rangle_\mathcal{H} \\
&= \sup_{f\in\mathcal{H}:\ \langle f,(S_\mu+\lambda\mathrm{Id})f\rangle_\mathcal{H}\le 1} \left\langle (S_\mu+\lambda\mathrm{Id})^{-1/2}(S_\mu+\lambda\mathrm{Id})^{1/2} f, m_\mu - m_\pi \right\rangle_\mathcal{H} \\
&= \sup_{f\in\mathcal{H}:\ \langle f,(S_\mu+\lambda\mathrm{Id})f\rangle_\mathcal{H}\le 1} \left\langle (S_\mu+\lambda\mathrm{Id})^{1/2} f, (S_\mu+\lambda\mathrm{Id})^{-1/2}(m_\mu - m_\pi) \right\rangle_\mathcal{H} \\
&\le \sup_{f\in\mathcal{H}:\ \langle f,(S_\mu+\lambda\mathrm{Id})f\rangle_\mathcal{H}\le 1} \|(S_\mu+\lambda\mathrm{Id})^{1/2} f\|_\mathcal{H} \cdot \|(S_\mu+\lambda\mathrm{Id})^{-1/2}(m_\mu - m_\pi)\|_\mathcal{H} \\
&= \|(S_\mu+\lambda\mathrm{Id})^{-1/2}(m_\mu - m_\pi)\|^2_\mathcal{H}.
\end{aligned}$$

The maximum is admitted at

$$f = \bar{f}_{\mu,\pi} = \left(\|(S_\mu+\lambda\mathrm{Id})^{-1/2}(m_\mu - m_\pi)\|_\mathcal{H}\right)^{-1} (S_\mu+\lambda\mathrm{Id})^{-1}(m_\mu - m_\pi).$$

For reasons of presentation and computational convenience, we instead work in the paper with the unnormalized witness

$$f_{\mu,\pi} = (S_\mu+\lambda\mathrm{Id})^{-1}(m_\mu - m_\pi).$$

This is without loss for our purposes, since SrMMD itself is not used elsewhere in the paper.

A.2 Proof of Theorem 3.4

Proof. The proof of the first claim is a direct application of [2, Theorem 8.3.1], where we only need to show that $\int_0^T \|\nabla f_{\mu_t,\pi}\|_{L_2^d(\mu_t)} \mathrm{d}t < \infty$. From Eq. (A.4), we have that $\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{MMD}^2(\mu_t, \pi) < 0$. This indicates that $\mathrm{MMD}^2(\mu_t, \pi)$ is monotonically decreasing in time along the trajectory. Hence, since $\mathrm{MMD}(\mu_0, \pi) < C_\mathrm{MMD}$, there is $\sup_{t\in[0,\infty)} \mathrm{MMD}(\mu_t, \pi) < C_\mathrm{MMD}$.

By definition of $f_{\mu_t,\pi}$ at any $t \in [0, T]$, we have

$$\begin{aligned}
\|\nabla f_{\mu_t,\pi}\|^2_{L_2^d(\mu_t)} &= \left\langle S_{\mu_t}(S_{\mu_t}+\lambda\mathrm{Id})^{-1}(m_{\mu_t} - m_\pi), (S_{\mu_t}+\lambda\mathrm{Id})^{-1}(m_{\mu_t} - m_\pi)\right\rangle_\mathcal{H} \\
&\le \lambda^{-1} \|m_{\mu_t} - m_\pi\|^2_\mathcal{H} = \lambda^{-1}\,\mathrm{MMD}(\mu_t, \pi)^2 \le \frac{C^2_\mathrm{MMD}}{\lambda}.
\end{aligned}$$

Therefore, we have proved that $\int_0^T \|\nabla f_{\mu_t,\pi}\|_{L_2^d(\mu_t)} \mathrm{d}t \le \frac{T C_\mathrm{MMD}}{\sqrt{\lambda}}$. By the converse part of [2, Theorem 8.3.1], $(\mu_t)_{t\in[0,T]}$ is an absolutely continuous curve with respect to W_2 in $\mathcal{P}_2(\mathbb{R}^d)$.

Next, to show uniqueness, we resort to [11]. To this end, we need to verify that hypothesis D1 in [11] is satisfied. Let $\mathcal{K} = B(0, R)$ be a ball of radius R in $\mathbb{R}^d$. Take any sequence $x_n \to x$ in $\mathcal{K}$. Then, for any $t \in [0, T]$,

$$\|\nabla f_{\mu_t,\pi}(x_n) - \nabla f_{\mu_t,\pi}(x)\| \le \|f_{\mu_t,\pi}\|_\mathcal{H} \cdot \|\nabla_1 k(x, \cdot) - \nabla_1 k(x_n, \cdot)\|_{\mathcal{H}^d} \to 0.$$

The last step holds from $\|f_{\mu_t,\pi}\|_{\mathcal{H}} \leq \frac{\|m_{\mu_t}-m_\pi\|_{\mathcal{H}}}{\lambda} < \infty$, and from Lemma B.7. Hence, D1(i) is satisfied. D1(iii) and D1(iv) are proved in Lemma B.6. Finally, D1(ii) is satisfied by the following derivations:

$$\begin{aligned}\|\nabla f_{\mu_t,\pi}(x)\| &= \left\|\nabla(S_{\mu_t}+\lambda\mathrm{Id})^{-1}(m_{\mu_t}-m_\pi)(x)\right\| \\ &\leq \sum_{j=1}^{d}\left|\left\langle \partial_{1,j}k(x,\cdot),\ (S_{\mu_t}+\lambda\mathrm{Id})^{-1}(m_{\mu_t}-m_\pi)\right\rangle_{\mathcal{H}}\right| \\ &\leq \sum_{j=1}^{d}\|\partial_{1,j}k(x,\cdot)\|_{\mathcal{H}}\cdot\frac{\|m_{\mu_t}-m_\pi\|_{\mathcal{H}}}{\lambda} \leq d\sqrt{M}(\|x\|+1)\cdot\frac{C_{\mathrm{MMD}}}{\lambda}. \end{aligned} \tag{A.1}$$

The last step holds by the upper bound on $|\partial_{1,j}\partial_{2,j}k(x,x)|$ in Assumption 1, as well as the upper bound on $\mathrm{MMD}(\mu_t,\pi)$.

Hence, by Theorem 4 of [11], there exists a trajectory solving the continuity inclusion on any finite time interval $[0,T]$. In the SrMMD flow setting, once existence is established, uniqueness follows separately from the Cauchy–Lipschitz theory for continuity equations, for instance from Theorem 2 of [11], since the velocity field satisfies the corresponding growth and local Lipschitz conditions as proved above. As the argument applies to every finite interval $[0,T]$, the trajectory cannot explode in finite time.

Finally, we are going to show that μ_t is absolutely continuous with respect to Lebesgue for all t. For any compact domain $B \subset \mathbb{R}^d$, since the vector field $\nabla f_{\mu_t,\pi}$ is locally Lipschitz and locally bounded as proved above, by [2, Lemma 8.1.4], the following ordinary differential equation admits a unique maximal solution in $[0,T]$.

$$\mathbf{T}(0,x)=x, \quad \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{T}(t,x)=\nabla f_{\mu_t,\pi}\left(\mathbf{T}(t,x)\right),$$

Moreover, the SrMMD flow iterate at time t has $\mu_t = (\mathbf{T}(t,\cdot))_{\#}\,\mu_0$. Under the local Lipschitz and linear-growth assumptions, the mapping $\mathbf{T}(t,\cdot)$ is locally bi-Lipschitz on bounded sets, as proved in Lemma B.9; hence $\mathbf{T}(t,\cdot)^{-1}$ maps Lebesgue-null sets to Lebesgue-null sets. Therefore, if μ_0 is absolutely continuous with respect to Lebesgue measure $\mathcal{L}^d$ on $\mathbb{R}^d$, then μ_t is also absolutely continuous with respect to Lebesgue measure $\mathcal{L}^d$ on $\mathbb{R}^d$.

□

A.3 Proof of Theorem 4.1

Proof. First, we consider the time-derivative, $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{MMD}^2(\mu_t,\pi)$. Let $(X_t)_{t\geq 0}$ be the stochastic process defined in Eq. (6) which satisfies $\mathrm{d}X_t = -\nabla f_{\mu_t,\pi}(X_t)\mathrm{d}t$. Then, consider any fixed test function $h \in \mathcal{H}$, differentiating along the characteristic and using the chain rule yields

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}m_{\mu_t}, h\right\rangle_{\mathcal{H}} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[h(X_t)] = \mathbb{E}\left[\nabla h(X_t)^\top\frac{\mathrm{d}X_t}{\mathrm{d}t}\right] = -\mathbb{E}\left[\nabla h(X_t)^\top\nabla f_{\mu_t,\pi}(X_t)\right].$$

The first equality holds from the continuity of the inner product in $\mathcal{H}$. The second equality holds by the following additional conditions on h (Lemma B.10):

$$\mathbb{E}\left[\int_0^T\left|\nabla h(X_t)^\top\frac{\mathrm{d}}{\mathrm{d}t}X_t\right|\mathrm{d}t\right] = \mathbb{E}\left[\int_0^T\left|\nabla h(X_t)^\top\nabla f_{\mu_t,\pi}(X_t)\right|\mathrm{d}t\right] < \infty. \tag{A.2}$$

Since $\mathrm{Law}(X_t)=\mu_t$, this becomes

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}m_{\mu_t}, h\right\rangle_{\mathcal{H}} = -\int\nabla h(x)^\top\nabla f_{\mu_t,\pi}(x)\,\mathrm{d}\mu_t(x). \tag{A.3}$$

Next, we will apply Eq. (A.3) with the fixed test function $h = m_{\mu_t}-m_\pi$ at time t. Since Eq. (A.3) holds for any test function in $\mathcal{H}$, it also holds for $h = m_{\mu_t}-m_\pi$. To see why $m_{\mu_t}-m_\pi$ satisfies the conditions outlined in Eq. (A.2), note that $\|\nabla(m_{\mu_t}-m_\pi)(x)\| = \|\boldsymbol{v}_{\mathrm{MMD}}[\mu_t](x)\| \leq dC_{\mathrm{MMD}}(\|x\|+1)$ proved in Eq. (B.1), and also $\|\nabla f_{\mu_t,\pi}(x)\| \leq d\sqrt{M}(\|x\|+1)\cdot\frac{C_{\mathrm{MMD}}}{\lambda}$ proved in Eq. (A.1).

Therefore, since $(\mu_t)_{t\in[0,T]}$ is an absolutely continuous curve in $\mathcal{P}_2(\mathbb{R}^d)$, we have verified Eq. (A.2). So we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{MMD}^2(\mu_t,\pi) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|m_{\mu_t}-m_\pi\|_{\mathcal{H}}^2 = \left\langle \frac{\mathrm{d}}{\mathrm{d}t}m_{\mu_t}, m_{\mu_t}-m_\pi\right\rangle_{\mathcal{H}}$$
$$= -\int \nabla(m_{\mu_t}-m_\pi)(x)^\top \nabla f_{\mu_t,\pi}(x)\,\mathrm{d}\mu_t(x).$$

Remark A.1 (No boundary condition). Therefore, the same dissipation identity as Appendix C.2 of [66] can be derived directly from the characteristic representation, without invoking integration by parts, and hereby avoiding a boundary condition.

By definition of the gradient covariance operator S_{μ_t} and by definition of $f_{\mu_t,\pi}$ in Proposition 3.3, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\,\mathrm{MMD}^2(\mu_t,\pi) = -\Big\langle \nabla(m_\mu - m_\pi), \nabla\left(S_{\mu_t}+\lambda\mathrm{Id}\right)^{-1}(m_\mu-m_\pi)\Big\rangle_{L_2^d(\mu_t)}$$
$$= -\Big\langle S_{\mu_t}(m_{\mu_t}-m_\pi), \left(S_{\mu_t}+\lambda\mathrm{Id}\right)^{-1}(m_{\mu_t}-m_\pi)\Big\rangle_{\mathcal{H}} < 0. \tag{A.4}$$

From Lemma B.4, the operator $S_{\mu_t}:\mathcal{H}\to\mathcal{H}$ is trace-class, Hilbert-Schmidt, and compact. Let $\{\varrho_{i,t}, e_{i,t}\}_{i\in\mathbb{N}}$ be the countable eigensystem of the operator S_{μ_t} and $\{e_{i,t}\}_{i\in\mathbb{N}}$ is an orthonormal system of $\mathcal{H}$.

Since $m_{\mu_t}-m_\pi \in \overline{\mathrm{Ran}(S_{\mu_t})}$, we may let $m_{\mu_t}-m_\pi = \sum_{i=0}^\infty m_i e_i$ with $\mathrm{MMD}^2(\mu_t,\pi) = \sum_{i=0}^\infty m_{i,t}^2$, which is finite by the proof of Theorem 3.4. Then, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\,\mathrm{MMD}^2(\mu_t,\pi) = -\sum_{i=0}^\infty \frac{\varrho_{i,t}}{\varrho_{i,t}+\lambda}m_{i,t}^2 = -\sum_{i=0}^\infty m_{i,t}^2 + \lambda\sum_{i=0}^\infty \frac{m_{i,t}^2}{\varrho_{i,t}+\lambda}$$
$$= -\,\mathrm{MMD}^2(\mu_t,\pi) + \lambda\sum_{i=0}^\infty \frac{m_{i,t}^2}{\varrho_{i,t}+\lambda} := -\,\mathrm{MMD}^2(\mu_t,\pi) + R(\lambda,t).$$

Since $|\frac{\lambda m_{i,t}^2}{\rho_{i,t}+\lambda}| \le m_{i,t}^2$, and $\sum_{i=0}^\infty m_{i,t}^2 < \infty$, by dominated convergence theorem (See, e.g., [34, Theorem 1.19]), we have $\lim_{\lambda\to 0+} R(\lambda,t) = \lim_{\lambda\to 0+}\sum_{i=0}^\infty \frac{\lambda m_{i,t}^2}{\varrho_{i,t}+\lambda} = 0$. Since this holds for all $t\in[0,T]$, we have $\sup_{t\in[0,T]}\lim_{\lambda\to 0+} R(\lambda,t) = 0$. Integrating the above differential equation from 0 to T yields

$$\mathrm{MMD}^2(\mu_T,\pi) = e^{-2T}\,\mathrm{MMD}^2(\mu_0,\pi) + 2e^{-2T}\int_0^T e^{2t}R(\lambda,t).$$

Under additional assumptions that $m_{\mu_t}-m_\pi \in \mathrm{Ran}\, S_{\mu_t}^{\alpha/2}$ for all $t\in[0,T]$, i.e., there exists $q_t\in\mathcal{H}$ such that $m_{\mu_t}-m_\pi = S_{\mu_t}^{\alpha/2}q_t$. Recall that $m_{\mu_t}-m_\pi = \sum_{i=0}^\infty m_{i,t}e_{i,t}$. Then, we can write the above equation as

$$R(\lambda,t) = \lambda\sum_{i=0}^\infty \frac{m_{i,t}^2}{\varrho_{i,t}+\lambda} = \lambda\sum_{i=0}^\infty\left(\frac{\varrho_{i,t}^\alpha}{\varrho_{i,t}+\lambda}\right)\varrho_{i,t}^{-\alpha}m_{i,t}^2 \le \lambda\lambda^{\alpha-1}\sum_{i=0}^\infty \varrho_{i,t}^{-\alpha}m_{i,t}^2 = \lambda^\alpha\|q_t\|_{\mathcal{H}}^2.$$

□

A.4 Proof of Theorem 4.3

Proof. We know that $\mu_{s+1} = (\mathrm{Id}-\gamma\nabla f_{\mu_s,\pi})_\#\mu_s$ where f_s is the witness function $f_{\mu_s,\pi} = (S_{\mu_s}+\lambda\mathrm{Id})^{-1}(m_{\mu_s}-m_\pi)$ and $\gamma>0$ is the step size. In the proof, we use $f_s := f_{\mu_s,\pi}$ to simplify notation. Consider $\rho_t = (\mathrm{I}-t\nabla f_s)_\#\mu_s$, which is an interpolation between the distributions within two time steps, such that $\rho_0=\mu_s$, and $\rho_\gamma = \mu_{s+1}$. First, we know that,

$$\mathrm{MMD}^2(\rho_t,\pi) = \iint k(x,z)\mathrm{d}\rho_t(x)\mathrm{d}\rho_t(z) - 2\iint k(x,z)\mathrm{d}\rho_t(x)\mathrm{d}\pi(z) + \text{const}$$
$$= \iint k(x-t\nabla f_{\mathfrak{s}}(x), z-t\nabla f_s(z))\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\mu_s(z) - 2\iint k(x-t\nabla f_{\mathfrak{s}}(x),z)\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\pi(z) + \text{const}.$$

Here, const represents the term that is independent of time t. So we have, for the first order derivative,

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\frac{1}{2}\,\mathrm{MMD}^2(\rho_t,\pi)=-\int\nabla f_{\mathfrak{s}}(x)^\top\left(\int\nabla_1 k(x,z)\mathrm{d}\mu_s(z)-\int\nabla_1 k(x,z)\mathrm{d}\pi(z)\right)d\mu_{\mathfrak{s}}(x)$$

$$=-\left\langle\nabla\left(m_{\mu_s}-m_\pi\right),\nabla\left(S_{\mu_s}+\lambda\mathrm{Id}\right)^{-1}\left(m_{\mu_s}-m_\pi\right)\right\rangle_{L_2^d(\mu_s)}$$

$$=-\left\langle S_{\mu_s}\left(m_{\mu_s}-m_\pi\right),\left(S_{\mu_s}+\lambda\mathrm{Id}\right)^{-1}\left(m_{\mu_s}-m_\pi\right)\right\rangle_{\mathcal{H}}.$$

This is the first-order term, which is the same as the derivations in the continuous-time analogue. Next, we define $\phi_t(x)=x-t\nabla f_{\mathfrak{s}}(x)$. We consider the second-order derivative,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{1}{2}\,\mathrm{MMD}^2(\rho_t,\pi)$$

$$=\underbrace{\left\|\int\nabla f_{\mathfrak{s}}(x)^\top\nabla_1 k(x-t\nabla f_{\mathfrak{s}}(x),\cdot)\mathrm{d}\mu_{\mathfrak{s}}(x)\right\|_{\mathcal{H}}^2}_{(i)}$$

$$+\underbrace{\int\nabla f_{\mathfrak{s}}(x)^\top\left(\int\mathbf{H}_1 k(x-t\nabla f_{\mathfrak{s}}(x),z)\mathrm{d}\rho_t(z)-\int\mathbf{H}_1 k(x-t\nabla f_{\mathfrak{s}}(x),z)\mathrm{d}\pi(z)\right)\nabla f_{\mathfrak{s}}(x)\mathrm{d}\mu_{\mathfrak{s}}(x)}_{(ii)}.$$

For the first term (i), we have

$$(i)=\iint\nabla f_{\mathfrak{s}}(x)^\top\nabla_1\nabla_2 k(\phi_t(x),\phi_t(y))\nabla f_{\mathfrak{s}}(y)\;\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\mu_{\mathfrak{s}}(y)$$

$$=\iint\left\langle\nabla f_{\mathfrak{s}}(x)\nabla f_{\mathfrak{s}}(y)^\top,\;\nabla_1\nabla_2 k(\phi_t(x),\phi_t(y))\right\rangle_F\;\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\mu_{\mathfrak{s}}(y)$$

$$\le\iint\left\|\nabla f_{\mathfrak{s}}(x)\nabla f_{\mathfrak{s}}(y)^\top\right\|_F\cdot\left\|\nabla_1\nabla_2 k(\phi_t(x),\phi_t(y))\right\|_F\;\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\mu_{\mathfrak{s}}(y)$$

$$\le\left(\iint\left\|\nabla f_{\mathfrak{s}}(x)\nabla f_{\mathfrak{s}}(y)^\top\right\|_F^2\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\mu_{\mathfrak{s}}(y)\right)^{\frac{1}{2}}\cdot\left(\iint\left\|\nabla_1\nabla_2 k(\phi_t(x),\phi_t(y))\right\|_F^2\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\mu_{\mathfrak{s}}(y)\right)^{\frac{1}{2}}$$

$$=\|\nabla f_s\|_{L_2^d(\mu_s)}^2\cdot\left(\iint\left\|\nabla_1\nabla_2 k(x,y)\right\|_F^2\mathrm{d}\rho_t(x)\mathrm{d}\rho_t(y)\right)^{\frac{1}{2}}.$$

Since for any $i,j\in[d]$,

$$\iint\left|\partial_{1,i}\partial_{2,j}k(x,y)\right|^2\mathrm{d}\rho_t(x)\mathrm{d}\rho_t(y)\le\iint\left|\partial_{1,i}\partial_{2,i}k(x,x)\right|\cdot\left|\partial_{1,j}\partial_{2,j}k(y,y)\right|\mathrm{d}\rho_t(x)\mathrm{d}\rho_t(y)$$

$$\le\left(\int M^2(1+\|x\|)^2\mathrm{d}\rho_t(x)\right)^{\frac{1}{2}}\cdot\left(\int M^2(1+\|y\|)^2\mathrm{d}\rho_t(y)\right)^{\frac{1}{2}}\le M^2 4C.$$

The second last inequality holds by the growth condition in Assumption 1, and the last inequality holds by the convexity of $\|\cdot\|^2$:

$$\int\|z\|^2\mathrm{d}\rho_t(z)\le(1-t\gamma^{-1})\int\|x\|^2\mathrm{d}\mu_{\mathfrak{s}}(x)+t\gamma^{-1}\int\|y\|^2\mathrm{d}\mu_{s+1}(y)\le C.$$

The last inequality holds by the uniform bound on the second moment assumed in the statement of the theorem. Therefore, we have proved that

$$(i)\le dM^2 4C\|\nabla f_s\|_{L_2^d(\mu_s)}^2.$$

For the second term (ii), we have

$$\iint\nabla f_{\mathfrak{s}}(x)^\top\mathbf{H}_1 k(x-\gamma\nabla f_{\mathfrak{s}}(x),z)\nabla f_{\mathfrak{s}}(x)\mathrm{d}\rho_t(z)\mathrm{d}\mu_{\mathfrak{s}}(x)$$

$$\le\iint\|\nabla f_{\mathfrak{s}}(x)\|^2\cdot\left\|\mathbf{H}_1 k(x-\gamma\nabla f_{\mathfrak{s}}(x),z)\right\|_{op}\;\mathrm{d}\mu_{\mathfrak{s}}(x)\mathrm{d}\rho_t(z)$$

$$\le\left(\int\|\nabla f_{\mathfrak{s}}(x)\|^2\;\mathrm{d}\mu_{\mathfrak{s}}(x)\right)\cdot\left(\int M(1+\|z\|^2)\rho_t(z)\right)$$

$$=\|\nabla f_s\|_{L_2^d(\mu_s)}^2(M+MC).$$

The second-to-last line holds under the growth conditions in Assumption 3. We can repeat the same derivations above to obtain an upper bound on the second term for (ii) as well. Combine the above upper bounds on (i) and (ii), then we can obtain

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{1}{2}\,\mathrm{MMD}^2(\rho_t,\pi) &\le d(4M^2C+2(M+MC))\|\nabla f_s\|^2_{L_2^d(\mu_s)} =: dC_0\|\nabla f_s\|^2_{L_2^d(\mu_s)}\\ &= dC_0\langle S_{\mu_s}(S_{\mu_s}+\lambda\mathrm{Id})^{-1}(m_{\mu_s}-m_\pi),(S_{\mu_s}+\lambda\mathrm{Id})^{-1}(m_{\mu_s}-m_\pi)\rangle_{\mathcal{H}}.\end{aligned}$$

The constant C_0 only depends on the other constants M, C in the assumptions. Putting the above together,

$$\begin{aligned}&\frac{1}{2}\,\mathrm{MMD}^2(\mu_{s+1},\pi)-\frac{1}{2}\,\mathrm{MMD}^2(\mu_s,\pi)\\ =\;&\frac{d}{dt}\Big|_{t=0}\frac{1}{2}\,\mathrm{MMD}^2(\rho_t,\pi)\,\gamma+\int_0^\gamma(\gamma-t)\frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{1}{2}\,\mathrm{MMD}^2(\rho_t,\pi)\,\mathrm{d}t\\ \le\;&-\gamma\left\langle S_{\mu_s}(m_{\mu_s}-m_\pi),(S_{\mu_s}+\lambda\mathrm{Id})^{-1}(m_{\mu_s}-m_\pi)\right\rangle_{\mathcal{H}}\\ &+\gamma^2dC_0\langle S_{\mu_s}(S_{\mu_s}+\lambda\mathrm{Id})^{-1}(m_{\mu_s}-m_\pi),(S_{\mu_s}+\lambda\mathrm{Id})^{-1}(m_{\mu_s}-m_\pi)\rangle_{\mathcal{H}}.\end{aligned}$$

Since $m_{\mu_s}-m_\pi\in\operatorname{Ran}S_{\mu_s}^{\alpha/2}$, there exists $q_s\in\mathcal{H}$ such that $m_{\mu_s}-m_\pi=S_{\mu_s}^{\alpha/2}q_s$. Recall that $m_{\mu_s}-m_\pi=\sum_{i=0}^\infty m_{i,s}e_{i,s}$, where $\{\varrho_{i,s},e_{i,s}\}_{i\in\mathbb{N}}$ is the eigensystem of the operator S_{μ_s}. Then, we can write the above equation as

$$\begin{aligned}&-\sum_{i=0}^\infty\frac{\gamma\varrho_{i,s}}{\varrho_{i,s}+\lambda}m_{i,s}^2+dC_0\sum_{i=0}^\infty\frac{\gamma^2\varrho_{i,s}}{(\varrho_{i,s}+\lambda)^2}m_{i,s}^2\\ =\;&-\gamma\sum_{i=0}^\infty m_{i,s}^2+\gamma\sum_{i=0}^\infty\frac{\lambda\varrho_{i,s}^\alpha}{\varrho_{i,s}+\lambda}\varrho_{i,s}^{-\alpha}m_{i,s}^2+dC_0\gamma^2\sum_{i=0}^\infty\left(\frac{\varrho_{i,s}^{\frac{1+\alpha}{2}}}{\varrho_{i,s}+\lambda}\right)^2\rho_{i,s}^{-\alpha}m_{i,s}^2\\ \overset{\text{(a)}}{\le}\;&-\gamma\,\mathrm{MMD}^2(\mu_s,\pi)+\gamma\lambda^\alpha\sum_{i=0}^\infty\rho_{i,s}^{-\alpha}m_{i,s}^2+dC_0\gamma^2\lambda^{\alpha-1}\|q_s\|_{\mathcal{H}}^2 \qquad \text{(A.5)}\\ =\;&-\gamma\,\mathrm{MMD}^2(\mu_s,\pi)+\gamma\lambda^\alpha\|q_s\|_{\mathcal{H}}^2+dC_0\gamma^2\lambda^{\alpha-1}\|q_s\|_{\mathcal{H}}^2,\end{aligned}$$

where (a) follows from Lemma B.1. The second term goes to 0 as $\lambda\to0$, and the third term goes to ∞ as $\lambda\to0$ because $\alpha<1$. To balance these two terms, we take λ such that $\gamma\lambda^\alpha\|q_s\|_{\mathcal{H}}^2=dC_0\gamma^2\lambda^{\alpha-1}\|q_s\|_{\mathcal{H}}^2$, which gives $\lambda=dC_0\gamma$. So we have

$$\mathrm{MMD}^2(\mu_{s+1},\pi)\le(1-2\gamma)\,\mathrm{MMD}^2(\mu_s,\pi)+4(dC_0)^\alpha\gamma^{1+\alpha}\|q_s\|_{\mathcal{H}}^2.$$

Since $0<\gamma<\frac{1}{2}$, after iterating s from 1 to S, we obtain

$$\mathrm{MMD}^2(\mu_S,\pi)\le(1-2\gamma)^S\,\mathrm{MMD}^2(\mu_0,\pi)+4(dC_0)^\alpha\gamma^{1+\alpha}\sum_{s=0}^{S-1}(1-2\gamma)^{S-1-s}\|q_s\|_{\mathcal{H}}^2.$$

Furthermore, if $\|q_s\|^2\le Q$ holds for all $0\le s\le S$, then

$$\begin{aligned}\mathrm{MMD}^2(\mu_S,\pi)&\le\exp(-2S\gamma)\,\mathrm{MMD}^2(\mu_0,\pi)+2(dC_0)^\alpha\gamma^{1+\alpha}Q\sum_{s=0}^\infty(1-2\gamma)^{S-1-s}\\ &=\exp(-2S\gamma)\,\mathrm{MMD}^2(\mu_0,\pi)+4(dC_0)^\alpha\gamma^\alpha Q.\end{aligned}$$

□

A.5 Proof of Proposition 5.1

Proof. Define the following operators:

$$D_X:\mathcal{H}\to\mathbb{R}^{Nd},\quad f\mapsto\frac{1}{\sqrt{N}}\left[\partial_1f(x^{(1)}),\cdots,\partial_df(x^{(N)})\right]^\top,$$

$$D_X^*:\mathbb{R}^{Nd}\to\mathcal{H},\quad v\mapsto\frac{1}{\sqrt{N}}\sum_{i=1}^N\sum_{l=1}^d v_{(i,l)}\,\partial_{1,l}k(x^{(i)},\cdot).$$

Then $S_{\hat{\mu}} = D_X^* D_X$ and $D_X D_X^* = \frac{1}{N} H_{XX}$. The witness function, by definition, admits the representation

$$f_{\hat{\mu},\pi} = (S_{\hat{\mu}} + \lambda \mathrm{Id})^{-1}(m_{\hat{\mu}} - m_\pi) = (D_X^* D_X + \lambda \mathrm{Id})^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} k(x^{(i)}, \cdot) - \mathbb{E}_\pi[k(Y, \cdot)] \right).$$

By the Woodbury identity, $(D_X^* D_X + \lambda \mathrm{Id})^{-1} = \frac{1}{\lambda}\mathrm{Id} - \frac{1}{\lambda} D_X^* (D_X D_X^* + \lambda \mathrm{Id})^{-1} D_X$, we obtain

$$f_{\hat{\mu},\pi} = \frac{1}{\lambda} \left(\frac{1}{N} \sum_{i=1}^{N} k(x^{(i)}, \cdot) - \mathbb{E}_\pi[k(Y, \cdot)] \right) - \frac{1}{\lambda} D_X^* (D_X D_X^* + \lambda \mathrm{Id})^{-1} D_X (m_{\hat{\mu}} - m_\pi).$$

Next, note that $(D_X D_X^* + \lambda \mathrm{Id})^{-1} = N\,(H_{XX} + N\lambda \mathrm{Id})^{-1}$. Moreover, for the mean embeddings,

$$D_X m_{\hat{\mu}} = \frac{1}{N} D_X [\mathbb{K}_X \mathbf{1}_N^\top] = \frac{1}{N\sqrt{N}} D_{XX} \mathbf{1}_N, \quad D_X m_\pi = \frac{1}{\sqrt{N}} \mathbb{E}_\pi[\mathbb{D}_X(Y)].$$

Putting these pieces together,

$$f_{\hat{\mu},\pi} = \frac{1}{N\lambda} \mathbf{1}_N^\top \mathbb{K}_X - \frac{1}{\lambda} \mathbb{E}_\pi[k(Y, \cdot)] - \frac{1}{\lambda} \mathbb{D}_X^\top (H_{XX} + N\lambda \mathrm{Id})^{-1} \left(\frac{1}{N} D_{XX} \mathbf{1}_N - \mathbb{E}_\pi[\mathbb{D}_X(Y)] \right),$$

which is the desired expression. □

Additional notations for Lemma A.2 : All orthogonal complements and closures are taken with respect to the RKHS $\mathcal{H}$. In particular, $\mathcal{G}^{\perp_\mathcal{H}}$ denotes the orthogonal complement of a subspace $\mathcal{G} \subseteq \mathcal{H}$, and $\overline{\mathcal{G}}$ denotes its closure in $\mathcal{H}$.

Lemma A.2. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ have full support on $\mathbb{R}^d$. Let $D_\mu : \mathcal{H} \to L_2^d(\mu)$ and $D_\mu^* : L_2^d(\mu) \to \mathcal{H}$ be the operators defined in the main text. Let $S_\mu := D_\mu^* D_\mu$. If the RKHS $\mathcal{H}$ contains constant functions, then $\overline{\mathrm{Ran}(S_\mu)} = \mathrm{Ker}(S_\mu)^{\perp_\mathcal{H}} = \mathrm{span}\{1\}^{\perp_\mathcal{H}}$; otherwise $\overline{\mathrm{Ran}(S_\mu)} = \{0\}^{\perp_\mathcal{H}}$. In both cases, $m_\mu - m_\pi \in \overline{\mathrm{Ran}(S_\mu)}$.*

Proof. Since $S_\mu = D_\mu^* D_\mu$, for any $f \in \mathcal{H}$ we have

$$\langle f, S_\mu f \rangle_\mathcal{H} = \|D_\mu f\|^2_{L_2^d(\mu)} = \mathbb{E}_{X \sim \mu}\big[\|\nabla f(X)\|^2\big].$$

Hence $f \in \mathrm{Ker}(S_\mu)$ if and only if $\nabla f = 0$ μ-almost surely. Since ∇f is continuous and μ has full support on $\mathbb{R}^d$, this implies $\nabla f(x) = 0$ for all $x \in \mathbb{R}^d$. Indeed, if $\nabla f(x_0) \neq 0$ for some $x_0 \in \mathbb{R}^d$, then by continuity of the derivative of f there exists an open ball $B(x_0, r)$ on which $\|\nabla f\| > 0$. Since μ has full support, $\mu(B(x_0, r)) > 0$, contradicting $\nabla f = 0$ μ-almost surely. Therefore $\nabla f = 0$ everywhere, and since $\mathbb{R}^d$ is connected, f is constant.

By the preceding argument, under the full-support assumption on μ, the null space of S_μ consists precisely of the constant functions that belong to $\mathcal{H}$. Thus, if $1 \in \mathcal{H}$, then $\mathrm{Ker}(S_\mu) = \mathrm{span}\{1\}$, whereas if $1 \notin \mathcal{H}$, then $\mathrm{Ker}(S_\mu) = \{0\}$. The latter case occurs, for instance, for the Gaussian kernel on $\mathbb{R}^d$ [78, Corollary 4.44]. Since S_μ is self-adjoint, $\overline{\mathrm{Ran}(S_\mu)} = \mathrm{Ker}(S_\mu)^{\perp_\mathcal{H}}$. If $1 \notin \mathcal{H}$, this gives $\overline{\mathrm{Ran}(S_\mu)} = \mathcal{H}$, and the claim is immediate. If $1 \in \mathcal{H}$, then $\langle m_\mu - m_\pi, 1 \rangle_\mathcal{H} = \int 1 \,\mathrm{d}\mu - \int 1 \,\mathrm{d}\pi = 1 - 1 = 0$, and hence $m_\mu - m_\pi \in \mathrm{span}\{1\}^{\perp_\mathcal{H}} = \overline{\mathrm{Ran}(S_\mu)}$. □

Additional notations for Lemma A.3 : For two normed vector spaces A, B, $A \cong B$ means that A and B are norm equivalent, i.e., their sets coincide, and the corresponding norms are equivalent. In other words, there are constants $c_1, c_2 > 0$ such that $c_1 \|h\|_A \leq \|h\|_B \leq c_2 \|h\|_A$ holds for all $h \in A$. For a function $f : \mathbb{T}^d \to \mathbb{R}$ defined on the d-dimensional torus, we write its Fourier series as $\widehat{f}_n = \int_{\mathbb{T}^d} f(x) \exp(-\mathrm{i} n^\top x) \mathrm{d}x$ for any $n \in \mathbb{Z}^d$.

Lemma A.3. *Let $\mathbb{T}^d = [0, 2\pi)^d$ be a d-dimensional torus, and let μ be a probability measure on $\mathbb{T}^d$ with Lebesgue density p satisfying $0 < c < p(x) < C < \infty$ a.e. on $\mathbb{T}^d$ for some universal constants c, C. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a periodic translation-invariant Matérn kernel, i.e., $k(x, x') = q(x - x')$, whose Fourier series satisfies $\widehat{q}_n = (1 + \|n\|^2)^{-r}$ for some $r > \max\{d/2, 1\}$. Let $\mathcal{H} = \mathcal{H}(\mathbb{T}^d)$ denote the RKHS induced by k on $\mathbb{T}^d$. Then, for $0 < \alpha \leq 1$,*

$$\mathrm{Ran}(S_\mu^{\alpha/2}) \cong H^{r(\alpha+1)-\alpha}(\mathbb{T}^d) \cap \mathrm{span}\{1\}^{\perp_\mathcal{H}},$$

where $H^\ell(\mathcal{X})$ denotes the Sobolev space of order ℓ on $\mathcal{X}$.

Proof. Denote $e_n(x) = e^{\mathrm{i}n^\top x}$ for all $n \in \mathbb{Z}^d$. Since $k(x, x') = q(x - x')$ is periodic and translation-invariant, we know that $\mathcal{H} \cong H^r(\mathbb{T}^d)$; (See, e.g., [54, Example on Sobolev spaces on the torus]). Moreover,

$$q(x) = \sum_{n\in\mathbb{Z}^d} \hat{q}_n e_n(x), \qquad \hat{q}_n = (1 + \|n\|^2)^{-r} > 0.$$

Therefore, the RKHS can be written as

$$\mathcal{H} = \Big\{ f = \sum_{n\in\mathbb{Z}^d} \hat{f}_n e_n(x) : \|f\|_{\mathcal{H}}^2 = \sum_{n\in\mathbb{Z}^d} \frac{|\hat{f}_n|^2}{\hat{q}_n} < \infty \Big\},$$

and $\phi_n = \hat{q}_n^{1/2} e_n$, $n \in \mathbb{Z}^d$ is an orthonormal basis of RKHS $\mathcal{H}$.

We first consider the Lebesgue measure $\mathfrak{m}$ on $\mathbb{T}^d$. Since $r \geq 1$, for any $f = \sum_n \hat{f}_n e_n$, $g = \sum_n \hat{g}_n e_n \in \mathcal{H}$, we have

$$\langle f, S_{\mathfrak{m}} g\rangle_{\mathcal{H}} = \int_{\mathbb{T}^d} \nabla f(x)^\top \nabla g(x) \mathrm{d}x = \sum_n \|n\|^2 \hat{f}_n \overline{\hat{g}_n}.$$

This also implies $S_{\mathfrak{m}}\phi_n = \tau_n \phi_n$, and $\tau_n = \|n\|^2 \hat{q}_n$. Therefore, $f \in \mathrm{Ran}(S_{\mathfrak{m}}^{\alpha/2})$ if and only if there exists $g = \sum_n b_n \phi_n$ such that

$$f = S_{\mathfrak{m}}^{\alpha/2} g = \sum_n \tau_n^{\alpha/2} b_n \phi_n = \sum_n \tau_n^{\alpha/2} b_n \hat{q}_n^{1/2} e_n = \sum_n \|n\|^\alpha \hat{q}_n^{\frac{1+\alpha}{2}} b_n e_n.$$

By the uniqueness of Fourier expansion, we have $\hat{f}_n = \|n\|^\alpha \hat{q}_n^{\frac{1+\alpha}{2}} b_n$. Since $g \in \mathcal{H}$, it requires $\sum_n |b_n|^2 < \infty$, i.e.,

$$\infty > \sum_{n\neq 0} \frac{|\hat{f}_n|^2}{\|n\|^{2\alpha} |\hat{q}_n|^{1+\alpha}} = \sum_{n\neq 0} \|n\|^{-2\alpha} (1 + \|n\|^2)^{r(1+\alpha)} |\hat{f}_n|^2,$$

and $\hat{f}_0 = 0$. Since for $n \neq 0$,

$$\|n\|^{-2\alpha} (1 + \|n\|^2)^{r(1+\alpha)} \asymp (1 + \|n\|^2)^{r(1+\alpha)-\alpha},$$

we obtain

$$\mathrm{Ran}(S_{\mathfrak{m}}^{\alpha/2}) = \Big\{ f \in \mathcal{H} : \hat{f}_0 = 0, \sum_{n\neq 0} (1 + \|n\|^2)^{r(1+\alpha)-\alpha} |\hat{f}_n|^2 < \infty \Big\}.$$

Since

$$H^\ell(\mathbb{T}^d) = \Big\{ f \in L_2(\mathbb{T}^d) : \sum_n (1 + \|n\|^2)^\ell |\hat{f}_n|^2 < \infty \Big\},$$

and $\hat{f}_0 = 0$ if and only if $\langle f, 1\rangle_{\mathcal{H}} = \hat{q}_0^{-1} \hat{f}_0 = 0$, if and only if $f \in \mathrm{span}\{1\}^{\perp_{\mathcal{H}}}$, therefore

$$\mathrm{Ran}(S_{\mathfrak{m}}^{\alpha/2}) \cong H^{r(\alpha+1)-\alpha}(\mathbb{T}^d) \cap \mathrm{span}\{1\}^{\perp_{\mathcal{H}}}.$$

Finally, we consider a general probability measure μ on $\mathbb{T}^d$. Recall that

$$\langle f, S_\mu g\rangle_{\mathcal{H}} = \int_{\mathbb{T}^d} \nabla f(x)^\top \nabla g(x) p(x) \mathrm{d}x.$$

Since $c \leq p \leq C$, we have

$$c\langle f, S_{\mathfrak{m}} f\rangle_{\mathcal{H}} \leq \langle f, S_\mu f\rangle_{\mathcal{H}} \leq C\langle f, S_{\mathfrak{m}} f\rangle_{\mathcal{H}}, \quad f \in \mathcal{H}.$$

Now consider the restrictions of S_μ and $S_{\mathfrak{m}}$ on $\mathcal{H}_0 := \{1\}^{\perp_{\mathcal{H}}}$. On $\mathcal{H}_0$, we have $c\, S_{\mathfrak{m}} \preceq S_\mu \preceq C\, S_{\mathfrak{m}}$. By the Löwner–Heinz inequality, one obtains $c^\alpha S_{\mathfrak{m}}^\alpha \preceq S_\mu^\alpha \preceq C^\alpha S_{\mathfrak{m}}^\alpha$, $\forall\, 0 < \alpha \leq 1$, still on $\mathcal{H}_0$. Then, an application of Douglas' lemma [29] to $S_\mu^{\alpha/2}$ and $S_{\mathfrak{m}}^{\alpha/2}$ on $\mathcal{H}_0$ shows that

$$\mathrm{Ran}(S_\mu^{\alpha/2}) \cong \mathrm{Ran}(S_{\mathfrak{m}}^{\alpha/2}), \quad \forall\, 0 < \alpha \leq 1. \tag{A.6}$$

Combining this with the previous result for $S_{\mathfrak{m}}$, we obtain

$$\mathrm{Ran}(S_\mu^{\alpha/2}) \cong H^{r(\alpha+1)-\alpha}(\mathbb{T}^d) \cap \mathrm{span}\{1\}^{\perp_{\mathcal{H}}}.$$

□

Lemma A.4. *Let $\mathbb{T}^d = [0, 2\pi)^d$ be a d-dimensional torus. Let k be a periodic Matérn kernel on a d-dimensional torus $\mathbb{T}^d$ whose RKHS is norm equivalent to a Sobolev space $H^r(\mathbb{T})$. Let μ be a distribution on $\mathbb{T}$ that is equivalent to the Lebesgue measure $\mathfrak{m}$; and also $\frac{\mathrm{d}\mu}{\mathrm{d}\mathfrak{m}} \in C^\infty(\mathbb{T})$. If $\frac{\mathrm{d}\pi}{\mathrm{d}\mu} - 1 \in H^{r\alpha-r-\alpha}(\mathbb{T})$, then $m_\mu - m_\pi \in \mathrm{Ran}(S_\mu^{\alpha/2})$.*

Proof. Following the same reasoning as in the lemma above, it suffices to first consider the case where μ is the Lebesgue measure m on $\mathbb{T}^d$. Write $e_n(x) = e^{\mathrm{i}n^\top x}$ for $n \in \mathbb{Z}^d$. Since $k(x, y) = q(x - y)$ is a periodic Matérn kernel whose RKHS is norm equivalent to $H^r(\mathbb{T}^d)$, we have

$$q(x) = \sum_{n\in\mathbb{Z}^d} \hat{q}_n e_n(x), \quad \hat{q}_n \asymp (1 + \|n\|^2)^{-r}.$$

We have that $\phi_n = \hat{q}_n^{1/2} e_n$, $n \in \mathbb{Z}^d$ is an orthonormal basis of RKHS $\mathcal{H}$. Define the following two operators:

$$\iota_\mathfrak{m} : \mathcal{H} \to L_2(\mathfrak{m}), \quad f \mapsto f, \quad \text{and} \quad \iota_\mathfrak{m}^* : L_2(\mathfrak{m}) \to \mathcal{H}, \quad \iota_\mathfrak{m}^* f(\cdot) = \int_\mathbb{T} k(x, \cdot) f(x) \mathrm{d}x.$$

We thus have

$$\begin{aligned}
[\widehat{m_\mu - m_\pi}]_n = \langle m_\mu - m_\pi, \phi_n\rangle_\mathcal{H} &= \left\langle \iota_\mathfrak{m}^* \left(\frac{\mathrm{d}\pi}{\mathrm{d}m} - 1\right), \phi_n \right\rangle_\mathcal{H} = \left\langle \frac{\mathrm{d}\pi}{\mathrm{d}\mathfrak{m}} - 1, \iota_\mathfrak{m}\phi_n \right\rangle_{L_2(\mathbb{T})} \\
&= \hat{q}_n^{1/2} \left\langle \frac{\mathrm{d}\pi}{\mathrm{d}\mathfrak{m}} - 1, e_n \right\rangle_{L_2(\mathbb{T})}.
\end{aligned}$$

By the proof of Lemma A.3, we know that a sufficient condition for $m_\mu - m_\pi \in \mathrm{Ran}(S_\mu^{\alpha/2})$ is that

$$\sum_n |[\widehat{m_\mu - m_\pi}]_n|^2 \left(\|n\|^2 \hat{q}_n\right)^{-\alpha} < \infty \iff \sum_n \left\langle \frac{\mathrm{d}\pi}{\mathrm{d}\mathfrak{m}} - 1, e_n \right\rangle_{L_2(\mathbb{T})}^2 (1 + \|n\|^2)^{-r(1-\alpha)-\alpha} < \infty.$$

By the definition of the Sobolev space $H^{r\alpha-r-\alpha}(\mathbb{T})$, the proof is concluded.

Next, we consider the more general case where μ is a distribution on $\mathbb{T}$ that is equivalent to the Lebesgue measure $\mathfrak{m}$; and also $\frac{\mathrm{d}\mu}{\mathrm{d}\mathfrak{m}} \in C^\infty(\mathbb{T})$. Following the same argument above, we can show that $m_\mu - m_\pi = \iota_\mathfrak{m}^*(\frac{\mathrm{d}\mu}{\mathrm{d}\mathfrak{m}} \cdot (\frac{\mathrm{d}\pi}{\mathrm{d}\mu} - 1))$. Since $\frac{\mathrm{d}\pi}{\mathrm{d}\mu} - 1 \in H^{r\alpha-r-\alpha}(\mathbb{T})$ and $\frac{\mathrm{d}\mu}{\mathrm{d}\mathfrak{m}} \in C^\infty(\mathbb{T})$, there is $(\frac{\mathrm{d}\mu}{\mathrm{d}\mathfrak{m}} \cdot (\frac{\mathrm{d}\pi}{\mathrm{d}\mu} - 1)) \in H^{r\alpha-r-\alpha}(\mathbb{T})$ and hence $m_\mu - m_\pi \in \mathrm{Ran}(S_\mathfrak{m}^{\alpha/2})$. The proof is concluded by that $\mathrm{Ran}(S_\mu^{\alpha/2}) \cong \mathrm{Ran}(S_\mathfrak{m}^{\alpha/2})$ proved in Eq. (A.6) above. □

A.6 Geometry of SrMMD flow

In this section, we expand on Remark 3.5 and discuss the geometry underlying the SrMMD flow.

First, we recall that, in the formal Riemannian interpretation of Wasserstein gradient flows, the tangent space $T_\mu\mathcal{P}_2(\mathbb{R}^d)$ at a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is identified with a subspace of vector fields in $L_2^d(\mu)$. More precisely, it is commonly taken as the $L_2^d(\mu)$-closure of gradients of smooth compactly supported functions [82]. The Wasserstein metric tensor is then given by the kinetic-energy inner product $g_\mu(v, w) := \langle v, w\rangle_{L_2^d(\mu)}$. Indeed, for a smooth functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ with first variation $\frac{\delta\mathcal{F}}{\delta\mu}$, an absolutely continuous curve satisfying $\partial_t\mu_t + \nabla \cdot (\mu_t v_t) = 0$ obeys the formal chain rule [21]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\mu_t) = \int_{\mathbb{R}^d} \nabla \frac{\delta\mathcal{F}}{\delta\mu_t}(x)^\top v_t(x) \,\mathrm{d}\mu_t(x).$$

Thus the gradient of $\mathcal{F}$ with respect to the optimal transport geometry, g_μ defined above, is represented by $\mathrm{grad}_W\mathcal{F}(\mu) = \nabla\frac{\delta\mathcal{F}}{\delta\mu}$, and the steepest descent velocity is $v_t = -\nabla\frac{\delta\mathcal{F}}{\delta\mu_t}$.

Now, consider the SrMMD flow. Recall the operator $D_\mu^* : L_2^d(\mu) \to \mathcal{H}$ defined in the main text. Define a Sobolev modified Riemannian metric tensor g_μ^λ with $\lambda > 0$: for two tangent vector fields $v, w \in T_\mu\mathcal{P}_2(\mathbb{R}^d)$,

$$g_\mu^\lambda(v, w) := \left\langle D_\mu^* v, D_\mu^* w \right\rangle_\mathcal{H} + \lambda\langle v, w\rangle_{L_2^d(\mu)}. \tag{A.7}$$

Theorem A.5. *Suppose that Assumptions 1 and 2 hold. Suppose $m_\mu - m_\pi \in \overline{\mathrm{Ran}(S_\mu)}$. The steepest descent of the MMD functional $\mu \mapsto \frac{1}{2}\,\mathrm{MMD}^2(\mu, \pi)$ under the metric g_μ^λ with $\lambda > 0$ yields the SrMMD velocity field $v = -\nabla \left(S_\mu + \lambda \mathrm{Id}\right)^{-1} (m_\mu - m_\pi)$.*

Proof. Let $\phi \in C_c^\infty(\mathbb{R}^d) \cap \mathcal{H}$ such that $\nabla\phi \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$. Consider an absolutely continuous curve $t \mapsto (\mathrm{Id} + t\nabla\phi)_{\#}\mu =: \eta_t$. Following the same derivation at the start of Appendix A.3, we obtain that

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \frac{1}{2}\,\mathrm{MMD}^2(\eta_t, \pi) = \int \nabla(m_\mu - m_\pi)(x)^\top \nabla\phi(x)\,\mathrm{d}\mu(x) = \langle m_\mu - m_\pi,\, S_\mu\phi\rangle_{\mathcal{H}}\,.$$

On the other hand, by definition in metric geometry, a gradient vector field $\mathrm{grad}_\mu\,\mathcal{E} = \nabla f$ (for some witness potential $f \in \mathcal{H}$) must satisfy the duality pairing: $g_\mu^\lambda(\nabla f, \nabla\phi) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \frac{1}{2}\,\mathrm{MMD}^2(\eta_t, \pi)$. From g_μ^λ defined above, we have

$$g_\mu^\lambda(\nabla f, \nabla\phi) = \langle S_\mu f, S_\mu\phi\rangle_{\mathcal{H}} + \lambda\,\langle f, S_\mu\phi\rangle_{\mathcal{H}} = \langle \left(S_\mu + \lambda\mathrm{Id}\right) f, S_\mu\phi\rangle_{\mathcal{H}}\,.$$

Comparing the two equations yields: $\langle \left(S_\mu + \lambda\mathrm{Id}\right) f, S_\mu\phi\rangle_{\mathcal{H}} = \langle m_\mu - m_\pi, S_\mu\phi\rangle_{\mathcal{H}}$. This equality must hold for all test functions $\phi \in \mathcal{H}$, meaning it holds for all elements $S_\mu\phi \in \overline{\mathrm{Ran}(S_\mu)}$. Since $m_\mu - m_\pi \in \overline{\mathrm{Ran}(S_\mu)}$ and we can freely pick $f \in \overline{\mathrm{Ran}(S_\mu)}$, we can drop the inner product with $S_\mu\phi$ to uniquely identify the operator equation:

$$\left(S_\mu + \lambda\mathrm{Id}\right) f = m_\mu - m_\pi.$$

Inverting the strictly positive-definite operator $(S_\mu + \lambda\mathrm{Id})$, we identify the optimal witness potential $f : f = \left(S_\mu + \lambda\mathrm{Id}\right)^{-1} (m_\mu - m_\pi)$. Consequently, the steepest descent vector field is $v = -\nabla f = -\nabla \left(S_\mu + \lambda\mathrm{Id}\right)^{-1} (m_\mu - m_\pi)$. □

B Auxiliary Results

Lemma B.1. *For any $\lambda > 0$ and $s \in [0, 1]$, we have*

$$\sup_{t \geq 0} \frac{t^s}{t + \lambda} \leq \lambda^{s-1}.$$

Proof. Since $a^s \leq a + 1$ for any $a \geq 0$ and $0 \leq s \leq 1$, the lemma follows from

$$\left(\frac{t}{\lambda}\right)^s \leq 1 + \frac{t}{\lambda} = \frac{t + \lambda}{\lambda}.$$

□

Lemma B.2. *Given two Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$. For any bounded linear operator $T : \mathcal{H}_1 \to \mathcal{H}_2$, T is Hilbert-Schmidt if and only if T^*T is trace-class. Moreover, $\mathrm{Tr}(T^*T) = \|T\|_{HS}^2$.*

Proof. We write $\mathbb{N}$ for either $\mathbb{N} = \{1, 2, \cdots\}$ or $\mathbb{N} = \{1, 2, \cdots, N\}$ for some fixed finite integer N.

First, assume that T is Hilbert-Schmidt. For any orthonormal basis $\{e_n\}_{n\in\mathbb{N}}$ of $\mathcal{H}_1$,

$$\sum_{n\in\mathbb{N}} \langle e_n, T^*Te_n\rangle_{\mathcal{H}_1} = \sum_{n\in\mathbb{N}} \|Te_n\|_{\mathcal{H}_1}^2 < \infty.$$

Since T^*T is a positive operator, it follows that T^*T is trace-class and

$$\mathrm{Tr}(T^*T) = \sum_{n\in\mathbb{N}} \|Te_n\|_{\mathcal{H}_1}^2 < \infty.$$

Conversely, assume that T^*T is trace-class. Since T^*T is positive, for any orthonormal basis $\{e_n\}_{n\in\mathbb{N}}$,

$$\mathrm{Tr}(T^*T) = \sum_{n\in\mathbb{N}} \langle e_n, T^*Te_n\rangle_{\mathcal{H}_1} = \sum_{n\in\mathbb{N}} \|Te_n\|_{\mathcal{H}_1}^2 < \infty.$$

This implies that

$$\|T\|_{\mathrm{HS}}^2 = \sum_{n\in\mathbb{N}} \|Te_n\|^2 < \infty,$$

i.e., T is Hilbert–Schmidt. □

Lemma B.3. *Suppose that the kernel k is continuously differentiable with respect to both of its arguments, and that it satisfies $\int \partial_{1,i}\partial_{2,i}k(x,x)\mathrm{d}\mu(x) < \infty$ for any $i \in [d]$. Then the operator $D_\mu : f \in \mathcal{H} \mapsto \nabla f$ is well defined, linear, and bounded. Furthermore, its adjoint operator $D_\mu^* : L_2^d(\mu) \to \mathcal{H}$ is also a bounded linear operator.*

Proof. From Corollary 4.36 of [78], an immediate consequence of the continuously differentiability of k is that all functions $f \in \mathcal{H}$ are differentiable, its derivative belongs to $\mathcal{H}$, and satisfies the reproducing property: $\partial_i f(x) = \langle f, \partial_{1,i}k(x,\cdot)\rangle_{\mathcal{H}}$ for any $i \in \{1,\dots,d\}$. Next, we are going to show that $\nabla f \in L_2^d(\mu)$. Too see why, notice that for any $i \in \{1,\dots,d\}$,

$$\int |\partial_i f(x)|^2 \,\mathrm{d}\mu(x) \le \|f\|_{\mathcal{H}}^2 \int \|\partial_{1,i}k(x,\cdot)\|_{\mathcal{H}}^2 \,\mathrm{d}\mu(x) = \|f\|_{\mathcal{H}}^2 \int \partial_{1,i}\partial_{2,i}k(x,x)\,\mathrm{d}\mu(x) < \infty.$$

As a result, $\mathrm{ran}(D_\mu) \subseteq L_2^d(\mu)$. Another immediate consequence of the above equation is that the operator D_μ is bounded.

Furthermore, consider another operator

$$D_\mu^* : L_2^d(\mu) \to \mathcal{H}, \quad g = (g_1,\dots,g_d) \mapsto \sum_{i=1}^d \int g_i(x)\partial_{1,i}k(x,\cdot)\mathrm{d}\mu(x) \in \mathcal{H}.$$

To see why D_μ^* is the adjoint operator of D_μ, one can easily verify that $\langle f, D_\mu^* g\rangle_{\mathcal{H}} = \langle D_\mu f, g\rangle_{L_2^d(\mu)}$. This mapping is well-defined: for each i, the function $x \mapsto g_i(x)\partial_{1,i}k(x,\cdot)$ is $\mathcal{H}$-valued and Bochner integrable with respect to μ. Indeed, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\left\| \int g_i(x)\partial_{1,i}k(x,\cdot)\mathrm{d}\mu(x) \right\|_{\mathcal{H}} &\le \int |g_i(x)| \left\|\partial_{1,i}k(x,\cdot)\right\|_{\mathcal{H}} \mathrm{d}\mu(x) \\ &\le \|g_i\|_{L_2(\mu)} \left(\int \|\partial_{1,i}k(x,\cdot)\|_{\mathcal{H}}^2 \,\mathrm{d}\mu(x) \right)^{1/2}.\end{aligned}$$

so D_μ^* is also a bounded linear operator $L_2^d(\mu) \to \mathcal{H}$. The proof is thus concluded. □

Lemma B.4. *Suppose that the kernel k is continuously differentiable with respect to both of its arguments, and that it satisfies $\int \partial_{1,i}\partial_{2,i}k(x,x)\mathrm{d}\mu(x) < \infty$ for any $i \in [d]$. The operator $S_\mu = D_\mu^* D_\mu : \mathcal{H} \to \mathcal{H}$ is trace-class, Hilbert-Schmidt, and compact.*

Proof. Following Lemma B.2, since $S_\mu = D_\mu^* D_\mu$, it suffices to prove that D_μ is Hilbert-Schmidt. Define $D_{\mu,i} : \mathcal{H} \to L_2(\mu)$ by $(D_{\mu,i}f)(x) = \partial_i f(x) = \langle f, \partial_{2,i}k(x,\cdot)\rangle_{\mathcal{H}}$ for each $i = 1, 2, \cdots, d$. Then, for any orthonormal basis $\{e_j\}_{j\ge1}$ of $\mathcal{H}$, and for any fixed i,

$$\|D_{\mu,i}\|_{\mathrm{HS}}^2 = \sum_{j=1}^\infty \|D_{\mu,i}e_j\|_{L_2(\mu)}^2 = \sum_{j=1}^\infty \int |\partial_i e_j(x)|^2 d\mu(x) = \int \sum_{j=1}^\infty |\langle e_j, \partial_{2,i}k(x,\cdot)\rangle|^2 d\mu(x).$$

By Parseval's identity, $\sum_{j=1}^\infty |\langle e_j, \partial_{2,i}k(x,\cdot)\rangle|^2 = \|\partial_{2,i}k(x,\cdot)\|_{\mathcal{H}}^2 = \partial_{1,i}\partial_{2,i}k(x,x)$. Therefore,

$$\|D_\mu\|_{\mathrm{HS}}^2 = \sum_{i=1}^d \|D_{\mu,i}\|_{\mathrm{HS}}^2 = \sum_{i=1}^d \int \partial_{1,i}\partial_{2,i}k(x,x)d\mu(x) < \infty.$$

□

Lemma B.5. *Suppose the kernel satisfies the tail condition $\lim_{\|x\|\to\infty} \|x\|^{d-1}\pi(x)\sqrt{k(x,x)} = 0$. For any vector-valued function $f : \mathbb{R}^d \to \mathbb{R}^d \in \mathcal{H}^d$, it satisfies $\int (f(x)^\top \nabla \log \pi(x) + \nabla \cdot f(x))\,\mathrm{d}\pi(x) = 0$.*

Proof. Note that

$$\int_{\mathbb{R}^d} (f(x)^\top \nabla \log \pi(x) + \nabla \cdot f(x)) \mathrm{d}\pi(x) = \int_{\mathbb{R}^d} (f(x)^\top \nabla \pi(x) + \nabla \cdot f(x)\pi(x)) \mathrm{d}x$$
$$= \lim_{R\to\infty} \int_{B_R} \nabla \cdot (\pi(x) f(x)) dx = \lim_{R\to\infty} \int_{\partial B_R} \pi(x) f(x)^\top n(x) \mathrm{d}\omega(x),$$

where $B_R := \{x : \|x\| \le R\}$ is a ball of radius R, $n(x)$ is the outward unit normal vector, and $\mathrm{d}\omega$ is the $(d-1)$-dimensional surface measure. Next, notice that $|f(x)^\top n(x)| \le \sum_{i=1}^d |f_i(x)| \le \sum_{i=1}^d \|f_i\|_{\mathcal{H}} \sqrt{k(x,x)} = \|f\|_{\mathcal{H}^d} \sqrt{k(x,x)}$. Therefore,

$$\int_{\partial B_R} \pi(x) f(x)^\top n(x) \mathrm{d}\omega(x) \le \int_{\partial B_R} \pi(x) \|f\|_{\mathcal{H}^d} \sqrt{k(x,x)} \mathrm{d}\omega(x)$$
$$\le \|f\|_{\mathcal{H}^d} \left| \int_{\partial B_R} 1 \, \mathrm{d}\omega(x) \right| \sup_{\|x\|=R} \pi(x) \sqrt{k(x,x)} = \|f\|_{\mathcal{H}^d} C_d R^{d-1} \sup_{\|x\|=R} \pi(x) \sqrt{k(x,x)}.$$

So we obtain

$$\left| \int_{\mathbb{R}^d} (f(x)^\top \nabla \log \pi(x) + \nabla \cdot f(x)) \mathrm{d}\pi(x) \right| \le \|f\|_{\mathcal{H}^d} C_d \lim_{\|x\|\to\infty} \|x\|^{d-1} \pi(x) \sqrt{k(x,x)}.$$

By the assumption in the statement of the lemma, the limit is 0, and the proof is concluded. □

Lemma B.6. *Suppose Assumption 2 holds. Let $\mathcal{K} = B(0,R)$ be a ball of radius R in $\mathbb{R}^d$. Let $\mu \in \mathcal{P}_2(\mathcal{K})$. Let $\boldsymbol{v}_{\mathrm{MMD}}[\mu]$ be the vector field of the MMD flow defined in Section 2. Then, there exists $l_{\mathcal{K}} > 0$ such that for any $\mu \in \mathcal{P}_2(\mathcal{K})$, we have $\mathrm{Lip}(\boldsymbol{v}_{\mathrm{MMD}}[\mu](\cdot); \mathcal{K}) \le l_{\mathcal{K}}$. Also, there exists $L_{\mathcal{K}} > 0$ such that for any $\mu, \nu \in \mathcal{P}_2(\mathcal{K})$, we have $\sup_{x\in\mathcal{K}} \|\boldsymbol{v}_{\mathrm{MMD}}[\mu](x) - \boldsymbol{v}_{\mathrm{MMD}}[\nu](x)\| \le \mathcal{L}_{\mathcal{K}} W_2(\mu,\nu)$. The same holds for the vector field of SrMMD flow $\nabla f_{\mu,\pi}$ in Proposition 3.3 as well, where $f_{\mu,\pi} = (S_\mu + \lambda \mathrm{Id})^{-1}(m_\mu - m_\pi)$.*

Proof. Recall from the main text that $\boldsymbol{v}_{\mathrm{MMD}}[\mu](\cdot) = \int \nabla_2 k(x,\cdot) \, \mathrm{d}(\mu-\pi)(x)$. Therefore, we have

$$\sup_{x_1,x_2\in\mathcal{K}} \frac{\|\boldsymbol{v}_{\mathrm{MMD}}[\mu](x_1) - \boldsymbol{v}_{\mathrm{MMD}}[\mu](x_2)\|}{\|x_1 - x_2\|}$$
$$= \sup_{x_1,x_2\in\mathcal{K}} \frac{\left\| \int (\nabla_2 k(x,x_1) - \nabla_2 k(x,x_2)) \mathrm{d}(\mu-\pi)(x) \right\|}{\|x_1 - x_2\|}$$
$$\le \sup_{x_1,x_2\in\mathcal{K}} \left\| \int k(x,\cdot) \mathrm{d}(\mu-\pi)(x) \right\|_{\mathcal{H}} \cdot \|\nabla_1 k(x_1,\cdot) - \nabla_1 k(x_2,\cdot)\|_{\mathcal{H}^d} \cdot \|x_1 - x_2\|^{-1}$$
$$\le 2 \sup_{x\in\mathcal{K}} \sqrt{k(x,x)} dB_R := \ell_{\mathcal{K}}.$$

The last inequality holds by Lemma B.7. So we have proved the first claim for $\boldsymbol{v}_{\mathrm{MMD}}$.

Next, let Γ be an optimal coupling between μ and ν. Then, for any $x_1 \in \mathcal{K}$,

$$\boldsymbol{v}_{\mathrm{MMD}}[\mu](x_1) - \boldsymbol{v}_{\mathrm{MMD}}[\nu](x_1) = \int \nabla_2 k(x,x_1) \, \mathrm{d}(\mu-\nu)(x)$$
$$= \int \big(\nabla_2 k(a,x_1) - \nabla_2 k(b,x_1)\big) \, \mathrm{d}\Gamma(a,b).$$

Therefore,

$$\sup_{x_1\in\mathcal{K}} \|\boldsymbol{v}_{\mathrm{MMD}}[\mu](x_1) - \boldsymbol{v}_{\mathrm{MMD}}[\nu](x_1)\| \le \sup_{x_1\in\mathcal{K}} \int \|\nabla_2 k(a,x_1) - \nabla_2 k(b,x_1)\| \, \mathrm{d}\Gamma(a,b)$$
$$\le \sup_{x_1\in\mathcal{K}} \sup_{a,b\in\mathcal{K}} \frac{\|\nabla_2 k(a,x_1) - \nabla_2 k(b,x_1)\|}{\|a-b\|} \int \|a-b\| \, \mathrm{d}\Gamma(a,b)$$
$$\le dB_R^2 W_2(\mu,\nu) =: \mathcal{L}_{\mathcal{K}} W_2(\mu,\nu).$$

Now, for each $i = 1, \dots, d$,

$$\sup_{x_1 \in \mathcal{K}} \sup_{a,b \in \mathcal{K}} \frac{|\partial_{2,i} k(a, x_1) - \partial_{2,i} k(b, x_1)|}{\|a - b\|} \leq \sup_{x_1 \in \mathcal{K}} \sup_{c \in \mathcal{K}} \|\nabla_1 \partial_{2,i} k(c, x_1)\| \leq \sqrt{d} B_R^2,$$

where we used the mean value theorem, and therefore,

$$\sup_{x_1 \in \mathcal{K}} \sup_{a,b \in \mathcal{K}} \frac{\|\nabla_2 k(a, x_1) - \nabla_2 k(b, x_1)\|}{\|a - b\|} \leq d B_R^2.$$

So we have proved the second claim for the $\boldsymbol{v}_{\mathrm{MMD}}$.

Next, we are going to prove that both claims hold for the vector field of the SrMMD flow as well.

$$\sup_{x_1, x_2 \in \mathcal{K}} \frac{\|\nabla f_{\mu,\pi}(x_1) - \nabla f_{\mu,\pi}(x_2)\|}{\|x_1 - x_2\|} \leq \|f_{\mu,\pi}\|_{\mathcal{H}} \cdot \|\nabla_1 k(x_1, \cdot) - \nabla_1 k(x_2, \cdot)\|_{\mathcal{H}^d} \cdot \|x_1 - x_2\|^{-1}$$
$$\leq 2\lambda^{-1} \sup_{x \in \mathcal{K}} \sqrt{k(x, x)} B_R.$$

And we also have,

$$\begin{aligned}
&\sup_{x_1 \in \mathcal{K}} \|\nabla f_{\mu,\pi}(x_1) - \nabla f_{\nu,\pi}(x_1)\| \\
&\leq \sup_{x_1 \in \mathcal{K}} \left\| \left\langle \nabla k(x_1, \cdot),\ (S_\mu + \lambda)^{-1}(m_\mu - m_\pi) - (S_\nu + \lambda)^{-1}(m_\nu - m_\pi) \right\rangle_{\mathcal{H}^d} \right\| \\
&\leq \sup_{x_1 \in \mathcal{K}} \|\nabla k(x_1, \cdot)\|_{\mathcal{H}^d} \cdot \left\| (S_\mu + \lambda)^{-1}(m_\mu - m_\pi) - (S_\nu + \lambda)^{-1}(m_\nu - m_\pi) \right\|_{\mathcal{H}}.
\end{aligned}$$

Note that

$$\begin{aligned}
&\left\| (S_\mu + \lambda)^{-1}(m_\mu - m_\pi) - (S_\nu + \lambda)^{-1}(m_\nu - m_\pi) \right\|_{\mathcal{H}} \\
&\leq \left\| (S_\mu + \lambda)^{-1}(m_\mu - m_\nu) \right\|_{\mathcal{H}} + \left\| \left((S_\mu + \lambda)^{-1} - (S_\nu + \lambda)^{-1} \right)(m_\nu - m_\pi) \right\|_{\mathcal{H}} \\
&\leq \lambda^{-1} \|m_\mu - m_\nu\|_{\mathcal{H}} + \left\| (S_\mu + \lambda \mathrm{Id})^{-1}(S_\nu - S_\mu)(S_\nu + \lambda)^{-1}(m_\nu - m_\pi) \right\|_{\mathcal{H}} \\
&\leq \lambda^{-1} \|m_\mu - m_\nu\|_{\mathcal{H}} + \lambda^{-2} \|S_\nu - S_\mu\|_{\mathrm{op}} \cdot \|m_\nu - m_\pi\|_{\mathcal{H}} \\
&\leq \lambda^{-1} \sqrt{d} B_R W_2(\mu, \nu) + 2\lambda^{-2} \sup_{x \in \mathcal{K}} \sqrt{k(x, x)}\, d^{3/2} B_R^2 W_2(\mu, \nu) \\
&= \frac{2 \sup_{x \in \mathcal{K}} \sqrt{k(x, x)}\, d^{3/2} B_R^2 + \lambda \sqrt{d} B_R}{\lambda^2} W_2(\mu, \nu).
\end{aligned}$$

In the above chain of derivations, we use the resolvent identity $(A + \lambda)^{-1} - (B + \lambda)^{-1} = (A + \lambda)^{-1}(B - A)(B + \lambda)^{-1}$. Moreover, if Γ is an optimal coupling between μ and ν, then

$$\|m_\mu - m_\nu\|_{\mathcal{H}} \leq \int \|k(x, \cdot) - k(y, \cdot)\|_{\mathcal{H}} \,\mathrm{d}\Gamma(x, y) \leq \int \sqrt{d} B_R \|x - y\| \,\mathrm{d}\Gamma(x, y) = \sqrt{d} B_R W_2(\mu, \nu).$$

The same argument yields $\|S_\mu - S_\nu\|_{\mathrm{op}} \leq 2d^{3/2} B_R^2 W_2(\mu, \nu)$ as well. Therefore, we have proved both claims for the vector field of SrMMD flow as well. □

Lemma B.7. *Suppose Assumption 2 holds with the constant B_R. Let $\mathcal{K} = B(0, R)$ be a ball of radius R in $\mathbb{R}^d$. Then, $\|k(x, \cdot) - k(y, \cdot)\|_{\mathcal{H}} \leq \sqrt{d} B_R \|x - y\|$, and $\|\nabla_1 k(x, \cdot) - \nabla_1 k(y, \cdot)\|_{\mathcal{H}^d} \leq d B_R \|x - y\|$.*

Proof. Fix $x, y \in \mathcal{K}$, and define $\gamma(t) = y + t(x - y)$ for $t \in [0, 1]$. Since $\mathcal{K}$ is convex, $\gamma(t) \in \mathcal{K}$ for all $t \in [0, 1]$. By the fundamental theorem of calculus for Hilbert-space-valued functions,

$$k(x, \cdot) - k(y, \cdot) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} k(\gamma(t), \cdot) \,\mathrm{d}t = \int_0^1 \sum_{i=1}^d (x_i - y_i)\, \partial_{1,i} k(\gamma(t), \cdot) \,\mathrm{d}t.$$

Therefore,

$$\|k(x, \cdot) - k(y, \cdot)\|_{\mathcal{H}} \leq \int_0^1 \sum_{i=1}^d |x_i - y_i|\, \|\partial_{1,i} k(\gamma(t), \cdot)\|_{\mathcal{H}} \,\mathrm{d}t \leq \sqrt{d}\, B_R \|x - y\|.$$

The same holds for $\|\nabla_1 k(x, \cdot) - \nabla_1 k(y, \cdot)\|_{\mathcal{H}^d}$ as well. □

Proposition B.8. *Suppose Assumptions 1 and 2 hold. Fix a finite time horizon $T \in [0, \infty)$. Suppose $\mathrm{MMD}(\mu_0, \pi) < C_{\mathrm{MMD}}$, and $\mu_0, \pi \in \mathcal{P}_2(\mathbb{R}^d)$. Then, $(\mu_t)_{t\geq 0}$ is well-defined, in the sense that the trajectory is a unique absolutely continuous curve in $\mathcal{P}_2(\mathbb{R}^d)$ and does not explode in finite time.*

Proof. To prove the existence of MMD gradient flow, we need to check that all four hypotheses of D1 in [11] are satisfied.

Since μ_t is the MMD flow, there is $\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{MMD}^2(\mu_t, \pi) = -\int_{\mathbb{R}^d} \|\nabla f_{\mu_t,\pi}(x)\|^2 \,\mathrm{d}\mu_t(x) \leq 0$, where $f_{\mu_t,\pi} = m_{\mu_t} - m_\pi$. This indicates that $\mathrm{MMD}^2(\mu_t, \pi)$ is monotonically decreasing in time along the trajectory. Hence, since $\mathrm{MMD}(\mu_0, \pi) < C_{\mathrm{MMD}}$, there is $\sup_{t\in[0,\infty)} \mathrm{MMD}(\mu_t, \pi) < C_{\mathrm{MMD}}$. Let $\mathcal{K} = B(0, R)$ be a ball of radius R in $\mathbb{R}^d$. Take any sequence $x_n \to x$ in $\mathcal{K}$.

$$\|\boldsymbol{v}_{\mathrm{MMD}}[\mu_t](x_n) - \boldsymbol{v}_{\mathrm{MMD}}[\mu_t](x)\| = \left\| \int \nabla_2 k(y, x) \,\mathrm{d}(\mu_t - \pi)(y) - \int \nabla_2 k(y, x_n) \,\mathrm{d}(\mu_t - \pi)(y) \right\|$$
$$\leq \|m_{\mu_t} - m_\pi\|_{\mathcal{H}} \cdot \|\nabla_1 k(x, \cdot) - \nabla_1 k(x_n, \cdot)\|_{\mathcal{H}^d} \to 0.$$

The last step holds from $\|m_{\mu_t} - m_\pi\|_{\mathcal{H}} \leq C_{\mathrm{MMD}}$ as proved above, and from Lemma B.7. Hence, D1(i) is satisfied.

D1(ii) is satisfied by the following derivations:

$$\|\boldsymbol{v}_{\mathrm{MMD}}[\mu_t](x)\| = \|\nabla(m_{\mu_t} - m_\pi)(x)\| \leq \sum_{j=1}^{d} \left|\langle \partial_{1,j} k(x, \cdot),\ (m_{\mu_t} - m_\pi)\rangle_{\mathcal{H}}\right|$$
$$\leq \sum_{j=1}^{d} \|\partial_{1,j} k(x, \cdot)\|_{\mathcal{H}} \cdot C_{\mathrm{MMD}} \leq d C_{\mathrm{MMD}}(\|x\| + 1). \tag{B.1}$$

D1(iii) and D1(iv) are checked in Lemma B.6. Therefore, it shows that the hypotheses (DI) of [11] are satisfied. Hence, by Theorem 4 of [11], there exists a trajectory solving the continuity inclusion on any finite time interval $[0, T]$.

In our MMD flow setting, once existence is established, uniqueness follows separately from the Cauchy–Lipschitz theory for continuity equations, for instance from Theorem 2 of [11], since the velocity field satisfies the corresponding growth and local Lipschitz conditions as proved above. As the argument applies on every finite interval $[0, T]$, the trajectory cannot explode in finite time. □

Lemma B.9 (Local bi-Lipschitz regularity of the characteristic flow)**.** *Let $T > 0$, and let $b : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ be a time-dependent vector field such that:*

1. *it is locally Lipschitz: for every $R > 0$, there exists $\ell_R \in L^1(0, T)$ such that for a.e. $t \in [0, T]$ and for all $x, y \in B_R$, $|b_t(x) - b_t(y)| \leq \ell_R(t)|x - y|$;*
2. *it has at most linear growth: there exists $m \in L^1(0, T)$ such that for a.e. $t \in [0, T)$ and all $x \in \mathbb{R}^d$, $|b_t(x)| \leq m(t)(1 + \|x\|)$.*

Let $X(t, x)$ be the unique global solution of the characteristic flow:

$$\partial_t X(t, x) = b_t(X(t, x)), \qquad X(0, x) = x.$$

Then, for every $R_0 > 0$ and every $t \in [0, T]$, the map $X_t : B_{R_0} \to \mathbb{R}^d, x \mapsto X(t, x)$, is bi-Lipschitz onto its image. More precisely, if $R_* := (1 + R_0) \exp(\int_0^T m(s)\,\mathrm{d}s) - 1$, then for all $x, y \in B_{R_0}$ and all $t \in [0, T]$,

$$\exp\Big(-\int_0^t \ell_{R_*}(s)\,\mathrm{d}s\Big)|x - y| \leq |X_t(x) - X_t(y)| \leq \exp\Big(\int_0^t \ell_{R_*}(s)\,\mathrm{d}s\Big)|x - y|.$$

In particular, X_t is locally bi-Lipschitz on bounded sets.

Proof. Fix $R_0 > 0$ and $x \in B_{R_0}$. By the linear growth assumption,

$$\|X(t, x)\| \leq \|x\| + \int_0^t m(s)\big(1 + \|X(s, x)\|\big)\,\mathrm{d}s.$$

Hence

$$1 + \|X(t,x)\| \le 1 + \|x\| + \int_0^t m(s)\big(1 + \|X(s,x)\|\big)\,\mathrm{d}s.$$

By Grönwall's inequality,

$$1 + \|X(t,x)\| \le (1 + \|x\|)\exp\Big(\int_0^t m(s)\,\mathrm{d}s\Big) \le (1 + R_0)\exp\Big(\int_0^T m(s)\,\mathrm{d}s\Big).$$

Therefore,

$$X(s,x) \in B_{R_*} \qquad \text{for all } s \in [0,T],\ x \in B_{R_0}.$$

Now fix $x, y \in B_{R_0}$. Since both trajectories remain in B_{R_*}, we have

$$\|X_t(x) - X_t(y)\| \le \|x - y\| + \int_0^t \ell_{R_*}(s)\,\|X_{\mathfrak{s}}(x) - X_{\mathfrak{s}}(y)\|\,\mathrm{d}s.$$

Another application of Grönwall's inequality yields

$$\|X_t(x) - X_t(y)\| \le \exp\Big(\int_0^t \ell_{R_*}(s)\,\mathrm{d}s\Big)\|x - y\|.$$

For the reverse bound, fix $0 \le s \le t$. Using the integral formulation of the ODE,

$$X_{\mathfrak{s}}(x) - X_{\mathfrak{s}}(y) = X_t(x) - X_t(y) - \int_s^t \big(b_r(X_r(x)) - b_r(X_r(y))\big)\,\mathrm{d}r.$$

Thus

$$\|X_{\mathfrak{s}}(x) - X_{\mathfrak{s}}(y)\| \le \|X_t(x) - X_t(y)\| + \int_s^t \ell_{R_*}(r)\,\|X_r(x) - X_r(y)\|\,\mathrm{d}r.$$

Applying Grönwall's inequality on the interval $[s,t]$, we obtain

$$\|X_{\mathfrak{s}}(x) - X_{\mathfrak{s}}(y)\| \le \exp\Big(\int_s^t \ell_{R_*}(r)\,\mathrm{d}r\Big)\|X_t(x) - X_t(y)\|.$$

Taking $s = 0$ gives

$$\|x - y\| \le \exp\Big(\int_0^t \ell_{R_*}(r)\,\mathrm{d}r\Big)\|X_t(x) - X_t(y)\|.$$

Equivalently,

$$\|X_t(x) - X_t(y)\| \ge \exp\Big(-\int_0^t \ell_{R_*}(r)\,\mathrm{d}r\Big)\|x - y\|.$$

Combining the two bounds proves the claim. □

Lemma B.10. *Let $(X_t)_{t\in[0,T]}$ be an absolutely continuous stochastic process in $\mathbb{R}^d$. Let $h \in C^1(\mathbb{R}^d)$ satisfy $\mathbb{E}\big[|h(X_0)|\big] < \infty$ and $\mathbb{E}\left[\int_0^T \left|\nabla h(X_t)^\top \dot{X}_t\right| \mathrm{d}t\right] < \infty$. Then the map $t \mapsto \mathbb{E}[h(X_t)]$ is absolutely continuous on $[0,T]$, and $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[h(X_t)] = \mathbb{E}\left[\nabla h(X_t)^\top \dot{X}_t\right]$ for a.e. $t \in [0,T]$.*

Proof. Since X_t is absolutely continuous and $h \in C^1(\mathbb{R}^d)$, the pathwise chain rule gives

$$h(X_t) - h(X_0) = \int_0^t \nabla h(X_s)^\top \dot{X}_s \mathrm{d}s$$

for every $t \in [0,T]$, almost surely. Taking expectations yields

$$\mathbb{E}[h(X_t)] - \mathbb{E}[h(X_0)] = \mathbb{E}\left[\int_0^t \nabla h(X_s)^\top \dot{X}_s \mathrm{d}s\right].$$

By the assumed integrability condition, Fubini's theorem applies, and therefore

$$\mathbb{E}[h(X_t)] - \mathbb{E}[h(X_0)] = \int_0^t \mathbb{E}\left[\nabla h(X_s)^\top \dot{X}_s\right] \mathrm{d}s.$$

Hence $t \mapsto \mathbb{E}[h(X_t)]$ is absolutely continuous. Differentiating the last display gives $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[h(X_t)] = \mathbb{E}\left[\nabla h(X_t)^\top \dot{X}_t\right]$ for a.e. $t \in [0,T]$. □

Proposition B.11. *Let the base kernel k be a translation invariant kernel $k(x,y) = \phi(\|x-y\|)$. Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is a four-times differentiable even function with locally bounded derivatives. Suppose there exists $M_0 > 0$ such that $\|\mathfrak{s}(x)\| \le M_0(1+\|x\|)$, $\|\mathbf{J}\mathfrak{s}(x)\|_F \le M_0(1+\|x\|)$ and that $\mathfrak{s}$ has locally bounded second order derivatives. Then, the Langevin Stein kernel k_π satisfies both Assumptions 1 and 2. If additionally, $\mathfrak{s}$ is Lipschitz and has a uniformly bounded Hessian, then the Langevin Stein kernel k_π also satisfies Assumption 3.*

Proof. Recall that the Stein kernel k_π is

$$k_\pi(x,y) := \mathfrak{s}(x)^\top \mathfrak{s}(y) k(x,y) + \mathfrak{s}(x)^\top \nabla_2 k(x,y) + \nabla_1 k(x,y)^\top \mathfrak{s}(y) + \nabla \cdot_1 \nabla_2 k(x,y).$$

Then,

$$\begin{aligned} |k_\pi(x,x)| &\le |\phi(0)|\|\mathfrak{s}(x)\|^2 + 2\|\mathfrak{s}(x)\||\phi'(0)| + d|\phi''(0)| \\ &\le \big(|\phi(0)|M_0^2 + 2M_0|\phi'(0)| + d|\phi''(0)|\big)(1+\|x\|^2). \end{aligned}$$

This proves that k_π satisfies the first growth condition in Assumption 1.

Denote $\mathbf{J}\mathfrak{s}(x) \in \mathbb{R}^{d\times d}$ as the Jacobian matrix of the score function. From the derivation in Appendix B.1 of [56], we have

$$\begin{aligned} \nabla_1\nabla_2 k_\pi(x,y) &= \big(\mathbf{J}\mathfrak{s}(x)^\top \mathfrak{s}(y)\big) \nabla_2 k(x,y)^\top + \mathfrak{s}(x)^\top \mathfrak{s}(y) \nabla_1\nabla_2 k(x,y) + \mathbf{J}\mathfrak{s}(y)^\top \mathbf{J}\mathfrak{s}(x) k(x,y) \\ &+ \nabla_1 k(x,y) \mathfrak{s}(x)^\top \mathbf{J}\mathfrak{s}(y) + \mathbf{H}_2 k(x,y) \mathbf{J}\mathfrak{s}(x) + \nabla_1 \mathbf{H}_2 k(x,y) \mathfrak{s}(x) + \nabla_2 \mathbf{H}_1 k(x,y) \mathfrak{s}(y) \\ &+ \mathbf{J}\mathfrak{s}(y)^\top \mathbf{H}_1 k(x,y) + \nabla_1\nabla_2 \operatorname{Tr}\left(\nabla_2\nabla_1 k(x,y)\right). \end{aligned}$$

Since ϕ is an even function, its odd derivatives at 0 equal to 0. So we have

$$\begin{aligned} \nabla_1\nabla_2 k_\pi(x,x) &= \|\mathfrak{s}(x)\|^2 \nabla_1\nabla_2 k(x,x) + \mathbf{J}\mathfrak{s}(x)^\top \mathbf{J}\mathfrak{s}(x) k(x,x) + 2\mathbf{H}_2 k(x,x) \mathbf{J}\mathfrak{s}(x) \\ &+ \nabla_1\nabla_2 \operatorname{Tr}\left(\nabla_2\nabla_1 k(x,x)\right) \\ &= -\|\mathfrak{s}(x)\|^2 \phi''(0)\mathrm{Id} + \mathbf{J}\mathfrak{s}(x)^\top \mathbf{J}\mathfrak{s}(x)\phi(0) + 2\phi''(0)\mathbf{J}\mathfrak{s}(x) + 4(d+2)\phi^{(4)}(0)\mathrm{Id}. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} \|\nabla_1\nabla_2 k_\pi(x,x)\|_F &\le d|\phi''(0)|\|\mathfrak{s}(x)\|^2 + |\phi(0)|\|\mathbf{J}\mathfrak{s}(x)\|_F^2 + 2|\phi''(0)|\|\mathbf{J}\mathfrak{s}(x)\|_F + 4(d+2)|\phi^{(4)}(0)| \\ &\le (|\phi''(0)|M_0^2 + |\phi(0)|M_0^2 + 2|\phi''(0)|M_0 + 4(d+2)|\phi^{(4)}(0)|)(1+\|x\|^2). \end{aligned}$$

This proves that k_π satisfies the second growth condition in Assumption 1. Next, since $\mathfrak{s}, \mathbf{J}\mathfrak{s}, \nabla^2\mathfrak{s}$ and $\phi, \phi', \phi'', \phi^{(4)}$ are all locally bounded, Assumption 2 is also satisfied.

The second claim is proved in Lemma 1 of [56], specifically, in Eq. (40)-(42). □

C Convergence Analysis of the Particle Implementation of SrMMD

For any $x \in \mathbb{R}^d$, we define the pointwise derivative operator $\mathscr{D}_x : \mathcal{H} \to \mathbb{R}^d$, $f \mapsto \nabla f(x)$, whose adjoint is $\mathscr{D}_x^* u = \sum_{i=1}^d u_i \partial_{1,i} k(x,\cdot)$. The derivative covariance operator S_μ can be written as $S_\mu = \int \mathscr{D}_x^* \mathscr{D}_x \mathrm{d}\mu(x)$. Throughout this section, $\mu_0, \pi \in \mathcal{P}_2(\mathbb{R}^d)$ and $\lambda > 0$. Under Assumption 4, the Newton–Leibniz formula gives $\|\partial_{1,i}k(x,\cdot) - \partial_{1,i}k(y,\cdot)\|_{\mathcal{H}} \le B\sum_{j=1}^d |x_j - y_j| \le \sqrt{d}B\|x-y\|$. Summing over i yields $\|\mathscr{D}_x - \mathscr{D}_y\|_{\mathrm{HS}} \le dB\|x-y\|$. Thus $\|\mathscr{D}_x\|_{\mathrm{HS}} \le \|\mathscr{D}_0\|_{\mathrm{HS}} + dB\|x\|$, and integration from 0 to x gives $\|k(x,\cdot)\|_{\mathcal{H}} \le \|k(0,\cdot)\|_{\mathcal{H}} + \|\mathscr{D}_0\|_{\mathrm{op}}\|x\| + dB\|x\|^2/2$. Since $\|\mathscr{D}_x^*\mathscr{D}_x\|_{\mathrm{HS}} \le \|\mathscr{D}_x\|_{\mathrm{HS}}^2$, the quadratic growth of the kernel features and derivative covariance ensures that m_μ and S_μ are well-defined for $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. The linear growth of the derivative evaluation operator implies at most linear growth of $\nabla f_{\mu,\pi}$, so each population update belongs to $\mathcal{P}_2(\mathbb{R}^d)$.

First, we show that the uniform fourth-moment bound and the global derivative bound in Assumption 4 imply the required moment bounds for the kernel and its mixed second derivatives.

Lemma C.1. *Under Assumption 4, there exists a constant $\sigma > 0$ such that, for all $s \ge 0$,*

$$\int \|k(x,\cdot) - m_{\mu_s}\|_{\mathcal{H}}^2 \,\mathrm{d}\mu_s(x) \le \sigma^2, \quad \sum_{i,j=1}^d \int |\partial_{1,i}\partial_{2,j}k(x,x)|^2 \,\mathrm{d}\mu_s(x) \le \sigma^2.$$

Proof. Set $R_4 = \sup_{s\geq 0}\int \|x\|^4 \,\mathrm{d}\mu_s(x) < \infty$ and $A = \|\mathscr{D}_0\|_{\mathrm{HS}}$. By Item 1 and the Newton–Leibniz formula, $\|\mathscr{D}_x\|_{\mathrm{HS}} \leq A + dB\|x\|$ and $\|k(x,\cdot) - k(0,\cdot)\|_{\mathcal{H}} \leq A\|x\| + dB\|x\|^2/2$. Cauchy–Schwarz inequality therefore give $\int \|k(x,\cdot) - m_{\mu_s}\|_{\mathcal{H}}^2 \,\mathrm{d}\mu_s(x) \leq \int \|k(x,\cdot) - k(0,\cdot)\|_{\mathcal{H}}^2 \,\mathrm{d}\mu_s(x) \leq 2A^2 R_4^{1/2} + \frac{d^2B^2}{2} R_4$. By the reproducing property and $\|\mathscr{D}_x^*\mathscr{D}_x\|_{\mathrm{HS}} \leq \|\mathscr{D}_x\|_{\mathrm{HS}}^2$, $\sum_{i,j=1}^d |\partial_{1,i}\partial_{2,j}k(x,x)|^2 = \|\mathscr{D}_x^*\mathscr{D}_x\|_{\mathrm{HS}}^2 \leq \|\mathscr{D}_x\|_{\mathrm{HS}}^4 \leq 8A^4 + 8d^4B^4\|x\|^4$, where the last inequality uses $(a+b)^4 \leq 8(a^4+b^4)$ for $a, b \geq 0$. Integrating this bound and choosing a common $\sigma^2 \geq 1$ proves the claim. □

Let $\sigma > 0$ be the constant in Lemma C.1, and choose the constant C in Theorem 5.2 to be at least $e \vee \sigma^2$.

Step 1: Bounding the one-step propagation error and proving monotonicity. First, we control the one-step error propagation of SrMMD descent.

Proposition C.2. *Suppose Assumption 4 holds, and let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ satisfy $\mathrm{MMD}(\mu,\pi) \leq \mathrm{MMD}(\mu_0,\pi) + 1$. Let $0 < t \leq 1/2$ and $4dBt(\mathrm{MMD}(\mu_0,\pi)+1) \leq \lambda$. Set $\rho_t = (\mathrm{Id} - t\nabla f_{\mu,\pi})_{\#}\mu$. Then we have*

$$\mathrm{MMD}^2(\rho_t,\pi) \leq (1-t)\,\mathrm{MMD}^2(\mu,\pi) + t\lambda\langle m_\mu - m_\pi, f_{\mu,\pi}\rangle_{\mathcal{H}}.$$

Proof. Since S_μ is positive semidefinite and $0 \leq u/(u+\lambda) \leq 1$ for $u \geq 0$, spectral calculus gives $\|S_\mu f_{\mu,\pi}\|_{\mathcal{H}}^2 = \langle S_\mu(S_\mu + \lambda\mathrm{Id})^{-1}(m_\mu - m_\pi), S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}} \leq \langle m_\mu - m_\pi, S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}}$ and

$$\|\nabla f_{\mu,\pi}\|_{L_2^d(\mu)}^2 = \langle (S_\mu + \lambda\mathrm{Id})^{-1}(m_\mu - m_\pi), S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}} \leq \lambda^{-1}\langle m_\mu - m_\pi, S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}}. \tag{C.1}$$

Write $m_\mu - m_\pi = \sum_{i=0}^\infty m_i e_i$, where $\{\varrho_i, e_i\}_{i\in\mathbb{N}}$ is an orthonormal eigensystem of S_μ. Then

$$\begin{aligned}
\|\nabla f_{\mu,\pi}\|_{L_2^d(\mu)}^2 &= \langle (S_\mu + \lambda\mathrm{Id})^{-1}(m_\mu - m_\pi), S_\mu(S_\mu + \lambda\mathrm{Id})^{-1}(m_\mu - m_\pi)\rangle_{\mathcal{H}} \\
&= \sum_{i=0}^\infty \frac{\varrho_i}{(\varrho_i+\lambda)^2} m_i^2 \overset{\text{(a)}}{\leq} \frac{1}{4\lambda}\sum_{i=0}^\infty m_i^2 \\
&= \frac{1}{4\lambda}\,\mathrm{MMD}^2(\mu,\pi) \leq \frac{1}{4\lambda}(\mathrm{MMD}(\mu_0,\pi)+1)^2,
\end{aligned} \tag{C.2}$$

where (a) uses $u/(u+\lambda)^2 = (u + \frac{\lambda^2}{u} + 2\lambda)^{-1} \leq (4\lambda)^{-1}$ at positive eigenvalues. At zero the ratio in Eq. (C.2) is zero, so the inequality holds for every $u \geq 0$. By Taylor's formula with integral remainder and the definition of push-forward,

$$\begin{aligned}
&m_{\rho_t} - m_\pi \\
&= \int k(x,\cdot)\mathrm{d}(\mathrm{Id} - t\nabla f_{\mu,\pi})_{\#}\mu(x) - m_\pi \\
&= \int k(x - t\nabla f_{\mu,\pi}(x),\cdot)\mathrm{d}\mu(x) - m_\pi \\
&= \int \Bigg(k(x,\cdot) - t\sum_{i=1}^d \partial_i f_{\mu,\pi}(x)\partial_{1,i}k(x,\cdot) \\
&\quad + t^2 \sum_{i,j=1}^d \partial_i f_{\mu,\pi}(x)\partial_j f_{\mu,\pi}(x)\int_0^1 (1-u)\partial_{1,i}\partial_{1,j}k(x - ut\nabla f_{\mu,\pi}(x),\cdot)\mathrm{d}u\Bigg)\mathrm{d}\mu(x) - m_\pi \\
&= m_\mu - m_\pi - t\sum_{i=1}^d \int \partial_i f_{\mu,\pi}(x)\partial_{1,i}k(x,\cdot)\mathrm{d}\mu(x) \\
&\quad + t^2 \sum_{i,j=1}^d \int \partial_i f_{\mu,\pi}(x)\partial_j f_{\mu,\pi}(x)\int_0^1 (1-u)\partial_{1,i}\partial_{1,j}k(x - ut\nabla f_{\mu,\pi}(x),\cdot)\mathrm{d}u\mathrm{d}\mu(x) \\
&= m_\mu - m_\pi - tS_\mu f_{\mu,\pi} \\
&\quad + t^2 \sum_{i,j=1}^d \int \partial_i f_{\mu,\pi}(x)\partial_j f_{\mu,\pi}(x)\int_0^1 (1-u)\partial_{1,i}\partial_{1,j}k(x - ut\nabla f_{\mu,\pi}(x),\cdot)\mathrm{d}u\mathrm{d}\mu(x),
\end{aligned}$$

where the last equality follows from the reproducing property and the definition of S_μ, since

$$\sum_{i=1}^{d} \int \partial_i f_{\mu,\pi}(x) \partial_{1,i} k(x,\cdot) \mathrm{d}\mu(x) = \sum_{i=1}^{d} \int \langle f_{\mu,\pi}, \partial_{1,i} k(x,\cdot)\rangle_{\mathcal{H}} \partial_{1,i} k(x,\cdot) \mathrm{d}\mu(x) = S_\mu f_{\mu,\pi}.$$

Therefore, by Item 1,

$$\begin{aligned}
&\|m_{\rho_t} - m_\mu + t S_\mu f_{\mu,\pi}\|_{\mathcal{H}} \\
&\le t^2 \sum_{i,j=1}^{d} \int |\partial_i f_{\mu,\pi}(x)||\partial_j f_{\mu,\pi}(x)| \int_0^1 (1-u) \|\partial_{1,i}\partial_{1,j} k(x - ut\nabla f_{\mu,\pi}(x), \cdot)\|_{\mathcal{H}} \mathrm{d}u \mathrm{d}\mu(x) \\
&\le \frac{Bt^2}{2} \int \sum_{i,j=1}^{d} |\partial_i f_{\mu,\pi}(x)||\partial_j f_{\mu,\pi}(x)| \mathrm{d}\mu(x) = \frac{Bt^2}{2} \int \left(\sum_{i=1}^{d} |\partial_i f_{\mu,\pi}(x)| \right)^2 \mathrm{d}\mu(x) \\
&\le \frac{dBt^2}{2} \int \|\nabla f_{\mu,\pi}(x)\|^2 \mathrm{d}\mu(x) = \frac{dBt^2}{2} \|\nabla f_{\mu,\pi}\|^2_{L_2^d(\mu)}.
\end{aligned}$$

The last inequality follows from Cauchy–Schwarz. Eq. (C.2) ensures the square integrability of $\nabla f_{\mu,\pi}$ and hence the integrability of the Taylor remainder. The linear term is integrable because $\int \|\mathscr{D}_x\|^2_{\mathrm{op}} \mathrm{d}\mu(x) < \infty$ and Cauchy–Schwarz applies to $\|\mathscr{D}_x\|_{\mathrm{op}} \|\nabla f_{\mu,\pi}(x)\|$. Thus Taylor's formula can be integrated as an identity in $\mathcal{H}$. Moreover,

$$\begin{aligned}
&\|m_\mu - m_\pi - t S_\mu f_{\mu,\pi}\|^2_{\mathcal{H}} \\
&= \|m_\mu - m_\pi\|^2_{\mathcal{H}} + \langle S_\mu f_{\mu,\pi}, -2t(m_\mu - m_\pi) + t^2 S_\mu (S_\mu + \lambda \mathrm{Id})^{-1}(m_\mu - m_\pi)\rangle_{\mathcal{H}} \\
&\le \mathrm{MMD}^2(\mu,\pi) + \langle S_\mu f_{\mu,\pi}, -2t(m_\mu - m_\pi) + t^2 (m_\mu - m_\pi)\rangle_{\mathcal{H}} \\
&= \mathrm{MMD}^2(\mu,\pi) - t(2-t)\langle S_\mu f_{\mu,\pi}, m_\mu - m_\pi\rangle_{\mathcal{H}}.
\end{aligned}$$

In particular, $\|m_\mu - m_\pi - t S_\mu f_{\mu,\pi}\|_{\mathcal{H}} \le \mathrm{MMD}(\mu,\pi) \le \mathrm{MMD}(\mu_0,\pi) + 1$. Adding the integral remainder to the linear Taylor polynomial, expanding the squared norm, and applying Cauchy–Schwarz to the cross term gives, with $\varepsilon_0 = \mathrm{MMD}(\mu_0,\pi)$,

$$\begin{aligned}
&\mathrm{MMD}^2(\rho_t,\pi) \\
&\le \mathrm{MMD}^2(\mu,\pi) - t(2-t)\langle m_\mu - m_\pi, S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}} \\
&\quad + dBt^2(\varepsilon_0+1)\|\nabla f_{\mu,\pi}\|^2_{L_2^d(\mu)} + \frac{(dB)^2 t^4}{4} \|\nabla f_{\mu,\pi}\|^4_{L_2^d(\mu)} \\
&\le \mathrm{MMD}^2(\mu,\pi) - t(2 - t - \frac{dBt(\varepsilon_0+1)}{\lambda} - \frac{d^2B^2t^3(\varepsilon_0+1)^2}{16\lambda^2})\langle m_\mu - m_\pi, S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}} \\
&\le \mathrm{MMD}^2(\mu,\pi) - t\langle m_\mu - m_\pi, S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}} = (1-t)\,\mathrm{MMD}^2(\mu,\pi) + t\lambda\langle m_\mu - m_\pi, f_{\mu,\pi}\rangle_{\mathcal{H}}.
\end{aligned}$$

For the second inequality, Eq. (C.1) controls the squared velocity norm, and the product of Eqs. (C.1) and (C.2) controls its square. For the third inequality, the factor in parentheses is at least $2 - \frac{1}{2} - \frac{1}{4} - \frac{1}{512} > 1$ by the assumptions on t and λ. The last equality follows from $S_\mu f_{\mu,\pi} = m_\mu - m_\pi - \lambda f_{\mu,\pi}$. Since $\langle m_\mu - m_\pi, S_\mu f_{\mu,\pi}\rangle_{\mathcal{H}} \ge 0$, the final inequality also gives $\mathrm{MMD}(\rho_t,\pi) \le \mathrm{MMD}(\mu,\pi)$. □

Lemma C.3. *Under the assumptions of Theorem 5.2,* $\mathrm{MMD}(\mu_s,\pi) \le \mathrm{MMD}(\mu_0,\pi)$ *for every* $s \ge 0$*, and, for* $0 \le s \le S$*,*

$$\mathrm{MMD}(\mu_s,\pi) \le \left(\mathrm{MMD}^2(\mu_0,\pi) + \lambda^\alpha Q\right)^{1/2} \left(1 + \frac{\gamma \max\{s-1,0\}}{\alpha}\right)^{-\alpha/2}.$$

Proof. Applying Proposition C.2 with $t = \gamma$ proves monotonicity by induction: $\mu_{s+1} = (\mathrm{Id} - \gamma\nabla f_{\mu_s,\pi})_{\#}\mu_s$, and bounding $\mathrm{MMD}(\mu_s,\pi)$ by its initial value verifies the proposition's hypothesis at each iteration. If $\mathrm{MMD}(\mu_s,\pi) = 0$ at some iteration, then $f_{\mu_s,\pi} = 0$ and every subsequent population measure equals μ_s, which gives the lemma's conclusion at those iterations. For iterations with a positive residual, we use the source condition as follows.

Recall that $m_{\mu_s} - m_\pi = \sum_{i=0}^{\infty} m_{i,s} e_{i,s}$, where $\{\varrho_{i,s}, e_{i,s}\}_{i\in\mathbb{N}}$ is an orthonormal eigensystem of S_{μ_s}. For $1 \le s \le S$, the source condition $m_{\mu_s} - m_\pi = S_{\mu_s}^{\alpha/2} q_s$ implies that $m_{i,s} = 0$ whenever $\varrho_{i,s} = 0$, and $\sum_{\varrho_{i,s}>0} \varrho_{i,s}^{-\alpha} m_{i,s}^2 \le \|q_s\|_{\mathcal{H}}^2 \le Q$. We obtain

$$\begin{aligned}
\mathrm{MMD}^2(\mu_s, \pi) &= \sum_{\varrho_{i,s}>0} m_{i,s}^2 = \sum_{\varrho_{i,s}>0} \left(\frac{\varrho_{i,s}}{\varrho_{i,s}+\lambda} m_{i,s}^2\right)^{\frac{\alpha}{1+\alpha}} \left(\left(\frac{\varrho_{i,s}+\lambda}{\varrho_{i,s}}\right)^{\alpha} m_{i,s}^2\right)^{\frac{1}{1+\alpha}} \\
&\le \left(\sum_{\varrho_{i,s}>0} \frac{\varrho_{i,s}}{\varrho_{i,s}+\lambda} m_{i,s}^2\right)^{\frac{\alpha}{1+\alpha}} \left(\sum_{\varrho_{i,s}>0} \left(\frac{\varrho_{i,s}+\lambda}{\varrho_{i,s}}\right)^{\alpha} m_{i,s}^2\right)^{\frac{1}{1+\alpha}} \\
&\le \left(\sum_{\varrho_{i,s}>0} \frac{\varrho_{i,s}}{\varrho_{i,s}+\lambda} m_{i,s}^2\right)^{\frac{\alpha}{1+\alpha}} \left(\sum_{\varrho_{i,s}>0} (\varrho_{i,s}^{\alpha}+\lambda^{\alpha}) \varrho_{i,s}^{-\alpha} m_{i,s}^2\right)^{\frac{1}{1+\alpha}} \\
&\le \langle m_{\mu_s} - m_\pi, S_{\mu_s} f_{\mu_s,\pi}\rangle_{\mathcal{H}}^{\frac{\alpha}{1+\alpha}} \left(\mathrm{MMD}^2(\mu_s,\pi) + \lambda^{\alpha} Q\right)^{\frac{1}{1+\alpha}}.
\end{aligned}$$

The first inequality is Hölder's inequality with exponents $(1+\alpha)/\alpha$ and $1+\alpha$. The second uses $(u+\lambda)^{\alpha} \le u^{\alpha} + \lambda^{\alpha}$ for $u \ge 0$ and $0 < \alpha \le 1$. The final inequality follows from the source condition in Theorem 5.2. Taking $t = \gamma$ and $\mu = \mu_s$ in Proposition C.2, solving the Hölder inequality for the inner product, and replacing the population MMD in its denominator by its initial value gives, for $1 \le s < S$,

$$\mathrm{MMD}^2(\mu_{s+1}, \pi) \le \mathrm{MMD}^2(\mu_s, \pi) - \gamma \frac{\mathrm{MMD}^{2+2/\alpha}(\mu_s, \pi)}{\left(\mathrm{MMD}^2(\mu_0, \pi) + \lambda^{\alpha} Q\right)^{1/\alpha}}.$$

We may use $(1-u)^{-1/\alpha} \ge 1 + u/\alpha$ for $0 \le u < 1$, or equivalently $1 - u \le (1 + u/\alpha)^{-\alpha}$, to obtain

$$\begin{aligned}
\mathrm{MMD}^2(\mu_{s+1}, \pi) &\le \mathrm{MMD}^2(\mu_s, \pi) \left(1 - \gamma \frac{\mathrm{MMD}^{2/\alpha}(\mu_s, \pi)}{\left(\mathrm{MMD}^2(\mu_0, \pi) + \lambda^{\alpha} Q\right)^{1/\alpha}}\right) \\
&\le \mathrm{MMD}^2(\mu_s, \pi) \left(1 + \gamma \frac{\mathrm{MMD}^{2/\alpha}(\mu_s, \pi)}{\alpha\left(\mathrm{MMD}^2(\mu_0, \pi) + \lambda^{\alpha} Q\right)^{1/\alpha}}\right)^{-\alpha}.
\end{aligned}$$

Whenever $\mathrm{MMD}(\mu_{s+1}, \pi) > 0$, the last inequality is equivalent to

$$\mathrm{MMD}^{-2/\alpha}(\mu_{s+1}, \pi) \ge \mathrm{MMD}^{-2/\alpha}(\mu_s, \pi) + \frac{\gamma}{\alpha(\mathrm{MMD}^2(\mu_0, \pi) + \lambda^{\alpha} Q)^{1/\alpha}}. \tag{C.3}$$

We repeat the recursion Eq. (C.3) from 1 to $s-1$. Monotonicity gives $\mathrm{MMD}^{-2/\alpha}(\mu_1, \pi) \ge (\mathrm{MMD}^2(\mu_0, \pi) + \lambda^{\alpha} Q)^{-1/\alpha}$, which supplies the initial term of this sum and proves the result for $1 \le s \le S$. □

Step 2: Establishing sampling bounds via auxiliary particles. Define the auxiliary particles by

$$\overline{x}_{s+1}^{(i)} = \overline{x}_s^{(i)} - \gamma \nabla f_{\mu_s,\pi}(\overline{x}_s^{(i)}), \quad \overline{x}_0^{(i)} = x_0^{(i)}.$$

Define the associated empirical distribution $\overline{\mu}_s = N^{-1} \sum_{i=1}^{N} \delta_{\overline{x}_s^{(i)}}$. For each fixed iteration s, the auxiliary particles are independent with law μ_s, since they are obtained from the independent initial samples by the same deterministic population updates. Therefore, by Lemma C.1,

$$\mathbb{E}\left[\|m_{\overline{\mu}_s} - m_{\mu_s}\|_{\mathcal{H}}^2\right] \vee \mathbb{E}\left[\|S_{\overline{\mu}_s} - S_{\mu_s}\|_{\mathrm{HS}}^2\right] \le \frac{\sigma^2}{N}.$$

Indeed, expanding the squared norm of either centered sample mean eliminates the cross terms by independence and gives $1/N$ times the centered second moment of one summand. For the operator term, the reproducing property gives $\|\mathscr{D}_x^* \mathscr{D}_x\|_{\mathrm{HS}}^2 = \sum_{i,j=1}^{d} |\partial_{1,i}\partial_{2,j} k(x,x)|^2$, and its centered second moment is $\int \|\mathscr{D}_x^* \mathscr{D}_x - S_{\mu_s}\|_{\mathrm{HS}}^2 \mathrm{d}\mu_s(x) = \int \|\mathscr{D}_x^* \mathscr{D}_x\|_{\mathrm{HS}}^2 \mathrm{d}\mu_s(x) - \|S_{\mu_s}\|_{\mathrm{HS}}^2 \le \sigma^2$.

Fix $0 < \delta < 1$, and $K \in \mathbb{N}$ to be specified in Step 4, and set $\xi = \sqrt{2(K+1)/(N\delta)}$. For each $0 \le s \le K$, Chebyshev's inequality gives $\mathbb{P}\left[\|m_{\overline{\mu}_s} - m_{\mu_s}\|_{\mathcal{H}} > \sigma\xi\right] \le 1/(N\xi^2)$ and $\mathbb{P}\left[\|S_{\overline{\mu}_s} - S_{\mu_s}\|_{\mathrm{HS}} > \sigma\xi\right] \le 1/(N\xi^2)$. The union bound gives

$$\mathbb{P}\left[\bigcup_{s=0}^{K}\{\|m_{\overline{\mu}_s} - m_{\mu_s}\|_{\mathcal{H}} > \sigma\xi\} \cup \bigcup_{s=0}^{K}\{\|S_{\overline{\mu}_s} - S_{\mu_s}\|_{\mathrm{HS}} > \sigma\xi\}\right] \le \frac{2(K+1)}{N\xi^2} = \delta. \tag{C.4}$$

On the complement of the union in Eq. (C.4), which has probability at least $1-\delta$, both $\|m_{\overline{\mu}_s} - m_{\mu_s}\|_{\mathcal{H}} \le \sigma\xi$ and $\|S_{\overline{\mu}_s} - S_{\mu_s}\|_{\mathrm{op}} \le \sigma\xi$ hold for every $0 \le s \le K$; the second inequality uses $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_{\mathrm{HS}}$. We compare the two particle systems on this complement.

For the comparison up to the chosen iteration, assume $\xi \le 1$ and $\sigma\xi \le (K+1)^{-\alpha/2}/4$. Step 4 relates K and N so that these two inequalities hold. Since $\hat{\mu}_0 = \overline{\mu}_0$, Eq. (C.4) and the triangle inequality give $\mathrm{MMD}(\hat{\mu}_0, \pi) \le \mathrm{MMD}(\mu_0, \pi) + 1$ on the complement of the union. Applying Proposition C.2 with $t = \gamma$ to successive empirical measures shows by induction that $\mathrm{MMD}(\hat{\mu}_s, \pi)$ is nonincreasing for all $s \ge 0$. Moreover, Jensen's inequality and Lemma C.1 give $\|S_{\mu_s}\|_{\mathrm{op}} \le \|S_{\mu_s}\|_{\mathrm{HS}} \le \sigma$, so $\xi \le 1$ implies $\|S_{\overline{\mu}_s}\|_{\mathrm{op}} \le 2\sigma$ for $s \le K$ on the complement of the union in Eq. (C.4).

Step 3: Establishing the error recursion. Let $\mathsf{e}_s = \left(\frac{1}{N}\sum_{i=1}^{N}\|x_s^{(i)} - \overline{x}_s^{(i)}\|^2\right)^{1/2}$. For a fixed iteration s, consider the interpolated particles defined by $x_t^{(i)} = (1-t)\overline{x}_s^{(i)} + t x_s^{(i)}$ and $\nu_t = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_t^{(i)}}$, where $0 \le t \le 1$. The induction hypothesis for $0 \le s \le S \wedge K$ is $\mathsf{e}_s \le 1 \wedge \frac{(K+1)^{-\alpha/2}}{4(\sqrt{2}\sigma + dB/2)}$. Taylor's formula at each $\overline{x}_s^{(i)}$, using the integral remainder in the proof of Proposition C.2, yields

$$\begin{aligned}
\|m_{\nu_t} - m_{\overline{\mu}_s}\|_{\mathcal{H}} &= \frac{1}{N}\left\|\sum_{i=1}^{N}\left(k(\overline{x}_s^{(i)} + t(x_s^{(i)} - \overline{x}_s^{(i)}), \cdot) - k(\overline{x}_s^{(i)}, \cdot)\right)\right\|_{\mathcal{H}} \\
&\le \left\|\frac{t}{N}\sum_{i=1}^{N}\mathscr{D}^*_{\overline{x}_s^{(i)}}(x_s^{(i)} - \overline{x}_s^{(i)})\right\|_{\mathcal{H}} + \frac{dB}{2}t^2\mathsf{e}_s^2 \\
&\le \sqrt{2}\sigma t\mathsf{e}_s + \frac{dB}{2}t^2\mathsf{e}_s^2 \le \frac{(K+1)^{-\alpha/2}}{4}.
\end{aligned}$$

The linear term is estimated by duality and Cauchy–Schwarz in the particle index, using the operator norm of the empirical derivative covariance from Eq. (C.4) and Lemma C.1. The remainder is bounded by $dBt^2\mathsf{e}_s^2/2$ by Item 1. The last inequality uses $t \le 1$, $\mathsf{e}_s \le 1$, and the induction hypothesis $\mathsf{e}_s \le 1 \wedge \frac{(K+1)^{-\alpha/2}}{4(\sqrt{2}\sigma + dB/2)}$. The triangle inequality and Eq. (C.4), on the complement of its union, then give $\mathrm{MMD}(\nu_t, \pi) \le \mathrm{MMD}(\mu_s, \pi) + (K+1)^{-\alpha/2}/2$ for every $0 \le t \le 1$.

Lemma C.4. *Under Assumption 4, along the straight-line interpolation between the auxiliary and interacting particles, for $0 \le \gamma \le 1$, we have*

$$\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\left\{x_t^{(i)} - \gamma\nabla f_{\nu_t,\pi}(x_t^{(i)})\right\}\right\|^2\right)^{1/2} \le \left(1 + \frac{5\gamma dB}{4\lambda}\,\mathrm{MMD}(\nu_t, \pi)\right)\mathsf{e}_s.$$

Proof. Direct differentiation gives $\frac{\mathrm{d}}{\mathrm{d}t}f_{\nu_t,\pi} = (S_{\nu_t} + \lambda\mathrm{Id})^{-1}(\frac{\mathrm{d}}{\mathrm{d}t}m_{\nu_t} - \frac{\mathrm{d}}{\mathrm{d}t}S_{\nu_t}f_{\nu_t,\pi})$, where

$$\frac{\mathrm{d}}{\mathrm{d}t}m_{\nu_t} = \frac{1}{N}\sum_{j=1}^{N}\sum_{j=1}^{N}\mathscr{D}_{x_t^{(j)}}(x_s^{(j)} - \overline{x}_s^{(j)}), \quad \frac{\mathrm{d}}{\mathrm{d}t}S_{\nu_t} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\mathrm{d}\mathscr{D}^*_{x_t^{(j)}}}{\mathrm{d}t}\mathscr{D}_{x_t^{(j)}} + \mathscr{D}_{x_t^{(j)}}\frac{\mathrm{d}\mathscr{D}_{x_t^{(j)}}}{\mathrm{d}t}\right).$$

The chain rule then gives

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left\{x_t^{(i)} - \gamma\nabla f_{\nu_t,\pi}(x_t^{(i)})\right\} \\
&= x_s^{(i)} - \overline{x}_s^{(i)} - \frac{\gamma}{N}\sum_{j=1}^{N} \mathscr{D}_{x_t^{(i)}}(S_{\nu_t} + \lambda\mathrm{Id})^{-1}\mathscr{D}^*_{x_t^{(j)}}(x_s^{(j)} - \overline{x}_s^{(j)}) \\
&\quad - \gamma\frac{\mathrm{d}\mathscr{D}_{x_t^{(i)}}}{\mathrm{d}t} f_{\nu_t,\pi} + \frac{\gamma}{N}\sum_{j=1}^{N} \mathscr{D}_{x_t^{(i)}}(S_{\nu_t} + \lambda\mathrm{Id})^{-1}\mathscr{D}^*_{x_t^{(j)}}\frac{\mathrm{d}\mathscr{D}_{x_t^{(j)}}}{\mathrm{d}t} f_{\nu_t,\pi} \\
&\quad + \frac{\gamma}{N}\sum_{j=1}^{N} \mathscr{D}_{x_t^{(i)}}(S_{\nu_t} + \lambda\mathrm{Id})^{-1}\frac{\mathrm{d}\mathscr{D}^*_{x_t^{(j)}}}{\mathrm{d}t}\mathscr{D}_{x_t^{(j)}} f_{\nu_t,\pi}.
\end{aligned}$$

By Minkowski's inequality,

$$\begin{aligned}
&\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\left\{x_t^{(i)} - \gamma\nabla f_{\nu_t,\pi}(x_t^{(i)})\right\}\right\|^2\right\}^{1/2} \\
&\le \underbrace{\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|x_s^{(i)} - \overline{x}_s^{(i)} - \frac{\gamma}{N}\sum_{j=1}^{N} \mathscr{D}_{x_t^{(i)}}(S_{\nu_t} + \lambda\mathrm{Id})^{-1}\mathscr{D}^*_{x_t^{(j)}}(x_s^{(j)} - \overline{x}_s^{(j)})\right\|^2\right\}^{1/2}}_{\text{(I)}} \\
&\quad + \underbrace{\gamma\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|-\frac{\mathrm{d}\mathscr{D}_{x_t^{(i)}}}{\mathrm{d}t} f_{\nu_t,\pi} + \frac{1}{N}\sum_{j=1}^{N} \mathscr{D}_{x_t^{(i)}}(S_{\nu_t} + \lambda\mathrm{Id})^{-1}\mathscr{D}^*_{x_t^{(j)}}\frac{\mathrm{d}\mathscr{D}_{x_t^{(j)}}}{\mathrm{d}t} f_{\nu_t,\pi}\right\|^2\right\}^{1/2}}_{\text{(II)}} \\
&\quad + \underbrace{\gamma\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\mathscr{D}_{x_t^{(i)}}(S_{\nu_t} + \lambda\mathrm{Id})^{-1}\frac{1}{N}\sum_{j=1}^{N}\frac{\mathrm{d}\mathscr{D}^*_{x_t^{(j)}}}{\mathrm{d}t}\mathscr{D}_{x_t^{(j)}} f_{\nu_t,\pi}\right\|^2\right\}^{1/2}}_{\text{(III)}}.
\end{aligned}$$

We first estimate (I). With the operator D_X introduced in Appendix A.5 at $x^{(i)} = x_t^{(i)}$, the $Nd \times Nd$ block matrix whose (i,j) block is $N^{-1}\mathscr{D}_{x_t^{(i)}}(S_{\nu_t}+\lambda\mathrm{Id})^{-1}\mathscr{D}^*_{x_t^{(j)}}$ equals $D_X(D_X^*D_X+\lambda\mathrm{Id})^{-1}D_X^* = D_XD_X^*(D_XD_X^* + \lambda\mathrm{Id})^{-1}$. Since $D_XD_X^*$ is positive semidefinite, the eigenvalues of this matrix are $u/(u+\lambda) \in [0,1]$ for $u \ge 0$. As $0 \le \gamma \le 1$, the matrix $\mathrm{Id} - \gamma D_X(D_X^*D_X + \lambda\mathrm{Id})^{-1}D_X^*$ also has eigenvalues in $[0,1]$ and hence operator norm at most one. Therefore, (I) $\le \mathsf{e}_s$.

For (II), Item 1 of Assumption 4 and the reproducing property give $\|\frac{\mathrm{d}\mathscr{D}_{x_t^{(i)}}}{\mathrm{d}t}\|_{\mathrm{op}} \le dB\|x_s^{(i)} - \overline{x}_s^{(i)}\|$. Indeed, differentiating $\mathscr{D}_{x_t^{(i)}} f = \nabla f(x_t^{(i)})$ along $t \mapsto x_t^{(i)}$ gives a sum involving $\partial_{1,k}\partial_{1,l}k(x_t^{(i)}, \cdot)$, whose $\mathcal{H}$-norm is bounded by B. Cauchy–Schwarz bounds each component by $\sqrt{d}B\|x_s^{(i)} - \overline{x}_s^{(i)}\|\|f\|_{\mathcal{H}}$; summing its square over the d components gives the factor dB. The two terms in (II) are obtained by applying $-\gamma[\mathrm{Id} - D_X(D_X^*D_X + \lambda\mathrm{Id})^{-1}D_X^*]$ to the vector with i-th component $(\frac{\mathrm{d}\mathscr{D}_{x_t^{(i)}}}{\mathrm{d}t})f_{\nu_t,\pi}$. The operator norm of the matrix in brackets is at most one. Moreover, $\|f_{\nu_t,\pi}\|_{\mathcal{H}} \le \mathrm{MMD}(\nu_t,\pi)/\lambda$. Consequently,

$$\text{(II)} \le \gamma dB\mathsf{e}_s\|f_{\nu_t,\pi}\|_{\mathcal{H}} \le \frac{\gamma dB}{\lambda}\mathsf{e}_s\,\mathrm{MMD}(\nu_t,\pi).$$

For (III), the definition of S_{ν_t} and $\|S_{\nu_t}^{1/2}(S_{\nu_t}+\lambda\mathrm{Id})^{-1}\|_{\mathrm{op}} \le 1/(2\sqrt{\lambda})$ give

$$\text{(III)} \le \frac{\gamma}{2\sqrt{\lambda}} \left\| \frac{1}{N}\sum_{j=1}^{N} \frac{\mathrm{d}\mathscr{D}^*_{x_t^{(j)}}}{\mathrm{d}t} \mathscr{D}_{x_t^{(j)}} f_{\nu_t,\pi} \right\|_{\mathcal{H}}.$$

The operator norm estimate follows by applying spectral calculus to $u/(u+\lambda)^2 \le 1/(4\lambda)$ for $u \ge 0$. Applying Item 1 to the differentiated adjoints and then using Cauchy–Schwarz gives

$$\begin{aligned}
&\left\| \frac{1}{N}\sum_{j=1}^{N} \frac{\mathrm{d}\mathscr{D}^*_{x_t^{(j)}}}{\mathrm{d}t} \mathscr{D}_{x_t^{(j)}} f_{\nu_t,\pi} \right\|_{\mathcal{H}} \le \frac{dB}{N}\sum_{j=1}^{N} \|x_s^{(j)} - \bar{x}_s^{(j)}\| \|\mathscr{D}_{x_t^{(j)}} f_{\nu_t,\pi}\| \\
&\le dB\mathsf{e}_s \|S_{\nu_t}^{1/2} f_{\nu_t,\pi}\|_{\mathcal{H}} \le \frac{dB}{2\sqrt{\lambda}}\mathsf{e}_s \,\mathrm{MMD}(\nu_t,\pi).
\end{aligned}$$

The last inequality uses $f_{\nu_t,\pi} = (S_{\nu_t}+\lambda\mathrm{Id})^{-1}(m_{\nu_t} - m_\pi)$ and the spectral calculation in Eq. (C.2), applied to the interpolated empirical measure. Multiplying by the outer resolvent factor in (III) gives

$$\text{(III)} \le \frac{\gamma dB}{4\lambda}\mathsf{e}_s \,\mathrm{MMD}(\nu_t,\pi).$$

Combining the estimates for (I), (II), and (III), we obtain

$$\left\{ \frac{1}{N}\sum_{i=1}^{N} \left\| \frac{\mathrm{d}}{\mathrm{d}t}\left\{ x_t^{(i)} - \gamma\nabla f_{\nu_t,\pi}(x_t^{(i)}) \right\} \right\|^2 \right\}^{1/2} \le \left\{ 1 + \frac{5\gamma dB}{4\lambda}\,\mathrm{MMD}(\nu_t,\pi) \right\} \mathsf{e}_s. \tag{C.5}$$

□

Proof of Theorem 5.2. Fix a deterministic integer K in Eq. (C.4). On the complement of the union in that equation, we first compare the velocities generated by the two witness functions at the auxiliary particles. By the definition of the witness functions,

$$\begin{aligned}
(S_{\bar{\mu}_s}+\lambda\mathrm{Id})(f_{\bar{\mu}_s,\pi} - f_{\mu_s,\pi}) &= (m_{\bar{\mu}_s} - m_\pi) - (S_{\bar{\mu}_s}+\lambda\mathrm{Id})f_{\mu_s,\pi} \\
&= m_{\bar{\mu}_s} - m_{\mu_s} + (S_{\mu_s} - S_{\bar{\mu}_s})f_{\mu_s,\pi},
\end{aligned}$$

where the last equality follows from $(S_{\mu_s}+\lambda\mathrm{Id})f_{\mu_s,\pi} = m_{\mu_s} - m_\pi$. Consequently,

$$f_{\bar{\mu}_s,\pi} - f_{\mu_s,\pi} = (S_{\bar{\mu}_s}+\lambda\mathrm{Id})^{-1}\big(m_{\bar{\mu}_s} - m_{\mu_s} + (S_{\mu_s} - S_{\bar{\mu}_s})f_{\mu_s,\pi}\big).$$

By the definition of $S_{\bar{\mu}_s}$, the root mean squared difference between the two witness velocities at the auxiliary particles satisfies

$$\begin{aligned}
\left\{ \frac{1}{N}\sum_{i=1}^{N} \|\nabla f_{\bar{\mu}_s,\pi}(\bar{x}_s^{(i)}) - \nabla f_{\mu_s,\pi}(\bar{x}_s^{(i)})\|^2 \right\}^{1/2} &= \|S_{\bar{\mu}_s}^{1/2}(f_{\bar{\mu}_s,\pi} - f_{\mu_s,\pi})\|_{\mathcal{H}} \\
&\le \frac{\|m_{\bar{\mu}_s} - m_{\mu_s}\|_{\mathcal{H}} + \|S_{\bar{\mu}_s} - S_{\mu_s}\|_{\mathrm{op}}\|f_{\mu_s,\pi}\|_{\mathcal{H}}}{2\sqrt{\lambda}} \\
&\le \frac{\sigma\xi}{2\sqrt{\lambda}}\left\{ 1 + \frac{\mathrm{MMD}(\mu_0,\pi)}{\lambda} \right\}.
\end{aligned}$$

The first equality follows from $\nabla f(x) = \mathscr{D}_x f$ and $S_{\bar{\mu}_s} = N^{-1}\sum_{i=1}^{N} \mathscr{D}^*_{\bar{x}_s^{(i)}}\mathscr{D}_{\bar{x}_s^{(i)}}$. The first inequality uses $\|S_{\bar{\mu}_s}^{1/2}(S_{\bar{\mu}_s}+\lambda\mathrm{Id})^{-1}\|_{\mathrm{op}} \le 1/(2\sqrt{\lambda})$, which follows from $u/(u+\lambda)^2 \le 1/(4\lambda)$ for $u \ge 0$. The last inequality follows from Eq. (C.4) on the complement of its union and from $\|f_{\mu_s,\pi}\|_{\mathcal{H}} \le \mathrm{MMD}(\mu_s,\pi)/\lambda \le \mathrm{MMD}(\mu_0,\pi)/\lambda$, using Lemma C.3 in the final step.

To apply Eq. (C.5), we separate the difference between the interacting particle updates from the difference between the witness velocities evaluated at the auxiliary particles. Since $x_0^{(i)} = \bar{x}_0^{(i)}$, we have $\mathsf{e}_0 = 0$. The update definitions give

$$\begin{aligned}
&x_{s+1}^{(i)} - \bar{x}_{s+1}^{(i)} \\
&= x_s^{(i)} - \gamma\nabla f_{\hat{\mu}_s,\pi}(x_s^{(i)}) - \bar{x}_s^{(i)} + \gamma\nabla f_{\mu_s,\pi}(\bar{x}_s^{(i)}) \\
&= \big(x_s^{(i)} - \gamma\nabla f_{\hat{\mu}_s,\pi}(x_s^{(i)})\big) - \big(\bar{x}_s^{(i)} - \gamma\nabla f_{\bar{\mu}_s,\pi}(\bar{x}_s^{(i)})\big) + \gamma\big(\nabla f_{\mu_s,\pi}(\bar{x}_s^{(i)}) - \nabla f_{\bar{\mu}_s,\pi}(\bar{x}_s^{(i)})\big).
\end{aligned}$$

Since $\nu_0 = \bar{\mu}_s$ and $\nu_1 = \hat{\mu}_s$, Newton–Leibniz formula gives

$$\left(x_s^{(i)} - \gamma \nabla f_{\hat{\mu}_s,\pi}(x_s^{(i)})\right) - \left(\bar{x}_s^{(i)} - \gamma \nabla f_{\bar{\mu}_s,\pi}(\bar{x}_s^{(i)})\right) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \left\{ x_t^{(i)} - \gamma \nabla f_{\nu_t,\pi}(x_t^{(i)}) \right\} \mathrm{d}t.$$

Taking the root mean squared norm, applying Minkowski's inequality, and then using Eq. (C.5), we obtain

$$\begin{aligned}
\mathsf{e}_{s+1} &\leq \int_0^1 \left\{ \frac{1}{N} \sum_{i=1}^N \left\| \frac{\mathrm{d}}{\mathrm{d}t} \left\{ x_t^{(i)} - \gamma \nabla f_{\nu_t,\pi}(x_t^{(i)}) \right\} \right\|^2 \right\}^{1/2} \mathrm{d}t \\
&\quad + \gamma \left\{ \frac{1}{N} \sum_{i=1}^N \|\nabla f_{\bar{\mu}_s,\pi}(\bar{x}_s^{(i)}) - \nabla f_{\mu_s,\pi}(\bar{x}_s^{(i)})\|^2 \right\}^{1/2} \\
&\leq \left\{ 1 + \frac{5\gamma dB}{4\lambda} \sup_{0 \leq t \leq 1} \mathrm{MMD}(\nu_t, \pi) \right\} \mathsf{e}_s + \frac{\gamma\sigma\xi}{2\sqrt{\lambda}} \left\{ 1 + \frac{\mathrm{MMD}(\mu_0, \pi)}{\lambda} \right\}.
\end{aligned}$$

For the interpolated measure, we estimate $\mathrm{MMD}(\nu_t, \pi)$ using its distance to the auxiliary empirical measure. Since $\nu_t = N^{-1} \sum_{i=1}^N \delta_{x_t^{(i)}}$ and $x_t^{(i)} = \bar{x}_s^{(i)} + t(x_s^{(i)} - \bar{x}_s^{(i)})$, the triangle inequality gives

$$\mathrm{MMD}(\nu_t, \pi) \leq \mathrm{MMD}(\mu_s, \pi) + \|m_{\bar{\mu}_s} - m_{\mu_s}\|_{\mathcal{H}} + \|m_{\nu_t} - m_{\bar{\mu}_s}\|_{\mathcal{H}}.$$

On the complement of the union in Eq. (C.4), $\|m_{\bar{\mu}_s} - m_{\mu_s}\|_{\mathcal{H}} \leq \sigma\xi \leq (K+1)^{-\alpha/2}/4$. Taylor's formula along the particle segments gives $\|m_{\nu_t} - m_{\bar{\mu}_s}\|_{\mathcal{H}} \leq (K+1)^{-\alpha/2}/4$ whenever $\mathsf{e}_s \leq 1 \wedge [(K+1)^{-\alpha/2}/(4(\sqrt{2\sigma} + dB/2))]$, by the linear-term and remainder calculations at the start of Step 3. Adding these two explicit terms in the triangle inequality gives $\mathrm{MMD}(\nu_t, \pi) \leq \mathrm{MMD}(\mu_s, \pi) + (K+1)^{-\alpha/2}/2$ for all $0 \leq t \leq 1$.

Integrating Eq. (C.5) and substituting the explicit estimates for the MMD of the interpolated measure and for the witness-velocity difference gives, for $s < \min\{S, K\}$,

$$\mathsf{e}_{s+1} \leq \left[1 + \frac{5\gamma dB}{4\lambda} \left\{ \mathrm{MMD}(\mu_s, \pi) + \frac{(K+1)^{-\alpha/2}}{2} \right\} \right] \mathsf{e}_s + \frac{\gamma\sigma\xi}{2\sqrt{\lambda}} \left\{ 1 + \frac{\mathrm{MMD}(\mu_0, \pi)}{\lambda} \right\}. \tag{C.6}$$

Step 4: Iterating the error recursion and choosing K. Throughout Step 4, $C \geq 1$ depends only on the parameters listed in Theorem 5.2 and may be increased from one estimate to the next. By Lemma C.3, $\mathrm{MMD}(\mu_s, \pi) \leq C(1+s)^{-\alpha/2}$ for $0 \leq s \leq S$. Comparing the decreasing function $(1+t)^{-\alpha/2}$ with its integral gives

$$\sum_{s=0}^{\min\{S,K\}-1} \mathrm{MMD}(\mu_s, \pi) \leq C \left(1 + \int_0^K (1+t)^{-\alpha/2} \mathrm{d}t \right) \leq C(K+1)^{1-\alpha/2}.$$

The last inequality uses $1 - \alpha/2 \geq 1/2$. We prove the induction hypothesis stated at the start of Step 3. The common initialization gives $\mathsf{e}_0 = 0$. Assume that this hypothesis holds through iteration $j-1$, where $1 \leq j \leq \min\{S, K\}$. Repeated substitution in Eq. (C.6), followed by $1 + u \leq \exp(u)$ for $u \geq 0$, gives

$$\begin{aligned}
\mathsf{e}_j &\leq \frac{\gamma\sigma\xi}{2\sqrt{\lambda}} \left(1 + \frac{\mathrm{MMD}(\mu_0, \pi)}{\lambda} \right) \sum_{i=0}^{j-1} \prod_{l=i+1}^{j-1} \left\{ 1 + \frac{5\gamma dB}{4\lambda} \left(\mathrm{MMD}(\mu_l, \pi) + \frac{(K+1)^{-\alpha/2}}{2} \right) \right\} \\
&\leq \frac{\gamma j \sigma\xi}{2\sqrt{\lambda}} \left(1 + \frac{\mathrm{MMD}(\mu_0, \pi)}{\lambda} \right) \exp\left\{ \frac{5\gamma dB}{4\lambda} \sum_{l=0}^{j-1} \left(\mathrm{MMD}(\mu_l, \pi) + \frac{(K+1)^{-\alpha/2}}{2} \right) \right\} \\
&\leq \frac{C(K+1)^{3/2}}{\sqrt{N\delta}} \exp\left\{ C(K+1)^{1-\alpha/2} \right\}.
\end{aligned}$$

An empty product equals one. The final inequality follows by summing the inequality in Lemma C.3 over the iterations and using $j \leq K$, $j(K+1)^{-\alpha/2} \leq (K+1)^{1-\alpha/2}$, and $\xi = \sqrt{2(K+1)/(N\delta)}$.

The induction hypothesis in Step 3 has right-hand side $1 \wedge [(K+1)^{-\alpha/2}/(4(\sqrt{2\sigma}+dB/2))]$, which is at least $(K+1)^{-\alpha/2}/\max\{1, 4(\sqrt{2\sigma}+dB/2)\}$. Squaring the expression obtained by iterating Eq. (C.6) and comparing it with the square of this explicit right-hand side gives, after increasing C, the sufficient condition

$$N\delta \geq C(K+1)^{3+\alpha} \exp\left\{C(K+1)^{1-\alpha/2}\right\}$$

for e_j to satisfy the induction hypothesis in Step 3. The power $3+\alpha$ comes from squaring the expression for e_j obtained from Eq. (C.6) and dividing by a constant multiple of $(K+1)^{-\alpha}$; the resulting exponential coefficient and fixed prefactor are included in C. This particle number requirement also ensures $\xi \leq 1$ and $\sigma\xi \leq (K+1)^{-\alpha/2}/4$, since their squared forms are $N\delta \geq 2(K+1)$ and $N\delta \geq 32\sigma^2(K+1)^{1+\alpha}$, respectively. The induction starts at $\mathsf{e}_0 = 0$ and proves the hypothesis in Step 3 for every $0 \leq s \leq \min\{S, K\}$.

At $t=1$, the interpolation satisfies $\nu_1 = \hat{\mu}_s$. Taylor's formula at the auxiliary particles, the triangle inequality, and Eq. (C.4) on the complement of its union give $\mathrm{MMD}(\hat{\mu}_s, \mu_s) \leq (K+1)^{-\alpha/2}/2$ for $0 \leq s \leq \min\{S, K\}$. If $S \leq K$, this covers every required iteration. If $S > K$, then for $K < s \leq S$, applying Proposition C.2 to the empirical and population sequences gives

$$\begin{aligned}\mathrm{MMD}(\hat{\mu}_s, \mu_s) &\leq \mathrm{MMD}(\hat{\mu}_s, \pi) + \mathrm{MMD}(\mu_s, \pi) \\ &\leq \mathrm{MMD}(\hat{\mu}_K, \pi) + \mathrm{MMD}(\mu_K, \pi) \\ &\leq \mathrm{MMD}(\hat{\mu}_K, \mu_K) + 2\,\mathrm{MMD}(\mu_K, \pi) \leq C(K+1)^{-\alpha/2}.\end{aligned}$$

The last inequality uses the estimate for the MMD between the particle and population measures at iteration K, together with Lemma C.3. Hence the particle number requirement on $N\delta$ obtained from Eq. (C.6) gives $\max_{0\leq s\leq S} \mathrm{MMD}(\hat{\mu}_s, \mu_s) \leq C(K+1)^{-\alpha/2}$ with probability at least $1-\delta$. The same constant applies to every final iteration S with a fixed source bound Q.

We choose K as follows: whenever $\log(N\delta) \geq 2C$,

$$K = \left\lfloor \left(\frac{\log(N\delta)}{2C}\right)^{2/(2-\alpha)} \right\rfloor.$$

Since $\log(N\delta) \geq 2C$, we have $K \geq 1$ and $K+1 \leq 2(\log(N\delta)/(2C))^{2/(2-\alpha)}$. Raising the latter inequality to the power $1-\alpha/2$ and using $2^{1-\alpha/2} \leq 2$, we obtain $C(K+1)^{1-\alpha/2} \leq \log(N\delta)$. The particle number requirement derived from Eq. (C.6) therefore holds for this K. Moreover, $K+1 \geq (\log(N\delta)/(2C))^{2/(2-\alpha)}$. Therefore,

$$\max_{0\leq s\leq S} \mathrm{MMD}(\hat{\mu}_s, \mu_s) \leq C(K+1)^{-\alpha/2} \leq C\left(\frac{\log(N\delta)}{2C}\right)^{-\alpha/(2-\alpha)}.$$

Absorbing the fixed factor $(2C)^{\alpha/(2-\alpha)}$ into the constant in Theorem 5.2 gives the stated rate. The definition of K uses the constant selected before separating the parameter ranges at $2C$. Because K is determined by N, δ and the fixed problem parameters, Eq. (C.4) applies for every final iteration S with success probability at least $1-\delta$.

For $1 \leq \log(N\delta) < 2C$, the lower endpoint follows from $N\delta \geq e$. Since $N\delta \geq \sigma^2$, Lemma C.1 and Chebyshev's inequality give $\mathbb{P}\left[\|m_{\hat{\mu}_0} - m_{\mu_0}\|_{\mathcal{H}} > 1\right] \leq \sigma^2/N \leq \delta$. On the complementary event, $\mathrm{MMD}(\hat{\mu}_0, \pi) \leq \mathrm{MMD}(\mu_0, \pi) + 1$, so Proposition C.2 gives empirical monotonicity. The triangle inequality and Lemma C.3 yield $\max_{0\leq s\leq S} \mathrm{MMD}(\hat{\mu}_s, \mu_s) \leq 2\,\mathrm{MMD}(\mu_0, \pi) + 1$. Since $[\log(N\delta)]^{-\alpha/(2-\alpha)} \geq (2C)^{-\alpha/(2-\alpha)}$, increasing the theorem's constant proves its conclusion in this range. Combining the two parameter ranges gives the conclusion with probability at least $1-\delta$.

□

D Additional Experimental Details

In this section, we provide additional details for reproducing the experiments in Section 6, together with supplementary experimental results. The experiments in Section 6 were run on a workstation equipped with an AMD Ryzen 9 5950X 16-Core Processor and an NVIDIA GeForce RTX 4090 GPU.

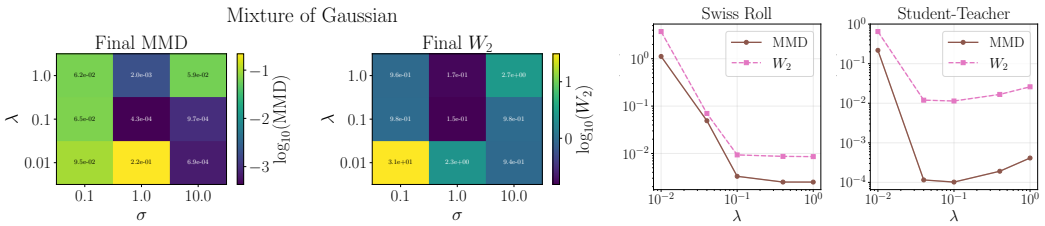

Figure 4: Ablation studies of SrMMD flow on Mixture of Gaussians, Swiss roll, and student-teacher network, with respect to kernel bandwidth σ and regularization parameter λ.

D.1 Toy examples

The mixture of Gaussian consists of four Gaussian distributions with mean $(\pm 2, \pm 2)$ and the same covariance matrix $1.2\mathrm{Id}$. The swiss roll distribution is taken from the SKLEARN package. For the gaussian mixture setting, the kernel mean embedding under a Gaussian kernel $k(x, y) = \exp(-(2\sigma^2)^{-1}\|x - y\|^2)$ admits a close-form expression [16]. Specifically, for a Gaussian distribution $\nu = \mathcal{N}(m, \Sigma)$:

$$m_\nu(x) = \det\left(\mathrm{Id} + \sigma^{-2}\Sigma\right)^{-1/2} \exp\left(-\frac{1}{2}(x - m)^\top \left(\Sigma + \sigma^2 \mathrm{Id}\right)^{-1}(x - m)\right).$$

In contrast, the kernel mean embedding for the Swiss roll is approximated as an empirical average over 500 i.i.d samples from the target.

In the left panel of Figure 4, we report ablation studies on the kernel bandwidth σ and the regularization parameter λ for the SrMMD flow on the Gaussian mixture benchmark. We also report an ablation study on λ for the Swiss roll benchmark, where the kernel is chosen as $k(x, x') = -\|x - x'\|$ and therefore does not involve a lengthscale parameter. For the mixture of Gaussians, the configuration achieving the smallest MMD and W_2 distance is $(\sigma, \lambda) = (1.0, 0.1)$. For the Swiss roll benchmark, the best-performing value is $\lambda = 0.1$. The hyperparameters of the MMD flow and DrMMD flow baselines are selected analogously by searching over the candidate sets $\{0.01, 0.03, 0.1, 0.3, 1.0, 3.0\}$ for σ and $\{0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0\}$ for λ. We omit the corresponding ablation plots, as these baselines are not the primary focus of this study.

D.2 Computational cost and low-rank approximations

We further evaluate the computational cost of SrMMD flow and the effect of low-rank approximations on the Gaussian mixture benchmark. We use $N = M = 500$ source and target particles in $d = 2$ dimensions, with Gaussian kernel length scale $\sigma = 1.0$, regularization parameter $\lambda = 0.1$, and step size $\gamma = 0.1$. We compare exact SrMMD flow with Nyström approximations [85] of ranks $r \in \{32, 64, 128, 256\}$ and random Fourier features (RFF) [70] with $R \in \{32, 64, 128, 256\}$ features, together with the vanilla MMD flow and DrMMD flow. All methods are run for 200 iterations over 10 random seeds, giving 110 runs in total. Performance is evaluated using MMD computed with the exact Gaussian kernel for all methods.

Low-rank approximations. In Table 2, we report the final MMD, time per iteration, and relative error of the final local velocity for each approximation. The time per iteration measures synchronized steady-state optimization time on the GPU after JAX compilation. The local velocity error compares the approximate velocity with the exact SrMMD velocity evaluated at the same particle configuration. Results are reported as mean $\pm$ standard deviation over 10 random seeds.

From Table 2, the rank-64 Nyström approximation achieves a final MMD comparable to exact SrMMD flow, while reducing the time per iteration by a factor of 2.19. The rank-32 approximation also reduces the cost relative to exact SrMMD, but has a higher final MMD and substantially larger variability across seeds. Ranks 128 and 256 closely match the final MMD of exact SrMMD, although their time per iteration is higher, suggesting that the additional cost of the approximation outweighs its benefit at this matrix size. The relative velocity error decreases from 0.3464 at rank 32 to 0.01165 at rank 64, and to approximately 10^{-4} and 10^{-6} at ranks 128 and 256, respectively.

Table 2: Ablation studies of Nyström and RFF approximations for SrMMD flow on the Gaussian mixture benchmark. Final MMD is computed with the exact kernel after 200 iterations. Results are reported as mean ± standard deviation over 10 random seeds.

Method	Rank/features	Time/iteration (s)	Final MMD	Relative velocity error
Exact SrMMD	–	0.003454 ± 0.000017	0.01093 ± 0.00087	0
Nyström	32	0.002453 ± 0.000060	0.01837 ± 0.01410	0.3464 ± 0.4350
Nyström	64	0.001578 ± 0.000005	0.01098 ± 0.00085	0.01165 ± 0.00813
Nyström	128	0.004154 ± 0.000009	0.01094 ± 0.00087	$(1.06 \pm 0.58) \times 10^{-4}$
Nyström	256	0.006333 ± 0.000019	0.01093 ± 0.00087	$(2.05 \pm 1.51) \times 10^{-6}$
RFF	32	0.000372 ± 0.000002	0.09479 ± 0.05341	0.9508 ± 0.0298
RFF	64	0.000410 ± 0.000001	0.05839 ± 0.01465	0.9204 ± 0.0466
RFF	128	0.000442 ± 0.000002	0.03344 ± 0.00533	0.8562 ± 0.0502
RFF	256	0.000992 ± 0.000014	0.02169 ± 0.00258	0.7541 ± 0.0599

For RFF, increasing the number of features from 32 to 256 reduces the final MMD from 0.09479 to 0.02169. With 256 features, the time per iteration is lower than that of exact SrMMD by a factor of 3.48. The initial relative Frobenius error of the kernel approximation decreases from 0.1176 ± 0.0818 to 0.0293 ± 0.0243, and the initial relative velocity error decreases from 0.282 ± 0.119 to 0.0810 ± 0.0295. However, the final relative velocity error remains 0.7541 ± 0.0599 with 256 features. Thus, increasing the number of random features improves the final MMD, but the resulting velocity approximation remains less accurate than Nyström on this benchmark.

Wall-clock time and memory. In Table 3, we compare the runtime and peak memory usage of exact SrMMD flow and selected approximations with the vanilla MMD flow and DrMMD flow. The total method time includes one-time JAX compilation and the full 200-iteration optimization. After 200 iterations, exact SrMMD flow achieves a final MMD approximately 10.2 and 13.9 times smaller than MMD flow and DrMMD flow, respectively, at a higher cost per iteration. The rank-64 Nyström approximation retains comparable final accuracy at a lower cost per iteration. RFF with 128 and 256 features reduces the time per iteration by factors of 7.82 and 3.48, respectively, while yielding final MMD values approximately 3.06 and 1.98 times those of exact SrMMD. Both configurations achieve smaller final MMD values than the vanilla MMD flow and DrMMD flow.

Table 3: Runtime, peak memory usage, and final MMD on the Gaussian mixture benchmark. Time per iteration is measured after JAX compilation; total time includes compilation and 200 optimization steps. Runtime and memory entries are means over 10 random seeds; final MMD is reported as mean ± standard deviation.

Method	Time/iteration (s)	Total time (s)	Peak CPU (MiB)	Peak GPU (MiB)	Final MMD
MMD	0.000050	0.1299	480.53	406	0.11187 ± 0.00539
DrMMD	0.000197	0.7690	689.07	456	0.15192 ± 0.00465
Exact SrMMD	0.003454	1.1337	646.50	482	0.01093 ± 0.00087
SrMMD, Nyström $r = 64$	0.001578	0.7998	705.50	504	0.01098 ± 0.00085
SrMMD, RFF $R = 128$	0.000442	0.7376	691.64	450	0.03344 ± 0.00533
SrMMD, RFF $R = 256$	0.000992	0.8603	694.79	450	0.02169 ± 0.00258

The rank-64 Nyström approximation does not reduce the measured peak memory usage relative to exact SrMMD. RFF reduces peak GPU memory from 482 MiB to 450 MiB, while its peak CPU memory is higher. These results therefore do not indicate a general memory advantage for the approximations at the scale considered here.

Comparison under a fixed time budget. We also compare all methods at a common post-compilation optimization-time horizon of 9.22 ms, determined by the shortest completed trajectory. In Table 4, we report the median MMD and the 25%–75% percentile range over 10 random seeds. The vanilla MMD flow achieves the smallest MMD under this short time budget. Among the SrMMD variants, RFF with 32 features performs best at this horizon, although its final MMD after 200 iterations is higher than that of the other SrMMD variants. Thus, the lower final MMD of exact

Table 4: MMD at the common post-compilation optimization-time horizon of 9.22 ms on the Gaussian mixture benchmark. Entries show the median and the 25%–75% percentile range over 10 random seeds. All MMD values are computed with the exact kernel.

Method	MMD at the common time horizon
MMD	$0.12660\ [0.11637, 0.13007]$
DrMMD	$0.15216\ [0.14928, 0.15482]$
Exact SrMMD	$0.95005\ [0.94722, 0.95088]$
SrMMD, Nyström $r = 32$	$0.93552\ [0.93146, 0.93757]$
SrMMD, Nyström $r = 64$	$0.83603\ [0.83244, 0.83971]$
SrMMD, Nyström $r = 128$	$0.95687\ [0.95407, 0.95768]$
SrMMD, Nyström $r = 256$	$0.98113\ [0.97857, 0.98199]$
SrMMD, RFF $R = 32$	$0.38136\ [0.32539, 0.41268]$
SrMMD, RFF $R = 64$	$0.38784\ [0.34902, 0.42402]$
SrMMD, RFF $R = 128$	$0.42884\ [0.40841, 0.44168]$
SrMMD, RFF $R = 256$	$0.72221\ [0.70681, 0.73046]$

SrMMD after a fixed number of iterations does not imply better performance under every wall-clock budget.

Empirical second moments. Finally, we examine the final empirical second moments as a numerical stability check. Across all 110 runs, the final empirical second moments are finite and range from 7.878 to 8.868. The mean $\pm$ standard deviation is 8.425 ± 0.170 for exact SrMMD, 8.424 ± 0.169 for the rank-64 Nyström approximation, and 8.382 ± 0.184 for RFF with 256 features.

D.3 Student–teacher network

The two-layer neural network used in this experiment is $\psi(z, \theta) = G\big(b^1 + W^1 \varsigma(W^0 z + b^0)\big)$, where ς here is the ReLU activation function and $G(x) = \exp(-x^2/4)$. Here, $\theta = (b^1, W^1, b^0, W^0) \in \mathbb{R}^d$ denotes the concatenation of all network parameters, with $d = 1 + 1 + 50 + 1 = 53$. The teacher distribution is $\pi = \mathcal{N}(0, \mathrm{Id})$ on $\mathbb{R}^{53}$, from which we observe $M = 10$ i.i.d. samples and then fixed througout. The student distribution is initialized with $N = 100$ particles sampled independently from $\mathcal{N}(0, 0.1\mathrm{Id})$.

As shown in the main text, training the mean-field studnet teacher network here amounts to MMD flow with kernel $k(\theta, \theta') = \mathbb{E}_{z \sim \mathbb{P}_{\mathrm{data}}}[\psi(z, \theta)^\top \psi(z, \theta')]$. Specifically, the squared discrepancy between the mean teacher and student networks can be written as:

$$\begin{aligned}
&\mathbb{E}_{z \sim \mathbb{P}_{\mathrm{data}}}\left[\left\|\Psi_\pi^{\mathsf{T}}(z) - \Psi_\mu^{\mathsf{S}}(z)\right\|^2\right] = \mathbb{E}_{z \sim \mathbb{P}_{\mathrm{data}}}\left[\left\|\mathbb{E}_{\theta \sim \pi}[\psi(z, \theta)] - \mathbb{E}_{\theta' \sim \mu}[\psi(z, \theta')]\right\|^2\right] \\
&\quad = \mathbb{E}_{\substack{\theta \sim \pi \\ \theta' \sim \pi}}[k(\theta, \theta')] + \mathbb{E}_{\substack{\theta \sim \mu \\ \theta' \sim \mu}}[k(\theta, \theta')] - 2\mathbb{E}_{\substack{\theta \sim \pi \\ \theta' \sim \mu}}[k(\theta, \theta')] \\
&\quad = \iint k(\theta, \theta')\, \mathrm{d}(\pi - \mu)(\theta)\, \mathrm{d}(\pi - \mu)(\theta') = \mathrm{MMD}^2(\mu, \pi).
\end{aligned} \tag{D.1}$$

In practice, the expectation with respect to $\mathbb{P}_{\mathrm{data}}$ is approximated using an empirical average over 1000 i.i.d. training samples, and performance is evaluated on a held-out set of 1000 independent validation samples. During training, the kernel $k(\theta, \theta')$ is approximated at each iteration by averaging over a randomly selected subset of 100 training samples.

In the right panel of Figure 4, we report an ablation study on the regularization parameter λ for the SrMMD flow. Since the kernel used in this experiment is defined through the network features, it does not involve a kernel lengthscale. The best-performing value is $\lambda = 0.1$. For the DrMMD flow baseline, we use the same value $\lambda = 0.1$, following the experimental results of Chen et al. [18].

D.4 Color transfer

All methods use a fixed Gaussian kernel $k_\sigma(x, y) = \exp(-0.5\|x - y\|^2)$ and use a fixed step size $\gamma = 0.01$. The regularization parameter λ is selected by searching over the pre-specified candidate sets $\{0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0\}$. The final selected regularization parameter for SrMMD flow is $\lambda = 0.01$. To transfer colors, for each source pixel with color c, we identify the nearest

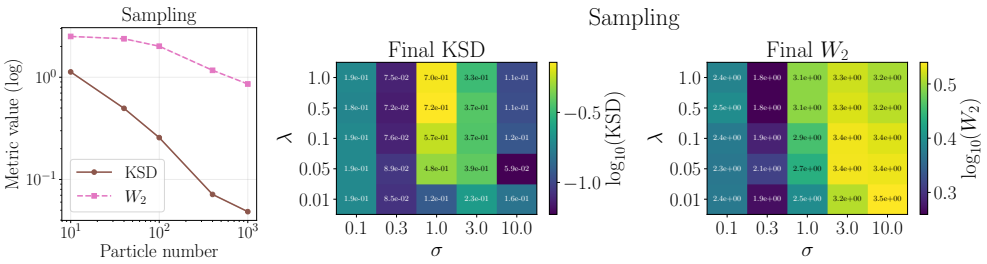

Figure 5: Ablation studies of SrMMD flow as a *sampling* method on the mixture of Gaussians, with respect to the number of particles (**Left**), lengthscale σ and regularization strength λ (**Middle & Right**).

particle in the initial particle set Y_0, $i = \arg\min_j \|c - Y_{0,j}\|^2$, and assign the pixel the color of the corresponding transported particle $Y_{T,i}$. The resulting image is shown in the right panel of Figure 2.

D.5 Sampling: Gaussian mixture and Bayesian logistic regression

Gaussian mixture Our Gaussian mixture setting is the same as the ones considered in [17, 15], which is a mixture of 10 Gaussians in $d = 2$ dimensions. In the sampling literature, this is the simplest setting where the target distribution π satisfies a Log-Sobolev or Poincaré inequality yet the corresponding constants increase exponentially with the distance between distinct modes. The score function is computed by taking the gradient of the log density. For SrMMD, KSD and HrMMD flow, we use the Stein kernel constructed from the Gaussian kernel as the base. For SVGD and R-SVGD, we use the Gaussian kernel. For a fair comparison, all these methods use the same kernel lengthscale $\sigma = 0.3$. For SrMMD, HrMMD, and KSD flows, we use the best tuned step sizes $\gamma = 0.01$ for KSD and $\gamma = 0.1$ for SrMMD and HrMMD. For SVGD and R-SVGD, we use the best-tuned step size $\gamma = 0.01$. All the step sizes are tuned from a pre-specified candidate set $\{0.01, 0.03, 0.1, 0.3, 1.0\}$.

In Figure 5, we provide an ablation study of our SrMMD flow. The left panel shows that both final KSD and W_2 decrease as the particle number N increases, indicating that our particle implementation benefits from a more accurate empirical approximation of the population descent scheme, as desired. The middle and right panel shows that the best configuration for SrMMD flow is $(\sigma, \lambda) = (0.3, 0.5)$.

Bayesian logistic regression We use four datasets from the UCI Machine Learning Repository [53]: BREAST CANCER ($d = 30$, 569 samples), IONOSPHERE ($d = 34$, 351 samples), GERMAN CREDIT ($d = 24$, 1000 samples), and COVTYPE ($d = 54$, 30,000 samples). Fro all the datasets, $\frac{2}{3}$ are used as training samples and $\frac{1}{3}$ are used as validation samples. For SrMMD and KSD flow, we use the Stein kernel constructed from the Gaussian kernel as the base. For SVGD, we use the Gaussian kernel. For a fair comparison, all these methods use the same kernel lengthscale $\sigma = 1.0$. We use the L-BFGS optimizer [59] for KSD flow as suggested by [56], with initial step size 0.01 and convergence tolerance 10^{-3}. SrMMD and SVGD use a fixed-step Euler update with step size $\gamma = 0.1$, selected from the candidate set $\{0.01, 0.03, 0.1, 0.3, 1.0\}$. All methods operate in the full-batch regime (no minibatching) and are run for 3,000 iterations until convergence is empirically observed.

D.6 Hybrid-regularized maximum mean discrepancy (HrMMD) flow

In this section, we discuss our Hybrid-regularized maximum mean discrepancy (HrMMD), which is another regularized variant of MMD with both L_2 norm and gradient norm regularization on the witness function. Specifically, for $\alpha \in [0, 1]$ and $\lambda > 0$, we define

$$\mathrm{HrMMD}(\mu, \pi) := \sup_{f \in \mathcal{H}:\, \alpha \|\nabla f\|^2_{L^d_2(\mu)} + (1-\alpha)\|f\|^2_{L_2(\mu)} + \lambda \|f\|^2_{\mathcal{H}} \le 1} \int f \,\mathrm{d}(\mu - \pi).$$

This formulation interpolates continuously between two previously discussed regularizations. When $\alpha = 1$, HrMMD reduces to our SrMMD in Definition 3.1. When $\alpha = 0$, only the L_2 norm remains and it reduces to DrMMD [18]. However, a slight distinction here remains since DrMMD penalizes the $L_2(\pi)$ norm while here it penalizes the $L_2(\mu)$ norm. HrMMD provides an interpolation of these

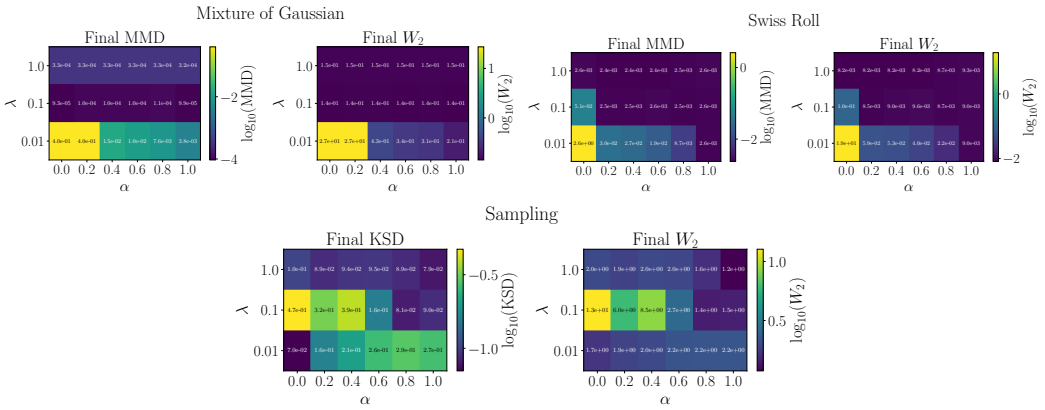

Figure 6: Ablation of HrMMD flow with respect to λ and α on mixture of Gaussians, Swiss roll and sampling.

two types of regularization, and it is also exactly the regularization used in the actual MMD GANs [3]. The HrMMD objective is the same as the Sobolev-Fisher discrepancy studied by [65], where it was subsequently used as the objective to be minimized in the Wasserstein–Fisher–Rao geometry.

The corresponding witness function admits the closed-form representation

$$f^{\mathrm{HrMMD}}_{\mu,\pi} = (\alpha S_\mu + (1-\alpha)\Sigma_\mu + \lambda \mathrm{Id})^{-1}(m_\mu - m_\pi),$$

where $S_\mu = D^*_\mu D_\mu$ is the gradient covariance operator defined in Section 4 in the main text, and $\Sigma_\mu = \mathbb{E}_{x\in\mu}[k(x,\cdot)\otimes k(x,\cdot)]$ is the covariance operator from $\mathcal{H}\to\mathcal{H}$ associated with the kernel k [67]. Similarly as our SrMMD, the witness function $f^{\mathrm{HrMMD}}_{\mu,\pi}$ of HrMMD also admits a tractable particle-level closed form expression. Let $\hat{\mu} = \frac{1}{N}\sum_{i=1}^N \delta_{x^{(i)}}$ be the empirical distribution of the particles, and let

$$[K_{XX}]_{ij} = k(x^{(i)}, x^{(j)}),\ [D_{XX}]_{(i,\ell),j} = \partial_{1,\ell}k(x^{(i)}, x^{(j)}),\ [H_{XX}]_{(i,\ell),(j,m)} = \partial_{1,\ell}\partial_{2,m}k(x^{(i)}, x^{(j)}).$$

We also denote

$$\mathbb{K}_X(z) = [k(x^{(1)}, z), \ldots, k(x^{(N)}, z)]^\top, \quad \mathbb{D}_X(z) = [\partial_{1,1}k(x^{(1)}, z), \ldots, \partial_{1,d}k(x^{(N)}, z)]^\top,$$

and let $\mathbf{1}_N \in \mathbb{R}^N$ be the vector of all ones. Then the empirical witness function can be written as

$$f^{(\alpha)}_{\hat{\mu},\hat{\pi}}(z) = \frac{1}{\lambda}\left(\frac{1}{N}\mathbf{1}_N^\top \mathbb{K}_X(z) - \mathbb{E}_\pi[k(Y,z)]\right) - \frac{1}{\lambda}\Big(\sqrt{1-\alpha}\,\beta_k^\top \mathbb{K}_X(z) + \sqrt{\alpha}\,\beta_d^\top \mathbb{D}_X(z)\Big),$$

where (β_k, β_d) is determined by the linear system

$$\left(\begin{bmatrix} (1-\alpha)K_{XX} & \sqrt{\alpha(1-\alpha)}\,D_{XX}^\top \\ \sqrt{\alpha(1-\alpha)}\,D_{XX} & \alpha H_{XX}\end{bmatrix} + N\lambda \mathrm{Id}\right)\begin{bmatrix}\beta_k \\ \beta_d\end{bmatrix} = \begin{bmatrix}\sqrt{1-\alpha}\,g \\ \sqrt{\alpha}\,r\end{bmatrix},$$

with

$$g = \frac{1}{N}K_{XX}\mathbf{1}_N - \mathbb{E}_{z\sim\pi}[\mathbb{K}_X(z)], \quad r = \frac{1}{N}D_{XX}\mathbf{1}_N - \mathbb{E}_{z\sim\pi}[\mathbb{D}_X(z)].$$

In the generative modeling setting, where π is represented by i.i.d. target samples $\{y^{(j)}\}_{j=1}^M$, all the expectations with respect to π above are replaced by empirical averages over these samples. In the sampling setting with a Stein kernel, the target mean embedding vanishes under Stein's identity, which further simplifies the expression.

In Figure 6, we report ablation studies for HrMMD flow with respect to α and λ on the mixture of Gaussians, Swiss roll, and sampling experiments. We vary the interpolation parameter α over $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and the regularization parameter λ over $\{0.01, 0.1, 1.0\}$. Overall, HrMMD flow tends to perform better when the gradient penalty is stronger. This suggests that the gradient penalty plays a more important role than the L_2 penalty, consistent with empirical observations in MMD GANs [3, 10]. Nevertheless, the optimum is not always attained at $\alpha = 1$, indicating that a small additional L_2 penalty can further improve performance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions: the SrMMD flow, its continuous- and discrete-time convergence guarantees, its relation to MMD/KSD flows, and its empirical evaluation.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses several limitations, including the weaker metric used for sampling guarantees, the source-condition regularity assumption, the saturation effect of Tikhonov regularization, excluded singular kernels, and the absence of a finite-particle convergence analysis for the implemented particle scheme.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main theoretical statements explicitly state their assumptions, including growth, local boundedness, moment, and source-condition assumptions. Proofs of the main results are provided in the appendix, together with supporting lemmas and propositions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main algorithms, baselines, metrics, and many hyperparameters are described in the main text and appendix. A fully anonymous github link to the code is also provided.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A fully anonymous github link to the code is also provided.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 and Appendix C describe the experimental settings, including the benchmarks, baselines, kernels, particle numbers, iteration counts, step sizes, regularization parameters, optimizers, metrics, and ablation studies.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental figures aggregate results over multiple random seeds where applicable and report median curves with shaded 25%–75% percentile regions. The captions specify the aggregation convention used for the plotted variability.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper describes the algorithms and experimental hyperparameters, but it does not currently report the hardware type, memory, per-run execution time, or total compute used for the experiments. These details should be added to the appendix or supplementary material if available.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is methodological and theoretical, uses synthetic or standard benchmark data, and does not involve human-subject experiments, private data, or high-risk deployment. To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is primarily foundational work on convergence and regularization for MMD/KSD gradient flows, and does not propose a deployed system or a model intended for direct real-world decision making. Any societal impacts are indirect and depend on downstream applications of sampling or generative modelling methods.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk pretrained models, scraped datasets, or other assets with substantial misuse potential. The experiments use synthetic examples and standard benchmark tasks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing methods, software packages, and benchmark datasets are credited through citations.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: A fully anonymous github link to the code is also provided, including usage instructions, environment information, and a license.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval or equivalent review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development, theory, and experiments do not use LLMs as an important, original, or non-standard component. Any use of LLMs, if present, is limited to writing, editing, or formatting and does not affect the scientific methodology.