Shedding a PAC-Bayesian Light on Adaptive Sliced-Wasserstein Distances

Ruben Ohana [*1] **Kimia Nadjahi** [*2] **Alain Rakotomamonjy** [3] **Liva Ralaivola** [3]

Abstract

The Sliced-Wasserstein distance (SW) is a computationally efficient and theoretically grounded alternative to the Wasserstein distance. Yet, the literature on its statistical properties – or, more accurately, its *generalization* properties – with respect to the distribution of slices, beyond the uniform measure, is scarce. To bring new contributions to this line of research, we leverage the PAC-Bayesian theory and a central observation that SW may be interpreted as an average risk, the quantity PAC-Bayesian bounds have been designed to characterize. We provide three types of results: i) PAC-Bayesian generalization bounds that hold on what we refer as *adaptive* Sliced-Wasserstein distances, i.e. SW defined with respect to arbitrary distributions of slices (among which data-dependent distributions), ii) a principled procedure to learn the distribution of slices that yields maximally discriminative SW, by optimizing our theoretical bounds, and iii) empirical illustrations of our theoretical findings.

1. Introduction

The Wasserstein distance is a metric between probability distributions and a key notion of the optimal transport framework (Villani, 2009; Peyré & Cuturi, 2019). Over the past years, it has received a lot of attention from the machine learning community because of its theoretical grounding and the increasing number of problems relying on the computation of distances between measures (Solomon et al., 2014; Frogner et al., 2015; Montavon et al., 2016; Kolouri et al., 2017; Courty et al., 2016; Schmitz et al., 2018), such as the learning of deep generative models (Arjovsky et al., 2017; Bousquet et al., 2017; Tolstikhin et al., 2017). As the measures μ and ν to be compared are usually unknown, the Wasserstein distance $W(\mu, \nu)$ is estimated through an "empirical" version $W(\mu_n, \nu_n)$, where $\mu_n \doteq \{x_1, \dots, x_n\}$ and $\nu_n \doteq \{y_1, \dots, y_n\}$ are i.i.d. samples from μ and ν, respectively (without loss of generality, samples will be assumed to have the same size n). Due to its unfavorable $\mathcal{O}(n^3 \log n)$ computational complexity, the Wasserstein distance scales badly on large datasets (Peyré & Cuturi, 2019) and alternatives have been devised to overcome this limitation, such as the Sinkhorn algorithm (Cuturi, 2013; Cuturi & Peyré, 2016), multi-scale (Oberman & Ruan, 2015) or sparse approximations approaches (Schmitzer, 2016).

The Sliced-Wasserstein distance (SW) (Rabin et al., 2012) is another computationally efficient alternative, which takes advantage of the closed-form and fast computation of the one-dimensional Wasserstein distance. For d-dimensional ($d > 1$) samples $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$, the computation of $\mathrm{SW}(\mu_n, \nu_n)$ is done by uniformly sampling m *projection directions* $\{\theta_1, \dots, \theta_m\}$ and by averaging the m one-dimensional Wasserstein distances $W(\{\langle \theta_j, x_1 \rangle, \dots, \langle \theta_j, x_n \rangle\}, \{\langle \theta_j, y_1 \rangle, \dots, \langle \theta_j, y_n \rangle\})$ for $j = 1, \dots, m$. SW has been analyzed theoretically (Bonnotte, 2013; Nadjahi et al., 2019; Bayraktar & Guo, 2021; Nadjahi et al., 2020b), refined to gain additional efficiency (Nadjahi et al., 2021) and to handle "nonlinear" projections (Kolouri et al., 2019a; 2020), and it has been successfully used in a variety of machine learning tasks (Bonneel et al., 2015; Kolouri et al., 2016; Carriere et al., 2017; Liutkus et al., 2019; Deshpande et al., 2018; Kolouri et al., 2018; 2019b; Nadjahi et al., 2020a; Bonet et al., 2021; Rakotomamonjy & Ralaivola, 2021). A direction to yet improve SW consists in adapting ρ, the distribution of $\{\theta_i\}_{i=1}^m$ in a data-dependent manner, as done by *maximum SW* (max-SW, (Deshpande et al., 2019)), which aims at finding a unique slice $\theta_\star$ (or equivalently, the Dirac measure $\delta_{\theta_\star}$) that maximizes the one-dimensional Wasserstein distance, or *distributional SW* (DSW) (Nguyen et al., 2021), which seeks for a maximally discriminative distribution on the unit sphere. These works fall into the class of what we refer as *adaptive Sliced-Wasserstein distances* and denote $\mathrm{SW}(\cdot, \cdot; \rho)$, overloading the $\mathrm{SW}(\cdot, \cdot)$ notation to make explicit the dependence on ρ.

A question of interest in adaptive SW, which has not been explicitly addressed in previous work, is whether one can learn a distribution $\rho^\star(\mu_n, \nu_n)$ from training data, such that $\mathrm{SW}_p^p(\mu, \nu; \rho^\star(\mu_n, \nu_n))$ is guaranteed to be highly discrim-

[*]Equal contribution [1]Flatiron Institute, USA [2]MIT, USA [3]Criteo AI Lab, France. Correspondence to: RO <rohana@flatironinstitute.org>, KN <knadjahi@mit.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

inative. In our work, we address this problem by measuring the "generalization" gap between $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ and $\mathrm{SW}_p^p(\mu, \nu; \rho)$. Bounds on this gap can be derived from existing results for max-SW (Lin et al., 2021; Niles-Weed & Rigollet, 2022). However, it is unclear how these bounds are able to accommodate distributions ρ that are not reduced to Dirac measures. To go that direction, we propose the first connection between adaptive SW and *PAC-Bayesian theory* and we derive a novel set of flexible PAC-Bayesian generalization bounds. Our bounds state that with probability $1-\delta$, the following holds for all measures (with non-discrete support) ρ on the d-dimensional unit sphere: $\mathrm{SW}(\mu, \nu; \rho) \geq \mathrm{SW}(\mu_n, \nu_n, \rho) - \varepsilon(n, \rho, \delta)$, where ε can be written explicitly and captures the properties of μ, ν, and allows us to control the tightness of the bound via ρ.

Three key reasons make the PAC-Bayesian theory (McAllester, 1999; Catoni, 2007; Alquier, 2021) particularly suited to characterize the generalization properties of adaptive SW. First, from a general perspective, the literature shows this framework allows the derivation of tight bounds that can be converted into effective learning procedures (Ambroladze et al., 2007; Laviolette et al., 2006; Germain et al., 2009; Zantedeschi et al., 2021). Second, PAC-Bayesian bounds deal with the generalization ability of learned distributions; while those distributions usually lie on spaces of predictors, the distributions ρ of interest in our case are the distributions of slices. Lastly, a key quantity of PAC-Bayesian bounds is the *average empirical risk* which, as we will show, can naturally be interpreted as $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$, our main focus.

The paper is organized as follows. In Section 2, we recall essential notions of Sliced-Wasserstein distances and PAC-Bayesian theory. We then delve into our contributions: *i)* a generic PAC-Bayesian bound for adaptive Sliced-Wasserstein distances and refinements to specific settings (Section 3), *ii)* a theoretically-grounded procedure to train a maximally discriminative Sliced-Wasserstein distances (Section 4) and *iii)* illustrations of the soundness of our theoretical results through numerical experiments, conducted on both toy and real-world datasets (Section 5).

Notations. Let $d \in \mathbb{N}^*$ with $\mathbb{N}^* \doteq \mathbb{N} \backslash \{0\}$. For $x, y \in \mathbb{R}^d$, $\langle x, y \rangle$ denotes the dot product between x and y, and $\|x\|$ is the Euclidean norm of x. For $\mathsf{X} \subset \mathbb{R}^d$, $\mathcal{P}(\mathsf{X})$ is the set of probability measures supported on X, and $\mathcal{P}_q(\mathsf{X})$ is the set of probability measures supported on X with finite moment of order q. $\mathcal{U}(\mathsf{X})$ is the uniform distribution on X, and δ_x is the Dirac measure with mass on $x \in \mathsf{X}$. For $\mu \in \mathcal{P}(\mathsf{X})$ and $n \in \mathbb{N}^*$, $\mu_n \doteq \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical measure supported on n samples $\{x_1, \ldots, x_n\}$ i.i.d. from μ. For $\mu \in \mathcal{P}(\mathbb{R})$, F_μ is the cumulative distribution function (c.d.f.) of μ and F_μ^{-1} is its quantile function.

2. Background

2.1. Sliced-Wasserstein Distances

Sliced-Wasserstein distances refer to a family of distances between probability measures, which was first introduced by (Rabin et al., 2012) to overcome the computational issues of the Wasserstein distance. We formally define the Wasserstein distance and SW, and explain why the latter can provide significant computational advantages over the former. In what follows, we fix $\mathsf{X} \subset \mathbb{R}^d$.

Definition 1 (Wasserstein distance)**.** *Let $p \in [1, +\infty)$. The Wasserstein distance of order p between $\mu, \nu \in \mathcal{P}(\mathsf{X})$ is*

$$\mathrm{W}_p^p(\mu, \nu) \doteq \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{X}} \|x-y\|^p \mathrm{d}\pi(x, y)\,,$$

where $\Pi(\mu, \nu) \subset \mathcal{P}(\mathsf{X} \times \mathsf{X})$ denotes the set of probability measures on $\mathsf{X} \times \mathsf{X}$, whose marginals with respect to the first and second variables are μ and ν respectively.

While W_p has been shown to possess appealing theoretical properties, e.g. it is a metric on $\mathcal{P}_p(\mathsf{X})$ which *metrizes* the weak convergence (Villani, 2009, Chapter 6), it is computationally too demanding in general. Indeed, consider two discrete distributions μ_n, ν_n, each supported on n samples. Computing $\mathrm{W}_p(\mu_n, \nu_n)$ means solving a linear program (Peyré & Cuturi, 2019, Section 3.1), whose solution is not analytically available in general, but can be approximated with standard solvers from linear programming and combinatorial optimization. However, such methods have a super-cubic cost in practice, and their worst-case computational complexity scales in $\mathcal{O}(n^3 \log n)$.

Nevertheless, if $\mu, \nu \in \mathcal{P}(\mathbb{R})$, $\mathrm{W}_p(\mu, \nu)$ admits an analytical expression which can be efficiently approximated (Peyré & Cuturi, 2019, Section 2.6): for any $\mu, \nu \in \mathcal{P}(\mathbb{R})$,

$$\mathrm{W}_p^p(\mu, \nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p \mathrm{d}t\,.$$

In particular, for $\mu_n = (1/n)\sum_{i=1}^n \delta_{x_i}$ and $\nu_n = (1/n)\sum_{i=1}^n \delta_{y_i}$ such that, $\forall i \in \{1, \ldots, n\}$, $x_i, y_i \in \mathbb{R}$,

$$\mathrm{W}_p^p(\mu_n, \nu_n) = \frac{1}{n}\sum_{i=1}^n |x_{(i)} - y_{(i)}|^p\,, \tag{1}$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$. Computing (1) thus consists in sorting the support points of μ_n and ν_n, which induces $\mathcal{O}(n \log n)$ operations.

Sliced-Wasserstein distances leverage the fast computation of $\mathrm{W}_p(\mu, \nu)$ for any $\mu, \nu \in \mathcal{P}(\mathbb{R})$ to efficiently compare distributions supported on medium to high-dimensional spaces. Their formal characterization is given below.

Definition 2 (Sliced-Wasserstein distance)**.** *Let $\mathbb{S}^{d-1} \doteq \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ be the unit sphere in $\mathbb{R}^d$. For*

$\theta \in \mathbb{S}^{d-1}$, denote by $\theta^* : \mathbb{R}^d \to \mathbb{R}$ the linear map such that for $x \in \mathbb{R}^d$, $\theta^*(x) \doteq \langle \theta, x \rangle$. Let $p \in [1, +\infty)$ and $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. The Sliced-Wasserstein distance of order p based on ρ *is defined for $\mu, \nu \in \mathcal{P}(\mathsf{X})$ as*

$$\mathrm{SW}_p^p(\mu, \nu; \rho) \doteq \int_{\mathbb{S}^{d-1}} W_p^p(\theta^*_\sharp \mu, \theta^*_\sharp \nu) \mathrm{d}\rho(\theta)\,, \qquad (2)$$

where for any measurable function f and $\xi \in \mathcal{P}(\mathbb{R}^d)$, $f_\sharp \xi$ is the push-forward measure *of ξ by f: for any measurable set $\mathrm{A} \subset \mathbb{R}$, $f_\sharp \xi(\mathrm{A}) \doteq \xi(f^{-1}(\mathrm{A}))$, $f^{-1}(\mathrm{A}) \doteq \{x \in \mathbb{R}^d : f(x) \in \mathrm{A}\}$.*

Computational complexity of SW. By (2), $\mathrm{SW}_p^p(\mu, \nu; \rho)$ is obtained by computing $\mathbb{E}[\mathrm{W}_p^p(\theta^*_\sharp \mu, \theta^*_\sharp \nu)]$ with $\mathbb{E}$ taken over $\theta \sim \rho$. This expectation is intractable in general, and commonly approximated with the Monte Carlo estimate

$$\widehat{\mathrm{SW}}_p^p(\mu, \nu; \rho) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{W}_p^p((\theta_j)^*_\sharp \mu, (\theta_j)^*_\sharp \nu)\,, \qquad (3)$$

where $\{\theta_j\}_{j=1}^m$ are i.i.d. samples from ρ. Note that for $\theta \in \mathbb{S}^{d-1}$, $\theta^*_\sharp \mu$ and $\theta^*_\sharp \nu$ are one-dimensional probability measures, which can be interpreted as projections of μ and ν along θ. To illustrate this, consider $\mu_n = (1/n) \sum_{i=1}^n \delta_{x_i}$ with $x_i \in \mathbb{R}^d$ for $i \in \{1, \dots, n\}$. By definition, $\theta^*_\sharp \mu_n = (1/n) \sum_{i=1}^n \delta_{\langle \theta, x_i \rangle}$. Therefore, computing (3) between μ_n and ν_n amounts to projecting $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ along $\theta_j \sim \rho$, then computing the one-dimensional Wasserstein distance using (1), for $j \in \{1, \dots, m\}$. This scheme requires $\mathcal{O}\big(m(dn + n \log n)\big)$ operations which is, in general, faster than computing $\mathrm{W}_p^p(\mu_n, \nu_n)$, especially for large n.

Theoretical properties of SW. Previous works have investigated theoretical properties of $\mathrm{SW}_p^p(\cdot, \cdot; \rho)$, to explain its empirical success (Bonnotte, 2013; Bayraktar & Guo, 2021; Nadjahi et al., 2019; Lin et al., 2021; Nguyen et al., 2021). However, most results apply to $\rho = \mathcal{U}(\mathbb{S}^{d-1})$ only (which corresponds to the original definition of SW, (Rabin et al., 2012)). In particular, whether (2) is a metric for any ρ has not been established: we show in Appendix A1.1 that $\mathrm{SW}_p^p(\cdot, \cdot; \rho)$ is always a pseudo-metric, and we discuss for which choices of ρ it satisfies all metric axioms.

Adaptive SW. Recent works have argued that the uniform distribution may not be the most relevant choice, depending on the task at hand. Instead, they proposed to learn ρ from the observed data. This strategy provides $\mathrm{SW}_p^p(\cdot, \cdot; \rho)$ with an actual degree of freedom ρ, and motivates the term *adaptive* Sliced-Wasserstein distance. Specifically, (Deshpande et al., 2019) and (Nguyen et al., 2021) solve a tailored optimization problem in ρ targetting a high discriminative power of ρ, in the sense that ρ puts more mass on the $\theta \in \mathbb{S}^{d-1}$ that maximize the separation of $\theta^*_\sharp \mu$ and $\theta^*_\sharp \nu$. The *maximum Sliced-Wasserstein distance* (max-SW, (Deshpande et al., 2019)) is defined as

$$\mathrm{maxSW}(\mu, \nu) \doteq \mathrm{SW}_p^p(\mu, \nu; \rho^\star_{\mathrm{maxSW}}(\mu, \nu)) \qquad (4)$$

$$\text{with} \quad \rho^\star_{\mathrm{maxSW}}(\mu, \nu) \doteq \underset{\delta_\theta : \theta \in \mathbb{S}^{d-1}}{\arg\sup}\, \mathrm{SW}_p^p(\mu, \nu; \delta_\theta)\,, \qquad (5)$$

while the *distributional Sliced-Wasserstein distance* (DSW, (Nguyen et al., 2021)) is given by

$$\mathrm{DSW}(\mu, \nu) \doteq \mathrm{SW}_p^p(\mu, \nu; \rho^\star_{\mathrm{DSW}}(\mu, \nu)) \qquad (6)$$

$$\text{with} \quad \rho^\star_{\mathrm{DSW}}(\mu, \nu) \doteq \underset{\substack{\rho \in \mathcal{P}(\mathbb{S}^{d-1}), \\ \mathbb{E}_{\theta, \theta' \sim \rho} |\theta^\top \theta'| \leq C}}{\arg\sup}\, \mathrm{SW}_p^p(\mu, \nu; \rho) \qquad (7)$$

where in (7), θ and θ' are independent and $C > 0$ is a hyperparameter. We have decoupled the search for the maximizing distances (4),(6) and the maximum arguments (5),(7) for reasons we clarify below.

While there exist statistical guarantees on the gap between $\mathrm{maxSW}(\mu, \nu)$ and $\mathrm{maxSW}(\mu_n, \nu_n)$ (Lin et al., 2021; Niles-Weed & Rigollet, 2022) (or between $\mathrm{DSW}(\mu, \nu)$ and $\mathrm{DSW}(\mu_n, \nu_n)$ (Nguyen et al., 2021)), there is no explicit theoretical argument on the error entailed by the learned distribution $\rho^\star_{\mathrm{maxSW}}(\mu_n, \nu_n)$ (or $\rho^\star_{\mathrm{DSW}}(\mu_n, \nu_n)$) considered on its own, outside the optimization procedure of max-SW (or DSW). Given new samples $\{x'_1, \dots, x'_n\}$ and $\{y'_1, \dots, y'_n\}$ from μ and ν, with empirical distributions μ'_n and ν'_n, there is no guarantee for $\mathrm{SW}_p^p(\mu'_n, \nu'_n; \rho^\star_{\mathrm{maxSW}}(\mu_n, \nu_n))$ to be high, or in other words, there is no argument ensuring the discriminative power of $\rho^\star_{\mathrm{maxSW}}(\mu_n, \nu_n)$. One way to palliate this lack of theory and to go one step further than the max-SW and DSW cases, is to derive general results relating $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ and $\mathrm{SW}_p^p(\mu, \nu; \rho)$, for families of distributions $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. This is what we bring in the present work in the form of a generalization bound rooted in the PAC-Bayesian theory.

2.2. PAC-Bayesian Theory

PAC-Bayesian theory aims at assessing the ability of learning algorithms to generalize to unseen data, by deriving *generalization bounds*. Let $\mathcal{X} \subset \mathbb{R}^q$, $q \in \mathbb{N}^*$, and $S_n \doteq \{z_i\}_{i=1}^n$ a dataset of i.i.d. samples from an unknown probability measure $\xi \in \mathcal{P}(\mathcal{X})$. Consider a learning algorithm whose outputs depend on the training data S_n and a vector of parameters $\omega \in \Omega$. The performance of such algorithm can be assessed via a *loss function* $\ell : \Omega \times \mathcal{X} \to \mathbb{R}_+$. Fix $\omega \in \Omega$. The *empirical ℓ-risk* $\hat{r}_\ell(\omega, S_n)$ and *true ℓ-risk* $r_\ell(\omega)$ are defined as,

$$\hat{r}_\ell(\omega, S_n) \doteq \frac{1}{n} \sum_{i=1}^{n} \ell(\omega, z_i) \qquad (8)$$

$$r_\ell(\omega) \doteq \mathbb{E}_{z \sim \xi}[\ell(\omega, z)] \qquad (9)$$

A key objective of a learning procedure is to optimize (e.g. minimize) the true risk (9), which in practice cannot be achieved, because ξ is unknown. Instead, one focuses on optimizing (8) over $\omega \in \Omega$, a sound strategy provided the

minimizer of (8) accurately estimates the minimizer of (9): this can be assessed via PAC-Bayesian bounds.

Let $\rho \in \mathcal{P}(\Omega)$. PAC-Bayesian theory analyzes the generalization ability of ρ by measuring the gap between the *average empirical ℓ-risk* $\mathbb{E}_{\omega\sim\rho}[\hat{r}_\ell(\omega, S_n)]$ and the *average true ℓ-risk* $\mathbb{E}_{\omega\sim\rho}[r_\ell(\omega)]$. A classical PAC-Bayesian bound was derived by (Catoni, 2003) and is recalled below.

Theorem 1 ((Catoni, 2003))**.** *Let $\rho_0 \in \mathcal{P}(\Omega)$ be a prior distribution. Assume that $0 \leq \ell \leq C$. For all $\lambda > 0$, for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$ (over the draw of the dataset S_n): $\forall \rho \in \mathcal{P}(\Omega)$,*

$$\mathbb{E}_{\omega\sim\rho}[r_\ell(\omega)] \leq \mathbb{E}_{\omega\sim\rho}[\hat{r}_\ell(\omega, S_n)] + \frac{\lambda C^2}{8n} + \frac{1}{\lambda}\left\{KL(\rho||\rho_0) + \log\frac{1}{\delta}\right\}, \tag{10}$$

where $KL(\rho||\rho_0)$ is the Kullback-Leibler divergence *between ρ and ρ_0 : if ρ is absolutely continuous with respect to ρ_0, $KL(\rho||\rho_0) \doteq \int \log\left(\rho(\mathrm{d}\theta)/\rho_0(\mathrm{d}\theta)\right)\rho(\mathrm{d}\theta)$.*

The literature on PAC-Bayes is rich of many other bounds, and we refer to (Alquier, 2021) for an extensive survey. In our work, we focus on Catoni's bound because it is generic (appropriate settings of λ give rise to other well-known bounds) as are the proof techniques used to derive it (Alquier, 2021, Section 2).

Applications. PAC-Bayesian bounds allow to control the true risk via a function depending on the empirical risk. For example, minimizing the left-hand side term of Catoni's bound (10) over $\rho \in \mathcal{P}(\Omega)$ yields a data-dependent distribution which guarantees the highest generalization ability (Alquier, 2021, Section 2.1.2). PAC-Bayesian theory was also applied for specific tasks, e.g. classification (McAllester, 1999), ranking (Ralaivola et al., 2010), density estimation (Higgs & Shawe-Taylor, 2010), deep learning (Dziugaite & Roy, 2017; Chérief-Abdellatif et al., 2022).

3. Generalization Bounds for Adaptive Sliced-Wasserstein Distances

In this section, we leverage the PAC-Bayesian framework to derive generalization bounds for adaptive Sliced-Wasserstein distances. Proofs are deferred to Appendix A2.

Before presenting our main results, we clarify the notion of generalization for adaptive SW. In practice, since one generally has access to data generated from unknown probability measures μ, ν, empirical estimates $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ are computed instead of $\mathrm{SW}_p^p(\mu, \nu; \rho)$. Besides, adaptive SW means that an algorithm is deployed to learn ρ from μ_n, ν_n, so that $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ is sufficiently discriminative (Section 2.1). In this context, the learning algorithm is said to generalize well if the distribution learned from μ_n, ν_n (denoted by $\rho(\mu_n, \nu_n)$) is such that $\mathrm{SW}_p^p(\cdot, \cdot; \rho(\mu_n, \nu_n))$ is discriminative, even on unseen data. More formally, given new samples $\{x'_1, \ldots, x'_n\}$ and $\{y'_1, \ldots, y'_n\}$ from μ and ν, with associated empirical distributions μ'_n and ν'_n, $\mathrm{SW}_p^p(\mu'_n, \nu'_n; \rho(\mu_n, \nu_n))$ should be large.

PAC-Bayes framework	Our framework
$\{z_i\}_{i=1}^n$	$\{(x_i, y_i)\}_{i=1}^n$
$\xi \in \mathcal{P}(\mathcal{X})$	$\mu \times \nu \in \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$
$\omega \in \Omega$	$\theta \in \mathbb{S}^{d-1}$
$\hat{r}_\ell(\omega, \{z_i\}_{i=1}^n)$	$\mathrm{W}_p^p(\theta_\sharp^* \mu_n, \theta_\sharp^* \nu_n)$
$\mathbb{E}_{\omega\sim\rho}[\hat{r}_\ell(\omega, \{z_i\}_{i=1}^n)]$	$\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$
$r_\ell(\omega)$	$\mathbb{E}_{(x_i,y_i)_{i=1}^n}[\mathrm{W}_p^p(\theta_\sharp^* \mu_n, \theta_\sharp^* \nu_n)]$
$\mathbb{E}_{\omega\sim\rho}[r_\ell(\omega)]$	$\mathbb{E}_{(x_i,y_i)_{i=1}^n}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]$

Table 1. Analogy between PAC-Bayes theory and our work.

Therefore, we measure generalization as the gap between $\mathrm{SW}_p^p(\mu, \nu; \rho)$ and $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ for any $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. We first derive a general bound on this gap, using PAC-Bayesian theory, then refine it to specific settings directed by conditions on the supports and the moments of μ and ν.

3.1. A Generic Generalization Bound

We establish a first generalization bound for adaptive SW, by combining statistical properties of adaptive SW and techniques from PAC-Bayesian theory.

Theorem 2. *Let $p \in [1, +\infty)$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Assume there exists a constant $\varphi_{\mu,\nu,p}$, possibly depending on μ, ν and p such that: $\forall \lambda > 0$, $\forall \theta \in \mathbb{S}^{d-1}$,*

$$\mathbb{E}\left[\exp\left(\lambda\left\{\mathrm{W}_p^p(\theta_\sharp^* \mu_n, \theta_\sharp^* \nu_n) - \mathbb{E}[\mathrm{W}_p^p(\theta_\sharp^* \mu_n, \theta_\sharp^* \nu_n)]\right\}\right)\right] \leq \exp(\lambda^2 \varphi_{\mu,\nu,p}\, n^{-1}), \tag{11}$$

where $\mathbb{E}$ is taken with respect to the support points of μ_n and ν_n. Additionally, assume there exists $\psi_{\mu,\nu,p} : \mathbb{N}^* \to \mathbb{R}_+$, possibly depending on μ, ν and p, such that, $\forall \rho \in \mathcal{P}(\mathbb{S}^{d-1})$,

$$\mathbb{E}\left|\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathrm{SW}_p^p(\mu, \nu; \rho)\right| \leq \psi_{\mu,\nu,p}(n).$$

Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$. Then, for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$: $\forall \rho \in \mathcal{P}(\mathbb{S}^{d-1})$,

$$\mathrm{SW}_p^p(\mu, \nu; \rho) \geq \mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \frac{\lambda}{n}\varphi_{\mu,\nu,p} - \frac{1}{\lambda}\left\{KL(\rho||\rho_0) + \log\left(\frac{1}{\delta}\right)\right\} - \psi_{\mu,\nu,p}(n). \tag{12}$$

Link with PAC-Bayesian theory. Theorem 2 can be interpreted as a novel PAC-Bayesian bound tailored to adaptive SW: the formal analogy between classical PAC-Bayesian framework and our work is summarized in Table 1. The key element is that $\mathrm{W}_p^p(\theta_\sharp^* \mu_n, \theta_\sharp^* \nu_n)$ for some $\theta \in \mathbb{S}^{d-1}$

	Bounded supports	Unbounded supports	
		Sub-Gaussianity (Def. 3)	Bernstein moments (Def. 4)
$\varphi_{\mu,\nu,p}$	Proposition 1	Proposition 3	Proposition 4
$\psi_{\mu,\nu,p}$	Proposition 2	(Manole et al., 2022)	

Table 2. Overview of the explicit forms of $\varphi_{\mu,\nu,p}$ and $\psi_{\mu,\nu,p}$ under different assumptions.

can be seen as an empirical risk (8), and consequently, the average empirical risk is exactly $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ (2). Nevertheless, we emphasize that Theorem 2 is not obtained by simply replacing the risks in Theorem 1 according to Table 1. Indeed, this would return an *upper* bound (in terms of $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$) for $\mathbb{E}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]$, while we propose a *lower* bound for $\mathrm{SW}_p^p(\mu, \nu; \rho)$ (12). Instead, we propose the following slight paradigm shift: while classical PAC-Bayesian theory aims at *minimizing* the average true risk (hence, the *upper* bounds), our goal is to *maximize* $\mathrm{SW}_p^p(\mu, \nu; \rho)$ over ρ (hence, the need of *lower* bounds). Therefore, Theorem 2 is established by first, adapting the elements of proof of Theorem 1 to establish a lower-bound for $\mathbb{E}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]$, then bounding from above $\mathbb{E}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]$ by $\mathrm{SW}_p^p(\mu, \nu; \rho)$.

Discussion. Since our bound holds for all $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ (12), it is therefore valid for $\rho^\star_{\mathrm{maxSW}}$ (5) and $\rho^\star_{\mathrm{DSW}}$ (7) computed by max-SW and DSW. However, our bound is vacuous for max-SW, because the KL penalty term is evaluated on a Dirac distribution. In that singular case, more informative generalization bounds can be deduced, using (Lin et al., 2021; Niles-Weed & Rigollet, 2022) instead of PAC-Bayes: we elaborate on this in Appendix A2.1.

We now clarify the role of each term involved in (12). The KL divergence and λ arise from adapting the proof techniques of Catoni's bound, so their influence on the generalization gap can be further illustrated with the examples in (Alquier, 2021, Section 2.1.3). More precisely, the KL divergence results from a change of measure inequality known as Donsker-Varadhan's lemma (Donsker & Varadhan, 1975). Previous work have applied other change of measure inequalities to derive PAC-Bayesian bounds in terms of other divergences than KL (Alquier & Guedj, 2018). Nevertheless, standard PAC-Bayesian bounds rely on the use of Donsker-Varadhan's lemma, hence the KL divergence. As we introduce the first connection between PAC-Bayesian theory and SW, we decided to use the most common technique.

Then, the quantities $\varphi_{\mu,\nu,p}$ and $\psi_{\mu,\nu,p}(n)$ capture the properties of SW, μ and ν. More precisely, $\varphi_{\mu,\nu,p}$ bounds the moment-generating function of a centered version of $\mathrm{W}_p^p(\theta^*_\sharp \mu_n, \theta^*_\sharp \nu_n)$ for any $\theta \in \mathbb{S}^{d-1}$, while $\psi_{\mu,\nu,p}$ reflects the *sample complexity* of $\mathrm{SW}_p^p(\cdot, \cdot; \rho)$ for any $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. To further illustrate this, we specialize our generic bound (12) under different settings: μ, ν have bounded supports, are sub-Gaussian, or satisfy a Bernstein-type moment condition. We present our results in the next sections and summarize them in Table 2.

3.2. Application to Measures with Bounded Support

We first consider distributions supported on a bounded domain. We derive $\varphi_{\mu,\nu,p}$ by applying similar arguments as in the proof of McDiarmid's inequality (McDiarmid, 1989), similarly to (Weed & Bach, 2019, Proposition 20).

Proposition 1. *Let $\mathsf{X} \subset \mathbb{R}^d$ such that X has a finite diameter Δ, i.e. $\Delta \doteq \sup_{(x,x') \in \mathsf{X}^2} \|x - x'\| < +\infty$. Let $p \in [1, +\infty)$, $\mu, \nu \in \mathcal{P}(\mathsf{X})$. Then, $\mu, \nu \in \mathcal{P}_p(\mathsf{X})$ and $\varphi_{\mu,\nu,p} = \Delta^{2p}/2$.*

Next, we adapt the proof of (Manole et al., 2022, Lemma B.3) to compute the explicit form of $\psi_{\mu,\nu,p}$ in this setting.

Proposition 2. *Let $\mu, \nu \in \mathcal{P}(\mathsf{X})$, where $\mathsf{X} \subset \mathbb{R}^d$ has a finite diameter Δ. Let $p \in [1, +\infty)$. Then, there exists a constant C such that, $\psi_{\mu,\nu,p}(n) = Cp\Delta^p n^{-1/2}$.*

By combining Propositions 1 and 2, we refine Theorem 2 to distributions supported on a bounded domain: the resulting bound is given in Appendix A2.4.

3.3. Application to Sub-Gaussian Measures

Next, we apply Theorem 2 to distributions with unbounded supports. To handle this case, we assume specific constraints on the moments on μ, ν, then derive $\varphi_{\mu,\nu,p}$ by using generalizations of McDiarmid's inequalities. More precisely, we assume that μ, ν are sub-Gaussian distributions.

Definition 3 (Sub-Gaussian distribution)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\sigma > 0$. μ is a* sub-Gaussian distribution *with* variance proxy σ^2 *if the following holds: for any $\theta \in \mathbb{S}^{d-1}$, for $\lambda \in \mathbb{R}$, $\int_{\mathbb{R}} \exp(\lambda t) \mathrm{d}(\theta^*_\sharp \mu)(t) \leq \exp(\lambda^2 \sigma^2 / 2)$.*

The next proposition results from applying the generalized McDiarmid's inequality for unbounded spaces with finite *sub-Gaussian diameter* (Kontorovich, 2014) (Appendix A2.5).

Proposition 3. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ such that μ, ν are sub-Gaussian with variance proxy σ^2, τ^2 respectively. Then, $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$ and $\varphi_{\mu,\nu,1} = \sigma^2 + \tau^2$.*

The last ingredient to specialize Theorem 2 is to derive $\psi_{\mu,\nu,p}$ for μ, ν satisfying either Definition 3. To this end, we

leverage the rate recently established in (Manole et al., 2022, Theorem 2), which shows that $\psi_{\mu,\nu,p}$ scales as $n^{-1/2}\log(n)$ if μ,ν are sub-Gaussian distributions. Our final bound is obtained by plugging Proposition 3 and the explicit formula of $\psi_{\mu,\nu,1}$ in Theorem 2. We present this result and its detailed proof in Appendix A2.7.

3.4. Bound for Measures with Bernstein moment conditions

We study a more general class of distributions: we consider the *Bernstein-type moment condition* below, which is milder than sub-Gaussian distributions.

Definition 4 (Bernstein condition)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\sigma^2, b > 0$. μ is said to satisfy the* (σ^2, b)-Bernstein condition *if for any $k \in \mathbb{N}$, $k \geq 2$, for any $\theta \in \mathbb{S}^{d-1}$, $\int_{\mathbb{R}} |t|^k \mathrm{d}(\theta_\sharp^* \mu)(t) \leq \sigma^2 k! b^{k-2}/2$.*

Definition 3 is strictly stronger than Definition 4: if $\mu \in \mathcal{P}(\mathbb{R}^d)$ verifies the (σ^2, b)-Bernstein condition, then μ belongs to the class of heavy-tailed distributions called *sub-exponential distributions* (Embrechts et al., 2013), which contains sub-Gaussian distributions. Hence, the class of sub-Gaussian distributions is smaller than the class of distributions characterized by Definition 4.

Consider μ,ν satisfying Definition 4. First, we leverage (Manole et al., 2022, Theorem 2) in that setting again, to show that $\psi_{\mu,\nu,p}$ scales as $n^{-1/2}\log(n)$ (Appendix A2.7). Then, we apply the Bernstein-type McDiarmid's inequality given in (Lei, 2020, Theorem 5.1) to establish Proposition 4.

Proposition 4. *Let $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ be two distributions satisfying the Bernstein condition with parameters (σ^2, b) and (τ^2, c) respectively. Let $\sigma_\star^2 = \max(\sigma^2, \tau^2)$, $b_\star = \max(b, c)$. Then, $\mu,\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and, for any $\lambda \in \mathbb{R}_+^*$ s.t. $\lambda < n/(2b_\star)$,*

$$\mathbb{E}\left[\exp\left(\lambda\left\{\mathrm{W}_1(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n) - \mathbb{E}[\mathrm{W}_1(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n)]\right\}\right)\right] \leq \exp(\lambda^2 \varphi_{\mu,\nu,1}(\lambda, n) n^{-1}) \,, \tag{13}$$

where $\varphi_{\mu,\nu,1}(\lambda, n) = 2\sigma_\star^2 n^{-1}(1 - 2b_\star \lambda n^{-1})^{-1}$.

We emphasize the following difference between equations (13) and (11): $\varphi_{\mu,\nu,1}$ is a function of $\lambda \in \Lambda \subset \mathbb{R}_+$ and $n \in \mathbb{N}^*$, while in Theorem 2, $\varphi_{\mu,\nu,p}$ is assumed to be a constant. Nevertheless, the proof of Theorem 2 can easily be adapted to derive a generic generalization bound assuming $\varphi_{\mu,\nu,p}$ depends on (λ, n): we give the corresponding statement in Theorem A3. Hence, by plugging Proposition 4 and (Manole et al., 2022, Theorem 2) in Theorem A3, we derive the generalization bound for distributions under the Bernstein moment condition: see Appendix A2.7.

Note that for μ,ν satisfying Definition 3 or Definition 4, we derived $\varphi_{\mu,\nu,p}$ for $p = 1$ only: the generalized McDiarmid's inequalities used in the proofs of Propositions 3 and 4 can be applied if W_p^p is Lipschitz (Kontorovich, 2014; Lei, 2020). This property is easily verified for $p = 1$, but not for $p > 1$. Hence, the derivation of $\varphi_{\mu,\nu,p}$ for $p > 1$ for such types of distributions with unbounded domains requires different proof techniques. We leave this problem for future work.

Algorithm 1 PAC-SW: Adaptive SW via PAC-Bayes bound optimization.

Input: dataset $\{(x_i, y_i)\}_{i=1}^n$, parameter λ, prior ρ_0, initialization $\rho^{(0)}$, number of iterations T, learning rate η
for $t \leftarrow 1$ to T **do**
  $\mathcal{L}(\rho^{(t-1)}) = \mathrm{SW}_p^p(\mu_n, \nu_n; \rho^{(t-1)}) - \mathrm{KL}(\rho^{(t-1)} || \rho_0)/\lambda$
  $\rho^{(t)} = \rho^{(t-1)} + \eta \nabla_\rho \mathcal{L}(\rho^{(t-1)})$
end for
return $\rho^{(T)}$

4. Optimization of Generalization Bounds for Adaptive SW

We develop a principled methodology to learn a highly discriminative Sliced-Wasserstein distance, by optimizing our PAC-Bayesian generalization bounds. The idea consists in making the lower bounds of $\mathrm{SW}_p^p(\mu,\nu;\rho)$ derived in Section 3 as tight as possible, in order to increase $\mathrm{SW}_p^p(\mu,\nu;\rho)$ while attaining a small generalization gap.

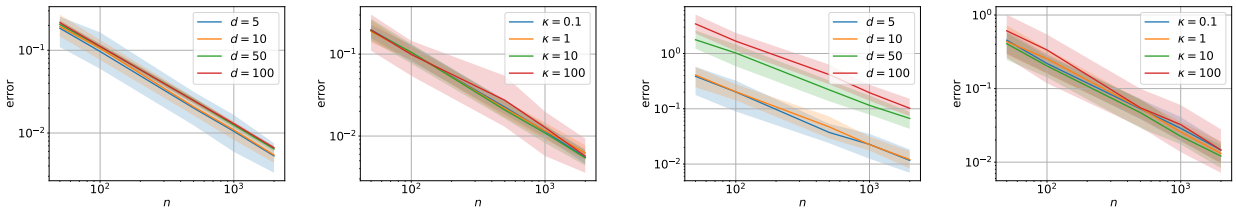
Given a training dataset $\{(x_i, y_i)\}_{i=1}^n$ and a prior $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$, our objective is to find $\rho^\star(\mu_n, \nu_n)$ such that,

$$\rho^\star(\mu_n, \nu_n) = \arg\sup_{\rho \in \mathcal{F}} \mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \frac{\mathrm{KL}(\rho || \rho_0)}{\lambda} \tag{14}$$

with $\mathcal{F}$ a family of probability measures supported on $\mathbb{S}^{d-1}$. The choice of $\mathcal{F}$ manages the complexity of solving (14): it should allow simple optimization, while being flexible to make $\rho^\star(\mu_n, \nu_n)$ expressive enough. We first propose to parameterize $\mathcal{F}$ as the class of *von Mises-Fisher distributions*.

Definition 5. *The* von Mises-Fisher distribution *vMF*(m, κ) *with* mean direction $\mathrm{m} \in \mathbb{S}^{d-1}$ *and* concentration parameter $\kappa \in \mathbb{R}_+^*$ *is a distribution on $\mathbb{S}^{d-1}$ whose density is defined for $\theta \in \mathbb{S}^{d-1}$ by vMF$(\theta; \mathrm{m}, \kappa) = C_{d/2}(\kappa)\exp(\kappa \mathrm{m}^\top \theta)$, $C_{d/2}(\kappa) = \kappa^{d/2-1}/\{(2\pi)^{d/2} I_{d/2-1}(\kappa)\}$, with $I_{d/2-1}$ the* modified Bessel function of the first kind at order $d/2 - 1$.

Intuitively, the higher κ, the more concentrated $\mathrm{vMF}(\mathrm{m}, \kappa)$ is around m. Our objective becomes finding $(\mathrm{m}^\star, \kappa^\star)$ such that $\mathrm{vMF}(\mathrm{m}^\star, \kappa^\star)$ maximizes (14) over $\mathcal{F} = \{\mathrm{vMF}(\mathrm{m}, \kappa), \mathrm{m} \in \mathbb{S}^{d-1}, \kappa \in \mathbb{R}_+^*\}$. Von Mises-Fisher distributions have been successfully deployed in several machine learning problems to effectively model spherical data (Hasnat et al., 2017; Kumar & Tsvetkov, 2018; Scott et al., 2021). Besides, one main advantage of using vMF is that both the KL divergence between $\rho = \mathrm{vMF}(\mathrm{m}, \kappa)$ and



(a) $\mu = \nu = \mathcal{U}([0,5]^d)$ (b) $\mu = \nu = \mathcal{N}(\mathbf{0}, \Sigma_d)$

Figure 1. $\mathrm{SW}_p^p(\mu_n, \nu_n; \mathrm{vMF}(\mathrm{m}, \kappa))$ vs. n. Results are averaged over 30 runs, on log-log scale, with 10th-90th percentiles.

$\rho_0 = \mathcal{U}(\mathbb{S}^{d-1})$ and its gradient with respect to (m, κ) admit an analytical formula (Davidson et al., 2018).

While the vMF parameterization is practical, as it yields an analytical objective, it may suffer from a lack of expressivity (e.g., vMF distributions are unimodal). To handle more complicated data, we also consider the parameterization proposed in (Nguyen et al., 2021): we solve (14) over $\mathcal{F} = \{\rho = f_\sharp \mathcal{U}(\mathbb{S}^{d-1}), f \text{ a neural network}\}$. In that case, the KL penalty term is intractable and we approximate it with the methodology in (Ghimire et al., 2021) – where approximation errors of the KL estimator are given in different scenarios.

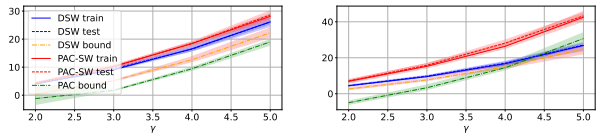
We approximate the solution of (14) via gradient ascent: our methodology is depicted in Algorithm 1, and specialized in Algorithm A2 for the vMF parameterization.

Tuning λ. In classical PAC-Bayesian theory, λ is usually set to $n^{1/2}$ so that all terms in the bound that depend on λ converge at the same rate to 0, as n grows to $+\infty$. Nevertheless, using $\lambda = n^\alpha$ with $\alpha \in (0,1)$, $\alpha \neq 1/2$ can be more useful in some specific settings. For instance, a common issue when optimizing PAC-Bayesian bounds is that the objective can be dominated by the KL term (Chérief-Abdellatif et al., 2022). To overcome this, one can downweight the KL term by using $\alpha > 1/2$, or more sophisticated schemes (Blundell et al., 2015). On the other hand, as shown in Section 3.2, 3.3 and 3.4, $\varphi_{\mu,\nu,p}$ depend on parameters related to the properties of μ, ν, which cannot be easily controlled in practice. Choosing $\lambda = n^\alpha$ with $\alpha < 1/2$ helps attenuate their influence on the objective (Haddouche et al., 2021).

5. Numerical Experiments

We conduct an empirical analysis to confirm our theoretical contributions and illustrate their consequences in practice, on both synthetic and real data. More details on our experimental setup are given in Appendix A3, and the code is available at https://github.com/rubenohana/PAC-Bayesian_Sliced-Wasserstein..

Illustration of our bounds. Our first set of experiments aims at empirically validating the rates of convergence in Section 3. We sample two sets of n i.i.d. samples from the same distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$. To illustrate our bound on both bounded and unbounded supports, we choose μ as a uniform or Gaussian distribution. We approximate $\mathrm{SW}_p^p(\mu_n, \nu_n; \mathrm{vMF}(\mathrm{m}, \kappa))$ with $\mathrm{m} \sim \mathcal{U}(\mathbb{S}^{d-1})$ and $\kappa > 0$ by its Monte Carlo estimate (3) over 1000 projection directions. Figure 1 plots the approximation error (which reduces to $\mathrm{SW}_p^p(\mu_n, \nu_n; \mathrm{vMF}(\mathrm{m}, \kappa))$, since the two datasets come from the same distribution) against n, for different d and κ. We observe that the error decays to 0 as n increases, and the convergence rate is slower as d and κ increase. This confirms our theoretical analysis: the higher d, the larger the diameter (resp., the sub-Gaussian diameter) when μ is uniform (resp., Gaussian), the larger $\varphi_{\mu,\nu,p}$ (Propositions 1 and 3). Besides, the higher κ, the larger $\mathrm{KL}(\mathrm{vMF}(\mathrm{m}, \kappa) || \mathcal{U}(\mathbb{S}^{d-1}))$.



(a) $d = 5$ (b) $d = 20$

Figure 2. PAC-SW and DSW between $\mu = \mathcal{N}(\mathbf{0}, \Sigma_d)$ and $\nu = \mathcal{N}(\gamma \mathbf{1}, \Sigma_d)$. The y-axis shows the distances or the associated objective functions (see legend). Results are averaged over 10 runs, and shown with 10th-90th percentiles.

Generalization ability of PAC-SW. Next, we study the generalization properties of PAC-SW, *i.e.* whether the adaptive SW computed by Algorithm 1 is discriminative, even on unseen data. We compare $\mu = \mathcal{N}(\mathbf{0}, \Sigma_d)$ and $\nu = \mathcal{N}(\gamma \mathbf{1}, \Sigma_d)$, with $\gamma > 0$, $\Sigma_d \in \mathbb{R}^{d \times d}$ symmetric positive semi-definite set at random, and $\mathbf{0}$ (resp., $\mathbf{1}$) the vector whose components are all equal to 0 (resp., 1). The higher γ, the more dissimilar μ and ν. We sample $n = 500$ samples from μ and ν and optimize $\rho^\star(\mu_n, \nu_n)$: the optimization is performed on the space of vMF distributions, using Adam (Kingma & Ba, 2015) with its default parameters. To analyze the generalization properties of $\rho^\star(\mu_n, \nu_n)$, we sample $l = 2000$ test points from μ, ν and compute $\mathrm{SW}_p^p(\mu_l, \nu_l; \rho^\star(\mu_n, \nu_n))$. We also compute the value of (14), to evaluate the tightness of our bound. Results for different values of d and γ are reported in Figure 2, and confirm the generalization ability of $\rho^\star(\mu_n, \nu_n)$.

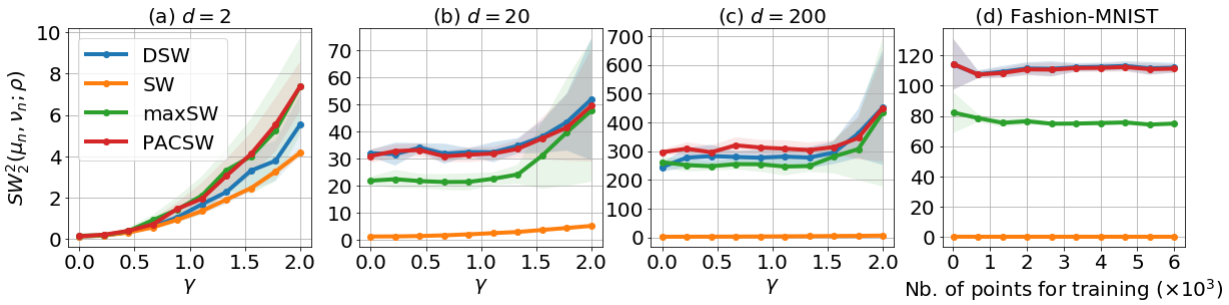


Figure 3. $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)$ with (a-c) $\mu = \mathcal{N}(\mathbf{0}, \Sigma_d)$, $\nu = \mathcal{N}(\gamma \mathbf{1}, \Sigma_d)$, $n = 1000$, against γ, (d) classes 4 and 5 of Fashion-MNIST, against n. ρ is learned on the train set, and we report values on the test set.

Comparison to existing instances of SW. In our previous experiment, we also implement a variant of DSW, which consists in solving (Nguyen et al., 2021, Definition 2) based on our vMF parameterization. Figure 2 shows that the gap between $\mathrm{SW}_p^p(\mu_n, \nu_n; \rho^\star_{\mathrm{DSW}}(\mu_n, \nu_n))$ and $\mathrm{SW}_p^p(\mu_m, \nu_m; \rho^\star_{\mathrm{DSW}}(\mu_n, \nu_n))$ is small, hence $\rho^\star_{\mathrm{DSW}}(\mu_n, \nu_n)$ generalizes well on that setup. DSW bound in Figure 2 corresponds to the associated objective function of (Nguyen et al., 2021, Definition 2).

Next, we compare the generalization properties of PAC-SW and DSW, with ρ parameterized as a neural network. We also evaluate max-SW and SW (*i.e.*, $\mathrm{SW}_p^p(\cdot, \cdot; \mathcal{U}(\mathbb{S}^{d-1}))$. We compute the Monte Carlo estimate with $m = 200$ and the learning rate η is taken as the best (*i.e.*, yielding the higher distance) out of $[10^{-3}, 10^{-2}, 10^{-1}, 1]$. Each run is averaged 10 times with standard deviations in shaded areas. On Figure 3(a-c), we measure the distance between two Gaussians, as in Figure 2. We observe that PAC-SW is always amongst the most discriminative distances, and since we evaluate distances on unseen data, this implies it has better generalization properties. On Figure 3(d), we consider a more complicated dataset: we measure the distance between 2 highly dissimilar classes of the Fashion-MNIST dataset (Xiao et al., 2017) (classes 4 (*coats*) and 5 (*sandals*)) for different number of training points. PAC-SW and DSW return higher values than max-SW and SW, illustrating they are able to better discriminate, even on test data.

Note that max-SW and DSW share a common feature: they learn a new distribution $\rho(\mu_n, \nu_n)$ every time they are called on new (μ_n, ν_n), *i.e.* they embed an optimization step. From here on, when we will refer to the generalization ability of max-DSW and DSW, it must be understood that a distribution $\rho^\star$ is learned from one sample pair (μ_n, ν_n) according to their respective induction principle, and $\rho^\star$ is used on test data to measure the generalization ability.

Generalization for generative modeling. In our previous experiments, we observed that DSW can generalize as well as PAC-SW. This encourages us to further explore the advantages of a high generalization ability on a more complicated setup. We consider a generative modeling task on MNIST data (Deng, 2012), and we train a deep neural network that

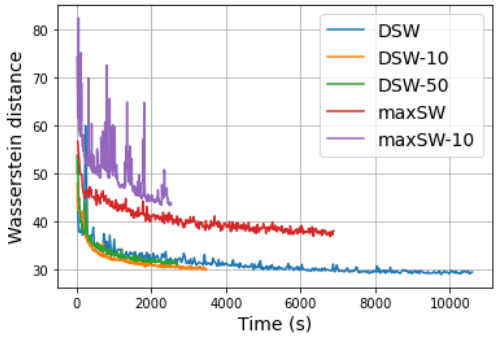


Figure 4. Evolution of the Wasserstein distance between a set of generated MNIST digits and the true MNIST test set with respect to training time.

uses DSW as a loss, in the flavor of (Deshpande et al., 2018; Nguyen et al., 2021). Usually, the distribution ρ is learned at each iteration, when the user receive new data. We conjecture that if the learned distribution generalizes well to unseen datasets, then gradients obtained from the distance between minibatches would still provide sufficient information to learn the generative model. As a consequence, we evaluate the robustness and generalization ability of the learned distribution using DSW updated only every 10 or 50 minibatches (denoted by -10 or -50 resp.). To train the model, we followed the same approach (architecture and optimizer) as the one described in (Nguyen et al., 2021). For each minibatch of size 512, the distribution ρ is learned by optimizing 100 projections over 100 iterations and the generative model is trained over 400 epochs. We also report results of a generative model trained with max-SW.

Figure 4 shows the evolution of the Wasserstein distance (WD) between generated data and the test set with respect to training time (measured after each epoch), for each distance and different update rate of the distribution ρ. We can observe that classical DSW yields a WD of 29 after $\sim 10^4$s. When learning ρ every 10 minibatch (DSW-10), we achieve similar a WD value with half the running time. When further reducing the frequency update of ρ (DSW-50), we converge faster but with a loss in quality of generation (WD ~ 32). While using max-SW as a loss yields a reasonable performance, computing $\rho^\star_{\mathrm{maxSW}}$ every 10 minibatches

leads to a very unstable learning and worst performances. Results for the PAC-SW loss and examples of generated digits can be found in Appendix A3.2.

6. Conclusion

We introduced a specific notion of generalization for adaptive SW, related to discriminative power, and leveraged the PAC-Bayesian framework to derive generalization bounds. We then developed a principled methodology to learn ρ from the observed data so as $\mathrm{SW}_p^p(\cdot, \cdot; \rho)$ is discriminative with high probability, thus, generalizes well. Our work, which presents the first connection between PAC-Bayes and SW, paves the way to interesting research directions. First, we will study possible refinements of our bounds, using other PAC-Bayes bounds than Catoni's. Then, we plan to further analyze why DSW generalizes well in our experiments, e.g. by investigating a potential connection between the optimization problem in (Nguyen et al., 2021) and ours. Finally, we would like to reduce the computational complexity of PAC-SW when ρ is parameterized as a neural network, since it suffers from slow execution times mainly because of the approximation of the KL term with (Ghimire et al., 2021).

References

Alquier, P. User-friendly introduction to PAC-Bayes bounds, 2021.

Alquier, P. and Guedj, B. Simpler pac-bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Ambroladze, A., Parrado-hernández, E., and Shawe-taylor, J. Tighter pac-bayes bounds. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bayraktar, E. and Guo, G. Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26(none):1 – 13, 2021. doi: 10.1214/21-ECP383.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1613–1622. JMLR.org, 2015.

Bonet, C., Courty, N., Septier, F., and Drumetz, L. Sliced-wasserstein gradient flows. *arXiv preprint arXiv:2110.10972*, 2021.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Bonnotte, N. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Paris 11, 2013.

Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J., and Schoelkopf, B. From optimal transport to generative modeling: the vegan cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

Carriere, M., Cuturi, M., and Oudot, S. Sliced wasserstein kernel for persistence diagrams. In *International conference on machine learning*, pp. 664–673. PMLR, 2017.

Catoni, O. A PAC-Bayesian approach to adaptive classification. preprint LPMA 840, 2003.

Catoni, O. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. Institute of Mathematical Statistics, 2007.

Chérief-Abdellatif, B.-E., Shi, Y., Doucet, A., and Guedj, B. On pac-bayesian reconstruction guarantees for vaes. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3066–3079. PMLR, 28–30 Mar 2022.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Cuturi, M. and Peyré, G. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1): 320–343, 2016.

Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. Hyperspherical variational auto-encoders. *34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Deshpande, I., Zhang, Z., and Schwing, A. G. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Elidan, G., Kersting, K., and Ihler, A. T. (eds.), *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017.

Embrechts, P., Klüppelberg, C., and Mikosch, T. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

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, 2015.

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a wasserstein loss. *arXiv preprint arXiv:1506.05439*, 2015.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. Pac-bayesian learning of linear classifiers. In *Proc. of the 26th International Conference on Machine Learning (ICML)*, pp. 353–360, 2009.

Ghimire, S., Masoomi, A., and Dy, J. Reliable estimation of kl divergence using a discriminator in reproducing kernel hilbert space. *Advances in Neural Information Processing Systems*, 34, 2021.

Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101330.

Hasnat, M., Bohné, J., Milgram, J., Gentric, S., Chen, L., et al. von mises-fisher mixture model-based deep learning: Application to face verification. *arXiv preprint arXiv:1706.04264*, 2017.

Higgs, M. and Shawe-Taylor, J. A pac-bayes bound for tailored density estimation. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory*, ALT'10, pp. 148–162, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3642161073.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kolouri, S., Zou, Y., and Rohde, G. K. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016.

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, 2017.

Kolouri, S., Rohde, G. K., and Hoffmann, H. Sliced wasserstein distance for learning gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436, 2018.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. K. Generalized sliced wasserstein distances. *arXiv preprint arXiv:1902.00434*, 2019a.

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

Kolouri, S., Nadjahi, K., Simsekli, U., and Shahrampour, S. Generalized sliced distances for probability distributions. *arXiv preprint arXiv:2002.12537*, 2020.

Kontorovich, A. Concentration in unbounded metric spaces and algorithmic stability. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 28–36, Bejing, China, 22–24 Jun 2014. PMLR.

Kumar, S. and Tsvetkov, Y. Von mises-fisher loss for training sequence to sequence models with continuous outputs. *arXiv preprint arXiv:1812.04616*, 2018.

Laviolette, F., Marchand, M., and Shah, M. A pac-bayes approach to the set covering machine. In Weiss, Y., Schölkopf, B., and Platt, J. (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2006.

Lei, J. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767 – 798, 2020. doi: 10.3150/19-BEJ1151.

Lin, T., Zheng, Z., Chen, E. Y., Cuturi, M., and Jordan, M. I. On projection robust optimal transport: Sample complexity and model misspecification. In *AISTATS*, pp. 262–270, 2021.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pp. 4104–4113. PMLR, 2019.

Manole, T., Balakrishnan, S., and Wasserman, L. Minimax confidence intervals for the Sliced Wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252 – 2345, 2022. doi: 10.1214/22-EJS2001.

McAllester, D. A. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

McDiarmid, C. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Montavon, G., Müller, K.-R., and Cuturi, M. Wasserstein training of restricted boltzmann machines. *Advances in Neural Information Processing Systems*, 29:3718–3726, 2016.

Nadjahi, K., Durmus, A., Şimşekli, U., and Badeau, R. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *arXiv preprint arXiv:1906.04516*, 2019.

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R., and Şimşekli, U. Approximate bayesian computation with the sliced-wasserstein distance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5470–5474. IEEE, 2020a.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Şimşekli, U. Statistical and topological properties of sliced probability divergences. *arXiv preprint arXiv:2003.05783*, 2020b.

Nadjahi, K., Durmus, A., Jacob, P., Badeau, R., and Simsekli, U. Fast approximation of the sliced-wasserstein distance using concentration of random projections. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

Niles-Weed, J. and Rigollet, P. Estimation of Wasserstein distances in the Spiked Transport Model. *Bernoulli*, 28(4):2663 – 2688, 2022. doi: 10.3150/21-BEJ1433.

Oberman, A. M. and Ruan, Y. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

Peyré, G. and Cuturi, M. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein Barycenter and Its Application to Texture Mixing. In Bruckstein, A. M., ter Haar Romeny, B. M., Bronstein, A. M., and Bronstein, M. M. (eds.), *Scale Space and Variational Methods in Computer Vision*, pp. 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-24785-9.

Rakotomamonjy, A. and Ralaivola, L. Differentially private sliced wasserstein distance. In *International Conference on Machine Learning*, pp. 8810–8820. PMLR, 2021.

Ralaivola, L., Szafranski, M., and Stempfel, G. Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes. *Journal of Machine Learning Research*, 11(65):1927–1956, 2010.

Rivasplata, O. Subgaussian random variables : An expository note. 2012.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1): 643–678, 2018.

Schmitzer, B. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2): 238–259, 2016.

Scott, T. R., Gallagher, A. C., and Mozer, M. C. von mises–fisher loss: An exploration of embedding geometries for supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10592–10602, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.01044.

Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pp. 306–314. PMLR, 2014.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4 A):2620–2648, 2019. ISSN 1350-7265. doi: 10.3150/18-BEJ1065.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zantedeschi, V., Viallard, P., Morvant, E., Emonet, R., Habrard, A., Germain, P., and Guedj, B. Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound. In *NeurIPS*, Online, France, 2021.

A1. Preliminaries

A1.1. Metric Properties of Sliced-Wasserstein Distances

Previous work have shown that for specific instances of $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$, $\mathrm{SW}_p(\cdot,\cdot;\rho) : \mathcal{P}_p(\mathbb{R}^d) \times \mathcal{P}_p(\mathbb{R}^d) \to \mathbb{R}_+$ is a metric, as it satisfies all metric axioms (positivity, symmetry, triangle inequality, identity of indiscernibles) (Bonnotte, 2013; Kolouri et al., 2019a; Nguyen et al., 2021; Niles-Weed & Rigollet, 2022). However, to the best of our knowledge, the metric properties of $\mathrm{SW}_p(\cdot,\cdot;\rho)$ for *any* $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ have not been established.

By adapting the proof techniques in (Bonnotte, 2013; Kolouri et al., 2019a), and due to the metric properties of the Wasserstein distance, one can show that symmetry, positivity and triangle inequality hold for any $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$, and that for any $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, $\mathrm{SW}_p(\mu,\mu;\rho) = 0$.

However, the reverse implication of the identity of indiscernibles, *i.e.*

$$\forall \mu,\nu \in \mathcal{P}_p(\mathbb{R}^d),\ \mathrm{SW}_p(\mu,\nu;\rho) = 0 \ \text{ implies } \ \mu = \nu\,, \tag{A1}$$

does not hold for any $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. For example, consider $\mu,\nu \in \mathcal{P}_p(\mathsf{X})$ with $\mathsf{X} \subset \mathbb{R}^d$, and μ different from ν. Suppose that $\rho \in \mathcal{P}(\Theta)$ with $\Theta \subset \mathbb{S}^{d-1}$ such that $\forall(\theta,x) \in \Theta \times \mathsf{X}$, $\langle\theta,x\rangle = 0$. In that case, for any $\theta \sim \rho$, $\theta^*_\sharp\mu = \theta^*_\sharp\nu = \delta_{\{0\}}$, and since W_p is a metric, $\mathrm{W}_p(\theta^*_\sharp\mu,\theta^*_\sharp\nu) = 0$. Therefore, $\mathrm{SW}^p_p(\mu,\nu;\rho) = \int_\Theta \mathrm{W}^p_p(\theta^*_\sharp\mu,\theta^*_\sharp\nu)\mathrm{d}\rho(\theta) = 0$, but $\mu \neq \nu$, so (A1) is not satisfied.

We conclude that for any $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$, $\mathrm{SW}_p(\cdot,\cdot;\rho)$ is a *pseudo-metric*, and if (A1) is satisfied, then it is a metric.

A1.2. Generalization Bounds for SW

We precise our argument in Section 1, which states that bounds on the generalization gap for SW distances can be established using existing results for max-SW.

Let $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. By applying the triangle inequality for $\mathrm{SW}_p(\cdot,\cdot;\rho)$, then by the definition of max-SW, we obtain,

$$\mathbb{E}|\mathrm{SW}_p(\mu_n,\nu_n;\rho) - \mathrm{SW}_p(\mu,\nu;\rho)| \leq \mathbb{E}[\mathrm{SW}_p(\mu_n,\mu;\rho)] + \mathbb{E}[\mathrm{SW}_p(\nu_n,\nu;\rho)] \tag{A2}$$

$$\leq \mathbb{E}[\mathrm{maxSW}(\mu_n,\mu)] + \mathbb{E}[\mathrm{maxSW}(\nu_n,\nu)]\,, \tag{A3}$$

where $\mathbb{E}$ is taken with respect to $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ i.i.d. from μ,ν respectively. We can then bound from above (A3), using the convergence rates established in (Lin et al., 2021, Section 3.2) or (Niles-Weed & Rigollet, 2022, Theorem 1). These rates vary depending on the properties of μ,ν: for instance, (Lin et al., 2021, Theorem 3.5) holds if μ,ν satisfy the Bernstein condition.

Nevertheless, we argue that the generalization bounds resulting from eq.(A2)-(A3) are not tight for an arbitrary $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. For instance, since we bound (A3) with (Lin et al., 2021; Niles-Weed & Rigollet, 2022), we obtain convergence rates that linearly depend on d for any ρ, due to the properties of maximum SW. However, if we consider $\rho = \mathcal{U}(\mathbb{S}^{d-1})$, it is known that $\mathbb{E}|\mathrm{SW}_p(\mu_n,\nu_n;\rho) - \mathrm{SW}_p(\mu,\nu;\rho)|$ converges to 0 at a dimension-free rate (Nadjahi et al., 2020b).

Another important drawback of such bounds is that the impact of ρ on the convergence rates is unclear. In Appendix A2.1, we will further explain why our generalization bounds derived from PAC-Bayesian theory are more flexible and informative for arbitrary ρ.

A2. Postponed Proofs for Section 3

A2.1. Proof of Theorem 2

Theorem 2 is obtained by adapting the proof techniques of Catoni's PAC-Bayesian bound (Catoni, 2003). First, we recall Donsker and Varadhan's variational formula, which plays a central role in the PAC-Bayesian framework.

Lemma A1 (Donsker and Varadhan's variational formula (Donsker & Varadhan, 1975))**.** *Let Θ be a set equipped with a σ-algebra and $\pi \in \mathcal{P}(\Theta)$. For any measurable, bounded function $h : \Theta \to \mathbb{R}$,*

$$\log \mathbb{E}_{\theta\sim\pi}\big[\exp(h(\theta))\big] = \sup_{\rho\in\mathcal{P}(\Theta)} \big[\mathbb{E}_{\theta\sim\rho}[h(\theta)] - \mathit{KL}(\rho||\pi)\big]$$

Proof of Theorem 2. Let $p \in [1, +\infty)$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Assume there exists $\varphi_{\mu,\nu,p}$ such that for any $\theta \in \mathbb{S}^{d-1}$ and $\lambda > 0$,

$$\mathbb{E}_{\mu,\nu}\left[\exp\left(\lambda\Big\{\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n) - \mathbb{E}_{\mu,\nu}[\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n)]\Big\}\right)\right] \leq \exp(\lambda^2\varphi_{\mu,\nu,p}n^{-1}) \ . \tag{A4}$$

Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$. By taking the expectation of (A4) with respect to ρ_0, then using Fubini's theorem to interchange the expectation over ρ_0 and the one over μ, ν, we obtain

$$\mathbb{E}_{\mu,\nu}\mathbb{E}_{\theta\sim\rho_0}\left[\exp\left(\lambda\big\{\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n) - \mathbb{E}_{\mu,\nu}[\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n)]\big\}\right)\right] \leq \exp(\lambda^2\varphi_{\mu,\nu,p}n^{-1}) \ . \tag{A5}$$

By definition of the Wasserstein distance between empirical, univariate distributions of (1) , one can prove that $\theta \mapsto \lambda\big\{\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n) - \mathbb{E}_{\mu,\nu}[\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n)]\big\}$ is a bounded real-valued function on $\mathbb{S}^{d-1}$. Therefore, we can apply Lemma A1 to rewrite (A5) as follows.

$$\begin{aligned}&\mathbb{E}_{\mu,\nu}\Big[\exp\Big(\sup_{\rho\in\mathcal{P}(\mathbb{S}^{d-1})}\big[\mathbb{E}_{\theta\sim\rho}\big[\lambda\big\{\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n) - \mathbb{E}_{\mu,\nu}[\mathrm{W}_p^p(\theta_\sharp^*\mu_n, \theta_\sharp^*\nu_n)]\big\}\big] - \mathrm{KL}(\rho||\rho_0)\big]\Big)\Big] \\ &\quad\leq \exp(\lambda^2\varphi_{\mu,\nu,p}n^{-1}) \ ,\end{aligned}$$

which, using the linearity of the expectation along with the definition of SW (2) , is equivalent to

$$\mathbb{E}_{\mu,\nu}\Big[\exp\Big(\sup_{\rho\in\mathcal{P}(\mathbb{S}^{d-1})}\big[\lambda\big\{\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathbb{E}_{\mu,\nu}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]\big\} - \mathrm{KL}(\rho||\rho_0)\big]\Big)\Big] \leq \exp(\lambda^2\varphi_{\mu,\nu,p}n^{-1}) \ ,$$

or,

$$\mathbb{E}_{\mu,\nu}\Big[\exp\Big(\sup_{\rho\in\mathcal{P}(\mathbb{S}^{d-1})}\big[\lambda\big\{\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathbb{E}_{\mu,\nu}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]\big\} - \mathrm{KL}(\rho||\rho_0)\big] - \lambda^2\varphi_{\mu,\nu,p}n^{-1}\Big)\Big] \leq 1. \tag{A6}$$

Let $s > 0$. By the Chernoff bound $\big(\mathbb{P}(X > a) = \mathbb{P}(e^{s.X} \geq e^{s.a}) \leq \mathbb{E}[e^{t.X}]e^{-t.a}\big)$

$$\begin{aligned}&\mathbb{P}_{\mu,\nu}\Big(\sup_{\rho\in\mathcal{P}(\mathbb{S}^{d-1})}\big[\lambda\big\{\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathbb{E}_{\mu,\nu}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]\big\} - \mathrm{KL}(\rho||\rho_0)\big] - \lambda^2\varphi_{\mu,\nu,p}n^{-1} > s\Big) \\ &\leq \mathbb{E}_{\mu,\nu}\Big[\exp\Big(\sup_{\rho\in\mathcal{P}(\mathbb{S}^{d-1})}\big[\lambda\big\{\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathbb{E}_{\mu,\nu}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]\big\} - \mathrm{KL}(\rho||\rho_0)\big] - \lambda^2\varphi_{\mu,\nu,p}n^{-1}\Big)\Big]\exp(-s) \\ &\leq 1 \cdot \exp(-s) = \exp(-s) \ ,\end{aligned}$$

where the last inequality follows from (A6).

Let $e^{-s} = \varepsilon$ such that $s = \log(1/\varepsilon)$. Then,

$$\mathbb{P}_{\mu,\nu}\Big(\exists\rho \in \mathcal{P}(\mathbb{S}^{d-1}),\ \lambda\big\{\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathbb{E}_{\mu,\nu}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)]\big\} - \mathrm{KL}(\rho||\rho_0) - \lambda^2\varphi_{\mu,\nu,p}n^{-1} > \log(1/\varepsilon)\Big) \leq \varepsilon \ . \tag{A7}$$

Taking the complement of (A7) and rearranging the terms yields

$$\begin{aligned}&\mathbb{P}_{\mu,\nu}\Big(\forall\rho \in \mathcal{P}(\mathbb{S}^{d-1}),\ \mathrm{SW}_p^p(\mu_n, \nu_n; \rho) < \mathbb{E}_{\mu,\nu}[\mathrm{SW}_p^p(\mu_n, \nu_n; \rho)] + \lambda^{-1}\big\{\mathrm{KL}(\rho||\rho_0) + \log(1/\varepsilon)\big\} + \lambda\varphi_{\mu,\nu,p}n^{-1}\Big) \\ &\quad\geq 1 - \varepsilon \ .\end{aligned}$$

Our final bound results from assuming there exists $\psi_{\mu,\nu,p}(n)$ such that,

$$\mathbb{E}_{\mu,\nu}|\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathrm{SW}_p^p(\mu, \nu; \rho)| \leq \psi_{\mu,\nu,p}(n) \ .$$

□

Comparison with Appendix A1.2. In our work, instead of bounding $\mathrm{SW}_p^p(\cdot,\cdot;\rho)$ by maxSW, we apply PAC-Bayesian theory directly on $\mathrm{SW}_p^p(\cdot,\cdot;\rho)$ for any ρ. As a result, our PAC-Bayes-inspired bounds are more flexible than bounds in Appendix A1.2, since their convergence rates adapt to the distribution ρ (via the KL divergence). However, when ρ is a Dirac measure, Theorem 2 become vacuous because of the KL term, as with most PAC-Bayesian bounds. In such cases, which include maxSW, the bounds in Appendix A1.2 are more informative.

As discussed in Section 3.4, in specific settings, $\varphi_{\mu,\nu,p}$ can be a function of $\lambda \in \mathbb{R}_+$ and $n \in \mathbb{N}^*$. In that case, a straightforward adaptation of the proof of Theorem 2 yields Theorem A3, which will be leveraged for distributions with Bernstein-type moment conditions (Definition 4).

Theorem A3. *Let $p \in [1,+\infty)$ and $\mu,\nu \in \mathcal{P}_p(\mathbb{R}^d)$. Let $\Lambda \subset \mathbb{R}_+^*$ and assume there exists $\varphi_{\mu,\nu,p} : \Lambda \times n \to \mathbb{R}_+$, possibly depending on μ,ν and p such that: $\forall \lambda \in \Lambda$, $\forall \theta \in \mathbb{S}^{d-1}$,*

$$\mathbb{E}\left[\exp\left(\lambda\left\{\mathrm{W}_p^p(\theta_\sharp^*\mu_n,\theta_\sharp^*\nu_n) - \mathbb{E}[\mathrm{W}_p^p(\theta_\sharp^*\mu_n,\theta_\sharp^*\nu_n)]\right\}\right)\right] \leq \exp(\lambda^2\varphi_{\mu,\nu,p}(\lambda,n)\,n^{-1})\ ,$$

where $\mathbb{E}$ is taken with respect to the support points of μ_n and ν_n. Additionally, assume there exists $\psi_{\mu,\nu,p} : \mathbb{N}^* \to \mathbb{R}_+$, possibly depending on μ,ν and p, such that, $\forall \rho \in \mathcal{P}(\mathbb{S}^{d-1})$,

$$\mathbb{E}\big|\mathrm{SW}_p^p(\mu_n,\nu_n;\rho) - \mathrm{SW}_p^p(\mu,\nu;\rho)\big| \leq \psi_{\mu,\nu,p}(n)\,.$$

Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$. Then, for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$: $\forall \rho \in \mathcal{P}(\mathbb{S}^{d-1})$,

$$\begin{aligned}\mathrm{SW}_p^p(\mu,\nu;\rho) \geq\ & \mathrm{SW}_p^p(\mu_n,\nu_n;\rho) - \frac{\lambda}{n}\varphi_{\mu,\nu,p}(\lambda,n)\\ & - \frac{1}{\lambda}\Big\{KL(\rho||\rho_0) + \log\Big(\frac{1}{\delta}\Big)\Big\} - \psi_{\mu,\nu,p}(n)\,.\end{aligned}$$

A2.2. Proof of Proposition 1

To prove Proposition 1, we leverage a concentration result that appears in the proof of McDiarmid's inequality (recalled in Theorem A4), and which relies on the *bounded differences property* (Definition A6).

Definition A6 (Bounded differences property)**.** *Let $\mathsf{X} \subset \mathbb{R}^d$, $n \in \mathbb{N}^*$ and $c = \{c_i\}_{i=1}^n \in \mathbb{R}^n$. A mapping $f : \mathsf{X}^n \to \mathbb{R}$ is said to satisfy the c-*bounded differences property *if for $i \in \{1,\dots,n\}$, $\{x_i\}_{i=1}^n \in \mathsf{X}^n$ and $x' \in \mathsf{X}$,*

$$|f(x_1,\dots,x_n) - f(x_1,\dots,x_{i-1},x',x_{i+1},\dots,x_n)| \leq c_i\ .$$

Theorem A4 ((McDiarmid, 1989))**.** *Let $(X_i)_{i=1}^n$ be a sequence of $n \in \mathbb{N}^*$ independent random variables with X_i valued in $\mathsf{X} \subset \mathbb{R}^d$ for $i \in \{1,\dots,n\}$. Let $c = \{c_i\}_{i=1}^n \in \mathbb{R}^n$ and $f : \mathsf{X}^n \to \mathbb{R}$ satisfying the c-bounded differences property. Then, for any $\lambda > 0$,*

$$\mathbb{E}\big[\exp(\lambda\{f - \mathbb{E}[f]\})\big] \leq \exp(\lambda^2\|c\|^2/8)\ .$$

The proof of Proposition 1 consists in applying Theorem A4 to a specific choice of f. To this end, we first show that the Wasserstein distance between univariate distributions satisfies the bounded differences property, assuming bounded supports.

Lemma A2. *Let $\mathsf{X} \subset \mathbb{R}$ be a bounded set with diameter $\Delta = \sup_{(x,x')\in\mathsf{X}^2}\|x-x'\| < +\infty$. Then, the mapping $f : (\mathsf{X}^2)^n \to \mathbb{R}_+$ defined for $w_{1:n} \doteq \{(u_i,v_i)\}_{i=1}^n \in (\mathsf{X}^2)^n$ as*

$$f(w_{1:n}) = W_p^p(\tilde{\mu}_n,\tilde{\nu}_n) \tag{A8}$$

where $\tilde{\mu}_n,\tilde{\nu}_n$ are the univariate empirical measures computed over $\{u_i\}_{i=1}^n,\{v_i\}_{i=1}^n$ respectively, satisfies the c-bounded differences property with $c_i = 2\Delta^p/n$ for $i \in \{1,\dots,n\}$.

Proof. For clarity purposes, we start by introducing some notations. Let $n \in \mathbb{N}^*$ and $w_{1:n} \doteq \{(u_j,v_j)\}_{j=1}^n \in (\mathsf{X}^2)^n$. Denote by $\tilde{\mu}_n,\tilde{\nu}_n$ the empirical distributions supported over $(u_j)_{j=1}^n,(v_j)_{j=1}^n \in \mathsf{X}^n$ respectively. Let $(u',v') \in \mathsf{X}^2$ and $i \in \{1,\dots,n\}$. Denote by $\tilde{\mu}_n'$ the empirical distribution supported on $(u_j')_{j=1}^n$ where $u_j' = u'$ if $j = i$, $u_j' = u_j$ otherwise, and by $\tilde{\nu}_n'$ the empirical distribution over $(v_j')_{j=1}^n$ where $v_j' = v'$ if $j = i$, $v_j' = v_j$ otherwise.

By definition of the Wasserstein distance between univariate distributions (1),

$$\mathrm{W}_p^p(\tilde{\mu}_n, \tilde{\nu}_n) - \mathrm{W}_p^p(\tilde{\mu}'_n, \tilde{\nu}'_n) = \frac{1}{n}\sum_{j=1}^n |u_{\sigma(j)} - v_{\tau(j)}|^p - \frac{1}{n}\sum_{j=1}^n |u'_{\sigma'(j)} - v'_{\tau'(j)}|^p$$

where $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ (respectively, $\sigma' : \{1, \dots, n\} \to \{1, \dots, n\}$) is the permutation s.t. for $j \in \{1, \dots, n\}$, $u_{\sigma(j)}$ (resp., $u'_{\sigma'(j)}$) is the j-th smallest value of $(u_j)_{j=1}^n$ (resp., $(u'_j)_{j=1}^n$). Let $\tau : \{1, \dots, n\} \to \{1, \dots, n\}$ (respectively, $\tau' : \{1, \dots, n\} \to \{1, \dots, n\}$) s.t. for $j \in \{1, \dots, n\}$, $v_{\tau(j)}$ (resp., $v'_{\tau'(j)}$) is the j-th smallest value of $(v_j)_{j=1}^n$ (resp., $(v'_j)_{j=1}^n$).

Therefore,

$$\begin{aligned}
\mathrm{W}_p^p(\tilde{\mu}_n, \tilde{\nu}_n) - \mathrm{W}_p^p(\tilde{\mu}'_n, \tilde{\nu}'_n) &\leq \frac{1}{n}\sum_{j=1}^n |u_{\sigma'(j)} - v_{\tau'(j)}|^p - \frac{1}{n}\sum_{j=1}^n |u'_{\sigma'(j)} - v'_{\tau'(j)}|^p \\
&= \frac{1}{n}\Big(|u_i - v_{\tau'\circ\sigma'^{-1}(i)}|^p - |u' - v'_{\tau'\circ\sigma'^{-1}(i)}|^p + |u_{\sigma'\circ\tau'^{-1}(i)} - v_i|^p - |u'_{\sigma'\circ\tau'^{-1}(i)} - v'|^p\Big) \\
&\leq \frac{2\Delta^p}{n}
\end{aligned}$$

We can use the same arguments to prove that $\mathrm{W}_p^p(\tilde{\mu}'_n, \tilde{\nu}'_n) - \mathrm{W}_p^p(\tilde{\mu}_n, \tilde{\nu}_n) \leq 2\Delta^p/n$. We conclude that,

$$\big|\mathrm{W}_p^p(\tilde{\mu}_n, \tilde{\nu}_n) - \mathrm{W}_p^p(\tilde{\mu}'_n, \tilde{\nu}'_n)\big| \leq \frac{2\Delta^p}{n} \ .$$

□

Remark 1. *Lemma A2 is an extension of (Weed & Bach, 2019, Proposition 20), which establishes a concentration bound for $W_p^p(\mu, \mu_n)$ around its expectation on any finite-dimensional compact space by exploiting McDiarmid's inequality along with the Kantorovich duality. We thus use similar arguments to prove Proposition 1 , except that we leverage the closed-form expression of the one-dimensional Wasserstein distance instead of the dual formulation since we compare univariate (projected) distributions.*

Proof of Proposition 1. Let $\mu, \nu \in \mathcal{P}(\mathsf{X})$ where $\mathsf{X} \subset \mathbb{R}^d$ has a finite diameter Δ. Let $\theta \in \mathbb{S}^{d-1}$. Then, $\theta^*_\sharp\mu, \theta^*_\sharp\nu$ are both supported on a bounded domain $\mathsf{X}_\theta \subset \mathbb{R}$ whose diameter is denoted by Δ_θ and satisfies $\Delta_\theta \leq \Delta$. Consider the mapping f defined as in (A8). Given Lemma A2, we can apply Theorem A4 to bound the moment-generating function of $f - \mathbb{E}f$: for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\big[\exp(\lambda\{f - \mathbb{E}[f]\})\big] &\leq \exp(\lambda^2 \sum_{i=1}^n (2\Delta_\theta^p/n)^2/8) \\
&\leq \exp(\lambda^2\Delta_\theta^{2p}/(2n)) \leq \exp(\lambda^2\Delta^{2p}/(2n)) \ ,
\end{aligned}$$

where the expectation is computed over n samples $w_{1:n} \doteq \{(u_i, v_i)\}_{i=1}^n \in (\mathsf{X}_\theta^2)^n$ i.i.d. from $\theta^*_\sharp\mu \times \theta^*_\sharp\nu$. We conclude by using the property of push-forward measures, which gives

$$\mathbb{E}_{w_{1:n}\sim(\theta^*_\sharp\mu\times\theta^*_\sharp\nu)^n}\big[\exp(\lambda\{f(w_{1:n}) - \mathbb{E}[f(w_{1:n})]\})\big] = \mathbb{E}_{z_{1:n}\sim(\mu\times\nu)^n}\big[\exp(\lambda\{f(\theta^*(z'_{1:n})) - \mathbb{E}[f(\theta^*(z'_{1:n}))]\})\big] \quad \text{(A9)}$$

where for $z_{1:n} \doteq \{(x_i, y_i)\}_{i=1}^n \in (\mathsf{X}^2)^n$, $\theta^*(z_{1:n}) \doteq \{(\langle\theta, x_i\rangle, \langle\theta, y_i\rangle)\}_{i=1}^n \in (\mathsf{X}_\theta^2)^n$.

□

A2.3. Proof of Proposition 2

Recent work have bounded $\mathbb{E}|\mathrm{SW}_p(\mu_n, \nu_n; \rho) - \mathrm{SW}_p(\mu, \nu; \rho)|$ or $\mathbb{E}|\mathrm{SW}_p(\mu, \mu_n; \rho)|$ for specific choices of $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ (Nadjahi et al., 2020b; Manole et al., 2022; Nguyen et al., 2021; Lin et al., 2021). These results do not exactly correspond to what Theorem 2 requires, *i.e.* a bound on $\mathbb{E}|\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathrm{SW}_p^p(\mu, \nu; \rho)|$. We bound the latter quantity in Proposition 2, by specifying the proof techniques in (Manole et al., 2022) for distributions with bounded supports, then generalizing a result in (Nadjahi et al., 2020b).

Lemma A3. *Let $\mathsf{X} \subset \mathbb{R}$ be a bounded set whose diameter is denoted by $\Delta < +\infty$. Let $\mu, \nu \in \mathcal{P}(\mathsf{X})$ and denote by μ_n, ν_n the empirical distributions supported over $n \in \mathbb{N}^*$ samples i.i.d. from μ, ν respectively. Let $p \in [1, +\infty)$. Then, there exists a constant C such that,*

$$\mathbb{E}\big|W_p^p(\mu_n, \nu_n) - W_p^p(\mu, \nu)\big| \leq Cp\Delta^p n^{-1/2} \ .$$

Proof. Lemma A3 is obtained by adapting the techniques used in the proof of (Manole et al., 2022, Lemma 6), then applying (Fournier & Guillin, 2015, Theorem 1). We provide the detailed proof for completeness.

Starting from the definition of $\mathrm{W}_p^p(\mu_n, \nu_n)$ (1), then using a Taylor expansion of $(x, y) \mapsto |x - y|^p$ around $(x, y) = (F_\mu^{-1}(t), F_\nu^{-1}(t))$, we obtain

$$\begin{aligned}
\mathrm{W}_p^p(\mu_n, \nu_n) &= \int_0^1 \big|F_{\mu_n}^{-1}(t) - F_{\nu_n}^{-1}(t)\big|^p \mathrm{d}t \\
&= \int_0^1 \big|F_\mu^{-1}(t) - F_\nu^{-1}(t)\big|^p \mathrm{d}t \qquad \text{(A10)} \\
&\quad + \int_0^1 p\,\mathrm{sgn}\big(\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big)\big|\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big|^{p-1}\big\{(F_{\mu_n}^{-1}(t) - F_\mu^{-1}(t)) - (F_{\nu_n}^{-1}(t) - F_\nu^{-1}(t))\big\}\mathrm{d}t
\end{aligned}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function, $\tilde{F}_{\mu_n}^{-1}(t)$ a real number between $F_{\mu_n}^{-1}(t)$ and $F_\mu^{-1}(t)$, and $\tilde{F}_{\nu_n}^{-1}(t)$ a real number between $F_{\nu_n}^{-1}(t)$ and $F_\nu^{-1}(t)$.

By definition, (A10) is exactly $\mathrm{W}_p^p(\mu, \nu)$ and we obtain

$$\begin{aligned}
&|\mathrm{W}_p^p(\mu_n, \nu_n) - \mathrm{W}_p^p(\mu, \nu)| \\
&= \Big|\int_0^1 p\,\mathrm{sgn}\big(\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big)\big|\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big|^{p-1}\big\{(F_{\mu_n}^{-1}(t) - F_\mu^{-1}(t)) - (F_{\nu_n}^{-1}(t) - F_\nu^{-1}(t))\big\}\mathrm{d}t\Big| \\
&\leq p\int_0^1 \big|\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big|^{p-1}\Big\{\big|F_{\mu_n}^{-1}(t) - F_\mu^{-1}(t)\big| + \big|F_{\nu_n}^{-1}(t) - F_\nu^{-1}(t)\big|\Big\}\mathrm{d}t \qquad \text{(A11)} \\
&\leq p \sup_{t\in(0,1)} \big|\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big|^{p-1}\Big\{\mathrm{W}_1(\mu_n, \mu) + \mathrm{W}_1(\nu_n, \nu)\Big\} \ , \qquad \text{(A12)}
\end{aligned}$$

where (A11) follows from the triangle inequality and (A12) results from the definition of the Wasserstein distance of order 1 between univariate distributions.

We then bound $\sup_{t\in(0,1)} \big|\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big|^{p-1}$ from above. By the definition of $\tilde{F}_{\mu_n}^{-1}(t), \tilde{F}_{\nu_n}^{-1}(t)$ for $t \in (0, 1)$, we distinguish the following four cases:

(i) $\tilde{F}_{\mu_n}^{-1}(t) \leq F_{\mu_n}^{-1}(t)$, $\tilde{F}_{\nu_n}^{-1}(t) \leq F_{\nu_n}^{-1}(t)$
(ii) $\tilde{F}_{\mu_n}^{-1}(t) \leq F_\mu^{-1}(t)$, $\tilde{F}_{\nu_n}^{-1}(t) \leq F_\nu^{-1}(t)$
(iii) $\tilde{F}_{\mu_n}^{-1}(t) \leq F_{\mu_n}^{-1}(t)$, $\tilde{F}_{\nu_n}^{-1}(t) \leq F_\nu^{-1}(t)$
(iv) $\tilde{F}_{\mu_n}^{-1}(t) \leq F_\mu^{-1}(t)$, $\tilde{F}_{\nu_n}^{-1}(t) \leq F_{\nu_n}^{-1}(t)$

Hence, using the definition of quantile functions and the fact that the supports of μ, ν are assumed to be bounded, we obtain

$$\sup_{t\in(0,1)} \big|\tilde{F}_{\mu_n}^{-1}(t) - \tilde{F}_{\nu_n}^{-1}(t)\big|^{p-1} \leq \Delta^{p-1} \ .$$

We conclude that,

$$|\mathrm{W}_p^p(\mu_n, \nu_n) - \mathrm{W}_p^p(\mu, \nu)| \leq p\Delta^{p-1}\Big\{\mathrm{W}_1(\mu_n, \mu) + \mathrm{W}_1(\nu_n, \nu)\Big\} \ .$$

and by linearity of the expectation,

$$\mathbb{E}|\mathrm{W}_p^p(\mu_n, \nu_n) - \mathrm{W}_p^p(\mu, \nu)| \leq p\Delta^{p-1}\Big\{\mathbb{E}[\mathrm{W}_1(\mu_n, \mu)] + \mathbb{E}[\mathrm{W}_1(\nu_n, \nu)]\Big\} \ . \qquad \text{(A13)}$$

Our final result follows from applying (Fournier & Guillin, 2015, Theorem 1). Since $\mu, \nu \in \mathcal{P}(\mathsf{X})$ where $\mathsf{X} \subset \mathbb{R}$ is a bounded set with finite diameter $\Delta < \infty$, then for any $q \geq 1$, the moment of μ (or ν) of order q is bounded by Δ^q. Therefore, the application of (Fournier & Guillin, 2015, Theorem 1) yields,

$$\mathbb{E}[\mathrm{W}_1(\mu_n, \mu)] \leq C' \Delta n^{-1/2} \ , \ \ \mathbb{E}[\mathrm{W}_1(\nu_n, \nu)] \leq C' \Delta n^{-1/2} \ . \tag{A14}$$

where C' is a constant. We conclude by plugging (A14) in (A13).

□

Proof of Proposition 2. Let $\theta \in \mathbb{S}^{d-1}$. Since we assume that $\mu, \nu \in \mathcal{P}(\mathsf{X})$ where $\mathsf{X} \subset \mathbb{R}^d$ is a bounded subset with finite diameter Δ, one can easily prove that $\theta^*_\sharp \mu, \theta^*_\sharp \nu$ are supported on a bounded domain with diameter $\Delta_\theta \leq \Delta < +\infty$. Therefore, by Lemma A3, there exists a constant C such that,

$$\mathbb{E}\big|\mathrm{W}_p^p(\theta^*_\sharp \mu_n, \theta^*_\sharp \nu_n) - \mathrm{W}_p^p(\theta^*_\sharp \mu, \theta^*_\sharp \nu)\big| \leq C p \Delta^p n^{-1/2} \ . \tag{A15}$$

Next, we adapt the proof techniques in (Nadjahi et al., 2020b, Theorem 4) to establish the following inequality: for *any* $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$,

$$\mathbb{E}|\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathrm{SW}_p^p(\mu, \nu; \rho)| \leq \int_{\mathbb{S}^{d-1}} \mathbb{E}|\mathrm{W}_p^p(\theta^*_\sharp \mu_n, \theta^*_\sharp \nu_n) - \mathrm{W}_p^p(\theta^*_\sharp \mu, \theta^*_\sharp \nu)| \mathrm{d}\rho(\theta) \ . \tag{A16}$$

Hence, by plugging (A15) in (A16), we obtain

$$\mathbb{E}|\mathrm{SW}_p^p(\mu_n, \nu_n; \rho) - \mathrm{SW}_p^p(\mu, \nu; \rho)| \leq C p \Delta^p n^{-1/2} \ .$$

□

A2.4. Final Bound for Bounded Supports

By incorporating Propositions 1 and 2 in Theorem 2, we obtain the following result. Corollary A1 corresponds to a specialization of our generic bound when considering distributions with bounded supports.

Corollary A1. *Let $p \in [1, +\infty)$ and assume a bounded diameter Δ. Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\delta > 0$. Then, with probability at least $1 - \delta$, for all $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\lambda > 0$, there exists a constant C such that,*

$$SW_p^p(\mu_n, \nu_n; \rho) \leq SW_p^p(\mu, \nu; \rho) + \{KL(\rho || \rho_0) + \log(1/\delta)\} \lambda^{-1} + \lambda \Delta^{2p} (2n)^{-1} + C p \Delta^p n^{-1/2}$$

A2.5. Proof of Proposition 3

When the supports of the distributions are not bounded, Lemma A2 does not hold true, thus preventing the use of McDiarmid's inequality. Hence, to compute $\varphi_{\mu,\nu,p}$, we may use extensions of McDiarmid's inequality which replace the finite-diameter constraint by conditions on the moments of the distributions.

In particular, Proposition 3 follows from applying (Kontorovich, 2014, Theorem 1), a concentration result based on the notion of *sub-Gaussian diameter*.

Definition A7 (Sub-Gaussian diameter (Kontorovich, 2014))**.** *Let η be a distance function and (X, η, μ) be the associated metric probability space. Consider a sequence of $n \in \mathbb{N}^*$ independent random variables $(X_i)_{i=1}^n$ with X_i distributed from μ for $i \in \{1, \ldots, n\}$. Let $\Xi(\mathsf{X})$ be the random variable defined by*

$$\Xi(\mathsf{X}) = \varepsilon \eta(X, X') \ ,$$

where X, X' are two independent realizations from μ and ε is a random variable valued in $\{-1, 1\}$ s.t. $p(\varepsilon = 1) = 1/2$ and ε is independent from X, X'. Additionally, suppose there exists $\sigma > 0$ s.t. for $\lambda \in \mathbb{R}$, $\mathbb{E}_\mu[\exp(\lambda X)] \leq \exp(\sigma^2 \lambda^2 / 2)$. The sub-Gaussian diameter *of (X, η, μ), denoted by $\Delta_{SG}(\mathsf{X})$, is defined as $\Delta_{SG}(\mathsf{X}) = \sigma\big(\Xi(\mathsf{X})\big)$.*

Note that $\Delta_{\mathrm{SG}} \leq \Delta$ (Kontorovich, 2014, Lemma 1). Since a set with infinite diameter may have a finite sub-Gaussian diameter, Theorem A5 relaxes the conditions of Theorem A4.

Theorem A5 (Theorem 1 (Kontorovich, 2014))**.** *Let $\mathsf{X} \subset \mathbb{R}^d$ and $\eta : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ be a distance function. Consider the metric probability space (X, η). For $n \in \mathbb{N}^*$, let X^n be the product probability space equipped with the product measure $\mu^n = \mu_1 \times \cdots \times \mu_n$, where $\mu_i = \mu$. Define the L_1* product metric *η^n for any $(x, x') \in \mathsf{X}^n \times \mathsf{X}^n$ as,*

$$\eta^n(x, x') = \sum_{i=1}^{n} \eta(x_i, x'_i) \ .$$

Let $f : \mathsf{X}^n \to \mathbb{R}$ s.t. f is 1-Lipschitz with respect to η^n, i.e. for any $(x, x') \in \mathsf{X}^n \times \mathsf{X}^n$, $|f(x) - f(x')| \leq \eta(x, x')$. Then, $\mathbb{E}[f] < +\infty$ and for $\lambda > 0$,

$$\mathbb{E}\left[\exp(\lambda\{f - \mathbb{E}[f]\})\right] \leq \exp(\lambda^2 n \Delta_{SG}(\mathsf{X})^2/2) \ .$$

As discussed in (Kontorovich, 2014), the sub-Gaussian distributions on $\mathbb{R}$ are precisely those for which $\Delta_{\text{SG}}(\mathbb{R}) < +\infty$. Proposition 3 then results from applying Theorem A5, as explained below.

Proof of Proposition 3. First, we prove that for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ such that μ is sub-Gaussian with parameter σ^2, then $\mu \in \mathcal{P}_1(\mathbb{R}^d)$. By definition, the first moment of μ is $\mathrm{m}_1(\mu) = \int_{\mathbb{R}^d} \|x\| \mathrm{d}\mu(x)$. For any $x \in \mathbb{R}^d$, we know that

$$\|x\| = \left(\sum_{k=1}^{d} |x_k|^2\right)^{1/2} \leq \sum_{k=1}^{d} |x_k|$$

Therefore, $\mathrm{m}_1(\mu)$ can be bounded from above as follows.

$$\begin{aligned}
\mathrm{m}_1(\mu) &\leq \int_{\mathbb{R}^d} \sum_{k=1}^{d} |x_k| \mathrm{d}\mu(x) \\
&\leq \sum_{k=1}^{d} \int_{\mathbb{R}^d} |x_k| \mathrm{d}\mu(x) && \text{(A17)} \\
&\leq \sum_{k=1}^{d} \int_{\mathbb{R}^d} |\left\langle \theta^k, x \right\rangle| \mathrm{d}\mu(x) \\
&\leq \sum_{k=1}^{d} \int_{\mathbb{R}} |t| \mathrm{d}(\theta^k)^\star_\sharp \mu(t) && \text{(A18)} \\
&\leq d\sqrt{2\pi\sigma^2} && \text{(A19)}
\end{aligned}$$

where for $k \in \{1, \ldots, d\}$, $\theta^k \in \mathbb{S}^{d-1}$ is defined as $(\theta^k)_i = 1$ if $i = k$, $(\theta^k)_i = 0$ otherwise. (A17) results from the linearity of the expectation, (A18) is obtained by applying the property of pushforward measures. (A19) follows from the sub-Gaussian assumption on μ (Definition 3) and (Rivasplata, 2012, Proposition 3.2). Since $\mathrm{m}_1(\mu) < \infty$ (A19), we conclude that $\mu \in \mathcal{P}_1(\mathbb{R}^d)$.

Now, consider the product metric space $(\mathbb{R}^2, \eta)$ where $\eta : \mathbb{R}^2 \to \mathbb{R}_+$ is the distance function defined for $w \doteq (u, v) \in \mathbb{R}^2$, $w' \doteq (u', v') \in \mathbb{R}^2$ as,

$$\eta(w, w') \doteq \|u - u'\| + \|v - v'\| = |u - u'| + |v - v'| \ .$$

Let $n \in \mathbb{N}^*$ and define $f : (\mathbb{R}^2)^n \to \mathbb{R}_+$ as: for any $w_{1:n} \doteq (w_i)_{i=1}^n \in (\mathbb{R}^2)^n$ such that $\forall i \in \{1, \ldots, n\}, w_i = (u_i, v_i) \in \mathbb{R}^2$,

$$f(w_{1:n}) = n\, \mathrm{W}_1(\tilde{\mu}_n, \tilde{\nu}_n) \ , \tag{A20}$$

where $\tilde{\mu}_n, \tilde{\nu}_n$ are the empirical distributions computed over $(u_i)_{i=1}^n, (v_i)_{i=1}^n$ respectively, *i.e.*, denoting by δ_x the Dirac measure at x,

$$\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{u_i}, \quad \tilde{\nu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{v_i} \ .$$

We prove that f is 1-Lipschitz with respect to the L_1 product metric η^n defined for any $w_{1:n} \doteq \{(u_i, v_i)\}_{i=1}^n \in (\mathbb{R}^2)^n$, $w'_{1:n} \doteq \{(u'_i, v'_i)\}_{i=1}^n \in (\mathbb{R}^2)^n$ as,

$$\eta^n(w_{1:n}, w'_{1:n}) = \sum_{i=1}^n \left\{ \|u_i - u'_i\| + \|v_i - v'_i\| \right\} . \tag{A21}$$

Since the Wasserstein distance satisfies the triangle inequality, one has

$$|\mathrm{W}_1(\tilde{\mu}_n, \tilde{\nu}_n) - \mathrm{W}_1(\tilde{\mu}'_n, \tilde{\nu}'_n)| \leq \mathrm{W}_1(\tilde{\mu}_n, \tilde{\mu}'_n) + \mathrm{W}_1(\tilde{\nu}_n, \tilde{\nu}'_n)$$

where $\tilde{\mu}_n, \tilde{\nu}_n$ are the empirical distributions supported on $(u_i)_{i=1}^n, (v_i)_{i=1}^n$ respectively, and $\tilde{\mu}'_n, \tilde{\nu}'_n$ are the empirical distributions supported on $(u'_i)_{i=1}^n, (v'_i)_{i=1}^n$ respectively. By definition of the Wasserstein distance between univariate discrete distributions (Peyré & Cuturi, 2019, Remark 2.28),

$$\begin{aligned}
\mathrm{W}_1(\tilde{\mu}_n, \tilde{\mu}'_n) &= \frac{1}{n} \sum_{i=1}^n \left| u_{(i)} - u'_{(i)} \right| \\
&\leq \frac{1}{n} \sum_{i=1}^n |u_i - u'_i| = \frac{1}{n} \sum_{i=1}^n \|u_i - u'_i\|
\end{aligned}$$

where $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(n)}$ and $u'_{(1)} \leq u'_{(2)} \leq \cdots \leq u'_{(n)}$. Analogously, $\mathrm{W}_1(\tilde{\nu}_n, \tilde{\nu}'_n) \leq \frac{1}{n} \sum_{i=1}^n \|v_i - v'_i\|$. Therefore,

$$|\mathrm{W}_1(\tilde{\mu}_n, \tilde{\nu}_n) - \mathrm{W}_1(\tilde{\mu}'_n, \tilde{\nu}'_n)| \leq \frac{1}{n} \sum_{i=1}^n \left\{ \|u_i - u'_i\| + \|v_i - v'_i\| \right\} . \tag{A22}$$

We conclude from (A20) and (A22) that f is 1-Lipschitz with respect to the product metric η^n, as defined in (A21).

Next, let $\theta \in \mathbb{S}^{d-1}$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ such that μ, ν are sub-Gaussian with respective variance proxy σ^2, τ^2. Consider the probability metric space $(\mathbb{R}^2, \eta, \theta^*_\sharp \mu \times \theta^*_\sharp \nu)$. By Definition A7 and the properties of the sum of independent sub-Gaussian random variables (Rivasplata, 2012, Theorem 2.7), the sub-Gaussian diameter of that space is $\Delta_{\mathrm{SG}}(\mathbb{R}^2) = \sqrt{2(\sigma^2 + \tau^2)}$.

We conclude the proof by applying Theorem A5 to f as defined in (A20), then reformulating the expectation over $\theta^*_\sharp \mu \times \theta^*_\sharp \nu$ as an expectation over $\mu \times \nu$ using the property of push-forward measures (see (A9)).

□

A2.6. Proof of Proposition 4

Proposition 4 results from the same arguments as in the proof of (Lei, 2020, Corollary 5.2). The latter result is obtained by applying a generalized McDiarmid's inequality, which we recall in Theorem A6.

Theorem A6 (Bernstein-type McDiarmid's inequality (Lei, 2020))**.** *Let $\mathsf{X} \subset \mathbb{R}^d$ and $X = (X_i)_{i=1}^n$ be a sequence of $n \in \mathbb{N}^*$ random variables i.i.d. from $\mu \in \mathcal{P}(\mathsf{X})$. Let $f : \mathsf{X}^n \to \mathbb{R}$ s.t. $\mathbb{E}|f| < \infty$. For $i \in \{1, \ldots, n\}$, let X'_i be an independent copy of X_i and $X'_{(i)} = (X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n)$. Assume for $i \in \{1, \ldots, n\}$, there exists $c_i, M > 0$ s.t. for $k \geq 2$,*

$$\mathbb{E}\left[|f(X) - f(X'_{(i)})|^k \mid X_{-i} \right] \leq c_i^2 k! M^{k-2}/2 , \tag{A23}$$

where $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Then, for $\lambda > 0$ s.t. $\lambda M < 1$,

$$\mathbb{E}\left[\exp\{\lambda(f - \mathbb{E}[f])\} \right] \leq \exp\left(\lambda^2 \|c\|^2 / \{2(1 - \lambda M)\} \right) .$$

Proof of Proposition 4. First, we justify why for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ s.t. μ satisfies the (σ^2, b)-Bernstein condition, $\mu \in \mathcal{P}_1(\mathbb{R}^d)$. By (A18), the first order moment of μ, $\mathrm{m}_1(\mu)$ can be bounded as,

$$\begin{aligned}
\mathrm{m}_1(\mu) &\leq \sum_{k=1}^d \int_{\mathbb{R}} |t| \mathrm{d}(\theta^k)^\star_\sharp \mu(t) \\
&\leq \sum_{k=1}^d \left\{ \int_{\mathbb{R}} |t|^2 \mathrm{d}(\theta^k)^\star_\sharp \mu(t) \right\}^{1/2} && \text{(A24)} \\
&\leq d\sigma && \text{(A25)}
\end{aligned}$$

where (A24) is obtained by applying Hölder's inequality, and (A25) results from Definition 4. Hence, $\mathrm{m}_1(\mu) < \infty$ and $\mu \in \mathcal{P}_1(\mathbb{R}^d)$.

The rest of the proof consists in applying Theorem A6 to $f : (\mathbb{R}^2)^n \to \mathbb{R}_+$ defined for any $w_{1:n} \doteq \{(u_i, v_i)\}_{i=1}^n \in (\mathbb{R}^2)^n$ as,

$$f(w_{1:n}) = \mathrm{W}_1(\tilde{\mu}_n, \tilde{\nu}_n) \tag{A26}$$

where $\tilde{\mu}_n, \tilde{\nu}_n$ are the empirical distributions of $(u_i)_{i=1}^n, (v_i)_{i=1}^n$ respectively.

For $i \in \{1, \dots, n\}$, let $(u_i', v_i') \in \mathbb{R}^2$. Denote by $\tilde{\mu}_n'$ the empirical distribution supported on $(u_1, \dots, u_{i-1}, u_i', u_{i+1}, \dots, u_n) \in \mathbb{R}^n$, and by $\tilde{\nu}_n'$ the empirical distribution supported on $(v_1, \dots, v_{i-1}, v_i', v_{i+1}, \dots, v_n) \in \mathbb{R}^n$. Then,

$$\begin{aligned}
|\mathrm{W}_1(\tilde{\mu}_n, \tilde{\nu}_n) - \mathrm{W}_1(\tilde{\mu}_n', \tilde{\nu}_n')| &\leq \mathrm{W}_1(\tilde{\mu}_n, \tilde{\mu}_n') + \mathrm{W}_1(\tilde{\nu}_n, \tilde{\nu}_n') && \text{(A27)}\\
&\leq \frac{1}{n}\big\{|u_i - u_i'| + \sum_{j=1,\dots,n, j\neq i} |u_j - u_j|\big\} + \frac{1}{n}\big\{|v_i - v_i'| + \sum_{j=1,\dots,n, j\neq i} |v_j - v_j|\big\} && \text{(A28)}\\
&\leq \frac{1}{n}\big\{|u_i - u_i'| + |v_i - v_i'|\big\} && \text{(A29)}
\end{aligned}$$

where (A27) follows from the fact that W_1 satisfies the triangle inequality, and (A28) results from the definition of the Wasserstein distance between univariate empirical distributions (Peyré & Cuturi, 2019, Remark 2.28).

Now, let $\mu \in \mathcal{P}(\mathbb{R}^d)$ (respectively, $\nu \in \mathcal{P}(\mathbb{R}^d)$) satisfy the (σ^2, b) (resp., (τ^2, c))-Bernstein condition (Definition 4). Let $\theta \in \mathbb{S}^{d-1}$ and consider $w_{1:n} = \{(u_i, v_i)\}_{i=1}^n \in (\mathbb{R}^2)^n$ i.i.d. from the product measure $\theta^*_\sharp \mu \times \theta^*_\sharp \nu$. We justify why f satisfies the conditions of Theorem A6.

First, we show that $\mathbb{E}|f|$ is finite, where the expectation $\mathbb{E}$ is computed over n i.i.d. samples $\{(u_i, v_i)\}_{i=1}^n$ from $\theta^*_\sharp \mu \times \theta^*_\sharp \nu_n$.

$$\begin{aligned}
\mathbb{E}|f| &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|u_i - v_i|] \leq \frac{1}{n}\sum_{i=1}^n \{\mathbb{E}|u_i| + \mathbb{E}|v_i|\}\\
&\leq \frac{1}{n}\sum_{i=1}^n \Big\{\mathbb{E}[|u_i|^2]^{1/2} + \mathbb{E}[|v_i|^2]^{1/2}\Big\} && \text{(A30)}\\
&\leq \sigma + \tau && \text{(A31)}
\end{aligned}$$

where (A30) results from Hölder's inequality, and (A31) directly follows from the definition of the Bernstein condition (Definition 4).

Besides, by using (A29) and the Bernstein condition Definition 4, one can show that

$$\mathbb{E}[|\mathrm{W}_1(\tilde{\mu}_n, \tilde{\nu}_n) - \mathrm{W}_1(\tilde{\mu}_n', \tilde{\nu}_n')|^k \mid u_{-i}, v_{-i}] \leq n^{-k} 2^{2(k-1)} [\sigma^2 b^{k-2} + \tau^2 c^{k-2}] k!$$

where the expectation is computed over $\{(u_i, v_i)\}_{i=1}^n$ i.i.d. from $\theta^*_\sharp \mu \times \theta^*_\sharp \nu$. In other words, f as defined in (A26) satisfies (A23) with, for $i \in \{1, \dots, 2n\}$, $c_i = 2\sigma_\star n^{-1}$ and $M = 4b_\star n^{-1}$, where $\sigma_\star = \max(\sigma, \tau)$ and $b_\star = \max(b, c)$. Our final result follows from applying Theorem A6 to f, then applying the property of push-forward measures to obtain the expectation with respect to $\mu \times \nu$ (see (A9)).

□

A2.7. Final Bound for Unbounded Supports

Before deriving the specialization of Theorem 2 for distributions with unbounded supports, we recall a useful bound on $\mathrm{SW}_p^p(\cdot, \cdot; \pi)$ with $\pi = \mathcal{U}(\mathbb{S}^{d-1})$ (Theorem A7), which can be generalized for SW based on any $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ by adapting the proof techniques in (Manole et al., 2022, Theorem 2).

Theorem A7 ((Manole et al., 2022)). *Let $p \geq 1$, $q > 2p$, $s \geq 1$ and $\pi = \mathcal{U}(\mathbb{S}^{d-1})$. Denote $\mathcal{P}_{p,q}(s) = \big\{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{S}^{d-1}} \mathbb{E}_\mu[|\theta^\top x|^q]^{p/q} \mathrm{d}\pi(\theta) \leq s\big\}$. Let $\mu, \nu \in \mathcal{P}_{p,q}(s)$. Then, there exists a constant $C(p, q) > 0$ depending on p, q such that,*

$$\mathbb{E}|SW_p^p(\mu_n, \nu_n; \pi) - SW_p^p(\mu, \nu; \pi)| \leq C(p, q) s \log(n)^{1/2} n^{-1/2}$$

We show that under the sub-Gaussian or the Bernstein moment condition assumptions, the assumptions in Theorem A7 are satisfied, thus allowing its application in these two settings. This yields Corollaries A2 and A3, which we state and prove hereafter.

Corollary A2. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. Assume that μ (respectively, ν) is sub-Gaussian with variance proxy σ^2 (resp., τ^2). Let $\sigma_\star^2 = \max(\sigma^2, \tau^2)$. Then, there exists $C'(p) > 0$ such that,*

$$\mathbb{E}|SW_p^p(\mu_n, \nu_n; \rho) - SW_p^p(\mu, \nu; \rho)| \leq C'(p)(4\sigma_\star^2)^p \log(n)^{1/2} n^{-1/2} \ .$$

Proof. Under the sub-Gaussian assumption on μ and ν, the moments of $\theta_\sharp^* \mu, \theta_\sharp^* \nu$ can be bounded for any $\theta \in \mathbb{S}^{d-1}$ as follows: for any $k \in \mathbb{N}^*$,

$$\mathbb{E}_\mu[|\langle \theta, x \rangle|^{2k}] \leq k!(4\sigma^2)^k \ , \quad \mathbb{E}_\nu[|\langle \theta, y \rangle|^{2k}] \leq k!(4\tau^2)^k \ .$$

We conclude that $\mu, \nu \in \mathcal{P}_{p,2(p+1)}(s)$ with $s = \{(p+1)!\}^{p/(2(p+1))}(4\sigma_\star^2)^p$ and $\sigma_\star^2 = \max(\sigma^2, \tau^2)$. The final result follows from applying Theorem A7.

□

Corollary A3. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. Assume that μ and ν satisfy the Bernstein condition, with parameters (σ^2, b) and (τ^2, c) respectively. Let $\sigma_\star^2 = \max(\sigma^2, \tau^2)$ and $b_\star = \max(b, c)$. Then, there exists $C'(p, q) > 0$ such that*

$$\mathbb{E}|SW_p^p(\mu_n, \nu_n; \rho) - SW_p^p(\mu, \nu; \rho)| \leq C'(p, q)\sigma_\star^{2p/q} b_\star^{p(q-2)/q} \log(n)^{1/2} n^{-1/2} \ .$$

Proof. Under the Bernstein condition on the moments of μ, ν, we can use the definition of the push-forward measures along with the Cauchy-Schwarz inequality and obtain for any $\theta \in \mathbb{S}^{d-1}$ and $k \in \mathbb{N}^*$,

$$\mathbb{E}_\mu[|\langle \theta, x \rangle|^{2k}] \leq \sigma^2 k! b^{k-2}/2 \ , \quad \mathbb{E}_\nu[|\langle \theta, y \rangle|^{2k}] \leq \tau^2 k! c^{k-2}/2 \ . \tag{A32}$$

Let $q > 2p$. By (A32), $\mu, \nu \in \mathcal{P}_{p,q}(s)$ with $s = (\sigma_\star^2 q!/2)^{p/q} b_\star^{p(q-2)/q}$. The application of Theorem A7 concludes the proof.

□

We can finally provide the refined bounds, assuming the distributions are either sub-Gaussian or satisfy the Bernstein condition. On the one hand, incorporating Proposition 3 and Corollary A2 in Theorem 2 gives us the following corollary.

Corollary A4. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Assume μ (resp., ν) is sub-Gaussian with variance proxy σ^2 (resp., τ^2). Let $\sigma_\star^2 \doteq \max(\sigma^2, \tau^2)$. Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\delta > 0$. Then, with probability at least $1 - \delta$, for all $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\lambda > 0$, there exists $C > 0$ such that*

$$\begin{aligned} SW_1(\mu_n, \nu_n; \rho) \leq SW_1(\mu, \nu; \rho) &+ \{KL(\rho||\rho_0) + \log(1/\delta)\}\lambda^{-1} \\ &+ \lambda(\sigma^2 + \tau^2)n^{-1} + C\sigma_\star^2 \log(n)^{1/2} n^{-1/2} \ . \end{aligned}$$

On the other hand, we leverage Proposition 4, Corollary A3 and Theorem A3 to derive the specified bound below.

Corollary A5. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Assume that μ and ν satisfy the Bernstein condition, with parameters (σ^2, b) and (τ^2, c) respectively. Let $\sigma_\star^2 = \max(\sigma^2, \tau^2)$ and $b_\star = \max(b, c)$. Let $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\delta > 0$. Then, with probability at least $1 - \delta$, for all $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\lambda > 0$ s.t. $\lambda < (2b_\star)^{-1}n$, for $q > 2$, there exists $C(q) > 0$ such that*

$$\begin{aligned} SW_1(\mu_n, \nu_n; \rho) \leq SW_1(\mu, \nu; \rho) &+ \{KL(\rho||\rho_0) + \log(1/\delta)\}\lambda^{-1} \\ &+ 2\lambda\sigma_\star^2(1 - 2b_\star\lambda n^{-1})^{-1}n^{-2} + C(q)\sigma_\star^{2/q} b_\star^{(q-2)/q} \log(n)^{1/2} n^{-1/2} \ . \end{aligned}$$

A3. Additional Experimental Details

All our numerical experiments presented in Section 5 can be reproduced using the code we provided in https://github.com/rubenohana/PAC-Bayesian_Sliced-Wasserstein.

Algorithm A2 PAC-Bayes bound optimization for vMF-based SW

Input: Datasets: $x_{1:n} = (x_i)_{i=1}^n$, $y_{1:n} = (y_i)_{i=1}^n$
SW order, number of slices: $p \in [1, +\infty)$, $n_S \in \mathbb{N}^*$
Bound parameter: $\lambda \in \mathbb{R}_+^*$
Number of iterations, learning rate: $T \in \mathbb{N}^*$, $\eta \in (0, 1)$
Initialized parameters: $(\mathrm{m}^{(0)}, \kappa^{(0)}) \in \mathbb{S}^{d-1} \times \mathbb{R}_+^*$

Output: Final parameters: $(\mathrm{m}^{(T)}, \kappa^{(T)})$
for $t \leftarrow 0$ to $T-1$ **do**
  $\rho^{(t)} \leftarrow \mathrm{vMF}(\mathrm{m}^{(t)}, \kappa^{(t)})$
  for $k \leftarrow 1$ to n_S **do**
    $\theta_k^{(t)} \sim \rho^{(t)}$ (Davidson et al., 2018, Algorithm 1)
  end for
  $\rho_n^{(t)} \leftarrow n_S^{-1} \sum_{k=1}^n \delta_{\theta_k^{(t)}}$
  $\mathcal{L}(x_{1:n}, y_{1:n}, \rho^{(t)}, \lambda) \leftarrow \mathrm{SW}_p^p(\mu_n, \nu_n; \rho_n^{(t)}) - \lambda^{-1}\mathrm{KL}(\rho^{(t)}||\rho^{(0)})$
  $\begin{bmatrix} \mathrm{m}^{(t+1)} \\ \kappa^{(t+1)} \end{bmatrix} \leftarrow \begin{bmatrix} \mathrm{m}^{(t)} \\ \kappa^{(t)} \end{bmatrix} + \eta \begin{bmatrix} \nabla_{\mathrm{m}} \mathcal{L}(x_{1:n}, y_{1:n}, \rho^{(t)}, \lambda) \\ \nabla_{\kappa} \mathcal{L}(x_{1:n}, y_{1:n}, \rho^{(t)}, \lambda) \end{bmatrix}$
end for
Return $(\mathrm{m}^{(T)}, \kappa^{(T)})$

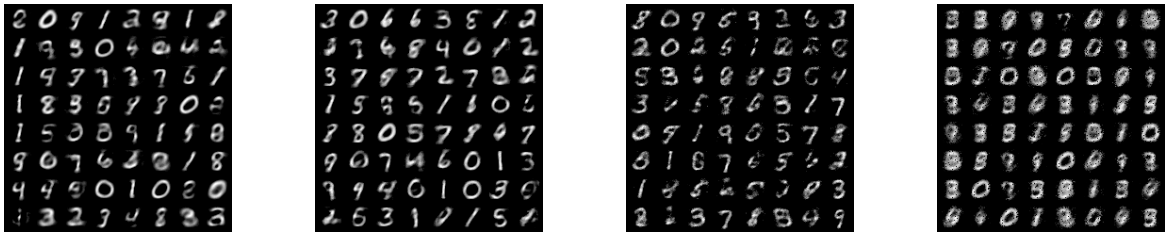


Figure A1. Examples of generated MNIST digits. Left to right: DSW, DSW-10, maxSW, maxSW-10.

A3.1. Details on the Algorithmic Procedure

For clarity, we specify Algorithm 1 when the optimization is performed over the space of von Mises-Fisher distributions (Definition 5). The procedure is detailed in Algorithm A2.

A3.2. Additional Results

Figure A1 displays additional qualitative results for the generative modeling experiment. We observe that the images generated by DSW have a better quality than the ones produced by maxSW, even if DSW is not optimized at every training iteration.

On Figure A2 are shown the results obtained on the generative modeling experiment of Section 5 using the PAC-SW loss. PAC-SW can be competitive with DSW, but takes more time to execute as the computation of the KL cost is more costly than the regularization term of DSW. However, we observe that the distribution of slices that we learn generalizes well.

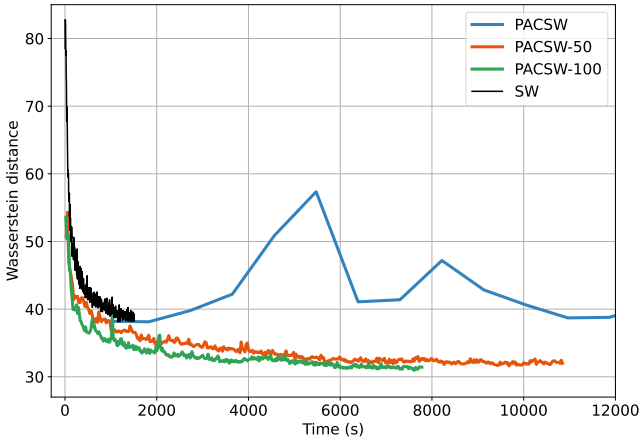


Figure A2. Generative modeling experiment when the slice distribution of PAC-SW is updated either at each iteration (PACSW), every 50 iterations (PACSW-50) or every 100 iterations (PACSW-100). Timing results of this experiment were obtained with a NVIDIA GPU A100 80 GB, compared to Figure 4 which was on a NVIDIA V100.