Gaussian-Smoothed Sliced Probability Divergences

Anonymous authors
Paper under double-blind review

Abstract

Gaussian smoothed sliced Wasserstein distance has been recently introduced for comparing probability distributions, while preserving privacy on the data. It has been shown that it provides performances similar to its non-smoothed (non-private) counterpart. However, the computational and statistical properties of such a metric have not yet been well-established. This work investigates the theoretical properties of this distance as well as those of generalized versions denoted as Gaussian-smoothed sliced divergences $\mathrm{G}_\sigma\mathrm{SD}_p$. We first show that smoothing and slicing preserve the metric property and the weak topology. To study the sample complexity of such divergences, we then introduce $\hat{\hat{\mu}}_n$ the double empirical distribution for the smoothed-projected μ. The distribution $\hat{\hat{\mu}}_n$ is a result of a double sampling process: one from sampling according to the origin distribution μ and the second according to the convolution of the projection of μ on the unit sphere and the Gaussian smoothing. We particularly focus on the Gaussian smoothed sliced Wasserstein distance $\mathrm{G}_\sigma\mathrm{SW}_p$ and prove that it converges with a rate $O(n^{-1/2p})$. We also derive other properties, including continuity, of different divergences with respect to the smoothing parameter. We support our theoretical findings with empirical studies in the context of privacy-preserving domain adaptation.

1 Introduction

Divergences for comparing two distributions have been shown to be important for achieving good performance in the contexts of generative modeling (Arjovsky et al., 2017; Salimans et al., 2018), domain adaptation (Long et al., 2015; Courty et al., 2016; Lee et al., 2019), and in computer vision (Bonneel et al., 2011; Solomon et al., 2015) among many more applications (Kolouri et al., 2017; Peyré & Cuturi, 2019; Nguyen et al., 2023). Examples of divergences that have proved useful for these tasks are the Maximum Mean Discrepancy (Gretton et al., 2012; Long et al., 2015; Sutherland et al., 2017), the Wasserstein distance (Monge, 1781; Kantorovich, 1942; Villani, 2009) or its variant the sliced Wasserstein distance (SW) (Kolouri et al., 2016; Bonneel & Coeurjolly, 2019; Kolouri et al., 2019b; Nguyen et al., 2021; 2022; 2024).

The SW distance has the advantage of being computationally efficient, since it uses a closed-form solution for distributions with support on $\mathbb{R}$, by computing the expectation of one-dimensional (1D) 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 (Kolouri et al., 2019a; Deshpande et al., 2019; Wu et al., 2019; Lee et al., 2019) and its statistical properties have been well-studied in Nadjahi et al. (2020).

Recently, Gaussian smoothed variants of the Wasserstein distance and the sliced Wasserstein distance have been introduced respectively in (Nietert et al., 2021) and in Rakotomamonjy & Ralaivola (2021). One main motivation behind these variants is to provide a privacy guarantee for the distribution comparison task as Gaussian smoothing is known to be a mechanism for achieving differential privacy (Dwork et al., 2014). While the properties of the Gaussian smoothed Wasserstein distance have been extensively studied by Nietert et al. (2021), the properties of the Gaussian smoothed sliced Wasserstein distance have not been fully investigated yet although they are known to be more computationally efficient.

In this work, we focus on the slicing of Gaussian-smoothed measure discrepancies by providing theoretical properties of more general divergences induced by some base distances or divergences for distributions defined

in $\mathbb{R}^d$. These base distances/divergences encompass Wasserstein, maximum mean discrepancy, Sinkhorn divergence. As for a main contribution, we first establish the topological properties of these divergences in term of a metrization of the weak topology and a semi-lower continuous property. Then we focus on the sample complexity of such divergences by introducing the *double empirical distribution* $\hat{\hat{\mu}}_n$ for the smoothed-projected origin distribution μ. The new empirical distribution is a result of double sampling process: one from sampling according to the origin distribution and the second according to the convolution of the projection of μ on the unit sphere and the Gaussian smoothing. The introducing of $\hat{\hat{\mu}}_n$ is inspired from the implementation part: we sample $X_1, \ldots, X_n$ from the raw distribution μ to define $\hat{\mu}_n$ then project it on the unit sphere and smooth this projection with a Gaussian distribution. This smoothing is a continuous measure that needs to be sampled. For that reason, we add a double sampling and then provide $\hat{\hat{\mu}}_n$. We particularly focus on the Gaussian smoothed sliced Wasserstein distance.

Given the importance of the noise level in the privacy/utility trade-off achieved by the divergence, we investigate an order relation and a continuity result with respect to the noise level. These properties are of high impact as it supports a computationally cheap warm-start/fine-tuning procedure when looking for a privacy/utility compromise of the divergence. Our theoretical study is backed by some numerical experiments on toy problems and on domain adaptation illustrating how owing to the topology induced by our metric and its continuity, differential privacy comes almost for free (without loss of performance) and multiple models with different level of privacy can be cheaply computed.

Comparison with previous works. Here we highlight the position of this work compared to the most linked previous ones, in particular Nadjahi et al. (2020) and Rakotomamonjy & Ralaivola (2021). The work of Nadjahi et al. (2020) is focused on sliced Wasserstein distance and its statistical properties, however our work is based on the properties of the Gaussian smoothed with general divergences (e.g. Wasserstein, MMD, Sinkhorn divergence). We argue that the properties cannot be directly derived from (Nadjahi et al., 2020), especially the sample complexity result. In Rakotomamonjy & Ralaivola (2021), the authors investigated the smoothed Wasserstein distance and their theoretical finding was principally on proving the metric property, whereas we further investigate sample and projection complexities and the continuity properties w.r.t. the smoothing noise level. We emphasize that the novelty of the present paper consists in the theoretical properties derived from the definition of the empirical measure $\hat{\hat{\mu}}_n$. The smoothing of the raw measures, from a theoretical point view, is a continuous measure (see Lemma 3.5) that needs to be sampled. This entails to define the second sampling step and construct $\hat{\hat{\mu}}_n$, an empirical version for the smoothing projection of μ. To the best of our knowledge, this work is the first introducing the double randomness in the case of smoothing optimal transport discrepancies. Recent works (Goldfeld et al., 2020; Nietert et al., 2021) addressed the smoothing Wasserstein an their theoretical results relied only on $\hat{\mu}_n$.

Layout of the paper. The paper is organized as follows: after introducing the notation and some background in Section 2, we detail the topological properties of Gaussian-smoothed sliced divergence in Section 3.1 while the double sampling process and its statistical properties are established in Section 3.2. The noise analyses are provided in Section 3.3. Experimental analyses for supporting the theory and showcasing the relevance of our divergences in domain adaptation are depicted in Section 4. Discussions on the perspectives and limitations are in Section 5. All the proofs of the theoretical results and some additional experiments are postponed to the appendices in the supplementary.

2 Preliminaries

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

Notation. 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. We denote $M_p(\mu) = \int_x \|x\|^p \mathrm{d}\mu(x)$. For two probability distributions μ and ν, we denote their convolution as $\mu * \nu \in \mathcal{P}(\mathbb{R}^d)$, namely $(\mu * \nu)(A) = \int_x \int_y \mathbf{1}_A(x+y)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$, 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.

Let $\Gamma : \mathbb{R} \to \mathbb{R}$ be the Gamma function expressed as $\Gamma(v) = \int_0^\infty t^{v-1}e^{-t}dt$ for $v > 0$. For $k \in \mathbb{N}$, $(\cdot)_k$ denoted the Pochhammer symbol, also known in the literature as a rising factorial, namely $(\alpha)_0 = 1, (\alpha)_1 = \alpha$, and $(\alpha)_k = \frac{\Gamma(\alpha+k)}{\Gamma(k)} = \alpha(\alpha+1)\cdots(\alpha+k-1)$, for $k \geq 1$. We denote by ${}_1F_1(\alpha, \gamma; z)$ the Kummers confluent hypergeometric function (Olver, 2010) and defined by ${}_1F_1(\alpha, \gamma; z) = \sum_{k=0}^\infty \frac{(\alpha)_k}{(\gamma)_k}\frac{z^k}{k!}$.

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

$$\mathrm{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^p \gamma(x, x') \mathrm{d}x\mathrm{d}x' \right)^{1/p}$$

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 calculated in closed-form owing to the cumulative distributions of μ and ν (Rachev & Rüschendorf, 1998). In practice for empirical distributions, the closed-form solution requires only the sorting of the samples, which makes it very efficient. Because of 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 one-dimensional space and then to compute the Wasserstein distance. This operation can be theoretically formalized through the use of the Radon transform, leading to the so-called sliced Wasserstein distance (Kolouri et al., 2016; Bonneel & Coeurjolly, 2019; Kolouri et al., 2019b; Nguyen et al., 2021).

Definition 2.1. For any $p \in [1, \infty)$ and two measures μ, $\nu \in \mathcal{P}_p(\mathbb{R}^d)$, the sliced Wasserstein distance (SW) reads as

$$\mathrm{SW}_p(\mu, \nu) \triangleq \left(\int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p(\mathcal{R}_\mathbf{u}\mu, \mathcal{R}_\mathbf{u}\nu) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \right)^{1/p}.$$

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 SW, replacing the Radon projected measures with their Gaussian-smoothed counterpart leads to the following definition:

Definition 2.2. The σ-Gaussian-smoothed p-Sliced Wasserstein distance between probability distributions μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$ writes as

$$\mathrm{G}_\sigma\mathrm{SW}_p(\mu, \nu) \triangleq \left(\int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \right)^{1/p},$$

where $\mathcal{N}_\sigma = \mathcal{N}(0, \sigma^2)$ is the zero-mean σ^2-variance Gaussian measure. It is important to note here that the smoothing (convolution) operation occurs after projection onto the one-dimensional space. Hence, assuming $X \sim \mu$, $Y \sim \nu$, for a given direction $\mathbf{u}$, we compute in the integral the one-dimensional Wasserstein distance between the probability laws of $\mathbf{u}^\top X + Z$ and $\mathbf{u}^\top Y + Z'$ where $Z, Z' \sim \mathcal{N}_\sigma$ are independent random variables. The metric properties of $\mathrm{G}_\sigma\mathrm{SW}_p$ for $p \geq 1$ have been discussed in a recent work (Rakotomamonjy & Ralaivola, 2021). 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 SW distance of the original distribution smoothed with a d-dimensional Gaussian $\mu * \mathcal{N}_{\sigma\mathbf{I}_d}$, where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix.

Gaussian-smoothed sliced divergence. The idea of slicing high-dimensional distributions before feeding them to a divergence between probability distributions can be extended to distances other than the Wasserstein distance. These sliced divergences have been studied by Nadjahi et al. (2020). Similarly, we can define a Gaussian-smoothed sliced divergence, given a divergence $\mathrm{D}_{\mathbb{R}^d} : \mathcal{P}_p(\mathbb{R}^d) \times \mathcal{P}_p(\mathbb{R}^d) \to \mathbb{R}^+$ for $d \geq 1$ as:

Definition 2.3. The σ-Gaussian-smoothed p-Sliced Divergence between probability distributions μ and ν in $\mathcal{P}_p(\mathbb{R}^d)$ associated to the *base divergence* $\mathrm{D} \triangleq \mathrm{D}_{\mathbb{R}}$, $p \geq 1$ is

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

Typical relevant divergences are the maximum mean discrepancy (MMD) (Gretton et al., 2012) or the Sinkhorn divergence (Genevay et al., 2018; Peyré & Cuturi, 2019). 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 analyze the properties of the Gaussian-smoothed sliced divergence, in terms 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 Rakotomamonjy & Ralaivola (2021) 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 $\mathrm{D}(\cdot,\cdot)$ under certain assumptions.

Theorem 3.1. *For any $\sigma > 0, p \geq 1$, the following properties hold:*

1. *if $\mathrm{D}(\cdot,\cdot)$ is non-negative (or symmetric), then $\mathrm{G}_\sigma\mathrm{SD}_p(\cdot,\cdot)$ is non-negative (or symmetric);*
2. *if $\mathrm{D}(\cdot,\cdot)$ satisfies the identity of indiscernibles, i.e. for $\mu', \nu' \in \mathcal{P}(\mathbb{R})$, $\mathrm{D}(\mu',\nu') = 0$ if and only if $\mu' = \nu'$, then this identity also holds for $\mathrm{G}_\sigma\mathrm{SD}_p(\cdot,\cdot)$ for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$;*
3. *if $\mathrm{D}(\cdot,\cdot)$ satisfies the triangle inequality then $\mathrm{G}_\sigma\mathrm{SD}_p(\cdot,\cdot)$ 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.

We further postponed to the appendix the proofs of the two other topological properties: (i) $\mathrm{G}_\sigma\mathrm{SD}$ metrizes the weak topology on $\mathcal{P}_p(\mathbb{R}^d)$ and (ii) $\mathrm{G}_\sigma\mathrm{SD}$ is lower semi-continuous with respect to the weak topology in $\mathcal{P}_p(\mathbb{R}^d)$.

Now, we establish under which conditions on the divergence D, the convergence of a sequence in $\mathrm{G}_\sigma\mathrm{SD}$ implies weak convergence in $\mathcal{P}_p(\mathbb{R}^d)$. We say that $\{\mu_k\}_{k\in\mathbb{N}}$ *converges weakly* to μ and write, $\mu_k \Rightarrow \mu$, if $\int f(x)d\mu_k(x) \to \int f(x)d\mu(x)$, as $k \to \infty$, for every f in the space of all bounded continuous real functions.

Theorem 3.2. *Let $\sigma > 0, p \geq 1$, $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, and $\{\mu_k \in \mathcal{P}_p(\mathbb{R}^d)\}_{k\in\mathbb{N}}$ a sequence of distributions. Assume that the divergence D is bounded and metrizes the weak topology on $\mathcal{P}(\mathbb{R})$. Then, $\lim_{k\to\infty} \mathrm{G}_\sigma\mathrm{SD}_p(\mu_k,\mu) = 0$ if and only if $\mu_k \Rightarrow \mu$.*

Note that Theorem 3.2 extends the results of Nadjahi et al. (2020) to Gaussian-smoothed distributions, as we retrieve them as a special case for $\sigma = 0$. In addition, based on Theorem 3.2 by Lin et al. (2021) and the above, we can also claim that the Gaussian-smoothed SWD metrizes the weak convergence.

Proposition 3.3. *Let $\sigma > 0, p \geq 1$ and assume that the base divergence D is lower semi-continuous w.r.t. the weak topology in $\mathcal{P}(\mathbb{R})$. Then, $\mathrm{G}_\sigma\mathrm{SD}_p$ is lower semi-continuous with respect to the weak topology in $\mathcal{P}_p(\mathbb{R}^d)$.*

When the base divergence D is equal to the Wasserstein distance W_p, that is lower semi-continuous (Villani, 2009), then Proposition 3.3 shows that the smoothed sliced Wasserstein distance is semi-lower continuous too.

3.2 Statistical properties

The next theoretical question we are interested in is about the incurred error when the true distribution μ is approximated by its empirical distribution $\hat{\mu}_n$. 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 the double empirical $\hat{\mathrm{G}}_\sigma\mathrm{SD}_p(\hat{\mu}_n, \hat{\nu}_n)$ (see Definition 3.6) to $\mathrm{G}_\sigma\mathrm{SD}_p(\mu, \nu)$ *(ii)* the convergence of $\widehat{\mathrm{G}_\sigma\mathrm{SD}}_p(\mu, \nu)$ (see (1)) to $\mathrm{G}_\sigma\mathrm{SD}_p(\mu, \nu)$, when approximating the expectation over the random projection with sample mean.

Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ be the empirical probability measures of independent observations. The smoothed Gaussian sliced divergence between $\hat{\mu}_n$ and $\hat{\nu}_n$ is given by

$$\mathrm{G}_\sigma\mathrm{SD}_p(\hat{\mu}_n, \hat{\nu}_n) = \left(\int_{\mathbb{S}^{d-1}} \mathrm{D}^p \left(\mathcal{R}_{\mathbf{u}}\hat{\mu}_n * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\hat{\nu}_n * \mathcal{N}_\sigma \right) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \right)^{1/p}.$$

Remark 3.4. Remark that for a fixed $\mathbf{u} \in \mathbb{S}^{d-1}$, the distributions $\mathcal{R}_{\mathbf{u}}\hat{\mu}_n * \mathcal{N}_\sigma$ and $\mathcal{R}_{\mathbf{u}}\hat{\nu}_n * \mathcal{N}_\sigma$ are *continuous*, in particular they are a mixture of Gaussian distributions centered on the projected samples with variance σ^2.

Lemma 3.5. *Conditionally on the samples* $\{X_i\}_{i=1,\dots,n}$ *and* $\{Y_i\}_{i=1,\dots,n}$*, one has:* $\mathcal{R}_{\mathbf{u}}\hat{\mu}_n * \mathcal{N}_\sigma = \frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mathbf{u}^\top X_i, \sigma^2)$ *and* $\mathcal{R}_{\mathbf{u}}\hat{\nu}_n * \mathcal{N}_\sigma = \frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mathbf{u}^\top Y_i, \sigma^2)$.

Note that we further need to sample with respect to the continuous mixture Gaussian measures in Lemma 3.5 in order to get a *fully* empirical measure version of $\mathrm{G}_\sigma\mathrm{SD}(\mu, \nu)$. To this end, we next define the *double* empirical divergence of $\mathrm{G}_\sigma\mathrm{SD}$.

3.2.1 Double empirical divergence of $\mathrm{G}_\sigma\mathrm{SD}$

Let $T_1^x, \dots, T_n^x$ and $T_1^y, \dots, T_n^y$ be i.i.d. observations of $\mathcal{R}_{\mathbf{u}}\hat{\mu}_n * \mathcal{N}_\sigma$ and $\mathcal{R}_{\mathbf{u}}\hat{\nu}_n * \mathcal{N}_\sigma$, respectively. Sampling i.i.d. $\{T_i^x\}_{i=1,\dots,n}$ is given by the following scheme: for $i = 1, \dots, n$, we first choose the component $\mathcal{N}(\mathbf{u}^\top X_i, \sigma^2)$ from the mixture $\frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mathbf{u}^\top X_i, \sigma^2)$ then we generate $T_i^x = \mathbf{u}^\top X_i + Z_i^x$, where $Z_i^x \sim \mathcal{N}_\sigma$. Hence, we set, for a given $\mathbf{u}$

$$\hat{\hat{\mu}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{T_i^x} = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{u}^\top X_i + Z_i^x} \text{ and } \hat{\hat{\nu}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{T_i^y} = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{u}^\top Y_i + Z_i^y}.$$

The measure $\hat{\hat{\mu}}_n \in \mathcal{P}(\mathbb{R})$ defines an empirical version of the continuous $\mathcal{R}_{\mathbf{u}}\hat{\mu}_n * \mathcal{N}_\sigma$ denoted as $\widehat{\mathcal{R}_{\mathbf{u}}\hat{\mu}_n * \mathcal{N}_\sigma}$ (similarly $\hat{\hat{\nu}}_n = \widehat{\mathcal{R}_{\mathbf{u}}\hat{\nu}_n * \mathcal{N}_\sigma}$). Using the aforementioned notation, we define.

Definition 3.6. The double empirical smoothed Gaussian sliced divergence reads as

$$\hat{\mathrm{G}}_\sigma\mathrm{SD}_p(\hat{\mu}_n, \hat{\nu}_n) \triangleq \left(\int_{\mathbb{S}^{d-1}} \mathrm{D}^p(\hat{\hat{\mu}}_n, \hat{\hat{\nu}}_n) u_d(\mathbf{u}) \mathrm{d}\mathbf{u} \right)^{1/p}.$$

Remark 3.7. *(i)* It is worth to comment the double randomnesses showing in the definition of $\hat{\mathrm{G}}_\sigma\mathrm{SD}_p(\hat{\mu}_n, \hat{\nu}_n)$: the first comes from sampling according to the original probability measure (μ or ν) whereas the second takes place from sampling according to the mixture $\frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mathbf{u}^\top X_i, \sigma^2)$.
(ii) The empirical measure of the convolution $\widehat{\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma}$ could be written as $\frac{1}{n}\sum_{i=1}^n \delta_{U_i^x + Q_i^x}$ allowing to sample *in a one shot* n i.i.d. samples $U_i^x + Q_i^x$ such that $U_i^x \sim \mathcal{R}_{\mathbf{u}}\mu$ and $Q_i^x \sim \mathcal{N}_\sigma$. From an empirical view, sampling according to $\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma$ is intractable. For that reason, our theoretical results and numerical experiments are based on $\hat{\hat{\mu}}_n, \hat{\hat{\nu}}_n$, and hence with respect to $\hat{\mathrm{G}}_\sigma\mathrm{SD}_p(\hat{\mu}_n, \hat{\nu}_n)$.

3.2.2 Sample complexity of $\mathrm{G}_\sigma\mathrm{SW}_p$

Herein, our goal is to quantify the error made when approximating $\mathrm{G}_\sigma\mathrm{SW}_p(\mu, \nu)$ with $\hat{\mathrm{G}}_\sigma\mathrm{SW}_p(\hat{\mu}_n, \hat{\nu}_n)$. More precisely, we are interested in establishing an order of the convergence rate of $\hat{\mathrm{G}}_\sigma\mathrm{SD}_p(\hat{\mu}_n, \hat{\nu}_n)$ towards $\mathrm{G}_\sigma\mathrm{SD}_p(\mu, \nu)$, according to the sample size n. This rate stands for the so-called *sample complexity*.

The convergence results in the sequel are given in expectation. Recall that the empirical distributions are derived from a double sampling process, which leads to consider a double expectations, wrt the origin distribution $\mathbf{E}_{\mu^{\otimes_n}}$ and wrt the sampling from the Gaussian smoothing $\mathbf{E}_{\mathcal{N}_\sigma^{\otimes_n}}$ where $\mu^{\otimes_n}$ and $\mathcal{N}_\sigma^{\otimes_n}$ are the n-fold product extensions of μ and $\mathcal{N}_\sigma$, respectively. We first consider the conditional expectation given the samples $X_1, \ldots, X_n$, i.e. $\mathbf{E}_{\mathcal{N}_\sigma^{\otimes_n}}[\cdot|X_1, \ldots, X_n]$, and then apply $\mathbf{E}_{\mu^{\otimes_n}}$. We denote by

$$\mathbf{E}_{\mu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}}[\cdot] = \mathbf{E}_{\mu^{\otimes_n}}\left[\mathbf{E}_{\mathcal{N}_\sigma^{\otimes_n}}[\cdot|X_1, \ldots, X_n]\right].$$

Next, we focus on the sample complexity for the special case of Gaussian-smoothed sliced Wasserstein distance.

Proposition 3.8. *Fix $\sigma > 0, p \geq 1$ and $\vartheta > \sqrt{2}$. For $X \sim \mu$, assume that $\int_0^\infty e^{\frac{2\xi^2}{\sigma^2\vartheta^2}} \mathbf{P}\big[\|X\| > \xi\big] d\xi < \infty$. Then,*

$$\mathbf{E}_{\mu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}}[\hat{\mathrm{G}}_\sigma \mathrm{SW}_p(\hat{\mu}_n, \mu)] \leq \Xi_{p,\sigma,\vartheta} \frac{1}{n^{1/2p}} + \Upsilon_{p,\sigma,\mu} \frac{(\log n)^{1/p}}{n^{1/p}},$$

where $\Xi_{p,\sigma,\vartheta} = \frac{2^{\frac{5}{2}-\frac{5}{4p}}}{\pi^{1/2p}} \sigma^{1-\frac{1}{4p}} \vartheta^{1+\frac{1}{p}} \big(\Gamma\big(p+\frac{1}{2}\big)\big(\sqrt{\frac{4\pi\sigma^2\vartheta^2}{\vartheta^2-2}} + 4\int_0^\infty e^{\frac{2\xi^2}{\sigma^2\vartheta^2}} \mathbf{P}[\|X\| > \xi] d\xi\big)\big)^{1/2p}$ *and* $\Upsilon_{p,\sigma,\mu} = \frac{2^{2-\frac{1}{2p}} C_p}{\pi^{1/2p}} \sigma^2 \big(\Gamma(p+\frac{1}{2}) \sum_{k=0}^\infty \frac{(-p)_k}{(\frac{1}{2})_k} \frac{(-1)^k}{(2\sigma^2)^k k!} M_{2k}(\mu)\big)^{1/p}$ *with C_p is a positive constant depending only on p.*

It is worth to note that for $p \in \mathbb{N}^*$, e.g. $p = 2$ (standard choice for numerical experiments), the (pseudo) confluent hypergeometric function $\sum_{k=0}^\infty \frac{(-p)_k}{(\frac{1}{2})_k} \frac{(-1)^k}{(2\sigma^2)^k k!} M_{2k}(\mu)$ is only depending on the $2k$-th moments of μ for $k = 1, \ldots, p$, since $(-p)_{(k)} = 0$ for $k \geq p+1$. Now, let us sketch the proof of Proposition 3.8: we first insert the proxy term of mixture Gaussian distribution $\frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mathbf{u}\top X_i, \sigma^2)$, then by an application of the triangle inequality on the Wasserstein distance we are faced to control two terms (i) $\mathrm{W}_p^p(\hat{\mu}_n, \frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mathbf{u}\top X_i, \sigma^2))$ and (ii) $\mathrm{W}_p^p(\frac{1}{n}\sum_{i=1}^n \mathcal{N}(\mathbf{u}\top X_i, \sigma^2), \mu)$. For (i) we get a standard order of $O(\frac{\log n}{n})$, which comes from a by-product of Fournier & Guillin (2015). For (ii), through a coupling via the maximal coupling using the total variation distance (Theorem 6.15 in Villani (2009)), we obtain the order $O(n^{-1/2})$. The control technique for (ii) was inspired from Goldfeld et al. (2020) and Nietert et al. (2021).

Remark 3.9. The condition $\int_0^\infty e^{\frac{2\xi^2}{\sigma^2\vartheta^2}} \mathbf{P}\big[\|X\| > \xi\big] \mathrm{d}\xi < \infty$ needs $\mathbf{P}\big[\|X\| > \xi\big]$ goes to 0 faster than $e^{-\kappa\xi^2}$ for $\kappa < 2/\sigma^2\vartheta^2$. This can be satisfied when $\|X\|$ is a ω-sub-gausssian ($\omega \geq 0$). Namely, $\mathbf{E}[e^{\eta^\top(X-\mathbf{E}[X])}] \leq e^{\frac{\omega\|\eta\|^2}{2}}$ for all $\eta \in \mathbb{R}^d$. If the parameter ω verifies $\omega < \sigma\vartheta/2$, then the latter condition holds.

Remark 3.10. Note that the sample complexity depends on the amount of smoothing through the moment of the Gaussian noise : the larger the amount of smoothing (and thus the privacy), the worse is the constant of the complexity. Hence, a trade-off on privacy and statistical estimation appears here as a reasonable guarantee on the differential privacy usually requires a large Gaussian variance.

Proposition 3.11. *Under the same conditions of Proposition 3.8, we have*

$$\mathbf{E}_{\mu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}} \mathbf{E}_{\nu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}}[\hat{\mathrm{G}}_\sigma \mathrm{SW}_p(\hat{\mu}_n, \hat{\nu}_n)] \leq 3^{1-\frac{1}{p}} \mathrm{G}_\sigma \mathrm{SW}_p(\mu, \nu) + 3\Xi_{p,\sigma,\vartheta} \frac{1}{n^{1/2p}} + 3^{1-\frac{1}{p}} (\Upsilon_{p,\sigma,\mu} + \Upsilon_{p,\sigma,\nu}) \frac{(\log n)^{1/p}}{n^{1/p}}$$

and

$$\mathrm{G}_\sigma \mathrm{SW}_p(\mu, \nu) \leq 3^{1-\frac{1}{p}} \mathbf{E}_{\mu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}} \mathbf{E}_{\nu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}}[\hat{\mathrm{G}}_\sigma \mathrm{SW}_p(\hat{\mu}_n, \hat{\nu}_n)] + 3\Xi_{p,\sigma,\vartheta} \frac{1}{n^{1/2p}} + 3^{1-\frac{1}{p}} (\Upsilon_{p,\sigma,\mu} + \Upsilon_{p,\sigma,\nu}) \frac{(\log n)^{1/p}}{n^{1/p}}.$$

Proof of Proposition 3.11 relies on a double application of triangle inequality satisfied by Wasserstein distance as follows: $\mathrm{W}_p(\hat{\mu}_n, \hat{\nu}_n) \leq \mathrm{W}_p(\hat{\mu}_n, \mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma) + \mathrm{W}_p(\mathcal{R}_\mathbf{u}\mu * \mathcal{N}_\sigma, \mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma) + \mathrm{W}_p(\mathcal{R}_\mathbf{u}\nu * \mathcal{N}_\sigma, \hat{\nu}_n)$, combined with Proposition 3.8. This gives a non sharp convergence result since we get the constant $3^{1-\frac{1}{p}}$ in front of $\mathbf{E}_{\mu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}} \mathbf{E}_{\nu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}}[\hat{\mathrm{G}}_\sigma \mathrm{SW}_p(\hat{\mu}_n, \hat{\nu}_n)]$ or $\mathrm{G}_\sigma \mathrm{SW}_p(\mu, \nu)$. However, when the power $p = 1$ we obtain a sharp convergence result with $O(n^{-1/2})$, namely

$$\mathbf{E}_{\mu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}} \mathbf{E}_{\nu^{\otimes_n}|\mathcal{N}_\sigma^{\otimes_n}}[|\hat{\mathrm{G}}_\sigma \mathrm{SW}(\hat{\mu}_n, \hat{\nu}_n) - \mathrm{G}_\sigma \mathrm{SW}(\mu, \nu)|] \leq 3\Xi_{1,\sigma,\vartheta} \frac{1}{\sqrt{n}} + (\Upsilon_{1,\sigma,\mu} + \Upsilon_{1,\sigma,\nu}) \frac{\log n}{n}$$

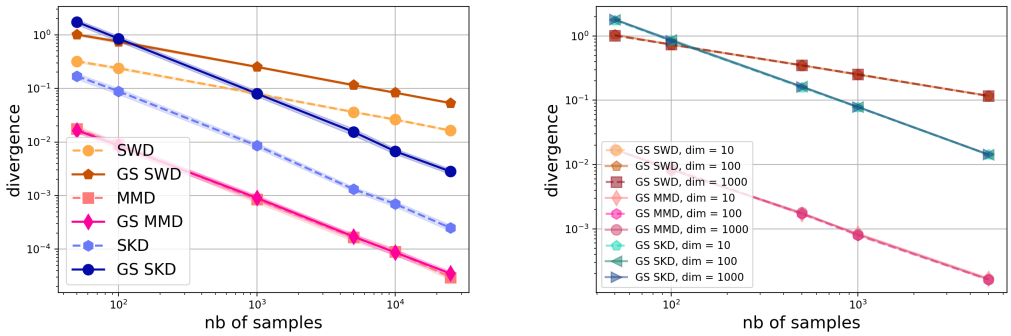

Figure 1: Measuring the divergence between two sets of samples in $\mathbb{R}^{50}$, of increasing size, randomly drawn from $\mathcal{N}(0, \mathbf{I})$. We compare three sliced divergences and their Gaussian-smoothed sliced versions with a $\sigma = 3$: (top) dimension has been set to $d = 50$; (bottom)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

Despite that our theoretical results hold only for Gaussian-smoothed sliced Wasserstein distance, our empirical results show that given other base divergences D, shows that the sample complexity of $\mathrm{G}_\sigma\mathrm{SD}^p$ is proportional to the one dimensional sample complexity of D^p ($p = 2$). Figure 1 provides an empirical illustration of this statement.

3.2.3 Projection complexity

To compute the Gaussian-smoothed sliced divergence, one may resort to a Monte Carlo scheme to numerically approximate the integral in $\mathrm{G}_\sigma\mathrm{SD}_p(\mu, \nu)$. Towards this, let define the following sum:

$$\widehat{\mathrm{G}_\sigma\mathrm{SD}}_p(\mu, \nu) = \Big(\frac{1}{L}\sum_{l=1}^{L} \mathrm{D}_p(\mathcal{R}_{\mathbf{u}_l}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}_l}\nu * \mathcal{N}_\sigma)\Big)^{1/p}, \tag{1}$$

where $\mathbf{u}_l$ is a random vector uniformly drawn from $\mathbb{S}^{d-1}$, for $l = 1, \ldots, L$. Theorem 3.12 shows that for a fixed dimension d, the root mean square error of Monte Carlo (MC) approximation is of order $O(\frac{1}{\sqrt{L}})$, which corresponds to the projection complexity. We denote by $u_d^{\otimes_L}$ and the L-fold product extensions of the uniform measure u_d on the unit sphere.

Proposition 3.12. *Let $\sigma > 0, p \geq 1$. Then the error related to the MC-estimation of $\mathrm{G}_\sigma\mathrm{SD}_p$ is bounded as follows*

$$\mathbf{E}_{u_d^{\otimes_L}}\big[\big|\widehat{\mathrm{G}_\sigma\mathrm{SD}}_p^{\,p}(\mu, \nu) - \mathrm{G}_\sigma\mathrm{SD}_p^p(\mu, \nu)\big|\big] \leq \frac{A(p, \sigma)}{\sqrt{L}},$$

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

The term $A^2(p, \sigma)$ corresponds to the variance of $\mathrm{D}^p(\mathcal{R}_{\mathbf{u}}\mu * \mathcal{N}_\sigma, \mathcal{R}_{\mathbf{u}}\nu * \mathcal{N}_\sigma)$ with respect to $\mathbf{u} \sim u_d$. It is worth to note that the precision 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.

Given the above results, we provide a finer analysis of $\mathrm{G}_\sigma\mathrm{SW}_p(\mu, \nu)$'s sample complexity. Towards this ends, for a fixed random projection $\mathbf{u}_l, (1 \leq l \leq L)$ we define $\hat{\hat{\mu}}_{n,l} = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{u}_l^\top X_i + Z_i^x}$ (similarly for $\hat{\hat{\nu}}_{n,l}$) and set

$$\widehat{\hat{\mathrm{G}}_\sigma\mathrm{SD}}_p(\hat{\mu}_n, \hat{\nu}_n) = \Big(\frac{1}{L}\sum_{l=1}^{L} \mathrm{W}_p^p(\hat{\hat{\mu}}_{n,l}, \hat{\hat{\nu}}_{n,l})\Big)^{1/p}$$

The overall complexity of $\mathrm{G}_\sigma\mathrm{SD}_p(\mu,\nu)$ consists in its approximation by sampling and projection of the origin probability measures μ,ν, i.e. through $\widehat{\hat{\mathrm{G}}_\sigma\mathrm{SD}}_p(\hat{\mu}_n,\hat{\nu}_n)$. By application of triangle inequality, one has

$$|\widehat{\hat{\mathrm{G}}_\sigma\mathrm{SW}}_p^{\,p}(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{G}_\sigma\mathrm{SW}_p^p(\mu,\nu)|\leq|\widehat{\hat{\mathrm{G}}_\sigma\mathrm{SW}}_p^{\,p}(\hat{\mu}_n,\hat{\nu}_n)-\hat{\mathrm{G}}_\sigma\mathrm{SW}_p^p(\hat{\mu}_n,\hat{\nu}_n)|+|\hat{\mathrm{G}}_\sigma\mathrm{SW}_p^p(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{G}_\sigma\mathrm{SW}_p(\mu,\nu)|.$$

The first term in the right-hand-side (RHS) of the latter decomposition can be controlled by Proposition 3.12 in the following way:

$$\mathbf{E}_{u_d^{\otimes_L}}\left[|\widehat{\hat{\mathrm{G}}_\sigma\mathrm{SW}}_p^{\,p}(\hat{\mu}_n,\hat{\nu}_n)-\hat{\mathrm{G}}_\sigma\mathrm{SW}_p^p(\hat{\mu}_n,\hat{\nu}_n)|\right]\leq\frac{\hat{A}(p,\sigma)}{\sqrt{L}}\triangleq\frac{\{\mathbf{V}_{\mathbf{u}\sim u_d}[\mathrm{W}_p^p(\hat{\mu}_n,\hat{\nu}_n)]\}^{1/2}}{\sqrt{L}}.$$

However we don't have a proper control for $p\geq 2$ of the second term in the RHS, $|\hat{\mathrm{G}}_\sigma\mathrm{SW}_p^p(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{G}_\sigma\mathrm{SW}_p(\mu,\nu)|$, as it can be seen from Proposition 3.11. For that reason, we derive an overall complexity in the case of $p=1$.

Corollary 3.13. *The sample and projection complexities of* $\mathrm{G}_\sigma\mathrm{SW}(\mu,\nu)$ *reads as* $\mathrm{complexity}(\mathrm{G}_\sigma\mathrm{SW})=O(n^{-1/2}+L^{-1/2})$. *If we consider the number of projections as* $L=\lfloor n^\beta\rfloor$ *for some* $\beta\in(0,1)$ *then the overall complexity* $\mathrm{complexity}(\mathrm{G}_\sigma\mathrm{SW}(\mu,\nu))=O(n^{-\beta/2})$.

3.3 Noise-level dependencies

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.

3.4 Order relation.

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

Proposition 3.14. *Let* $\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d)$ *and consider the noise levels* σ_1,σ_2 *such that* $0\leq\sigma_1\leq\sigma_2<\infty$. *Assume that the base divergence* D *satisfies* $\mathrm{D}(\mu'*\mathcal{N}_{\sigma_2},\nu'*\mathcal{N}_{\sigma_2})\leq\mathrm{D}(\mu'*\mathcal{N}_{\sigma_1},\nu'*\mathcal{N}_{\sigma_1})$, *for any* $\mu',\nu'\in\mathcal{P}(\mathbb{R})$. *Then,* $\mathrm{G}_{\sigma_2}\mathrm{SD}^p(\mu,\nu)\leq\mathrm{G}_{\sigma_1}\mathrm{SD}^p(\mu,\nu)$.

Note that the assumption for the base divergence inequality holds for the Gaussian-smoothed Wasserstein distance Nietert et al. (2021). 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 3.14, we show some specific properties of the metric with respect to the noise level σ.

Proposition 3.15. $\mathrm{G}_\sigma\mathrm{SD}^p(\mu,\nu)$ *is decreasing with respect to* σ *and we have* $\lim_{\sigma\to 0}\mathrm{G}_\sigma\mathrm{SD}^p(\mu,\nu)=\mathrm{D}^p(\mu,\nu)$.

The proof of Proposition 3.15 comes straightforwardly from Proposition 3.14 by taking $\sigma_2=\sigma$ and letting $\sigma_1\to 0$. This property interestingly states that the $\mathrm{G}_\sigma\mathrm{SD}^p$ recovers the sliced divergence when the noise level vanishes. We end up this section by providing a relation between Gaussian-smoothed sliced Wasserstein distances under two noise levels.

Proposition 3.16. *Let* $0\leq\sigma_1\leq\sigma_2$ *be two noise levels. Then, one has* $\mathrm{G}_{\sigma_2}\mathrm{SW}_p(\mu,\nu)\leq\mathrm{G}_{\sigma_1}\mathrm{SW}_p(\mu,\nu)$ *and*

$$|\,\mathrm{G}_{\sigma_1}\mathrm{SW}_p(\mu,\nu)-\mathrm{G}_{\sigma_2}\mathrm{SW}_p(\mu,\nu)|\leq(2^{1-\frac{1}{p}}-1)\,\mathrm{G}_{\sigma_2}\mathrm{SW}_p(\mu,\nu)+2^{\frac{5}{2}}(\sigma_2^2-\sigma_1^2),$$

in particular for $p=1$, $|\,\mathrm{G}_\sigma\mathrm{SW}(\mu,\nu)-\mathrm{G}_{\sigma_2}\mathrm{SW}(\mu,\nu)|\leq 2^{\frac{5}{2}}(\sigma_2^2-\sigma_1^2)$.

3.4.1 Continuity

Now we analyze the continuity properties of some $\mathrm{G}_\sigma\mathrm{SD}^p(\mu,\nu)$ w.r.t. the noise level.

Proposition 3.17. *For any two distributions* μ *and* ν *for which the sliced Wasserstein is well-defined, the Gaussian-smoothed sliced Wasserstein distance is continuous w.r.t. to* σ.

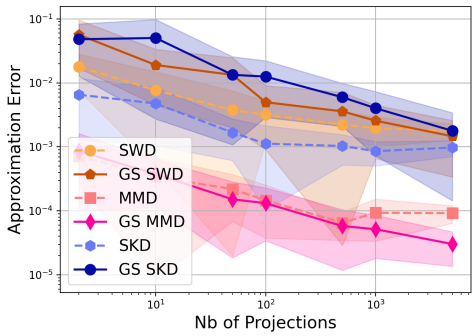

Figure 2: Absolute difference between the approximated Monte Carlo approximation of all divergences compared to the true one (evaluated with $10,000$ number of projections). The two sets of 500 samples in $\mathbb{R}^{50}$ are randomly drawn from $\mathcal{N}(0, \mathbf{I})$. The Gaussian-smoothed sliced divergences are parameterized with $\sigma = 3$.

Proposition 3.18. *Assume that the kernel defining the maximum mean discrepancy* (MMD) *divergence is bounded. Then the Gaussian-smoothed sliced* G_σMMD *is continuous w.r.t. to* σ.

The above propositions show that most distribution divergences are continuous with respect to σ under mild conditions.

4 Numerical Experiments

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

4.1 Supporting the theoretical results

Sample complexity. The first experiment (see Figure 1) analyzes the sample complexity of different base divergences. It shows that the sample complexity stays similar to the one of their original and sliced counterparts up to a constant (see Proposition 3.8). 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 sliced versions, named as GS SWD, GS SKD and GS MMD. On the top plot, we can see that all Gaussian-smoothed sliced divergences preserve the complexity rate with just a slight to moderate overhead. The worst difference is for Sinkhorn divergence, while MMD almost comes for free in term of complexity. From the bottom 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.

Two other experiments on the sample complexity and identity of indiscernibles are also reported in the supplementary material.

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 2 plots the approximation error between the true expectation of the sliced divergences (computed for a number of

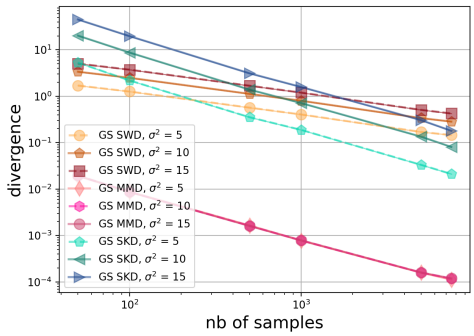

Figure 3: Measuring the divergence between two sets of samples in $\mathbb{R}^{50}$ drawn from $\mathcal{N}(0, \mathbf{I})$. We plot the sample complexity for different Gaussian-smoothed sliced divergence at different level of noises.

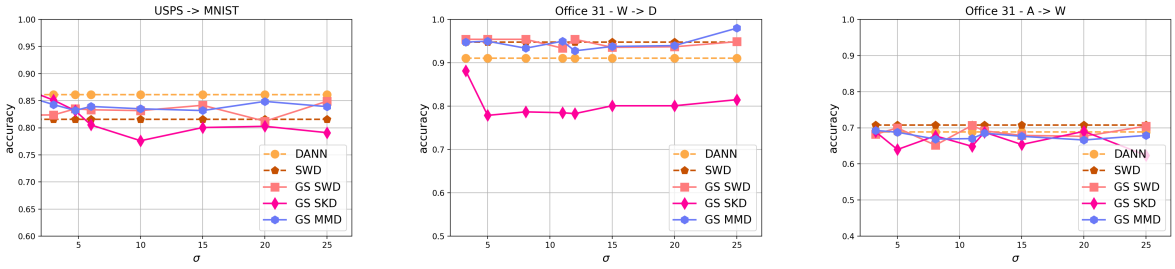

Figure 4: Domain adaptation performances using different divergences on distributions with respect to the Gaussian smoothing. (Left) USPS to MNIST. (Middle) Office-31 Webcam to DSLR. (Right) Office-31 Amazon to Webcam.

$L = 10,000$ 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.

Performance path on the impact of the noise parameter. Since the Gaussian smoothing parameter σ is key in a privacy preserving context, as it impacts on the level of privacy of the Gaussian mechanism, we have analyzed its impact on 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 3 shows these sample complexities. The first very interesting point to note is that the smoothing parameter has almost no effect on the GS MMD sample complexity. For the GS SWD and GS SKD 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 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 sliced SWD in our Proposition 3.8, 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 $\mathrm{G}_\sigma\mathrm{SW}$

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 $\mathbf{X}_t$,

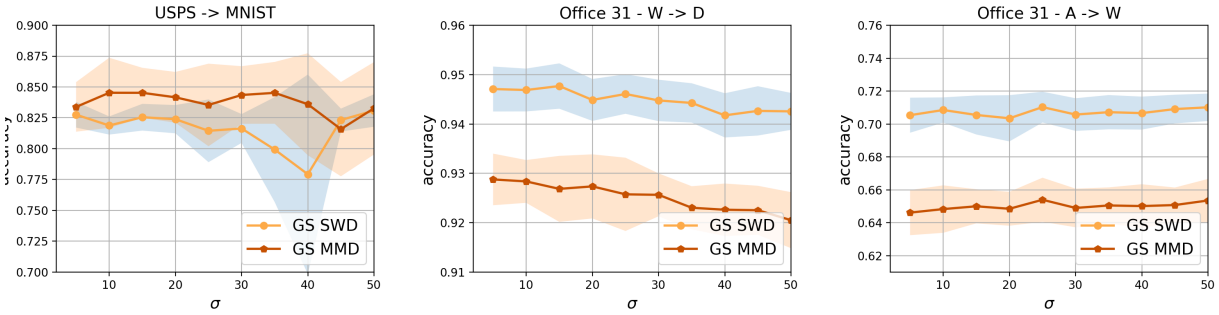

Figure 5: Domain adaptation performances using different divergences on distributions with respect to the Gaussian smoothing using **one-epoch-fine-tuned** models. (Left) USPS to MNIST. (Middle) Office-31 Webcam to DSLR. (Right) Office-31 Amazon to Webcam.

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

$$\min_{g,h} \left\{ \mathcal{L}_c(h(g(\mathbf{X}_s)), \mathbf{y}_s) + \mathcal{D}(g(\mathbf{X}_s), g(\mathbf{X}_t)) \right\}$$

where $\mathcal{L}_c$ 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 stochastic gradient descent, similarly to many approaches that use sliced Wasserstein distance as a distribution distance Lee et al. (2019). Note that, in practice, using a smoothed divergence preserves the privacy of the target samples as shown by (Rakotomamonjy & Ralaivola, 2021).

When performing such model adaptation, a privacy/utility trade-off that has to be handled. In practice, one would prefer the most private model while not hurting its performance. Hence, one would seek the largest noise level $\sigma > 0$ to use while preserving accuracy on target domain. Hence, it is useful to evaluate how the model performs on a range of noise level (hence, privacy level). This can be computationally expensive at it requires to fully train several models on hundreds of epochs. Instead, we leverage on the continuity of our G_σSD to employ a fine-tuning strategy: we train a domain adaptation model for the largest desired value of σ (over the full number of epochs) and when σ is decreased, we just fine-tune the lasted model by training on only one epoch.

Our experiments evaluate the studied Gaussian-smoothed sliced divergences in classical unsupervised domain adaptation. We have considered two datasets: a handwritten digit recognition (USPS/MNIST) and Office 31 datasets.

In our first analysis, we have compared our G_σSD performances with non-smoothed divergences. The first one is the sliced Wasserstein distance (SWD) Lee et al. (2019) and the second one is the Jenssen-Shannon approximation based on adversarial approach, known as DANN Ganin & Lempitsky (2015). 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. Here for each noise value σ, we have trained the model from scratch for 100 epochs. Results are depicted in Figure 4. 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.

In the second analysis, we have studied the privacy/utility trade-off when fine-tuning models, using only one epoch, for decreasing values of σ. Results are shown in Figure 5. They highlight that depending on the data and the used smoothed divergence, performance varies between one percent for Office 31 to four percent for

USPS to MNIST. Note that except for the largest value of σ, we are training a model using only one epoch instead of a hundred. A very large gain in complexity is thus achieved for swiping the full range of noise level. Hence depending on the importance this slight drop in performance will have, it is worth using a large value of σ and preserving strong privacy or go through a validation procedure of several (cheaply obtained) models.

5 Conclusion

This work provided the properties of Gaussian-smoothed sliced divergences for comparing distributions. We derived several theoretical results related to their topological and statistical properties and showed, under mild conditions on their base divergences, the smoothing and slicing operations preserves the metric property. From a statistical point of view, we introduced the double empirical distribution and focused on the sample complexity of the smoothed sliced Wasserstein distance and we proved that it converges with a rate $O(n^{-1/2})$. We furhter analyzed the behavior of these divergences on domain adaptation problems and confirm the fact that using those divergences yields only to slight loss of performances while preserving privacy. Note that in the obtained bound we use upper bound of higher moments of the smoothing distribution. An important direction for future research is considering non Gaussian smoothing distribution enjoying this property.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Nicolas Bonneel and David Coeurjolly. Spot: Sliced partial optimal transport. 38(4), 2019. ISSN 0730-0301.

Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Trans. Graph.*, 30(6):158:1–158:12, 2011. ISSN 0730-0301.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G. Schwing. 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.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015. ISSN 1432-2064.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Leonid V. Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 2:227–229, 1942.

Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein kernels for probability distributions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5258–5267, 2016.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4): 43–59, July 2017. ISSN 1053-5888.

Soheil Kolouri, Phillip E. Pope , Charles E. Martin, and Gustavo K. Rohde. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2019a.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 261–272. Curran Associates, Inc., 2019b.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10285–10295, 2019.

Tianyi Lin, Zeyu Zheng, Elynn Chen, Marco Cuturi, and Michael Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pp. 262–270. PMLR, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

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

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Şimşekli. Statistical and topological properties of sliced probability divergences. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

Khai Nguyen, Tongzheng Ren, Huy Nguyen, Litu Rout, Tan Nguyen, and Nhat Ho. Hierarchical sliced wasserstein distance. *arXiv preprint arXiv:2209.13570*, 2022.

Khai Nguyen, Dang Nguyen, and Nhat Ho. Self-attention amortized distributional projection optimization for sliced wasserstein point-cloud reconstruction. In *International Conference on Machine Learning*, pp. 26008–26030. PMLR, 2023.

Khai Nguyen, Tongzheng Ren, and Nhat Ho. Markovian sliced wasserstein distances: Beyond independent projections. *Advances in Neural Information Processing Systems*, 36, 2024.

Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications. In Marina Meila and Tong Zhang (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.

Frank W. J. Olver. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trendső in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

Svetlozar T. Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*. Mass Transportation Problems. Springer, 1998. ISBN 9780387983509.

Alain Rakotomamonjy and Liva Ralaivola. Differentially private sliced wasserstein distance. In Marina Meila and Tong Zhang (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.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

Justin Solomon, Fernando d de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, 2015.

Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR (Poster)*, 2017.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, 2009. ISBN 9783540710509.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3713–3722, 2019.