Statistical and Topological Properties of Gaussian Smoothed Sliced Probability Divergences

Alain Rakotomamonjy* Criteo AI Lab

Mokhtar Z. Alaya* LMAC, UTC

Maxime Bérar LITIS, URN

Gilles Gasso LITIS, INSA-Rouen

Abstract

Gaussian smoothed sliced Wasserstein distance has been recently introduced for comparing probability distributions, while preserving privacy on the data. It has been shown, in applications such as domain adaptation, to provide performances similar to its non-private (non-smoothed) counter-part. However, the computational and statistical properties of such a metric is not yet been well-established. In this paper, we analyze the theoretical properties of this distance as well as those of generalized versions denoted as Gaussian smoothed sliced divergences. We show that smoothing and slicing preserve the metric property and the weak topology. We also provide results on the sample complexity of such divergences. Since, the privacy level depends on the amount of Gaussian smoothing, we analyze the impact of this parameter on the divergence. We support our theoretical findings with empirical studies of Gaussian smoothed and sliced version of Wassertein distance, Sinkhorn divergence and maximum mean discrepancy (MMD). In the context of privacy-preserving domain adaptation, we confirm that those Gaussian smoothed sliced Wasserstein and MMD divergences perform very well while ensuring data privacy.

1 Introduction

Divergences for comparing two distributions have been shown to be important for achieving good performances in the contexts of generative modeling [1, 27], domain adaptation [18, 5, 17], and in computer vision [3, 28] among many more applications [14, 24]. Examples of divergences that have been proven to be useful for those tasks are maximum mean discrepancy [11, 18, 29], Wasserstein distance [19, 12, 30] or its variant the sliced Wasserstein distance (SWD) [2, 15, 21, 13].

Sliced Wasserstein distance has the advantage of being computationally efficient as it exploits a closed-form solution for distributions with support on $\mathbb{R}$, by computing the expectation of one-dimensional random projections of distributions in $\mathbb{R}^d$. Owing to this efficiency and the resulting scalability, this distance has been successfully applied in several applications ranging from generative models to domain adaptation [16, 7, 31, 17] and its statistical property has been well-studied [20].

A differentially private variant of sliced Wasserstein distance has been recently introduced in [26], for comparing distributions in relation to sensitive applications in which training data can not be disclosed. Privacy through a so-called Gaussian mechanism is induced by adding Gaussian noise to each 1D projection of each distribution, leading to the so-called Gaussian-smoothed sliced Wasserstein distance. This relationship between Gaussian smoothing and privacy has also been mentioned by Nietert et al. [22] as future work to address, while they analyzed the structural and statistical behavior of Gaussian smoothed Wasserstein distances.

However, up to now, theoretical properties of this Gaussian smoothed sliced Wasserstein distance are not well fully understood except for its metric properties [26]. In this work, we investigate those theoretical properties and the one of more general Gaussian smoothed sliced divergences. Indeed, given a base distance or divergence for distributions in $\mathbb{R}^d$, we can introduce its related Gaussian smoothed sliced divergence. Specifically, the theoretical properties of interest are the metric property and the underlying topology. From a statistical point of view, we seek at understanding the relationship between the sample complex-

Preliminary work. Under review.
* Equal contribution

ity of the base divergence and its Gaussian smoothed sliced version. Regarding privacy, the role of the Gaussian smoothing is of primary importance as it induces the privacy level achieved by the divergence. Hence, we also provide an analysis on its impact with respect to the standard deviation of the Gaussian noise. For supporting our theoretical study, we provide some numerical experiments on toy problem, and we also provide some numerical study on domain adaptation illustrating how owing to the topology induced by our metric, differential privacy comes almost for free (without loss of performances) in this context.

The paper is organized as follows, after introducing the notations and some background in Section 2, we detail the topological properties of Gaussian smoothed sliced divergence in Section 3.1 while the statistical properties are established in Section 3.2. Experimental analyses for supporting the theory and showcasing the relevance of our divergences in a domain adaptation situation are depicted in Section 4. Discussions on the perspectives and limitations are in Section 5.

2 Preliminaries

For the reader's convenience, we provide a brief summary of the standard notations and the definitions that will be used throughout the paper.

Notations. For $d \in \mathbb{N}^*$, let $\mathcal{P}(\mathbb{R}^d)$ be the set of Borel probability measures on $\mathbb{R}^d$ and $\mathcal{P}_p(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$, those with finite moment of order p, i.e., $\mathcal{P}_p(\mathbb{R}^d) \triangleq \{\mu \in \mathcal{P} : \int \|x\|^p d\mu(x) < +\infty\}$, where $\|\cdot\|$ is the Euclidean norm and $\langle \cdot, \cdot \rangle$ is the Euclidean inner-product. For two probability distributions μ and ν, we denote their convolution as $\mu * \nu \in \mathcal{P}(\mathbb{R}^d)$ and by definition, we have $(\mu * \nu)(A) = \int_x \int_y \mathbf{1}_A(x+y) d\mu(x) d\nu(y)$, where $\mathbf{1}_A(\cdot)$ is the indicator function over A. Given two independent random variables $X \sim \mu$ and $Y \sim \nu$, we remind that $X + Y \sim \mu * \nu$.

The d-dimensional unit-sphere is noted as $\mathbb{S}^{d-1} \triangleq \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$. We denote by u_d the uniform distribution on $\mathbb{S}^{d-1}$ and we use $\delta(\cdot)$ to denote the Kronecker delta function. We note as $\mathbf{E}_\mu f$ the expectation of the function f with respect to μ. Hence, the characteristic function of a probability distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ is $\varphi_\mu(t) = \mathbf{E}_\mu[e^{iX^\top t}]$. Given this definition, similarly to the Fourier transform, the characteristic function of the convolution of two probability distributions has the following form $\varphi_{\nu * \mu}(t) = \varphi_\nu(t) \cdot \varphi_\mu(t)$.

Sliced Wasserstein Distance. We remind in this paragraph several measures of similarity between two distributions. The Wasserstein distance of order $p \in [1, \infty)$ between two measures in $\mathcal{P}_p(\mathbb{R}^d)$ is given by the relaxation of the optimal transport problem, and it is defined as

$$W_p^p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^p \gamma(x, x') dx dx'$$

where $\Pi(\mu, \nu) \triangleq \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) | \pi_{1\#}\gamma = \mu, \pi_{2\#}\gamma = \nu\}$ and π_1, π_2 are the marginal projectors of γ on each of its coordinates. When $d = 1$, the Wasserstein distance can be computed in a closed-form owing to the cumulative distributions of μ and ν [25]. Note that the superscript in W_p^p refers to the power p. In practice for empirical distributions, the closed-form solution needs just the sorting of samples, which makes it very efficient. Due to this efficiency, efforts have been devoted to derive a metric for high-dimensional distributions based on 1D Wasserstein distance. The main idea is to project high-dimensional probability distributions onto a random 1-dimensional space and then to compute the Wasserstein distance. That operation can be theoretically formalized through the use of the Radon transform, leading to the so-called sliced Wasserstein distance [2, 15, 21, 13].

Definition 1. *For any $p \in [1, \infty)$ and two measures μ, $\nu \in \mathcal{P}_p(\mathbb{R}^d)$, the sliced Wasserstein distance (SWD) reads as*

$$SWD_p^p(\mu, \nu) \triangleq \int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}_\mathbf{u}\mu, \mathcal{R}_\mathbf{u}\nu) u_d(\mathbf{u}) d\mathbf{u}.$$

where $\mathcal{R}_\mathbf{u}$ is the Radon transform of a probability distribution, namely $\mathcal{R}_\mathbf{u}\mu(\cdot) = \int_{\mathbb{R}^d} \mu(\mathbf{s}) \delta(\cdot - \mathbf{s}^\top \mathbf{u}) d\mathbf{s}$.

In practice, the integral is approximated through a Monte-Carlo simulation leading to a sum of 1D Wasserstein distances over a fixed number of random directions $\mathbf{u}$.

Gaussian Smoothed Sliced Wasserstein Distance. Based on this definition of SWD, replacing the Radon projected measures with their Gaussian-smoothed counterpart leads to the following definition:

Definition 2. *The σ-Gaussian smoothed p-Sliced Wasserstein distance between probability distributions μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$ is*

$$G_\sigma SWD_p^p(\mu, \nu) \triangleq \int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma) u_d(\mathbf{u}) d\mathbf{u}.$$

It is important to note here that the smoothing (convolution) operation occurs after projection onto the one-dimensional space. Hence, assuming $X \sim \mu$, $Y \sim \nu$ in the integral, for a given $\mathbf{u}$, we compute the 1D Wasserstein distance between the probability laws of $\mathbf{u}^\top X + Z$ and $\mathbf{u}^\top Y + Z'$ with $Z, Z' \sim \mathcal{N}_\sigma$ being independent random variables. The metric properties of $G_\sigma \mathrm{SWD}_p^p$ for $p \geq 1$, of this Gaussian smoothed sliced Wasserstein distance have been discussed in a recent work [26]. This latter work has also shown, in the context of differential privacy, the importance of convolving the Radon projected distribution with a Gaussian instead of computing the sliced Wasserstein distance of the original distribution smoothed with a d-dimensional Gaussian $\mu * \mathcal{N}_\sigma$.

Gaussian Smoothed Sliced Divergence. The idea of slicing high-dimensional distributions before feeding them to a divergence between probability distributions can be extended to other distance than Wasserstein distance. Those sliced divergences have been studied by [20]. In a similar way, we can define a Gaussian smoothed sliced divergence, given a divergence $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^+$ for $d \geq 1$ as:

Definition 3. *The σ-Gaussian smoothed p-Sliced Divergence between probability distributions μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$ associated to the divergence $D \triangleq D_\mathbb{R}$, $p \geq 1$ is*

$$G_\sigma SD^p(\mu, \nu) \triangleq \int_{\mathbb{S}^{d-1}} D^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma) u_d(\mathbf{u}) d\mathbf{u}.$$

where the superscript p refers to a power.

Typical relevant divergence is the maximum mean discrepancy (MMD) [11] or the Sinkhorn divergence [10, 24]. In Section 4, we report empirical findings based on these divergences as well as on the Wasserstein distance.

3 Theoretical Properties

In this section, we will analyze the properties of the Gaussian smoothed sliced divergence, in term of topological and statistical properties and the influence of the Gaussian smoothing parameter σ on the distance.

3.1 Topology

It has already been shown in [26] that the Gaussian smoothed sliced Wasserstein is a metric on $\mathcal{P}(\mathbb{R}^d)$. In the next, we extend these results to any divergence $D(\cdot, \cdot)$ under some assumptions.

Theorem 1. *For any $p \in [1, \infty)$ and $\sigma > 0$, the following properties hold:*

1. *if $D(\cdot, \cdot)$ is non-negative (or symmetric), then $G_\sigma SD^p(\cdot, \cdot)$ is non-negative (or symmetric);*
2. *if the base divergence $D(\cdot, \cdot)$ satisfies the identity of indiscernibles, for $\mu', \nu' \in \mathcal{P}(\mathbb{R})$, $D(\mu', \nu') = 0$ if and only if $\mu' = \nu'$, then this identity also holds for $G_\sigma SD^p(\cdot, \cdot)$ for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$;*
3. *if the $D(\cdot, \cdot)$ satisfies the triangle inequality then its Gaussian smoothed sliced version $G_\sigma SD^p(\cdot, \cdot)$ satisfies the triangle inequality.*

The above theorem shows that under mild hypotheses over the base divergence D, as being a metric for instance, the metric property of its Gaussian smoothed sliced version naturally derives. As exposed in the appendix, the more involved property to prove is the identity of indiscernibles.

Now, we establish under which conditions on the divergence D, the convergence of a sequence in $G_\sigma SD^p$ implies weak convergence in $\mathcal{P}(\mathbb{R}^d)$.

Theorem 2. *Let $\sigma \geq 0, p \in [1, \infty)$, $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, the sequence of distributions $\{\mu_k \in \mathcal{P}_p(\mathbb{R}^d)\}_{k \geq 1}$. Assume that the divergence D metricizes the weak topology. Then, $\lim_{k\to\infty} G_\sigma SD^p(\mu_k, \mu) = 0$ if and only if $\{\mu_k\}_k$ converges weakly to μ* i.e., *if for any f in the set of bounded and continuous functions, $\mu_k \to \mu$ if $\int_{\mathbb{R}^d} f(x) d\mu_k(x) \to \int_{\mathbb{R}^d} f(x) d\mu(x)$.*

Proof. By using results from [20], we know that if D metricizes the weak topology for $\mathcal{P}(\mathbb{R})$ then the weak convergence in $\mathcal{P}(\mathbb{R})$ is equivalent to the convergence under D. Hence, we have

$$G_\sigma SD^p(\mu_k, \mu) \to 0 \Leftrightarrow \mu_k * \mathcal{N}_\sigma \to \mu * \mathcal{N}_\sigma$$

then, using the convolution property of characteristic function gives

$$\varphi_{\mu_k}(t)\varphi_{\mathcal{N}_\sigma}(t) \to \varphi_\mu(t)\varphi_{\mathcal{N}_\sigma}(t) \quad \forall t.$$

This means that for all t, $\varphi_{\mu_k}(t) \to \varphi_\mu(t)$ which concludes the proof, owing to the one-to-one correspondence between characteristic functions. □

3.2 Statistical properties

The next theoretical question we are interested in is about the error we made when the true distribution μ is approximated by its empirical distribution $\hat{\mu}$. Such a case is common in practical applications where only (high-dimensional) empirical samples are at disposal. Specifically, we are interested in quantifying two key properties of empirical Gaussian smoothed divergence: *(i)* the convergence of $G_\sigma SD^p(\hat{\mu}_n, \hat{\nu}_n)$ to $G_\sigma SD^p(\mu, \nu)$ *(ii)* the convergence of $\widehat{G_\sigma SD}^p(\mu, \nu)$ to $G_\sigma SD^p(\mu, \nu)$, i.e., when approximating the expectation over the random projection with sample mean.

3.2.1 Sample complexity

Herein, our goal is to quantify the error made when approximating $G_\sigma SD^p(\mu, \nu)$ with $G_\sigma SD^p(\hat{\mu}_n, \hat{\nu}_n)$, where $\hat{\mu}_n, \hat{\nu}_n$ are the empirical counterparts of μ, ν defined over n samples. More precisely, we are interested in establishing an order of the convergence rate of $G_\sigma SD^p(\hat{\mu}_n, \hat{\nu}_n)$ towards $G_\sigma SD^p(\mu, \nu)$, according to the number of samples n. This rate stands for the so-called *sample complexity.*

Theorem 3. *Fix $p \in [1, \infty)$ and assume that for any $\mu' \in \mathcal{P}(\mathbb{R})$ with empirical measure $\hat{\mu}_n^{'}$, $\mathbf{E}[D^p(\hat{\mu}_n^{'}, \mu')] \leq \alpha_n(p)$. Then, for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ with empirical measure $\hat{\mu}_n$,*

$$\mathbf{E}[G_\sigma SD^p(\hat{\mu}_n, \mu)] \leq \alpha_n(p).$$

Additionally, if D^p is a pseudo-metric (non-negative, symmetric with triangle inequality), then

$$\mathbf{E}[|G_\sigma SD^p(\hat{\mu}_n, \hat{\nu}_n) - G_\sigma SD^p(\mu, \nu)|] \leq 2\alpha_n(p).$$

Proof. We have

$$\begin{aligned}
&\mathbf{E}[G_\sigma SD^p(\hat{\mu}_n, \mu)] \\
&\quad = \mathbf{E}\Big[\int_{\mathbb{S}^{d-1}} D^p(\mathcal{R}_\mathbf{u}\hat{\mu}_n * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma) u_d(\mathbf{u}) \mathrm{d}\mathbf{u}\Big] \\
&\quad \leq \int_{\mathbb{S}^{d-1}} \mathbf{E}\big[D^p(\mathcal{R}_\mathbf{u}\hat{\mu}_n * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma) u_d(\mathbf{u}) \mathrm{d}\mathbf{u}\big] \\
&\quad \leq \int_{\mathbb{S}^{d-1}} \alpha_n(p) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} = \alpha_n(p).
\end{aligned}$$

The triangle inequality entails that, $G_\sigma SD^p(\hat{\mu}_n, \hat{\nu}_n) \leq G_\sigma SD^p(\hat{\mu}_n, \mu) + G_\sigma SD^p(\mu, \nu) + G_\sigma SD^p(\nu, \hat{\nu}_n)$, which entails, by taking expectation with respect to $\hat{\mu}_n, \hat{\nu}_n$,

$$\begin{aligned}
&\mathbf{E}\big[|G_\sigma SD^p(\hat{\mu}_n, \hat{\nu}_n) - G_\sigma SD^p(\mu, \nu)|\big] \\
&\qquad \leq \mathbf{E}\big[G_\sigma SD^p(\hat{\mu}_n, \mu)\big] + \mathbf{E}\big[G_\sigma SD^p(\nu, \hat{\nu}_n)\big] \\
&\qquad \leq \mathbf{E}\big[G_\sigma SD^p(\hat{\mu}_n, \mu)\big] + \mathbf{E}\big[G_\sigma SD^p(\hat{\nu}_n, \nu)\big] \\
&\qquad \leq 2\alpha_n(p),
\end{aligned}$$

which completes the proof. □

Remark 1. *Given any base divergence D^p, Theorem 3 shows that the sample complexity of $G_\sigma SD^p$ is proportional to the one dimensional sample complexity of D^p.*

Next, we focus on the sample complexity for the special case of Gaussian smoothed sliced Wasserstein distance. We also provide the convergence rate of $G_\sigma SWD^p(\hat{\mu}_n, \hat{\nu}_n)$ towards $G_\sigma SWD^p(\mu, \nu)$.

Proposition 1. *For any $p, q \in [1, \infty)$ such that $q > p$, consider $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$ with its empirical measure $\hat{\mu}_n$. Then, the following holds*

$$\mathbf{E}[|G_\sigma SWD^p(\hat{\mu}_n, \hat{\nu}_n) - G_\sigma SWD^p(\mu, \nu)|] \leq \alpha_n(p, q, \sigma)$$

where

$$\begin{aligned}
&\alpha_n(p, q, \sigma) \\
&\quad = 2C_{p,q} \times \begin{cases} 2^{p(q-1)/q}(M_q(\mu, \nu) + M_q(\mathcal{N}_\sigma))^{p/q}\mathbf{1}_{q \in 2\mathbb{N}^*}, \\ 2^{p(q-1)/q} M_q(\mu, \nu))^{p/q}\mathbf{1}_{q \in 2\mathbb{N}+1}, \end{cases} \\
&\qquad\qquad \times \begin{cases} n^{-1/2}\mathbf{1}_{q > 2p}, \\ n^{-1/2}\log(n)\mathbf{1}_{q = 2p} \\ n^{-(q-p)/q}\mathbf{1}_{q \in (p, 2p)} \end{cases}
\end{aligned}$$

and where $M_q(\mu, \nu) = M_q(\mu) + M_q(\nu)$, $C_{p,q}$ is a positive constant depending only p, q, and $M_q(\mathcal{N}_\sigma)$ stands for the q-th moment of $\mathcal{N}_\sigma$, that is

$$M_q(\mathcal{N}_\sigma))\mathbf{1}_{q \in 2\mathbb{N}^*} = \frac{(2q)!}{2^q q!}\sigma^{2q} = 1 \cdot 2 \cdot 3 \cdots (2q-1)\sigma^{2q}.$$

The latter theoretical results show that empirical Gaussian smoothed Wasserstein distance converges at a rate of order $n^{-1/2}$ in the best scenario. It is worth also noting that the sample complexity depends on the amount of smoothing through the moment of the Gaussian noise : the larger the amount of smoothing, the worse is the constant of the complexity.

3.2.2 Projection complexity

To compute the Gaussian smoothed sliced divergence, one may resort to a Monte Carlo scheme to numerically approximate the integral in $G_\sigma SD^p(\mu, \nu)$. Towards this, let define the following sum:

$$\widehat{G_\sigma SD}^p(\mu, \nu) = \frac{1}{L}\sum_{l=1}^{L} D^p(\mathcal{R}_{\mathbf{u}_l}\hat{\mu}_n * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}_l}\hat{\nu}_n * \mathcal{N}_\sigma),$$

where $\mathbf{u}_l$ is a random vector uniformly drawn from $\mathbb{S}^{d-1}$, for $l = 1, \ldots, L$. Theorem 4 shows that for a fixed dimension d, the root mean square error of Monte Carlo approximation is of order $O\left(\frac{1}{\sqrt{L}}\right)$, which corresponds to the projection complexity.

Theorem 4. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and fix $p \in [1, \infty)$. Then, the error related to the Monte Carlo estimation of $G_\sigma SD^p$ is bounded as follows*

$$\mathbf{E}[|\widehat{G_\sigma SD}^p(\mu, \nu) - G_\sigma SD^p(\mu, \nu)|] \leq \frac{A(p, \sigma)}{\sqrt{L}},$$

where $A^2(p, \sigma) = \int_{\mathbb{S}^{d-1}} \big(D^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma) - \bar{\vartheta}_p\big)^2 du_d(\mathbf{u})$, with $\bar{\vartheta}_p = \int_{\mathbb{S}^{d-1}} D^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma) du_d(\mathbf{u})$.

The term $A^2(p, \sigma)$ corresponds to the variance of $D^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma)$ with respect to $\mathbf{u} \sim u_d$ drawn according to the uniform distribution over the unit-sphere $\mathbb{S}^{d-1}$. It is worth to note that the precision

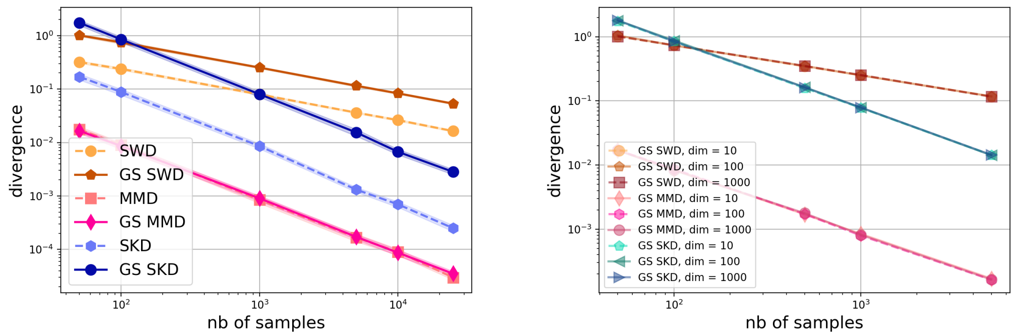

Figure 1: Measuring the divergence between two sets of samples in $\mathbb{R}^{50}$, of increasing size, randomly drawn from $\mathcal{N}(0, \mathbf{I})$. We compare three sliced divergences and their Gaussian smoothed versions with a $\sigma = 3$. (left) dimension has been set to $d = 50$. (right) sample complexity with different dimensions. This plot confirms that the complexity is dimension-independent.

of the Monte Carlo scheme approximation depends on the number of projections L and the variance of the evaluations of the divergence D^p. The estimation error decreases at the rate $L^{-1/2}$ according to the number of projections used to compute the smoothed sliced divergence.

Remark 2. *Given the above results, we can provide a finer analysis of the Gaussian smoothed SWD sample complexity. For any $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$, the overall complexity of the Gaussian smoothed sliced Wasserstein distance is bounded by the sample and projection complexities, that is,*

$$\text{complexity}(G_\sigma SWD^p(\mu, \nu)) = O\Big(\alpha_n(p, q, \sigma) + \frac{A(p, \sigma)}{\sqrt{L}}\Big).$$

If we consider the number of projections as $L = \lfloor n^\beta \rfloor$ for some $\beta \in (0, 1)$ then the overall complexity $\text{complexity}(G_\sigma SD^p(\mu, \nu)) = O(n^{-\beta/2})$. *We further mention that complexity is "interestingly" independent of the dimension d.*

3.3 Noise-level dependencies

When considered in sensitive applications requiring privacy preserving, the parameter σ of the Gaussian smoothing function $\mathcal{N}_\sigma$ may significantly influence the attained privacy level. Hence, we provide theoretical results analyzing the effect of the noise level σ on the induced Gaussian smoothed sliced divergence.

Order relation. We first show that the noise level tends to reduce the difference between two distributions as measured using $G_\sigma \mathrm{SD}^p(\mu, \nu)$ provided the base divergence D satisfies some mild assumptions.

Proposition 2. *Let μ and ν two distributions in $\mathcal{P}(\mathbb{R}^d)$ and consider the noise levels σ_1, σ_2 such that $0 \le \sigma_1 \le \sigma_2 < \infty$. Assume that the base divergence D satisfies*

$$D^p(\mu' * \mathcal{N}_{\sigma_2}, \nu' * \mathcal{N}_{\sigma_2}) \le D^p(\mu' * \mathcal{N}_{\sigma_1}, \nu' * \mathcal{N}_{\sigma_1}),$$

for any $\mu', \nu' \in \mathcal{P}(\mathbb{R})$. Then,

$$G_{\sigma_2} SD^p(\mu, \nu) \le G_{\sigma_1} SD^p(\mu, \nu).$$

Proof. For all $\mathbf{u} \in \mathbb{S}^{d-1}$ we have $\mathcal{R}_\mathbf{u}\mu, \mathcal{R}_\mathbf{u}\nu \in \mathcal{P}(\mathbb{R})$. By application of the inequality of noise level satisfied by D_p^p in one dimension we get

$$D^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_{\sigma_2}, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_{\sigma_2}) \le D^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_{\sigma_1}, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_{\sigma_1}).$$

Then, computing the expectation over the projections $\mathbf{u}$ since the divergence is non-negative concludes the proof. □

Note that the assumption for the base divergence inequality holds for the Gaussian smoothed Wasserstein distance [22]. While we conjecture that it holds also for smoothed Sinkhorn and MMD, we leave the proofs for future works.

Based on the property in Proposition 2, we can show some specific properties of the metric with respect to the noise level σ.

Proposition 3. *$G_\sigma SD^p(\mu, \nu)$ is decreasing with respect to σ and we have*

$$\lim_{\sigma \to 0} G_\sigma SD^p(\mu, \nu) = D^p(\mu, \nu).$$

Proof. The proof comes straightforwardly from Proposition 2 by taking $\sigma_2 = 0$ and letting $\sigma_1 \to 0$. □

This property interestingly states that the $G_\sigma \mathrm{SD}^p$ recovers the sliced divergence when the noise level vanishes. We end up this section by providing a relation between Gaussian smoothed sliced Wasserstein

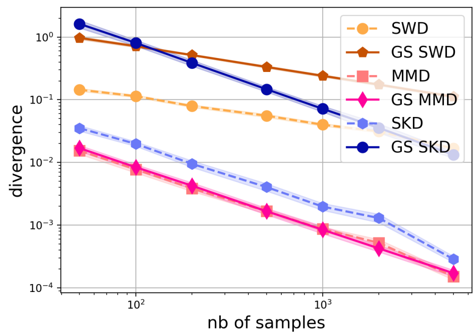

Figure 2: Measuring the divergence between two sets of samples drawn iid from the CIFAR10 dataset. We compare three sliced divergences and their Gaussian smoothed versions with a $\sigma = 3$.

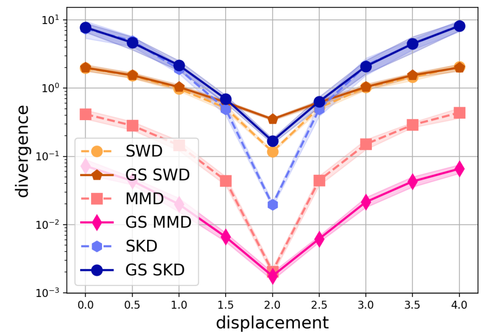

Figure 3: Measuring the divergence between two sets of samples in $\mathbb{R}^{50}$, one with mean $2\mathbf{1}_d$ and the other with mean $s\mathbf{1}_d$ with increasing s. We compare three sliced divergences and their Gaussian smoothed version with a $\sigma = 3$.

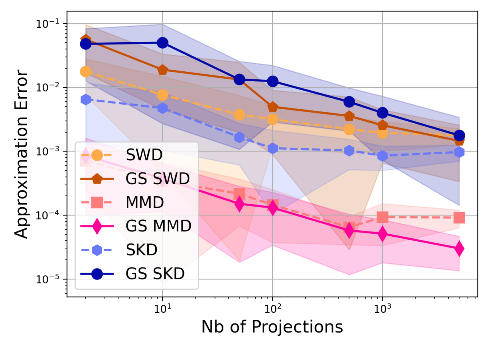

Figure 4: Absolute difference between the approximated Monte-carlo approximation of all divergences compared to the true one (evaluated with $10,000$ number of projections). The two sets of 500 samples in $\mathbb{R}^{50}$ are randomly drawn from $\mathcal{N}(0, \mathbf{I})$. The Gaussian smoothed divergences are parameterized with $\sigma = 3$.

distances under two noise levels. Proof of Proposition 4 is postponed to the appendix.

Proposition 4. *Let $0 \leq \sigma_1 \leq \sigma_2$ be two noise levels. Then, one has*

$$G_{\sigma_1} SWD^p(\mu, \nu) \leq 2^{p-1} G_{\sigma_2} SWD^p(\mu, \nu) + \frac{2\pi^{d/2}}{\Gamma(d/2)} 2^{3p/2} (\sigma_2^2 - \sigma_1^2)^{2p},$$

where $\Gamma : \mathbb{R} \to \mathbb{R}$ is the Gamma function expressed as $\Gamma(v) = \int_0^\infty t^{v-1} e^{-t} dt$.

The above proposition allows to control the variation of the $G_\sigma SWD$ divergence with respects to the amount of Gaussian smoothing.

Remark 3. *All these properties hold for the population case. When considering empirical approximation of the true distribution, it may not hold due to the impact of $\mathcal{N}_\sigma$ over the sample complexity.*

4 Numerical Experiments

In this section, we report on a serie of experiments that support the theoretical results established in the previous section. We also highlight the usefulness of the findings in a context of privacy- preserving domain adaptation problem.

4.1 Supporting the theorical results

Sample complexity The first experiment (see Figure 1) analyzes the sample complexity of the different Gaussian smoothed sliced divergences. It shows that the sample complexity stays similar to the one of their original and sliced counterparts up to a constant (see Theorem 3). For this purpose, we have considered samples in $\mathbb{R}^d$ randomly drawn from a Normal distribution $\mathcal{N}(0, \mathbf{I})$. For the Sinkhorn divergence, the entropy regularization has been set to 0.1 and for MMD, we used a Gaussian kernel for which the bandwidth has been set to the mean of all pairwise distances between samples. The number of projections has been fixed to $L = 50$ and we perform 20 runs per experiment. For the first study, the convergence rate has been evaluated by increasing the samples number up to 25,000 with fixed dimension $d = 50$. For the second one, we vary both the dimension and the number of samples.

Figure 1 shows the sample complexity of some sliced divergences, respectively noted as SWD, SKD and MMD for Sliced Wasserstein distance, Sinkhorn divergence and Maximum Mean discrepancy) and their

Gaussian-smoothed version, named as GS SWD, GS SKD and GS MMD. On the left plot, we can see that all Gaussian smoothed divergences preserve the complexity rate with just a slight to moderate overhead. The worst difference is for Sinkhorn divergence, while smoothed MMD almost comes for free in term of complexity. From the right plot where sample complexities for different dimensions d are given, we confirm the finding that Gaussian smoothing keeps the independence of the convergence rate to the dimension of sliced divergences. We have also evaluated the sample complexity for the CIFAR dataset by sampling sets of increasing size. Results reported in Figure 2 confirms the findings obtained from the toy dataset.

Identity of indescernibles The second experiment aims at checking whether our divergences converge towards a small value when the distributions to be compared are the same. For this, we consider samples from distributions μ and ν chosen as normal distributions with respectively mean $2 \times \mathbf{1}_d$ and $s\mathbf{1}_d$ with varying s (noted as the displacement). Results are depicted in Figure 3. We can see that all methods are able to attain their minimum when $s = 2$. Interestingly, the gap between the Gaussian smoothed and non-smoothed divergences for Wasserstein and Sinkhorn is almost indiscernible as the distance between distribution increases.

Projection complexity We have also investigated the impact of the number of projections when estimating the distance between two sets of 500 samples drawn from the same distribution, $\mathcal{N}(0, \mathbf{I})$. Figure 4 plots the approximation error between the true expectation of the sliced divergences (computed for a number of $L = 10,000$ projections) and its approximated versions. We remark that, for all methods, the error ranges within 10-fold when approximating with 50 projections and decreases with the number of projections.

Impact of the noise parameter. Since the Gaussian smoothing parameter is key in a privacy preserving context, as it impacts on the level of privacy of the Gaussian mechanism, we have analyzed its impact of the smoothed sliced divergence. We have reproduced the experiment for the sample complexity but with different values of σ. The number of projections has been set to 50. Figure 5 shows these sample complexities. The first very interesting point to note is that the smoothing parameter has almost no effect on the MMD sample complexity. For the Gaussian smoothed SWD and Sinkhorn divergences, instead, the smoothing tends to increase the divergence at fixed number of samples. Another interpretation is that to achieve a given value of divergence, one needs more far samples

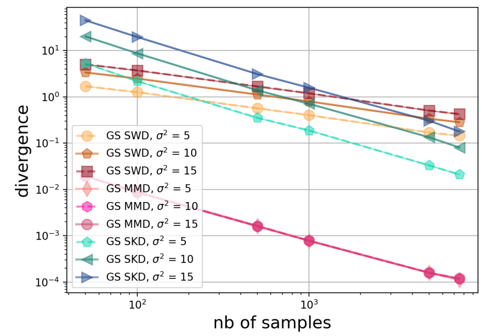

Figure 5: Measuring the divergence between two sets of samples in $\mathbb{R}^{50}$ drawn from $\mathcal{N}(0, \mathbf{I})$. We plot the sample complexity for different Gaussian smoothed divergence at different level of noises.

when the smoothing is larger ($i.e$ for getting a given divergence value at $\sigma = 5$, one needs almost 10-fold more samples for $\sigma = 15$). This overhead of samples needed when smoothing increases is properly described, for the Gaussian smoothed SWD in our Proposition 1, as the sample complexity depends on the moments of the Gaussian.

As for conclusion from these analyses, we highlight that the Gaussian smoothed Sliced MMD seems to present several strong benefits : its sample complexity does not depend on the dimension and seems to be the best one among the divergence we considered. More interestingly, it is not impacted by the amount of Gaussian smoothing and thus not impacted by a desired privacy level.

4.2 Domain adaptation with Gaussian Smoothed Sliced Divergence

As an application, we have considered the problem of unsupervised domain adaptation for a classification task. In this context, given source examples $\mathbf{X}_s$ and their label $\mathbf{y}_s$ and unlabeled target examples $\mathbf{X}_t$, our goal is to design a classifier $h(\cdot)$ learned from the source examples that generalizes well on the target ones. A classical approach consists in learning a representation mapping $g(\cdot)$ that leads to invariant latent representations, invariance being measured as a distance between empirical distributions of mapped source and target samples. Formally, this leads to the following problem

$$\min_{g,h} L_c(h(g(\mathbf{X}_s)), \mathbf{y}_s) + \mathcal{D}(g(\mathbf{X}_s), g(\mathbf{X}_t))$$

where L_c can be the cross-entropy loss or a quadratic loss and $\mathcal{D}$ a divergence between empirical distributions, in our case, $\mathcal{D}$ will be any Gaussian smoothed sliced divergence. We solve this problem through

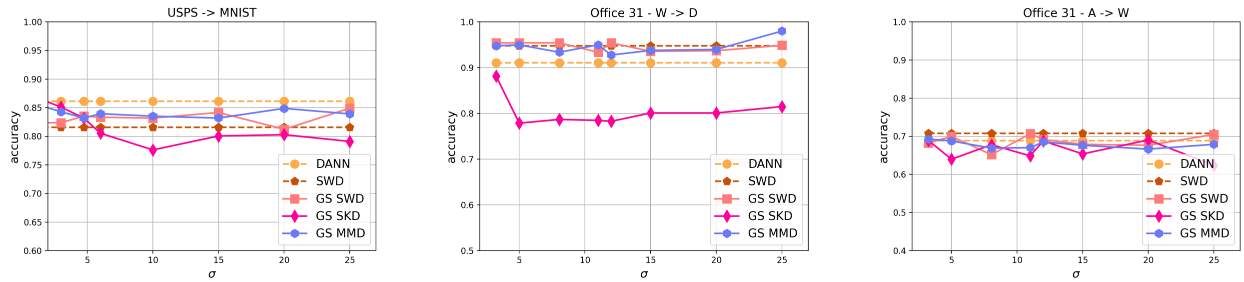

Figure 6: Domain adaptation performances using different divergences on distributions with respect to the Gaussian smoothing. (left) USPS to MNIST. (middle) Office-31 Webcam to DSLR. (right) Office-31 Amazon to Webcam.

stochastic gradient descent, similarly to many approaches that use Sliced Wasserstein Distance as a distribution distance [17]. Note that, in practice, using a smoothed divergence preserves the privacy of the target samples as shown by Rakotomamonjy & Ralaivola [26].

Our experiments evaluate the studied Gaussian-smoothed sliced divergences in classical unsupervised domain adaptation. We have considered twi datasets: a handwritten digit recognition (USPS/MNIST) and Office 31 datasets. Our goal is to analyze how our divergences perform compared with non-smoothed divergences. The first one is the Sliced Wasserstein Distance (SWD) [17] and the second one is the Jenssen-Shannon approximation based on adversarial approach, known as DANN [9]. For all methods and for each dataset, we used the same neural network architecture for representation mapping and for classification. Approaches differ only on how distance between distributions have been computed.

Results are depicted in Figure 6. For the two problems, we can see that performances obtained with the Gaussian smoothed sliced Wasserstein or MMD divergences are similar to those obtained with DANN or SWD across all ranges of noise. The smoothed version of Sinkhorn is less stable and induces a slight loss of performance. Owing to the metric property and the induced weak topology, the privacy preservation comes almost without loss of performance in this domain adaptation context.

5 Conclusion

In this study, we have analyzed the properties of Gaussian smoothed sliced divergence for comparing distributions as they play a crucial role in a privacy preserving context. We have derived several theoretical results related to their topological and statistical properties. More precisely, we have shown that under mild condition on their base divergence, the smoothing and slicing operations preserves metric property. From a statistical point of view, we have shown that sample complexity does not depend on the dimension of the problem and follows a similar complexity than their sliced version, although some overhead may have to be paid due to the smoothing. We have illustrated those theoretical findings through some experimental analyses on toy problem. We have also analyzed the behavior of our divergence on domain adaptation problems and confirm the fact that using those divergences yields only to slight loss of performances while preserving privacy. One lesson we have also learnt is that Gaussian smoothed sliced MMD seems to present several strong benefits in term of sample complexity.

References

[1] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223, International Convention Centre, Sydney, Australia, 2017. PMLR.

[2] Bonneel, N. and Coeurjolly, D. Spot: Sliced partial optimal transport. 38(4), 2019.

[3] Bonneel, N., van de Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using lagrangian mass transport. *ACM Trans. Graph.*, 30(6):158:1–158:12, 2011.

[4] Bonnotte, N. *Unidimensional and Evolution Methods for Optimal Transportation*. Theses, Université Paris Sud - Paris XI ; Scuola normale superiore (Pise, Italie), December 2013.

[5] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adapta-

tion. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[6] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300. Curran Associates, Inc., 2013.

[7] Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced Wasserstein distance and its use for gans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10640–10648, 2019.

[8] Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015.

[9] Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

[10] Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

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

[12] Kantorovich, L. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 2:227–229, 1942.

[13] Kolouri, S., Zou, Y., and Rohde, G. K. Sliced Wasserstein kernels for probability distributions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5258–5267, 2016.

[14] Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, July 2017.

[15] Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced Wasserstein distances. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 261–272. Curran Associates, Inc., 2019.

[16] Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019.

[17] Lee, C.-Y., Batra, T., Baig, M. H., and Ulbricht, D. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10285–10295, 2019.

[18] Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

[19] Monge, G. Mémoire sur la théotie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pp. 666–704, 1781.

[20] Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[21] Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-Wasserstein and applications to generative modeling. *arxiv preprint 2002.07367*, 2020.

[22] Nietert, S., Goldfeld, Z., and Kato, K. From smooth wasserstein distance to dual sobolev norm: Empirical approximation and statistical applications. *arXiv preprint arXiv:2101.04039*, 2021.

[23] Nietert, S., Goldfeld, Z., and Kato, K. Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8172–8183. PMLR, 18–24 Jul 2021.

[24] Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[25] Rachev, S. and Rüschendorf, L. *Mass Transportation Problems: Volume I: Theory*. Mass Transportation Problems. Springer, 1998.

[26] Rakotomamonjy, A. and Ralaivola, L. Differentially private sliced wasserstein distance. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8810–8820. PMLR, 18–24 Jul 2021.

[27] Salimans, T., Zhang, H., A.Radford, and Metaxas, D. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

[28] Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, 2015.

[29] Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A. J., and Gretton, A. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR (Poster)*, 2017.

[30] Villani, C. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009.

[31] Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P., and Gool, L. V. Sliced wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3713–3722, 2019.

A Additional definitions

A.1 Maximum Mean Discrepancy

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the reproducing kernel of a reproducing kernel Hilbert space $\mathcal{H}$. The metric on distance denoted as maximum mean discrepancy between μ and ν belonging to $\mathcal{P}(\mathcal{X})$ is defined as:

$$MMD(\mu, \nu) = \left\| \int k(\cdot, x) d\mu(x) - \int k(\cdot, x) d\nu(x) \right\|_{\mathcal{H}}.$$

For empirical distributions, one can estimate the MMD using biased or unbiased formulations as given by Gretton et al. [11]: For empirical distributions, one can estimate the MMD using biased or unbiased formulations as given by Gretton et al. [11]:

$$MMD(\hat{\mu}, \hat{\nu}) = \Bigg[\frac{1}{n^2} \sum_{i,j} k(x_i, x_j) + \frac{1}{m^2} \sum_{i,j} k(y_i, y_j) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j) \Bigg]^{\frac{1}{2}}$$

A.2 Sinkhorn Divergence and Gaussian Smoothed Sliced Sinkhorn Divergence

Let define the entropic regularized Wasserstein distance [6] between distributions μ and ν as

$$W^p_{p,\lambda}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \gamma(x, y) dx dy + \lambda H(\gamma | \mu \otimes \nu).$$

where the set $\Pi(\mu, \nu)$ is defined as in Section 2. The term $H(\cdot|\cdot)$ is the relative entropy regularization of the transport plan with respect to the product measure $\mu \otimes \nu$, and is given by

$$H(\gamma | \mu \otimes \nu) = \iint \log\Big(\frac{\mathrm{d}\gamma(x, y)}{\mathrm{d}\mu \otimes \mathrm{d}\nu(x, y)} \Big) \mathrm{d}\gamma(x, y).$$

The related regularization parameter is $\lambda \geq 0$. Then, the Sinkhorn divergence is defined as

$$SKD_\lambda(\mu, \nu) = W^p_{p,\lambda}(\mu, \nu) - \frac{1}{2} W^p_{p,\lambda}(\mu, \mu) - \frac{1}{2} W^p_{p,\lambda}(\nu, \nu).$$

Accordingly the Gaussian Smoothed Sliced Sinkhorn Divergence is expressed as

$$G_\sigma SKD^p_{p,\lambda}(\mu, \nu) = \int_{\mathbb{S}^{d-1}} SKD^p_\lambda(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma) u_d(\mathbf{u}) d\mathbf{u}.$$

B Proofs

B.1 Proof of Theorem 1

• *Non-negativity (or symmetry).* The non-negativity (or symmetry) follows directly from the non-negativity (or symmetry) of D^p, see Definition 3.
• *Identity property.* For the identity property, if the base divergence D^p satisfies the identity property in one dimensional measures, then for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\mathbf{u} \in \mathbb{S}^{d-1}$, one has that $D^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma) = 0$, hence, by Definition 3, $G_\sigma SD^p(\mu, \mu) = 0$. Let us now prove the fact that for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d), G_\sigma SD^p(\mu, \mu) = 0$ entails $\mu = \nu$ a.s. On one hand, $G_\sigma SD^p(\mu, \mu) = 0$ gives the fact that $D^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma) = 0$ for u_d-almost every $\mathbf{u} \in \mathbb{S}^{d-1}$, hence $\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma = \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma$ for u_d-almost every $\mathbf{u} \in \mathbb{S}^{d-1}$. Following the techniques in proof of Proposition 5.1.2 in [4], for any measure $\eta \in \mathcal{P}(\mathbb{R}^m)$ (with $m \geq 1$), $\mathcal{F}[\eta](\cdot)$ stands for the Fourier transform of $\mathbf{s}$ and is given as $\mathcal{F}[\eta](\mathbf{v}) = \int_{\mathbb{R}^m} e^{-i\mathbf{s}^\top \mathbf{v}} \mathrm{d}\eta(\mathbf{v})$ for any $\mathbf{v} \in \mathbb{R}^m$. Then

$$\begin{aligned}
\mathcal{F}[\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma](v) &= \int_{\mathbb{R}} e^{-ivt} \mathrm{d}(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma)(t) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-i(r+t)v} \mathrm{d}\mathcal{R}_{\mathbf{u}}\mu(r) \mathrm{d}\mathcal{N}_\sigma(t) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}} e^{-i(\langle \mathbf{u}, s \rangle + t)v} \mathrm{d}\mu(\mathbf{s}) \mathrm{d}\mathcal{N}_\sigma(t) \\
&= \int_{\mathbb{R}} e^{-itv} \mathrm{d}\mathcal{N}_\sigma(t) \int_{\mathbb{R}^d} e^{-i(\langle \mathbf{u}, s \rangle)v} \mathrm{d}\mu(\mathbf{s}) \\
&= \mathcal{F}[\mathcal{N}_\sigma](v) \mathcal{F}[\mu](v\mathbf{u}).
\end{aligned}$$

Since for u_d-almost every $\mathbf{u} \in \mathbb{S}^{d-1}, \mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma = \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma$, and hence $\mathcal{F}[\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma] = \mathcal{F}[\mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma] \Leftrightarrow \mathcal{F}[\mathcal{N}_\sigma]\mathcal{F}[\mu] = \mathcal{F}[\mathcal{N}_\sigma]\mathcal{F}[\nu] \Leftrightarrow \mathcal{F}[\mu] = \mathcal{F}[\nu]$. Since the Fourier transform is injective, we conclude that $\mu = \nu$.
• *Triangle inequality.* Assume that D^p is a metric and let $\mu, \nu, \eta \in \mathcal{P}(\mathbb{R}^d)$. We then have

$$\begin{aligned}
&G_\sigma SD(\mu, \nu) \\
&= \Big\{ \int_{\mathbb{S}^{d-1}} D^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \Big\}^{1/p} \\
&\leq \Big\{ \int_{\mathbb{S}^{d-1}} \Big(D(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\eta * \mathcal{N}_\sigma) + D(\mathcal{R}_{\mathbf{u}}\eta * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma) \Big)^p u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \Big\}^{1/p} \\
&\underbrace{\leq}_{(\star)} \Big\{ \int_{\mathbb{S}^{d-1}} \Big(D^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\eta * \mathcal{N}_\sigma) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \Big\}^{1/p} + \Big\{ \int_{\mathbb{S}^{d-1}} D^p(\mathcal{R}_{\mathbf{u}}\eta * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma) \Big)^p u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \Big\}^{1/p} \\
&= G_\sigma SD(\mu, \eta) + G_\sigma SD(\eta, \nu),
\end{aligned}$$

where inequality in $(\star)$ follows from the application of Minkowski inequality.

B.2 Proof of Proposition 1

Let us first upper bound the k-th moment of $M_k(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma)$, for all $k \geq 1$. For all $\mathbf{u} \in \mathbb{S}^{d-1}$, one has

$$\begin{aligned} M_k(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma) &= \int_{\mathbb{R}} |t|^k \mathrm{d}(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma)(t) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} |r+t|^k \mathrm{d}\mathcal{R}_{\mathbf{u}}\mu(r) \mathrm{d}\mathcal{N}_\sigma(t) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} |\langle \mathbf{u}, s\rangle + t|^k \mathrm{d}\mu(\mathbf{s}) \mathrm{d}\mathcal{N}_\sigma(t). \end{aligned}$$

Using the elementary inequality $(a+b)^k \leq 2^{k-1}(a^k + b^k)$ for $k \geq 1, a \geq 0$, and $b \geq 0$, we obtain

$$\begin{aligned} &M_k(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma) \\ &\leq 2^{k-1} \int_{\mathbb{R}^d} \int_{\mathbb{R}} \big(|\langle \mathbf{u}, s\rangle|^k + |t|^k\big) \mathrm{d}\mu(s) \mathrm{d}\mathcal{N}_\sigma(t) \\ &\leq 2^{k-1} \Big(\|\mathbf{u}\| \int_{\mathbb{R}^d} \|s\|^k \mathrm{d}\mu(s) + \int_{\mathbb{R}} |t|^k \mathrm{d}\mathcal{N}_\sigma(t) \Big) \\ &\leq 2^{k-1} \Big(\int_{\mathbb{R}^d} \|s\|^k \mathrm{d}\mu(s) + \int_{\mathbb{R}} |t|^k \mathrm{d}\mathcal{N}_\sigma(t) \Big) \\ &= 2^{k-1}(M_k(\mu)) + M_k(\mathcal{N}_\sigma)). \end{aligned}$$

We then use the following result:

Lemma 1 (see proof of Theorem 1 in [8]). *Let $\eta \in \mathcal{P}(\mathbb{R})$ and let $p \geq 1$. Assume that $M_q(\eta) < \infty$ for some $q > p$. There exists a constant $C_{p,q}$ depending only on p, q such that, for all $n \geq 1$,*

$$\mathbf{E}[W_p^p(\hat{\eta}_n, \eta)] \leq C_{p,q} M_q(\eta)^{p/q} \begin{cases} n^{-1/2}\mathbf{1}_{q>2p}, \\ n^{-1/2}\log(n)\mathbf{1}_{q=2p} \\ n^{-(q-p)/q}\mathbf{1}_{q\in(p,2p)}. \end{cases}$$

Let us fix $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ with $q > p \geq 1$ an empirical measure $\hat{\mu}_n$. Then, one has

$$M_q(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma) \leq 2^{q-1}(M_q(\mu)) + M_q(\mathcal{N}_\sigma)) < \infty.$$

By Lemma 1, we obtain

$$\begin{aligned} &\mathbf{E}[G_\sigma SWD_p^p(\hat{\mu}_n, \mu)] \\ &= \mathbf{E}\Big[\int_{\mathbb{S}^{d-1}} W_p^p(\mathcal{R}_{\mathbf{u}}\hat{\mu}_n * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \Big] \\ &\leq C_{p,q} \begin{cases} n^{-1/2}\mathbf{1}_{q>2p}, \\ n^{-1/2}\log(n)\mathbf{1}_{q=2p} \\ n^{-(q-p)/q}\mathbf{1}_{q\in(p,2p)}. \end{cases} \\ &\qquad \times \int_{\mathbb{S}^{d-1}} M_q(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma)^{p/q} u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \\ &\leq C_{p,q} \begin{cases} n^{-1/2}\mathbf{1}_{q>2p}, \\ n^{-1/2}\log(n)\mathbf{1}_{q=2p} \\ n^{-(q-p)/q}\mathbf{1}_{q\in(p,2p)}. \end{cases} \\ &\times \begin{cases} \big(2^{q-1}(M_q(\mu)) + M_q(\mathcal{N}_\sigma))\big)^{p/q}\mathbf{1}_{q\in 2\mathbb{N}^*}, \\ \big(2^{q-1}(M_q(\mu))\big)^{p/q}\mathbf{1}_{q\in 2\mathbb{N}+1}. \end{cases} \end{aligned}$$

On the other hand, since $W_p(\cdot,\cdot)$ is a metric, by applying Theorem 3, we obtain the following:

$$\begin{aligned} &\mathbf{E}[|G_\sigma SWD^p(\hat{\mu}_n, \hat{\nu}_n) - G_\sigma SWD^p(\mu,\nu)|] \\ &\leq 2C_{p,q} \begin{cases} n^{-1/2}\mathbf{1}_{q>2p}, \\ n^{-1/2}\log(n)\mathbf{1}_{q=2p} \\ n^{-(q-p)/q}\mathbf{1}_{q\in(p,2p)}. \end{cases} \\ &\times \begin{cases} \big(2^{q-1}(M_q(\mu,\nu)) + M_q(\mathcal{N}_\sigma))\big)^{p/q}\mathbf{1}_{q\in 2\mathbb{N}^*} \\ \big(2^{q-1}(M_q(\mu,\nu))\big)^{p/q}\mathbf{1}_{q\in 2\mathbb{N}+1}. \end{cases} \end{aligned}$$

B.3 Proof of Theorem 4

Using Holder's inequality, we have

$$\begin{aligned} &\mathbf{E}_{\mathbf{u}\sim u_d}\big[\big|\widehat{G_\sigma SD}^p(\mu,\nu) - G_\sigma SD^p(\mu,\nu)\big|\big] \\ &\leq \Big(\mathbf{E}_{\mathbf{u}\sim u_d}\big[\big|\widehat{G_\sigma SD}^p(\mu,\nu) - G_\sigma SD^p(\mu,\nu)\big|^2\big]\Big)^{1/2} \\ &= \Big(\mathbf{V}_{\mathbf{u}\sim u_d}\big[\big|\widehat{G_\sigma SD}^p(\mu,\nu)\big|\big]\Big)^{1/2} \\ &= \Big(\mathbf{V}_{\mathbf{u}\sim u_d}\big[\big|G_\sigma SD^p(\mu,\nu)\big|\big]\Big)^{1/2} \\ &= \frac{A(p,\sigma)}{\sqrt{L}}. \end{aligned}$$

B.4 Proof of Proposition 4

The proof follows the same lines in proof of Lemma 1 in [23]. First, we have that $\mathcal{N}_{\sigma_2} = \mathcal{N}_{\sigma_1} * \mathcal{N}_{\sqrt{\sigma_2^2-\sigma_1^2}}$. Setting the following random variables: $X_{\mathbf{u}} \sim \mathcal{R}_{\mathbf{u}}\mu, Y_{\mathbf{u}} \sim \mathcal{R}_{\mathbf{u}}\nu, Z_X \sim \mathcal{N}_{\sigma_1}, Z_Y \sim \mathcal{N}_{\sigma_1}, Z'_X \sim \mathcal{N}_{\sqrt{\sigma_2^2-\sigma_1^2}}, Z'_Y \sim \mathcal{N}_{\sqrt{\sigma_2^2-\sigma_1^2}}$. The sliced Wasserstein distance $\mathrm{W}_p^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_{\sigma_2}, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_{\sigma_2})$ is given as a minimization over couplings $(X_{\mathbf{u}}, Z_X, Z'_X)$ and $(Y_{\mathbf{u}}, Z_Y, Z'_Y)$. Using the inequality $\mathbf{E}[|X|^p] - 2^{p-1}\mathbf{E}[|Y|^p] \leq 2^{p-1}\mathbf{E}[|X+Y|^p]$ for any random variables $X, Y \in \mathbb{L}_p$ integrable, we obtain,

$$\begin{aligned} &2^{p-1}\mathbf{E}\big[|(X_{\mathbf{u}} + Z_X) - (Y_{\mathbf{u}} + Z_Y) + (Z'_X + Z'_Y)|^p\big] \\ &\geq \mathbf{E}\big[|(X_{\mathbf{u}} + Z_X) - (Y_{\mathbf{u}} + Z_Y)|^p\big] \\ &\qquad - 2^{p-1}\mathbf{E}\big[|(Z'_X + Z'_Y)|^p\big]\big). \end{aligned}$$

Hence,

$$\begin{aligned} &2^{p-1}\mathrm{W}_p^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_{\sigma_2}, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_{\sigma_2}) \\ &\geq \inf \Big(\mathbf{E}\big[|(X_{\mathbf{u}} + Z_X) - (Y_{\mathbf{u}} + Z_Y)|^p\big] \\ &\qquad - 2^{p-1}\mathbf{E}\big[|(Z'_X + Z'_Y)|^p\big]\big)\Big) \\ &\geq \mathrm{W}_p^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_{\sigma_1}, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_{\sigma_1}) \\ &\qquad - 2^{p-1}\sup \mathbf{E}\big[|(Z'_X + Z'_Y)|^p\big] \\ &\geq \mathrm{W}_p^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_{\sigma_1}, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_{\sigma_1}) - 2^p \sup \mathbf{E}\big[|(Z'_X)|^p\big]. \end{aligned}$$

Therefore,

$$\begin{aligned} 2^{p-1}G_{\sigma_2}SWD_p^p(\mu,\nu) \geq G_{\sigma_1}SWD_p^p(\mu,\nu)| \\ - 2^p u_d(\mathbb{S}^{d-1}) \sup \mathbf{E}\big[|(Z'_X)|^p\big], \end{aligned}$$

hence,

$$
\begin{aligned}
G_{\sigma_1} SWD_p^p(\mu,\nu)| \leq 2^{p-1} & G_{\sigma_2} SWD_p^p(\mu,\nu) \\
& + 2^p u_d(\mathbb{S}^{d-1}) \sup \mathbf{E}\big[|(Z'_X)|^p\big].
\end{aligned}
$$

Recall that if $Z \sim \mathcal{N}_\sigma$

$$
\mathbf{E}[|Z|^p] = \frac{2^p \Gamma((p+1)/2)}{\Gamma(1/2)} \sigma^{2p} \leq 2^{p/2} \sigma^{2p}.
$$

and $u_d(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ then

$$
G_{\sigma_1} SWD_p^p(\mu,\nu) \leq 2^{p-1} G_{\sigma_2} SWD_p^p(\mu,\nu) + \frac{2\pi^{d/2}}{\Gamma(d/2)} 2^{3p/2} (\sigma_2^2 - \sigma_1^2)^{2p}.
$$