GENERALIZED SLICED PROBABILITY METRICS

Soheil Kolouri[1], *Kimia Nadjahi*[2], *Shahin Shahrampour*[3], *Umut Şimşekli*[4,5]

1: Vanderbilt University, 2: Telecom Paris, Institut Polytechnique de Paris, 3: Northeastern University, 4: INRIA, 5: École Normale Supérieure

ABSTRACT

Sliced probability metrics have become increasingly popular in machine learning, and they play a quintessential role in various applications, including statistical hypothesis testing and generative modeling. However, in a practical setting, the convergence behavior of the algorithms built upon these distances have not been well established, except for a few specific cases. In this paper, we introduce a new family of sliced probability metrics, namely Generalized Sliced Probability Metrics (GSPMs), based on the idea of slicing high-dimensional distributions into a set of their one-dimensional marginals. We show that GSPMs are true metrics, and they are related to the Maximum Mean Discrepancy (MMD). Exploiting this relationship, we consider GSPM-based gradient flows and show that, under mild assumptions, the gradient flow converges to the global optimum. Finally, we demonstrate that various choices of GSPMs lead to new positive definite kernels that could be used in the MMD formulation while providing a unique integral geometric interpretation. We illustrate the application of GSPMs in gradient flows.

Index Terms— Sliced probability metrics, gradient flows

1. INTRODUCTION

Measuring the discrepancy between probability distributions is at the heart of statistics and machine learning problems. A classic example in statistics is the hypothesis testing in higher dimensions, which has attracted a plethora of interest in recent years [1–3]. Similarly, in generative modeling, leveraging probability metrics and discrepancy measures as an alternative to the adversarial networks, used in Generative Adversarial Networks (GANs), has become an exciting topic [4–7]. Notably, variations of the Wasserstein distances and the Maximum Mean Discrepancy (MMD) have enjoyed ample attention from the community and have incited many enthralling works.

There are specific challenges with measuring the discrepancy between two high-dimensional probability distributions, including the high computational cost (e.g., for Wasserstein distances [8]) and the growing *sample complexity*, *i.e.*, in the sense of the dependence of convergence rate of a given metric between a measure and its empirical counterpart on the number of samples [9]. The community has tackled these challenges from different angles in recent years. One of the thought-provoking approaches is via slicing high-dimensional distributions over their one-dimensional marginals and comparing their marginal distributions [10, 11]. The idea of slicing distributions has been successfully used in, for instance, sliced-Wasserstein distances in various applications [12–17]. More recently, Kolouri et al. [18] extended the idea of linear slicing, used in sliced-Wasserstein distances, to non-linear slicing of high-dimensional distributions and showed an interesting connection between the popular idea of adversarial networks (as in GANs) and their proposed max-generalized-sliced-Wasserstein (max-GSW) distance. However, in a practical setting, the convergence behavior of the algorithms built upon these distances have not been well established, except for a few specific cases.

In this paper, we first point out that the idea of generalized slices presented in [18] is not particular to Wasserstein distances and can thus be extended to any metric. We then leverage this idea to introduce a broad family of probability metrics, which we denote as *Generalized Sliced Probability Metrics* (GSPMs). We provide a geometric interpretation of these metrics and show their connection to the well-celebrated MMDs. Finally, following the work of Arbel et al. [19], we identify sufficient regularity conditions under which kernels based on GSPM lead to global convergence of gradient flows; hence, making the proposed distance a suitable choice for applications dealing with probability flows and implicit generative modeling.

2. BACKGROUND

Let μ and ν be probability measures defined on a measurable space, $\mathcal{X}$, with corresponding densities p and q. In addition, let $(\mathcal{X}, d)$ denote a metric space. Let $\mathcal{F}$ be a class of real-valued bounded measurable functions on $\mathcal{X}$. Then, the slice of a probability measure μ, with respect to $f \in \mathcal{F}$, is the pushforward measure $f_{\#}\mu$. We use the equivalent terminology that the slice of a d-dimensional probability density function p ($d \geq 2$), with respect to a function $f \in \mathcal{F}$, is a one-dimensional probability density function given as,

$$p_f(\cdot) \quad = \quad \int_{\mathcal{X}} \delta(\cdot - f)\mathrm{d}\mu = \int_{\mathcal{X}} \delta(\cdot - f(x))p(x)\mathrm{d}x \qquad (1)$$

where δ is a one-dimensional Dirac function. Intuitively, p_f is the distribution of $f(x)$, for i.i.d samples x from p. A fundamental question here is when the set of all such one-dimensional slices (or marginals), $\mathcal{P}_{\mathcal{F}} = \{p_f \ : \ \forall f \in \mathcal{F}\}$, characterizes the distribution p. The answer to this question lies in the theory of the *Radon transform*, which we briefly review next.

Generalized Radon Transform (GRT). The field of generalized Radon transform focuses on the question of when one can recover the probability distribution p from the set of its marginal distributions, *i.e.*, $\mathcal{P}_{\mathcal{F}}$. The main application of the generalized Radon transform is in tomography, where p is an object to be imaged and $\mathcal{P}_{\mathcal{F}}$ is a set of tomographic measurements from the object. In the remainder of this paper, we assume $\mathcal{X} \subseteq \mathbb{R}^d$. The classical Radon transform states that p can be reconstructed from $\mathcal{P}_{\mathcal{F}}$ when $\mathcal{F} = \{f_\theta(x) = \langle x, \theta \rangle : \forall x \in \mathcal{X}, \forall \theta \in \mathbb{S}^{(d-1)}\}$, where $\mathbb{S}^{(d-1)} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$ is the d-dimensional unit sphere. Hence, the set $\mathcal{P}_{\mathcal{F}}$ fully characterizes p. The generalized Radon transform extends the classical Radon transform by considering more sophisticated classes of functions $\mathcal{F}$ [18].

3. GENERALIZED SLICED PROBABILITY METRICS

In this section, we show that any probability metric between two one-dimensional probability measures can be extended to higher di-

mensions via the concept of generalized slicing. Note that, comparing one-dimensional distributions (*i.e.*, slices) is often computationally preferable (for instance for the Wasserstein distance), or in some cases, is required by construction (as for the Cramér distance). Let $\xi(\cdot,\cdot)$ be a metric for one-dimensional probability measures. Let $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R} : \theta \in \Omega_\theta\}$ be a set of "defining functions", and denote by ρ a probability measure on Ω_θ with positive density $\omega(\theta) > 0$, such that $\mathrm{d}\rho(\theta) = \omega(\theta)\mathrm{d}\theta$. Then, for any two probability measures μ and ν defined on $\mathcal{X} \subseteq \mathbb{R}^d$ with respective densities p and q, we introduce the *Generalized Sliced Probability Metic* (GSPM), which extends $\xi(\cdot,\cdot)$ to $\mathbb{R}^d$ via,

$$\zeta_{\mathcal{F},\rho}(p,q) \triangleq \left(\int_{\Omega_\theta} \xi^r(p_{f_\theta}, q_{f_\theta})\mathrm{d}\rho(\theta)\right)^{\frac{1}{r}} \tag{2}$$

where $r \geq 1$. Let us first show that GSPM is a metric. Non-negativity and symmetry immediately follow from non-negativity and symmetry of $\xi(\cdot,\cdot)$, while triangle inequality follows from the Minkowski inequality (see Supplementary material [20]). Finally, the identity of indiscernibles states that, $\zeta_{\mathcal{F},\rho}(p,q) = 0$ if and only if (iff) $p = q$. The forward proof is straightforward: $p = q$ results in $p_{f_\theta} = q_{f_\theta}$ for all $\theta \in \Omega_\theta$, and since ξ is a metric $\xi(p_{f_\theta}, q_{f_\theta}) = 0$. The backward proof, however, requires the injectivity of GRT. Meaning that, when $p_{f_\theta} = q_{f_\theta}\ \forall\theta \in \Omega_\theta$, we can conclude that $p = q$ iff the GRT is injective. Hence, if GRT is injective then the associated GSPM provides a metric between μ and ν, which is induced by the one-dimensional probability metric $\xi(\cdot,\cdot)$. Otherwise, GSPMs are pseudo-metrics, which still could be very useful. For instance, Kolouri et al. [18] used the Wasserstein distance as ξ, and showed that setting $\mathcal{F}$ to be the parametric class of neural networks with a fixed architecture, leads to a pseudo-metric that can successfully be used in generative modeling.

We remark here that the probability measure ρ defined on Ω_θ could in practice be learned to provide a new metric learning schema over a set of distributions. However, we leave the topic of metric learning as a potential application of GSPMs to be pursued in future.

4. GSPMS AND MMDS

The seminal work by Gretton et al. [1,21] on MMD provides a framework for efficient comparison of probability distributions. MMD belongs to the class of *integral probability metrics* [22], and has become a popular tool to compare distributions in a wide variety of applications, e.g., generative modeling [6,23], and gradient flows [19]. In practice, MMD is defined with respect to a Reproducing Kernel Hilbert Space (RKHS), with a unique kernel. Like other kernel methods, the choice of kernel type is often crucial and yet application-dependent, with the most common choice being the *radial basis function* (RBF) kernel. In what follows, we show that an interesting subset of GSPMs are equivalent to MMDs, and lead to novel positive(-definite) kernels that have unique integral geometric interpretations.

Consider Equation (2) for the special case of $\xi(p_{f_\theta}, q_{f_\theta}) = \|Ap_{f_\theta} - Aq_{f_\theta}\|_2$ and $r = 2$, where A is a positive(-definite) linear operator. The positive assumption enforces ξ to be a norm (*i.e.*, the weighted Euclidean norm). If A is positive semi-definite, then ξ would become a pseudo-metric, and as a consequence $\zeta_{\mathcal{F},\rho}$ also becomes a pseudo-metric. Given a linear operator, A, we can write

$$\zeta^2_{\mathcal{F},\rho}(p,q) = \int_{\Omega_\theta} \|Ap_{f_\theta} - Aq_{f_\theta}\|_2^2 \mathrm{d}\rho(\theta)\ . \tag{3}$$

We focus on practical settings where we only observe i.i.d. samples $\{x_i\}_{i=1}^N$ and $\{y_j\}_{j=1}^M$ from p and q respectively. Substituting the empirical distribution in Equation (1) gives us the empirical slices as $\hat{p}_{f_\theta}(t) = \frac{1}{N}\sum_{i=1}^N \delta(t - f_\theta(x_i))$ and $\hat{q}_{f_\theta}(t) = \frac{1}{M}\sum_{j=1}^M \delta(t - f_\theta(y_j))$. Using a common trick-of-trade in statistics, and without the loss of generality, we consider a smoothened version of the empirical slices via a RBF, ϕ_σ, where σ identifies the radius of the RBF ($\phi_{\sigma=0}(\cdot) = \delta(\cdot)$). Note that using ϕ_σ is equivalent to assuming smoothness priors on the slices.

By plugging in the (smoothened) empirical sliced distributions into (3), we obtain:

$$\begin{aligned}
&\zeta^2_{\mathcal{F},\rho}(\hat{p},\hat{q}) = \\
&\frac{1}{N^2}\sum_{ij} \underbrace{\int_{\Omega_\theta} \langle A\phi_\sigma(\cdot - f_\theta(x_i)), A\phi_\sigma(\cdot - f_\theta(x_j))\rangle \mathrm{d}\rho(\theta)}_{k(x_i,x_j)} + \\
&\frac{1}{M^2}\sum_{ij} \underbrace{\int_{\Omega_\theta} \langle A\phi_\sigma(\cdot - f_\theta(y_i)), A\phi_\sigma(\cdot - f_\theta(y_j))\rangle \mathrm{d}\rho(\theta)}_{k(y_i,y_j)} - \\
&\frac{2}{MN}\sum_{ij} \underbrace{\int_{\Omega_\theta} \langle A\phi_\sigma(\cdot - f_\theta(x_i)), A\phi_\sigma(\cdot - f_\theta(y_j))\rangle \mathrm{d}\rho(\theta)}_{k(x_i,y_j)}
\end{aligned} \tag{4}$$

Equation (4) is also the squared MMD with the particular kernel shown there-in. Note that one can use the Monte-Carlo integral approximation to obtain an algorithmic way of calculating the kernel for any feasible $\mathcal{F}$, ϕ_σ, and A. We now argue that these family of kernels are positive definite (PD). Indeed,

$$k_\theta(x_i,x_j) = \langle A\phi_\sigma(\cdot - f_\theta(x_i)), A\phi_\sigma(\cdot - f_\theta(x_j))\rangle \tag{5}$$

is a dot-product kernel, which is by definition PD, and summation/integration of PD kernels results in a PD kernel. Therefore,

$$k(x_i,x_j) = \int_{\Omega_\theta} k_\theta(x_i,x_j)\mathrm{d}\rho(\theta) \tag{6}$$

is a PD kernel. Below, we study some special cases of the GSPMs based on $\xi(p_{f_\theta}, q_{f_\theta}) = \|Ap_{f_\theta} - Aq_{f_\theta}\|_2$, and their equivalent MMD form based on kernels.

Sliced-ℓ_2: A is the identity operator. When $A = id(\cdot)$, the GSPM is a generalized-sliced ℓ_2 distance between the two distributions, which was used in [24]. Following the work of Knop et al. [24], we assume that the RBF is a Gaussian, $\phi_\sigma(t) = \mathcal{N}(0, \frac{\sigma}{2})(t)$, which simplifies the kernel in Eq. (5) to:

$$k_\theta(x_i,x_j) = \mathcal{N}(f_\theta(x_i) - f_\theta(x_j), \sigma)(0) \tag{7}$$

The geometric interpretation of Equation (7) is quite interesting. First note that $f_\theta : \mathcal{X} \to \mathbb{R}$ therefore, the pre-image of a scalar in the range of f_θ is a *hyper-surface* in $\mathcal{X}$. This means that all points living on a hyper-surface would be projected to the same scalar in the range of f_θ (*i.e.*, iso-hyper-surface). Therefore, while x_i and x_j could be far away from one another (in a Euclidean sense), as long as they live on the same or nearby iso-hyper-surfaces they will be considered to be similar (with respect to f_θ).

As a special case, Knop et al. [24] used linear slices (*i.e.*, $f_\theta(x) = \langle x, \theta\rangle$), chose ρ as the uniform distribution over $\mathbb{S}^{(d-1)}$, and showed that when ϕ_σ is the Gaussian function, then Equation (6) has a closed form, which we provide in the supplement. Recall that in these derivations, we started by fixing a slicing operation (linear slices), and used a specific distance, *i.e.* ℓ_2 distance, and that we know

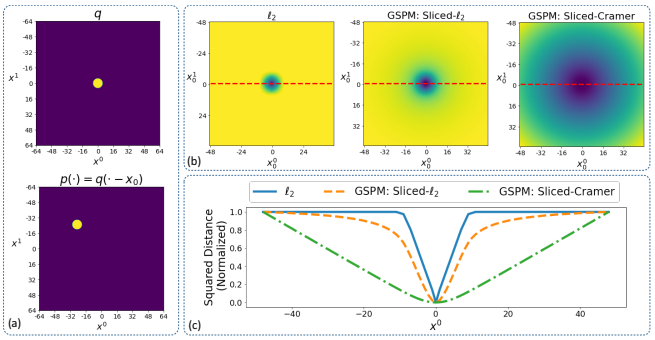

Fig. 1: Consider q to be a uniform distribution on a disc of radius R, and $p(\cdot) = q(\cdot - x_0)$ be a translated version of this distribution by x_0: panel (a). Panel (b) shows the distance between the two distributions p and q as a function of $x_0 = [x_0^0, x_0^1]^T$ for ℓ_2, sliced-ℓ_2, and sliced-Cramér. Panel (c) plots the distances for $x_0^1 = 0$ as a function of x_0^0. As expected the ℓ_2 distance carries no gradient information when $\|x_0\|_2 > 2R$, *i.e.* when p and q do not overlap, while sliced-ℓ_2 and sliced-Cramér both have non-zero gradients for $\forall x_0 \neq [0,0]^T$. In this experiment we chose linear slices, *i.e.* $f_\theta(x) = \langle x, \theta \rangle$. Both sliced-$\ell_2$ and sliced-Cramér guarantee convergence of gradient flows. This experiment is similar to Example 1 in [7].

the geometric meaning of both of these steps and their implications. Then, we ended up with a novel PD kernel that defines an MMD, which inherits these geometric properties. We note that the distance used here (and in [24]) is a sliced-ℓ_2.

Sliced-Cramér: A is the cumulative integral operator. We now study the specific case of the GSPM with the Cramér distance. This latter distance measures the ℓ_2 distance between the cumulative distribution functions (CDFs) of two probability densities. Therefore, setting A to be the cumulative integral operator (which is PD),

$$Ap_{f_\theta}(t) = \int_{-\infty}^{t} p_{f_\theta}(\tau)\mathrm{d}\tau.$$

would yield the (2-)Cramér distance $\xi(p_{f_\theta}, q_{f_\theta}) = \|Ap_{f_\theta} - Aq_{f_\theta}\|_2$ [25] between the two one-dimensional probability distributions, p_{f_θ} and q_{f_θ}. This distance has recently been used in several publications [26, 27], and shares some common characteristics with the Wasserstein distances. In fact, the 1-Cramér distance and the 1-Wasserstein distance are equivalent. It is straightforward to show that $k_\theta(x_i, x_j)$ for this choice of A is unbounded. Note that $A\phi_\sigma$ is the CDF of an RBF, therefore its integral is unbounded. However, assuming that the integral domain is bounded, *i.e.* $[-T, T]$, we can find closed form solutions for $k_\theta(\cdot,\cdot)$. For instance, for $\phi_{\sigma=0}(\cdot) = \delta(\cdot)$ we have that $A\phi_0$ is a step function and $k_\theta(x_i, x_j) = T - \max(f_\theta(x_i), f_\theta(x_j))$.

The boundedness assumption enforces us to use kernels ϕ_σ with a bounded range (hence, Gaussian kernels will not be allowed in this setting). Our experiments indicate that smoothstep functions, often used in computer graphics, are well-suited candidates for $A\phi_\sigma$.

5. GSPM GRADIENT FLOWS

Gradient flows have become increasingly popular in implicit generative modeling [18, 19, 28], where the aim is to minimize the following functional in the space of probability measures with bounded second-order moments:

$$p^\star = \arg\min_p \zeta^2_{\mathcal{F},\rho}(p, q). \tag{8}$$

First, to demonstrate the favorable characteristic of GSPMs in gradient flows consider the problem in (8), where the target distribution, q, is a uniform density on a two-dimensional disc of radius R,

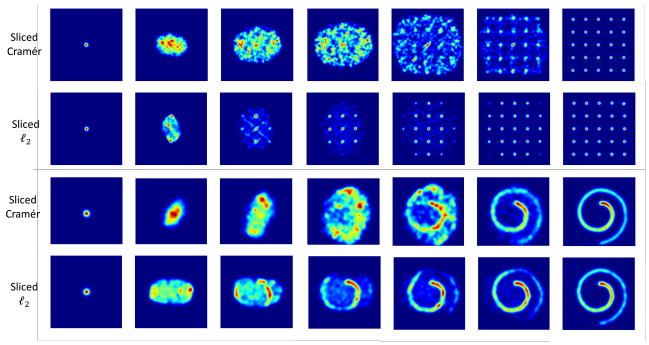

Fig. 2: Gradient flows on the 25-Gaussians and Swiss Roll distributions using GSPM-flows with sliced-ℓ_2 and sliced-Cramér distances.

$p(\cdot) = q(\cdot - x_0)$ is the shifted version of q, and the optimization is performed with respect to x_0 to minimize the distance between p and q. For this simple setting, we calculate the distances between p and q as a function of x_0 for ℓ_2 and the two GSPMs presented in Section 4. Figure 1 shows the distributions p and q on the left and the distances on the right. As expected, it can be seen that when $\|x_0\|_2 > 2R$, *i.e.*, p and q do not overlap, the ℓ_2 distance provides no gradient information, while sliced-ℓ_2 and sliced-Cramér both guarantee a converging flow. Below, we exploit the connections that we developed between GSPMs and MMD (as detailed in Section 4), and build up on the recent results given in [19] to provide theoretical guarantees for *global convergence* of gradient flow algorithms based on GSPMs.

We present the *GSPM-flows*, that aim at generating a path of measures $(p_t)_{t\geq 0}$ which minimizes the squared GSPM between an initial measure p_0 and a target measure q as t goes to infinity. In particular, we will consider the continuity equation, where the divergence of flux is provided by the gradient [29]:

$$\partial_t p_t = -\mathrm{div}(vp_t) = -\nabla_{\mathcal{W}} \zeta^2_{\mathcal{F},\rho}(p_t, q)/2, \tag{9}$$

where $\nabla_{\mathcal{W}}$ denotes a notion of a gradient in the Wasserstein space [29] (*i.e.*, the space of probability measures with bounded second-order moments), $\mathrm{div}(\cdot)$ denotes the divergence operator, and v is the velocity field obtained from [19]:

$$v(x, p) = \nabla_x \left(\int k(z, x)p(z)\mathrm{d}z - \int k(z, x)q(z)\mathrm{d}z \right),$$

where $k(\cdot,\cdot)$ is defined in (6). The partial differential equation (PDE) representation (9) has important practical implications, since such PDEs are often associated with a McKean-Vlasov (MV) process [30], which can be used for developing practical algorithms. In particular, associated to the continuity equation, we can define a MV process $(X_t)_{t\geq 0}$ as a solution to the following differential equation:

$$\mathrm{d}X_t = v(X_t, p_t)\mathrm{d}t, \qquad X_0 \sim p_0, \tag{10}$$

where X_t denotes the state of the process at time t. Here, X_t evolves through the *drift* function v, which requires the knowledge of p_t, *i.e.* the density function of X_t. The probability density functions of $(X_t)_t$ solve the continuity equation, therefore, solving the optimization problem (9) reduces to simulating (10). The exact simulation of (10) is, however, intractable due to (i) the process is continuous-time, so it needs to be discretized, (ii) the drift depends on the density p_t, which might not be available in practice. We will focus on the discretization of the process first, then we will develop a particle-based approach to alleviate the second problem. In order to discretize

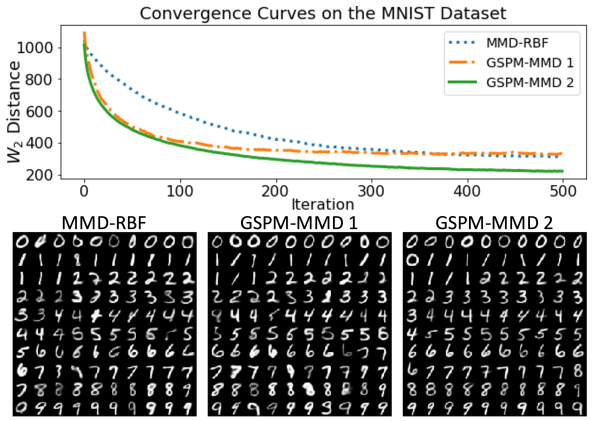

Fig. 3: Comparison of GSPM-flows on the MNIST dataset. The top panel shows the 2-Wasserstein distance between the source and target distributions, while the bottom panel visualizes the target when $N = 100$.

(10), we consider the noisy Euler-Maruyama (EM) scheme, proposed in [19], given as:

$$X_{n+1} = X_n + \eta v(X_n + \beta_n U_n, p_n), \tag{11}$$

where $\eta > 0$ is a step-size, $n \in \mathbb{N}$ denotes the iteration number, p_n is the density of X_n, $\beta_n > 0$ an inverse temperature variable, and U_n a standard Gaussian variable. If $\beta_n = 0$ for all n, this scheme reduces to the standard EM discretization, whereas a positive β_n would drive the scheme to explore the space in a more efficient way. Below we identify sufficient regularity conditions, on the defining function f_θ and the smoothing function ϕ_σ, to ensure convergence of GSPM-flows and their discretization (11), and state our main result.

Condition 1. *A is a linear, bounded, PSD operator with the corresponding operator norm $\|A\|_{op}$.*

Condition 2. *There exists a constant G_f, such that $\forall\theta \in \Omega_\theta$, $\|\nabla f_\theta(x)\| \le G_f$ for all $x \in \mathcal{X}$ and*

$$\|\nabla f_\theta(x) - \nabla f_\theta(y)\| \le G_f\|x - y\|\,, \quad \forall x, y \in \mathcal{X}. \tag{12}$$

Condition 3. *There exists a constant G_ϕ, such that the following inequalities hold: $|\phi_\sigma(\cdot)| \le G_\phi$, $|\phi'_\sigma(\cdot)| \le G_\phi$, $|\phi_\sigma(t) - \phi_\sigma(t')| \le G_\phi|t - t'|$, and $|\phi'_\sigma(t) - \phi'_\sigma(t')| \le G_\phi|t - t'|$.*

Theorem 1. *Let p_0 be a distribution with finite second-order moment. Then, under Conditions 1,2,3, there exists a unique $(X_t)_{t\ge0}$ solving (10) such that the density functions of $(X_t)_{t\ge0}$ constitute the unique solution of (9). Furthermore, let $(X_n)_{n\in\mathbb{N}}$ be the iterates obtained by (11). If $\sum_{i=1}^n \beta_i^2 \to \infty$ as $n \to \infty$, then the following holds:*

$$\zeta(p_n, q) \le \zeta(p_0, q)e^{-2\lambda^2\eta(1-3\eta L)\sum_{i=0}^n \beta_i^2}, \tag{13}$$

where p_n is the density of X_n, $L = (G_f^2 + G_f)G_\phi^2\|A\|_{op}^2$, and $\lambda = \left(2d\|A\|_{op}^2 G_\phi^2 G_f^2(1 + G_f^2)\right)^{1/2}$.

The proof is given in the supplement [20]. This result shows that, with sufficiently regular f_θ and ϕ_σ, the noisy EM scheme (11) can achieve the global optimum, where the convergence rate depends on the structure of f_θ and ϕ_σ. Theorem 1 demonstrates the benefits of the proposed gradient flow, however, the discretization scheme (11) is still intractable due to the dependency of v on p_n. In order to obtain a practical algorithm, we finally consider a *particle system* that serves as an approximation to the original system (11):

$$X_{n+1}^i = X_n^i + \eta v(X_n^i + \beta_n U_n^i, \hat{p}_n), \tag{14}$$

where $i \in \{1, \dots, N\}$ is the particle index and $\hat{p}_n = \frac{1}{N}\sum_{j=1}^N \delta_{X_n^i}$ denotes the *empirical distribution* of $\{X_n^i\}_{i=1}^N$. Here, the idea is to approximate p_n by $\hat{p}_n$ by evolving N different particles at the same time. Similar schemes have proven to be successful in generative modeling [28] and Bayesian machine learning [31].

6. NUMERICAL EXPERIMENTS

We first perform a numerical experiment with synthetic datasets to demonstrate the performance of the proposed GSPM-MMD kernels. To simplify the presentation, in our first experiment, we assumed the noise $\beta_n = 0$ for all n (*i.e.*, the standard EM discretization). We consider two target distributions in $\mathbb{R}^2$, namely the Swiss Roll and the 25-Gaussians distributions. The source distribution is initialized with N samples from a Gaussian distribution. Figure 2 shows the datasets and the flow (calculated using GSPM-MMD), we use Kernel Density Estimation (KDE) for better visualization. We calculate the gradient flow updates (see Equation (14)) to match the source and target distributions. We used sliced-ℓ_2 and sliced-Cramér to calculate the flows, and for its simplicity, the slices are linear $f_\theta(x) = \langle x, \theta\rangle$. Figure 2 shows the calculated flows. The full numerical analysis corresponding to the gradient-flows shown in Figure 2, including the sensitivity to the choice of σ, and the comparison with MMD-RBF (MMD with RBF kernel) is in the supplementary material [20].

To show the effectiveness of the proposed distances in higher dimensions, we designed the following experiment. We first learn a simple convolutional auto-encoder, with an added classifier on its bottleneck ensuring a discriminative space embedding, to embed the MNIST dataset into a 16-dimensional space. Then, we solve the gradient flow problem in the embedded space with $N = 100$ particles initialized from a Gaussian distribution. We compare gradient-flows based on MMD-RBF, sliced-ℓ_2, and sliced-Cramér. We use the 2-Wasserstein distance, $\mathcal{W}_2(p_t, q)$, as an objective measure of convergence. The experiments were repeated 10 times and the average performance for each method is reported in Figure 3 (top panel). Once convergence is reached, we sort the particles according to the output of the classifier and feed them to the decoder network to visualize the corresponding digits for each method: see Figure 3 (bottom panel). We note that the same σ was used for all three methods, with a fixed step-size η. Linear slicing was used for GSPM-flows. We conclude that, similar to Figure 1, the GSPM-flows with sliced-Cramér, seems to achieve a superior performance as compared to others. We refer to the supplementary material for more extensive experimental results, including comparison between linear and nonlinear slices, as well as studying the effect of noise, β.

7. CONCLUSION

We introduced GSPMs, which calculate the expected distances between slices of two input distributions, and showed that a subset of the proposed distances is equivalent to MMD with novel kernels defined in this work. Furthermore, we applied GSPMs in the domain of gradient flows and identified sufficient regularity conditions on the building elements of our proposed distances for guaranteeing global convergence of the gradient flow. Finally, we provided numerical experiments demonstrating the benefits of slicing distributions, and showed the consistent behavior of GSPM-flows on various datasets.

8. REFERENCES

[1] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[2] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi, "On wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, pp. 47, 2017.

[3] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton, "Fast two-sample testing with analytic representations of probability measures," in *Advances in Neural Information Processing Systems*, 2015, pp. 1981–1989.

[4] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 2015, pp. 258–267.

[5] Shakir Mohamed and Balaji Lakshminarayanan, "Learning in implicit generative models," *arXiv preprint arXiv:1610.03483*, 2016.

[6] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos, "Mmd gan: Towards deeper understanding of moment matching network," in *Advances in Neural Information Processing Systems*, 2017, pp. 2203–2213.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[8] Cédric Villani, *Optimal transport: old and new*, vol. 338, Springer, 2009.

[9] Aude Genevay, Lenaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré, "Sample complexity of sinkhorn divergences," in *Proc. AISTATS'19*, 2019.

[10] Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde, "Sliced wasserstein auto-encoders," in *International Conference on Learning Representations*, 2019.

[11] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau, "Asymptotic guarantees for learning generative models with the sliced-wasserstein distance," in *Advances in Neural Information Processing Systems*, 2019, pp. 250–260.

[12] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot, "Wasserstein barycenter and its application to texture mixing," in *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2011, pp. 435–446.

[13] Soheil Kolouri, Yang Zou, and Gustavo K Rohde, "Sliced wasserstein kernels for probability distributions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5258–5267.

[14] Mathieu Carriere, Marco Cuturi, and Steve Oudot, "Sliced wasserstein kernel for persistence diagrams," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 664–673.

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

[16] Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann, "Sliced wasserstein distance for learning gaussian mixture models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3427–3436.

[17] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli, "Approximate bayesian computation with the sliced-wasserstein distance," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 5470–5474.

[18] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K Rohde, "Generalized sliced wasserstein distances," in *Advances in Neural Information Processing Systems*, 2019.

[19] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton, "Maximum mean discrepancy gradient flow," in *Advances in Neural Information Processing Systems*, 2019.

[20] "Supplementary material," `https://www.di.ens.fr/~simsekli/icassp2022/`, 2021.

[21] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola, "A kernel method for the two-sample-problem," in *Advances in neural information processing systems*, 2007, pp. 513–520.

[22] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, et al., "Equivalence of distance-based and rkhs-based statistics in hypothesis testing," *The Annals of Statistics*, vol. 41, no. 5, pp. 2263–2291, 2013.

[23] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf, "Wasserstein auto-encoders," in *International Conference on Learning Representations*, 2018.

[24] Szymon Knop, Jacek Tabor, Przemysław Spurek, Igor Podolak, Marcin Mazur, and Stanisław Jastrzebski, "Cramer-wold autoencoder," 2018.

[25] Harald Cramér, "On the composition of elementary errors: First paper: Mathematical deductions," *Scandinavian Actuarial Journal*, vol. 1928, no. 1, pp. 13–74, 1928.

[26] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos, "The cramer distance as a solution to biased wasserstein gradients," *arXiv preprint arXiv:1705.10743*, 2017.

[27] Soheil Kolouri, Nicholas A. Ketz, Andrea Soltoggio, and Praveen K. Pilly, "Sliced cramer synaptic consolidation for preserving deeply learned representations," in *International Conference on Learning Representations*, 2020.

[28] A. Liutkus, U. Şimşekli, S. Majewski, A. Durmus, and F.-R. Stoter, "Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions," in *International Conference on Machine Learning (ICML)*, 2019, pp. 4104–4113.

[29] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[30] Vladimir I Bogachev, Nicolai V Krylov, Michael Röckner, and Stanislav V Shaposhnikov, *Fokker-Planck-Kolmogorov Equations*, vol. 207, American Mathematical Soc., 2015.

[31] Qiang Liu and Dilin Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," in *Advances in neural information processing systems*, 2016, pp. 2378–2386.