Sliced-Wasserstein distance for large-scale machine learning : theory, methodology and extensions

Kimia Nadjahi

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAT050

Thèse de doctorat

Sliced-Wasserstein Distance for Large-Scale Machine Learning: Theory, Methodology and Extensions

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Signal, Images, Automatique et robotique

Thèse présentée et soutenue à Palaiseau, le 23 Novembre 2021, par

KIMIA NADJAHI

Composition du Jury :

Julie Delon Professeur, Université de Paris (MAP5)	Présidente
Nicolas Courty Professeur, Université de Bretagne Sud (IRISA)	Rapporteur
Gabriel Peyré Directeur de recherche, CNRS, École Normale Supérieure (DMA)	Rapporteur
Laetitia Chapel Maître de conférences, Université de Bretagne Sud (IRISA)	Examinatrice
Marco Cuturi Professeur, ENSAE (CREST) et Google Brain	Examinateur
Justin Solomon Associate Professor, MIT (CSAIL)	Examinateur
Roland Badeau Professeur, Télécom Paris (LTCI)	Directeur de thèse
Alain Durmus Maître de conférences, ENS Paris-Saclay (CMLA)	Invité
Umut Şimşekli Chargé de recherche, Inria (SIERRA) et École Normale Supérieure (DI ENS)	Invité

À mes parents, pour leur courage et leur dévouement

Abstract

Many methods for statistical inference and generative modeling rely on a *probability divergence* to effectively compare two probability distributions. This divergence should be carefully chosen as its theoretical properties and practical implications strongly affect the performance of the associated algorithm. In that context, the *Wasserstein distance*, which emerges from *optimal transport* (OT) theory, has been an interesting choice due to its theoretical guarantees. However, it suffers from important computational and statistical limitations, which have severly hindered its use to problems with large amounts of high-dimensional data.

In recent years, several workarounds have been proposed to alleviate these issues thus enable the use of OT in machine learning (ML) applications. In particular, the *Sliced-Wasserstein distance* (SW) is an alternative OT metric, which compares two distributions by computing the expected Wasserstein distance between their one-dimensional linear projections. SW has been increasingly popular since it can offer significant computational advantages over the Wasserstein distance, especially on large-scale problems, and has been successfully applied in many practical tasks, such as classification, Bayesian inference, and implicit generative modeling. Nevertheless, there has been little work regarding the theoretical guarantees of such SW-based methods. This thesis further explores the use of the Sliced-Wasserstein distance in modern statistical and ML problems, with a twofold objective: on the one hand, provide new theoretical insights to understand in depth the empirical behavior of existing SW-based algorithms; on the other hand, design novel tools inspired by SW to extend its applicability and offer increased scalability.

We first focus on the estimators obtained by minimizing SW, which form the basis of several recently proposed generative modeling methods. We prove a set of asymptotic properties of these estimators, as well as a central limit theorem which characterizes their asymptotic distribution and exhibits a dimension-free convergence rate. We also develop a novel likelihood-free inference technique, SW-ABC, by incorporating SW in the *Approximate Bayesian Computation* framework. We prove asymptotical guarantees on the convergence of the posterior distribution returned by SW-ABC, and illustrate the advantages of our algorithm in practice, on synthetic data and an image denoising problem.

By definition, SW is an expectation over random projections, which is intractable in general and commonly estimated with a simple Monte Carlo procedure. This approximation induces an error that can potentially degrade the performance of SW-based algorithms on high-dimensional settings. To overcome this issue, we introduce the *Generalized Sliced-Wasserstein distances* (GSW), which extends the definition of SW by considering *nonlinear* projections of the distributions. We study the metric axioms of GSW using the theory of the *Radon transform*, and show that GSW can yield better results than SW on generative modeling applications.

We then adopt another perspective to address the issues due to the Monte Carlo approximation: we leverage the *concentration of measure* phenomenon, which states under mild assumptions that one-dimensional linear projections of a high-dimensional random vector are approximately Gaussian. Based on this result, we develop a simple deterministic approximation for SW, which can lead to a significant computational time reduction over Monte Carlo. We derive nonasymptotical guarantees for our methodology under a weak dependence condition, validate them on synthetic experiments, and illustrate the proposed approximation on image generation.

Inspired by the growing success of SW, new instances of *sliced probability divergences* (SPDs) have been deployed in statistical and ML applications, but their theoretical implications have not been well established. To bridge this gap, we introduce the first general definition of SPDs and investigate their statistical and topological properties. Our theoretical analysis sheds light to the consequences of slicing: in particular, we prove that the sample complexity of any SPD does not depend on the data dimension, but can be impacted by an additional error term due to the Monte Carlo approximation. We then apply our general results to specific SPDs in order to gain a better understanding of them.

Résumé

De nombreuses méthodes d'inférence statistique et de modélisation générative reposent sur une *divergence* pour comparer de façon pertinente deux distributions de probabilité. Cette divergence doit être choisie avec soin puisque ses propriétés théoriques et répercussions pratiques affectent fortement les performances de l'algorithme en question. Dans ce contexte, la *distance de Wasserstein*, qui découle de la théorie du *transport optimal*, est un choix intéressant de par ses garanties théoriques. Cependant, elle présente d'importantes limites computationnelle et statistique qui l'empêchent de traiter efficacement de grandes quantités de données à haute dimension.

Plusieurs méthodes visant à atténuer ces problèmes ont été proposées ces dernières années, permettant ainsi l'utilisation du transport optimal pour l'apprentissage automatique. En particulier, la *distance de Sliced-Wasserstein* (SW) est une métrique alternative qui compare deux distributions en calculant la distance de Wasserstein moyenne entre leurs projections linéaires unidimensionnelles. SW est de plus en plus utilisée en raison de sa capacité à fournir des avantages calculatoires significatifs par rapport à la distance de Wasserstein, surtout pour les problèmes à grande échelle, et a fait ses preuves dans de nombreuses tâches pratiques, telles que la classification, l'inférence bayésienne et la modélisation générative implicite. Néanmoins, peu de travaux ont étudié les garanties théoriques des méthodes basées sur SW. Cette thèse explore plus en profondeur l'utilisation de SW pour des problèmes modernes de statistique et d'apprentissage automatique, avec un double objectif : d'une part, apporter de nouvelles connaissances théoriques permettant une compréhension approfondie des algorithmes basés sur SW; d'autre part, concevoir de nouveaux outils inspirés de SW afin d'élargir son champ d'applications et offrir une meilleure scalabilité.

Nous nous focalisons d'abord sur les estimateurs obtenus en minimisant SW, qui constituent la base de plusieurs méthodes de modélisation générative. Nous prouvons un ensemble de propriétés asymptotiques pour ces estimateurs, ainsi qu'un théorème central limite qui caractérise leur distribution asymptotique avec un taux de convergence indépendent de la dimension des données. Nous développons également une nouvelle technique d'inférence qui n'utilise pas la vraisemblance, appelée SW-ABC, en incorporant SW dans un algorithme de type *Approximate Bayesian Computation*. Nous prouvons des garanties asymptotiques sur la convergence de la distribution *a posteriori* calculée par SW-ABC, et illustrons les avantages de notre algorithme en pratique, sur des données synthétiques et un problème de débruitage d'image.

Par définition, SW est une espérance sur des projections aléatoires, qui est difficile à calculer en général et couramment estimée par une méthode de Monte Carlo simple. Cette approximation entraîne une erreur susceptible de dégrader les performances des algorithmes basés sur SW en grande dimension. Afin de pallier à ce problème, nous introduisons les *distances de Sliced-Wasserstein généralisées* (SWG), en étendant la définition des SW de façon à y inclure les *projections non linéaires* des distributions. Nous étudions dans quelle mesure SWG vérifient les axiomes d'une distance en utilisant la théorie autour de la *transformation de Radon*, et montrons que SWG peut fournir de meilleurs résultats que SW dans des applications de modélisation générative.

Nous adoptons ensuite une autre perspective pour résoudre les problèmes dus à l'estimation par Monte Carlo : nous tirons parti du phénomène de *concentration de mesure* qui affirme, selon des hypothèses modérées, que les projections linéaires unidimensionnelles d'une variable aléatoire à grande dimension suivent une loi presque gaussienne. A partir de ce résultat, nous proposons une nouvelle formule simple et déterministe pour calculer SW bien plus rapidement qu'avec Monte Carlo. Nous prouvons des garanties non-asymptotiques pour notre méthodologie sous une condition de dépendance faible, les validons sur des expériences synthétiques, puis utilisons notre nouvelle approximation pour la génération d'images.

Motivées par le succès grandissant de SW, de nouvelles *divergences "sliced"* (DS) ont été déployées pour des applications statistiques et d'apprentissage, mais leurs implications théoriques restent méconnues. Nous introduisons alors la première définition des DS et étudions leurs propriétés statistiques et topologiques. Notre analyse théorique met en lumière les conséquences du *slicing* : en particulier, nous prouvons que l'erreur d'approximation de tout DS par des échantillons ne dépend pas de la dimension des données, mais peut être affectée par l'approximation par Monte Carlo. Nous appliquons ensuite nos résultats à des DS spécifiques pour mieux les comprendre.

Remerciements

En premier lieu, je souhaite remercier mes directeurs de thèse, Roland, Alain et Umut. Je mesure pleinement ma chance d'avoir pu travailler avec chacun d'entre vous ces trois dernières années. Roland, je te suis avant tout très reconnaissante d'avoir accepté de diriger ma thèse et d'avoir veillé à ce qu'elle se déroule dans les meilleures conditions possibles du début à la fin. Merci également pour ton soutien et tous les retours que tu as pris le temps de me donner dès que j'avais besoin d'un avis, que ce soit pour ma recherche ou certaines formalités administratives. Alain, merci de m'avoir poussée à être la plus minutieuse possible dans mes preuves mathématiques, mes expériences numériques, mais aussi mes notations avec le fameux fichier `def.tex`. J'admire sincèrement ta rigueur scientifique, dont les exigences m'ont permis de sortir de ma zone de confort et d'améliorer mon esprit critique. Je tiens également à te remercier d'avoir été aussi présent et impliqué, notamment en période de rush : ta patience et tes conseils m'ont considérablement aidée à apprivoiser mon stress à l'approche de certaines deadlines. Umut, I am forever grateful to you for offering me the opportunity to pursue a PhD under your supervision. Thank you for teaching me so much about conducting a research project and for introducing me to many great people. Thank you, also, for sharing my obsession with details and my love for awkward jokes. More importantly, thank you for encouraging and pushing me, because you knew I could do it long before I did, while caring about my feelings. I couldn't have dreamt of a better mentor.

Merci à mes rapporteurs de thèse, Gabriel Peyré et Nicolas Courty, d'avoir pris le temps de relire attentivement mon manuscrit. I sincerely thank Julie Delon, Laetitia Chapel, Marco Cuturi and Justin Solomon, for accepting to be part of the jury. It has been an honor to defend my thesis in front of all of you, whose work and passion I greatly admire. Merci Marco d'avoir également accepté, un an et demi plus tôt, de compter parmi les membres du jury pour ma soutenance à mi-parcours ; et bien avant cela, de m'avoir accompagnée dans la procédure pour partir en stage de master au Japon, et de m'y avoir fait découvrir le transport optimal.

Many thanks to all the people I have been very fortunate to collaborate with during my PhD: Soheil Kolouri, Gustavo K. Rohde, Valentin De Bortoli, Lénaïc Chizat, Shahin Shahrampour, Tudor Manole, Ruben Ohana and Pierre E. Jacob. Valentin, je garde un bon souvenir de notre collaboration, qui nous a été particulièrement fructueuse. J'espère que nous aurons d'autres occasions de travailler ensemble à l'avenir. Soheil, I feel lucky to have worked with you on several research projects. Your immense enthusiasm and creativity made our collaboration even more enjoyable and have been a great source of inspiration for me. Pierre, merci d'avoir inspiré et relu avec intérêt mes premiers articles de thèse, puis d'avoir accepté de faire partie du jury de ma soutenance à mi-parcours. Je te suis très reconnaissance pour nos discussions autour de la recherche académique, ainsi que pour tes conseils qui m'ont poussée à me dépasser. Collaborer avec toi fut une belle manière de conclure ma thèse, et j'espère que nous pourrons un jour creuser les

autres idées que nous avons eues.

Merci à tous mes collègues du laboratoire LTCI de Télécom Paris, qui ont participé à créer une ambiance de travail conviviale, et à la chaire DSAIDIS d'avoir financé mon contrat doctoral. J'ai une pensée particulière à tous les membres et visiteurs du bureau B412 et de son annexe, grâce auxquels j'ai pu démarrer ma thèse dans un environnement amical et bienveillant. Merci à Amaury et Nidham pour ces soirées passées autour d'une IPA à discuter de la thèse sans filtre – je suis heureuse de vous avoir rencontrés et de vous compter parmi mes amis.

Je tiens à exprimer ma gratitude envers Claire Boyer et Julie Josse pour m'avoir offert l'opportunité de travailler avec elles après mon doctorat.

Enfin, je souhaite adresser mes sincères remerciements à Romain Laroche, avec qui j'ai écrit mon tout premier article scientifique à Montréal, quelques mois avant de commencer ma thèse à Paris. Merci Romain de m'avoir redonné confiance en ma capacité à faire de la recherche.

J'aimerais remercier mes amis, à qui je n'ai pas forcément parlé de ma thèse en détail, mais qui y ont contribué en vivant avec moi des moments de joie et d'insouciance.

Merci Karen et Anthony, les voisins de la rue Barrault chez qui je pouvais me réfugier pour déguster des tagliatelles au citron, discuter plus ou moins savamment de psychologie et rire (puis me battre avec Edward) pendant une partie de jeu de société.

Merci à mes amis de l'Ensimag, sur lesquels je pouvais compter pour aller prendre des bières et décompresser : ne perdez jamais le goût de la fête. Maxime, mon acolyte emo/indie, merci de partager ma passion pour les concerts, et donc aussi mon désespoir quand tout s'est arrêté. J'ai hâte de retourner au Primavera avec toi et le reste de la team pour se laisser à nouveau emporter. Dr. Bentriou et Dr. Mezghani, je vous remercie de m'avoir montré la voie, même si Dr. Mezghani perdait souvent la sienne le dimanche matin. Merci d'avoir partagé votre sagesse et de me faire autant rire.

Bien entendu, je n'oublie pas d'où je viens. Merci à Chloé, mon amie d'enfance également docteure, qui comprenait donc parfaitement ce que je traversais. Tant que la tradition Bookmas persiste, j'ai confiance en l'avenir. Edward, merci d'être présent dans ma vie depuis aussi longtemps et de toujours répondre à l'appel pour en fêter les grandes étapes. Je ne me lasserai pas de tes gâteaux, de nos débats sur la société ni de nos blagues, surtout quand elles ne font rire que nous.

Merci Guillaume pour tout le réconfort que tu as su m'apporter, même dans les périodes difficiles. Je me sens extrêmement chanceuse d'avoir rencontré un être aussi sensible et lumineux que toi.

Manon, je n'oublierai pas toutes ces soirées pendant lesquelles nous avons pu vivre des instants parfois absurdes, souvent hilarants : Paris était une fête avec toi. Merci pour ta présence rassurante, ton écoute et tes analyses pertinentes qui m'ont permis plus d'une fois d'ouvrir les yeux.

Merci Canelle de n'avoir jamais raté une occasion de me hyper et de m'avoir communiqué ta belle énergie quand j'en avais besoin. J'attends avec impatience le prochain volet de nos aventures sur la route, mais en attendant, continuons de nous retrouver à Paris ou à Nantes pour écouter des podcasts et débattre du sens de la vie tout en buvant des tisanes aux plantes mystérieuses.

Mes derniers remerciements s'adressent à ma famille, en particulier à ma sœur, mon frère, mon père et ma mère. Mes accomplissements n'auraient jamais été possibles sans les sacrifices et la générosité de mes parents : merci d'avoir tout entrepris, quitte à vous oublier vous-mêmes, pour favoriser mon épanouissement et me permettre de faire de longues études. Et bravo maman pour ton troisième doctorat.

Contents

Summary of Notations

Throughout this thesis, we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with associated expectation operator $\mathbb{E}$, on which all the random variables are defined.

$\mathcal{B}(\mathsf{Y})$	Borel set of Y
$\mathcal{P}(\mathsf{Y})$	Set of probability distributions supported on Y
$\mathcal{P}_p(\mathsf{Y})$	Set of distributions of $\mathcal{P}(\mathsf{Y})$ with finite p'th moment
δ_y	Dirac measure with mass on y
Leb_d	Lebesgue measure on $\mathbb{R}^d$
$\mathcal{N}(\mathbf{m}, \Sigma)$	Gaussian distribution with mean $\mathbf{m}$ and covariance matrix Σ
$\mu \otimes \nu$	Product measure of two probability distributions μ and ν
$\hat{\mu}_n$	Empirical distribution computed over $n \in \mathbb{N}^*$ samples i.i.d. from the probability distribution μ, $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$
F_μ	Cumulative distribution function of the probability distribution μ
F_μ^{-1}	Quantile function of the probability distribution μ
$\mathcal{F}[\mu]$	Fourier transform of the probability distribution μ
$f_\sharp \mu$	Push-forward measure of the probability measure μ by the measurable function f
$x \sim \mu$	x is a sample drawn from the probability distribution μ
$\|x\|$	Euclidean norm of the vector x
$\langle x, y \rangle$	Euclidean inner-product between the vectors x and y
$\mathrm{card}(\mathsf{X})$	Cardinal of the set X
$\mathrm{diam}(\mathsf{X})$	Diameter of the compact set X
$\mathrm{Tr}(\mathbf{A})$	Trace of the matrix $\mathbf{A}$
$\mathrm{Det}(\mathbf{A})$	Determinant of the matrix $\mathbf{A}$
$\|\mathbf{A}\|_F$	Frobenius norm of the matrix $\mathbf{A}$
$\mathbf{0}$	Vector in $\mathbb{R}^d$ whose d components are all equal to 0
$\mathbf{I}_d$	Identity matrix of size $d \times d$
$\nabla_z f$	Gradient of the function f with respect to its variable z
$\mathcal{F}[f]$	Fourier transform of the function f
$\|f\|_\infty$	Supremum norm of the function f
$\mathbb{S}^{d-1}$	Unit sphere on $\mathbb{R}^d$, $\mathbb{S}^{d-1} = \left\{\theta \in \mathbb{R}^d \,:\, \|\theta\| = 1\right\}$
$\mathcal{L}^p(\mathsf{X}, \mu)$	Set of functions whose p'th power is absolutely integrable with respect to the measure μ, $\mathcal{L}^p(\mathsf{X}, \mu) = \left\{f \,:\, \mathsf{X} \to \mathbb{R},\ \int_{\mathsf{X}} \lvert f(x)\rvert^p \,\mathrm{d}\mu(x) < \infty\right\}$
$\mathbb{M}(\mathsf{X})$	Set of real-valued measurable functions on X
$\mathbb{M}_b(\mathsf{X})$	Set of bounded functions of $\mathbb{M}(\mathsf{X})$
$\mathrm{B}_d(\mathbf{0}, R)$	Open ball in $\mathbb{R}^d$ of radius $R > 0$ centered around $\mathbf{0} \in \mathbb{R}^d$, $\mathrm{B}_d(\mathbf{0}, R) = \left\{x \in \mathbb{R}^d \,:\, \|x\| < R\right\}$

Abbreviations

ABC	Approximate Bayesian Computation
CDF	Cumulative Distribution Function
GMM	Gaussian Mixture Model
GSW	Generalized Sliced-Wasserstein (distance)
IGM	Implicit Generative Modeling
IPMs	Integral Probability Metrics
KL	Kullback-Leibler (divergence)
max-GSW	Maximum Generalized Sliced-Wasserstein (distance)
max-SW	Maximum Sliced-Wasserstein (distance)
ML	Machine Learning
MMD	Maximum Mean Discrepancy
OT	Optimal Transport
SPD	Sliced Probability Divergence
SSD	Sliced-Sinkhorn Divergences
SW	Sliced-Wasserstein (distance)
i.i.d.	independent and identically distributed
s.t.	such that
w.r.t.	with respect to

Chapter 1

Introduction

The significant breakthroughs achieved in the field of *machine learning* (ML) over the last decades have, in turn, raised many technical questions. In particular, building a successful machine learning algorithm remains an important problem, which is even more challenging given the evergrowing collection of data in various forms as well as the limits of available computational resources: the success of an algorithm is not only measured by the accuracy of the returned results, but also by its computational requirements and processing speed. As a result, many efforts have been made in machine learning research to improve existing methods so that, for example, they produce more accurate results, require less hyperparameters tuning, or execute faster, especially on large datasets.

The performance of machine learning algorithms depends on many design elements taking into account different factors, such as the formalism of the problem, the nature of data, and the computational resources at hand. This thesis focuses on a crucial component for many algorithms, that is the *probability divergence* used to compare two distributions. To further clarify our motivation, we consider a concrete problem, namely *generative modeling*, and illustrate the importance of having an "appropriate" divergence within this framework.

1.1 Motivation: Probability Divergences and Generative Models

The goal of generative modeling is to reproduce new data points, by learning a *probabilistic model* that best describes how the input dataset is generated. Specifically, *generative models* receive as input a set of $n \in \mathbb{N}^*$ observations, generally assumed to be independent and identically distributed (i.i.d.) samples from an unknown probability distribution $\mu_\star$, and aim at learning $\mu_\star$. Many different techniques can be employed to reach this objective, thus defining a variety of generative models.

A common practice consists in fitting $\mu_\star$ with a parametric distribution: one defines a *statistical model*, $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\}$, and finds the optimal parameters $\theta_\star \in \Theta$ such that $\mu_{\theta_\star} \in \mathcal{M}$ yields the most accurate approximation of $\mu_\star$. For instance, *Gaussian mixture models* (GMMs) propose to fit $\mu_\star$ with a mixture of $K \in \mathbb{N}^*$ Gaussian distributions (Figure 1.1). The parameters θ then refer to the mean and covariance matrix of each Gaussian, and mixing coefficients $\{\pi_k\}_{k=1}^K$, where π_k is the probability for the observations $\{x_i\}_{i=1}^n$ to be sampled from the k'th Gaussian distribution. While GMMs define simple and powerful generative models with many applications, their performance can be limited by the fact that a mixture of Gaussians does not necessarily describe well

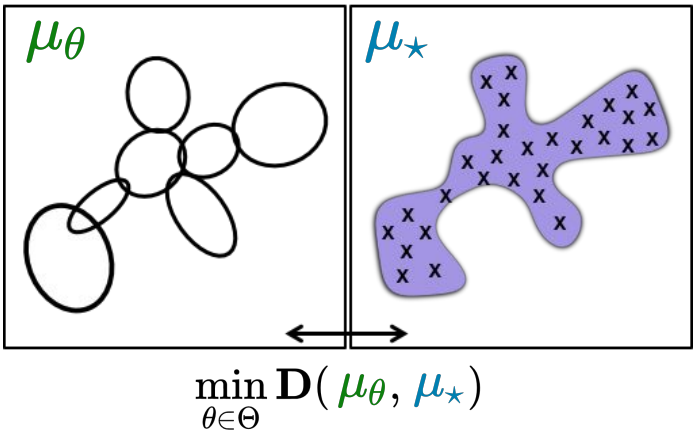


Figure 1.1: Illustration of Gaussian mixture models: the goal is to fit the true unknown distribution of the observations, $\mu_\star$, with a mixture of $K = 7$ Gaussian distributions, μ_θ. Figure courtesy of Soheil Kolouri (adapted from one of his presentations).

the structure of the observations.

To address the drawbacks of such "explicit" generative models, e.g. GMMs, the *implicit generative modeling* (IGM) methodology builds on the recent advances in *deep learning* and has attracted considerable attention within the ML community. The strategy, illustrated in Figure 1.2, consists in first choosing a distribution ζ from which it is easy to sample (e.g., a standard Gaussian distribution), then apply nonlinear differentiable operators which define a *(deep) neural network* with parameters θ, denoted by T_θ. The transformed distribution then corresponds to μ_θ. Once the neural network is trained, it can map $z \sim \zeta$, received as input, to new data points. The most common methods include generative adversial networks (GANs, Goodfellow et al. [2014]) and auto-encoders [Kingma and Welling, 2014]. Despite their ability to generate samples of high quality, such approaches inherit from the drawbacks of deep learning: designing and training a neural network can be highly difficult and time-consuming. More precisely, defining an implicit generative model amounts to choosing, for example, the nature of the nonlinear operators, the number of hidden layers, the latent distribution ζ or its objective function. To this day, why some architectures perform better than others is still an open question. On the other hand, the performance of implicit generative models also depends on how the similarity between $\mu_\theta \in \mathcal{M}$ and $\mu_\star$ is measured. We expand on this aspect in the remainder of this section, since the general intention of this thesis is to study a specific distance between probability distibutions.

As we described above, IGM defines a methodology to address the general problem of *density fitting*: the goal is to find a parametric distribution μ_θ that best approximates the true underlying distribution $\mu_\star$. In practical terms, this task requires having a tool that effectively measures how close one distribution is to another. *Probability divergences* are suited for the job, since they compute a certain notion of distance between two distributions. More formally, let X be a *Polish space*, *i.e.* a topological space that is separable and completely metrizable, and denote by $\mathcal{P}(\mathsf{X})$ the class of probability distributions supported on X; we will use the notation $\mathbf{D} : \mathcal{P}(\mathsf{X}) \times \mathcal{P}(\mathsf{X}) \to \mathbb{R}_+ \cup \{+\infty\}$ to refer to a generic divergence used to compare any two distributions in $\mathcal{P}(\mathsf{X})$. The objective of

density fitting is a *minimum distance estimation* problem, written as

$$\min_{\theta \in \Theta} \mathbf{D}(\mu_\theta, \mu_\star) \ .$$

In implicit generative modeling, μ_θ in (1.1) corresponds to the output of a deep neural network. Since there are several options available for the choice of $\mathbf{D}$, which will be presented later on, the question becomes, "*what probability divergence should one use in their generative models?*".
As a first answer, we provide general informal guidelines by enumerating some of the main desirable properties for $\mathbf{D}$.

(P1) $\mathbf{D}$ is a "proper" distance function, *i.e.* it verifies the *metric axioms*.

(P2) $\mathbf{D}$ satisfies a certain notion of *continuity*, which can be described as follows: if a sequence of distributions $(\mu_k)_{k\in\mathbb{N}}$ "gets closer" to μ as k grows to infinity, then $\mathbf{D}(\mu_k, \mu)$ shrinks to 0.

(P3) $\theta \mapsto \mathbf{D}(\mu_\theta, \mu_\star)$ is *differentiable* and has a *unique minimizer*.

(P4) $\mathbf{D}(\mu, \nu)$ can be *effectively estimated from samples* drawn from μ and ν.

We will give the formal statement for the metric axioms **(P1)** and the continuity **(P2)** in Section 2.1. **(P4)** is motivated by the fact that in most problems in data sciences, one has access to finite sets of observations instead of the underlying probability distributions. Therefore, any distribution ξ is only accessible through an empirical approximation, which usually corresponds to a discrete measure $\hat{\xi}_n$ computed over the sequence of $n \in \mathbb{N}^*$ random variables $\{Y_i\}_{i=1}^n$ i.i.d. from ξ, given by

$$\hat{\xi}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i} \ ,$$

with δ_Y being the Dirac measure with mass on the point Y. In that case, the chosen probability divergence $\mathbf{D}$ must have specific practical features in order to appropriately handle the empirical approximations, which will be discussed in Section 2.1.

Overall, properties **(P1)** to **(P4)** reflect that $\mathbf{D}$ should be a sufficiently regular and computationally practical divergence, so that the associated generative model is easy to use, somewhat robust and able to capture relevant information regarding the geometry of the problem.

1.2 Classical Probability Divergences

We now present traditional choices of probability divergences and explain their implications in terms of the criteria listed in the previous section. If $\mathbf{D}$ satisfies **(P2)**, we say that $\mathbf{D}$ is *weakly continuous* and report the associated mathematical definition in Section 2.1.

f-divergences. These divergences were introduced in [Rényi, 1961] and also referred to as *φ-divergences*, *Ciszár divergences* [Ciszár, 1967] or *Ali-Silvey distances* [Ali and Silvey, 1966]. They include widely known instances which have played a crucial role in various areas such as probability theory, statistics and information theory. For example,

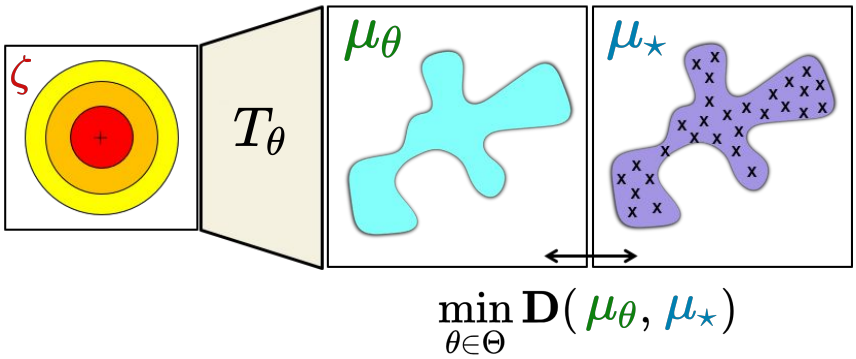


Figure 1.2: Illustration of implicit generative modeling: ζ is a distribution easy to sample from (e.g., a standard Gaussian distribution), which is then mapped to a more complicated distribution μ_θ by applying the parametric map T_θ (e.g., a deep neural network). The goal is to fit the true unknown distribution of the observations, $\mu_\star$, with μ_θ. Figure courtesy of Soheil Kolouri (adapted from one of his presentations).

the χ^2-*divergence* has commonly been used for adaptive importance sampling [Cornebise et al., 2008]; the *Kullback-Leibler divergence* (KL) is related to the notions of *mutual information* and *relative entropy* [Cover and Thomas, 2006, Section 2.3] and is the core component of *variational inference* [Blei et al., 2017] and more recently, of *variational auto-encoders* [Kingma and Welling, 2014].

In the context of generative modeling, f-divergences suffer from important drawbacks. First, they do not metrize weak convergence, which might severely affect the quality of the comparisons made. This is illustrated by the fact that KL evaluated between μ and ν is infinite when the two distributions μ and ν are supported on domains that do not overlap. On the other hand, approximating a f-divergence, denoted by $\mathbf{D}_f$, from a set of samples is not computationally simple: the straightforward estimator $\mathbf{D}_f(\hat{\mu}_n, \hat{\nu}_n)$ obtained by plugging the empirical measures in place of μ and ν does not converge to $\mathbf{D}_f(\mu, \nu)$ as n increases in general. Several workarounds have then been developed, but they solve provably hard problems, induce a sample complexity with slow convergence rates, or rely on strong structural assumptions: see the discussion in [Rubenstein et al., 2019].

Integral Probability Metrics (IPMs). Introduced by Müller [1997], the class of IPMs provides important advantages over f-divergences for generative modeling: they satisfy almost all metric axioms, are weakly continuous under mild assumptions, and can easily be estimated from the available samples. More details on these aspects are provided in Section 2.2.

A popular example of IPMs is given by the *Maximum Mean Discrepancy* (MMD, Gretton et al. [2012]), whose attractive analytical and computational properties make it an interesting probability divergence for many applications, especially statistical hypothesis testing [Gretton et al., 2012] and generative modeling [Li et al., 2015, Dziugaite et al., 2015, Sutherland et al., 2017, Bińkowski et al., 2018, Arbel et al., 2019]. However, the training and performance of these generative models are highly sensitive to the kernel function defining MMD and its parametrization. For example, the Gaussian RBF kernel is a very traditional choice, but its bandwith parameter determines the statistical

efficiency of the associated MMD and is not easy to tune appropriately [Li et al., 2015]. Besides, its derivatives decay exponentially which cause important stability issues when training deep neural networks, as discussed in Arbel et al. [2019].

A second important instance of IPMs corresponds to the *Wasserstein distance of order 1* [Villani, 2008, Theorem 5.10]. This metric also falls into the category of *optimal transport metrics*, which is presented next.

Optimal transport divergences. *Optimal transport* (OT) is a mathematical theory on a specific optimization problem, which has been extensively studied and leveraged in various applied fields such as economics, combinatorial optimization and more recently, data sciences. Specific formulations of the OT problem define a family of powerful probability divergences: the *Wasserstein distances*. These metrics, which will be formally presented in Section 2.4, are able to capture key information for comparing two distributions μ and ν, since they rely on a cost function that carries relevant geometric properties of the supports of μ and ν. If the cost function c is the p'th power of the Euclidean distance, *i.e.* for any $x, y \in \mathbb{R}^d$, $c : (x, y) \mapsto \|x - y\|^p$ with $p \in [1, +\infty)$, then the resulting Wasserstein distance is known as the *Wasserstein distance of order* p and denoted by $\mathbf{W}_p$. This divergence satisfy all metric axioms and is weakly continuous when evaluated on the space of probability distributions with finite p'th moment [Villani, 2008, Chapter 6]. Therefore, the Wasserstein distance of order p can effectively compare any two distributions even when their supports do not overlap, as opposed to f-divergences, and does not require tedious hyperparameter tuning, unlike MMD.

However, it is well known that Wasserstein distances suffer from important computational and statistical limitations, especially in high dimensions. Indeed, their computation is expensive, apart in some special settings presented in Section 2.4.1, for example when comparing two probability distributions supported on $\mathbb{R}$. Let us assume the following typical scenario in machine learning: one would like to compare two distributions μ and ν supported on $\mathbb{R}^d$, but only observe n samples drawn from them. In general, Wasserstein distances are not analytically available but can be estimated with approximate solvers, which tend to have a super-cubic cost in practice and executes in $\mathcal{O}(n^3 \log(n))$ in the worst case. This implies that a single evaluation of Wasserstein distances is computationally demanding, especially on large-scale datasets.

Besides, the Wasserstein distance computed from the empirical approximations $\hat{\mu}_n$ and $\hat{\nu}_n$ has been shown to converge to the true value, *i.e.* the Wasserstein distance between μ and ν, with a rate in $\mathcal{O}(n^{-1/d})$: see our discussion in Section 2.4.2. Therefore, for high values of the ambient dimension d, the estimates computed from n samples are reasonably accurate provided that n is sufficiently large, which induces an important computational complexity according to the previous paragraph.

Hence, the computational and statistical limitations of Wasserstein distances have considerably prevented their application in data sciences for a long time, in particular in generative modeling, which was studied solely from a theoretical perspective [Bassetti et al., 2006]. Nevertheless, in recent years, number of studies have introduced various methods to alleviate the practical issues of Wasserstein distances and thus managed to expand the domain of applicability of OT. These techniques define the active research topic of *Computational Optimal Transport* [Peyré and Cuturi, 2019], and their development was made possible by the progress in optimization and large-scale machine learning. For instance, Wasserstein distances have recently been incorporated within the IGM framework, leading to the formulation of novel effective models [Arjovsky et al., 2017, Bousquet et al., 2017, Gulrajani et al., 2017, Tolstikhin et al., 2018].

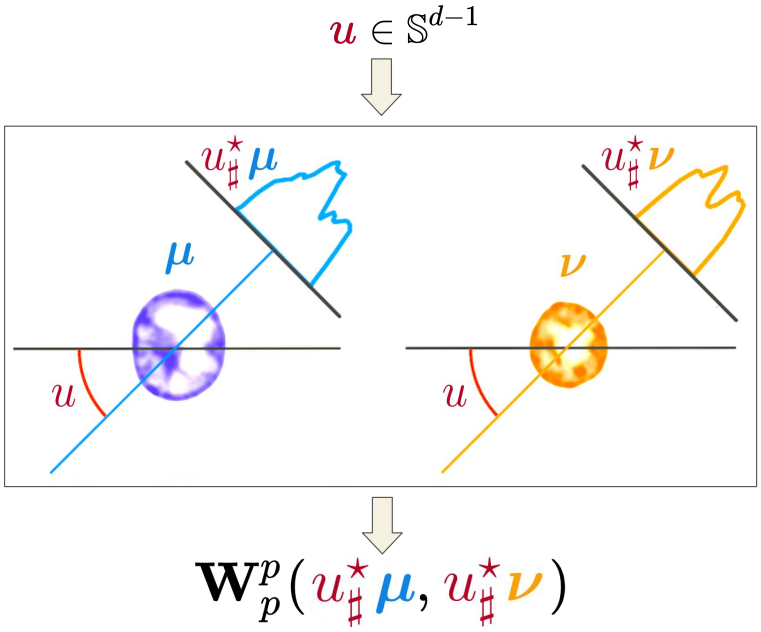


Figure 1.3: Illustration of the Sliced-Wasserstein distance of order $p \in [1, +\infty)$. The two distributions to compare, μ and ν, are projected along any direction u on the unit sphere $\mathbb{S}^{d-1}$. This gives the univariate distributions denoted by $u_\sharp^\star \mu$ and $u_\sharp^\star \nu$, which are then compared with the Wasserstein distance of order p. SW is finally defined as $\mathbb{E}[\mathbf{W}_p^p(u_\sharp^\star \mu, u_\sharp^\star \nu)]$ where the expectation is computed over u uniformly distributed on $\mathbb{S}^{d-1}$. Figure courtesy of Soheil Kolouri (adapted from [Kolouri et al., 2017]).

The field of computational optimal transport also gave rise to the formulation of alternative metrics to the Wasserstein distance, such as *Sinkhorn divergences*, which emerge from the regularization of the OT problem [Cuturi, 2013], and the *Sliced-Wasserstein distance*. We dedicate the next section to the latter distance, as it constitutes the main focus of this thesis, and will discuss the former in Section 2.5.

1.3 Focus on the Sliced-Wasserstein Distance

While regularized OT and Sinkhorn divergences have strongly help OT theory gain immense interest from the machine learning community, another alternative to classical OT has become increasingly popular in the last few years: the Sliced-Wasserstein distance (SW).

In this section, we explain the definition of SW and its practical implications; the formal statements will be given in Section 2.6. We then discuss related work to review the existing theoretical properties and applications of SW, as well as motivate why this metric is the central object of this thesis. In what follows, $\mathbf{SW}_p$ with $p \in [1, +\infty)$ denotes the *Sliced-Wasserstein of order p* (Definition 2.9).

SW was first introduced by Rabin et al. [2012] and Bonneel et al. [2015] as a practical

alternative to the Wasserstein distance to speed up the computation of barycenters of measures. The idea behind SW is to offer computational efficiency by leveraging the analytical form of the Wasserstein distance between univariate distributions: SW compares two multivariate probability distributions by first, obtaining a family of representations in $\mathbb{R}$ for each distribution through projections, then computing the expected value of the Wasserstein distance between these univariate representations. We illustrate the Sliced-Wasserstein distance in Figure 1.3 and provide its definition in the caption of that figure.

The expectation that defines SW does not admit an analytical expression in general and is thus commonly estimated with a simple Monte Carlo average based on $L \in \mathbb{N}^*$ samples. In other words, computing SW amounts to solving a finite number of one-dimensional OT problems, which can conveniently be done in closed-form. This approximation technique has a complexity in $\mathcal{O}\big(Ldn + Ln\log(n)\big)$, thus can provide significant computational benefits over Wasserstein distances.

Many methods building on the Sliced-Wasserstein distance have been developed to address various applied problems in a computationally efficient way, such as in machine learning, imaging and statistics. Barycenters of measures based on SW have successfully been used on several image processing tasks, *i.e.* color transfer, texture synthesis and texture mixing [Rabin et al., 2012, Bonneel et al., 2015]. Very recently, these barycenters have been extended to the case where SW compares more than two measures [Cohen et al., 2021].

Novel kernel functions based on SW have also been proposed, and their benefits have been illustrated in topological data analysis and pattern recognition tasks, e.g., classification and clustering [Kolouri et al., 2016, Carrière et al., 2017].

Lastly, an important body of literature is dedicated to applying SW in generative modeling applications: in [Karras et al., 2017], SW serves as a score to evaluate the performance of GANs; [Kolouri et al., 2018] build a novel GMM which learns the parameters of Gaussian distributions by minimizing SW; finally, SW forms the basis of several new IGM models, including auto-encoders and GANs [Deshpande et al., 2018, Liutkus et al., 2019, Wu et al., 2019, Kolouri et al., 2019b, Dai and Seljak, 2021].

On the theoretical side, SW has been shown to satisfy the metric axioms [Bonnotte, 2013, Proposition 5.1.2], is always bounded above by the Wasserstein distance [Bonnotte, 2013, Proposition 5.1.3], and both metrics are equivalent when computed between compactly supported probability measures [Bonnotte, 2013, Theorem 5.1.5]. The equivalence between the Wasserstein distance and SW has further been studied in a very recent study [Bayraktar and Guo, 2021]. A mathematical analysis of gradient flows based on SW is provided in [Bonnotte, 2013, Chapter 5] and has been used to derive the theoretical guarantees of the IGM methodology proposed in [Liutkus et al., 2019]. Finally, recent work observed the statistical benefits of SW in practice, and consequently, derived a concentration inequality to bound $|\mathbf{SW}_2(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SW}_2(\mu, \nu)|$ with high probability for the specific case where μ, ν are Gaussian [Deshpande et al., 2019].

Hence, the Sliced-Wasserstein distance has established itself as a powerful practical metric for data sciences, due to its connection to OT, computational efficiency and flexibility. Besides, SW shares certain properties with the Wasserstein distance, while providing statistical advantages on very specific settings. The contributions to the theoretical analysis of SW are nevertheless limited, which contrasts with the availability of

numerous empirical studies. This imbalance forms the starting point of this thesis.

1.4 Outline and Contributions

Computational optimal transport has become a very active and popular field within the machine learning community in recent years, and the growing literature on the subject has demonstrated that SW is a practical tool with broad application. Motivated by the reported empirical success of SW and its increasing popularity, this thesis further studies this metric, with particular focus on its theoretical guarantees, relevance to approximate inference and generative modeling, and extensions. The broad objective is to conduct a thorough analysis of the theoretical and empirical implications of SW, which helps unlock its full potential on modern machine learning problems. More precisely, our contributions revolve around the following research directions.

Objective 1. While the empirical performance of SW has been analyzed on a wide range of practical tasks, there has been little work regarding its theoretical guarantees. This implies that most SW-based methods are not sufficiently theoretically grounded. Our first goal is then to bridge this gap by investigating the theoretical properties of SW and their repercussions in practice.

Objective 2. Since the literature on SW is quite recent, we believe that there are other applied problems where this metric finds relevance. We thus explore how SW can be used to design novel methods for tasks that have not been addressed in prior work, such as Bayesian inference.

Objective 3. For the sake of providing a balanced analysis, we examine the known major limitation of SW, caused by its Monte Carlo approximation: the computational efficiency induced by this method is offset by an approximation error, which can be important depending on the problem at hand. We further illustrate this issue and develop new techniques to mitigate it, by leveraging existing tools from other fields.

Objective 4. Our last objective aims at better understanding an essential component of SW, that is the *slicing* operation: we step back from the OT paradigm and study from a theoretical perspective the consequences of slicing *any* divergence other than the Wasserstein distance. This analysis will allow us to broaden the reach of SW, as well as sharpen its theoretical analysis.

We now present an overview of the organization of this thesis: for each chapter, we explain the motivation, present related work, and summarize our key findings which are in line with the aforementioned objectives.

Chapter 3: Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance

Let us consider the *minimum distance estimation* (MDE, Wolfowitz [1957], Basu et al. [2011]) problem, formally defined as

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{D}(\hat{\mu}_n, \mu_\theta) \,, \tag{1.1}$$

where $\mathbf{D}$ denotes a divergence between probability measures, Θ is the parameter space, μ_θ is a probability measure indexed by $\theta \in \Theta$, and $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{Y_i}$ is the empirical distribution of a set of i.i.d. observations $\{Y_i\}_{i=1}^n$. MDE has been extremely useful in statistical inference to infer the parameters of a distribution whose analytical expression is unknown [Basu et al., 2011], and have inspired the formulation of several IGM strategies. In that context, optimal transport metrics have become increasingly popular due to their attractive theoretical properties: the *minimum Wasserstein estimator* [Bassetti et al., 2006, Bernton et al., 2019], obtained by replacing $\mathbf{D}$ in (1.1) with $\mathbf{W}_p$, forms the basis of popular IGM algorithms [Arjovsky et al., 2017, Genevay et al., 2017, Tolstikhin et al., 2018]. Motivated by its practical success, the theoretical properties of this estimator have been studied [Bousquet et al., 2017, Liu et al., 2017] and very recently, Bernton et al. [2019] have derived a set of asymptotic properties, including the asymptotic distribution of the estimator when the distributions are supported on $\mathbb{R}$.

Due to the computational limitations induced by $\mathbf{W}_p$, the computational complexity of minimum Wasserstein estimators rapidly becomes excessive with the increasing problem dimension. To avoid this problem, several practical alternatives to the Wasserstein distance have been proposed, and in particular, the Sliced-Wasserstein distance has been an increasingly popular metric, including in IGM [Deshpande et al., 2018, Liutkus et al., 2019, Wu et al., 2019, Kolouri et al., 2019b]. However, since the theoretical properties of these estimators had not been established, the techniques based on them are not sufficiently theoretically grounded.

Summary of our contributions in Chapter 3. To further motivate the use of SW in statistical inference, we investigate the asymptotic properties of the *minimum Sliced-Wasserstein estimators*, which are obtained by replacing $\mathbf{D}$ in (1.1) with $\mathbf{SW}_p$. Our theoretical contributions are summarized below.

1. We prove that *convergence under SW implies weak convergence of probability measures* on general domains.

2. We show, under some assumptions, that *minimum SW estimators exist, are measurable, and are consistent* in the sense that as the number of observations n increases, in well-specified models, the estimates will converge to the parameters that generated the observed dataset. Similar consistency guarantees hold for misspecified models.

3. We finally derive a *central limit theorem* which characterizes the asymptotic distribution of minimum SW estimators, and establishes a convergence rate of $\sqrt{n}$ for *any* finite dimension.

Our work is inspired by [Bernton et al., 2019] and the adaptation of their techniques was made possible by the identification of novel properties regarding the topology induced by SW: for example, we established for the first time that convergence in SW implies weak convergence of probability measures, which generalizes the results given in [Bonnotte, 2013]. Besides, our CLT is stronger than the analogous one derived in [Bernton et al., 2019] for minimum Wasserstein estimators, since ours is not restricted to one-dimensional data, and our convergence rate in $\sqrt{n}$ is valid for any dimension.

Our theoretical findings are supported with experiments that we conducted on synthetic and real data. We first consider a classical statistical inference problem, where the statistical models are characterized by a Gaussian or a multidimensional α-stable distribution. In both models, the experiments validate our consistency and CLT results. We also illustrate that, as expected, the minimum SW estimators have significantly better computational properties as compared to the minimum Wasserstein estimators, especially on high-dimensional problems. Finally, we consider the deep generative model in [Deshpande et al., 2018] and conduct an empirical analysis on the MNIST dataset, which confirms the consistent behavior of minimum SW estimators in IGM applications.

Chapter 4: Approximate Bayesian Computation with the Sliced-Wasserstein Distance

We consider the problem of estimating the posterior distribution of some model parameters $\theta \in \Theta$ given n data points $y_{1:n} = (y_1, \dots, y_n)$. By the Bayes' theorem, this distribution has a closed-form expression which depends on the likelihood $\pi(y_{1:n}|\theta)$. For many statistical models of interest, $\pi(y_{1:n}|\theta)$ cannot be numerically evaluated in a reasonable amount of time, which prevents the application of classical likelihood-based approximate inference methods.

Nevertheless, in various settings, it is possible to generate data from the likelihood given any parameter value θ. This generative setting gave rise to the popular likelihood-free framework for approximate inference, called *Approximate Bayesian Computation* [Tavaré et al., 1997, Beaumont et al., 2002], which has proven useful in various practical applications, e.g. in ecology [Wood, 2010] and biology [Tanaka et al., 2006]. ABC approximates the exact posterior $\pi(\theta|y_{1:n})$ from the parameter values for which the synthetic data $z_{1:m}$ generated from the likelihood are close enough to the observations $y_{1:n}$. Closeness is usually measured with a discrepancy measure between the two datasets reduced to some "summary statistics" (e.g., empirical mean or empirical covariance). The quality of the approximate posterior distribution highly depends on these summaries, and finding relevant statistics is a non-trivial and tedious task.

Recently, discrepancy measures that view data sets as empirical probability distributions to eschew the construction of summary statistics have been proposed for ABC. Examples include the Kullback-Leibler divergence [Jiang et al., 2018], maximum mean discrepancy [Park et al., 2016], and Wasserstein distance (WABC, Bernton et al. [2019]). While the Wasserstein distance seems like a relevant choice of discrepancy in that context thanks to its strong theoretical properties, its computational and statistical issues can strongly affect the performance of WABC applied to high-dimensional data.

Summary of our contributions for Chapter 4. Motivated by the computational efficiency of SW and its successful performance in generative settings, we develop a novel framework for likelihood-free approximate Bayesian inference, which estimates the posterior by retaining the parameter values for which

$$\mathbf{SW}_p(\hat{\mu}_n, \hat{\nu}_m) \leq \varepsilon \,, \tag{1.2}$$

where $\hat{\mu}_n$ denotes the empirical distributions of the observations $y_{1:n}$, $\hat{\nu}_m$ is the empirical distribution of the samples $z_{1:n}$ drawn from the likelihood, and $\varepsilon > 0$ is a tolerance threshold. This results in a novel ABC method, called Sliced-Wasserstein ABC (SW-ABC), which does not require choosing summary statistics. Besides, SW-ABC is more

efficient than WABC on high-dimensional settings, thanks to the computational and statistical benefits of SW.

We show that SW-ABC comes with guarantees on the convergence of the resulting posterior, under two asymptotic regimes.

1. On the one hand, we prove that for a fixed set of observations $y_{1:n}$, the posterior approximated by SW-ABC converges to the true posterior as ε goes to 0, under specific assumptions on the density used to generate synthetic data.

2. On the other hand, we study the SW-ABC posterior when the value of ε is kept fixed and the number of observations grows. We show that, as n goes to $+\infty$, the approximate posterior converges to the prior distribution on θ restricted to a specific subset of Θ that depends on ε.

We then illustrate on a synthetical problem the superior empirical performance of SW-ABC against existing ABC techniques based on other divergences. We also demonstrate the flexibility and computational advantages of our methodology by designing a novel algorithm for image denoising, which corresponds to a combination of SW-ABC with a widely used technique for this task, Non-Local Means [Buades et al., 2005].

Chapter 5: Generalized Sliced Wasserstein Distances

By definition, the Sliced-Wasserstein distance measures the dissimilarity between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ by comparing "linear" projections of μ and ν along all possible directions on $\mathbb{S}^{d-1}$. These projections correspond to the push-forward measures $u^\star_\sharp \mu$ and $u^\star_\sharp \nu$ for any $u \in \mathbb{S}^{d-1}$, and are actually closely related to the Radon transform [Rabin et al., 2012, Proposition 6], which is widely used in tomography [Radon, 1917, Helgason, 2011]. SW does not admit an analytical formula in general and is thus commonly estimated by Monte Carlo: in practice, one approximates the expectation over $u \sim \boldsymbol{\sigma}$ in (2.20) with an average over a finite number of directions $\{u_l\}_{l=1}^L$, where for $l' \in \{1, \dots, L\}$, $u'_l \sim \boldsymbol{\sigma}$.

Previous empirical studies have reported that this Monte Carlo strategy might degrade the performance of SW-based algorithms on high-dimensional settings because of the induced approximation error, and propose to change the nature of the projections to overcome this problem. For instance, in [Rowland et al., 2019, Wu et al., 2019], SW is estimated with a Monte Carlo average based on a finite number of *orthogonal* projection directions. An alternative OT metric called the "*maximum Sliced-Wasserstein distance*" is introduced in [Deshpande et al., 2019], and extends SW by replacing the expectation with a maximum operator so that one retains the "most informative" projection direction. The information returned by a direction $u \in \mathbb{S}^{d-1}$ is measured by $\mathbf{W}_p(u^\star_\sharp \mu, u^\star_\sharp \nu)$: the larger this Wasserstein distance, the more informative u. Paty and Cuturi [2019] generalizes this idea by considering k-dimensional projections of μ, ν with $k \in \{1, \dots, d\}$: the goal is then to find the most informative *subspace* on which μ and ν are being projected. While these methods reduce the computational cost by requiring a lower number of projections, they incur an additional cost due to the resolution of a non-convex optimization problem over manifolds.

Summary of our contributions in Chapter 5. In this chapter, we take an alternative route to alleviate the inefficiencies caused by the Monte Carlo approximation of SW, by assuming that the linear nature of the projections might not guarantee an efficient

evaluation of SW. Indeed, in very high-dimensional settings, the data often lives in a thin manifold, so the number of randomly chosen linear projections required to capture the structure of the data distribution grows very quickly.

We introduce a novel class of divergences between probability distributions, called *generalized Sliced-Wasserstein distances* (GSW). GSW are defined as SW, except they compute different types of projections: to compare μ and ν, one chooses a function $g^u : \mathbb{R}^d \to \mathbb{R}$ defined for any $u \in \mathbb{R}^q$ ($q \in \mathbb{N}^*$), and collect the pushforward measures $(g^u)_\sharp\mu, (g^u)_\sharp\nu$ for any $u \in \mathbb{R}^q$. GSW is then defined as the expected value of the Wasserstein distance between these one-dimensional representations. As the name suggests, GSW generalizes the concept behind SW: when $g^u = u^\star$, its definition boils down to the Sliced-Wasserstein distance. We also use these non-linear projections to extend the maximum Sliced-Wasserstein distance and introduce the *maximum generalized Sliced-Wasserstein distances* (max-GSW).

Analogously to (max-)SW, our definition of (max-)GSW is connected with the Radon transform: by using the theory of the *generalized Radon transform* (GRT, Beylkin [1984]), we identify some regularity conditions g^u needs to satisfy to ensure that the resulting generalized distance is well-defined. We also prove that GSW and max-GSW verify all metric axioms if and only if the GRT they rely on is injective; otherwise, they are pseudo-metrics. This result helps us identify useful instances of g^u that guarantee the injectivity of the associated GRT.

Then, we demonstrate that GSW and max-GSW can outperform SW and max-SW in several generative modeling applications, with both synthetic and real data: the inherent non-linearity of the one-dimensional representations used in (max-)GSW seems to capture the complex structure of high-dimensional distributions with much less projections. Besides, to ensure an automatic tuning of g^u, we propose to define it as a neural network. This scheme brings practical advantages and an interesting perspective on adversarial generative modeling, showing that such algorithms contain an implicit stage for learning projections with different cost functions than ours.

Chapter 6: Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections

The motivations of this chapter are the same as in Chapter 5: although SW has been shown to offer important computational and theoretical advantages over the Wasserstein distance, its practical performance can be limited by the error induced by the commonly used Monte Carlo approximation. To ensure that this approximation error is reasonably small, one might need to choose a large number of projections L, which inevitably increases the computational complexity of SW. We thus aim at developing an alternative method to overcome this issue.

We adopt a different perspective to approximate SW by leveraging *concentration results on random projections*: under relatively mild conditions, the typical distribution of low-dimensional projections of high-dimensional random variables is close to some Gaussian law [Sudakov, 1978, Diaconis and Freedman, 1984]. This result has recently been illustrated with a bound in terms of the Wasserstein distance [Reeves, 2017, Theorem 1]: let $\{X_i\}_{i=1}^d$ be a sequence of real random variables with distribution μ_d, such that $X_1, \dots, X_d$ are independent with finite fourth-order moments; then, $\mathbb{E}[\mathbf{W}_2^2(u_\sharp^\star\mu, \mathcal{N}_\mu)^2]$

goes to zero as d increases, where $\mathcal{N}_\mu$ denotes a univariate Gaussian distribution whose variance depends on μ_d, and the expectation is taken with respect to a Gaussian variable u. This result has very recently been used to bound the "maximum-sliced distance" between any probability measure and its Gaussian approximation [Goldt et al., 2021].

Summary of our contributions in Chapter 6. We develop a novel technique that approximates SW with a simple *deterministic* formula, which builds on [Reeves, 2017, Theorem 1]. As opposed to Monte Carlo, our methodology does not rely on a finite set of random projections, thus eliminates the need of tuning the hyperparameter L and can lead to a significant computational time reduction.

The formulation of our approximate SW is supported by the following findings.

1. We define an alternative SW whose projection directions are drawn from the same Gaussian distribution as in [Reeves, 2017], instead of uniformly on $\mathbb{S}^{d-1}$. We establish its relation with the original SW, and in particular, we prove that the two distances are equal when $p = 2$.

2. We use this alternative SW and [Reeves, 2017, Theorem 1] to bound the absolute difference between SW applied to any two probability measures μ_d, ν_d on $\mathbb{R}^d$ and the Wasserstein distance between the univariate Gaussians $\mathcal{N}_{\mu_d}$, $\mathcal{N}_{\nu_d}$. We explain why the mean parameters of μ_d and ν_d should necessarily be zero for the absolute difference to decrease as d grows.

3. We show that the requirement on the mean parameters is not a practical problem, by proving the following result: SW between μ_d, ν_d can be equivalently written as the sum of the *difference between their means* and the SW between the *centered versions* of μ_d, ν_d.

Based on the aforementioned results, we introduce a novel estimate of SW, defined as

$$\widehat{\mathbf{SW}}_2^2(\mu_d, \nu_d) = \frac{1}{d}\|\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d}\|^2 + \mathbf{W}_2^2(\mathcal{N}_{\bar{\mu}_d}, \mathcal{N}_{\bar{\nu}_d}) \ , \tag{1.3}$$

where for $\xi \in \{\mu_d, \nu_d\}$, $\mathbf{m}_\xi$ is the mean parameter of ξ and $\bar{\xi}$ denotes the centered version of ξ. Since the Wasserstein distance between Gaussian distributions admits a closed-form solution, (1.3) is very easy to compute, and faster than the Monte Carlo estimate obtained with a large number of projections.

We derive nonasymptotical guarantees on the error induced by our approach: we define a weak dependence condition that is weaker than the one in [Doukhan and Neumann, 2007], which is a notion commonly used in statistics, and prove that under this condition, $|\mathbf{SW}_2(\mu_d, \nu_d) - \widehat{\mathbf{SW}}_2(\mu_d, \nu_d)|$ goes to zero as d grows to $+\infty$.

The nonasymptotical guarantees of our approximate SW, as well as its computational efficiency, are then validated with experiments conducted on synthetic data. We finally leverage our theoretical insights to design a novel adversarial framework for a typical generative modeling problem, namely image generation. As compared to generative models based on SW estimated with Monte Carlo, our framework produces images of higher quality with further computational benefits.

Chapter 7: Statistical and Topological Properties of Sliced Probability Divergences

Recent years have witnessed the formulation of novel sliced probability divergences (SPDs), such as Sliced Gromov-Wasserstein [Vayer et al., 2019] or Sliced-Cramér [Kolouri et al., 2020a], due to the success of the Sliced-Wasserstein distance in generative modeling. We identify two reasons why "slicing" a divergence is beneficial. First, some probability divergences are only defined to compare measures supported on one-dimensional spaces, for instance the Cramér distance [Cramér, 1928]. The slicing operation thus extends these divergences to multivariate distributions [Knop et al., 2020, Kolouri et al., 2020a]. Then, slicing leverages the computational advantages available in one dimension to define divergences achieving computational efficiency on multivariate settings [Rabin et al., 2012, Deshpande et al., 2019, Paty and Cuturi, 2019, Kolouri et al., 2019b, Vayer et al., 2019].

Even though various sliced divergences have successfully been deployed in practical applications, their theoretical properties have not yet been well understood. Indeed, the literature on such divergences has largely been devoted to the study of SW. Besides, some properties of SW have only been characterized for specific settings, in particular its statistical benefits observed in practice [Deshpande et al., 2018, 2019].

Summary of our contributions in Chapter 7. We conduct a theoretical analysis on sliced probability divergences, in order to bridge the gap between theory and practice and gain insight on what the slicing operation itself is bringing. To this end, we formulate the first general definition of the class of SPDs, which includes existing instances in the literature and enables us to adopt a general point of view. Specifically, we consider a generic base divergence $\boldsymbol{\Delta}$ between one-dimensional probability measures, and define its sliced version, denoted by $\mathbf{S}\boldsymbol{\Delta}$, which operates on multivariate settings.

We first prove several results on the topology induced by $\mathbf{S}\boldsymbol{\Delta}$, whose statements can be summarized as follows.

1. If $\boldsymbol{\Delta}$ is a metric, so is $\mathbf{S}\boldsymbol{\Delta}$. In other words, slicing preserves the metric properties.

2. If the convergence in $\boldsymbol{\Delta}$ implies the weak convergence of measures (or conversely), then slicing preserves this property, *i.e.* the convergence in $\mathbf{S}\boldsymbol{\Delta}$ implies the weak convergence of measures (or conversely).

3. If $\boldsymbol{\Delta}$ is an integral probability metric, $\boldsymbol{\Delta}$ and $\mathbf{S}\boldsymbol{\Delta}$ are strongly equivalent under specific sufficient conditions which we identify.

We also study the statistical properties of $\mathbf{S}\boldsymbol{\Delta}$.

4. We show that the sample complexity of $\mathbf{S}\boldsymbol{\Delta}$ is proportional to the sample complexity of $\boldsymbol{\Delta}$ for one-dimensional measures. Therefore, the sample complexity of *any* SPDs does not depend on the dimension d.

5. We also derive a bound on the error made when estimating divergences with Monte Carlo, which is the most common practice: we prove that this approximation scheme induces an additional variance term in the complexity of computing the sliced divergence.

Hence, our results demonstrate that while SPDs can offer important statistical benefits thanks to the dimension-free rate in their sample complexity, the Monte Carlo strategy induces an error that might affect the overall complexity of computing SPDs in practice, especially on high-dimensional settings. Our results confirm the recent empirical observations reported in [Deshpande et al., 2019], which motivated Chapters 5 and 6, and provide a better understanding for them.

We demonstrate the applicability of our general theoretical results by applying them to specific instances of SPDs. For example, we establish a novel result on the topology induced by the Sliced-Cramér distance. We also derive a sample complexity result for SW which has never been shown before, under different assumptions on the measures to be compared.

We then introduce the sliced version of Sinkhorn divergences and demonstrate its statistical and computational advantages. Indeed, by combining our general results with recent work [Genevay et al., 2019, Mena and Niles-Weed, 2019], we derive the sample complexity of *Sliced-Sinkhorn divergences*, and obtain rates which, as opposed the sample complexity of Sinkhorn divergences, do not depend on d nor on the regularization parameter ε. We also show that this sliced divergence improves the worst-case computational complexity bounds of Sinkhorn divergences in $\mathbb{R}^d$.

Finally, we support our theory by conducting numerical experiments on synthetic and real data: we consider examples of sliced divergences (Sliced-MMD, Sliced-Wasserstein distance, Sliced-Sinkhorn divergences) and provide an empirical analysis of their topological, statistical or computational properties, which agrees with our theoretical results.

1.5 List of Publications

The contributions that we described in Section 1.4 led to the following publications.

- **Kimia Nadjahi**, Alain Durmus, Umut Şimşekli, and Roland Badeau. Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. *(Spotlight presentation)*

- Soheil Kolouri*, **Kimia Nadjahi***, Umut Şimşekli, Roland Badeau, and Gustavo Rohde. Generalized Sliced Wasserstein Distances. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. *(Star '$*$' means equal contribution)*

- **Kimia Nadjahi**, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli. Approximate Bayesian Computation with the Sliced-Wasserstein Distance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. *(Won the Best Student Paper Award)*

- **Kimia Nadjahi**, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Şimşekli. Statistical and Topological Properties of Sliced Probability Divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. *(Spotlight presentation)*

- **Kimia Nadjahi**, Alain Durmus, Pierre E. Jacob, Roland Badeau, and Umut Şimşekli. Fast Approximation of the Sliced-Wasserstein Distance Using Concen-

tration of Random Projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Chapter 2

Technical Background

This chapter provides a set of technical results to specify the notions addressed in Chapter 1, which will be useful for the remainder of the thesis. We first describe some of the theoretical properties that help evaluate the advantages of a probability divergence when used in generative modeling applications, specifically its *metric axioms*, *metrization of weak convergence* and *sample complexity*.

Then, we examine these properties for specific instances of probability divergences, namely the general class of *Integral Probability Metrics*, with a special focus on the *Maximum Mean Discrepancy*, and the *Wasserstein distance*. To properly introduce this latter metric, we give some background on *optimal transport theory*.

Finally, we explain the computational and statistical limitations of the Wasserstein distance to motivate the use of alternatives, including *Sinkhorn divergences* and *the Sliced-Wasserstein distance*. We define these two divergences, which rely on fundamentally different approaches, and present their known benefits over classical optimal transport metrics.

2.1 General Properties on Probability Divergences

In this section, we give the formal statement of desirable properties for a probability divergence in the context of generative modeling, starting with the metric axioms.

Definition 2.1 (Metric axioms)**.** *Let* X *be a Polish space and consider a divergence* $\mathbf{D}$ *on the space of probability distributions* $\mathcal{P}(\mathsf{X})$*.* $\mathbf{D}$ *is a* metric *if it takes finite values, i.e.* $\mathbf{D} : \mathcal{P}(\mathsf{X}) \times \mathcal{P}(\mathsf{X}) \to \mathbb{R}_+$*, and satisfies the following axioms.*

1. Symmetry: *For any* $\mu, \nu \in \mathcal{P}(\mathsf{X})$*,* $\mathbf{D}(\mu, \nu) = \mathbf{D}(\nu, \mu)$*.*
2. Triangle inequality: *For any* $\mu, \nu, \xi \in \mathcal{P}(\mathsf{X})$*,* $\mathbf{D}(\mu, \nu) \leq \mathbf{D}(\mu, \xi) + \mathbf{D}(\xi, \nu)$*.*
3. Identity of indiscernibles: *For any* $\mu, \nu \in \mathcal{P}(\mathsf{X})$*,* $\mathbf{D}(\mu, \nu) = 0$ *if and only if* $\mu = \nu$*.*

If $\mathbf{D}$ *only verifies some of these axioms, then it is said to be a* pseudo-metric*.*

To clarify the property of continuity described in **(P2)** (page 17), we define the *weak convergence of measures*, an important topological notion that characterizes a certain type of convergence for sequences of probability distributions [Billingsley, 1999, Problem 1.11, Chapter 1].

Definition 2.2 (Weak convergence). *Let* X *be a Polish space,* $(\mu_k)_{k\in\mathbb{N}}$ *a sequence of probability measures supported on* X*, and* $\mu \in \mathcal{P}(\mathsf{X})$*. We say that* μ_k converges weakly *to a probability measure* μ *on* X*, and write* $(\mu_k)_{k\in\mathbb{N}} \xrightarrow{w} \mu$ *(or* $\mu_k \xrightarrow{w} \mu$*), if for any continuous and bounded function* $f : \mathsf{X} \to \mathbb{R}$*,*

$$\lim_{k\to+\infty} \int_{\mathsf{X}} f \, \mathrm{d}\mu_k = \int_{\mathsf{X}} f \, \mathrm{d}\mu \, .$$

Another type of convergence for probability distributions can be characterized through a probability divergence: consider a Polish space X and $\mathbf{D} : \mathcal{P}(\mathsf{X}) \times \mathcal{P}(\mathsf{X}) \to \mathbb{R}_+ \cup \{+\infty\}$; using the same notations as in Definition 2.2, we say that $(\mu_k)_{k\in\mathbb{N}}$ *converges to* μ *under* $\mathbf{D}$ if $\lim_{k\to\infty} \mathbf{D}(\mu_k, \mu) = 0$.

Depending on the choice of $\mathbf{D}$ and X, convergence under $\mathbf{D}$ can imply weak convergence, and conversely. If weak convergence implies convergence under $\mathbf{D}$, *i.e.* property **(P2)** (page 17) is true, then we say that $\mathbf{D}$ is *weakly continuous*. If the reverse implication also holds, *i.e.* convergence under $\mathbf{D}$ is equivalent to weak convergence, and $\mathbf{D}$ is additionally a metric, then $\mathbf{D}$ is said to *metrize the weak convergence in* $\mathcal{P}(\mathsf{X})$.

Finally, property **(P4)** (page 17) is directly related to the *statistical efficiency* of a divergence. Indeed, consider that one has access to sets of samples drawn from unknown or intractable distributions, which is typically the case in data sciences. Then, from a practical point of view, a probability divergence $\mathbf{D}$ is useful in that context if it is able to appropriately handle the empirical approximations. Above all, this means that the evaluation of $\mathbf{D}$ from samples, or equivalently from empirical measures, must be easy to implement. Then, the error induced by this empirical approximation should be sufficiently small: given two unknown distributions $\mu, \nu \in \mathcal{P}(\mathsf{X})$ and their respective empirical instantiations $\hat{\mu}_n, \hat{\nu}_n$, the deviation $|\mathbf{D}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{D}(\mu, \nu)|$ must shrink to 0 as n increases with a reasonably fast convergence rate. This rate is commonly referred to as the *sample complexity* of $\mathbf{D}$.

2.2 Integral Probability Metrics

We now study the family of *Integral Probability Metrics*, introduced by Müller [1997] and characterized through the generic formula recalled in Definition 2.3.

Definition 2.3 (Integral Probability Metrics). *Let* Y *be a measurable space and denote by* $\mathbb{M}(\mathsf{Y})$ *the set of real-valued measurable functions on* Y*. Let* $\mathsf{F} \subset \mathbb{M}(\mathsf{Y})$ *and* $\mathcal{P}_{\mathsf{F}}(\mathsf{Y})$ *be the subset of measures in* $\mathcal{P}(\mathsf{Y})$ *characterized as*

$$\mathcal{P}_{\mathsf{F}}(\mathsf{Y}) = \{\mu \in \mathcal{P}(\mathsf{Y}) \, : \, \forall f \in \mathsf{F}, \, \int_{\mathsf{Y}} |f(y)| \, \mathrm{d}\mu(y) < +\infty\} \, .$$

The Integral Probability Metric associated with F*, denoted by* $\boldsymbol{\gamma}_{\mathsf{F}}$*, is defined for any* $\mu, \nu \in \mathcal{P}_{\mathsf{F}}(\mathsf{Y})$ *as*

$$\boldsymbol{\gamma}_{\mathsf{F}}(\mu, \nu) = \sup_{f\in\mathsf{F}} \left| \int_{\mathsf{Y}} f(y) \mathrm{d}(\mu - \nu)(y) \right| \, . \tag{2.1}$$

If μ *or* ν *does not belong to* $\mathcal{P}_{\mathsf{F}}(\mathsf{Y})$*, we set* $\boldsymbol{\gamma}_{\mathsf{F}}(\mu, \nu) = +\infty$.

IPMs offer several theoretical guarantees which suggest that their deployment in generative modeling applications is relevant. First, they are pseudo-metrics [Sriperumbudur et al., 2009]: they are non-negative, symmetric, verify the triangle inequality and for any $\mu \in \mathcal{P}_{\mathsf{F}}(\mathsf{Y})$, $\gamma_{\mathsf{F}}(\mu, \mu) = 0$. Besides, $\boldsymbol{\gamma}_{\mathsf{F}}$ metrizes weak convergence, provided that the span of F is dense in the space of continous and bounded functions on Y endowed with the supremum norm [Ambrosio et al., 2005, Section 5.1].

Finally, any IPM admits an empirical estimate which is consistent: consider the empirical distributions $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, where $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ are two sets of $n \in \mathbb{N}^*$ samples i.i.d. from μ and ν respectively, and denote by

$$\widehat{\boldsymbol{\gamma}}_{\mathsf{F}}(\hat{\mu}_n, \hat{\nu}_n) = \sup_{f \in \mathsf{F}} \left| \frac{1}{n} \sum_{i=1}^n \big(f(x_i) - f(y_i)\big) \right| . \tag{2.2}$$

Then, $\widehat{\boldsymbol{\gamma}}_{\mathsf{F}}(\hat{\mu}_n, \hat{\nu}_n)$ converges to $\boldsymbol{\gamma}_{\mathsf{F}}(\mu, \nu)$ as n grows to $+\infty$, under some mild assumptions on μ, ν and their empirical instances [Sriperumbudur et al., 2012, Lemma 3.1]. Nevertheless, for arbitrary F, the empirical estimate $\widehat{\boldsymbol{\gamma}}_{\mathsf{F}}(\hat{\mu}_n, \hat{\nu}_n)$ is not always simple to compute nor converges to $\boldsymbol{\gamma}_{\mathsf{F}}(\mu, \nu)$ sufficiently fast. An important example of IPMs which admits a consistent and statistically efficient empirical estimator is the *Maximum Mean Discrepancy* (MMD, Gretton et al. [2012]), which is defined below.

Definition 2.4 (Maximum Mean Discrepancy)**.** *Let H be a reproducing kernel Hilbert space (RKHS) for real-valued functions on a measurable space Y, and F be the unit ball in H. Then, γ_{F} (Definition 2.3) defines the* MMD in RKHS*: for any $\mu, \nu \in \mathcal{P}_{\mathsf{F}}(\mathsf{Y})$,*

$$\mathbf{MMD}(\mu, \nu; \mathsf{F}) = \sup_{f \in \mathsf{F},\ i.e.\ f:\mathsf{Y}\to\mathbb{R}, \|f\|_{\mathsf{H}} \leq 1} \left| \int_{\mathsf{Y}} f(y) \mathrm{d}(\mu - \nu)(y) \right| \tag{2.3}$$

By [Gretton et al., 2012, Lemma 6], the above definition can equivalently be written in terms of the kernel k associated to H,

$$\begin{aligned} \mathbf{MMD}^2(\mu, \nu; \mathsf{F}) = &\int_{\mathsf{Y}\times\mathsf{Y}} k(x, x') \mathrm{d}(\mu \otimes \mu)(x, x') + \int_{\mathsf{Y}\times\mathsf{Y}} k(y, y') \mathrm{d}(\nu \otimes \nu)(y, y') \\ &- 2 \int_{\mathsf{Y}\times\mathsf{Y}} k(x, y) \mathrm{d}(\mu \otimes \nu)(x, y) \end{aligned} \tag{2.4}$$

MMD has been shown to metrize weak convergence when Y is a compact space or $\mathsf{Y} = \mathbb{R}^d$, under specific conditions on its defining kernel k [Sriperumbudur et al., 2010]. One of the most popular kernels is the *Gaussian (RBF) kernel*, defined for any $x, y \in \mathbb{R}^d$ as $k(x, y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$, where $\sigma > 0$ is called the *bandwidth parameter*. According to the aforementioned general result, MMD based on the Gaussian kernel metrizes weak convergence on compact spaces.

When F is the unit ball of an RKHS, the solution of (2.2) is unique and available in closed form, thus defines an empirical estimator for MMD that is easy to implement [Sriperumbudur et al., 2012, Theorem 2.4]. This estimate, denoted by $\widehat{\mathbf{MMD}}(\cdot, \cdot; \mathsf{F})$, is consistent and exhibits a rate that does not depend on the data dimension d: by [Sriperumbudur et al., 2012, Corollary 3.5],

$$\left| \widehat{\mathbf{MMD}}(\hat{\mu}_n, \hat{\nu}_n; \mathsf{F}) - \mathbf{MMD}(\mu, \nu; \mathsf{F}) \right| = \mathcal{O}(n^{-1/2}) . \tag{2.5}$$

The analytical properties of MMD have been especially useful for statistical hypothesis testing [Gretton et al., 2008, Fukumizu et al., 2008, Gretton et al., 2012] and generative modeling [Li et al., 2015, Dziugaite et al., 2015, Sutherland et al., 2017, Bińkowski et al., 2018, Arbel et al., 2019]. However, these studies also demonstrate that the theoretical guarantees and practical implications induced by MMD are strongly affected by the choice of its kernel function. For instance, the constant in the sample complexity (2.5) depends on $\sup_{x\in\mathsf{Y}} \sqrt{k(x,x)}$ [Sriperumbudur et al., 2012, Corollary 3.5]. The influence of the kernel function on the training and performance of MMD-based IGM methods has not been fully understood, and the study of this aspect is an active research topic: while [Arbel et al., 2019] provided some answers by identifying the settings where the gradient estimators used in MMD GANs are unbiased, it is still unclear why some kernel choices yields better results than others in their experiments.

We conclude this section by presenting another important instance of IPMs: when $\mathsf{F} = \{f : \mathsf{Y} \to \mathbb{R} \,:\, \|f\|_{\mathrm{Lip}} \leq 1\}$, where

$$\|f\|_{\mathrm{Lip}} = \sup_{x,y\in\mathsf{Y}, x\neq y} \frac{|f(x) - f(y)|}{\|x - y\|} \,, \tag{2.6}$$

γ_{F} is then known as the *Wasserstein distance of order 1* [Villani, 2008, Theorem 5.10]. This metric is directly related to the field of *optimal transport theory*, as we explain in the next sections.

2.3 Optimal Transport

We provide the basics of optimal transport theory in order to give a better understanding of the roots of the Wasserstein distance.

OT was first formulated by Monge [1781] and defines a mathematical formalism which, intuitively, aims at finding a way to move the probability mass from one distribution to another with least effort. This effort is quantified by means of a *cost function*, which operates as follows: denote by $\mu \in \mathcal{P}(\mathsf{X})$ the *source distribution* and by $\nu \in \mathcal{P}(\mathsf{Y})$ the *target distribution*, where X and Y are two Polish spaces. Then, $c : \mathsf{X}\times\mathsf{Y} \to \mathbb{R}_+\cup\{\infty\}$ defines the function that returns for any $(x, y) \in \mathsf{X} \times \mathsf{Y}$ the cost of transporting x to y. This cost function is typically chosen as a distance on $\mathsf{X} \times \mathsf{Y}$, so that it evaluates how far x and y are from each other: the smaller $c(x, y)$, the closer x is to y, so the least effort.

On the other hand, the idea of transporting one distribution to another is described by the notion of *push-forward*. To illustrate how this mathematical operator works, we consider the specific case of discrete measures: let $f : \mathsf{X} \to \mathsf{Y}$ be a continuous map, and $\xi \in \mathcal{P}(\mathsf{X})$ supported over n points, *i.e.* $\xi = n^{-1}\sum_{i=1}^{n} \delta_{x_i}$. Then, the push-forward operator associated to f, denoted by $f_\sharp$, takes ξ as input and returns a probability distribution on Y characterized by

$$f_\sharp\xi = \frac{1}{n}\sum_{i=1}^{n} \delta_{f(x_i)} \,.$$

In other words, $f_\sharp$ moves $\xi \in \mathcal{P}(\mathsf{X})$ towards a new distribution on Y, by applying f to its support points $\{x_i\}_{i=1}^n$. This operation can be generalized to *continuous* distributions: in this case, f moves each elementary mass of the input distribution. We give the formal definition of the push-forward operator in Definition 2.5.

Definition 2.5 (Push-forward operator)**.** *Let* X, Y *be two Polish spaces,* $f : \mathsf{X} \to \mathsf{Y}$ *a continuous map, and* $\xi \in \mathcal{P}(\mathsf{X})$*. We denote by* $f_\sharp : \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{Y})$ *the push-forward operator associated to* f*. The evaluation of* $f_\sharp$ *at* ξ *yields a probability distribution supported on* Y*, denoted by* $f_\sharp\xi$ *and characterized by the following two properties, which are equivalent.*

1. *For any measurable set* A *in* Y*,* $f_\sharp\xi(\mathsf{A}) = \xi(f^{-1}(\mathsf{A}))$ *where*
$$f^{-1}(\mathsf{A}) = \{x \in \mathsf{X} : \ f(x) \in \mathsf{A}\} \ .$$
2. *For any* $g \in \mathcal{C}(\mathsf{Y})$*,* $\int_{\mathsf{Y}} h(y)\mathrm{d}f_\sharp\xi(y) = \int_{\mathsf{X}} g \circ f(x)\mathrm{d}\xi(x)$ *.*

$f_\sharp\xi$ *can be referred to as the* push-forward measure of ξ by f*.*

The OT problem as formulated by Monge consists in finding the map $T : \mathsf{X} \to \mathsf{Y}$ that transports the source distribution $\mu \in \mathcal{P}(\mathsf{X})$ to its target $\nu \in \mathcal{P}(\mathsf{Y})$ (via the push-forward operator $T_\sharp$) with minimal total cost (measured by the cost function c). Formally, the corresponding optimization problem is given by

$$\min_T \int_{\mathsf{X}} c\big(x, T(x)\big)\mathrm{d}\mu(x) \quad s.t. \quad T_\sharp\mu = \nu \ . \tag{2.7}$$

Monge's formulation leads to a nonconvex optimization problem (2.7), which boils down to a combinatorial assignment problem in the discrete case [Peyré and Cuturi, 2019, Section 2.2]. Such problems are difficult to solve in general, and feasible solutions do not always exist.

To overcome the issues induced by the resolution of Monge's OT problem, Kantorovich [1942] introduced a relaxed version of (2.7) such that the transportation becomes probabilistic: the probability mass of any source point can be split into smaller masses which are then assigned to different target points, whereas the Monge problem performs a one-to-one assignment. To allow for mass splitting, Kantorovich proposes to use *couplings* instead of deterministic transport maps, *i.e.* the feasible set of solutions now corresponds to the set of joint distributions on $\mathsf{X}\times\mathsf{Y}$ whose first and second marginals are given by μ and ν respectively. The condition on the marginals reflects the *conservation of total probability mass* when carried from μ to ν. Kantorovich's formulation of OT is finally given by

$$\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{Y}} c(x,y)\mathrm{d}\pi(x,y) \ , \tag{2.8}$$

$$\text{with,} \quad \Pi(\mu,\nu) = \big\{\pi \in \mathcal{P}(\mathsf{X}\times\mathsf{Y}) \ : \text{for any measurable sets } \mathsf{A}\subset\mathsf{X},\ \mathsf{B}\subset\mathsf{Y},\ \pi(\mathsf{A}\times\mathsf{Y}) = \mu(\mathsf{A}),\ \pi(\mathsf{X}\times\mathsf{B}) = \nu(\mathsf{B})\big\} \ .$$

The solution to (2.8) has been shown to always exist, provided that the following conditions are satisfied.

(i) X and Y are compact metric spaces, and the cost function c is continuous [Santambrogio, 2015, Theorem 1.4] or lower semi-continuous and bounded from below [Santambrogio, 2015, Theorem 1.5],

(ii) or alternatively, X and Y are Polish spaces and c is lower semi-continuous [Santambrogio, 2015, Theorem 1.7].

These theoretical guarantees provides a significant advantage over the Monge problem, which can be ill-posed. Furthermore, Kantorovich's formulation led to the definition of powerful probability divergences, namely Wasserstein distances.

2.4 Wasserstein Distances

Wasserstein distances define a class of probability divergences which compare two distributions by solving a transport problem, and have gradually become a useful tool for many other applications than OT theory. In this section, we introduce the mathematical definition of Wasserstein distances, specify some of their key theoretical properties and present two special settings where they are easily computable. Then, we explain the main limitations of these OT metrics, which makes them impractical when deployed in modern machine learning applications.

2.4.1 Definition and elementary properties

We start by defining the *Wasserstein distance of order p* between any two distributions. In what follows, we will assume for ease of reading that the compared distributions are supported on the same space, *i.e.* $\mathsf{X} = \mathsf{Y}$.

Definition 2.6 (Wasserstein distances)**.** *Let* X *be a Polish space equipped with a distance* ρ*, and* $p \in [1, +\infty)$*. The Wasserstein distance of order* p *is defined for any* $\mu, \nu \in \mathcal{P}(\mathsf{X})$ *as*

$$\mathbf{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{X}} \rho(x, y)^p \mathrm{d}\pi(x, y) \right)^{1/p} . \tag{2.9}$$

One can easily see from Definition 2.6 that Wasserstein distances correspond to specific instances of the Kantorovich transport problem recalled in (2.8). Hence, comparing two distributions μ and ν via the Wasserstein distance amounts to solving a transport problem where the source and target distributions are given by μ and ν; or conversely, by ν and μ, since this order has no importance in Kantorovich's formulation.

By definition, Wasserstein distances leverage the information captured by the cost function c on the geometry of the supports of μ and ν in order to optimally move the probability mass from μ to ν. The comparisons made via this OT metric are then conceptually more powerful than with traditional divergences, e.g. f-divergences which perform pointwise comparisons of the probability mass [Rényi, 1961].

The benefits of Wasserstein distances are further confirmed by the following theoretical results: denote by $\mathcal{P}_p(\mathsf{X})$ the set of probability measures on X with finite p'th moment, *i.e.*

$$\mathcal{P}_p(\mathsf{X}) = \left\{ \mu \in \mathcal{P}(\mathsf{X}) \ : \ \int_{\mathsf{X}} \rho(x_0, x)^p \mathrm{d}\mu(x) < +\infty, \ \text{for some } x_0 \in \mathsf{X} \right\} .$$

Then, $\mathbf{W}_p$ is a metric on $\mathcal{P}_p(\mathsf{X})$ [Villani, 2008, Chapter 6] which metrizes the weak convergence, *i.e.* the weak convergence of probability measures supported on $\mathcal{P}_p(\mathsf{X})$ is equivalent to convergence under $\mathbf{W}_p$ [Villani, 2008, Theorem 6.9]. These properties explain why the Wasserstein distance can effectively compare any two distributions, even when their supports do not overlap, as opposed to f-divergences, e.g. KL, and some instances of IPMs: see [Arjovsky et al., 2017, Example 1] for an illustration of this advantage.

There are some special cases where computing the Wasserstein distance, *i.e.* solving the corresponding Kantorovich problem, is easy and reasonably cheap. We present two of these settings, which provide closed-form solutions and are of significant important in this thesis.

Gaussian distributions. Denote by $\mathcal{N}(\mathbf{m}, \Sigma)$ the Gaussian distribution on $\mathbb{R}^d$ with mean $\mathbf{m} \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ symmetric positive-definite. The Wasserstein distance of order 2 between two Gaussians, also known as the *Wasserstein-Bures metric* [Dowson and Landau, 1982], is given by

$$\mathbf{W}_2^2\{\mathcal{N}(\mathbf{m}_1, \Sigma_1), \mathcal{N}(\mathbf{m}_2, \Sigma_2)\} = \|\mathbf{m}_1 - \mathbf{m}_2\|^2 + \mathrm{Tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right] , \quad (2.10)$$

where Tr denotes the trace operator.

Besides, if $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$, (2.10) can be written in the following simpler form

$$\mathbf{W}_2^2\{\mathcal{N}(\mathbf{m}_1, \Sigma_1), \mathcal{N}(\mathbf{m}_2, \Sigma_2)\} = \|\mathbf{m}_1 - \mathbf{m}_2\|^2 + \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_F^2 , \quad (2.11)$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Univariate distributions. Consider $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, and denote by F_μ^{-1} and F_ν^{-1} the quantile functions of μ and ν respectively. By [Rachev and Rüschendorf, 1998, Theorem 3.1.2.(a)],

$$\mathbf{W}_p^p(\mu, \nu) = \int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right|^p \mathrm{d}t . \quad (2.12)$$

The analytical formula in (2.12) can be efficiently approximated by replacing the integral with a Monte Carlo estimate. The first approximation scheme we consider is given by,

$$\mathbf{W}_p^p(\mu, \nu) \approx \frac{1}{K}\sum_{k=1}^K \left|\tilde{F}_\mu^{-1}(t_k) - \tilde{F}_\nu^{-1}(t_k)\right|^p , \quad (2.13)$$

where $\{t_k\}_{k=1}^K$ are uniform and independent samples from $[0, 1]$ and for $\xi \in \{\mu, \nu\}$, $\tilde{F}_\xi^{-1}$ is a linear interpolation of $\bar{F}_\xi^{-1}$ which denotes either the exact quantile function of ξ if ξ is discrete, or an approximation by a Monte Carlo procedure. This last option is justified by the Glivenko-Cantelli theorem [Loève, 1977].

The second approximation is given by,

$$\mathbf{W}_p^p(\mu, \nu) \approx \frac{1}{K}\sum_{k=1}^K \left|s_k - \tilde{F}_\nu^{-1}(\tilde{F}_\mu(s_k))\right|^p , \quad (2.14)$$

where $\{s_k\}_{i=1}^K$ are uniform and independent samples from μ and for $\xi \in \{\mu, \nu\}$, $\tilde{F}_\xi$ (resp. $\tilde{F}_\xi^{-1}$) is a linear interpolation of $\bar{F}_\xi$ (resp. $\bar{F}_\xi^{-1}$) which denotes either the exact cumulative distribution function (resp. quantile function) of ξ if ξ is discrete or an approximation by a Monte Carlo procedure.

Let us finally mention the convenient case of univariate discrete measures: if $\mu = n^{-1}\sum_{i=1}^n \delta_{x_i}$ and $\nu = n^{-1}\sum_{i=1}^n \delta_{y_i}$, where $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ are two sets of $n \in \mathbb{N}^*$ observations taking values in $\mathbb{R}$, then (2.12) can simply be calculated by sorting $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$. In this case,

$$\mathbf{W}_p^p(\mu, \nu) = \frac{1}{n}\sum_{i=1}^n |x_{(i)} - y_{(i)}|^p , \quad (2.15)$$

where $x_{(1)} \leq \ldots \leq x_{(n)}$ and $y_{(1)} \leq \ldots \leq y_{(n)}$.

However, apart from favorable singular cases such as the ones discussed above, computing Wasserstein distances entails several issues in practice, which are especially problematic in data sciences. We elaborate on these practical limitations in the next section.

2.4.2 Practical aspects and limitations

When evaluated between two multivariate empirical distributions, the Wasserstein distance admits a discrete formulation which can be rewritten as a linear program [Peyré and Cuturi, 2019, Section 3.1] and the solution is not analytically available in general. Standard solvers from linear programming and combinatorial optimization (e.g., the simplex method) are then particularly relevant in that context, since they can be used to compute $\mathbf{W}_p(\hat{\mu}_n, \hat{\nu}_n)$ [Peyré and Cuturi, 2019, Chapter 3]. While they participated in the spread of OT, these methods tend to have a super-cubic cost in practice, and their worst-case computational complexity scales in $\mathcal{O}(n^3 \log(n))$. Therefore, Wasserstein distances usually require important computational resources when deployed in applications that deal with medium-to-large volumes of data, such as machine learning problems.

Additionally, several prior studies have demonstrated the statistical limitations of the Wasserstein distance by deriving its sample complexity: the convergence rate of $\mathbf{W}_p(\hat{\mu}_n, \hat{\nu}_n)$ to $\mathbf{W}_p(\mu, \nu)$ is in $\mathcal{O}(n^{-1/d})$. This result was first established by Dudley [1969] for $p = 1$ and compactly supported measures, and has been extended and sharpened in subsequent work [Dereich et al., 2013, Boissard and Gouic, 2014, Fournier and Guillin, 2015]. Their contribution show that in general, the convergence rate of $\mathbf{W}_p(\mu, \hat{\mu}_n)$ to zero (consequently, of $\mathbf{W}_p(\hat{\mu}_n, \hat{\nu}_n)$ to $\mathbf{W}_p(\mu, \nu)$, by the triangle inequality) degrades exponentially in the ambient dimension d, meaning that for high data dimensions, n must be very large for $\mathbf{W}_p(\hat{\mu}_n, \hat{\nu}_n)$ to yield an accurate approximation of $\mathbf{W}_p(\mu, \nu)$. This requirement might be unrealistic or too restrictive, since increasing n inevitably increases the complexity of computing $\mathbf{W}_p(\hat{\mu}_n, \hat{\nu}_n)$, let alone that it might be difficult to collect sufficiently many samples in some practical settings. We note however that a recent study drew a more optimistic conclusion by considering measures that are intrinsically lower-dimensional: in this specific case, the rate can be reasonably fast as it depends on that intrinsic dimension [Weed and Bach, 2019].

Because of these computational and statistical limitations, deploying the Wasserstein distance to address practical tasks in data sciences can result in highly inefficient algorithms. Nevertheless, these issues have in turn paved the way for novel research directions within the machine learning community, which taken altogether, define a whole new field called *computational optimal transport* (COT).

The next two sections are precisely dedicated to the main classes of probability divergences that have emerged from COT: *Sinkhorn divergences*, which stem from *regularized optimal transport*, and the *Sliced-Wasserstein distance*, which is the central object of this thesis. Thanks to their favorable computational and statistical properties, these divergences serve as practical alternatives to the Wasserstein distance and have inspired the design of novel efficient techniques in data sciences.

2.5 Regularized Optimal Transport, Sinkhorn Divergences

Popularized by Cuturi [2013], *regularized optimal transport* is one of the main methodologies that have allowed the application of OT in high-dimensional applied problems. Its guiding principle consists in adding a penalty term to the Kantorovich problem in order to approximate its solution. This penalty can be any strictly convex function, and this choice of regularization leads to different practical consequences [Peyré and Cuturi, 2019, Remark 4.10]. We will only describe the most common practice, which is *entropic penalization* and thereby defines the following regularized version of (2.9),

$$\mathbf{W}_{p,\varepsilon}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{\mathsf{X}\times\mathsf{X}} \rho(x,y)^p \mathrm{d}\pi(x,y) + \varepsilon \mathbf{H}(\pi|\mu\otimes\nu) \right\}, \tag{2.16}$$

where $\varepsilon > 0$ acts as a regularization parameter and $\mathbf{H}(\pi|\mu\otimes\nu)$ denotes the relative entropy (or Kullback-Leibler divergence) of the transport plan π with respect to $\mu\otimes\nu$: if π is absolutely continuous with respect to $\mu\otimes\nu$,

$$\mathbf{H}(\pi|\mu\otimes\nu) = \int_{\mathsf{X}\times\mathsf{X}} \log\Big(\frac{\mathrm{d}\pi(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)}\Big)\mathrm{d}\pi(x,y)\,, \tag{2.17}$$

otherwise, $\mathbf{H}(\pi|\mu\otimes\nu) = +\infty$.

This entropic regularization provides a number of desirable properties and has thereupon contributed significantly to the renewed interest in OT for data sciences, e.g. in supervised learning [Frogner et al., 2015], computer vision [Solomon et al., 2015], domain adaptation [Courty et al., 2017, Redko et al., 2019], dictionary or embeddings learning [Schmitz et al., 2018, Frogner et al., 2019] and generative modeling [Genevay et al., 2018].

The first major advantage is that the solutions of the Kantorovich problem can be efficiently approximated via this regularization. Indeed, when μ and ν are two discrete distributions based on n points, the penalized problem in (2.16) can be efficiently solved by applying an iterative numerical solver called *Sinkhorn's algorithm* [Franklin and Lorenz, 1989], where each iteration consists in matrix-vector products. Such operations induce a complexity in $\mathcal{O}(n^2)$, which can further be improved for example by implementing them on GPU since they can be carried out in parallel, or by leveraging the structure of the cost function: see [Peyré and Cuturi, 2019, Section 4.3] for the explanation of such acceleration techniques. Therefore, Sinkhorn's algorithm executes considerably faster than the approximate solvers used in classical OT, and has been shown to converge to the solution of the unregularized Kantorovich problem with an accuracy of δ in $\mathcal{O}(n^2\|c\|_\infty^2 \log(n)\delta^{-2})$ operations [Dvurechensky et al., 2018]. Several algorithms for regularized optimal transport have then been developed and can achieve improved convergence rates or a superior empirical performance, e.g. a better computational efficiency [Altschuler et al., 2017, Dvurechensky et al., 2018, Seguy et al., 2018, Abid and Gower, 2018, Lin et al., 2019].

On the other hand, the entropic regularization of OT yields an alternative divergence to the Wasserstein distance that is better suited for data sciences: $\mathbf{W}_{p,\varepsilon}(\mu,\nu)$ is convex as a function of (μ,ν) for $\varepsilon \geq 0$, and is always smooth provided that $\varepsilon > 0$, with a gradient known in closed-form [Feydy et al., 2019]. Therefore, adding an entropic penalty has been especially useful in defining a reliable loss function which, as opposed to the unregularized Wasserstein distance, can efficiently be differentiated in general: a detailed discussion on this aspect accompanied with concrete examples is provided in [Peyré and Cuturi, 2019, Section 9.1].

Nevertheless, this divergence does not verify all metric axioms since the entropic penalty term introduces the following bias: for any distribution μ, $\mathbf{W}_{p,\varepsilon}(\mu,\mu) \neq 0$. The family of *Sinkhorn divergences* has then been introduced to fix this problem, while inheriting from the benefits of entropic regularized OT.

Definition 2.7 (Sinkhorn divergences)**.** *Let X be a Polish space equipped with a distance ρ. Let $p \in [1,+\infty)$ and $\varepsilon > 0$. The Sinkhorn divergence is defined for any $\mu,\nu \in \mathcal{P}(\mathsf{X})$ as*

$$\overline{\mathbf{W}}_{p,\varepsilon}(\mu,\nu) = \mathbf{W}_{p,\varepsilon}(\mu,\nu) - \frac{1}{2}\big\{\mathbf{W}_{p,\varepsilon}(\mu,\mu) + \mathbf{W}_{p,\varepsilon}(\nu,\nu)\big\}\,, \tag{2.18}$$

where $\mathbf{W}_{p,\varepsilon}$ is the regularized OT cost given by (2.16).

Sinkhorn divergences have been proved to be smooth and convex, similarly to $\mathbf{W}_{p,\varepsilon}$, but they are also symmetric positive definite and metrize the weak convergence [Feydy et al., 2019]. Regarding their practical implications, computing $\overline{\mathbf{W}}_{p,\varepsilon}$ is not much more expensive than for $\mathbf{W}_{p,\varepsilon}$, and can be done efficiently on GPU [Feydy et al., 2019, Section 3].

Another important feature of Sinkhorn divergences is that they interpolate between the Wasserstein distance (when $\varepsilon \to 0$) and MMD (when $\varepsilon \to +\infty$) [Ramdas et al., 2017]. Hence, Sinkhorn divergences achieve a trade-off between these two divergences, which is clearly reflected in their sample complexity recalled below.

Theorem 2.8 (Theorem 3 in Genevay et al. [2019])**.** *Let X be a compact set of $\mathbb{R}^d$ with diameter denoted by D_{X}, and consider $\mu,\nu \in \mathcal{P}(\mathsf{X})$. The sample complexity of the Sinkhorn divergence is given by*

$$\mathbb{E}\left|\overline{\mathbf{W}}_{p,\varepsilon}(\hat{\mu}_n,\hat{\nu}_n) - \overline{\mathbf{W}}_{p,\varepsilon}(\mu,\nu)\right| \leq K_{d,D_{\mathsf{X}}}\big(1+\varepsilon^{-\lfloor d/2\rfloor}\big)\exp\left(\frac{k_{D_{\mathsf{X}}}}{\varepsilon}\right) n^{-1/2}\,, \tag{2.19}$$

where $K_{d,D_{\mathsf{X}}}$ is a constant that depends on d and D_{X}, and $k_{D_{\mathsf{X}}}$ is a constant that depends on D_{X}.

Theorem 2.8 has very recently been refined to the case where μ,ν are subgaussian [Mena and Niles-Weed, 2019]. In both [Genevay et al., 2019] and [Mena and Niles-Weed, 2019], the convergence rate is consistent with the aforementioned interpolation property: when ε goes to infinity, it scales in $n^{-1/2}$ and is thus comparable to the MMD case (2.5); when ε shrinks to 0, the rate gets significantly slower and, analogously to the Wasserstein distance, further degrades as the dimension d increases.

2.6 Sliced-Wasserstein Distance

Another important class of alternative divergence emerging for computational optimal transport relies on the use of *low-dimensional projections* of probability distributions. This line of work was initiated by Rabin et al. [2012] and Bonneel et al. [2015], which designed the *Sliced-Wasserstein distance* in order to speed up the computation of Wasserstein barycenters. We give the formal definition of SW in Definition 2.9, which we illustrate in Figure 1.3 (page 20).

Definition 2.9 (Sliced-Wasserstein distance)**.** *Let $\mathsf{X} \subset \mathbb{R}^d$ be a Polish space endowed with the Euclidean distance and denote by $\mathbb{S}^{d-1} = \big\{\theta \in \mathbb{R}^d \,:\, \|\theta\| = 1\big\}$ the unit sphere in $\mathbb{R}^d$. For any $u \in \mathbb{S}^{d-1}$, denote by $u^\star : \mathsf{X} \to \mathbb{R}$ the linear form given by $u^\star(x) = \langle u, x\rangle$. Let*

Algorithm 1: Monte Carlo approximation of SW

Input: Two sets of observations $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, number of projection directions L, order p.

SW $= 0$
for $l = 1, \dots, L$ **do**
  Sample: $u_l \sim \boldsymbol{\sigma}$
  for $i = 1, \dots, n$ **do**
    Project: $x'_i = \langle u_l, x_i \rangle$, $y'_i = \langle u_l, y_i \rangle$
  Sort: $x'_{(1)} \leq x'_{(2)} \leq \cdots \leq x'_{(n)}$, $y'_{(1)} \leq y'_{(2)} \leq \cdots \leq y'_{(n)}$
  SW $=$ SW $+ (1/n) \sum_{i=1}^n |x'_{(i)} - y'_{(i)}|^p$
SW $= (\text{SW}/L)^{1/p}$
return SW

$p \in [1, +\infty)$. *The Sliced-Wasserstein distance of order p is defined for any $\mu, \nu \in \mathcal{P}_p(\mathsf{X})$ as*

$$\mathbf{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p(u_\sharp^\star \mu, u_\sharp^\star \nu) \mathrm{d}\boldsymbol{\sigma}(u) \ , \tag{2.20}$$

where $\boldsymbol{\sigma}$ is the uniform distribution on $\mathbb{S}^{d-1}$, and for any $u \in \mathbb{S}^{d-1}$, $u_\sharp^\star = (u^\star)_\sharp$ denotes the push-forward operator associated to $u^\star$.

In other words, SW measures the dissimilarity between $\mu, \nu \in \mathcal{P}_p(\mathsf{X})$ by computing $\mathbb{E}[\mathbf{W}_p^p(u_\sharp^\star \mu, u_\sharp^\star \nu)]$, where $\mathbb{E}$ is taken with respect to u uniformly distributed on $\mathbb{S}^{d-1}$. The push-forward measures $u_\sharp^\star \mu$ and $u_\sharp^\star \nu$ are univariate since they correspond to "projections" of μ and ν along the direction $u \in \mathbb{S}^{d-1}$. In particular, when $\xi \in \{\mu, \nu\}$ is approximated by the empirical measure $\hat{\xi}_n = (1/n) \sum_{i=1}^n \delta_{x_i}$, where $\{x_i\}_{i=1}^n$ are i.i.d. samples from ξ, then

$$u_\sharp^\star \hat{\xi}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\langle u, x_i \rangle} \ .$$

This means that empirical approximations of $u_\sharp^\star \mu$ and $u_\sharp^\star \nu$ can simply be obtained by projecting the available samples from μ and ν along u.

Besides, the expectation that defines SW can easily be approximated with a standard Monte Carlo method: one draws $L \in \mathbb{N}^*$ samples i.i.d. from $\boldsymbol{\sigma}$, denoted by $\{u_l\}_{l=1}^L$, and approximates SW (2.20) with

$$\mathbf{SW}_{p,L}^p(\mu, \nu) = \frac{1}{L} \sum_{l=1}^L \mathbf{W}_p^p\big(u_{l\sharp}^\star \mu, u_{l\sharp}^\star \nu\big) \ , \tag{2.21}$$

where for $l \in \{1, \dots, L\}$, $u_{l\sharp}^\star = \{(u_l)^\star\}_\sharp$ is the push-forward operator associated to $(u_l)^\star$, the linear form introduced in Definition 2.9. The Monte Carlo estimate of SW thus requires solving finitely many OT problems in $\mathbb{R}$, which is convenient given our discussion in Section 2.4.1. This approximation method is summarized in Algorithm 1, and has a complexity in $\mathcal{O}\big(Ldn + Ln\log(n)\big)$ due to the projecting and sorting operations.

The computational benefits of SW over the Wasserstein distance have encouraged its use in various practical applications, thus making it an increasingly popular divergence over the last few years. We present hereafter the main contributions on the theoretical properties of SW.

First, [Bonnotte, 2013, Chapter 5] provides a collection of useful results on SW, including the following: SW is a distance on $\mathcal{P}_p(\mathbb{R}^d)$ [Bonnotte, 2013, Proposition 5.1.2], and its relation with the Wasserstein distance of order p has been rigorously established as follows.

Theorem 2.10 (Proposition 5.1.3 and Theorem 5.1.5 of Bonnotte [2013])**.** *Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. There exists a positive constant $c_{d,p} \leq 1$ depending on d and p such that*

$$\mathbf{SW}_p^p(\mu, \nu) \leq c_{d,p} \mathbf{W}_p^p(\mu, \nu) \ . \tag{2.22}$$

Besides, denote by $\mathrm{B}_d(\mathbf{0}, R) = \left\{x \in \mathbb{R}^d \ : \ \|x\| < R\right\}$ the open ball in $\mathbb{R}^d$ of radius $R > 0$ centered around $\mathbf{0} \in \mathbb{R}^d$, and assume that μ and ν are supported on $\mathrm{B}_d(\mathbf{0}, R)$. Then, there exists a constant $C_{d,p} > 0$ such that

$$\mathbf{W}_p^p(\mu, \nu) \leq \frac{C_{d,p}}{c_{d,p}} R^{p-1/(d+1)} \mathbf{SW}_p(\mu, \nu)^{1/(d+1)} \tag{2.23}$$

Very recently, Bayraktar and Guo [2021] conducted a thorough theoretical analysis on the equivalence between the Wasserstein distance and SW, which nicely complements Theorem 2.10. Their results also concern a variant of SW, the *maximum Sliced-Wasserstein distance* [Deshpande et al., 2019], which we described in Section 1.4 (page 25) and define in Chapter 5.

Finally, only a few studies have investigated the sample complexity of SW to justify the statistical efficiency of SW-based methods. If μ, ν are two isotropic Gaussian distributions supported on $\mathbb{R}^d$, then $|\mathbf{SW}_2(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SW}_2(\mu, \nu)|$ can be bounded with high probability using the concentration inequality derived in [Deshpande et al., 2019, Claim 1], which shows that $\mathbf{SW}_2$ offers a "polynomial" sample complexity. A very recent study derived two upper bounds on $\mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \mu)]$, which demonstrate that the rate in the sample complexity of $\mathbf{SW}_p$ cannot be worse than $n^{-1/2p}$ and can achieve $n^{-1/2}$, under specific assumptions on the regularity of μ as measured by a certain functional [Manole et al., 2019, Proposition 1].

To conclude, let us mention that the complete review of the literature on SW provided in Sections 1.3 and 1.4 confirms the relevance of this metric in data sciences, but also emphasizes a severe lack of theoretical insights. As explained in Section 1.4, one of the objectives of this thesis is then to bridge this gap between theory and practice: in particular, the next chapter aims at making SW-based generative models more theoretically grounded, by analyzing the asymptotic properties of the estimators obtained by minimizing SW.

Chapter 3

Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance

This chapter is based on [Nadjahi et al., 2019].

Minimum expected distance estimation (MEDE) algorithms have been widely used for probabilistic models with intractable likelihood functions, and have become increasingly popular due to their use in implicit generative modeling. Emerging from computational optimal transport, the Sliced-Wasserstein distance has become a popular choice in MEDE thanks to its simplicity and computational benefits. While several studies have reported empirical success on generative modeling with SW, the theoretical properties of such estimators have not yet been established.

In this chapter, we investigate the asymptotic properties of estimators that are obtained by minimizing SW. We first show that convergence in SW implies weak convergence of probability measures in general Wasserstein spaces. Then we show that estimators obtained by minimizing SW (and also an approximate version of SW) are asymptotically consistent. We finally prove a central limit theorem, which characterizes the asymptotic distribution of the estimators and establish a convergence rate of $\sqrt{n}$, where n denotes the number of observed data points. We illustrate the validity of our theory on both synthetic data and neural networks.

3.1 Introduction

Minimum distance estimation (MDE) is a generalization of maximum-likelihood inference, where the goal is to minimize a distance between the empirical distribution of a set of i.i.d. observations $Y_{1:n} = (Y_1, \dots, Y_n)$ and a family of distributions indexed by a parameter θ. The problem is formally defined as follows [Wolfowitz, 1957, Basu et al., 2011],

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} \, \mathbf{D}(\hat{\mu}_n, \mu_\theta) \, , \tag{3.1}$$

where $\mathbf{D}$ denotes a distance (or a divergence in general) between probability measures, μ_θ denotes a probability measure indexed by θ, Θ denotes the parameter space, and

$$\hat{\mu}_n = \frac{1}{n} \sum\nolimits_{i=1}^{n} \delta_{Y_i} \tag{3.2}$$

denotes the empirical measure of $Y_{1:n}$. When $\mathbf{D}$ is chosen as the Kullback-Leibler divergence, this formulation coincides with the maximum likelihood estimation (MLE, Basu et al. [2011]).

While MDE provides a fruitful framework for statistical inference, when working with generative models, solving the optimization problem in (3.1) might be intractable since it might be impossible to evaluate the probability density function associated with μ_θ. Nevertheless, in various settings, even if the density is not available, one can still generate samples from the distribution μ_θ, and such samples turn out to be useful for making inference. More precisely, under such settings, a natural alternative to (3.1) is the minimum *expected* distance estimator [Bernton et al., 2019], which is defined as

$$\hat{\theta}_{n,m} = \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}\left[\mathbf{D}(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] \ , \tag{3.3}$$

where

$$\hat{\mu}_{\theta,m} = \frac{1}{m}\sum\nolimits_{i=1}^{m} \delta_{Z_i} \tag{3.4}$$

denotes the empirical distribution of $Z_{1:m}$, that is a sequence of i.i.d. random variables with distribution μ_θ. This algorithmic framework has computationally favorable properties since one can replace the expectation with a simple Monte Carlo average in practical applications.

In the context of MDE, distances that are based on optimal transport have become increasingly popular due to their computational and theoretical properties [Arjovsky et al., 2017, Tolstikhin et al., 2018, Genevay et al., 2018, Patrini et al., 2018, Adler and Lunz, 2018]. For instance, if we replace the distance $\mathbf{D}$ in (3.3) with the Wasserstein distance, we obtain the minimum expected Wasserstein estimator [Bassetti et al., 2006, Bernton et al., 2019]. In the classical statistical inference setting, the typical use of such an estimator is to infer the parameters of a measure whose density does not admit an analytical closed-form formula [Basu et al., 2011]. On the other hand, in the implicit generative modeling (IGM) setting, this estimator forms the basis of two popular IGM strategies: Wasserstein generative adversarial networks [Arjovsky et al., 2017] and Wasserstein auto-encoders [Tolstikhin et al., 2018]. These methods are related to each other according to [Genevay et al., 2017], and fall within the IGM framework explained in Section 1.1: the goal is to find the best parametric *transport map* T_θ, such that T_θ transforms a simple distribution μ (e.g., standard Gaussian or uniform) to a potentially complicated data distribution $\hat{\mu}_n$, by minimizing the Wasserstein distance between the transported distribution $\mu_\theta = T_{\theta\sharp}\mu$ and $\hat{\mu}_n$. In practice, θ is typically chosen as a neural network, for which it is often impossible to evaluate the induced density μ_θ. However, one can easily generate samples from μ_θ by first generating a sample from μ and then applying T_θ to that sample, making minimum expected distance estimation (3.3) feasible for this setting. Motivated by its practical success, the theoretical properties of this estimator have been recently taken under investigation [Bousquet et al., 2017, Liu et al., 2017] and very recently, Bernton et al. [2019] have established the consistency (for the general setting) and the asymptotic distribution (for one dimensional setting) of this estimator.

Even though estimation with the Wasserstein distance has served as a fertile ground for many generative modeling applications, except for the case when the measures are supported on $\mathbb{R}^1$, the computational complexity of minimum Wasserstein estimators rapidly becomes excessive with the increasing problem dimension, and developing accurate and efficient approximations is a highly non-trivial task. Therefore, there have been

several attempts to use more practical alternatives to the Wasserstein distance [Cuturi, 2013, Genevay et al., 2018], and in this context, the Sliced-Wasserstein distance has been an increasingly popular alternative to the Wasserstein distance. While several studies have reported empirical success on generative modeling with SW, the theoretical properties of such estimators have not yet been fully established, as we discussed in Section 1.3.

In this chapter, we investigate the asymptotic properties of estimators given in (3.1) and (3.3) when $\mathbf{D}$ is replaced with $\mathbf{SW}_p$. We first prove that convergence in SW implies weak convergence of probability measures defined on general domains, which generalizes the results given in [Bonnotte, 2013]. Then, by using similar techniques to the ones in [Bernton et al., 2019], we show that the estimators defined by (3.1) and (3.3) are consistent, meaning that as the number of observations n increases the estimates will get closer to the data-generating parameters. We finally prove a central limit theorem (CLT) in the multidimensional setting, which characterizes the asymptotic distribution of these estimators and establishes a convergence rate of $\sqrt{n}$. The CLT that we prove is stronger than the one derived in [Bernton et al., 2019] in the sense that it is not restricted to the one-dimensional setting, as opposed to [Bernton et al., 2019].

We support our theory with experiments that are conducted on both synthetic and real data. We first consider a more classical statistical inference setting, where we consider a Gaussian model and a multidimensional α-stable model whose density is not available in closed-form. In both models, the experiments validate our consistency and CLT results. We further observe that, especially for high-dimensional problems, the estimators obtained by minimizing SW have significantly better computational properties when compared to the ones obtained by minimizing the Wasserstein distance, as expected. In the IGM setting, we consider the neural network-based generative modeling algorithm proposed in [Deshpande et al., 2018] to show that our results also hold in the real data setting as well.

3.2 Asymptotic Guarantees for Minimum Sliced-Wasserstein Estimators

We first clarify the mathematical formalism for the problem of parameter inference in purely generative models. Let $(Y_k)_{k\in\mathbb{N}}$ be a sequence of random variables associated with observations, where each observation takes value in $\mathsf{Y} \subset \mathbb{R}^d$. We assume that these observations are i.i.d. from $\mu_\star \in \mathcal{P}(\mathsf{Y})$, and we consider the statistical model $\mathcal{M} = \{\mu_\theta \in \mathcal{P}(\mathsf{Y}),\ \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^{d_\theta}$ is the parametric space. For all $\theta \in \Theta$, we can generate i.i.d. samples $(Z_k)_{k\in\mathbb{N}^*} \in \mathsf{Y}^{\mathbb{N}^*}$ from μ_θ, but the associated likelihood is numerically intractable. The empirical approximation of μ_θ based on $m \in \mathbb{N}^*$ samples is then given by $\hat{\mu}_{\theta,m} = m^{-1}\sum_{i=1}^m \delta_{Z_i}$.

Throughout this chapter, we assume that the following conditions hold:

(**C1**) Y, endowed with the Euclidean distance ρ, is a Polish space,

(**C2**) Θ, endowed with the distance ρ_Θ, is a Polish space,

(**C3**) Θ is a σ-compact space, *i.e.* the union of countably many compact subspaces,

(**C4**) Parameters are identifiable, *i.e.* $\mu_\theta = \mu_{\theta'}$ implies $\theta = \theta'$.

We endow $\mathcal{P}(\mathsf{Y})$ with the Lévy-Prokhorov distance $\mathbf{d}_{\mathcal{P}}$, which metrizes the weak convergence by [Billingsley, 1999, Theorem 6.8] since Y is assumed to be a Polish space.

We define the *minimum Sliced-Wasserstein estimator* (MSWE) *of order* p as the estimator obtained by plugging $\mathbf{SW}_p$ in place of $\mathbf{D}$ in (3.1). Similarly, we define the *minimum expected Sliced-Wasserstein estimator* (MESWE) *of order* p as the estimator obtained by plugging $\mathbf{SW}_p$ in place of $\mathbf{D}$ in (3.3). In the rest of the chapter, MSWE and MESWE will be denoted by $\hat{\theta}_n$ and $\hat{\theta}_{n,m}$ respectively.

In what follows, we present the asymptotic properties that we derived for MSWE and MESWE, namely their existence, consistency and measurability. We also formulate a CLT that characterizes the asymptotic distribution of MSWE and establishes a convergence rate for any dimension. All the proofs for these results are provided in Sections 3.5.3 to 3.5.7.

Note that since the Sliced-Wasserstein distance is defined as an average of one-dimensional Wasserstein distances, some proofs are inevitably similar to the proofs done in [Bernton et al., 2019]. However, the adaptation of these techniques to the SW case is made possible by the identification of novel properties regarding the topology induced by the SW distance, which we establish for the first time in this study: see Sections 3.2.1 and 3.5.1.

3.2.1 Topology induced by the Sliced-Wasserstein distance

We begin this section by a useful result which we believe is interesting on its own and implies that the topology induced by $\mathbf{SW}_p$ on $\mathcal{P}_p(\mathbb{R}^d)$ is finer than the weak topology induced by the Lévy-Prokhorov metric $\mathbf{d}_{\mathcal{P}}$.

Theorem 3.1. *Let $p \in [1, +\infty)$. The convergence in $\mathbf{SW}_p$ implies the weak convergence in $\mathcal{P}(\mathbb{R}^d)$, i.e. if $(\mu_k)_{k\in\mathbb{N}}$ is a sequence of measures in $\mathcal{P}_p(\mathbb{R}^d)$ satisfying*

$$\lim_{k\to+\infty} \mathbf{SW}_p(\mu_k, \mu) = 0$$

with $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, then $(\mu_k)_{k\in\mathbb{N}} \xrightarrow{w} \mu$.

The property that convergence in $\mathbf{SW}_p$ implies weak convergence has already been proven in [Bonnotte, 2013] for *compact* domains only. While the implication of weak convergence is one of the most crucial requirements that a distance metric should satisfy, to the best of our knowledge, this implication has not been proved for general domains before. In [Bonnotte, 2013], the main proof technique was based on showing that $\mathbf{SW}_p$ is equivalent to $\mathbf{W}_p$ in compact domains, whereas we follow a different path and use the Lévy characterization: see Section 3.5.2 for the detailed proof.

3.2.2 Existence and consistency of MSWE and MESWE

In our next set of results, we will show that both MSWE and MESWE are consistent, in the sense that, when the number of observations n increases, the estimators will converge to a parameter $\theta_\star$ that minimizes the ideal problem $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$. Before we make this argument more precise, let us first present the assumptions that will imply our results.

A1. *The map $\theta \mapsto \mu_\theta$ is continuous from (Θ, ρ_Θ) to $(\mathcal{P}(\mathsf{Y}), \mathbf{d}_{\mathcal{P}})$, i.e. for any sequence $(\theta_n)_{n\in\mathbb{N}}$ in Θ satisfying $\lim_{n\to+\infty} \rho_\Theta(\theta_n, \theta) = 0$, then $(\mu_{\theta_n})_{n\in\mathbb{N}} \xrightarrow{w} \mu_\theta$.*

A 2. *The data-generating process is such that* $\lim_{n\to+\infty} \mathbf{SW}_p(\hat{\mu}_n, \mu_\star) = 0$, $\mathbb{P}$*-almost surely.*

A3. *There exists* $\epsilon > 0$*, such that setting* $\epsilon_\star = \inf_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)$*, the set*

$$\Theta_\epsilon^\star = \{\theta \in \Theta : \mathbf{SW}_p(\mu_\star, \mu_\theta) \leq \epsilon_\star + \epsilon\}$$

is bounded.

These assumptions are mostly related to the identifiability of the statistical model and the regularity of the data generating process. They are arguably mild assumptions, analogous to those that have already been considered in the literature [Bernton et al., 2019]. Note that, without Theorem 3.1, the formulation and use of **A**2 in our proofs would not be possible. In the next result, we establish the consistency of MSWE.

Theorem 3.2 (Existence and consistency of MSWE)**.** *Assume* ***A****1, **A****2 and **A****3. There exists* $\mathsf{E} \in \mathcal{F}$ *with* $\mathbb{P}(\mathsf{E}) = 1$ *such that, for all* $\omega \in \mathsf{E}$*,*

$$\lim_{n\to+\infty} \inf_{\theta\in\Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta), \quad \textit{and} \tag{3.5}$$

$$\limsup_{n\to+\infty} \operatorname{argmin}_{\theta\in\Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \subset \operatorname{argmin}_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)\ , \tag{3.6}$$

where $\hat{\mu}_n$ *is defined by* (3.2)*.*

Besides, for all $\omega \in \mathsf{E}$*, there exists* $n(\omega)$ *such that, for all* $n \geq n(\omega)$*, the set* $\operatorname{argmin}_{\theta\in\Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ *is non-empty.*

Our proof technique is similar to the one given in [Bernton et al., 2019]. This result shows that, when the number of observations goes to infinity, the estimate $\hat{\theta}_n$ will converge to a global minimizer of the problem $\min_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)$. In our next result, we prove a similar property for MESWEs as $\min(m, n)$ goes to infinity. In order to increase clarity, and without loss of generality, in this setting, we consider m as a function of n such that $\lim_{n\to+\infty} m(n) = +\infty$.

Now, we derive an analogous version of Theorem 3.2 for MESWE. For this result, we need to introduce another continuity assumption.

A4. *If* $\lim_{n\to+\infty} \rho_\Theta(\theta_n, \theta) = 0$*, then* $\lim_{n\to+\infty} \mathbb{E}\left[\mathbf{SW}_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n,n})|Y_{1:n}\right] = 0$*.*

The next theorem establishes the consistency of MESWE.

Theorem 3.3 (Existence and consistency of MESWE)**.** *Assume* ***A****1, **A****2, **A****3 and **A****4. Let* $(m(n))_{n\in\mathbb{N}^*}$ *be an increasing sequence satisfying* $\lim_{n\to+\infty} m(n) = +\infty$*. There exists a set* $\mathsf{E} \subset \Omega$ *with* $\mathbb{P}(\mathsf{E}) = 1$ *such that, for all* $w \in \mathsf{E}$*,*

$$\lim_{n\to+\infty} \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] = \inf_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta), \quad \textit{and} \tag{3.7}$$

$$\limsup_{n\to+\infty} \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \subset \operatorname{argmin}_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)\ , \tag{3.8}$$

where $\hat{\mu}_n$ *and* $\hat{\mu}_{\theta,m(n)}$ *are defined by* (3.2) *and* (3.4) *respectively.*

Besides, for all $\omega \in \mathsf{E}$*, there exists* $n(\omega)$ *such that, for all* $n \geq n(\omega)$*, the set* $\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)})|Y_{1:n}\right]$ *is non-empty.*

Similar to Theorem 3.2, this theorem shows that, when the number of observations goes to infinity, the estimator obtained with the expected distance will converge to a global minimizer.

3.2.3 Convergence of MESWE to MSWE

Since in practical applications, we can only use a finite number of generated samples $Z_{1:m}$, we analyze the case where the observations $Y_{1:n}$ are kept fixed while the number of generated samples increases, *i.e.* $m \to +\infty$. We show in this scenario that MESWE converges to MSWE, assuming the latter exists. Before deriving this result, we formulate a technical assumption below.

A5. *For some $\epsilon > 0$ and $\epsilon_n = \inf_{\theta\in\Theta} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$, the set*

$$\Theta_{\epsilon,n} = \{\theta \in \Theta : \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) \leq \epsilon_n + \epsilon\}$$

is bounded almost surely.

Theorem 3.4 (MESWE converges to MSWE as $m \to +\infty$)**.** *Assume **A**1, **A**4 and **A**5. Then,*

$$\lim_{m\to+\infty} \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] = \inf_{\theta\in\Theta} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) \tag{3.9}$$

$$\limsup_{m\to+\infty} \operatorname{argmin}_{\theta\in\Theta}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] \subset \operatorname{argmin}_{\theta\in\Theta}\mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) \tag{3.10}$$

There exists m^* such that, for any $m \geq m^*$, $\operatorname{argmin}_{\theta\in\Theta}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right]$ is a non-empty set.

This result shows that MESWE would be indeed promising in practice, as one get can more accurate estimations by increasing m.

3.2.4 Measurability of MSWE and MESWE

The measurability of the MSWE and MESWE follows from the application of [Brown and Purves, 1973, Corollary 1], also used in [Bassetti et al., 2006, Bernton et al., 2019], and which we recall in Theorem 3.18.

Theorem 3.5 (Measurability of the MSWE)**.** *Assume **A**1. For any $n \geq 1$ and $\epsilon > 0$, there exists a Borel measurable function $\hat{\theta}_{n,\epsilon} : \Omega \to \Theta$ that satisfies: for any $\omega \in \Omega$,*

$$\hat{\theta}_{n,\epsilon}(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\Theta} \ \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta), & \text{if this set is non-empty,} \\ \{\theta \in \Theta \ : \ \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \epsilon_\star + \epsilon\}, & \text{otherwise.} \end{cases}$$

where $\epsilon_\star = \inf_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)$.

We prove an analogous result to Theorem 3.5 in Section 3.5.6, which establishes the measurability of MESWE.

3.2.5 Rate of convergence and the asymptotic distribution

In our last set of theoretical results, we investigate the asymptotic distribution of MSWE and we establish a rate of convergence. We now suppose that we are in the *well-specified setting*, *i.e.* there exists $\theta_\star$ in the interior of Θ such that $\mu_{\theta_\star} = \mu_\star$, and we consider additional assumptions, as stated below.

For any $u \in \mathbb{S}^{d-1}$ and $t \in \mathbb{R}$, we define $F_\theta(u,t) = \int_{\mathsf{Y}} \mathbb{1}_{(-\infty,t]}(\langle u, y\rangle)\mathrm{d}\mu_\theta(y)$. Note that for any $u \in \mathbb{S}^{d-1}$, $F_\theta(u,\cdot)$ is the cumulative distribution function (CDF) associated to the measure $u^\star_\sharp\mu_\theta$.

A6. *For all $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\inf_{\theta \in \Theta:\ \rho_\Theta(\theta,\theta_\star) \geq \epsilon} \mathbf{SW}_1(\mu_{\theta_\star}, \mu_\theta) > \delta \ .$$

Let $\mathcal{L}^1(\mathbb{S}^{d-1} \times \mathbb{R}, \boldsymbol{\sigma} \otimes \mathrm{Leb}_1)$ denote the class of functions that are absolutely integrable on the domain $\mathbb{S}^{d-1} \times \mathbb{R}$ w.r.t. the product measure $\boldsymbol{\sigma} \otimes \mathrm{Leb}_1$, where Leb_1 denotes the Lebesgue measure on $\mathbb{R}$.

A7. *Assume that there exists a measurable function $D_\star = (D_{\star,1}, \ldots, D_{\star,d_\theta}) : \mathbb{S}^{d-1} \times \mathbb{R} \mapsto \mathbb{R}^{d_\theta}$ such that for each $i = 1, \ldots, d_\theta$, $D_{\star,i} \in \mathcal{L}^1(\mathbb{S}^{d-1} \times \mathbb{R}, \boldsymbol{\sigma} \otimes \mathrm{Leb}_1)$ and*

$$\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |F_\theta(u,t) - F_{\theta_\star}(u,t) - \langle \theta - \theta_\star, D_\star(u,t) \rangle| \, \mathrm{d}t \mathrm{d}\boldsymbol{\sigma}(u) = \boldsymbol{\epsilon}(\rho_\Theta(\theta, \theta_\star)) \ ,$$

where $\boldsymbol{\epsilon} : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies $\lim_{t \to 0} \boldsymbol{\epsilon}(t) = 0$. Besides, $\{D_{\star,i}\}_{i=1}^{d_\theta}$ are linearly independent in $\mathcal{L}^1(\mathbb{S}^{d-1} \times \mathbb{R}, \boldsymbol{\sigma} \otimes \mathrm{Leb}_1)$.

For any $u \in \mathbb{S}^{d-1}$, and $t \in \mathbb{R}$, define

$$\hat{F}_n(u,t) = n^{-1} \, \mathrm{card} \, \{ i \in \{1, \ldots, n\} \ : \ \langle u, Y_i \rangle \leq t \} \ ,$$

where card denotes the cardinality of a set, and for any $u \in \mathbb{S}^{d-1}$, $\hat{F}_n(u, \cdot)$ is the CDF associated to the measure $u^\star_\sharp \hat{\mu}_n$.

A8. *There exists a random element $G_\star : \mathbb{S}^{d-1} \times \mathbb{R} \mapsto \mathbb{R}$ such that the stochastic process $\sqrt{n}(\hat{F}_n - F_{\theta_\star})$ converges weakly in $\mathcal{L}^1(\mathbb{S}^{d-1} \times \mathbb{R}, \boldsymbol{\sigma} \otimes \mathrm{Leb}_1)$ to $G_\star$.*

Under mild assumptions on the tails of $u^\star_\sharp \mu_\star$ for any $u \in \mathbb{S}^{d-1}$, we believe that one can prove that **A**8 holds in general by extending [Dede, 2009, Proposition 3.5] and [del Barrio et al., 1999, Theorem 2.1a].

We can now formulate our central limit theorem based on these assumptions.

Theorem 3.6. *Assume **A**1, **A**2, **A**3, **A**6, **A**7 and **A**8. Then, the asymptotic distribution of the goodness-of-fit statistic is given by,*

$$\sqrt{n} \inf_{\theta \in \Theta} \mathbf{SW}_1(\hat{\mu}_n, \mu_\theta) \xrightarrow{w} \inf_{\theta \in \Theta} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t) \rangle| \, \mathrm{d}t \mathrm{d}\boldsymbol{\sigma}(u)$$

as $n \to +\infty$, where $\hat{\mu}_n$ is defined by (3.2).

Theorem 3.7. *Assume **A**1, **A**2, **A**3, **A**6, **A**7 and **A**8. Suppose also that the random map $\theta \mapsto \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t) \rangle| \, \mathrm{d}t \mathrm{d}\boldsymbol{\sigma}(u)$ has a unique infimum almost surely. Then, MSWE with $p = 1$ satisfies,*

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{w} \mathrm{argmin}_{\theta \in \Theta} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t) \rangle| \, \mathrm{d}t \mathrm{d}\boldsymbol{\sigma}(u)$$

as $n \to +\infty$, where $\hat{\theta}_n$ is defined by (3.1) with $\mathbf{SW}_1$ in place of $\mathbf{D}$.

These results show that the estimator and the associated goodness-of-fit statistics will converge to a random variable in distribution, where the rate of convergence is $\sqrt{n}$. Note that $G_\star$ is defined as a random element (see **A**8), therefore we cannot claim that

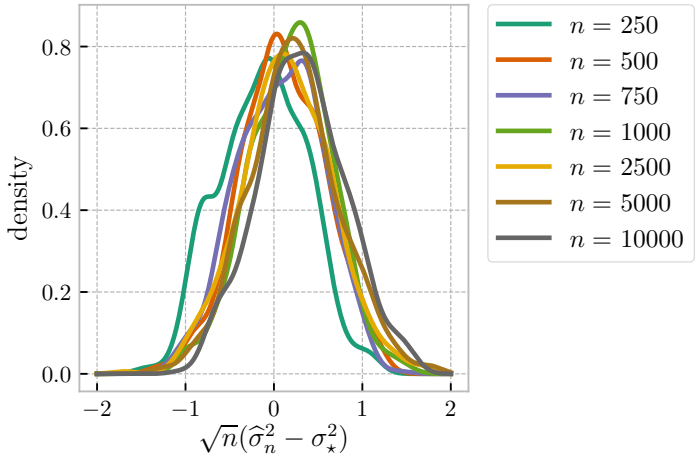


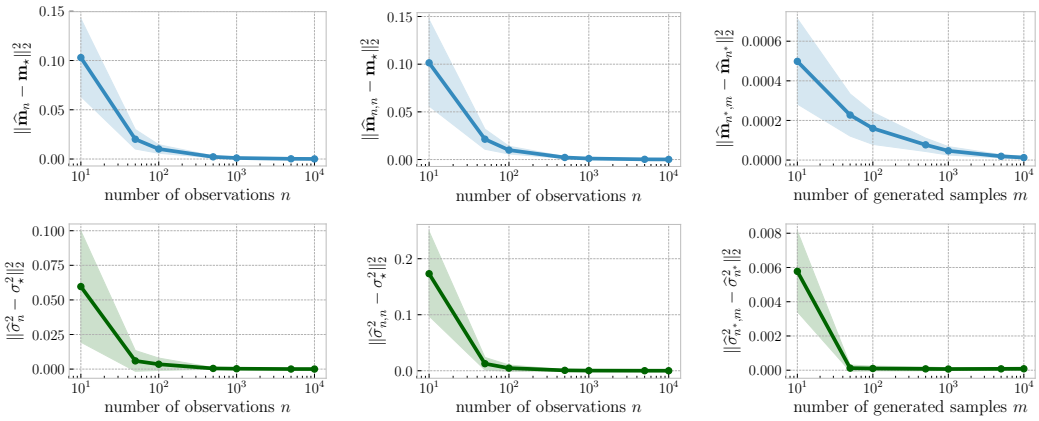
Figure 3.1: Probability density estimates of the MSWE $\hat{\sigma}_n^2$ of order 1, centered and rescaled by $\sqrt{n}$, on the 10-dimensional Gaussian model for different values of n.

the convergence in distribution derived in Theorem 3.6 and 3.7 implies the convergence in probability.

This CLT is also inspired by [Bernton et al., 2019], where they identified the asymptotic distribution associated to the minimum Wasserstein estimator. However, since $\mathbf{W}_p$ admits an analytical form only when $d = 1$, their result is restricted to the scalar case, and in their conclusion, Bernton et al. [2019] conjecture that the rate of the minimum Wasserstein estimators would depend negatively on the dimension of the observation space. On the contrary, since $\mathbf{SW}_p$ is defined in terms of one-dimensional $\mathbf{W}_p$ distances, we circumvent the curse of dimensionality and our result holds for any finite dimension. While the perceived computational burden has created a pessimism in the machine learning community about the use of Wasserstein-based methods in large dimensional settings, which motivated the rise of regularized optimal transport, we believe that our findings provide another interesting counter-example to this conception.

3.3 Experiments

We conduct experiments on synthetic and real data to empirically confirm our theorems. We explain in Section 3.5.8 the optimization methods used to find the estimators. Specifically, we can use stochastic iterative optimization algorithm (e.g., stochastic gradient descent). Note that, since we calculate (expected) SW with Monte Carlo approximations over a finite set of projections (and a finite number of *'generated datasets'*), MSWE and MESWE fall into the category of *doubly stochastic algorithms*. Our experiments on synthetic data actually show that using only one random projection and one randomly generated dataset at each iteration of the optimization process is enough to illustrate our theorems. We provide the code to reproduce the experiments[1].



(a) MSWE vs. n (b) MESWE vs. $n = m$ (c) MESWE vs. m

Figure 3.2: Min. SW estimation on Gaussians in $\mathbb{R}^{10}$. Figure 3.2a and Figure 3.2b show the mean squared error between $(\mathbf{m}_\star, \sigma_\star^2) = (\mathbf{0}, 1)$ and MSWE $(\hat{\mathbf{m}}_n, \hat{\sigma}_n^2)$ (resp. MESWE $(\hat{\mathbf{m}}_{n,n}, \hat{\sigma}_{n,n}^2)$) for n from 10 to 10 000, illustrating Theorems 3.2 and 3.3. Figure 3.2c shows the error between $(\hat{\mathbf{m}}_n, \hat{\sigma}_n^2)$ and $(\hat{\mathbf{m}}_{n,m}, \hat{\sigma}_{n,m}^2)$ for $n = 2000$ observations and m from 10 to 10 000, to illustrate Theorem 3.4. Results are averaged over 100 runs, the shaded areas represent the standard deviation.

3.3.1 Multivariate Gaussian distributions

We consider the task of estimating the parameters of a 10-dimensional Gaussian distribution using our SW estimators: we are interested in the model

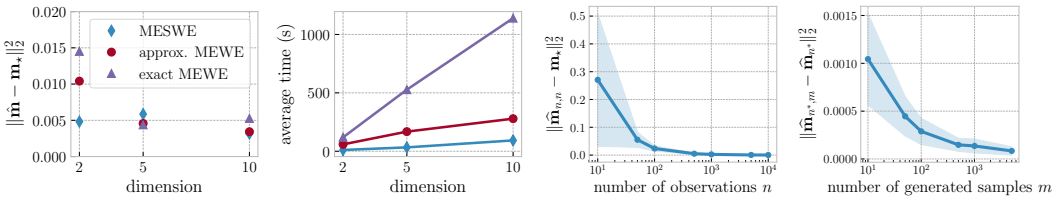
$$\mathcal{M} = \left\{ \mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{I}) \ : \ \mathbf{m} \in \mathbb{R}^{10}, \ \sigma^2 > 0 \right\} ,$$

and we draw i.i.d. observations with $(\mathbf{m}_\star, \sigma_\star^2) = (\mathbf{0}, 1)$. The advantage of this simple setting is that the density of the generated data has a closed-form expression, which makes MSWE tractable.

We empirically verify our central limit theorem: for different values of n, we compute 500 times MSWE of order 1 using one random projection, then we estimate the density of $\hat{\sigma}_n^2$ with a kernel density estimator. Figure 3.1 shows the distributions centered and rescaled by $\sqrt{n}$ for each n, and confirms the convergence rate that we derived (Theorem 3.7).

To illustrate the consistency property in Theorem 3.2, we approximate MSWE of order 2 for different numbers of observed data n using one random projection and we report for each n the mean squared error between the estimate mean and variance and the data-generating parameters $(\mathbf{m}_\star, \sigma_\star^2)$. We proceed the same way to study the consistency of MESWE (Theorem 3.3), which we approximate using one random projection and one generated dataset $z_{1:m}$ of size $m = n$ for different values of n. We also verify the convergence of MESWE to MSWE (Theorem 3.4): we compute these estimators on a fixed set of $n = 2000$ observations for different m, and we measure the error between them for each m. Results are shown in Figure 3.2. We see that our estimators indeed converge to $(\mathbf{m}_\star, \sigma_\star^2)$ as the number of observations increases (Figures 3.2a, 3.2b), and on a fixed observed dataset, MESWE converges to MSWE as we generate more samples (Figure 3.2c).

[1] See our GitHub repository: https://github.com/kimiandj/min_swe.



(a) Comparison of MESWE and MEWE (b) MESWE (c) MESWE (n^*=100)

Figure 3.3: Min. SW estimation for the location parameter of multivariate elliptically contoured stable distributions. Figure 3.3a compares the quality of the estimation provided by SW and Wasserstein-based estimators as well as their average computational time, for different values of dimension d. Figure 3.3b and Figure 3.3c illustrate, for $d = 10$, the consistency of MESWE $\hat{\mathbf{m}}_{n,m}$ and its convergence to the MSWE $\hat{\mathbf{m}}_n$. Results are averaged over 100 runs, the shaded area represent the standard deviation.

3.3.2 Multivariate elliptically contoured stable distributions

We focus on parameter inference for a subclass of multivariate stable distributions, called *elliptically contoured stable distributions* and denoted by $\mathcal{E}\alpha\mathcal{S}_c$ [Nolan, 2013]. Stable distributions refer to a family of heavy-tailed probability distributions that generalize Gaussian laws and appear as the limit distributions in the generalized central limit theorem [Samorodnitsky and Taqqu, 1994]. These distributions have many attractive theoretical properties and have been proven useful in modeling financial [Mandelbrot, 2013] data or audio signals [Şimşekli et al., 2015, Leglaive et al., 2017]. While special univariate cases include Gaussian, Lévy and Cauchy distributions, the density of stable distributions has no general analytic form, which restricts their practical application, especially for the multivariate case.

If $Y \in \mathbb{R}^d \sim \mathcal{E}\alpha\mathcal{S}_c(\mathbf{\Sigma}, \mathbf{m})$, then its joint characteristic function is defined for any $\mathbf{t} \in \mathbb{R}^d$ as

$$\mathbb{E}\left[\exp(i\mathbf{t}^T Y)\right] = \exp\left(-(\mathbf{t}^T\mathbf{\Sigma}\mathbf{t})^{\alpha/2} + i\mathbf{t}^T\mathbf{m}\right) ,$$

where $\mathbf{\Sigma}$ is a positive definite matrix (akin to a correlation matrix), $\mathbf{m} \in \mathbb{R}^d$ is a location vector (equal to the mean if it exists) and $\alpha \in (0, 2)$ controls the thickness of the tail. Even though their densities cannot be evaluated easily, it is straightforward to sample from $\mathcal{E}\alpha\mathcal{S}_c$ [Nolan, 2013], and we explain the sampling method in Section 3.5.8. Therefore, it is particularly relevant to apply MESWE instead of MSWE here.

To demonstrate the computational advantage of MESWE over the minimum expected Wasserstein estimator (MEWE, Bernton et al. [2019]), we consider observations in $\mathbb{R}^d$ i.i.d. from $\mathcal{E}\alpha\mathcal{S}_c(\mathbf{I}, \mathbf{m}_\star)$ where each component of $\mathbf{m}_\star$ is 2 and $\alpha = 1.8$, and

$$\mathcal{M} = \left\{\mathcal{E}\alpha\mathcal{S}_c(\mathbf{I}, \mathbf{m}) \ : \ \mathbf{m} \in \mathbb{R}^d\right\} .$$

The Wasserstein distance on multivariate data is either computed exactly by solving the linear program, or approximated by solving a regularized version of this problem with Sinkhorn's algorithm (Section 2.5). The MESWE is approximated using 10 random projections and 10 sets of generated samples. Then, following the approach in [Bernton et al., 2019], we use the gradient-free optimization method Nelder-Mead [Nelder and Mead, 1965] to minimize the Wasserstein and SW distances. We report on Figure 3.3a the mean squared error between each estimate and $\mathbf{m}_\star$, as well as their average computational time for different values of dimension d. We see that MESWE provides the same

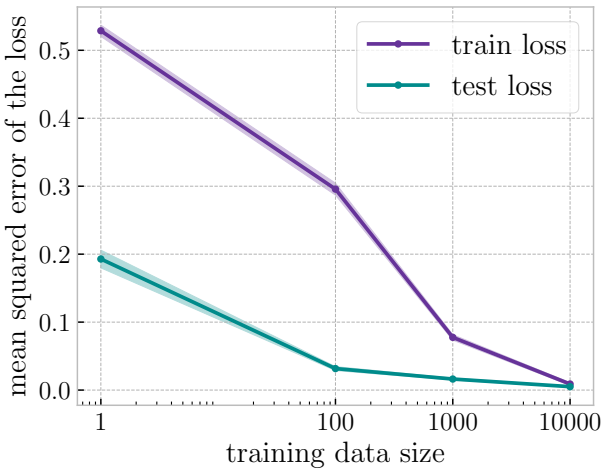


Figure 3.4: Mean-squared error between the training (test) loss for $(n, m) \in \{(1, 1), (100, 20), (1000, 40), (10\,000, 60)\}$ and the training (test) loss for $(n, m) = (60\,000, 200)$ on MNIST using the SW generator. We trained for 20 000 iterations with the ADAM optimizer Kingma and Ba [2015].

quality of estimation as its Wasserstein-based counterparts while considerably reducing the computational time, especially in higher dimensions.

We focus on this model in $\mathbb{R}^{10}$ and we illustrate the consistency of the MESWE $\hat{\mathbf{m}}_{n,m}$, approximated with one random projection and one generated dataset, the same way as for our previous Gaussian model: see Figure 3.3b. To confirm the convergence of $\hat{\mathbf{m}}_{n,m}$ to the MSWE $\hat{\mathbf{m}}_n$, we fix $n = 100$ observations and we compute the mean squared error between the two approximate estimators (using one random projection and one generated dataset) for different values of m (Figure 3.3c). Note that the MSWE is approximated with the MESWE obtained for a large enough value of m: $\hat{\mathbf{m}}_n \approx \hat{\mathbf{m}}_{n,10\,000}$.

3.3.3 High-dimensional real data using GANs

Finally, we run experiments on image generation using the Sliced-Wasserstein Generator (SWG), an alternative GAN formulation based on the minimization of the SW distance [Deshpande et al., 2018]. The goal is to optimize the neural network parameters such that the generated images are close to the observed ones. Deshpande et al. [2018] proposes to minimize the SW distance between μ_θ and the real data distribution over θ as the generator objective, and train on MESWE in practice.

For our experiments, we design a neural network with the fully-connected configuration given in [Deshpande et al., 2018, Appendix D] and we use the MNIST dataset [LeCun and Cortes, 2010], made of 60 000 training images and 10 000 test images of size 28×28. Our training objective is MESWE of order 2 approximated with 20 random projections and 20 different generated datasets. We study the consistent behavior of the MESWE by training the neural network on different sizes n of training data and different numbers m of generated samples, and by comparing the final training loss and test loss to the ones obtained when learning on the whole training dataset ($n = 60\,000$) and $m = 200$. Results are averaged over 10 runs and shown on Figure 3.4, where the shaded areas correspond to the standard deviation over the runs. We observe that our

results confirm Theorem 3.3.

3.4 Conclusion

The Sliced-Wasserstein distance has been an attractive metric choice for learning in generative models, where the densities cannot be computed directly. In this chapter, we investigated the asymptotic properties of estimators that are obtained by minimizing SW and the expected SW. We showed that (i) convergence in SW implies weak convergence of probability measures in general Wasserstein spaces, (ii) the estimators are consistent, and (iii) the estimators converge to a random variable in distribution with a rate of $\sqrt{n}$. We validated our mathematical results on both synthetic data and neural networks.

We would like to point out that, in all of our experiments, the random projections used in the Monte Carlo estimate of SW were picked uniformly on $\mathbb{S}^{d-1}$: see Section 3.5.8 for more details. The sampling on $\mathbb{S}^{d-1}$ directly impacts the quality of the resulting approximation of SW, and might induce variance in practice when learning generative models: this aspect is addressed in the next chapters, specifically in Chapters 5 to 7. On the theoretical side, studying the asymptotic properties of SW-based estimators obtained with a *finite* number of projections is an interesting question (e.g., their behavior might depend on the sampling method or the number of projections used). We leave this study for future research.

3.5 Appendix: Postponed Proofs and Experimental Details

3.5.1 Preliminary theoretical results

We first recall the definition of *epi-convergence* and gather technical results regarding the *lower semi-continuity* of (expected) Sliced-Wasserstein distances, which will be needed in our proofs.

Definition 3.8 (Epi-convergence). *Let Θ be a metric space and $f : \Theta \to \mathbb{R}$. Consider a sequence $(f_k)_{k\in\mathbb{N}}$ of functions from Θ to $\mathbb{R}$. We say that the sequence $(f_k)_{k\in\mathbb{N}}$ epi-converges to a function $f : \Theta \to \mathbb{R}$, and write $(f_k)_{k\in\mathbb{N}} \xrightarrow{e} f$, if for each $\theta \in \Theta$,*

$$\liminf_{k\to\infty} f_k(\theta_k) \geq f(\theta) \quad \textit{for every sequence } (\theta_k)_{n\in\mathbb{N}} \textit{ s.t. } \lim_{k\to+\infty} \theta_k = \theta \,,$$

$$\textit{and} \quad \limsup_{k\to\infty} f_k(\theta_k) \leq f(\theta) \quad \textit{for a sequence } (\theta_k)_{n\in\mathbb{N}} \textit{ s.t. } \lim_{k\to+\infty} \theta_k = \theta \,.$$

An equivalent and useful characterization of epi-convergence is given in [Rockafellar et al., 2009, Proposition 7.29], which we paraphrase in Proposition 3.10 after recalling the definition of lower semi-continuous functions.

Definition 3.9 (Lower semi-continuity). *Let Θ be a metric space and $f : \Theta \to \mathbb{R}$. We say that f is lower semi-continuous (l.s.c.) on Θ if for any $\theta_0 \in \Theta$,*

$$\liminf_{\theta\to\theta_0} f(\theta) \geq f(\theta_0)$$

Proposition 3.10 (Characterization of epi-convergence via minimization, Proposition 7.29 of Rockafellar et al. [2009]). *Let Θ be a metric space and $f : \Theta \to \mathbb{R}$ be a l.s.c. function. The sequence $(f_k)_{k\in\mathbb{N}}$, with $f_k : \Theta \to \mathbb{R}$ for any $n \in \mathbb{N}$, epi-converges to f if and only if*

(a) $\liminf_{k\to\infty} \inf_{\theta\in\mathsf{K}} f_k(\theta) \geq \inf_{\theta\in\mathsf{K}} f(\theta)$ *for every compact set* $\mathsf{K} \subset \Theta$;

(b) $\limsup_{k\to\infty} \inf_{\theta\in\mathsf{O}} f_k(\theta) \leq \inf_{\theta\in\mathsf{O}} f(\theta)$ *for every open set* $\mathsf{O} \subset \Theta$.

[Rockafellar et al., 2009, Theorem 7.31], paraphrased below, gives asymptotic properties for the infimum and argmin of epiconvergent functions and will be useful to prove the existence and consistency of our estimators.

Theorem 3.11 (Inf and argmin in epiconvergence, Theorem 7.31 of Rockafellar et al. [2009]). *Let* Θ *be a metric space,* $f : \Theta \to \mathbb{R}$ *be a l.s.c. function and* $(f_k)_{k\in\mathbb{N}}$ *be a sequence with* $f_k : \Theta \to \mathbb{R}$ *for any* $n \in \mathbb{N}$. *Suppose* $(f_k)_{k\in\mathbb{N}} \xrightarrow{e} f$ *with* $-\infty < \inf_{\theta\in\Theta} f(\theta) < \infty$.

(a) It holds $\lim_{k\to\infty} \inf_{\theta\in\Theta} f_k(\theta) = \inf_{\theta\in\Theta} f(\theta)$ *if and only if for every* $\eta > 0$ *there exists a compact set* $\mathsf{K} \subset \Theta$ *and* $N \in \mathbb{N}$ *such for any* $k \geq N$,

$$\inf_{\theta\in\mathsf{K}} f_k(\theta) \leq \inf_{\theta\in\Theta} f_k(\theta) + \eta \; .$$

(b) In addition, $\limsup_{k\to\infty} \operatorname{argmin}_{\theta\in\Theta} f_k(\theta) \subset \operatorname{argmin}_{\theta\in\Theta} f(\theta)$.

Now, we derive novel topological results regarding SW.

Lemma 3.12 (Lower semi-continuity of $\mathbf{SW}_p$). *Let* $p \in [1, \infty)$. *The Sliced-Wasserstein distance of order* p *is lower semi-continuous on* $\mathcal{P}_p(\mathsf{Y}) \times \mathcal{P}_p(\mathsf{Y})$ *endowed with the topology of weak convergence, i.e. for any sequences* $(\mu_k)_{k\in\mathbb{N}}$ *and* $(\nu_k)_{k\in\mathbb{N}}$ *of* $\mathcal{P}_p(\mathsf{Y})$ *which converge weakly to* $\mu \in \mathcal{P}_p(\mathsf{Y})$ *and* $\nu \in \mathcal{P}_p(\mathsf{Y})$ *respectively, we have*

$$\mathbf{SW}_p(\mu, \nu) \leq \liminf_{k\to+\infty} \mathbf{SW}_p\left(\mu_k, \nu_k\right) .$$

Proof. First, by the continuous mapping theorem, if a sequence $(\mu_k)_{k\in\mathbb{N}}$ of elements of $\mathcal{P}_p(\mathsf{Y})$ converges weakly to μ, then for any continuous function $f : \mathsf{Y} \to \mathbb{R}$, $(f_\sharp\mu_k)_{k\in\mathbb{N}}$ converges weakly to $f_\sharp\mu$. In particular, for any $u \in \mathbb{S}^{d-1}$, $u^\star_\sharp\mu_k \xrightarrow{w} u^\star_\sharp\mu$ since $u^\star$ is a bounded linear form thus continuous.

Let $p \in [1, \infty)$. We introduce the two sequences $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ of elements of $\mathcal{P}_p(\mathsf{Y})$ such that $\mu_k \xrightarrow{w} \mu$ and $\nu_k \xrightarrow{w} \nu$. We show that for any $u \in \mathbb{S}^{d-1}$,

$$\mathbf{W}_p^p(u^\star_\sharp\mu, u^\star_\sharp\nu) \leq \liminf_{k\to+\infty} \mathbf{W}_p^p(u^\star_\sharp\mu_k, u^\star_\sharp\nu_k) \; . \tag{3.11}$$

Indeed, if (3.11) holds, then the proof is completed using the definition of the Sliced-Wasserstein distance and Fatou's Lemma. Let $u \in \mathbb{S}^{d-1}$. For any $k \in \mathbb{N}$, let $\gamma_k \in \mathcal{P}(\mathbb{R} \times \mathbb{R})$ be an optimal transference plan between $u^\star_\sharp\mu_k$ and $u^\star_\sharp\nu_k$ for the Wasserstein distance of order p, which exists by [Villani, 2008, Theorem 4.1], *i.e.* $\mathbf{W}_p^p(u^\star_\sharp\mu_k, u^\star_\sharp\nu_k) = \int_{\mathbb{R}\times\mathbb{R}} |a - b| \, \mathrm{d}\gamma_k(a, b)$. Note that by [Villani, 2008, Lemma 4.4] and Prokhorov's Theorem, $(\gamma_k)_{k\in\mathbb{N}}$ is sequentially compact in $\mathcal{P}(\mathbb{R} \times \mathbb{R})$ for the topology associated with the weak convergence. Now, consider a subsequence $(\gamma_{\phi_1(k)})_{k\in\mathbb{N}}$ where $\phi_1 : \mathbb{N} \to \mathbb{N}$ is increasing such that

$$\begin{aligned} \lim_{k\to+\infty} \int_{\mathbb{R}\times\mathbb{R}} |a - b|^p \, \mathrm{d}\gamma_{\phi_1(k)}(a, b) &= \lim_{k\to+\infty} \mathbf{W}_p^p(u^\star_\sharp\mu_{\phi_1(k)}, u^\star_\sharp\nu_{\phi_1(k)}) \\ &= \liminf_{k\to+\infty} \mathbf{W}_p^p(u^\star_\sharp\mu_k, u^\star_\sharp\nu_k) \; . \end{aligned} \tag{3.12}$$

Since $(\gamma_k)_{k\in\mathbb{N}}$ is sequentially compact, $(\gamma_{\phi_1(k)})_{k\in\mathbb{N}}$ is sequentially compact as well, so there exists an increasing function $\phi_2 : \mathbb{N} \to \mathbb{N}$ and a probability distribution $\gamma \in \mathcal{P}(\mathbb{R}\times\mathbb{R})$ such that $(\gamma_{\phi_2(\phi_1(k))})_{k\in\mathbb{N}}$ converges weakly to γ. Then, we obtain by (3.12),

$$\begin{aligned}\int_{\mathbb{R}\times\mathbb{R}} \|a-b\|^p \,\mathrm{d}\gamma(a,b) &= \lim_{k\to+\infty} \int_{\mathbb{R}\times\mathbb{R}} \|a-b\|^p \,\mathrm{d}\gamma_{\phi_2(\phi_1(k))}(a,b) \\ &= \liminf_{k\to+\infty} \mathbf{W}_p^p(u_\sharp^\star\mu_k, u_\sharp^\star\nu_k) \ .\end{aligned}$$

If we show that $\gamma \in \Gamma(u_\sharp^\star\mu, u_\sharp^\star\nu)$, it will conclude the proof of (3.11) by definition of the Wasserstein distance (Definition 2.6). But for any continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$, since for any $k \in \mathbb{N}$, $\gamma_k \in \Gamma(\mu_k, \nu_k)$, and $(\mu_k)_{k\in\mathbb{N}}, (\nu_k)_{k\in\mathbb{N}}$ converge weakly to μ and ν respectively, we have

$$\begin{aligned}\int_{\mathbb{R}\times\mathbb{R}} f(a)\mathrm{d}\gamma(a,b) &= \lim_{k\to+\infty} \int_{\mathbb{R}\times\mathbb{R}} f(a)\mathrm{d}\gamma_{\phi_2(\phi_1(k))}(a,b) \\ &= \lim_{k\to+\infty} \int_{\mathbb{R}} f(a)\mathrm{d}u_\sharp^\star\mu_{\phi_2(\phi_1(k))}(a) \\ &= \int_{\mathbb{R}} f(a)\mathrm{d}u_\sharp^\star\mu(a) \ ,\end{aligned}$$

and similarly,

$$\int_{\mathbb{R}\times\mathbb{R}} f(b)\mathrm{d}\gamma(a,b) = \int_{\mathbb{R}} f(b)\mathrm{d}u_\sharp^\star\nu(a) \ .$$

This shows that $\gamma \in \Gamma(u_\sharp^\star\mu, u_\sharp^\star\nu)$ and therefore, (3.11) is true. We conclude by applying Fatou's Lemma.

□

By a direct application of Lemma 3.12, we obtain the following corollary.

Corollary 3.13. *Assume **A**1. Then, $(\mu,\theta) \mapsto \mathbf{SW}_p(\mu,\mu_\theta)$ is lower semi-continuous in $\mathcal{P}_p(\mathsf{Y})\times\Theta$.*

Next, we establish analogous properties for the expected SW distance.

Lemma 3.14 (Lower semi-continuity of $\mathbb{E}\mathbf{SW}_p$)**.** *Let $p \in [1,\infty)$ and $m \in \mathbb{N}^*$. Denote for any $\mu \in \mathcal{P}_p(\mathsf{Y})$, $\hat{\mu}_m = (1/m)\sum_{i=1}^m \delta_{Z_i}$, where $Z_{1:m}$ are i.i.d. samples from μ. Then, the map $(\nu,\mu) \mapsto \mathbb{E}\left[\mathbf{SW}_p(\nu,\hat{\mu}_m)\right]$ is lower semi-continuous on $\mathcal{P}_p(\mathsf{Y})\times\mathcal{P}_p(\mathsf{Y})$ endowed with the topology of weak convergence.*

Proof. We consider two sequences $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ of probability measures in Y, such that $(\mu_k)_{k\in\mathbb{N}} \xrightarrow{w} \mu$ and $(\nu_k)_{k\in\mathbb{N}} \xrightarrow{w} \nu$, and we fix $m \in \mathbb{N}^*$.

By Skorokhod's representation theorem, there exists a probability space $(\tilde{\Omega},\tilde{\mathcal{F}},\tilde{\mathbb{P}})$, a sequence of random variables $(\tilde{X}_k^1,\dots,\tilde{X}_k^m)_{k\in\mathbb{N}}$ and a random variable $(\tilde{X}^1,\dots,\tilde{X}^m)$ defined on $\tilde{\Omega}$ such that for any $k \in \mathbb{N}$ and $i \in \{1,\dots,m\}$, $\tilde{X}_k^i$ has distribution μ_k, $\tilde{X}^i$ has distribution μ and $(\tilde{X}_k^1,\dots,\tilde{X}_k^m)_{k\in\mathbb{N}^*}$ converges to $(\tilde{X}^1,\dots,\tilde{X}^m)$, $\tilde{\mathbb{P}}$-almost surely. We then show that the sequence of (random) empirical distributions $(\hat{\mu}_{k,m})_{k\in\mathbb{N}}$ defined by $\hat{\mu}_{k,m} = (1/m)\sum_{i=1}^m \delta_{\tilde{X}_k^i}$, weakly converges to $\hat{\mu}_m = (1/m)\sum_{i=1}^m \delta_{\tilde{X}^i}$, $\tilde{\mathbb{P}}$-almost surely. Note that it is sufficient to show that for any deterministic sequence $(x_k^1,\dots,x_k^m)_{k\in\mathbb{N}^*}$ which converges to $(x^1,\dots,x^m)$, *i.e.* $\lim_{k\to+\infty}\max_{i\in\{1,\dots,m\}}\rho(x_k^i,x^i) = 0$, then the sequence of empirical distributions $(\hat{\nu}_{k,m})_{k\in\mathbb{N}}$ defined by $\hat{\nu}_{k,m} = (1/m)\sum_{i=1}^m \delta_{x_k^i}$, weakly converges to $\hat{\nu}_m = (1/m)\sum_{i=1}^m \delta_{x^i}$. Since the Lévy-Prokhorov metric $\mathbf{d}_{\mathcal{P}}$ metrizes

the weak convergence by [Billingsley, 1999, Theorem 6.8], we only need to show that $\lim_{k\to+\infty} \mathbf{d}_{\mathcal{P}}(\hat{\nu}_{k,m}, \hat{\nu}_m) = 0$. More precisely, since for any probability measure ζ_1 and ζ_2,

$$\begin{aligned}&\mathbf{d}_{\mathcal{P}}(\zeta_1, \zeta_2)\\ &\quad = \inf\left\{\epsilon > 0 \,:\, \text{for any } \mathsf{A} \in \mathcal{B}(\mathsf{Y}),\, \zeta_1(\mathsf{A}) \leq \zeta_2(\mathsf{A}^\epsilon) + \epsilon \text{ and } \zeta_2(\mathsf{A}) \leq \zeta_1(\mathsf{A}^\epsilon) + \epsilon\right\} ,\end{aligned}$$

where $\mathcal{B}(\mathsf{Y})$ is the Borel σ-field of (Y, ρ) and for any $\mathsf{A} \in \mathcal{B}(\mathsf{Y})$, $\mathsf{A}^\epsilon = \{x \in \mathsf{Y} \,:\, \rho(x, y) < \epsilon \text{ for any } y \in \mathsf{A}\}$, we get

$$\mathbf{d}_{\mathcal{P}}(\hat{\nu}_{k,m}, \hat{\nu}_m) \leq 2 \max_{i\in\{1,\ldots,m\}} \rho(x_k^i, x^i) ,$$

and therefore $\lim_{k\to+\infty} \mathbf{d}_{\mathcal{P}}(\hat{\nu}_{k,m}, \hat{\nu}_m) = 0$, so that, $(\hat{\nu}_{k,m})_{k\in\mathbb{N}}$ weakly converges to $\hat{\nu}_m$.

Finally, we have that $\hat{\mu}_{k,m} = (1/m)\sum_{i=1}^m \delta_{\tilde{X}_k^i}$, weakly converges to $\hat{\mu}_m = (1/m)\sum_{i=1}^m \delta_{\tilde{X}^i}$ $\tilde{\mathbb{P}}$-almost surely and we obtain the final result using the lower semi-continuity of the Sliced-Wasserstein distance derived in Lemma 3.12 and Fatou's lemma,

$$\tilde{\mathbb{E}}\left[\mathbf{SW}_p(\nu, \hat{\mu}_m)\right] \leq \tilde{\mathbb{E}}\left[\liminf_{i\to\infty} \mathbf{SW}_p(\nu_i, \hat{\mu}_{m,i})\right] \leq \liminf_{i\to\infty} \tilde{\mathbb{E}}\left[\{\mathbf{SW}_p(\nu_i, \hat{\mu}_{m,i})\right] ,$$

where $\tilde{\mathbb{E}}$ is the expectation corresponding to $\tilde{\mathbb{P}}$.

□

The following corollary is a direct consequence of Lemma 3.14.

Corollary 3.15. *Assume **A**1. Then, $(\nu, \theta) \mapsto \mathbb{E}[\mathbf{SW}_p(\nu, \hat{\mu}_{\theta,m})|Y_{1:n}]$ is lower semi-continuous on $\mathcal{P}(\mathsf{Y}) \times \Theta$.*

3.5.2 Proof of Theorem 3.1

Lemma 3.16. *Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence of probability measures on $\mathbb{R}^d$ and μ a measure in $\mathbb{R}^d$ such that*

$$\lim_{k\to\infty} \mathbf{SW}_1(\mu_k, \mu) = 0 .$$

Then, there exists an increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that the subsequence $(\mu_{\phi(k)})_{k\in\mathbb{N}}$ converges weakly to μ.

Proof. By definition, we have that

$$\lim_{k\to\infty} \int_{\mathbb{S}^{d-1}} \mathbf{W}_1(u_\sharp^\star \mu_k, u_\sharp^\star \mu) \mathrm{d}\boldsymbol{\sigma}(u) = 0 .$$

Therefore, by [Bogachev, 2007, Theorem 2.2.5], there exists an increasing mapping $\phi : \mathbb{N} \to \mathbb{N}$ such that for $\boldsymbol{\sigma}$-almost every ($\boldsymbol{\sigma}$-a.e.) $u \in \mathbb{S}^{d-1}$,

$$\lim_{k\to\infty} \mathbf{W}_1(u_\sharp^\star \mu_{\phi(k)}, u_\sharp^\star \mu) = 0 .$$

By [Villani, 2008, Theorem 6.9], it implies that for $\boldsymbol{\sigma}$-a.e. $u \in \mathbb{S}^{d-1}$,

$$(u_\sharp^\star \mu_{\phi(k)})_{k\in\mathbb{N}} \xrightarrow{w} u_\sharp^\star \mu .$$

Lévy's characterization [Kallenberg, 1997, Theorem 4.3] gives that, for $\boldsymbol{\sigma}$-a.e. $u \in \mathbb{S}^{d-1}$ and any $s \in \mathbb{R}$,

$$\lim_{k\to\infty} \Phi_{u_\sharp^\star \mu_{\phi(k)}}(s) = \Phi_{u_\sharp^\star \mu}(s) ,$$

where, for any distribution $\nu \in \mathcal{P}(\mathbb{R}^p)$, Φ_ν denotes the characteristic function of ν and is defined for any $v \in \mathbb{R}^p$ as

$$\Phi_\nu(v) = \int_{\mathbb{R}^p} \mathrm{e}^{\mathrm{i}\langle v,w\rangle} \mathrm{d}\nu(w) \; .$$

Then, we can conclude that for Lebesgue-almost every $z \in \mathbb{R}^d$,

$$\lim_{k\to\infty} \Phi_{\mu_{\phi(k)}}(z) = \Phi_\mu(z) \; . \tag{3.13}$$

We can now show that $(\mu_{\phi(k)})_{k\in\mathbb{N}} \xrightarrow{w} \mu$, *i.e.* for any $f : \mathbb{R}^d \to \mathbb{R}$ continuous with compact support,

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu_n(z) = \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu(z) \; . \tag{3.14}$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function with compact support and $\sigma > 0$. Consider the function f_σ defined for any $x \in \mathbb{R}^d$ as

$$f_\sigma(x) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} f(x-z) \exp\left(-\frac{\|z\|^2}{2\sigma^2}\right) \mathrm{dLeb}_d(z) = f * g_\sigma(x) \; ,$$

where $*$ denotes the convolution product and g_σ is the density of the d-dimensional Gaussian with zero mean and covariance matrix $\sigma^2 \mathbf{I}_d$. We first show that (3.14) holds with f_σ in place of f. Since for any $z \in \mathbb{R}^d$,

$$\mathbb{E}\left[\exp(\mathrm{i}\,\langle G, z\rangle)\right] = \exp\left(\mathrm{i}\,\langle \mathtt{m}, z\rangle + \frac{\|z\|^2}{2\sigma^2}\right)$$

if G is a d-dimensional Gaussian random variable with zero mean and covariance matrix $(1/\sigma^2)\,\mathrm{I}_d$, then by Fubini's theorem, we get for any $k \in \mathbb{N}$,

$$\begin{aligned}
\int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu_{\phi(k)}(z) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w) g_\sigma(z-w) \mathrm{d}w \mathrm{d}\mu_{\phi(k)}(z) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)(2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \mathrm{e}^{i\langle z-w,x\rangle} g_{1/\sigma}(x) \mathrm{d}x \mathrm{d}w \mathrm{d}\mu_{\phi(k)}(z) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (2\pi\sigma^2)^{-d/2} f(w) \mathrm{e}^{-i\langle w,x\rangle} g_{1/\sigma}(x) \Phi_{\mu_{\phi(k)}}(x) \mathrm{d}x \mathrm{d}w \\
&= (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_{\mu_{\phi(k)}}(x) \mathrm{d}x \; ,
\end{aligned} \tag{3.15}$$

where $\mathcal{F}[f](x) = \int_{\mathbb{R}^d} f(w) \mathrm{e}^{\mathrm{i}\langle w,x\rangle} \mathrm{d}w$ denotes the Fourier transform of f, which exists since f is assumed to have a compact support. In an analogous manner, we prove that

$$\int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu(z) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_\mu(x) \mathrm{d}x \; . \tag{3.16}$$

Now, using that $\mathcal{F}[f]$ is bounded by $\int_{\mathbb{R}^d} |f(w)| \mathrm{d}w < +\infty$ since f has compact support, we obtain that, for any $k \in \mathbb{N}$ and $x \in \mathbb{R}^d$,

$$\left|\mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_{\mu_{\phi(k)}}(x)\right| \leq g_{1/\sigma}(x) \int_{\mathbb{R}^d} |f(w)| \mathrm{d}w \; .$$

By (3.13), (3.15), (3.16) and Lebesgue's Dominated Convergence Theorem, we obtain

$$\lim_{k\to\infty}\int_{\mathbb{R}^d}(2\pi\sigma^2)^{-d/2}\mathcal{F}[f](x)g_{1/\sigma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x=\int_{\mathbb{R}^d}(2\pi\sigma^2)^{-d/2}\mathcal{F}[f](x)g_{1/\sigma}(x)\Phi_{\mu}(x)\mathrm{d}x$$

$$\lim_{k\to\infty}\int_{\mathbb{R}^d}f_\sigma(z)\mathrm{d}\mu_{\phi(k)}(z)=\int_{\mathbb{R}^d}f_\sigma(z)\mathrm{d}\mu(z)\;. \tag{3.17}$$

We can now complete the proof of (3.14). For any $\sigma>0$, we have

$$\left|\int_{\mathbb{R}^d}f(z)\mathrm{d}\mu_{\phi(k)}(z)-\int_{\mathbb{R}^d}f(z)\mathrm{d}\mu(z)\right|$$
$$\leq 2\sup_{z\in\mathbb{R}^d}|f(z)-f_\sigma(z)|+\left|\int_{\mathbb{R}^d}f_\sigma(z)\mathrm{d}\mu_{\phi(k)}(z)-\int_{\mathbb{R}^d}f_\sigma(z)\mathrm{d}\mu(z)\right|\;.$$

Therefore, by (3.17), for any $\sigma>0$, we get

$$\limsup_{k\to+\infty}\left|\int_{\mathbb{R}^d}f(z)\mathrm{d}\mu_{\phi(k)}(z)-\int_{\mathbb{R}^d}f(z)\mathrm{d}\mu(z)\right|\leq 2\sup_{z\in\mathbb{R}^d}|f(z)-f_\sigma(z)|\;.$$

Finally, [Folland, 1999, Theorem 8.14-b] implies that

$$\lim_{\sigma\to 0}\sup_{z\in\mathbb{R}^d}|f_\sigma(z)-f(z)|=0\;,$$

which concludes the proof.

□

Proof of Theorem 3.1. Now, assume that

$$\lim_{k\to\infty}\mathbf{SW}_p(\mu_k,\mu)=0 \tag{3.18}$$

and that $(\mu_k)_{k\in\mathbb{N}}$ does not converge weakly to μ. Therefore, $\lim_{k\to\infty}\mathbf{d}_{\mathcal{P}}(\mu_k,\mu)\neq 0$, where $\mathbf{d}_{\mathcal{P}}$ denotes the Lévy-Prokhorov metric, and there exists $\epsilon>0$ and a subsequence $(\mu_{\psi(k)})_{k\in\mathbb{N}}$ with $\psi:\mathbb{N}\to\mathbb{N}$ increasing, such that for any $k\in\mathbb{N}$,

$$\mathbf{d}_{\mathcal{P}}(\mu_{\psi(k)},\mu)>\epsilon \tag{3.19}$$

In addition, by Hölder's inequality, we know that $\mathbf{W}_1(\mu_k,\mu)\leq\mathbf{W}_p(\mu_k,\mu)$, thus $\mathbf{SW}_1(\mu_k,\mu)\leq\mathbf{SW}_p(\mu_k,\mu)$, and by (3.18), $\lim_{k\to\infty}\mathbf{SW}_1(\mu_{\psi(k)},\mu)=0$. Then, according to Lemma 3.16, there exists a subsequence $(\mu_{\phi(\psi(k))})_{k\in\mathbb{N}}$ with $\phi:\mathbb{N}\to\mathbb{N}$ increasing, such that

$$\mu_{\phi(\psi(k))}\xrightarrow{w}\mu$$

which is equivalent to $\lim_{k\to\infty}\mathbf{d}_{\mathcal{P}}(\mu_{\phi(\psi(k))},\mu)=0$, thus contradicts (3.19). We conclude that (3.18) implies $(\mu_k)_{k\in\mathbb{N}}\xrightarrow{w}\mu$.

□

3.5.3 Proof of Theorem 3.2

This result is proved analogously to [Bernton et al., 2019, Theorem 2.1]. The key step is to show that the function $\theta\mapsto\mathbf{SW}_p(\hat{\mu}_n,\mu_\theta)$ epi-converges to $\theta\mapsto\mathbf{SW}_p(\mu_\star,\mu_\theta)$ $\mathbb{P}$-almost surely, and then apply [Rockafellar et al., 2009, Theorem 7.31], recalled in Theorem 3.11.

Proof of Theorem 3.2. First, by **A**1 and Corollary 3.13, the map $\theta \mapsto \mathbf{SW}_p(\mu, \mu_\theta)$ is l.s.c. on Θ for any $\mu \in \mathcal{P}_p(\mathsf{Y})$. Therefore, by **A**3, there exists $\theta_\star \in \Theta$ such that $\mathbf{SW}_p(\mu_\star, \mu_{\theta_\star}) = \epsilon_\star$, and the set $\Theta_\epsilon^\star$ is non-empty as it contains $\theta_\star$, closed by lower semi-continuity of $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$, and bounded. $\Theta_\epsilon^\star$ is thus compact, and we conclude again by lower semi-continuity that the set $\operatorname{argmin}_{\theta\in\Theta}\mathbf{SW}_p(\mu_\star, \mu_\theta)$ is non-empty [Aliprantis et al., 1999, Theorem 2.43].

Consider the event given by **A**2, $\mathsf{E} \in \mathcal{F}$ such that $\mathbb{P}(\mathsf{E}) = 1$ and for any $\omega \in \mathsf{E}$, $\lim_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) = 0$. Then, we prove that $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$ epi-converges to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$ $\mathbb{P}$-almost surely using the characterization in [Rockafellar et al., 2009, Proposition 7.29], *i.e.* we verify that, for any $\omega \in \mathsf{E}$, the two conditions below hold: for every compact set $\mathsf{K} \subset \Theta$ and every open set $\mathsf{O} \subset \Theta$,

$$\begin{aligned} \liminf_{n\to\infty} \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) &\geq \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \\ \limsup_{n\to\infty} \inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) &\leq \inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \ . \end{aligned} \tag{3.20}$$

We fix ω in E. Let $\mathsf{K} \subset \Theta$ be a compact set. By lower semi-continuity of $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$, there exists $\theta_n = \theta_n(\omega) \in \mathsf{K}$ such that for any $n \in \mathbb{N}$,

$$\inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) = \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_{\theta_n}) \ . \tag{3.21}$$

We consider the subsequence $(\hat{\mu}_{\phi(n)})_{n\in\mathbb{N}}$ where $\phi : \mathbb{N} \to \mathbb{N}$ is increasing such that $\mathbf{SW}_p(\hat{\mu}_{\phi(n)}(\omega), \mu_{\theta_{\phi(n)}})$ converges to $\liminf_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_{\theta_n})$, which is equal to

$$\liminf_{n\to\infty} \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$$

by (3.21). Since K is compact, there also exists an increasing function $\psi : \mathbb{N} \to \mathbb{N}$ such that, for $\bar{\theta} \in \mathsf{K}$, $\lim_{n\to\infty} \rho_\Theta(\theta_{\psi(\phi(n))}, \bar{\theta}) = 0$. Therefore,

$$\begin{aligned} \liminf_{n\to\infty} \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) &= \lim_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_{\phi(n)}(\omega), \mu_{\theta_{\phi(n)}}) \\ &= \lim_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_{\psi(\phi(n))}(\omega), \mu_{\theta_{\psi(\phi(n))}}) \\ &= \liminf_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_{\psi(\phi(n))}(\omega), \mu_{\theta_{\psi(\phi(n))}}) \\ &\geq \mathbf{SW}_p(\mu_\star, \mu_{\bar{\theta}}) \qquad (3.22) \\ &\geq \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \ , \end{aligned}$$

where (3.22) is obtained by lower semi-continuity since $\hat{\mu}_{\psi(\phi(n))}(\omega) \xrightarrow{w} \mu_\star$ by **A**2 and Theorem 3.1, and $\mu_{\theta_{\psi(\phi(n))}} \xrightarrow{w} \mu_{\bar{\theta}}$ by **A**1. We conclude that the first condition in (3.20) holds.

Now, we fix $\mathsf{O} \subset \Theta$ open. There exists a sequence $(\theta_n)_{n\in\mathbb{N}}$ in O such that the sequence $\{\mathbf{SW}_p(\mu_\star, \mu_{\theta_n})\}_{n\in\mathbb{N}}$ converges to $\inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\mu_\star, \mu_\theta)$ as $n \to +\infty$, and for $n \in \mathbb{N}$, $\inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_{\theta_n})$, by definition of the infimum. Then,

$$\begin{aligned} &\limsup_{n\to\infty} \inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \\ &\leq \limsup_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_{\theta_n}) \\ &\leq \limsup_{n\to\infty} \left(\mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathbf{SW}_p(\mu_\star, \mu_{\theta_n})\right) \quad \text{(by the triangle inequality)} \\ &\leq \limsup_{n\to\infty} \mathbf{SW}_p(\mu_\star, \mu_{\theta_n}) \quad \text{(by } \mathbf{A}2\text{)} \\ &= \inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \quad \text{(by definition of } (\theta_n)_{n\in\mathbb{N}}\text{)} \ . \end{aligned}$$

This shows that the second condition in (3.20) holds, and hence, the sequence of functions $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ epi-converges to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$.

Next, we apply [Rockafellar et al., 2009, Theorem 7.31]. First, by [Rockafellar et al., 2009, Theorem 7.31(b)], (3.6) immediately follows from the epi-convergence of $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$. We then show that [Rockafellar et al., 2009, Theorem 7.31(a)] can be applied by proving that, for any $\eta > 0$, there exists a compact set $\mathsf{B} \subset \Theta$ and $N \in \mathbb{N}$ such that, for all $n \geq N$,

$$\inf_{\theta \in \mathsf{B}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) + \eta \; . \tag{3.23}$$

In fact, we simply show that there exists a compact set $\mathsf{B} \subset \Theta$ and $N \in \mathbb{N}$ such that, for all $n \geq N$,

$$\inf_{\theta \in \mathsf{B}} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \; .$$

On the one hand, the second condition in (3.20) gives us

$$\limsup_{n \to \infty} \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta \in \Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta) = \epsilon_\star \; .$$

We deduce that there exists $n_{\epsilon/4}(\omega)$ such that, for $n \geq n_{\epsilon/4}(\omega)$,

$$\inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \epsilon_\star + \epsilon/4 \; ,$$

where ϵ is given by **A**3. As $n \geq n_{\epsilon/4}(\omega)$, the set $\widehat{\Theta}_{\epsilon/2} = \{\theta \in \Theta : \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \epsilon_\star + \frac{\epsilon}{2}\}$ is non-empty since it contains θ^* defined as

$$\mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_{\theta^*}) = \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \; .$$

On the other hand, by **A**2, there exists $n_{\epsilon/2}(\omega)$ such that, for $n \geq n_{\epsilon/2}(\omega)$,

$$\mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) \leq \frac{\epsilon}{2} \; . \tag{3.24}$$

Let $n \geq n_*(\omega) = \max\{n_{\epsilon/4}(\omega), n_{\epsilon/2}(\omega)\}$ and $\theta \in \widehat{\Theta}_{\epsilon/2}$. By the triangle inequality,

$$\begin{aligned} \mathbf{SW}_p(\mu_\star, \mu_\theta) &\leq \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \\ &\leq \epsilon_\star + \epsilon \qquad \text{(since } \theta \in \widehat{\Theta}_{\epsilon/2} \text{ and by (3.24))} \end{aligned}$$

This means that, when $n \geq n_*(\omega)$, $\widehat{\Theta}_{\epsilon/2} \subset \Theta^\star_\epsilon$, and since $\inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ is attained in $\widehat{\Theta}_{\epsilon/2}$, we have

$$\inf_{\theta \in \Theta^\star_\epsilon} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \; . \tag{3.25}$$

As shown in the first part of the proof $\Theta^\star_\epsilon$ is compact and then by [Rockafellar et al., 2009, Theorem 7.31(a)], (3.5) is a direct consequence of (3.23)-(3.25) and the epi-convergence of $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$.

Finally, by the same arguments used in this proof for $\operatorname{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)$, the set $\operatorname{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ is non-empty for $n \geq n_*(\omega)$.

□

3.5.4 Proof of Theorem 3.3

This result is proved analogously to [Bernton et al., 2019, Theorem 2.4]. The key step is to show that the function $\theta \mapsto \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)})|Y_{1:n}]$ epi-converges to $\theta \mapsto \mathbb{E}[\mathbf{SW}_p(\mu_\star, \mu_\theta)|Y_{1:n}]$, and then apply [Rockafellar et al., 2009, Theorem 7.31], which we recall in Theorem 3.11.

Proof of Theorem 3.3. Since we assume **A**1 and **A**3, we can apply the same reasoning as in the proof of Theorem 3.2 to show that $\operatorname{argmin}_{\theta\in\Theta}\mathbf{SW}_p(\mu_\star, \mu_\theta)$ is a non-empty set.

Next, consider the event given by **A**2, $\mathsf{E} \in \mathcal{F}$ such that $\mathbb{P}(\mathsf{E}) = 1$ and for any $\omega \in \mathsf{E}$, $\lim_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) = 0$. Then, we prove that $\theta \mapsto \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)})|Y_{1:n}]$ epi-converges to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$ $\mathbb{P}$-almost surely using the characterization of [Rockafellar et al., 2009, Proposition 7.29], *i.e.* we verify that, for any $\omega \in \mathsf{E}$, the two conditions below hold: for every compact set $\mathsf{K} \subset \Theta$ and for every open set $\mathsf{O} \subset \Theta$,

$$\begin{aligned} \liminf_{n\to+\infty} \inf_{\theta\in\mathsf{K}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] &\geq \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \\ \limsup_{n\to+\infty} \inf_{\theta\in\mathsf{O}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] &\leq \inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \end{aligned} \tag{3.26}$$

We fix ω in E. Let $\mathsf{K} \subset \Theta$ be a compact set. By **A**1 and Corollary 3.15, the mapping $\theta \mapsto \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})|Y_{1:n}]$ is l.s.c., so there exists $\theta_n = \theta_n(\omega) \in \mathsf{K}$ such that for any $n \in \mathbb{N}$,

$$\inf_{\theta\in\mathsf{K}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] = \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)})\middle|Y_{1:n}\right] .$$

We consider the subsequence $(\hat{\mu}_{\phi(n)})_{n\in\mathbb{N}}$ where $\phi : \mathbb{N} \to \mathbb{N}$ is increasing such that $\mathbb{E}[\mathbf{SW}_p(\hat{\mu}_{\phi(n)}(\omega), \hat{\mu}_{\theta_{\phi(n)},m(\phi(n))})|Y_{1:n}]$ converges to

$$\liminf_{n\to\infty} \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_{n,m(n)}})|Y_{1:n}] = \liminf_{n\to\infty} \inf_{\theta\in\mathsf{K}} \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})|Y_{1:n}] .$$

Since K is compact, there also exists an increasing function $\psi : \mathbb{N} \to \mathbb{N}$ such that, for any $\bar{\theta} \in \mathsf{K}$, $\lim_{n\to\infty} \rho_\Theta(\theta_{\psi(\phi(n))}, \bar{\theta}) = 0$. Therefore,

$$\begin{aligned} &\liminf_{n\to\infty} \inf_{\theta\in\mathsf{K}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \\ &= \lim_{n\to\infty} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_{\phi(n)}(\omega), \hat{\mu}_{\theta_{\phi(n)},m(\phi(n))})\middle|Y_{1:n}\right] \\ &= \lim_{n\to\infty} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_{\psi(\phi(n))}(\omega), \hat{\mu}_{\theta_{\psi(\phi(n))},m(\psi(\phi(n)))})\middle|Y_{1:n}\right] \\ &= \liminf_{n\to\infty} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_{\psi(\phi(n))}(\omega), \hat{\mu}_{\theta_{\psi(\phi(n))},m(\psi(\phi(n)))})\middle|Y_{1:n}\right] \\ &\geq \liminf_{n\to\infty} \left\{\mathbf{SW}_p(\hat{\mu}_{\psi(\phi(n))}(\omega), \mu_{\theta_{\psi(\phi(n))}}) - \mathbb{E}\left[\mathbf{SW}_p(\mu_{\theta_{\psi(\phi(n))}}, \hat{\mu}_{\theta_{\psi(\phi(n))},m(\psi(\phi(n)))})\middle|Y_{1:n}\right]\right\} \end{aligned} \tag{3.27}$$

$$\begin{aligned} &\geq \liminf_{n\to\infty} \mathbf{SW}_p(\hat{\mu}_{\psi(\phi(n))}(\omega), \mu_{\theta_{\psi(\phi(n))}}) \\ &\quad - \limsup_{n\to\infty} \mathbb{E}\left[\mathbf{SW}_p(\mu_{\theta_{\psi(\phi(n))}}, \hat{\mu}_{\theta_{\psi(\phi(n))},m(\psi(\phi(n)))})\middle|Y_{1:n}\right] \\ &\geq \mathbf{SW}_p(\mu_\star, \mu_{\bar{\theta}}) \\ &\geq \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \end{aligned} \tag{3.28}$$

where (3.27) follows from the triangle inequality, and (3.28) is obtained on one hand by lower semi-continuity since $\hat{\mu}_{\psi(\phi(n))}(\omega) \xrightarrow{w} \mu_\star$ by **A**2 and Theorem 3.1 and $\mu_{\theta_{\psi(\phi(n))}} \xrightarrow{w} \mu_{\bar{\theta}}$ by **A**1, and on the other hand by **A**4 which gives

$$\limsup_{n\to\infty} \mathbb{E}[\mathbf{SW}_p(\mu_{\theta_{\psi(\phi(n))}}, \hat{\mu}_{\theta_{\psi(\phi(n))},m(\psi(\phi(n)))})|Y_{1:n}] = 0 \,.$$

We conclude that the first condition in (3.26) holds.

Now, we fix $\mathsf{O} \subset \Theta$ open. By definition of the infimum, there exists a sequence $(\theta_n)_{n\in\mathbb{N}}$ in O such that $\{\mathbf{SW}_p(\mu_\star, \mu_{\theta_n})\}_{n\in\mathbb{N}}$ converges to $\inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\mu_\star, \mu_\theta)$ as $n \to +\infty$, and $\inf_{\theta\in\mathsf{O}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \leq \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)})\middle|Y_{1:n}\right]$ for any $n \in \mathbb{N}$. Then,

$$\begin{aligned}
&\limsup_{n\to\infty} \inf_{\theta\in\mathsf{O}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \\
&\leq \limsup_{n\to\infty} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)})\middle|Y_{1:n}\right] \\
&\leq \limsup_{n\to\infty} \left\{\mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathbf{SW}_p(\mu_\star, \mu_{\theta_n}) + \mathbb{E}\left[\mathbf{SW}_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n,m(n)})\middle|Y_{1:n}\right]\right\} \qquad (3.29) \\
&= \limsup_{n\to\infty} \mathbf{SW}_p(\mu_\star, \mu_{\theta_n}) \quad \text{(by } \mathbf{A}2 \text{ and } \mathbf{A}4\text{)} \\
&= \inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\mu_\star, \mu_\theta) \quad \text{(by definition of } (\theta_n)_{n\in\mathbb{N}}\text{)}
\end{aligned}$$

where (3.29) follows from the triangle inequality. This shows that the second condition in (3.26) holds, hence $\theta \mapsto \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right]$ epi-converges to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$.

We now apply [Rockafellar et al., 2009, Theorem 7.31]. First, by [Rockafellar et al., 2009, Theorem 7.31(b)], (3.8) immediately follows from the epi-convergence of $\theta \mapsto \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right]$ to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$. Next, we show that [Rockafellar et al., 2009, Theorem 7.31(a)] holds by finding, for any $\eta > 0$, a compact set $\mathsf{B} \subset \Theta$ and $N \in \mathbb{N}$ such that, for all $n \geq N$,

$$\inf_{\theta\in\mathsf{B}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \leq \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] + \eta \,.$$

In fact, we simply show that there exists a compact set $\mathsf{B} \subset \Theta$ and $N \in \mathbb{N}$ such that, for all $n \geq N$,

$$\inf_{\theta\in\mathsf{B}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] = \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \,.$$

On the one hand, the second condition in (3.26) gives us

$$\limsup_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \leq \inf_{\theta\in\Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta) = \epsilon_\star \,.$$

We deduce that there exists $n_{\epsilon/6}(\omega)$ such that, for $n \geq n_{\epsilon/6}(\omega)$,

$$\inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \leq \epsilon_\star + \frac{\epsilon}{6},$$

with ϵ from **A**3. When $n \geq n_{\epsilon/6}(\omega)$, $\widehat{\Theta}_{\epsilon/3} = \{\theta \in \Theta : \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})|Y_{1:n}] \leq \epsilon_\star + \frac{\epsilon}{3}\}$ is a non-empty set as it contains θ^* defined as $\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta^*,m(n)})\middle|Y_{1:n}\right] = \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right]$.

On the other hand, by **A**2, there exists $n_{\epsilon/3}(\omega)$ such that, for $n \geq n_{\epsilon/3}(\omega)$,

$$\mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) \leq \frac{\epsilon}{3} . \tag{3.30}$$

Finally, by **A**4, there exists $n'_{\epsilon/3}(\omega)$ such that, for $n \geq n'_{\epsilon/3}(\omega)$,

$$\mathbb{E}\left[\mathbf{SW}_p(\mu_\theta, \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \leq \frac{\epsilon}{3} . \tag{3.31}$$

Let $n \geq n_*(\omega) = \max\{n_{\epsilon/6}(\omega), n_{\epsilon/3}(\omega), n'_{\epsilon/3}(\omega)\}$ and $\theta \in \widehat{\Theta}_{\epsilon/3}$. By the triangle inequality,

$$\begin{aligned}
&\mathbf{SW}_p(\mu_\star, \mu_\theta) \\
&\leq \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] + \mathbb{E}\left[\mathbf{SW}_p(\mu_\theta, \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] \\
&\leq \epsilon_\star + \epsilon \quad \text{(since } \theta \in \widehat{\Theta}_{\epsilon/3} \text{ and by (3.30) and (3.31))}
\end{aligned}$$

This means that, when $n \geq n_*(\omega)$, $\widehat{\Theta}_{\epsilon/3} \subset \Theta^\star_\epsilon$ with $\Theta^\star_\epsilon$ as defined in **A**3, and since $\inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right]$ is attained in $\widehat{\Theta}_{\epsilon/3}$, we obtain

$$\inf_{\theta\in\Theta^\star_\epsilon} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] = \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right] . \tag{3.32}$$

By [Rockafellar et al., 2009, Theorem 7.31(a)], (3.7) is a direct consequence of (3.32) and the epi-convergence of $\theta \mapsto \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right]$ to $\theta \mapsto \mathbf{SW}_p(\mu_\star, \mu_\theta)$.

Finally, by the same arguments used in this proof for $\operatorname{argmin}_{\theta\in\Theta}\mathbf{SW}_p(\mu_\star, \mu_\theta)$, the set $\operatorname{argmin}_{\theta\in\Theta}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})\middle|Y_{1:n}\right]$ is non-empty for $n \geq n_*(\omega)$.

□

3.5.5 Proof of Theorem 3.4

Here again, the result follows from applying [Rockafellar et al., 2009, Theorem 7.31] paraphrased in Theorem 3.11.

Proof of Theorem 3.4. First, by **A**1 and Corollary 3.13, the map $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$ is l.s.c. on Θ. Therefore, there exists $\theta_n \in \Theta$ such that $\mathbf{SW}_p(\hat{\mu}_n, \mu_{\theta_n}) = \epsilon_n$. The set $\Theta_{\epsilon,n}$ with the ϵ from **A**5 is non-empty as it contains θ_n, closed by lower semi-continuity of $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$, and bounded. $\Theta_{\epsilon,n}$ is thus compact, and we conclude again by lower semi-continuity that the set $\operatorname{argmin}_{\theta\in\Theta}\mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$ is non-empty [Aliprantis et al., 1999, Theorem 2.43].

Then, we prove that $\theta \mapsto \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right]$ epi-converges to $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$ as $m \to \infty$ using the characterization in [Rockafellar et al., 2009, Proposition 7.29], *i.e.* we verify that for every compact set $\mathsf{K} \subset \Theta$ and every open set $\mathsf{O} \subset \Theta$,

$$\begin{aligned}
\liminf_{m\to\infty} \inf_{\theta\in\mathsf{K}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] &\geq \inf_{\theta\in\mathsf{K}} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) \\
\limsup_{m\to\infty} \inf_{\theta\in\mathsf{O}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] &\leq \inf_{\theta\in\mathsf{O}} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) .
\end{aligned} \tag{3.33}$$

Let $\mathsf{K} \subset \Theta$ be a compact set. By **A**1 and Corollary 3.15, for any $m \in \mathbb{N}$, the map $\theta \mapsto \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}]$ is l.s.c., so there exists $\theta_m \in \mathsf{K}$ such that

$$\inf_{\theta\in\mathsf{K}} \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}] = \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta_m,m})|Y_{1:n}] .$$

We consider the subsequence $\{\hat{\mu}_{\theta_{\phi(m)},\phi(m)}\}_{m\in\mathbb{N}}$ where $\phi:\mathbb{N}\to\mathbb{N}$ is increasing such that $\mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta_{\phi(m)},\phi(m)})|Y_{1:n}]$ converges to

$$\liminf_{m\to\infty}\ \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta_m,m})|Y_{1:n}] = \liminf_{m\to\infty}\inf_{\theta\in\mathsf{K}}\mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}]\ .$$

Since K is compact, there also exists an increasing function $\psi:\mathbb{N}\to\mathbb{N}$ such that for $\bar{\theta}\in\mathsf{K}$, $\lim_{m\to\infty}\rho_\Theta(\theta_{\psi(\phi(m))},\bar{\theta})=0$. Therefore,

$$\begin{aligned}
&\liminf_{m\to\infty}\inf_{\theta\in\mathsf{K}}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}\right]\\
&=\lim_{m\to\infty}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta_{\phi(m)},\phi(m)})\middle|Y_{1:n}\right]\\
&=\lim_{m\to\infty}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta_{\psi(\phi(m))},\psi(\phi(m))})\middle|Y_{1:n}\right]\\
&=\liminf_{m\to\infty}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta_{\psi(\phi(m))},\psi(\phi(m))})\middle|Y_{1:n}\right]\\
&\geq\liminf_{m\to\infty}[\mathbf{SW}_p(\hat{\mu}_n,\mu_{\theta_{\psi(\phi(m))}})-\mathbb{E}\left[\mathbf{SW}_p(\mu_{\theta_{\psi(\phi(m))}},\hat{\mu}_{\theta_{\psi(\phi(m))},\psi(\phi(m))})\middle|Y_{1:n}\right]] \qquad (3.34)\\
&\geq\liminf_{m\to\infty}\mathbf{SW}_p(\hat{\mu}_n,\mu_{\theta_{\psi(\phi(m))}})-\limsup_{m\to\infty}\mathbb{E}\left[\mathbf{SW}_p(\mu_{\theta_{\psi(\phi(m))}},\hat{\mu}_{\theta_{\psi(\phi(m))},\psi(\phi(m))})\middle|Y_{1:n}\right]\\
&\geq\mathbf{SW}_p(\hat{\mu}_n,\mu_{\bar{\theta}}) \qquad (3.35)\\
&\geq\inf_{\theta\in\mathsf{K}}\mathbf{SW}_p(\hat{\mu}_n,\mu_\theta)
\end{aligned}$$

where (3.34) results from the triangle inequality and (3.35) is obtained by **A**4 on one hand and by lower semi-continuity on the other hand since $\mu_{\theta_{\psi(\phi(n))}}\xrightarrow{w}\mu_{\bar{\theta}}$ by **A**1. We conclude that the first condition in (3.33) holds.

Now, we fix $\mathsf{O}\subset\Theta$ open. There exists a sequence $(\theta_m)_{m\in\mathbb{N}}$ in O such that the sequence $\{\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta_m,m})\}_{m\in\mathbb{N}}$ converges to $\inf_{\theta\in\mathsf{O}}\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})$, and for any $m\in\mathbb{N}$, $\inf_{\theta\in\mathsf{O}}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}\right]\leq\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\mu_{\theta_m,m})|Y_{1:n}\right]$, by definition of the infimum. Then,

$$\begin{aligned}
&\limsup_{m\to\infty}\inf_{\theta\in\mathsf{O}}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}\right]\\
&\leq\limsup_{m\to\infty}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta_m,m})|Y_{1:n}\right]\\
&\leq\limsup_{m\to\infty}[\mathbf{SW}_p(\hat{\mu}_n,\mu_{\theta_m})+\mathbb{E}\left[\mathbf{SW}_p(\mu_{\theta_m},\hat{\mu}_{\theta_m,m})|Y_{1:n}\right]] \qquad (3.36)\\
&\leq\limsup_{m\to\infty}\mathbf{SW}_p(\hat{\mu}_n,\mu_{\theta_m}) \quad (\text{by } \mathbf{A}4)\\
&=\inf_{\theta\in\mathsf{O}}\mathbf{SW}_p(\hat{\mu}_n,\mu_\theta) \quad (\text{by definition of } (\theta_m)_{m\in\mathbb{N}})
\end{aligned}$$

where (3.36) results from applying the triangle inequality. This shows that the second condition in (3.33) holds, hence the sequence of functions $\theta\mapsto\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}\right]$ epi-converges to $\theta\mapsto\mathbf{SW}_p(\hat{\mu}_n,\mu_\theta)$.

Now, we apply [Rockafellar et al., 2009, Theorem 7.31]. By [Rockafellar et al., 2009, Theorem 7.31(b)], (3.10) immediately follows from the epi-convergence of $\theta\mapsto\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}\right]$ to $\theta\mapsto\mathbf{SW}_p(\hat{\mu}_n,\mu_\theta)$. Next, we show that [Rockafellar et al., 2009, Theorem 7.31(a)] holds by finding for any $\eta>0$ a compact set $\mathsf{B}\subset\Theta$ and $N\in\mathbb{N}$ such that, for all $n\geq N$,

$$\inf_{\theta\in\mathsf{B}}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}\right]\leq\inf_{\theta\in\Theta}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n,\hat{\mu}_{\theta,m})|Y_{1:n}\right]+\eta\ .$$

In fact, we simply show that there exists a compact set $\mathsf{B} \subset \Theta$ and $N \in \mathbb{N}$ such that $\inf_{\theta\in\mathsf{B}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] = \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right]$ for all $n \geq N$. On one hand, the second condition in (3.33) gives us

$$\limsup_{m\to\infty} \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] \leq \inf_{\theta\in\Theta} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) = \epsilon_n \ .$$

We deduce that there exists $m_{\epsilon/4}$ such that, for $m \geq m_{\epsilon/4}$,

$$\inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] \leq \epsilon_n + \frac{\epsilon}{4} \ , \tag{3.37}$$

with ϵ from **A**5. When $m \geq m_{\epsilon/4}$, $\Theta_{\epsilon/2} = \{\theta \in \Theta : \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}] \leq \epsilon_n + \frac{\epsilon}{2}\}$ is a non-empty set, as it contains θ^* defined as

$$\mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta^*,m})|Y_{1:n}] = \inf_{\theta\in\Theta} \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}] \ .$$

On the other hand, by **A**4, there exists $m_{\epsilon/2}$ such that, for $m \geq m_{\epsilon/2}$,

$$\mathbb{E}\left[\mathbf{SW}_p(\mu_\theta, \hat{\mu}_{\theta,m})|Y_{1:n}\right] \leq \frac{\epsilon}{2} \ . \tag{3.38}$$

Let $\theta \in \Theta_{\epsilon/2}$ and $m \geq m_* = \max\{m_{\epsilon/4}, m_{\epsilon/2}\}$. By the triangle inequality,

$$\begin{aligned}\mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) &\leq \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] + \mathbb{E}\left[\mathbf{SW}_p(\mu_\theta, \hat{\mu}_{\theta,m})|Y_{1:n}\right] \\ &\leq \epsilon_n + \epsilon \quad (\text{since } \theta \in \Theta_{\epsilon/2} \text{ and by (3.38)})\end{aligned}$$

Therefore, when $m \geq m_*$, $\Theta_{\epsilon/2} \subset \Theta_{\epsilon,n}$, and since $\inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right]$ is attained in $\Theta_{\epsilon/2}$,

$$\inf_{\theta\in\Theta_{\epsilon,n}} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] = \inf_{\theta\in\Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right] \ . \tag{3.39}$$

By [Rockafellar et al., 2009, Theorem 7.31(a)], (3.9) is a direct consequence of (3.39) and the epiconvergence of $\theta \mapsto \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m})|Y_{1:n}\right]$ to $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$.

Finally, by the same arguments used in this proof for $\operatorname{argmin}_{\theta\in\Theta}\mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$, the set $\operatorname{argmin}_{\theta\in\Theta}\mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})|Y_{1:n}\right]$ is non-empty for $m \geq m_*$.

□

3.5.6 Proof of Theorem 3.5

Let us first formally establish measurability for MESWE.

Theorem 3.17 (Measurability of the MESWE)**.** *Assume* ***A****1. For any $n \geq 1$, $m \geq 1$ and $\epsilon > 0$, there exists a Borel measurable function $\hat{\theta}_{n,m,\epsilon} : \Omega \to \Theta$ that satisfies for any $\omega \in \Omega$,*

$$\hat{\theta}_{n,m,\epsilon}(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\Theta} \ \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m})|Y_{1:n}\right], \ \textit{if this set is non-empty,} \\ \left\{\theta \in \Theta \ : \ \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m})|Y_{1:n}\right] \leq \epsilon_* + \epsilon\right\}, \ \textit{otherwise,} \end{cases}$$

where $\epsilon_* = \inf_{\theta\in\Theta} \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m})|Y_{1:n}]$.

We prove the measurability of MSWE and MESWE by verifying the conditions of [Brown and Purves, 1973, Corollary 1], which are recalled below.

Theorem 3.18 (Corollary 1 in Brown and Purves [1973]). *Let* U, V *be Polish spaces and* f *be a real-valued Borel measurable function defined on a Borel subset* D *of* $\mathsf{U} \times \mathsf{V}$. *We denote by* $\mathrm{proj}(\mathsf{D})$ *the set defined as*

$$\mathrm{proj}(\mathsf{D}) = \{u \ : \ \text{there exists } v \in \mathsf{V}, \ (u,v) \in \mathsf{D}\} \ .$$

Suppose that for each $u \in \mathrm{proj}(\mathsf{D})$, *the section* $\mathsf{D}_u = \{v \in V, (u,v) \in \mathsf{D}\}$ *is* σ*-compact and* $f(u,\cdot)$ *is lower semi-continuous with respect to the relative topology on* D_u. *Then,*

1. *The sets* $\mathrm{proj}(\mathsf{D})$ *and* $\mathsf{I} = \{u \in \mathrm{proj}(\mathsf{D}), \text{ for some } v \in \mathsf{D}_u, \ f(u,v) = \inf f_u\}$ *are Borel*

2. *For each* $\epsilon > 0$, *there is a Borel measurable function* ϕ_ϵ *satisfying, for* $u \in \mathrm{proj}(\mathsf{D})$,

$$\begin{aligned} f(u, \phi_\epsilon(u)) &= \inf_{\mathsf{D}_u} f_u, && \text{if } \ u \in \mathsf{I}, \\ &\leq \epsilon + \inf_{\mathsf{D}_u} f_u, && \text{if } \ u \notin \mathsf{I}, \quad \text{and} \quad \inf_{\mathsf{D}_u} f_u \neq -\infty \\ &\leq -\epsilon^{-1}, && \text{if } \ u \notin \mathsf{I}, \quad \text{and} \quad \inf_{\mathsf{D}_u} f_u = -\infty \ . \end{aligned}$$

Proof of Theorems 3.5 and 3.17. We start by proving Theorem 3.5. The empirical measure $\hat{\mu}_n(\omega)$ depends on $\omega \in \Omega$ only through $y = (y_1, \ldots, y_n) \in \mathsf{Y}^n$, so we can consider it as a function on Y^n rather than on Ω. We introduce $\mathsf{D} = \mathsf{Y}^n \times \Theta$. Since Y is Polish, Y^n $(n \in \mathbb{N}^*)$ endowed with the product topology is Polish. For any $y \in \mathsf{Y}^n$, the set $\mathsf{D}_y = \{\theta \in \Theta, \ (y,\theta) \in \mathsf{D}\} = \Theta$ is assumed to be σ-compact.

The map $y \mapsto \hat{\mu}_n(y)$ is continuous for the weak topology (see the proof of Lemma 3.14), as well as the map $\theta \mapsto \mu_\theta$ according to **A**1. We deduce by Corollary 3.13 that the map $(\mu, \theta) \mapsto \mathbf{SW}_p(\mu, \mu_\theta)$ is l.s.c. for the weak topology. Since the composition of a lower semi-continuous function with a continuous function is l.s.c., the map $(y, \theta) \mapsto \mathbf{SW}_p(\hat{\mu}_n(y), \mu_\theta)$ is l.s.c. for the weak topology, thus measurable and for any $y \in \mathsf{Y}^n$, $\theta \mapsto \mathbf{SW}_p(\hat{\mu}_n(y), \mu_\theta)$ is l.s.c. on Θ. A direct application of Theorem 3.18 finalizes the proof.

Theorem 3.17 can be proved via the same methodology: we verify that we can apply Theorem 3.18 using Corollary 3.15 instead of Corollary 3.13.

□

3.5.7 Proof of Theorems 3.6 and 3.7

The proof of Theorem 3.6 and Theorem 3.7 consists in showing that the conditions of [Pollard, 1980, Theorems 4.2] and [Pollard, 1980, Theorem 7.2] respectively are satisfied: conditions (i), (ii) and (iii) follow from **A**6, **A**7 and **A**8.

3.5.8 Additional details on Section 3.3

Sampling schemes. We first explain the methods that we used to generate i.i.d. samples from the uniform distribution on $\mathbb{S}^{d-1}$ (required for the Monte Carlo estimate of SW (2.21)) and multivariate elliptically contoured stable distributions (for Section 3.3.2).

- **Uniform sampling on the sphere.** To sample from $\mathbb{S}^{d-1}$, we form the d-dimensional vector $\mathbf{s}$ by drawing each of its d components from the standard normal distribution $\mathcal{N}(0,1)$ and we normalize it, *i.e.*

$$\mathbf{s}' = \frac{\mathbf{s}}{\|\mathbf{s}\|_2} \ ,$$

so that $\mathbf{s}'$ lies on the sphere.

- **Sampling from multivariate elliptically contoured stable distributions.** We recall that if $Y \in \mathbb{R}^d$ is α-stable and elliptically contoured, *i.e.* $Y \sim \mathcal{E}\alpha\mathcal{S}_c(\mathbf{\Sigma}, \mathbf{m})$, then its joint characteristic function is defined as, for any $\mathbf{t} \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(i\mathbf{t}^T Y)] = \exp\left(-(\mathbf{t}^T \mathbf{\Sigma} \mathbf{t})^{\alpha/2} + i\mathbf{t}^T \mathbf{m}\right) , \tag{3.40}$$

where $\mathbf{\Sigma}$ is a positive definite matrix (akin to a correlation matrix), $\mathbf{m} \in \mathbb{R}^d$ is a location vector (equal to the mean if it exists) and $\alpha \in (0, 2)$ controls the thickness of the tail. Elliptically contoured stable distributions are scale mixtures of multivariate Gaussian distributions [Samorodnitsky and Taqqu, 1994, Proposition 2.5.2], whose densities are intractable, but can easily be simulated [Nolan, 2013]: let $A \sim \mathcal{S}_{\alpha/2}(\beta, \gamma, \delta)$ be a one-dimensional positive $(\alpha/2)$-stable random variable with $\beta = 1$, $\gamma = 2\cos(\frac{\pi\alpha}{4})^{2/\alpha}$ and $\delta = 0$, and $G \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Then,

$$Y = \sqrt{A}G + \mathbf{m}$$

has (3.40) as characteristic function.

Optimization methods. By definition, computing the M(E)SWE and implies minimizing the (expected) Sliced-Wasserstein distance over the set of parameters, which is in general computationnally intractable. We then resort to numerical methods in our experiments to approximate these two estimators, as we detail below.

- **Multivariate Gaussian distributions.** We derive the explicit gradient expressions of the approximate $\mathbf{SW}_2^2$ distance with respect to the mean and scale parameters $\mathbf{m}$ and σ^2, and we use the ADAM stochastic optimization method with the default parameter settings suggested in [Kingma and Ba, 2015]. For the MSWE, we use (2.14) to approximate the one-dimensional Wasserstein distance, and we evaluate directly the Gaussian density of the generated samples, utilizing the fact that the projection of a Gaussian of parameters $(\mathbf{m}, \sigma^2\mathbf{I})$ along $u \in \mathbb{S}^{d-1}$ is a 1D normal distribution of parameters $(\langle u, \mathbf{m}\rangle, \sigma^2\langle u, u\rangle)$. In this case, the gradient of the approximate $\mathbf{SW}_2^2$ between $\mu = \mathcal{N}(\mathbf{m}, \sigma^2\mathbf{I})$ and the empirical distribution associated to n samples drawn by $\mathcal{N}(\mathbf{m}_\star, \sigma_\star^2\mathbf{I})$, denoted by $\hat{\nu}$, is given by,

$$\begin{aligned}\nabla_{\mathbf{m}}\mathbf{SW}_2^2(\mu, \hat{\nu}) =& (1/\operatorname{card}(\mathsf{U})\operatorname{card}(\mathsf{S})) \sum_{u\in\mathsf{U}, s\in\mathsf{S}} \Bigg(\left| s - \tilde{F}^{-1}_{u^\star_\sharp\hat{\nu}}(\tilde{F}_{u^\star_\sharp\mu}(s)) \right|^2 \\ & \mathcal{N}(s; \langle u, \mathbf{m}\rangle, \sigma^2\|u\|^2) \frac{s - \langle u, \mathbf{m}\rangle}{\sigma^2\|u\|^2} u \Bigg) ,\end{aligned}$$

$$\begin{aligned}\nabla_{\sigma^2}\mathbf{SW}_2^2(\mu, \hat{\nu}) =& (1/\operatorname{card}(\mathsf{U})\operatorname{card}(\mathsf{S})) \sum_{u\in\mathsf{U}, s\in\mathsf{S}} \Bigg(\left| s - \tilde{F}^{-1}_{u^\star_\sharp\hat{\nu}}(\tilde{F}_{u^\star_\sharp\mu}(s)) \right|^2 \\ & \mathcal{N}(s; \langle u, \mathbf{m}\rangle, \sigma^2\|u\|^2) \frac{1}{2\sigma^2} \left(\frac{(s - \langle u, \mathbf{m}\rangle)^2}{\sigma^2\|u\|^2} - 1 \right) \Bigg) ,\end{aligned}$$

where $\mathsf{U} \subset \mathbb{S}^{d-1}$ is a finite set of random projections picked uniformly on $\mathbb{S}^{d-1}$, S is a finite subset in $\mathbb{R}$, and for any $s \in \mathsf{S}$, $\mathcal{N}(s; \langle u, \mathbf{m}\rangle, \sigma^2\|u\|^2)$ denotes the density function of the Gaussian of parameters $(\langle u, \mathbf{m}\rangle, \sigma^2\|u\|^2)$ evaluated at s.

For the MESWE, we use (2.13) and evaluate the empirical distribution of generated samples instead of their normal density. Therefore, the gradient of the approximate $\mathbf{SW}_2^2$ between the empirical distributions corresponding to one generated dataset of m samples drawn from $\mathcal{N}(\mu, \sigma^2\mathbf{I})$ and n samples drawn from $\mathcal{N}(\mu_\star, \sigma_\star^2\mathbf{I})$, respectively denoted by $\hat{\mu}$ and $\hat{\nu}$, is obtained with,

$$\nabla_{\mathbf{m}}\mathbf{SW}_2^2(\hat{\mu},\hat{\nu}) = \frac{-2}{\mathrm{card}(\mathsf{U}).K}\sum_{u\in\mathsf{U}}\sum_{k=1}^{K}\left|\tilde{F}^{-1}_{u^\star_\sharp\hat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\sharp\hat{\nu}}(t_k)\right| u \ ,$$

$$\nabla_{\sigma^2}\mathbf{SW}_2^2(\hat{\mu},\hat{\nu}) = \frac{1}{\mathrm{card}(\mathsf{U}).K}\sum_{u\in\mathsf{U}}\sum_{k=1}^{K}\left|\tilde{F}^{-1}_{u^\star_\sharp\hat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\sharp\hat{\nu}}(t_k)\right| \frac{\langle u, \mathbf{m}\rangle - \tilde{F}^{-1}_{u^\star_\sharp\hat{\mu}}(t_k)}{\sigma^2} \ .$$

- **Multivariate elliptically contoured stable distributions.** When comparing MESWE to MEWE, we approximate these estimators using the derivative-free optimization method Nelder-Mead (Nelder and Mead [1965], implemented in `Scipy`), following the approach in [Bernton et al., 2019].

  When illustrating the theoretical properties of MESWE, we proceed in the same way as for the multivariate Gaussian experiment: we compute the explicit gradient expression of the approximate $\mathbf{SW}_2^2$ distance with respect to the location parameter $\mathbf{m}$, and we use the ADAM stochastic optimization method with the default settings. Equation (3.41) gives the formula of the gradient of the approximate $\mathbf{SW}_2^2$ between the empirical distributions of one generated dataset of m samples drawn from $\mathcal{E}\alpha\mathcal{S}_c(\mathbf{I}, \mathbf{m})$ and n samples drawn from $\mathcal{E}\alpha\mathcal{S}_c(\mathbf{I}, \mathbf{m}_\star)$, respectively denoted by $\hat{\mu}$ and $\hat{\nu}$, with respect to $\mathbf{m}$.

$$\nabla_{\mathbf{m}}\mathbf{SW}_2^2(\hat{\mu},\hat{\nu}) = \frac{-2}{\mathrm{card}(\mathsf{U}).K}\sum_{u\in\mathsf{U}}\sum_{k=1}^{K}\left|\tilde{F}^{-1}_{u^\star_\sharp\hat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\sharp\hat{\nu}}(t_k)\right| u \ . \tag{3.41}$$

- **High-dimensional real data using GANs.** We use the ADAM optimizer provided by TensorFlow GPU.

Computing infrastructure. The experiment comparing the computational time of MESWE and MEWE was conducted on a daily-use laptop (CPU intel core i7, 1.90GHz $\times$ 8 and 16GB of RAM). The neural network experiment was run on a cluster with 4 relatively modern GPUs.

Chapter 4

Approximate Bayesian Computation with the Sliced-Wasserstein Distance

This chapter is based on [Nadjahi et al., 2020a].

Approximate Bayesian Computation (ABC) is a popular method for approximate inference in generative models with intractable but easy-to-sample likelihood. It constructs an approximate posterior distribution by finding parameters for which the simulated data are close to the observations in terms of summary statistics. These statistics are defined beforehand and might induce a loss of information, which has been shown to deteriorate the quality of the approximation. To overcome this problem, a Wasserstein-based ABC technique has recently been proposed, and compares the datasets via the Wasserstein distance between their empirical distributions, but does not scale well to the dimension or the number of samples.

In this chapter, we propose a new ABC technique, called *Sliced-Wasserstein ABC* and relying on the Sliced-Wasserstein distance, which has better computational and statistical properties. We derive two theoretical results showing the asymptotical consistency of our approach, and we illustrate its advantages on synthetic data and an image denoising task.

4.1 Introduction

Consider the problem of estimating the posterior distribution of some model parameters $\theta \in \mathbb{R}^{d_\theta}$ given n data points $y_{1:n} \in \mathsf{Y}^n$. This distribution has a closed-form expression given by the Bayes' theorem up to a multiplicative constant,

$$\pi(\theta|y_{1:n}) \propto \pi(y_{1:n}|\theta)\pi(\theta) \ .$$

For many statistical models of interest, the likelihood $\pi(y_{1:n}|\theta)$ cannot be numerically evaluated in a reasonable amount of time, which prevents the application of classical likelihood-based approximate inference methods. Nevertheless, in various settings, even if the associated likelihood is numerically intractable, one can still generate synthetic data given any model parameter value. This generative setting gave rise to an alternative framework of likelihood-free inference techniques. Among them, *Approximate Bayesian Computation* methods [Tavaré et al., 1997, Beaumont et al., 2002] have become a popular choice and have proven useful in various practical applications, e.g.

[Peters and Sisson, 2006, Tanaka et al., 2006, Wood, 2010]. The core idea of ABC is to bypass calculation of the likelihood by using simulations: the exact posterior is approximated by retaining the parameter values for which the synthetic data are close enough to the observations. Closeness is usually measured with a discrepancy measure between the two datasets reduced to some 'summary statistics' (e.g., empirical mean or empirical covariance). While summaries allow a practical and efficient implementation of ABC, especially in high-dimensional data spaces, the quality of the approximate posterior distribution highly depends on them and constructing sufficient statistics is a non-trivial task. Summary statistics can be designed by hand using expert knowledge, which can be tedious especially in real-world applications, or in an automated way, for instance see [Fearnhead and Prangle, 2012].

Recently, discrepancy measures that view data sets as empirical probability distributions to eschew the construction of summary statistics have been proposed for ABC. Examples include the Kullback-Leibler divergence [Jiang et al., 2018], maximum mean discrepancy [Park et al., 2016], and Wasserstein distance [Bernton et al., 2019]. As we discussed in Chapters 1 and 2, this latter distance emerging from optimal transport theory has attracted abundant attention in statistics and machine learning, due to its strong theoretical properties and applications on many domains. In particular, we recall that it has the ability of making meaningful comparisons even between probability measures with non-overlapping supports, unlike KL. However, the computational complexity of the Wasserstein distance rapidly becomes a challenge when the dimension of the observations is large, and several numerical methods have been proposed during the past few years to speed-up this computation. In particular, Wasserstein-ABC (WABC, Bernton et al. [2019]) introduces a different computational approach to those presented in Chapter 2: the Wasserstein distance is estimated with a novel approximation based on the Hilbert space-filling curve and termed the *Hilbert distance*, which is computationally efficient but accurate for small dimensions only. Besides, under a general setting, the Wasserstein distance suffers from a curse of dimensionality in the sense that the error made when approximating it from samples grows exponentially fast with the data space dimension (Section 2.4.2). These computational and statistical issues can strongly affect the performance of WABC applied to high-dimensional data.

Building on the computational efficiency of SW and its successful performance in generative settings, as demonstrated by prior studies (reviewed in Section 1.3) and our previous chapter, we develop a novel ABC framework that uses SW as the data discrepancy measure. This defines a likelihood-free method which does not require choosing summary statistics and is efficient even with high-dimensional observations, thus overcoming the limitations of WABC. We derive asymptotical guarantees on the convergence of the resulting ABC posterior, and we illustrate the superior empirical performance of our methodology by applying it on a synthetical problem and an image denoising task.

4.2 Background on Approximate Bayesian Computation

In this chapter, we consider the same purely generative modeling framework as in Chapter 3, so we keep the formalism and notations presented in Section 3.1. Additionally, we assume that conditions **(C1)**, **(C2)** and **(C4)** hold. However, instead of using minimum distance estimation to perform parameter inference in such models, we focus on another class of approximate inference methods, called Approximation Bayesian Computation

Algorithm 2: Vanilla ABC.

Input: observations $y_{1:n}$, number of iterations T, data discrepancy measure $\mathbf{D}$, summary statistics s, tolerance threshold $\varepsilon > 0$.

for $t = 1, \dots, T$ **do**

  repeat

    $\theta \sim \pi(\cdot)$ and $z_{1:m} \sim \mu_\theta$ i.i.d.

  until $\mathbf{D}\big(s(y_{1:n}), s(z_{1:m})\big) \leq \varepsilon$;

  $\theta^{(t)} = \theta$

return $\theta^{(1)}, \dots, \theta^{(T)}$

algorithms.

ABC methods are used to approximate the posterior distribution in generative models when the likelihood is numerically intractable but easy to sample from. The basic and simplest ABC algorithm is an acceptance-rejection method [Tavaré et al., 1997], which iteratively draws a candidate parameter θ' from a prior distribution π, and 'synthetic data' $z_{1:m} = (z_i)_{i=1}^m$ from $\mu_{\theta'}$, and keeps θ' if $z_{1:m}$ is close enough to the observations $y_{1:n} = (y_i)_{i=1}^n$. Specifically, the acceptance rule is

$$\mathbf{D}\big(s(y_{1:n}), s(z_{1:m})\big) \leq \varepsilon \,, \tag{4.1}$$

where $\mathbf{D}$ is a data discrepancy measure taking non-negative values, ε is a tolerance threshold, and $s : \sqcup_{n \in \mathbb{N}^*} \mathsf{Y}^n \to \mathbb{R}^{d_s}$ with small d_s is a summary statistics. The algorithm is summarized in Algorithm 2 and returns samples of θ that are distributed from:

$$\pi^{\varepsilon}_{y_{1:n}}(\theta) = \frac{\pi(\theta) \int_{\mathsf{Y}^m} \mathbb{1}\{\mathbf{D}\big(s(y_{1:n}), s(z_{1:m})\big) \leq \varepsilon\} \mathrm{d}\mu_\theta(z_{1:m})}{\int_\Theta \mathrm{d}\pi(\theta) \int_{\mathsf{Y}^m} \mathbb{1}\{\mathbf{D}\big(s(y_{1:n}), s(z_{1:m})\big) \leq \varepsilon\} \mathrm{d}\mu_\theta(z_{1:m})} \tag{4.2}$$

The choice of $s(\cdot)$ directly impacts the quality of the resulting approximate posterior: if the statistics are *sufficient statistics*, $\pi^{\varepsilon}_{y_{1:n}}(\theta)$ converges to the true posterior $\pi(\theta|y_{1:n})$ as $\varepsilon \to 0$, otherwise, the limiting distribution is at best $\pi(\theta|s(y_{1:n}))$ [Sisson et al., 2018, Frazier et al., 2018]. Wasserstein-ABC has then been proposed to avoid this loss of information.

Wasserstein distance and ABC. Wasserstein-ABC [Bernton et al., 2019] is a variant of ABC (2) that uses $\mathbf{W}_p$, $p \in [1, +\infty)$ between the empirical distributions of the observed and synthetic data, in place of the discrepancy measure $\mathbf{D}$ between summaries. To make this method scalable to any dataset size, Bernton et al. [2019] introduces a new approximation of the Wasserstein distance, called the Hilbert distance, which extends the idea behind the computation of $\mathbf{W}_p$ in 1D to higher dimensions, by sorting samples according to their projection obtained via the Hilbert space-filling curve. This alternative can be computed in $\mathcal{O}(n \log(n))$, but yields accurate approximations only for low dimensions, as emphasized in [Bernton et al., 2019]. The same work also uses a second approximation, the *swapping distance*, based on an iterative greedy swapping algorithm. However, each iteration requires n^2 operations, and there is no guarantee of convergence to $\mathbf{W}_p$.

4.3 Sliced-Wasserstein ABC

Given the computational and statistical issues caused by the Wasserstein distance, we state that the ABC framework can benefit from using an alternative computational

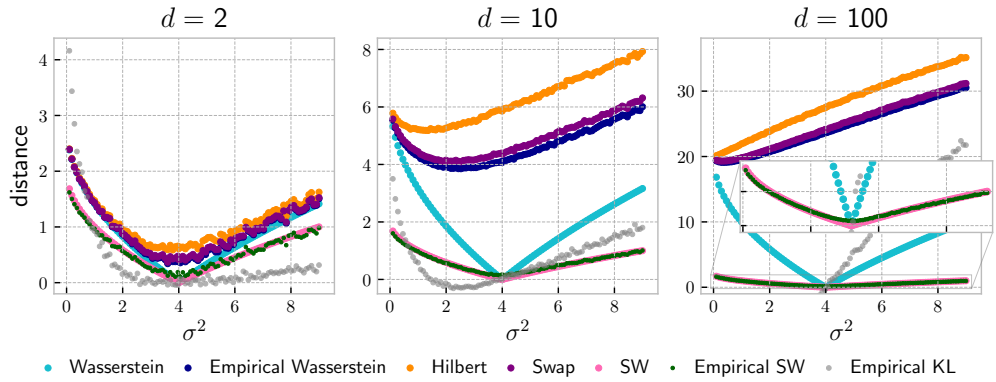


Figure 4.1: Comparison of OT distances and KL between data generated from d-dimensional Gaussian distributions μ_σ vs. $\mu_{\sigma_\star}$, $\sigma_\star^2 = 4$, with 1000 i.i.d draws. SW is approximated with 100 random projections.

OT metric, especially on high-dimensional settings. We consider the Sliced-Wasserstein distance based on the computational efficiency of its Monte Carlo estimate, as well as its statistical advantages over the Wasserstein distance and the Hilbert and swapping approximations.

We illustrate the latter statistical aspect on the task of estimating the scaling factor of the covariance matrix in a multivariate Normal model, as in the supplementary material of [Bernton et al., 2019]. For any $\sigma > 0$, denote by μ_σ the d-dimensional Gaussian distribution with zero-mean and covariance matrix $\sigma^2 \mathbf{I}_d$. Observations are assumed i.i.d. from $\mu_{\sigma_\star}$ with $\sigma_\star^2 = 4$, and we draw the same number of i.i.d. data from μ_σ for 100 values of σ^2 equispaced between 0.1 and 9. We then compute $\mathbf{W}_2$ and $\mathbf{SW}_2$ between the empirical distributions of the samples, and the swapping and Hilbert approximations presented in Bernton et al. [2019], for $d \in \{2, 10, 100\}$ and 1000 observations. We know that $\mathbf{W}_2$ between two Gaussian measures has an analytical formula, which boils down in our setting to

$$\mathbf{W}_2^2(\mu_{\sigma_\star}, \mu_\sigma) = d(\sigma_\star - \sigma)^2 \,, \tag{4.3}$$

and we approximate the exact SW using a Monte Carlo approximation of

$$\mathbf{SW}_2^2(\mu_{\sigma_\star}, \mu_\sigma) = \mathbf{W}_2^2(\mu_{\sigma_\star}, \mu_\sigma) \int_{\mathbb{S}^{d-1}} u^T u \, \mathrm{d}\boldsymbol{\sigma}(u) \,, \tag{4.4}$$

This formula (4.4) follows from Definition 2.9 and (4.3). We approximate KL with the estimator proposed for KL-based ABC (KL-ABC, Jiang et al. [2018]).

Figure 4.1 shows the distances plotted against σ^2 for each d. When the dimension increases, we observe that (i) as pointed out in Bernton et al. [2019], the quality of the approximation of empirical Wasserstein returned by Hilbert and swapping rapidly deteriorates, and (ii) SW, approximated using densities or samples, is the only approximate metric that attains its minimum at $\sigma_\star^2$. This curse of dimensionality can be a limiting factor for the performance of WABC and KL-ABC in high dimensions.

Motivated by the practical success of SW regardless of the dimension value in the previous experiment, we propose a variant of ABC based on SW, referred to as *Sliced-Wasserstein ABC* (SW-ABC). Our method is similar to WABC in the sense that it compares empirical distributions, but instead of $\mathbf{W}_p$, we choose the discrepancy measure to be $\mathbf{SW}_p$, $p \in [1, +\infty)$. The usage of SW allows the method to scale better to the data

size and dimension. The resulting posterior distribution, called the SW-ABC posterior, is thus defined in (4.2) with $\mathbf{D}$ replaced by $\mathbf{SW}_p$.

4.4 Theoretical Study

In this section, we analyze the asymptotic behavior of the SW-ABC posterior under two different regimes. Our first result concerns the situation where the observations $y_{1:n}$ are fixed, and ε goes to zero. We prove that the SW-ABC posterior is *asymptotically consistent* in the sense that it converges to the true posterior, under specific assumptions on the density used to generate synthetic data.

Proposition 4.1. *Let $p \in [1, +\infty)$. Suppose that μ_θ has a density f_θ w.r.t. the Lebesgue measure such that f_θ is continuous and there exists $\mathcal{N}_\Theta \subset \Theta$ satisfying $\pi(\mathcal{N}_\Theta) = 0$ and*

$$\sup_{\theta \in \Theta \setminus \mathcal{N}_\Theta} f_\theta(y_{1:n}) < \infty \, .$$

In addition, assume that there exists $\bar{\varepsilon} > 0$ such that,

$$\sup_{\theta \in \Theta \setminus \mathcal{N}_\Theta} \sup_{z_{1:m} \in \mathsf{A}^{\bar{\varepsilon}}} f_\theta(z_{1:m}) < \infty \, ,$$

where $\mathsf{A}^{\bar{\varepsilon}} = \{z_{1:m} \, : \, \mathbf{SW}_p(y_{1:n}, z_{1:m}) \leq \bar{\varepsilon}\}$. Then, with $y_{1:n}$ fixed, the SW-ABC posterior converges to the true posterior as ε goes to 0, in the sense that, for any measurable $\mathsf{B} \subset \Theta$,

$$\lim_{\varepsilon \to 0} \pi^\varepsilon_{y_{1:n}}(\mathsf{B}) = \pi(\mathsf{B} | y_{1:n}) \, ,$$

where $\pi^\varepsilon_{y_{1:n}}$ is defined by (4.2).

The proof of Proposition 4.1 is provided in Section 4.7 and consists in applying [Bernton et al., 2019, Proposition 3.1].

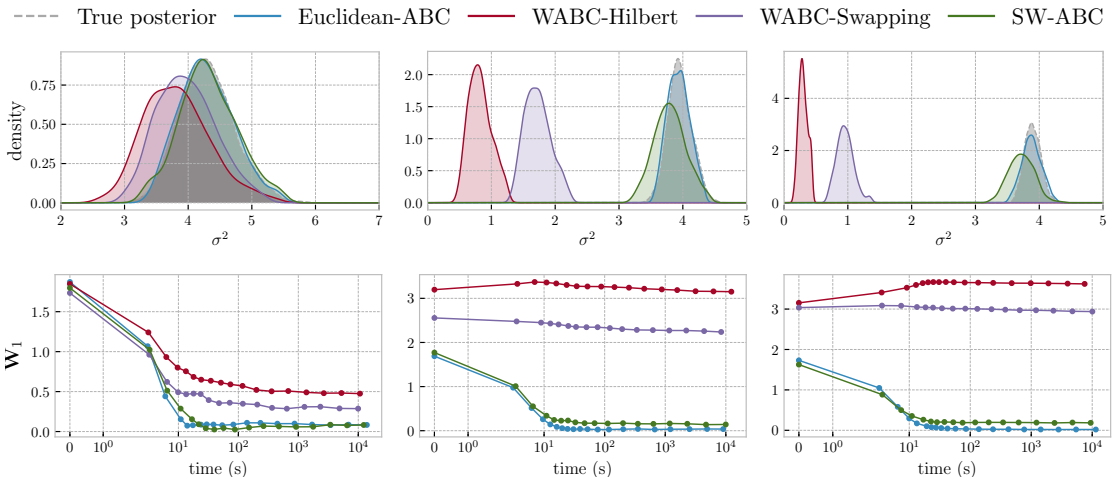
Next, we study the limiting SW-ABC posterior when the value of ε is fixed and the number of observations increases, *i.e.* $n \to \infty$. We suppose that the size m of the synthetic dataset grows to αn with $\alpha > 0$, such that m can be written as a function of n, $m(n)$, satisfying $\lim_{n \to \infty} m(n) = \infty$. We show that, under this setting and appropriate additional conditions, the resulting approximate posterior converges to the prior distribution on θ restricted to the region $\{\theta \in \Theta \, : \, \mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\}$.

Proposition 4.2. *Let $p \in [1, +\infty)$, $\varepsilon > 0$ and $(m(n))_{n \in \mathbb{N}^*}$ be an increasing sequence satisfying $\lim_{n \to \infty} m(n)/n = \alpha$, for $\alpha > 0$. Assume that the statistical model $\mathscr{M}_\Theta$ is well specified, i.e. there exists $\theta_\star \in \Theta$ such that $\mu_\star = \mu_{\theta_\star}$, and that almost surely the following holds.*

$$\lim_{n \to \infty} \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta, m(n)}) = \mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \, , \tag{4.5}$$

where $\hat{\mu}_n$, $\hat{\mu}_{\theta,m(n)}$ denote the empirical distributions of the observations $y_{1:n}$ and synthetic data $z_{1:m(n)}$ respectively. Then, the SW-ABC posterior converges to the restriction of the prior π on the region $\{\theta \in \Theta \, : \, \mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\}$ as $n \to \infty$, i.e. for any $\theta \in \Theta$,

$$\begin{aligned} \lim_{n \to \infty} \pi^\varepsilon_{\mathbf{y}_{1:n}}(\theta) &= \pi(\theta | \mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon) \\ &\propto \pi(\theta) \mathbb{1}\{\mathbf{SW}_p(\mu_{\theta_\star}, \mu_\theta) \leq \varepsilon\} \, . \end{aligned}$$



(a) $d = 2$ (b) $d = 10$ (c) $d = 20$

Figure 4.2: Comparison of SMC-ABC strategies in the multivariate Gaussians problem. Each strategy uses 1000 particles and are run for 3 hours max. First row shows ABC and true posteriors of σ^2, second row reports $\mathbf{W}_1$-distance to true posterior vs. time. SW is approximated with its MC estimate over 100 random projections.

Proposition 4.2 follows from the application of [Jiang et al., 2018, Theorem 1] to $\mathbf{SW}_p$ and the required conditions. Note that condition (4.5) is a mild assumption, e.g. is fulfilled if Y is compact and separable: in this case, for any $\nu \in \mathcal{P}_p(\mathsf{Y})$ and its empirical instantiation $\hat{\nu}_n$, $\lim_{n\to\infty} \mathbf{W}_p(\nu, \hat{\nu}_n) = 0$ ν-almost surely [Weed and Bach, 2019], then $\lim_{n\to\infty} \mathbf{SW}_p(\nu, \hat{\nu}_n) = 0$ ν-almost surely [Bonnotte, 2013, Proposition 5.1.3], and (4.5) follows by applying the triangle inequality.

4.5 Experiments

4.5.1 Synthetic experiments

As a first set of experiments, we investigate the performance of SW-ABC on a synthetical setting where the posterior distribution is analytically available. We consider $n = 100$ observations $(y_i)_{i=1}^n$ i.i.d. from a d-dimensional Gaussian $\mathcal{N}(\mathbf{m}_\star, \sigma_\star^2 \mathbf{I}_d)$, with $\mathbf{m}_\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\sigma_\star^2 = 4$. The parameter θ is σ^2 for which the prior distribution is assigned to be an inverse gamma distribution $\mathcal{IG}(1, 1)$. Therefore, the posterior distribution of σ^2 given $(y_i)_{i=1}^n$ and $\mathbf{m}_\star$ is an inverse gamma distribution as well, whose parameters are given by

$$\mathcal{IG}\big(1 + \frac{nd}{2}, 1 + \frac{1}{2}\sum_{i=1}^{n} \|y_i - \mathbf{m}_\star\|^2\big) \,.$$

We compare SW-ABC against ABC using the Euclidean distance between sample variances (Euclidean-ABC), WABC with the Hilbert distance, WABC with the swapping distance and KL-ABC. Each ABC approximation was obtained using the *sequential Monte Carlo sampler-based ABC* method [Toni et al., 2009], which is computationally more efficient than vanilla ABC (Algorithm 2) and implemented in the package `pyABC` [Klinger et al., 2018]. We provide the code to reproduce our empirical results[1].

[1] See our GitHub repository: `https://github.com/kimiandj/slicedwass_abc`

Figure 4.2 reports for $d \in \{2, 10, 20\}$, the resulting ABC posteriors and $\mathbf{W}_1$ to the true posterior (computed with the POT package [Flamary et al., 2021]) vs. time. Due to the poor performance of the estimator of KL between two empirical distributions proposed in [Jiang et al., 2018] (see Figure 4.1), KL-ABC fails at approximating well the posterior in these experiments. Hence, we excluded those results from Figure 4.2 for clarity. Euclidean-ABC provides the most accurate approximation, as expected since it relies on statistics that are sufficient in our setting. WABC performs poorly with high-dimensional observations, contrary to SW-ABC, which approximates well the posterior for each dimension value and is as fast.

4.5.2 Application to image denoising

We now evaluate our approach on a real application, namely image denoising. We consider a widely used algorithm for this task, the *Non-Local Means algorithm* (NL-means, Buades et al. [2005]), and we present a novel variant of it derived from SW-ABC.

We formally define the denoising problem as follows. Let $\mathbf{u} \in \mathbb{R}^{M \times N}$, denote a clean gray-level image. We observe a corrupted version of this image, $\mathbf{v} = \mathbf{u} + \mathbf{w}$, where $\mathbf{w}$ is some noise in $\mathbb{R}^{M \times N}$. The goal is to restore $\mathbf{u}$ from $\mathbf{v}$. We focus on denoising methods that consider 'patch-based representations' of images, e.g. NL-means. Specifically, let $r \in \mathbb{N}$ be a patch size and $I = \{1, \dots, M\} \times \{1, \dots, N\}$ the set of pixel positions. For $\mathbf{i} \in I$, $\mathbf{u}'(\mathbf{i})$ denotes the pixel value at position $\mathbf{i}$ in image $\mathbf{u}'$, and $P_\mathbf{i}$ is a $(2r+1) \times (2r+1)$ window in $\mathbf{v}$ centered at $\mathbf{i}$: for $\mathbf{k} \in \{-r, \dots, r\}^2$, $P_\mathbf{i}(\mathbf{k}) = \mathbf{v}(\mathbf{i} + \mathbf{k})$, where $\mathbf{v}$ is extended to $\mathbb{Z}^2$ by periodicity. Let $D \subset I$ be a dictionary of positions, and $\phi : I \to D$ such that, for $\mathbf{i} \in I$,

$$\phi(\mathbf{i}) = \operatorname{argmin}_{\mathbf{j} \in D} \|P_\mathbf{i} - P_\mathbf{j}\|_2 \ ,$$

i.e. $\phi(\mathbf{i})$ is the position in D of the most similar patch to $P_\mathbf{i}$. For $\mathbf{j} \in D$, an estimator of $P_\mathbf{j}$ is given by $\hat{P}_\mathbf{j} = \mathbb{E}_{\pi(\mathbf{i}|(P_\mathbf{k})_{\mathbf{k} \in \phi^{-1}(\mathbf{j})})\tilde{\pi}(\mathbf{l})}[P_{\mathbf{i}+\mathbf{l}}]$, $\tilde{\pi}$ being the uniform distribution on $\phi^{-1}(\mathbf{j})$. In practice, it is approximated with a Monte Carlo scheme,

$$\hat{P}_\mathbf{j} \approx (Tn)^{-1} \sum_{t=1}^{T} \sum_{s=1}^{S} P_{\mathbf{i}^{(t)}+\mathbf{l}^{(s)}} \ , \tag{4.6}$$

where $\mathbf{i}^{(t)} \sim \pi(\mathbf{i}^{(t)}|(P_\mathbf{k})_{\mathbf{k} \in \phi^{-1}(\mathbf{j})})$, $\mathbf{l}^{(s)} \sim \tilde{\pi}(\mathbf{l})$, and $\mathbf{i}$, $\mathbf{l}$ are mutually independent. Finally, we construct an estimate $\hat{\mathbf{u}}$ of $\mathbf{u}$ as follows: for any $\mathbf{i} \in I$,

$$\hat{\mathbf{u}}(\mathbf{i}) = \sum_{\mathbf{k}, \|\mathbf{k}-\mathbf{i}\|_\infty \leq r} \hat{P}_{\phi(\mathbf{k})}(\mathbf{i} - \mathbf{k}) \ (2r+1)^{-2} \ .$$

The classical NL-means estimator corresponds to the case where $D = I$ (thus $\phi = \mathrm{Id}$) and for any $\mathbf{i} \in I$ and $P \in \mathbb{R}^{(2r+1)\times(2r+1)}$, $\pi(\mathbf{i}, P) \propto \mathbb{1}_W(\mathbf{i}) \mathrm{e}^{-\|P - P_\mathbf{i}\|^2/(2\sigma^2)}$, where W is a search window.

In our work, we assume that the likelihood $\pi(P|\mathbf{i})$ is not available, but we observe for $j \in D$, $(P_{\mathbf{k}_\ell})_{\ell=1}^m$ ($\mathbf{k}_\ell \in \phi^{-1}(\mathbf{j})$) i.i.d. from $\pi(\cdot|\mathbf{i})$. By replacing $\pi(\mathbf{i}|(P_{\mathbf{k}_\ell})_{\ell=1}^m)$ in (4.6) by the SW-ABC posterior, we obtain the proposed denoising method, called *the SW-ABC NL-means algorithm*. We provide the Python implementation of our algorithm[2].

We compare our approach with the classical NL-means. We consider one of the standard image denoising datasets [Fan et al., 2019], called CBSD68 [Martin et al., 2001]

[2] See https://vdeborto.github.io/publication/sw_abc

	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	$\sigma = 50$
NL-means	**30.43**	**26.32**	24.22	21.99
SW-ABC	27.09	26.26	**24.86**	**22.56**

Table 4.1: Comparison of NL-means and SW-ABC on the image denoising task in terms of average PSNR (dB). For each σ, we fine-tuned the hyperparameters of NL-means and reported the best result.

and consisting of 68 colored images of size 321×481. We first convert the images to gray scale, then manually corrupt each of them with a Gaussian noise with standard deviation σ, and try to recover the clean image. The quality of the output images is evaluated with the Peak Signal to Noise Ratio (PSNR) measure,

$$\mathrm{PSNR} = -10 \log_{10} \left(\frac{\|\mathbf{u} - \hat{\mathbf{u}}\|_2^2}{255^2 NM} \right) .$$

In our experiments, we use a dictionary of 1000 patches picked uniformly at random, we set $T = S = m = 10$, $r = 3$, $W = \{-10, \dots, 10\}^2$, $\varepsilon = (2r+1)^2$, and we compute SW with a MC scheme over $L = 100$ projections.

We report the average PSNR values for different values of the noise level σ in Table 4.1. We observe that for small σ, NL-means provides more accurate results, whereas when σ becomes larger SW-ABC outperforms NL-means, thanks to the patch representation and the use of SW.

On the other hand, another important advantage of SW-ABC becomes prominent in the computation time: the proposed approach takes $\approx 6s$ on a standard laptop computer per image whereas the classical NL-means algorithm takes $\approx 30s$. Indeed, the computational complexity of SW-ABC NL-means is upper-bounded by $\mathrm{card}(D)TSC_{\mathrm{SW}}$, where $C_{\mathrm{SW}} = Lm\log(m)$ is the cost of computing SW, and for the naïve implementation of NL-means, it is given by $NM\,\mathrm{card}(W)(2r+1)^2$. We can observe that SW-ABC has a lower computational complexity since $\mathrm{card}(D) \ll NM$ in practice. We note that the computation time of NL-means can be improved by certain acceleration techniques, which can be directly used to improve the speed of SW-ABC NL-means as well.

Finally, in Figure 4.3, we illustrate the performance of SW-ABC on two 512×512 images for visual inspection. The results show that the injected noise is successfully removed by the proposed approach.

4.6 Conclusion

In this chapter, we explored other applications where the Sliced-Wasserstein distance can be useful, and proposed a novel ABC method, SW-ABC, based on this metric. We derived asymptotic guarantees for the convergence of the SW-ABC posterior to the true posterior under different regimes, and we evaluated our approach on a synthetical and an image denoising problem. Our results showed that SW-ABC provides an accurate approximation of the posterior, even with high-dimensional data spaces and a small number of samples.

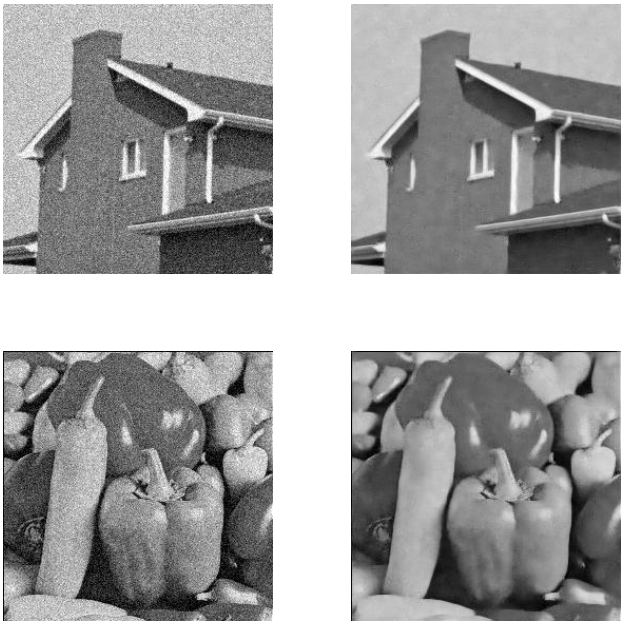


Figure 4.3: Visualization of the results. For each couple, the left one is the noisy image ($\sigma = 20$) and the right one is the output of SW-ABC.

4.7 Appendix: Proof of Proposition 4.1

Proof of Proposition 4.1. The proof consists in applying [Bernton et al., 2019, Proposition 3.1], which establishes the conditions for the data discrepancy measure to yield an ABC posterior that converges to the true posterior in the asymptotic regime we consider. This amounts to verify that:

(i) For any $\mathbf{y}_{1:n}$ and $\mathbf{z}_{1:m}$, with respective empirical distributions $\hat{\mu}_n$ and $\hat{\mu}_{\theta,m}$, $\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) = 0$ if and only if $\hat{\mu}_n = \hat{\mu}_{\theta,m}$.

(ii) $\mathbf{SW}_p$ is continuous in the sense that, if $(\mathbf{z}^k_{1:m})_{k\in\mathbb{N}}$ converges to $\mathbf{z}_{1:m}$ in the metric ρ, then, for any empirical distribution $\hat{\mu}_n$, $\lim_{k\to\infty} \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}^k_{\theta,m}) = \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$, where $\hat{\mu}^k_{\theta,m}$ is the empirical measure of $\mathbf{z}^k_{1:m}$.

Condition (i) follows from the fact that $\mathbf{SW}_p$ is a distance [Bonnotte, 2013, Proposition 5.1.2]. Now, let $\mathbf{y}' \in \mathsf{Y}$ and $\psi : \mathsf{Y} \to \mathbb{R}$ be a continuous function such that for any $\mathbf{y} \in \mathsf{Y}$, $|\psi(\mathbf{y})| \leq K\big(1 + \rho(\mathbf{y}', \mathbf{y})^p\big)$ with $K \in \mathbb{R}$. Since $(\mathbf{z}^k_{1:m})_{k\in\mathbb{N}}$ converges to $\mathbf{z}_{1:m}$ in the metric ρ and ψ is continuous, we get that $\lim_{k\to\infty} \int \psi \, \mathrm{d}\hat{\mu}^k_{\theta,m} = \int \psi \, \mathrm{d}\hat{\mu}_{\theta,m}$. This implies that $\hat{\mu}^k_{\theta,m}$ weakly converges to $\hat{\mu}_{\theta,m}$ in $\mathcal{P}_p(\mathsf{Y})$ [Villani, 2008, Definition 6.7], which, by [Villani, 2008, Theorem 6.8], is equivalent to $\lim_{k\to\infty} \mathbf{W}_p(\hat{\mu}^k_{\theta,m}, \hat{\mu}_{\theta,m}) = 0$. By applying the triangle inequality and [Bonnotte, 2013, Proposition 5.1.3], there exists $C \geq 0$ such that, for any empirical measure $\hat{\mu}_n$,

$$\begin{aligned} |\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}^k_{\theta,m}) - \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})| &\leq \mathbf{SW}_p(\hat{\mu}^k_{\theta,m}, \hat{\mu}_{\theta,m}) \\ &\leq C^{1/p} \, \mathbf{W}_p(\hat{\mu}^k_{\theta,m}, \hat{\mu}_{\theta,m}) \, . \end{aligned}$$

We conclude that $\lim_{k\to\infty} \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}^k_{\theta,m}) = \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$, making condition (ii) applicable.

□

Chapter 5

Generalized Sliced Wasserstein Distances

This chapter is based on [Kolouri et al., 2019a].

The Wasserstein distance and its practical alternatives have recently attracted a lot of attention from the machine learning community. The Sliced-Wasserstein distance, specifically, was shown to have similar theoretical properties to the Wasserstein distance, while being much simpler to compute thanks to its Monte Carlo approximation. SW has therefore been used in various applications, including generative modeling and general supervised/unsupervised learning, but its performance might suffer from the error induced by the Monte Carlo estimation. This limitation has thus motivated the formulation of alternatives, such as the *maximum Sliced-Wasserstein distance* (max-SW).

In this chapter, we propose another method to address this issue, by defining a novel family of probability divergences that extends the idea behind SW. We first clarify the mathematical connection between SW and the Radon transform, then leverage the *generalized* Radon transform to formulate the class of *generalized Sliced-Wasserstein distances* (GSW). We also formulate a generalization of max-SW, called the *maximum generalized Sliced-Wasserstein distances* (max-GSW). We identify some conditions on the generalized Radon transform under which GSW and max-GSW satisfy the metric axioms. Finally, we compare the empirical performance of the proposed distances on different implicit generative modeling problems to illustrate their advantages over SW.

5.1 Introduction

To compare two probability distributions μ and ν supported on $\mathbb{R}^d$ with the Sliced-Wasserstein distance, one needs to collect linear projections of μ and ν along all possible directions on $\mathbb{S}^{d-1}$, which is done by computing the push-forward measures $\theta^\star_\sharp\mu$ and $\theta^\star_\sharp\nu$ for any $\theta \in \mathbb{S}^{d-1}$. As we will detail in the next section, these push-forward measures are closely related to the Radon transform [Rabin et al., 2012, Proposition 6], which is widely used in tomography [Radon, 1917, Helgason, 2011]. In practice, unless an analytical formula is known, the integral that defines SW (Definition 2.9) is usually approximated with a Monte Carlo strategy, which computes an average over a finite number of directions uniformly picked at random on $\mathbb{S}^{d-1}$. Intuitively, the linear nature of these projections does not guarantee an efficient approximation of the Sliced-Wasserstein distance: since in very high-dimensional settings, the data often lives in a thin manifold, one might

have to sample a very large number of directions to effectively capture the structure of the data distribution, which can be very expensive according to our discussion on the complexity of SW in Section 2.6. Reducing the number of required projections would thus result in a significant performance improvement.

To alleviate the inefficiencies caused by the linear projections, several attempts have recently been made: linear projections can be combined with orthogonal coupling in Monte Carlo estimation to increase computational efficiency and estimation quality [Rowland et al., 2019, Wu et al., 2019]. Alternatively, Deshpande et al. [2019] extended SW to the "*maximum Sliced-Wasserstein distance*", where the integral over $\mathbb{S}^{d-1}$ in (2.20) is replaced by a maximum operator so that one retains the "most informative" projection direction. In this context, the information returned by a direction $\theta \in \mathbb{S}^{d-1}$ is measured by $\mathbf{W}_p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu)$: the larger this Wasserstein distance, the more informative θ. This idea is also reflected in another study [Paty and Cuturi, 2019], which considers k-dimensional projections of μ, ν with $k \in \{1, \dots, d\}$, and aims at finding the most informative *subspace* on which μ and ν are being projected. While these methods reduce the computational cost by requiring a lower number of projections, they incur an additional cost due to the resolution of a non-convex optimization problem over manifolds.

In this chapter, we address the computational limitations of the Sliced-Wasserstein distance by taking an alternative route: we allow the projections of the compared distributions μ and ν to be *non-linear*. More precisely, we use the theory of the *generalized Radon transform* [Beylkin, 1984] to extend the definition of SW to an entire class of probability divergences, which we call *generalized Sliced-Wasserstein distances* (GSW). We then show that, similar to [Deshpande et al., 2019], we can formulate a metric that relies on the most informative projection instead of infinitely many projections. We aptly call this distance the *maximum generalized Sliced-Wasserstein distance* (max-GSW). We prove that replacing the *linear projections* with *non-linear projections* can still yield a valid metric on the space of probability distributions: we identify general conditions under which GSW and max-GSW satisfy the metric axioms recalled in Definition 2.1.

As instances of non-linear projections, we first investigate projections with *polynomial kernels*, which meet all the conditions that we identified. However, we observe that the memory complexity required by such projections has a combinatorial growth with respect to the dimension of the problem, which hinders their applications to modern ML problems, such as IGM. This motivates us to consider a *neural-network-based projection scheme*, where we observe that fully connected or convolutional networks with leaky ReLU activations fulfill the crucial conditions for the resulting GSW to be a pseudo-metric.

Due to their inherent non-linearity, GSW and max-GSW are expected to capture the complex structure of high-dimensional distributions by using much less projections. Besides, the use of deep learning techniques additionally allows the projections to be data-adaptive. For these reasons, we expect our metrics to reduce the iteration complexity in a significant amount. We verify this intuition in our experiments, where we illustrate the superior performance of the proposed generalized distances in IGM problems, with both synthetic and real data.

5.2 Background on the Radon transform

We review in this section the Radon transform and explain how it enables the definition of SW and max-SW, accordingly to [Rabin et al., 2012]. Then, we present an extension of that transform, namely the generalized Radon transform, which is a fundamental tool in this chapter.

5.2.1 Definition of the Radon transform

Before formally defining the Radon transform, we present some useful notations. Denote by $\mathcal{L}^1(\mathbb{R}^d, \text{Leb}_d) = \left\{f : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } \int_{\mathbb{R}^d} |f(x)| \, \text{dLeb}_d(x) < \infty\right\}$ the class of functions that are absolutely integrable on $\mathbb{R}^d$ w.r.t. the Lebesgue measure on $\mathbb{R}^d$, Leb_d, and $\mathcal{L}^1(\mathbb{S}^{d-1} \times \mathbb{R}, \boldsymbol{\sigma} \otimes \text{Leb}_1)$ the class of absolutely integrable functions on the domain $\mathbb{S}^{d-1} \times \mathbb{R}$ w.r.t. $\boldsymbol{\sigma} \otimes \text{Leb}_1$.

The Radon transform, introduced in [Radon, 1917] and denoted by $\mathcal{R}$, maps a function in $\mathcal{L}^1(\mathbb{R}^d, \text{Leb}_d)$ to the infinite set of its integrals over the hyperplanes of $\mathbb{R}^d$, *i.e.* $\mathcal{R} : \mathcal{L}^1(\mathbb{R}^d, \text{Leb}_d) \to \mathcal{L}^1(\mathbb{S}^{d-1} \times \mathbb{R}, \boldsymbol{\sigma} \otimes \text{Leb}_1)$, and is defined as follows.

Definition 5.1 (Radon transform). *Let $I \in \mathcal{L}^1(\mathbb{R}^d, \text{Leb}_d)$. The Radon transform associated to I is a function defined for any $(t, \theta) \in \mathbb{R} \times \mathbb{S}^{d-1}$ as*

$$\mathcal{R}I(t, \theta) = \int_{\mathbb{R}^d} I(x)\delta(t - \langle x, \theta \rangle) \mathrm{d}x \; . \tag{5.1}$$

Note that by definition, each hyperplane in $\mathbb{R}^d$ can be written as

$$\mathsf{H}_{t,\theta} = \left\{x \in \mathbb{R}^d \; : \; \langle x, \theta \rangle = t\right\} \; , \tag{5.2}$$

where $t \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$. Therfore, for a fixed $\theta \in \mathbb{S}^{d-1}$, $\mathcal{R}I(\cdot, \theta) : \mathbb{R} \to \mathbb{R}$ integrates the input function I over all hyperplanes in $\mathbb{R}^d$ that are orthogonal to θ, and (5.1) can be rewritten as

$$\mathcal{R}I(t, \theta) = \int_{\mathsf{H}_{t,\theta}} I(x) \mathrm{d}x \; . \tag{5.3}$$

The Radon transform has been shown to be invertible, which allows to recover a function I out of its projections along hyperplanes, $\mathcal{R}I$ [Natterer, 1986, Helgason, 2011]. In particular, the *filtered back-projection* method defines the inverse formula of the transform when $I \in \mathcal{L}^1(\mathbb{R}^2, \text{Leb}_2)$ and has been extensively used for image reconstruction, for example in tomographic imaging [Deans, 2007].

On the other hand, Definition 5.1 can be extended so that the Radon transform applies to measures instead of functions in $\mathcal{L}^1(\mathbb{R}^d, \text{Leb}_d)$. The formal statement [Bonneel et al., 2015, Definition 6] is recalled below.

Definition 5.2 (Radon transform of measures). *Denote by $\mathcal{C}_0(\mathsf{A})$ the space of continuous functions on a set A that tend to 0 at infinity. Let $\mu \in \mathcal{P}(\mathbb{R}^d)$. The Radon transform associated to μ, denoted by $\mathcal{R}\mu$, is defined through the following characterization: for any $f \in \mathcal{C}_0(\mathbb{R} \times \mathbb{S}^{d-1})$,*

$$\int_{\mathbb{R} \times \mathbb{S}^{d-1}} f(t, \theta) \mathrm{d}(\mathcal{R}\mu)(t, \theta) = \int_{\mathbb{R}^d} (\mathcal{R}^* f)(x) \mathrm{d}\mu(x) \; , \tag{5.4}$$

where $\mathcal{R}^* : \mathcal{C}_0(\mathbb{R} \times \mathbb{S}^{d-1}) \to \mathcal{C}_0(\mathbb{R}^d)$ *is the* back-projection operator *defined as*

$$\mathcal{R}^* f(x) = \int_{\mathbb{S}^{d-1}} f(\langle x, \theta \rangle, \theta) \mathrm{d}\theta \ . \tag{5.5}$$

Hence, $\mathcal{R}\mu : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R} \times \mathbb{S}^{d-1})$.

Note that the Radon transform of measures can actually be defined on the general space of *Radon measures* supported on $\mathbb{R}^d$, instead of the subset of probability measures $\mathcal{P}(\mathbb{R}^d)$: the reason why we consider $\mu \in \mathcal{P}(\mathbb{R}^d)$ in Definition 5.2 is that we aim at clarifying the link between the Sliced-Wasserstein distance and the Radon transform.

5.2.2 Link between Radon transform and Sliced-Wasserstein distance

We first present an important result originally established in [Bonneel et al., 2015, Proposition 6], which shows that Radon transforms of measures are equivalent to specific push-forward measures.

Proposition 5.3. *Let* $\mu \in \mathcal{P}(\mathbb{R}^d)$. *Then, for any* $\theta \in \mathbb{S}^{d-1}$,

$$\mathcal{R}\mu(\cdot, \theta) = \theta^\star_\sharp \mu \ , \tag{5.6}$$

where $\theta^\star_\sharp$ *denotes the push-forward operator (Definition 2.5) associated to the linear form* $\theta^\star(x) = \langle \theta, x \rangle$.

Therefore, the one-dimensional representations of $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ computed by the Sliced-Wasserstein distance, and respectively denoted by $\theta^\star_\sharp \mu$ and $\theta^\star_\sharp \nu$ for $\theta \in \mathbb{S}^{d-1}$, are actually obtained via the Radon transform: by Proposition 5.3, the definition of SW of order $p \in [1, +\infty)$ between μ and ν (Definition 2.9) can equivalently be formulated as

$$\mathbf{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\mathcal{R}\mu(\cdot, \theta), \mathcal{R}\nu(\cdot, \theta)\big) \mathrm{d}\boldsymbol{\sigma}(\theta) \ . \tag{5.7}$$

In practice, $\mathbf{SW}_p(\mu, \nu)$ is approximated with a simple Monte Carlo scheme (Section 2.6), which may lead to underestimating the actual dissimilarity between μ and ν, especially when these two distributions are supported on a high-dimensional space. To further illustrate this phenomenon, let us consider $\mu = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\nu = \mathcal{N}(\mathbf{m}, \mathbf{I}_d)$ with $\mathbf{m} \in \mathbb{S}^{d-1}$. Then, their projected representations are univariate Gaussians, given for any $\theta \in \mathbb{S}^{d-1}$ by,

$$\mathcal{R}\mu(\cdot, \theta) = \mathcal{N}(0, 1) \ , \quad \text{and} \quad \mathcal{R}\nu(\cdot, \theta) = \mathcal{N}(\langle \theta, \mathbf{m} \rangle, 1) \ . \tag{5.8}$$

It is therefore clear that $\mathbf{W}_2(\mathcal{R}\mu(\cdot, \theta), \mathcal{R}\nu(\cdot, \theta))$ achieves its maximum value when $\theta = \mathbf{m}$, and is equal to zero when θ is orthogonal to $\mathbf{m}$. On the other hand, an application of Hoeffding's inequality gives the following concentration inequality for any $\mathbf{m}' \in \mathbb{S}^{d-1}$,

$$\mathbb{P}(|\left\langle \theta, \mathbf{m}' \right\rangle| \leq \varepsilon) \geq 1 - 2e^{-\frac{d\varepsilon^2}{2}} \ , \tag{5.9}$$

which implies that for a high dimension d, $\theta \sim \boldsymbol{\sigma}$ is likely to be nearly orthogonal to $\mathbf{m}$, so $\mathbf{W}_2(\mathcal{R}\mu(\cdot, \theta), \mathcal{R}\nu(\cdot, \theta))$ is almost null with high probability.

To remedy the inaccuracies caused by the Monte Carlo estimation, one can pick projection directions that return discriminant information between μ and ν, instead of

uniformly sampling them on $\mathbb{S}^{d-1}$. This idea was for instance used in [Rowland et al., 2019, Wu et al., 2019], where the Monte Carlo samples $\{\theta_l\}_{l=1}^L$ forms a set of orthogonal vectors, or in [Deshpande et al., 2018, Section 3.2], where implements a GAN to find discriminant projections.

A similarly flavored but less heuristic approach consists in using the *maximum Sliced-Wasserstein distance* (max-SW), an alternative OT metric defined in [Deshpande et al., 2019] as follows.

Definition 5.4 (Maximum Sliced-Wasserstein distance)**.** *Let $p \in [1, +\infty)$. The maximum Sliced-Wasserstein distance of order p is defined for $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ as*

$$\text{max–}\mathbf{SW}_p(\mu, \nu) = \max_{\theta \in \mathbb{S}^{d-1}} \mathbf{W}_p(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) \ , \tag{5.10}$$

where for any $\theta \in \mathbb{S}^{d-1}$, $\theta^\star_\sharp = (\theta^\star)_\sharp$ denotes the push-forward operator associated to the linear form $\theta^\star : \mathbb{R}^d \to \mathbb{R}$ given by $\theta^\star(x) = \langle \theta, x \rangle$.

By Proposition 5.3, (5.10) is equivalent to

$$\text{max–}\mathbf{SW}_p(\mu, \nu) = \max_{\theta \in \mathbb{S}^{d-1}} \mathbf{W}_p(\mathcal{R}\mu(\cdot, \theta), \mathcal{R}\nu(\cdot, \theta)) \ . \tag{5.11}$$

Since $\mathbf{W}_p$ is a distance (Section 2.4.1), one can show that max–$\mathbf{SW}_p$ is also a distance: we will prove in Section 5.3.1 that the metric axioms hold for the class of maximum Generalized Sliced-Wasserstein distances, which contains max-SW as a special case.

5.2.3 Generalized Radon transform

The generalized Radon transform (GRT) extends the original idea of the classical Radon transform presented in Section 5.2.1 from integration over hyperplanes of $\mathbb{R}^d$ to integration over *hypersurfaces*, *i.e.* $(d-1)$-dimensional manifolds [Beylkin, 1984, Denisyuk, 1994, Ehrenpreis, 2003, Gel'fand et al., 1969, Kuchment, 2006, Homan and Zhou, 2017].

According to (5.2), any hyperplane of $\mathbb{R}^d$ can alternatively be interpreted as a level set of the function $g \in \mathbb{R}^d \times \mathbb{S}^{d-1} \to \mathbb{R}$ given by $g(x, \theta) = \langle x, \theta \rangle$. Therefore, to generalize the Radon transform, one can simply consider another function g, which is then referred to as the *defining function* and characterized as follows.

Definition 5.5 (Defining function)**.** *Consider a function $g : \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\}) \to \mathbb{R}$, where $\mathsf{X} \subset \mathbb{R}^d$ and $q \in \mathbb{N}^*$. We say that g is a* defining function *if it satisfies the four conditions below.*

(D1) *g is a real-valued C^∞ function on $\mathsf{X} \times (\mathbb{R}^n \backslash \{0\})$*

(D2) *g is homogeneous of degree 1 with respect to its second variable, i.e.*

$$\forall (x, \theta) \in \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\}), \ \forall \lambda \in \mathbb{R}, \ \ g(x, \lambda\theta) = \lambda g(x, \theta) \ .$$

(D3) *g is non-degenerate in the sense that*

$$\forall (x, \theta) \in \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\}), \ \ \frac{\partial g}{\partial x}(x, \theta) \neq 0 \ .$$

(D4) *The mixed Hessian of g is strictly positive, i.e.*

$$\text{Det}\left[\left(\frac{\partial^2 g}{\partial x^i \partial \theta^j}\right)_{i,j}\right] > 0 \ .$$

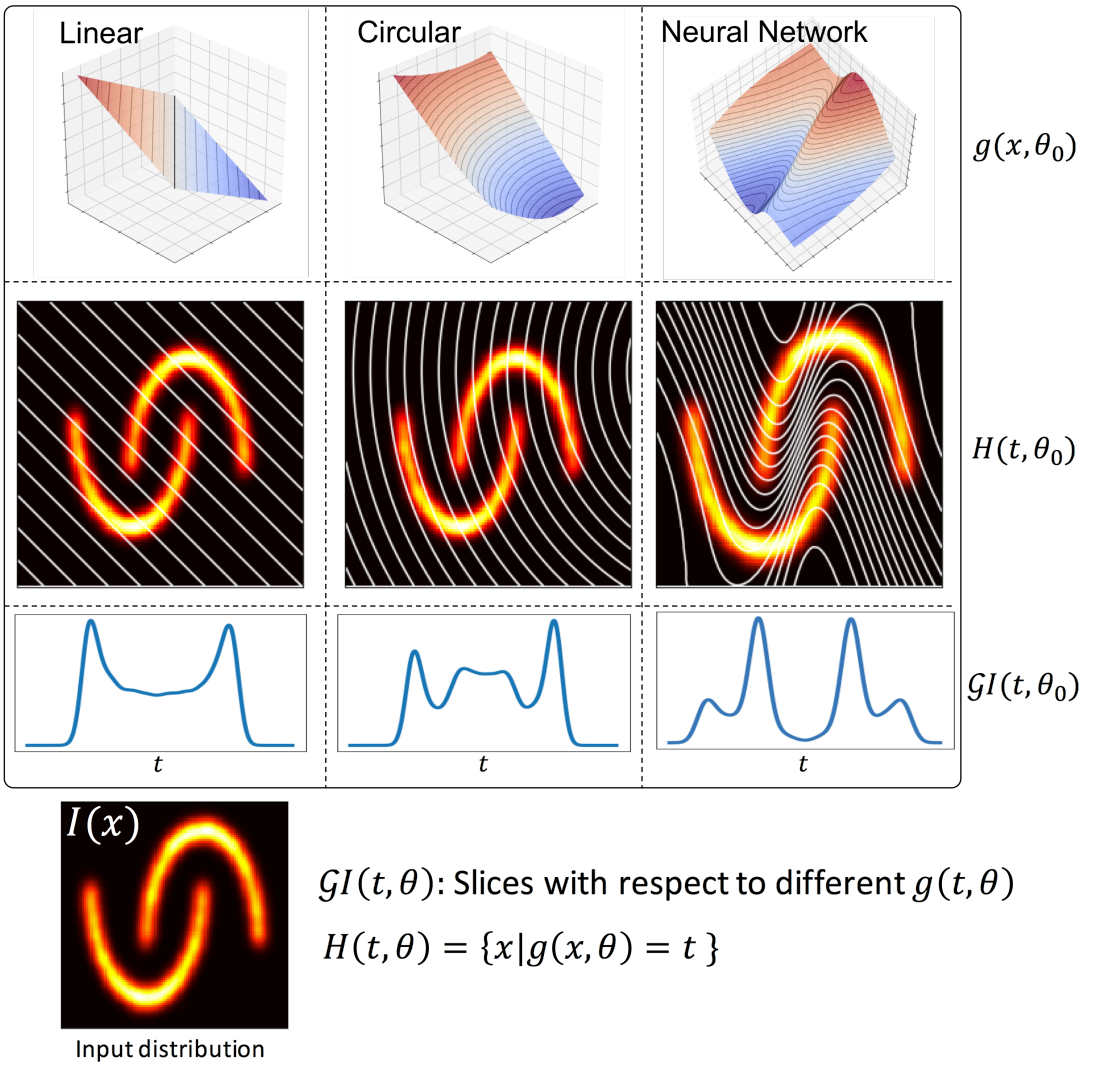


Figure 5.1: Illustration of the application of the standard or generalized Radon transform for the Half Moons distribution.

Then, given a defining function g, the generalized Radon transform associated to g and applied to $I \in \mathcal{L}^1(\mathbb{R}^d, \text{Leb}_d)$ integrates I over the hypersurfaces characterized by the level sets of g, *i.e.*

$$\tilde{\mathsf{H}}_{t,\theta} = \{x \in \mathsf{X} : \ g(x, \theta) = t\} \ . \tag{5.12}$$

Definition 5.6 (Generalized Radon transform). *Consider a defining function $g : \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\}) \to \mathbb{R}$, with $\mathsf{X} \subset \mathbb{R}^d$ and $q \in \mathbb{N}^*$. Let $I \in \mathcal{L}^1(\mathbb{R}^d, \text{Leb}_d)$. The generalized Radon transform of I based on g is a function defined for any $(t, \theta) \in \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\})$ as*

$$\mathcal{G}I(t, \theta; g) = \int_{\mathsf{X}} I(x)\delta(t - g(x, \theta))\mathrm{d}x \ . \tag{5.13}$$

According to Definition 5.6, the standard Radon transform (Definition 5.1) is indeed a special case of the GRT, as it is obtained with $g(x, \theta) = \langle x, \theta \rangle$ for $(x, \theta) \in \mathbb{R}^d \times \mathbb{S}^{d-1}$. GRT has various applications, including thermoacoustic tomography, where the hypersurfaces correspond to spheres, and electrical impedance tomography, which requires integration over hyperbolic surfaces.

One can also use GRT to project high-dimensional distributions along hypersurfaces and obtain one-dimensional representations: Definition 5.6 can be extended so that GRT receives probability distributions as input, analogously to Definition 5.2. In that case, the resulting projections of $\mu \in \mathcal{P}(\mathbb{R}^d)$ acquired via the GRT based on a defining function g correspond to $(g^\theta)_\sharp \mu$ for any $\theta \in \mathbb{R}^q \backslash \{\mathbf{0}\}$, where $(g^\theta)_\sharp$ is the push-forward operator of g^θ, defined for any $x \in \mathbb{R}^d$ as $g^\theta(x) = g(x, \theta)$. Figure 5.1 illustrates the application of different Radon transforms on the "Half Moons" dataset and demonstrates that the output projections highly depend on the defining function. This encourages us to investigate the consequences of using the generalized Radon transform instead of the standard one in the definition of SW (5.7).

5.3 Generalized Sliced-Wasserstein Distances

We propose in this chapter to extend the definition of the Sliced-Wasserstein distance to formulate new optimal transport metrics, which we call Generalized Sliced-Wasserstein distances. These are obtained using the same procedure as for SW, except that the one-dimensional representations are acquired through nonlinear projections via the generalized Radon transform. We also extend the concept of max-SW to the class of maximum generalized Sliced-Wasserstein distances (max-GSW).

In this section, we formally define (maximum) Generalized Sliced-Wasserstein distances and establish the conditions under which they satisfy the metric axioms. We then provide examples of defining functions such that these conditions are met, and present an alternative implementation of GSW inspired by neural networks.

5.3.1 Definition and theoretical properties

Following the definition of SW in terms of the Radon transform (5.7), we define Generalized Sliced-Wasserstein distances using the generalized Radon transform as follows.

Definition 5.7 (Generalized Sliced-Wasserstein distances)**.** *Consider a defining function $g : \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\}) \to \mathbb{R}$, with $\mathsf{X} \subset \mathbb{R}^d$ and $q \in \mathbb{N}^*$. Let $p \in [1, +\infty)$. The Generalized Sliced-Wasserstein distance of order p based on g is defined for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ as*

$$\mathbf{GSW}_p^p(\mu, \nu) = \int_{\mathbb{R}^q \backslash \{\mathbf{0}\}} \mathbf{W}_p^p\big(\mathcal{G}\mu(\cdot, \theta; g), \mathcal{G}\nu(\cdot, \theta; g)\big) \mathrm{d}\boldsymbol{\sigma}_q(\theta) \,, \tag{5.14}$$

where $\boldsymbol{\sigma}_q$ denotes the uniform distribution on $\mathbb{R}^q \backslash \{\mathbf{0}\}$.

We also formulate the maximum Generalized Sliced-Wasserstein distance, which generalizes the maximum Sliced-Wasserstein distance defined in (5.11).

Definition 5.8 (Maximum Generalized Sliced-Wasserstein distances)**.** *Consider a defining function $g : \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\}) \to \mathbb{R}$, with $\mathsf{X} \subset \mathbb{R}^d$ and $q \in \mathbb{N}^*$. Let $p \in [1, +\infty)$. The Maximum Generalized Sliced-Wasserstein distance of order p based on g is defined for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ as*

$$\textit{max}\text{-}\mathbf{GSW}_p(\mu, \nu) = \max_{\theta \in \Omega_\theta} \mathbf{W}_p\big(\mathcal{G}\mu(\cdot, \theta; g), \mathcal{G}\nu(\cdot, \theta; g)\big) \,, \tag{5.15}$$

where for any $\theta \in \mathbb{S}^{d-1}$, $\Omega_\theta \subset \mathbb{R}^q \backslash \{\mathbf{0}\}$ is a compact set of feasible parameters for $\theta \mapsto g(\cdot, \theta)$.

We point out that max-GSW must rely on a compact set Ω_θ so that (5.15) is not ill-defined. For instance, for the specific case of max-SW where $g(\cdot, \theta) = \theta^\star(\cdot)$, we have $\Omega_\theta = \mathbb{S}^{d-1}$.

In the next proposition, we show that GSW and max-GSW are, indeed, distances on $\mathcal{P}_p(\mathbb{R}^d)$ if and only if the underlying GRT is injective. The proof of this result is given in Section 5.6.

Proposition 5.9. *Let $p \in [1, +\infty)$. The Generalized Sliced-Wasserstein (or maximum Generalized Sliced-Wasserstein) distance of order p based on the defining function g satisfies all metric axioms if and only if the generalized Radon transform associated to g is injective.*

According to Proposition 5.9, the injectivity of GRT is sufficient and necessary for GSW to be a metric. In this respect, our result brings a different perspective on [Bonnotte, 2013, Proposition 5.1.2]: since the standard Radon transform is injective, SW is indeed a distance.

Remark 5.10. *If the chosen generalized Radon transform is not injective, then we can only say that the resulting GSW and max-GSW are pseudo-metrics: they still satisfy non-negativity, symmetry, the triangle inequality, and* $\mathbf{GSW}_p(\mu, \mu) = 0$, *max-*$\mathbf{GSW}_p(\mu, \mu) = 0$.

5.3.2 Injectivity of the generalized Radon transform

We have shown in Proposition 5.9 that the injectivity of the GRT is crucial for the resulting GSW and max-GSW to be distances between probability measures. We now enumerate some of the known defining functions that lead to injective GRTs.

The investigation of the sufficient and necessary conditions guaranteeing the injectivity of GRTs is a long-standing topic [Beylkin, 1984, Homan and Zhou, 2017, Uhlmann, 2003, Ehrenpreis, 2003]. The *circular defining function*, supported on $\mathbb{R}^d \times \mathbb{S}^{d-1}$ and given by

$$g(x, \theta) = \|x - r\theta\| \tag{5.16}$$

with $r \in \mathbb{R}+$, provides an injective GRT [Kuchment, 2006]. *Homogeneous polynomials with an odd degree* also yield an injective GRT [Rouvière, 2015], and are defined as

$$g(x, \theta) = \sum_{|\alpha|=m} \theta_\alpha x^\alpha \,, \tag{5.17}$$

where we use the multi-index notation $\alpha = (\alpha_1, \ldots, \alpha_{d_\alpha}) \in \mathbb{N}^{d_\alpha}$, $|\alpha| = \sum_{i=1}^{d_\alpha} \alpha_i$, and $x^\alpha = \prod_{i=1}^{d_\alpha} x_i^{\alpha_i}$. The summation in (5.17) iterates over all possible multi-indices α, such that $|\alpha| = m$, where m denotes the degree of the polynomial and $\theta_\alpha \in \mathbb{R}$. The parameter set for homogeneous polynomials is then set to $\Omega_\theta = \mathbb{S}^{d_\alpha - 1}$. We can observe that choosing $m = 1$ reduces to the linear case $g(x, \theta) = \langle x, \theta \rangle$, since the set of multi-indices with $|\alpha| = 1$ becomes

$$\{(\alpha_1, \ldots, \alpha_d) \,:\, \alpha_i = 1 \text{ for a single } i \in \mathbb{N}^*, 1 \leq i \leq d, \text{ and } \alpha_j = 0, \quad \forall j \neq i\} \,,$$

and contains d elements.

Algorithm 3: Approximation of GSW

Input: Two sets of observations $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, number of projection directions L, order p, defining function g supported on $\mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\})$.

GSW $= 0$

for $l = 1, \dots, L$ **do**

  Sample: $\theta_l \sim \boldsymbol{\sigma}_q$

  for $i = 1, \dots, n$ **do**

    Project: $x'_i = g(x_i, \theta_l)$, $y'_i = g(y_i, \theta_l)$

  Sort: $x'_{(1)} \leq x'_{(2)} \leq \cdots \leq x'_{(n)}$, $y'_{(1)} \leq y'_{(2)} \leq \cdots \leq y'_{(n)}$

  GSW $=$ GSW $+ (1/n) \sum_{i=1}^n |x'_{(i)} - y'_{(i)}|^p$

GSW $= (\text{GSW}/L)^{1/p}$

return GSW

While the polynomial defining functions form an interesting alternative to linear projections, their memory complexity d_α grows exponentially with the dimension of the data and the degree of the polynomial, hence deteriorates their potential in modern machine learning problems. As a remedy, inspired by the current success of neural networks, a natural task in our context would be to come up with a neural network, which would yield a valid GSW or max-GSW, when used as the defining function in the GRT.

As a neural network-based defining function, we propose a multi-layer fully connected network with *leaky ReLU* activations. Under this specific network architecture, one can easily show that the corresponding defining function satisfies **(D1)** to **(D4)** on $(\mathsf{X}\backslash\{\mathbf{0}\}) \times (\mathbb{R}^q\backslash\{\mathbf{0}\})$, where $\mathsf{X} \subset \mathbb{R}^d$ and $q \in \mathbb{N}^*$ is the number of parameters of the network. On the other hand, proving the injectivity of the associated GRT is highly non-trivial, so the GSW associated with this particular defining function is a pseudo-metric, as we discussed in Remark 5.10. However, as illustrated in Section 5.4, this neural network-based defining function still performs well in practice, and the non-differentiability of the leaky ReLU function at $\mathbf{0}$ does not seem to be a big issue in practice.

With a neural network as the defining function, minimizing max-GSW between two distributions is analogical to adversarial learning, where the adversary network's goal is to distinguish the two distributions. In the max-GSW case, the adversary network, *i.e.* the defining function, seeks optimal parameters that maximize the GSW distance between the input distributions.

5.4 Numerical Implementation and Experiments

We conduct several experiments to compare the empirical performance of GSW and max-SW for different choices of defining functions. To this end, we first explain how we compute GSW and max-GSW in practice.

5.4.1 Implementation of generalized Sliced-Wasserstein distances

Consider $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, which we wish to compare using GSW. In most machine learning applications, we do not have access to the distributions, but to two sets of $n \in \mathbb{N}^*$ i.i.d. samples from μ and ν, which are respectively denoted by $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^n$. This implies that one can only compute $\mathbf{GSW}_p(\hat{\mu}_n, \hat{\nu}_n)$, which is an approximation

Algorithm 4: Approximation of max-GSW

Input: Two sets of observations $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, number of projection directions L, order p, defining function g, maximum number of iterations T, step size ρ.

Randomly initialize $\theta_0 \in \Omega_\theta$
for $t = 0, \dots, T-1$ **do**
  for $i = 1, \dots, n$ **do**
    Project: $x'_i = g(x_i, \theta_t)$, $y'_i = g(y_i, \theta_t)$
  Sort: $x'_{(1)} \leq x'_{(2)} \leq \cdots \leq x'_{(n)}$, $y'_{(1)} \leq y'_{(2)} \leq \cdots \leq y'_{(n)}$
  $\mathrm{W} = (1/n) \sum_{i=1}^n |x'_{(i)} - y'_{(i)}|^p$
  Perform one gradient ascent step: $\theta' = \theta_l + \rho \nabla_\theta \mathrm{W}$
  Project θ' on Ω_θ: $\theta_{l+1} = \mathrm{proj}_{\Omega_\theta}(\theta')$
for $i = 1, \dots, n$ **do**
  Project: $x'_i = g(x_i, \theta_T)$, $y'_i = g(y_i, \theta_T)$
Sort: $x'_{(1)} \leq x'_{(2)} \leq \cdots \leq x'_{(n)}$, $y'_{(1)} \leq y'_{(2)} \leq \cdots \leq y'_{(n)}$
$\mathrm{mGSW} = (1/n) \sum_{i=1}^n |x'_{(i)} - y'_{(i)}|^p$
$\mathrm{mGSW} = (\mathrm{mGSW})^{1/p}$
return mGSW

of $\mathbf{GSW}_p(\mu, \nu)$. Besides, the integral in (5.14) is generally intractable, hence will be estimated.

Similarly to SW, we will use a simple Monte Carlo scheme and the analytical expression of the Wasserstein distance between univariate distributions (2.15) to compute the following estimate of $\mathbf{GSW}_p(\hat{\mu}_n, \hat{\nu}_n)$,

$$\widehat{\mathbf{GSW}}_p(\hat{\mu}_n, \hat{\nu}_n) = \left(\frac{1}{Ln} \sum_{l=1}^{L} \sum_{i=1}^{n} |g(x_{(i)}, \theta_l) - g(y_{(i)}, \theta_l)|^p \right)^{1/p}, \tag{5.18}$$

where for any sequence of vectors $\{z_i\}_{i=1}^n$ and $\theta \in \mathbb{R}^q \backslash \{\mathbf{0}\}$, $\{g(z_{(i)}, \theta)\}_{i=1}^n$ is the sorted sequence of projections, *i.e.* $g(z_{(1)}, \theta) \leq g(z_{(2)}, \theta) \leq \cdots \leq g(z_{(n)}, \theta)$. The procedure to approximate GSW is summarized in Algorithm 3.

To compute max–$\mathbf{GSW}_p(\hat{\mu}_n, \hat{\nu}_n)$, we employ a numerical optimization method similar to the expectation-maximization (EM) algorithm, which repeats the following: (a) given $\theta \in \Omega_\theta$, $\{g(x_i, \theta)\}_{i=1}^n$ and $\{g(y_i, \theta)\}_{i=1}^n$ are sorted to compute the one-dimensional Wasserstein distance, once again according to (2.15), (b) θ is updated with a projected gradient ascent step. Once the convergence is reached (for example, by setting a maximum number of iterations), the algorithm returns an nearly-optimal projection direction θ^*, which is then used to approximate max–$\mathbf{GSW}_p(\hat{\mu}_n, \hat{\nu}_n)$ as follows

$$\text{max–}\widehat{\mathbf{GSW}}_p(\hat{\mu}_n, \hat{\nu}_n) = \frac{1}{n} \sum_{i=1}^{n} |g(x_{(i)}, \theta^*) - g(y_{(i)}, \theta^*)|^p \,.$$

The whole procedure is summarized in Algorithm 4. Note that the gradient with respect to θ is computed via automatic differentiation, and the gradient ascent can be replaced with any iterative optimization method, such as Adam [Kingma and Ba, 2015].

Our EM-like method finds the optimal θ by optimizing the actual Wasserstein distance between the projected distributions, as opposed to the heuristic approaches pro-

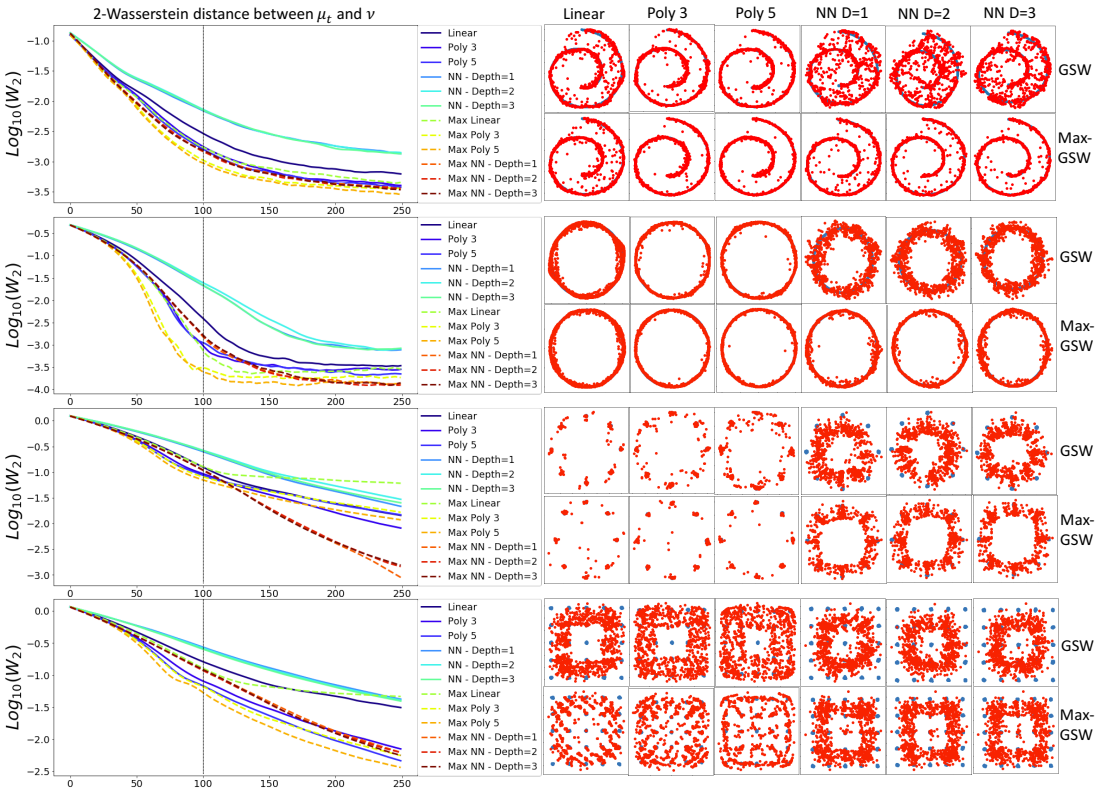


Figure 5.2: Evaluation of the performance of GSW on the flows experiment when using different defining functions and four synthetic datasets as the target. "Linear" and "Max Linear" refer to SW and max-SW respectively.

posed in [Deshpande et al., 2018, Kolouri et al., 2018], where the pseudo-optimal slice is found with perceptrons or penalized linear discriminant analysis [Wang et al., 2011].

5.4.2 Experiments

Now that we have clarified the numerical implementation of (max-)GSW, we present the experiments that we conducted to evaluate the performance of our metrics in generative modeling applications. Our empirical results can be reproduced with our open source code[1].

To study the effects of the defining function on the practical performance of GSW and max-GSW, we consider the following flows problem

$$\min_{\mu} \mathbf{D}(\mu, \nu) \ , \tag{5.19}$$

where $\mathbf{D}$ denotes an instance of (max-)GSW, ν is a target distribution and μ is the source distribution. The solution of (5.19) is approximated using the following iterative optimization scheme: first, μ is initialized as the empirical distribution associated to i.i.d. observations from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$; then, these observations are updated by performing gradient descent on $\mathbf{D}(\mu, \nu)$.

The target ν corresponds to the distribution of i.i.d. samples from one of these four well-known distributions: "25-Gaussians", "8-Gaussians", "Swiss Roll" and "Circle". We

[1] See our GitHub repository: https://github.com/kimiandj/gsw

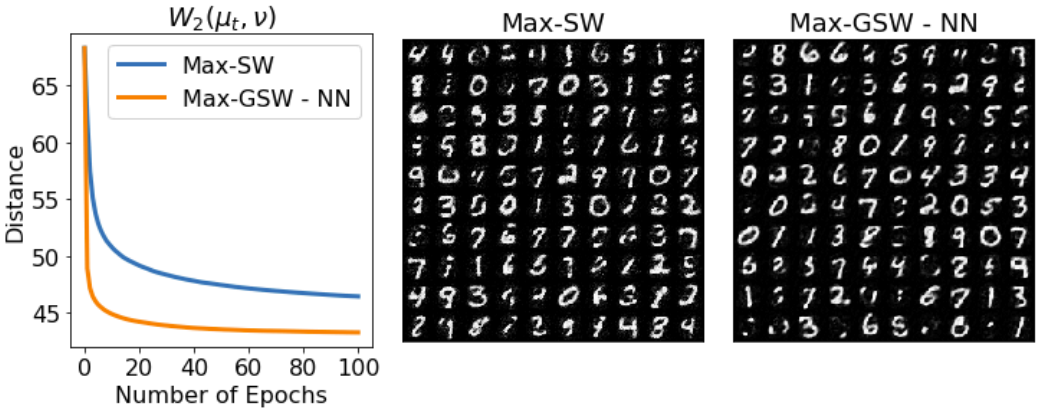


Figure 5.3: Comparison of max-GSW and max-SW on the flows experiment applied to the MNIST dataset.

compare different instances of GSW and max-GSW, characterized by different choices of defining functions, namely "linear" (in that case, GSW and max-GSW are equivalent to SW and max-SW respectively), homogeneous polynomials of degree 3 and 5, and the neural networks described in Section 5.3.2 with 1 to 3 hidden layers. We used the exact same optimization scheme for all instances, and kept only $L = 1$ projection when approximating GSW with (5.18). In order to easily compare the results produced by the different flows, we computed the Wasserstein distance of order 2 between μ_t and ν (by solving the corresponding linear program), where μ_t denotes the source distribution at iteration t of the optimization procedure. We repeated each experiment 100 times and report the average Wasserstein distance (computed with the POT implementation [Flamary et al., 2021]) for all target datasets in Figure 5.2. We also show in that same figure a snapshot of μ_{100} and ν for all datasets.

We observe that (i) max-GSW outperforms GSW, of course at the cost of an additional optimization, and (ii) while the choice of the defining function $g(\cdot, \theta)$ is data-dependent, one can see that the homogeneous polynomials are often among the top performers for all datasets. Specifically, SW is always outperformed by GSW with polynomial projections ("Poly 3" and "Poly 5" in Figure 5.2) and by all the variants of max-GSW. Besides, max-SW is consistently outperformed by max-GSW based on neural networks. The only variant of GSW that is outperformed by SW is GSW with a neural network-based defining function, which was expected because of the inherent complexity of approximating the integral over a very large domain (5.14) with a simple Monte Carlo average. Similarly to max-SW, max-GSW replaces sampling with optimization to circumvent this issue.

We then move to more realistic datasets, by running the same experiment for the MNIST dataset [LeCun and Cortes, 2010]: we solve (5.19), where μ is initialized to the distribution of 100 random images of dimension 784, and ν is associated to the training set of MNIST. Given the high-dimensional nature of the problem, we cannot use the homogeneous polynomials due to memory constraints caused by the combinatorial growth of the coefficients, as discussed in Section 5.3.2. Therefore, we only compare max-SW against max-GSW whose defining function is a 3-layer neural network. We report the Wasserstein distance of order 2 between each μ_t (the 100 images) and ν (the training set of MNIST) in Figure 5.3, where t is the number of training epochs. We observe that

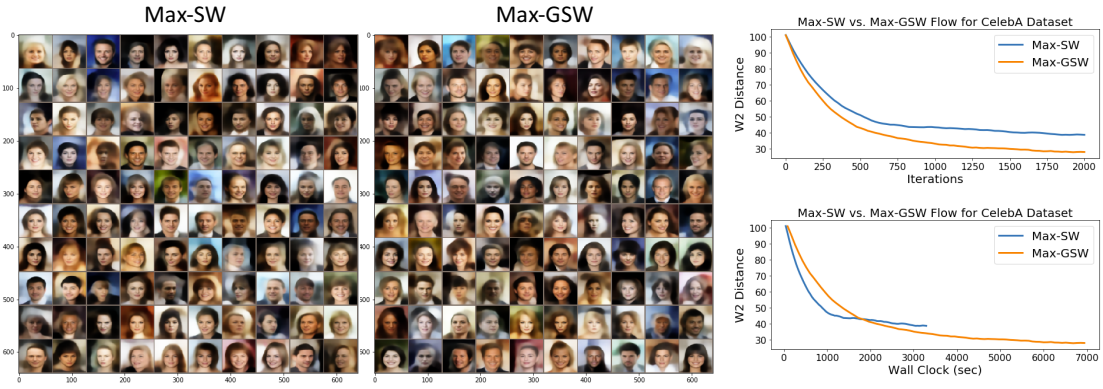


Figure 5.4: Comparison of max-GSW and max-SW on the flows experiment applied to the CelebA dataset.

with the proposed generalized approach, the error is decreasing significantly faster when compared to max-SW. We also show the generated images and observe that max-GSW produces "crisper" images than max-SW.

Finally, we considered a larger dataset, namely CelebA [Liu et al., 2015]. Since the dimension is very large, we ran a pre-trained auto-encoder to find a latent space of dimension 256 for the dataset, then solved the flows problem on this lower-dimensional space. We compared max-SW against max-GSW based on a 3-layer neural network, by measuring the Wasserstein distance of order 2 between the target and optimized source distributions in the latent space. Figure 5.4 shows the results of this experiment: max-GSW finds a better solution than max-SW in fewer iterations, but each iteration takes more time because of the neural network training. We also report the generated images in Figure 5.4 and observe that the quality of the images produced by max-GSW is slightly better. Hence, although max-GSW seems like an interesting alternative to max-SW as it produces better results, the practitionner should also be careful about the fact that it has important computational implications and might execute more slowly.

5.5 Conclusion

We introduced a novel family of probability divergences, which extends the concept behind the Sliced-Wasserstein and maximum Sliced-Wasserstein distances: while SW and max-SW measure the dissimilarity between two distributions by comparing their projections on hyperplanes, we propose to compare projections on hypersurfaces instead. The resulting divergences, called the Generalized Sliced-Wasserstein and maximum Sliced-Wasserstein distances, are characterized through a general version of the Radon transform, whose standard version was originally used to define SW. We proved that GSW and max-GSW satisfy all metric axioms if and only if the generalized Radon transform they are based on is injective. We then explained how to implement GSW and max-GSW between any two empirical measures, and demonstrated the superior performance of our generalized divergences over SW in several generative modeling applications.

5.6 Appendix: Proof of Proposition 5.9

The proof of Proposition 5.9 consists in verifying that GSW and max-GSW satisfy all the metric axioms and relies on the fact that $\mathbf{W}_p$ is a metric.

Proof of Proposition 5.9. Let $p \in [1, +\infty)$. Consider a defining function $g : \mathsf{X} \times (\mathbb{R}^q \backslash \{\mathbf{0}\}) \to \mathbb{R}$, with $\mathsf{X} \subset \mathbb{R}^d$ and $q \in \mathbb{N}^*$. We denote by Ω_θ a compact set of feasible parameters for $g(\cdot, \theta)$.

Non-negativity. Since the Wasserstein distance is a distance and is thus non-negative, we can easily prove that GSW and max-GSW distances satisfy non-negativity as well. Indeed, consider any defining function g. By using the definition of GSW and max-GSW, and the fact that $\mathbf{W}_p(\mu', \nu') \geq 0$ for any two probability distributions μ', ν', we obtain for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$\begin{aligned}\mathbf{GSW}_p(\mu, \nu) &= \left(\int_{\mathbb{R}^q \backslash \{\mathbf{0}\}} \mathbf{W}_p^p\big(\mathcal{G}\mu(\cdot, \theta; g), \mathcal{G}\nu(\cdot, \theta; g)\big) \mathrm{d}\boldsymbol{\sigma}_q(\theta) \right)^{\frac{1}{p}} \\ &\geq \left(\int_{\mathbb{R}^q \backslash \{\mathbf{0}\}} (0)^p \mathrm{d}\boldsymbol{\sigma}_q(\theta) \right)^{\frac{1}{p}} = 0 \ ,\end{aligned}$$

$$\begin{aligned}\text{and,} \quad \max\text{–}\mathbf{GSW}_p(\mu, \nu) &= \max_{\theta \in \Omega_\theta} \mathbf{W}_p\big(\mathcal{G}\mu(\cdot, \theta; g), \mathcal{G}\nu(\cdot, \theta; g)\big) \\ &= \mathbf{W}_p\big(\mathcal{G}\mu(\cdot, \theta^*; g), \mathcal{G}\nu(\cdot, \theta^*; g)\big) \\ &\geq 0 \ ,\end{aligned}$$

where $\theta^* = \mathrm{argmax}_{\theta \in \Omega_\theta} \mathbf{W}_p(\mathcal{G}\mu(\cdot, \theta; g), \mathcal{G}\nu(\cdot, \theta; g))$.

Symmetry. Since the Wasserstein distance is symmetric, we have for any two distributions μ', ν', $\mathbf{W}_p(\mu', \nu') = \mathbf{W}_p(\nu', \mu')$. In particular, we can write for all $\theta \in \mathbb{R}^q \backslash \{\mathbf{0}\}$,

$$\mathbf{W}_p(\mathcal{G}\mu(\cdot, \theta; g), \mathcal{G}\nu(\cdot, \theta; g)) = \mathbf{W}_p(\mathcal{G}\nu(\cdot, \theta; g), \mathcal{G}\mu(\cdot, \theta; g)) \ , \tag{5.20}$$

$$\text{and} \quad \max_{\theta \in \Omega_\theta} \mathbf{W}_p(\mathcal{G}\mu(\cdot, \theta; g), \mathcal{G}\nu(\cdot, \theta; g)) = \max_{\theta \in \Omega_\theta} \mathbf{W}_p(\mathcal{G}\nu(\cdot, \theta; g), \mathcal{G}\mu(\cdot, \theta; g)) \ . \tag{5.21}$$

The symmetry of GSW and max-GSW directly follows from (5.20) and (5.21) respectively.

Triange inequality. We now prove that GSW and max-GSW satisfy the triangle inequality. Let μ_1, μ_2 and μ_3 in $\mathcal{P}_p(\mathbb{R}^d)$. Since the Wasserstein distance satisfies the triangle inequality, the following holds for any $\theta \in \mathbb{R}^q \backslash \{\mathbf{0}\}$.

$$\begin{aligned}\mathbf{W}_p(\mathcal{G}\mu_1(\cdot, \theta; g), \mathcal{G}\mu_3(\cdot, \theta; g)) \leq \mathbf{W}_p(\mathcal{G}\mu_1(\cdot, \theta; g), \mathcal{G}\mu_2(\cdot, \theta; g)) \\ + \mathbf{W}_p(\mathcal{G}\mu_2(\cdot, \theta; g), \mathcal{G}\mu_3(\cdot, \theta; g)) \ .\end{aligned}$$

Therefore, we can write

$$
\begin{aligned}
&\mathbf{GSW}_p(\mu_1,\mu_3)\\
&= \left(\int_{\mathbb{R}^q\setminus\{\mathbf{0}\}} \mathbf{W}_p^p(\mathcal{G}\mu_1(\cdot,\theta;g),\mathcal{G}\mu_3(\cdot,\theta;g))\mathrm{d}\boldsymbol{\sigma}_q(\theta)\right)^{\frac{1}{p}}\\
&\leq \left(\int_{\mathbb{R}^q\setminus\{\mathbf{0}\}} \left\{\mathbf{W}_p(\mathcal{G}\mu_1(\cdot,\theta;g),\mathcal{G}\mu_2(\cdot,\theta;g)) + \mathbf{W}_p(\mathcal{G}\mu_2(\cdot,\theta;g),\mathcal{G}\mu_3(\cdot,\theta;g))\right\}^p \mathrm{d}\boldsymbol{\sigma}_q(\theta)\right)^{\frac{1}{p}}\\
&\leq \Big(\int_{\mathbb{R}^q\setminus\{\mathbf{0}\}} \mathbf{W}_p^p(\mathcal{G}\mu_1(\cdot,\theta;g),\mathcal{G}\mu_2(\cdot,\theta;g))\mathrm{d}\boldsymbol{\sigma}_q(\theta)\Big)^{\frac{1}{p}}\\
&\qquad\qquad + \Big(\int_{\mathbb{R}^q\setminus\{\mathbf{0}\}} \mathbf{W}_p^p(\mathcal{G}\mu_2(\cdot,\theta;g),\mathcal{G}\mu_3(\cdot,\theta;g))\mathrm{d}\boldsymbol{\sigma}_q(\theta)\Big)^{\frac{1}{p}} \qquad (5.22)
\end{aligned}
$$

where (5.22) follows from the application of the Minkowski inequality in $\mathcal{L}^p(\mathbb{R}^q\setminus\{\mathbf{0}\},\boldsymbol{\sigma}_q)$. We conclude that GSW satisfies the triangle inequality.

Now, denote by $\theta^* = \mathrm{argmax}_{\theta\in\Omega_\theta}\mathbf{W}_p(\mathcal{G}\mu_1(\cdot,\theta;g),\mathcal{G}\mu_3(\cdot,\theta;g))$; then,

$$
\begin{aligned}
&\text{max–}\mathbf{GSW}_p(\mu_1,\mu_3)\\
&= \max_{\theta\in\Omega_\theta}\ \mathbf{W}_p(\mathcal{G}\mu_1(\cdot,\theta;g),\mathcal{G}\mu_3(\cdot,\theta;g))\\
&= \mathbf{W}_p(\mathcal{G}\mu_1(\cdot,\theta^*;g),\mathcal{G}\mu_3(\cdot,\theta^*;g))\\
&\leq \mathbf{W}_p(\mathcal{G}\mu_1(\cdot,\theta^*;g),\mathcal{G}\mu_2(\cdot,\theta^*;g)) + \mathbf{W}_p(\mathcal{G}\mu_2(\cdot,\theta^*;g),\mathcal{G}\mu_3(\cdot,\theta^*;g))\\
&\leq \max_{\theta\in\Omega_\theta}\mathbf{W}_p(\mathcal{G}\mu_1(\cdot,\theta;g),\mathcal{G}\mu_2(\cdot,\theta;g)) + \max_{\theta\in\Omega_\theta}\mathbf{W}_p(\mathcal{G}\mu_2(\cdot,\theta;g),\mathcal{G}\mu_3(\cdot,\theta;g))\\
&\leq \text{max–}\mathbf{GSW}_p(\mu_1,\mu_2) + \text{max–}\mathbf{GSW}_p(\mu_2,\mu_3)\ ,
\end{aligned}
$$

hence, max-GSW also satisfies the triangle inequality.

Identity of indiscernibles. Since for any distribution μ', $\mathbf{W}_p(\mu',\mu') = 0$, then by definition of GSW and max-GSW, for any $\mu\in\mathcal{P}_p(\mathbb{R}^d)$,

$$
\mathbf{GSW}_p(\mu,\mu) = 0, \quad \text{and} \quad \text{max–}\mathbf{GSW}_p(\mu,\mu) = 0\ . \qquad (5.23)
$$

Now, assume for $\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d)$, $\mathbf{GSW}_p(\mu,\nu) = 0$ and max–$\mathbf{GSW}_p(\mu,\nu) = 0$. Both statements are equivalent to $\mathcal{G}\mu(\cdot,\theta;g) = \mathcal{G}\nu(\cdot,\theta;g)$ for almost all $\theta\in\mathbb{R}^q\setminus\{\mathbf{0}\}$. Therefore, GSW and max-GSW satisfy the identity of indiscernibles if and only if $\mathcal{G}\mu(\cdot,\theta;g) = \mathcal{G}\nu(\cdot,\theta;g)$ implies $\mu=\nu$, *i.e.* the GRT is injective.

□

Chapter 6

Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections

This chapter is based on [Nadjahi et al., 2021].

The Sliced-Wasserstein distance is being increasingly used in machine learning applications as an alternative to the Wasserstein distance and offers significant computational and statistical benefits. Since it is defined as an expectation over random projections, SW is commonly approximated by Monte Carlo. We adopt a new perspective to approximate SW by making use of the *concentration of measure phenomenon*: under mild assumptions, one-dimensional projections of a high-dimensional random vector are approximately Gaussian. Based on this observation, we develop a simple deterministic approximation for SW. Our method does not require sampling a number of random projections, and is therefore both accurate and easy to use compared to the usual Monte Carlo approximation. We derive nonasymptotical guarantees for our approach, and show that the approximation error goes to zero as the dimension increases, under a weak dependence condition on the data distribution. We validate our theoretical findings on synthetic datasets, and illustrate the proposed approximation on a generative modeling problem.

6.1 Introduction

The Sliced-Wasserstein distance is a practical alternative optimal transport metric, as it exploits the analytical form of the Wasserstein distance between univariate distributions. As a reminder, its definition, which is formally given in Definition 2.9, reads as follows: consider two random variables X and Y in $\mathbb{R}^d$ with respective distributions μ and ν, and denote by $\theta^\star_\sharp\mu, \theta^\star_\sharp\nu$ the univariate distributions of the projections of X, Y along $\theta \in \mathbb{R}^d$; SW then compares μ and ν by computing $\mathbb{E}[\mathbf{W}_p^p(\theta^\star_\sharp\mu, \theta^\star_\sharp\nu)]$, where the expectation $\mathbb{E}$ is taken with respect to θ uniformly distributed on the unit sphere, and $\mathbf{W}_p$ is the Wasserstein distance of order $p \geq 1$ (Definition 2.6).

In practice, this expectation is typically estimated by Monte Carlo: one uniformly draws L projection directions $\{\theta_l\}_{l=1}^L$ and approximates SW with $L^{-1}\sum_{l=1}^L \mathbf{W}_p^p\big(\theta^\star_{l\sharp}\mu, \theta^\star_{l\sharp}\nu\big)$.

Since the Wasserstein distance between univariate distributions can easily be computed in closed form, this scheme leads to significant computational benefits as compared to the Wasserstein distance, provided that L is not chosen too large. As we discussed in previous chapters, SW has been successfully applied in several practical tasks, and has been shown to offer nice theoretical properties as well. For instance, even though the sample complexity of Wasserstein grows exponentially with the data dimension (Section 2.4.2), the sample complexity of SW does not depend on the dimension [Nadjahi et al., 2020b]. This latter study, which will be presented in detail in Chapter 7, also demonstrated with a theoretical error bound that the quality of the Monte Carlo estimate of SW depends on the number of projections and the variance of the one-dimensional Wasserstein distances (Theorem 7.9). In other words, to ensure that the induced approximation error is reasonably small, one might need to choose a large value for L, which inevitably increases the computational complexity of SW. Alternative approaches have been proposed to overcome this issue, and mainly consist in picking more "informative" projection directions: e.g., SW based on orthogonal projections [Wu et al., 2019, Meng et al., 2019], maximum SW [Deshpande et al., 2019], generalized SW distances (Chapter 5) and distributional SW distances [Nguyen et al., 2021].

In this chapter, we adopt a different perspective and leverage *concentration results on random projections* to approximate SW: previous work showed that, under relatively mild conditions, the typical distribution of low-dimensional projections of high-dimensional random variables is close to some Gaussian law [Sudakov, 1978, Diaconis and Freedman, 1984]. Recently, this phenomenon has been illustrated with a bound in terms of the Wasserstein distance [Reeves, 2017]: let $\{X_i\}_{i=1}^d$ be a sequence of real random variables with distribution μ_d, such that $X_1, \dots, X_d$ are independent with finite fourth-order moments; then, $\mathbb{E}[\mathbf{W}_2^2(\theta_\sharp^\star \mu_d, \mathcal{N}_{\mu_d})^2]$ goes to zero as d increases, where $\mathcal{N}_{\mu_d}$ is a univariate Gaussian distribution whose variance depends on μ_d and the expectation is taken with respect to a Gaussian variable θ. This result has very recently been used to bound the "maximum-sliced distance" between any probability measure and its Gaussian approximation [Goldt et al., 2021]. In our work, we use it to design a novel technique that estimates SW with a simple *deterministic* formula. As opposed to Monte Carlo, our method does not depend on a finite set of random projections, therefore it eliminates the need of tuning the hyperparameter L and can lead to a significant computational time reduction. Besides, our proposal is quite different from the aforementioned variants of SW which consist in selecting "informative" projection directions: these alternatives are defined as optimization problems whose resolution is challenging (e.g., [Nguyen et al., 2021, Section 3.2]) and are then computed by finding an approximate solution. This incurs an additional computational cost and estimation error, while our method directly approximates SW (thus, does not define an alternative distance) via simple deterministic operations, does not rely on any hyperparameters, and comes with theoretical guarantees on its induced error.

The important steps to formulate our approximate SW are summarized as follows. We first define an alternative SW whose projection directions are drawn from the same Gaussian distribution as in [Reeves, 2017], instead of uniformly on $\mathbb{S}^{d-1}$, and establish its relation with the original SW. By combining this property with [Reeves, 2017, Theorem 1], we bound the absolute difference between SW applied to any two probability measures μ_d, ν_d on $\mathbb{R}^d$, and the Wasserstein distance between the univariate Gaussians $\mathcal{N}_{\mu_d}$, $\mathcal{N}_{\nu_d}$. Then, we explain why the mean parameters of μ_d and ν_d should be zero for the approximation error to decrease as d grows. Nevertheless, we show that it is not a limiting factor, by exploiting the following decomposition of SW: SW between μ_d, ν_d can

be equivalently written as the sum of the *difference between their means* and the SW between the *centered versions* of μ_d, ν_d.

Our approach then consists in estimating SW between the centered versions with the Wasserstein term between Gaussian approximations to meet the zero-means condition, and recover SW between the original measures via the aforementioned property. Since the Wasserstein distance between Gaussian distributions admits a closed-form solution, our approximate SW is very easy to compute, and faster than the Monte Carlo estimate obtained with a large number of projections. We derive nonasymptotical guarantees on the error induced by our approach. Specifically, we define a weak dependence condition under which the error is shown to go to zero with increasing d. Our theoretical results are then validated with experiments conducted on synthetic data. Finally, we leverage our theoretical insights to design a novel adversarial framework for a typical generative modeling problem in machine learning, and illustrate its advantages in terms of accuracy and computational time, over generative models based on the Monte Carlo estimate of SW. Our empirical results can be reproduced with our open source code[1].

6.2 Background on Central Limit Theorems for Random Projections

There is a rich literature on the typical behavior of one-dimensional random projections of high-dimensional vectors. To be more specific, let $(\theta_i)_{i\in\mathbb{N}^*}$ be i.i.d. standard one-dimensional Gaussian random variables and $(X_i)_{i\in\mathbb{N}^*}$ be a sequence of one-dimensional random variables. Denote for any $d \in \mathbb{N}^*$, $\theta_{1:d} = \{\theta_i\}_{i=1}^d$ and $X_{1:d} = \{X_i\}_{i=1}^d$. Several central limits theorems ensure that, under relatively mild conditions, the sequence of distributions of $d^{-1/2}\langle\theta_{1:d}, X_{1:d}\rangle \in \mathbb{R}$ given $\theta_{1:d} \in \mathbb{R}^d$ converges in distribution to a Gaussian random variable in probability. This line of work goes back to [Sudakov, 1978, Diaconis and Freedman, 1984], whose contributions has then been sharpened and generalized in [Hall and Li, 1993, von Weizsäcker, 1997, Anttila et al., 2003, Bobkov, 2003, Klartag, 2007, Meckes, 2010, Dümbgen and Del Conte-Zerial, 2013, Leeb, 2013].

In particular, a recent study [Reeves, 2017] gives a quantitative version of this phenomenon. More precisely, denote for any $d \in \mathbb{N}^*$ by μ_d^X, the distribution of $X_{1:d}$ (*i.e.*, the joint distribution of $X_1, X_2, \ldots, X_d$), and $\boldsymbol{\gamma}_d$ the zero-mean Gaussian distribution with covariance matrix $(1/d)\mathbf{I}_d$. Assume that for any $d \in \mathbb{N}^*$, $\mu_d^X \in \mathcal{P}_2(\mathbb{R}^d)$. Then, [Reeves, 2017, Theorem 1] shows that there exists a universal constant $C \geq 0$ such that

$$\int_{\mathbb{R}^d} \mathbf{W}_2^2\left(\theta_\sharp^\star \mu_d^X, \mathcal{N}\left(0, d^{-1}\mathtt{m}_2(\mu_d^X)\right)\right) \mathrm{d}\boldsymbol{\gamma}_d(\theta) \leq C\Xi_d(\mu_d^X)\,, \quad \text{with} \tag{6.1}$$

$$\Xi_d(\mu_d^X) = d^{-1}\{\alpha(\mu_d^X) + \left(\mathtt{m}_2(\mu_d^X)\beta_1(\mu_d^X)\right)^{1/2} + \mathtt{m}_2(\mu_d^X)^{1/5}\beta_2(\mu_d^X)^{4/5}\}\,, \tag{6.2}$$

$$\mathtt{m}_2(\mu_d^X) = \mathbb{E}\left[\|X_{1:d}\|^2\right]\,, \tag{6.3}$$

$$\alpha(\mu_d^X) = \mathbb{E}\left[\left|\,\|X_{1:d}\|^2 - \mathtt{m}_2(\mu_d^X)\right|\right]\,, \tag{6.4}$$

$$\beta_q(\mu_d^X) = \mathbb{E}^{\frac{1}{q}}[\left|\langle X_{1:d}, X'_{1:d}\rangle\right|^q]\,, \tag{6.5}$$

where $q \in \{1, 2\}$ and $(X'_i)_{i\in\mathbb{N}^*}$ is an independent copy of $(X_i)_{i\in\mathbb{N}^*}$. A formal statement of this result is also given for completeness in Section 6.6.1.

[1] See our GitHub repository: https://github.com/kimiandj/fast_sw

6.3 Approximate Sliced-Wasserstein Distance Based on Concentration of Random Projections

We develop a novel method to approximate the Sliced-Wasserstein distance of order 2, by extending the bound in (6.1) and deriving novel properties for SW. We then derive nonasymptotical guarantees of the corresponding approximation error, which ensure that our estimate is accurate for high-dimensional data under a weak dependence condition.

6.3.1 Sliced-Wasserstein distance with Gaussian projections

First, to enable the use of (6.1) for the analysis of SW, we introduce a variant of $\mathbf{SW}_p$ (Definition 2.9) whose projections are drawn from the Gaussian distribution considered in (6.1), instead of uniformly on the sphere. The Sliced-Wasserstein distance of order $p \in [1, +\infty)$ based on Gaussian projections is defined for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ as

$$\widetilde{\mathbf{SW}}_p^p(\mu, \nu) = \int_{\mathbb{R}^d} \mathbf{W}_p^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu) \mathrm{d}\boldsymbol{\gamma}_d(\theta) \; . \tag{6.6}$$

In the next proposition, we establish a simple mathematical relation between traditional SW and the newly introduced one: we prove that $\widetilde{\mathbf{SW}}_p$ is equal to $\mathbf{SW}_p$ up to a proportionality constant that only depends on the data dimension d and the order p.

Proposition 6.1. *Let $p \in [1, +\infty)$. Then, $\widetilde{\mathbf{SW}}_p$ (6.6) is related to $\mathbf{SW}_p$ (2.20) as follows: for any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,*

$$\widetilde{\mathbf{SW}}_p(\mu, \nu) = \left(\frac{2}{d}\right)^{1/2} \left\{\frac{\Gamma(d/2 + p/2)}{\Gamma(d/2)}\right\}^{1/p} \mathbf{SW}_p(\mu, \nu) \; ,$$

where Γ is the Gamma function.

Since (6.1) only applies to the Wasserstein distance of order 2, we will focus on SW of that same order in the rest of the chapter. In this case, SW with Gaussian projections is *equal* to the original SW. Indeed, we can show that the constant $(2/d)^{1/2} \{\Gamma(d/2 + p/2) \; / \; \Gamma(d/2)\}^{1/p}$ defined in Proposition 6.1 is equal to 1 when $p = 2$, by using the property $\Gamma(d/2 + 1) = (d/2)\Gamma(d/2)$.

6.3.2 Approximate Sliced-Wasserstein distance

Our next result is an easy consequence of (6.1) and Proposition 6.1, and shows that, for any $\mu_d, \nu_d \in \mathcal{P}_2(\mathbb{R}^d)$, the difference between $\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}$ and $\mathbf{SW}_2(\mu_d, \nu_d)$ in absolute value is bounded from above by $\Xi_d(\mu_d) + \Xi_d(\nu_d)$ (6.2).

Theorem 6.2. *There exists a universal constant $C > 0$ such that for any $\mu_d, \nu_d \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\left|\mathbf{SW}_2(\mu_d, \nu_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}\right| \leq C\big(\Xi_d(\mu_d) + \Xi_d(\nu_d)\big)^{1/2} \; , \tag{6.7}$$

where, for $\xi_d \in \{\mu_d, \nu_d\}$, $\Xi_d(\xi_d)$ and $\mathtt{m}_2(\xi_d)$ are defined in (6.2) and (6.3) respectively.

Since $\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}$ has a closed-form solution given by (2.10), it provides a computationally efficient approximation of $\mathbf{SW}_2(\mu_d, \nu_d)$ whose accuracy is quantified by Theorem 6.2. Next, we identify settings where this approximation

is accurate, by analyzing the error $\Xi_d(\mu_d) + \Xi_d(\nu_d)$.

Our first observation is that μ_d and ν_d should have zero means for the error to go to zero as $d \to +\infty$, and we develop a novel approximation of SW that takes into account this constraint. Going back to the definition of $\Xi_d(\mu_d^X)$ in (6.2), setting $\bar{X}_i = X_i - \mathbb{E}[X_i]$ and $\bar{X}_i' = X_i' - \mathbb{E}[X_i']$, we get

$$\mathtt{m}_2(\mu_d^X) = \mathbb{E}[\|\bar{X}_{1:d}\|^2] + \|\mathbb{E}[X_{1:d}]\|^2 \tag{6.8}$$

$$\beta_2^2(\mu_d^X) = \mathbb{E}\left[\left\langle \bar{X}_{1:d}, \bar{X}_{1:d}' \right\rangle^2\right] + 4\mathbb{E}\left[\left\langle \mathbb{E}[X_{1:d}], \bar{X}_{1:d} \right\rangle^2\right] + \|\mathbb{E}[X_{1:d}]\|^4 \ . \tag{6.9}$$

By Equations (6.8) and (6.9), since in practice the norm of the mean $\mathbb{E}[X_{1:d}]$ is expected to increase linearly with $d^{1/2}$ at least, so are $\mathtt{m}_2(\mu_d^X)$ and $\beta_2(\mu_d^X)$ as functions of d. As a consequence, $\Xi_d(\mu_d^X)$ cannot be shown to converge to 0 as $d \to \infty$ in this setting, but only to be bounded. However, if the data are centered, the norm of the mean is zero, thus $\Xi_d(\mu_d^X)$ might be decreasing. Therefore, we derive a convenient formula to compute $\mathbf{SW}_2(\mu_d, \nu_d)$ from $\mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d)$ where for any $\xi_d \in \mathcal{P}_2(\mathbb{R}^d)$, $\bar{\xi}_d$ is the centered version of ξ_d, *i.e.* the push-forward measure of ξ_d by $x \mapsto x - \mathbf{m}_{\xi_d}$ with $\mathbf{m}_{\xi_d} = \int_{\mathbb{R}^d} y \, \mathrm{d}\xi_d(y)$. This result is the last ingredient to formulate our approximation of SW.

Proposition 6.3. *Let $\mu_d, \nu_d \in \mathcal{P}_2(\mathbb{R}^d)$ with respective means $\mathbf{m}_{\mu_d}, \mathbf{m}_{\nu_d}$. Then, the Sliced-Wasserstein distance of order 2 can be decomposed as*

$$\mathbf{SW}_2^2(\mu_d, \nu_d) = \mathbf{SW}_2^2(\bar{\mu}_d, \bar{\nu}_d) + \frac{1}{d}\|\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d}\|^2 \ . \tag{6.10}$$

Based on Proposition 6.3, instead of approximating $\mathbf{SW}_2(\mu_d, \nu_d)$ directly with the term $\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}$, we propose estimating $\mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d)$ with $\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_d))\}$ and then using (6.10). This strategy yields our final approximation of SW, which is defined for any $\mu_d, \nu_d \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$\widehat{\mathbf{SW}}_2^2(\mu_d, \nu_d) = \mathbf{W}_2^2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_d))\} + \frac{1}{d}\|\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d}\|^2 \ , \tag{6.11}$$

where for $\xi_d \in \{\bar{\mu}_d, \bar{\nu}_d\}$, $\mathtt{m}_2(\xi_d)$ is defined in (6.3). Note that (6.11) can be simplified since by (2.10),

$$\mathbf{W}_2^2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_d))\} = d^{-1}(\mathtt{m}_2(\bar{\mu}_d)^{1/2} - \mathtt{m}_2(\bar{\nu}_d)^{1/2})^2 \ .$$

Besides, if μ_d and ν_d are both supported on a finite set of points, $\widehat{\mathbf{SW}}_2(\mu_d, \nu_d)$ has a closed-form expression: given $\xi_d = n^{-1}\sum_{j=1}^n \delta_{x^{(j)}} \in \mathcal{P}_2(\mathbb{R}^d)$ with $x^{(j)} \in \mathbb{R}^d$ for $j \in \{1, \dots, n\}$, we then have

$$\mathbf{m}_{\xi_d} = \frac{1}{n}\sum_{j=1}^n x^{(j)} \ , \quad \text{and} \quad \mathtt{m}_2(\xi_d) = \frac{1}{n}\sum_{j=1}^n \|x^{(j)}\|^2 \ .$$

The associated computational complexity is therefore in $\mathcal{O}(dn)$, which constitutes a significant benefit of our methodology over the traditional Monte Carlo estimation. Indeed, as explained in Section 2.6, computing $\mathbf{SW}_{p,L}$ (2.21) between two empirical distributions amounts to projecting sets of n observations in $\mathbb{R}^d$ along L directions, and sorting the projected data. The resulting computational complexity is $\mathcal{O}(Ldn + Ln\log n)$, meaning that the Monte Carlo estimate is more expensive when d, n and L increase, and it is

often unclear how L should be chosen in order to control the approximation error.

Hence, we introduced an alternative technique to estimate SW which does not rely on a finite set of random projections, as opposed to the commonly used Monte Carlo technique. Our approach thus eliminates the need for practitioners to tune the number of projections L, but also to sort the projected data. As a consequence, it is more efficient to compute $\widehat{\mathbf{SW}}_2(\mu_d, \nu_d)$ than $\mathbf{SW}_{2,L}(\mu_d, \nu_d)$ for large L. We illustrate this latter point with empirical results in Section 6.4.

6.3.3 Error analysis under weak dependence

We have discussed in Section 6.3.2 why centering the data is important to ensure that the approximation error goes to zero with increasing d. Next, we introduce a weak dependence condition under which the error is guaranteed to decrease as d increases.

We first consider a setting mentioned in [Reeves, 2017] where $\mu_d = \mu^{(1)} \otimes \cdots \otimes \mu^{(d)}$ and $\nu_d = \nu^{(1)} \otimes \cdots \otimes \nu^{(d)}$, $\otimes$ denoting the tensor product of measures, and $\mu^{(j)}, \nu^{(j)} \in \mathcal{P}_4(\mathbb{R})$ for $j \in \{1, \ldots, d\}$. We prove in this case that $\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathfrak{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathfrak{m}_2(\nu_d))\}$ converges to $\mathbf{SW}_2(\mu_d, \nu_d)$ at a rate of $d^{-1/8}$. This result is reported in Section 6.6.5, and can be interpreted as an extension of [Reeves, 2017, Corollary 3] for SW.

We emphasize that the assumptions of this first setting severely restrict the scope of application of our approximation method: in several statistical and machine learning tasks, the random variables of interest $\{X_i\}_{i=1}^d$ are not independent from each other (e.g. for image data, each X_i typically represents the value of a pixel at a certain position, thus depends on the neighboring pixels). Therefore, we relax this independence condition by considering a concept of "weak dependence" inspired by [Doukhan and Neumann, 2007] and properly defined in Definition 6.4.

Definition 6.4. *Let $(X_j)_{j\in\mathbb{N}^*}$ be a stationary sequence of one-dimensional random variables with mean zero, i.e. X_i and X_j have the same distribution for any $i, j \in \mathbb{N}^*$ and $\mathbb{E}[X_1] = 0$. We say that $(X_j)_{j\in\mathbb{N}^*}$ is* fourth-order weakly dependent *if there exist some constant $K \geq 0$ and a nonincreasing sequence of real coefficients $\{\rho(n)\}_{n\in\mathbb{N}}$ such that, for any $i, j \in \mathbb{N}^*$, $i \leq j$,*

$$|\mathrm{Cov}(X_i^2, X_j^2)| \leq K\rho(j-i)\,, \qquad |\mathrm{Cov}(X_i, X_j)| \leq K\rho(j-i)\,. \tag{6.12}$$

In addition, the sequence $\{\rho(n)\}_{n\in\mathbb{N}}$ satisfies $\sum_{n=0}^{+\infty} \rho(n) \leq \rho_\infty < +\infty$.

Intuitively, in practical applications, the weak dependence condition would essentially require the components of the observations not to exhibit strong correlations; yet, they are allowed to depend on each other. Furthermore, since our weak dependence condition is weaker than the one introduced in [Doukhan and Neumann, 2007, Theorem 1], it is satisfied by the various examples of models described in [Doukhan and Neumann, 2007, Section 5]. We present some of them below, to illustrate Definition 6.4 more clearly.

1) *Gaussian processes* and *associated processes* [Doukhan and Louhichi, 1999, Section 3.1], provided that they are stationary.

2) *Bernoulli shifts*: $X_t = H(\varepsilon_t, \ldots, \varepsilon_{t-r})$ for $t \in \mathbb{N}^*$, where $H : \mathbb{R}^{r+1} \to \mathbb{R}$ is a measurable function and $(\varepsilon_i)_{i\in\mathbb{N}^*}$ is a sequence of i.i.d. real random variables. A simple example of such process is given by *moving-average models*.

3) *Autoregressive models*, defined as $X_t = f(X_{t-1}, \dots, X_{t-r}) + \varepsilon_t$ for $t \in \mathbb{N}^*$, where $(\varepsilon_i)_{i\in\mathbb{N}^*}$ a sequence of i.i.d. real random variables with $\mathbb{E}\,|\varepsilon_1| < \infty$, and

$$|f(u_1, \dots, u_r) - f(v_1, \dots, v_r)| \leq \sum_{i=1}^{r} a_i\,|u_i - v_i|$$

for some $a_1, \dots, a_r \geq 0$ such that $\left(\sum_{i=1}^r a_i\right)^{1/r} < 1$.

We then consider a sequence of fourth-order weakly dependent random variables $(X_j)_{j\in\mathbb{N}^*}$, and prove that $\Xi_d(\mu_d^X)$ goes to zero as $d \to \infty$, with a rate of convergence depending on $\{\rho(n)\}_{n\in\mathbb{N}}$. This result is given in Section 6.6.6, and helps us refine Theorem 6.2 under this weak dependence condition: the next corollary establishes that the error approaches 0 at a rate of $d^{-1/8}$.

Corollary 6.5. *Let $(X_j)_{j\in\mathbb{N}^*}$ and $(Y_j)_{j\in\mathbb{N}^*}$ be sequences of random variables which are fourth-order weakly dependent. Set for any $d \in \mathbb{N}^*$, $X_{1:d} = \{X_j\}_{j=1}^d$ and $Y_{1:d} = \{Y_j\}_{j=1}^d$, and denote by μ_d, ν_d the distributions of $X_{1:d}$, $Y_{1:d}$ respectively. Then, there exists a universal constant $C > 0$ such that*

$$\left|\mathbf{SW}_2(\mu_d, \nu_d) - \mathbf{W}_2(\mathcal{N}(0, d^{-1}\mathfrak{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathfrak{m}_2(\nu_d)))\right| \leq C d^{-1/8}\,.$$

Hence, by replacing the independence condition of the first setting with weak dependence, we broaden the scope of application whilst guaranteeing that the approximation error goes to zero as d increases. We finally note that in these two settings, the data are required to have zero mean, which is automatically verified with our approximation method since we estimate SW between the centered distributions: see (6.11).

6.4 Experiments

6.4.1 Synthetic experiments

The goal of these experiments is to illustrate our theoretical results derived in Section 6.3. In each setting, we generate two sets of d-dimensional samples, denoted by $\{x^{(j)}\}_{j=1}^n$ and $\{y^{(j)}\}_{j=1}^n$ with $n = 10^4$ and $x^{(j)}, y^{(j)} \in \mathbb{R}^d$ for $j \in \{1, \dots, n\}$. We then approximate SW between their empirical distributions in $\mathcal{P}_2(\mathbb{R}^d)$, given by $\mu_d = n^{-1}\sum_{j=1}^n \delta_{x^{(j)}}$ and $\nu_d = n^{-1}\sum_{j=1}^n \delta_{y^{(j)}}$.

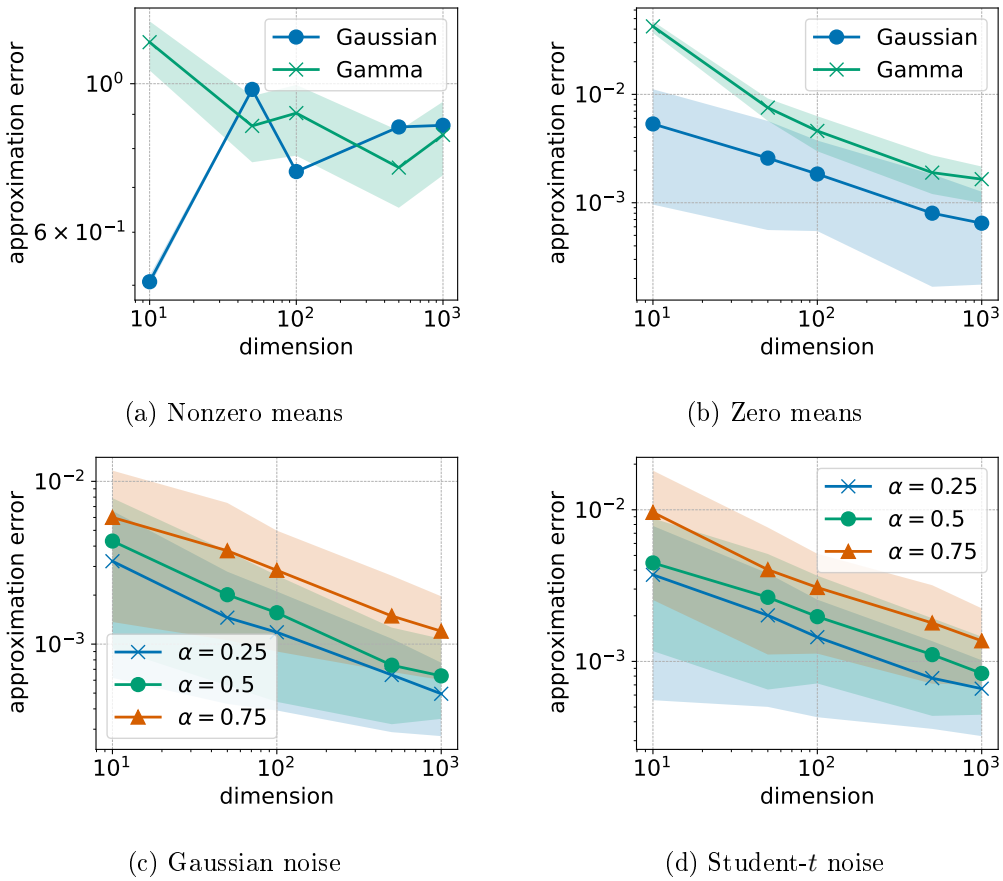
First, we analyze the consequences of centering data. Here, $\{x^{(j)}\}_{j=1}^n$ and $\{y^{(j)}\}_{j=1}^n$ are n independent samples from Gaussian or Gamma distributions: see Section 6.6.7 for more details. We compute

$$|\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathfrak{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathfrak{m}_2(\nu_d))\} - \mathbf{SW}_2(\mu_d, \nu_d)|$$

on the one hand, and

$$|\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathfrak{m}_2(\bar{\mu}_d)), \mathcal{N}(0, d^{-1}\mathfrak{m}_2(\bar{\nu}_d))\} - \mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d)|$$

on the other hand. In the Gaussian case, the exact value of $\mathbf{SW}_2$ is known (6.73), while for the Gamma distributions, it is approximated with Monte Carlo based on 2×10^4 random projections. Figures 6.1a and 6.1b show that the error goes to zero as d increases if the data are centered. This confirms our analysis provided in Section 6.3.2 about the



(a) Nonzero means

(b) Zero means

(c) Gaussian noise

(d) Student-t noise

Figure 6.1: Analysis of the approximation according to the dimension: in Figures 6.1a and 6.1b, data have independent components; in Figures 6.1c and 6.1d, they are stationary AR(1) processes. Errors are averaged over 100 runs and reported on log-log scale with their 10th-90th percentiles.

influence of the mean, and in Section 6.3.3 on sequences of independent random variables.

Next, we consider autoregressive processes of order one (AR(1)). An AR(1) process is defined as $X_1 = \varepsilon_1$ and, for $t \in \mathbb{N}^*$, $X_t = \alpha X_{t-1} + \varepsilon_t$, where $\alpha \in [0, 1]$ and $(\varepsilon_i)_{i \in \mathbb{N}^*}$ is an i.i.d. sequence of real random variables with $\mathbb{E}[\varepsilon_1] = 0$ and finite second-order moment. If $\alpha < 1$, the process has a stationary distribution and $(X_j)_{j \in \mathbb{N}^*}$ satisfies the weak dependence condition in its stationary regime [Doukhan and Neumann, 2008]. In practice, we generate a sample by using this recursion formula for $10^4 + d$ steps, and keeping the last d samples. The discarded samples correspond to a "burn-in" phase which helps reaching the stationary solution of the process. We generate $\{x^{(j)}\}_{j=1}^n$ and $\{y^{(j)}\}_{j=1}^n$ using the same distribution for the noise (either a Gaussian or Student's t-distribution, as described in Section 6.6.7). This means that both datasets come from the same distribution, thus the exact value of $\mathbf{SW}_2$ is zero. We plot on Figures 6.1c and 6.1d the approximation error according to $d \in [10, 10^3]$ for different values of α. The error converges to zero with increasing d, which is consistent with Corollary 6.5.

Note that Figure 6.1 exhibits rate of convergence that are better than the one in $d^{-1/8}$ derived in Section 6.3.3: in Figure 6.1b, the slope is approximately -0.45 (Gaussian) and -0.7 (Gamma), and in Figures 6.1c and 6.1d, it is on average -0.35. This suggests

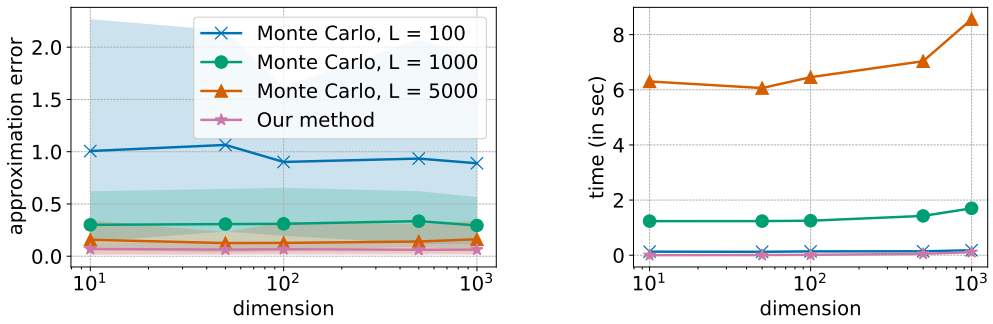


Figure 6.2: Comparison of different methods to approximate SW, according to their accuracy (left) and computation time (right). The datasets contain n samples of dimension d independently drawn from Gamma distributions, with $d \in [10^1, 10^3]$ and $n = 10^4$. Results are averaged over 100 runs.

that our theoretical bounds might be improved, and we further investigate this aspect for the Gaussian case: we consider the case where $\{x^{(j)}\}_{j=1}^n, \{y^{(j)}\}_{j=1}^n$ are n independent samples from Gaussian distributions with diagonal covariance matrices, and we prove that $\mathbb{E}|\mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_d))\} - \mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d)|$ goes to 0 as $dn \to +\infty$ with a convergence rate in $d^{-1/2}n^{-1/2}$. We provide the complete statement and formal proof in Section 6.6.7 (Proposition 6.9). This result is consistent with Figure 6.1b, and is a first encouraging step towards the following research direction: we can study if our proofs and the ones in Reeves [2017] can be refined when assuming additional structure on the distributions (e.g., sub-Gaussian and sub-exponential), in order to identify the settings under which our current bounds are tight or can be improved.

Finally, we compare our approximation scheme against the standard Monte Carlo estimation, in terms of accuracy and computation time. We use the same setting as in Figure 6.1b, where the n samples are independently drawn from Gamma distributions. We compute $\widehat{\mathbf{SW}}_2(\mu_d, \nu_d)$ (6.11) and $\mathbf{SW}_{2,L}(\mu_d, \nu_d)$ (2.21) with $L \in \{100, 1000, 5000\}$, and we compare each approximation with $\mathbf{SW}_{2,2\times10^4}(\mu_d, \nu_d)$, which we consider as the exact value of SW. Figure 6.2 reports the approximation error and computation time of each scheme for $d \in [10, 10^4]$, and shows that our method is more accurate and faster than Monte Carlo. In particular, when $d = 10^3$, the average computation time of our technique is 0.02s, while the second best approximation (Monte Carlo with $L = 5000$) takes more than 8s. Besides, we observe that Monte Carlo is very sensitive to the hyperparameters, since it loses accuracy when L decreases and gets slower as L and d increase. This observation is consistent with the computational complexity of $\mathbf{SW}_{2,L}$ recalled at the end of Section 6.3.2. On the other hand, our approximation scheme is extremely efficient even for large d and n, since it is based on a simple deterministic formula which does not require projecting and sorting data along random directions.

6.4.2 Image generation

Finally, we leverage our theoretical insights to design a novel method for a typical generative modeling application. The problem consists in tuning a neural network that takes as input k-dimensional samples from a reference distribution (e.g., uniform or Gaussian), to generate images of dimension $d > k$. During the training phase, the parameters of the network are updated by iteratively minimizing a dissimilarity measure between the

dataset to fit and the generated images.

In [Deshpande et al., 2018], the dissimilarity measure is Monte Carlo SW approximated with 10^4 random projections, and the resulting generative model is called the *Sliced-Wasserstein generator* (SWG). This model performs well on moderately high-dimensional image datasets (e.g., 28×28 for MNIST images [LeCun and Cortes, 2010]). However, for very large dimensions (e.g., $64 \times 64 \times 3$ for the CelebA dataset [Liu et al., 2015]), Monte Carlo SW requires more than 10^4 random projections to capture relevant information, which leads to very expensive training iterations and potential memory issues.

To offer better scalability, SWG can be augmented with a discriminator network [Deshpande et al., 2018, Section 3.2] that aims at finding a lower-dimensional space in which the two projected datasets are clearly distinguishable. The intuition behind this heuristic is that the more distinct the two datasets are from each other, the fewer projection directions Monte Carlo SW requires to provide useful information. The training then consists in optimizing the generator's and discriminator's objective functions in an alternating fashion.

Our novel approach builds on SWG and modifies the saddle-point problem in [Deshpande et al., 2018, Section 3.2]: motivated by the gain in accuracy and time illustrated in Figure 6.2 on high-dimensional datasets, we propose to replace Monte Carlo SW with our approximate SW (6.11) in the generator's objective; then, to make sure that our approximation is accurate, we regularize the discriminator's objective:

$$\max_{\psi} \; L(\psi) + \lambda_1 \left\| \mathrm{Cov}[d'_\psi(X)] \right\|_F^2 + \lambda_1 \left\| \mathrm{Cov}[d'_\psi(g_\phi(Z))] \right\|_F^2 \tag{6.13}$$

$$+ \lambda_2 \, \mathbb{E}\left[\|d'_\psi(X)\|^{-2} \right] + \lambda_2 \, \mathbb{E}\left[\|d'_\psi(g_\phi(Z))\|^{-2} \right] \tag{6.14}$$

where L is the discriminator's loss used in SWG, g_ϕ and d'_ψ are the generator's last layer and the discriminator's penultimate layer respectively (parameterized by ϕ, ψ), X and Z are the random variables corresponding to the images to fit and the generator's input, Cov denotes the covariance matrix, $\| \cdot \|_F$ the Frobenius norm, and $\lambda_1, \lambda_2 \geq 0$. The regularization in (6.13) enforces the weak dependence condition (Corollary 6.5), while (6.14) prevents the network to converge to $d'_\psi = 0$. We call this generative adversarial network *regularized deterministic SWG* (reg-det-SWG).

To investigate the consequences of (i) regularizing the discriminator, and (ii) replacing the Monte Carlo SW with our approximation, we design another model, called *regularized SWG* (reg-SWG): similarly to SWG, the generator minimizes $\mathbf{SW}_{2,10^4}$, but the discriminator's objective is regularized as in (6.13), (6.14).

We then compare reg-det-SWG against SWG and reg-SWG, by training the models on MNIST and CelebA and measuring their respective training time and Fréchet Inception Distances (FID, Heusel et al. [2017]): see Table 6.1. We used the same network architectures for all methods, and tuned (λ_1, λ_2) via cross-validation: more details on the experimental setup are given in Section 6.6.8. First, we observe that the regularized models produce images of higher quality, since reg-SWG and reg-det-SWG return lower FID values than SWG. The FID of reg-SWG and reg-det-SWG are close for both datasets, thus the two models seem to yield similar performances. Hence, we report in Figure 6.3 the images generated by SWG and reg-det-SWG only.

The training process is more expensive when regularizing the discriminator: the average running time per epoch is higher for the regularized models. We also observe that

Dataset	Model	FID	$T_{\mathbf{SW}}$ (s/epoch)		$T_{\mathbf{tot}}$ (s/epoch)	
			GPU	CPU	GPU	CPU
MNIST	SWG	22.41 ± 2.34	1.3	1.4×10^2	4.5	2.7×10^2
	Reg-SWG	15.53 ± 0.88	1.1	1.1×10^2	6.5	3.0×10^2
	Reg-det-SWG	15.72 ± 0.57	0.07	0.2	5.3	1.5×10^2
CelebA	SWG	31.04 ± 2.78	10.1	2.7×10^3	3.9×10^2	1.6×10^4
	Reg-SWG	24.14 ± 0.48	10.0	2.7×10^3	4.4×10^2	2.0×10^4
	Reg-det-SWG	23.65 ± 0.93	1.3	2.6	4.2×10^2	1.7×10^4

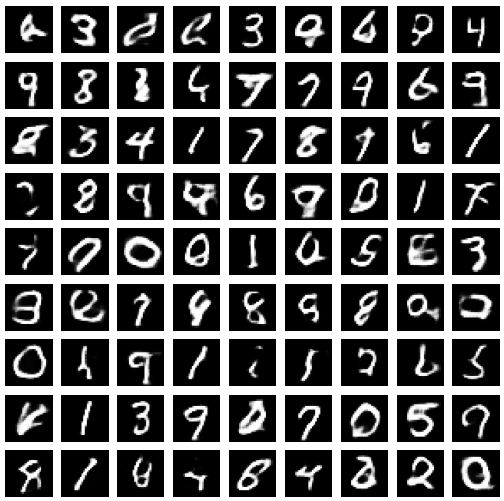
Table 6.1: Results obtained after training generative models on MNIST and CelebA, averaged over 5 runs. FID are reported with their standard deviation (the lower FID, the better). T_{SW} denotes the average time per epoch for approximating SW. T_{tot} is the average running time per epoch.

reg-det-SWG is faster than reg-SWG, which is consistent with the fact that our approximation method is faster than Monte Carlo on high-dimensional settings. To further illustrate this point, we reported the average time spent in computing the generative loss per epoch, *i.e.* $\mathbf{SW}_{2,10^4}$ for SWG and reg-SWG, and $\widehat{\mathbf{SW}}_2$ for reg-det-SWG: see column T_{SW} in Table 6.1. On GPU, reg-det-SWG is at least 15 times faster than SWG and reg-SWG on MNIST, and 6 times faster on CelebA.

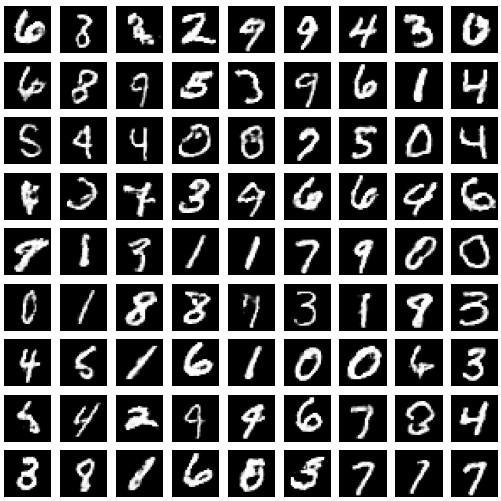
Note that the models were trained using PyTorch, thus Monte Carlo SW benefits from a GPU-accelerated implementation of the sorting operation (with the function `torch.sort`). We also reported the computation times when models are trained on CPU. In this case, computing $\widehat{\mathbf{SW}}_2$ takes at most less than 3s per epoch, whereas the Monte Carlo estimation executes in several minutes (e.g., approximately 45min on CelebA). As a result, the total training time is almost the same for reg-det-SWG and SWG on CelebA, and the lowest for reg-det-SWG on MNIST. Our approximation method then fosters the development of models to speed up existing machine learning algorithms on CPU, which is useful when powerful hardware resources are not available, or when their use is deliberately avoided for environmental purposes.

6.5 Conclusion

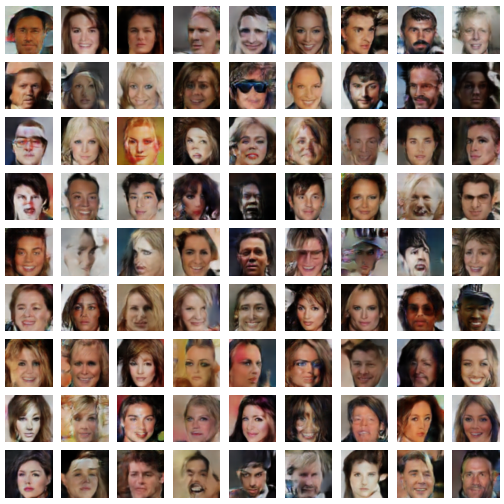
We presented a novel method to approximate the Sliced-Wasserstein distance of order 2, which relies on the concentration-of-measure phenomenon for random projections. The resulting method computes SW with simple deterministic operations, which are computationally efficient even on high-dimensional settings and do not require any hyperparameters. We proved nonasymptotical guarantees showing that, under a weak dependence condition, the approximation error goes to zero as the dimension increases. Our theoretical findings are then illustrated with experiments on synthetic datasets. Motivated by the computational efficiency and accuracy of our approximate SW, we finally designed a novel approach for image generation that leverages our theoretical insights. As compared to generative models based on SW estimated with Monte Carlo, our framework produces images of higher quality with further computational benefits. This encourages the use of our approximate SW on other algorithms that rely on Monte Carlo SW, e.g. autoencoders [Kolouri et al., 2019b] or normalizing flows [Dai and Seljak, 2021].



(a) SWG (FID = 19.52)



(b) Reg-det-SWG (FID = 14.87)



(c) SWG (FID = 27.75)

(d) Reg-det-SWG (FID = 22.87)

Figure 6.3: Images generated after training on MNIST (top row) and CelebA (bottom row). For each model, the images are associated with the lowest FID obtained over 5 runs.

The weak dependence condition can be inappropriate to describe the underlying geometry of real data in ML applications, and in that case, approximating SW with our method seems inadequate. To overcome this problem, we encourage practitioners to resort to models where real data are represented by features that can be made weakly dependent. This strategy has proven successful in our image generation experiment: the reg-det-SWG model uses our approximation to compare two sets of features (instead of the raw images) whose covariance matrices are regularized to enforce weak dependence. Since many ML techniques make use of features and regularizers, we believe that our methodology is not restrictive and can then be applied to other standard problems than image generation. Besides, our weak dependence condition in Definition 6.4 is weaker

than the one in [Doukhan and Neumann, 2007], which is a notion commonly used in statistics.

Our empirical results on synthetic data show that the approximation error goes to zero with a faster convergence rate than the one we proved. Then, the main current limitation of our framework is that our theoretical convergence rate in $d^{-1/8}$ might be slower than necessary. We proved that the overall approximation error is upper-bounded by a term in $d^{-1/2}$ when comparing Gaussians with diagonal covariance matrices, and the improvement of our error bounds for other specific distributions is left for future work. On the other hand, the extension of our methodology to variants of SW is another challenging future research direction. To the best of our knowledge, the literature on the concentration of measure phenomenon focuses on linear random projections, therefore the derivation of deterministic approximations for SW based on nonlinear projections seems highly nontrivial. A more promising direction would be to generalize our approach to SW based on k-dimensional linear projection by leveraging the bound in [Reeves, 2017, Theorem 1] for $k > 1$.

6.6 Appendix: Postponed Proofs and Experimental Details

6.6.1 Conditional central limit theorem for Gaussian projections

We give the formal statement of the result presented in Section 6.2, corresponding to [Reeves, 2017, Theorem 1] for the special case of one-dimensional projections.

Theorem 6.6 ([Reeves, 2017, Theorem 1]). *There exists a constant C such that for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d} \mathbf{W}_2^2\left(\theta_\sharp^\star \mu, \mathcal{N}\big(0, d^{-1}\mathtt{m}_2(\mu)\big)\right) \mathrm{d}\boldsymbol{\gamma}_d(\theta) \tag{6.15}$$

$$\leq C d^{-1}\left\{\alpha(\mu) + \big(\mathtt{m}_2(\mu)\beta_1(\mu)\big)^{1/2} + \mathtt{m}_2(\mu)^{1/5}\beta_2(\mu)^{4/5}\right\}, \tag{6.16}$$

where

$$\mathtt{m}_2(\mu) = \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x)\,, \tag{6.17}$$

$$\alpha(\mu) = \int_{\mathbb{R}^d} \big|\,\|x\|^2 - \mathtt{m}_2(\mu)\big|\mathrm{d}\mu(x)\,, \tag{6.18}$$

$$\beta_q(\mu) = \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \big|\langle x, x'\rangle\big|^q \,\mathrm{d}(\mu\otimes\mu)(x,x')\right)^{\frac{1}{q}}\,, \tag{6.19}$$

with $q \in \{1, 2\}$.

6.6.2 Proof of Proposition 6.1

Proof of Proposition 6.1. Let $\theta \in \mathbb{R}^d$ and write $\theta = r\bar{\theta}$, $r \geq 0$ and $\bar{\theta} \in \mathbb{S}^{d-1}$. Then, we get

$$\mathbf{W}_p^p(\theta_\sharp^\star\mu, \theta_\sharp^\star\nu) = \mathbf{W}_p^p\big((r\bar{\theta})_\sharp^\star\mu, (r\bar{\theta})_\sharp^\star\nu\big) \tag{6.20}$$

$$= \int_0^1 \big|F^{\leftarrow}_{(r\bar{\theta})_\sharp^\star\mu}(t) - F^{\leftarrow}_{(r\bar{\theta})_\sharp^\star\nu}(t)\big|^p \mathrm{d}t\,, \tag{6.21}$$

where (6.21) results from (2.12): $F_{\tilde{\mu}}$ and $F^{\leftarrow}_{\tilde{\mu}}$ denote the cumulative distribution and quantile function respectively, of a one-dimensional probability measure $\tilde{\mu}$, *i.e.* $F_{\tilde{\mu}}(s) = \tilde{\mu}((-\infty, s])$ and $F^{\leftarrow}_{\tilde{\mu}}(t) = \inf\{s' \in \mathbb{R} \,:\, F_{\tilde{\mu}}(s') \geq t\}$ for $s \in \mathbb{R}$ and $t \in [0,1]$. For any $r > 0$ and $\theta \in \mathbb{S}^{d-1}$, we get

$$F_{(r\bar{\theta})^\star_\sharp \mu}(s) = \big((r\bar{\theta})^\star_\sharp \mu\big)\{(-\infty, s]\} \tag{6.22}$$

$$= \big(\bar{\theta}^\star_\sharp \mu\big)\{(-\infty, s/r]\} = F_{\bar{\theta}^\star_\sharp \mu}(s/r)\ , \tag{6.23}$$

which easily implies that $F^{\leftarrow}_{(r\bar{\theta})^\star_\sharp \mu}(t) = r F^{\leftarrow}_{\bar{\theta}^\star_\sharp \mu}(t)$. Therefore, using this property in (6.21), we obtain,

$$\mathbf{W}_p^p(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) = \int_0^1 \big| r F^{\leftarrow}_{\bar{\theta}^\star_\sharp \mu}(t) - r F^{\leftarrow}_{\bar{\theta}^\star_\sharp \nu}(t) \big|^p \mathrm{d}t \tag{6.24}$$

$$= r^p\, \mathbf{W}_p^p(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu)\ . \tag{6.25}$$

By applying a d-spherical change of variables in the definition of $\widetilde{\mathbf{SW}}_p$ (6.6) and plugging (6.25),

$$\widetilde{\mathbf{SW}}_p^p(\mu, \nu) = \int_{\mathbb{R}_+} \int_{\mathbb{S}^{d-1}} r^p \mathbf{W}_p^p\big(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu\big)\ (2\pi)^{-\frac{d}{2}} d^{\frac{d}{2}} e^{-\frac{d}{2}\|r\bar{\theta}\|^2} r^{d-1} \mathrm{d}\bar{\theta}\mathrm{d}r \tag{6.26}$$

$$= (2\pi)^{-\frac{d}{2}} d^{\frac{d}{2}} \int_{\mathbb{R}_+} r^{p+d-1} e^{-\frac{d}{2}r^2} \left(\int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu\big) \mathrm{d}\bar{\theta} \right) \mathrm{d}r\ . \tag{6.27}$$

Since the surface area of $\mathbb{S}^{d-1}$ is equal to $2\pi^{\frac{d}{2}}\Gamma(d/2)^{-1}$ [Huber, 1982], and by definition of SW (2.20),

$$\int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu\big) \mathrm{d}\bar{\theta} = 2\pi^{\frac{d}{2}}\Gamma(d/2)^{-1} \mathbf{SW}_p^p(\mu, \nu)\ .$$

Besides, by applying the change of variables $t = (d/2)^{1/2} r$,

$$\begin{aligned} &\int_{\mathbb{R}_+} r^{p+d-1} e^{-\frac{d}{2}r^2} \mathrm{d}r \\ &= 2^{(p+d)/2} d^{-(p+d)/2} \int_{\mathbb{R}_+} t^{p+d-1} e^{-t^2} \mathrm{d}t \\ &= 2^{(p+d)/2-1} d^{-(p+d)/2}\ \Gamma\big((d+p)/2\big)\ . \end{aligned}$$

We finally obtain,

$$\widetilde{\mathbf{SW}}_p^p(\mu, \nu) = \left(\frac{2}{d}\right)^{p/2} \frac{\Gamma\big(d/2 + p/2\big)}{\Gamma(d/2)}\ \mathbf{SW}_p^p(\mu, \nu)\ . \tag{6.28}$$

□

6.6.3 Proof of Theorem 6.2

Proof of Theorem 6.2. By the triangle inequality, for any $\theta \in \mathbb{R}^d$,

$$\big|\mathbf{W}_2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}\big| \tag{6.29}$$

$$\leq \mathbf{W}_2\{\theta^\star_\sharp \mu_d, \mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d))\} + \mathbf{W}_2\{\theta^\star_\sharp \nu_d, \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \tag{6.30}$$

Therefore, taking the integral with respect to $\boldsymbol{\gamma}_d$,

$$\int_{\mathbb{R}^d} \Big(\mathbf{W}_2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta) \tag{6.31}$$

$$\leq \int_{\mathbb{R}^d} \Big(\mathbf{W}_2\{\theta^\star_\sharp \mu_d, \mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d))\} + \mathbf{W}_2\{\theta^\star_\sharp \nu_d, \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta) \tag{6.32}$$

$$\leq 2 \left\{ \int_{\mathbb{R}^d} \mathbf{W}_2^2\{\theta^\star_\sharp \mu_d, \mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d))\} \mathrm{d}\boldsymbol{\gamma}_d(\theta) + \int_{\mathbb{R}^d} \mathbf{W}_2^2\{\theta^\star_\sharp \nu_d, \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \mathrm{d}\boldsymbol{\gamma}_d(\theta) \right\} , \tag{6.33}$$

where (6.33) follows from $(a+b)^2 \leq 2(a^2+b^2)$. Then, we apply Theorem 6.6 to bound (6.33), and we conclude there exists a universal constant $C > 0$ such that

$$\int_{\mathbb{R}^d} \Big(\mathbf{W}_2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta) \tag{6.34}$$

$$\leq C\big(\Xi_d(\mu_d) + \Xi_d(\nu_d)\big) \tag{6.35}$$

Using $|\|a\| - \|b\|| \leq \|a - b\|$ in $\mathcal{L}^2(\mathbb{R}^d, \boldsymbol{\gamma}_d)$ gives

$$\Big| \Big\{ \int_{\mathbb{R}^d} \mathbf{W}_2^2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) \mathrm{d}\boldsymbol{\gamma}_d(\theta) \Big\}^{1/2}$$

$$- \Big\{ \int_{\mathbb{R}^d} \mathbf{W}_2^2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \mathrm{d}\boldsymbol{\gamma}_d(\theta) \Big\}^{1/2} \Big| \tag{6.36}$$

$$\leq \Big\{ \int_{\mathbb{R}^d} \Big(\mathbf{W}_2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta) \Big\}^{1/2} \tag{6.37}$$

$$\leq C^{1/2}\big(\Xi_d(\mu_d) + \Xi_d(\nu_d)\big)^{1/2} \tag{6.38}$$

By (6.6) and Proposition 6.1,

$$\int_{\mathbb{R}^d} \mathbf{W}_2^2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) \mathrm{d}\boldsymbol{\gamma}_d(\theta) = \widetilde{\mathbf{SW}}_2^2(\mu_d, \nu_d) = \mathbf{SW}_2^2(\mu_d, \nu_d) .$$

We then obtain the final result by rewritting (6.36) as

$$\big| \mathbf{SW}_2(\mu_d, \nu_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \big| .$$

□

6.6.4 Proof of Proposition 6.3

Proof of Proposition 6.3. This result follows from an analogous translation property of the Wasserstein distance: by [Peyré and Cuturi, 2019, Remark 2.19], $\mathbf{W}_2$ can factor out translations; in particular, for any $\xi, \xi' \in \mathcal{P}_2(\mathbb{R}^d)$ with respective means $\mathbf{m}_\xi, \mathbf{m}_{\xi'}$ and centered versions $\bar{\xi}, \bar{\xi}'$,

$$\mathbf{W}_2^2(\xi, \xi') = \mathbf{W}_2^2(\bar{\xi}, \bar{\xi}') + \|\mathbf{m}_\xi - \mathbf{m}_{\xi'}\|^2 . \tag{6.39}$$

By using (6.39) in the definition of SW of order 2 (2.20), we obtain for any $\mu_d, \nu_d \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathbf{SW}_2^2(\mu_d, \nu_d) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_2^2(\theta^\star_\sharp \bar{\mu}_d, \theta^\star_\sharp \bar{\nu}_d) \mathrm{d}\boldsymbol{\sigma}(\theta) + \int_{\mathbb{S}^{d-1}} |\mathbf{m}_{\theta^\star_\sharp \mu_d} - \mathbf{m}_{\theta^\star_\sharp \nu_d}|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{6.40}$$

$$= \mathbf{SW}_2^2(\bar{\mu}_d, \bar{\nu}_d) + \int_{\mathbb{S}^{d-1}} |\mathbf{m}_{\theta^\star_\sharp \mu_d} - \mathbf{m}_{\theta^\star_\sharp \nu_d}|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) . \tag{6.41}$$

By the properties of push-forward measures, $\mathbf{m}_{\theta^\star_\sharp \xi} = \langle \theta, \mathbf{m}_\xi \rangle$ for any $\theta \in \mathbb{S}^{d-1}$ and $\xi \in \mathcal{P}_2(\mathbb{R}^d)$. The second term of (6.41) can thus be reformulated as

$$\int_{\mathbb{S}^{d-1}} |\mathbf{m}_{\theta^\star_\sharp \mu_d} - \mathbf{m}_{\theta^\star_\sharp \nu_d}|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{6.42}$$

$$= \int_{\mathbb{S}^{d-1}} |\langle \theta, \mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d} \rangle|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{6.43}$$

$$= (\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d})^\top \left(\int_{\mathbb{S}^{d-1}} \theta\theta^\top \mathrm{d}\boldsymbol{\sigma}(\theta) \right) (\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d}) \tag{6.44}$$

$$= (1/d) \, \|\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d}\|^2 \, , \tag{6.45}$$

where the last equation results from $\int_{\mathbb{S}^{d-1}} \theta\theta^\top \mathrm{d}\boldsymbol{\sigma}(\theta) = (1/d)\mathbf{I}_d$. The final result is obtained by incorporating (6.45) in (6.41).

□

6.6.5 Error analysis under independence

This section gives a detailed analysis of the error bound under the first setting discussed in Section 6.3.3: we consider sequences of independent random variables which have zero means and finite fourth-order moments, and we derive an upper bound for Ξ_d in the next proposition.

Proposition 6.7. *Let $(X_j)_{j\in\mathbb{N}^*}$ be a sequence of independent random variables with zero means and $\mathbb{E}[X_j^4] < +\infty$ for $j \in \mathbb{N}^*$. Set for any $d \in \mathbb{N}^*$, $X_{1:d} = \{X_j\}_{j=1}^d$ and let μ_d be the distribution of $X_{1:d}$. Then, we have*

$$\Xi_d(\mu_d) \leq d^{-1/2} \Big\{ \max_{1\leq j\leq d} \mathrm{Var}[X_j^2] \Big\}^{1/2} + \Big\{ d^{-1/4} + d^{-2/5} \Big\} \max_{1\leq j\leq d} \mathrm{Var}[X_j] \, . \tag{6.46}$$

Proof of Proposition 6.7. Given the definition of $\Xi_d(\mu_d)$ (6.2), the proof consists in bounding $\mathtt{m}_2(\mu_d)$, $\alpha(\mu_d)$ and $\beta_q(\mu_d)$ for $q \in \{1, 2\}$.

Since for any $j \in \{1, \ldots, d\}$, $\mathbb{E}[X_j] = 0$, then $\mathrm{Var}\big[X_j\big] = \mathbb{E}[X_j^2]$ and

$$\mathtt{m}_2(\mu_d) = \sum_{j=1}^d \mathbb{E}[X_j^2] = \sum_{j=1}^d \mathrm{Var}[X_j] \leq d \max_{1\leq j\leq d} \mathrm{Var}[X_j] \tag{6.47}$$

To bound $\alpha(\mu_d)$, we first use the Cauchy–Schwarz inequality.

$$\alpha(\mu_d) \leq \left\{ \int_{\mathbb{R}^d} \big(\|x_{1:d}\|^2 - \mathtt{m}_2(\mu_d) \big)^2 \mathrm{d}\mu_d(x_{1:d}) \right\}^{1/2} \tag{6.48}$$

Besides, $\int_{\mathbb{R}^d} \big(\|x_{1:d}\|^2 - \mathtt{m}_2(\mu_d) \big)^2 \mathrm{d}\mu_d(x_{1:d}) = \mathrm{Var}\big[\|X_{1:d}\|^2\big]$, and since the d components of $X_{1:d}$ are assumed to be pairwise independent, $\mathrm{Var}\big[\|X_{1:d}\|^2\big] = \sum_{j=1}^d \mathrm{Var}\big[X_j^2\big]$. We conclude that

$$\alpha(\mu_d) \leq \left(\sum_{j=1}^d \mathrm{Var}\big[X_j^2\big] \right)^{1/2} \leq \big(d \max_{1\leq j\leq d} \mathrm{Var}[X_j^2] \big)^{1/2} \, . \tag{6.49}$$

Finally, we bound $\beta_q(\mu_d)$ for $q \in \{1, 2\}$ by bounding $\beta_2(\mu_d)$ then using the fact that $\beta_1(\mu_d) \leq \beta_2(\mu_d)$ by the Cauchy–Schwarz inequality. Denote by $X'_{1:d}$ an independent copy of $X_{1:d}$.

$$\langle X_{1:d}, X'_{1:d} \rangle^2 = \Big(\sum_{j=1}^{d} X_j X'_j \Big)^2 = \sum_{j=1}^{d} {X_j}^2 {X'_j}^2 + 2 \sum_{i<j} X_i X'_i X_j X'_j \,. \tag{6.50}$$

Since $X_{1:d}$ and $X'_{1:d}$ are independent on one hand, and they both are sequences of d independent random variables with zero means on the other hand, we have

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} (\langle x_{1:d}, x'_{1:d} \rangle)^2 \mathrm{d}(\mu_d \otimes \mu_d)(x_{1:d}, x'_{1:d}) \tag{6.51}$$

$$= \sum_{j=1}^{d} \mathbb{E}\big[{X_j}^2\big] \mathbb{E}\big[{X'_j}^2\big] = \sum_{j=1}^{d} \mathbb{E}\big[X_j^2\big]^2 = \sum_{j=1}^{d} \mathrm{Var}\big[X_j\big]^2 \,. \tag{6.52}$$

Therefore, $\beta_2(\mu_d) \leq (\sum_{j=1}^{d} \mathrm{Var}[X_j]^2)^{1/2} \leq (d \max_{1\leq j\leq d} \mathrm{Var}[X_j]^2)^{1/2}$. Since $X_{1:d}$ has finite second and fourth-order moments,

$$\max_{1\leq j\leq d} \mathrm{Var}[X_j], \ \max_{1\leq j\leq d} \mathrm{Var}[X_j^2] < \infty \,,$$

and we get

$$\mathtt{m}_2(\mu_d) \leq d \max_{1\leq j\leq d} \mathrm{Var}[X_j], \tag{6.53}$$

$$\alpha(\mu_d) \leq d^{1/2} (\max_{1\leq j\leq d} \mathrm{Var}[X_j^2])^{1/2}, \tag{6.54}$$

$$\beta_1(\mu_d), \, \beta_2(\mu_d) \leq d^{1/2} \max_{1\leq j\leq d} \mathrm{Var}[X_j] \,. \tag{6.55}$$

The final result is obtained by bounding $\Xi(\mu_d)$ using (6.53), (6.54) and (6.55).

□

Note that the setting considered in Proposition 6.7 was mentioned in [Reeves, 2017] to illustrate the conditions of [Reeves, 2017, Corollary 3]. We derived an explicit upper bound of Ξ_d under this setting for completeness, showing that $\Xi_d(\mu_d)$ goes to zero as $d \to \infty$, which we can then use to refine the convergence rate in Theorem 6.2, as we explained in Section 6.3.3.

6.6.6 Error analysis under weak dependence

We now analyze the error under the weak dependence condition introduced in Definition 6.4. Specifically, the proposition below gives the formal statement of the result mentioned before Corollary 6.5: we consider a sequence of fourth-order weakly dependent random variables, and we prove that $\Xi(\mu_d)$ goes to zero as $d \to \infty$, with a convergence rate that depends on $\{\rho(n)\}_{n\in\mathbb{N}^*}$.

Proposition 6.8. *Let $(X_j)_{j\in\mathbb{N}^*}$ be a sequence of random variables which is fourth-order weakly dependent. Set for any $d \in \mathbb{N}^*$, $X_{1:d} = \{X_j\}_{j=1}^d$ and denote by μ_d the distribution of $X_{1:d}$. Then, there exists a universal constant $C > 0$ such that*

$$\Xi_d(\mu_d) \leq C \Big\{ d^{-1/2} \big(\rho(0) + 2\rho_\infty\big)^{1/2} + d^{-1/4} \rho(0)^{1/2} \big(\rho(0)^2 + 2\rho_\infty \max_{1\leq k\leq d-1} \rho(k)\big)^{1/4} + d^{-2/5} \rho(0)^{1/5} \big(\rho(0)^2 + 2\rho_\infty \max_{1\leq k\leq d-1} \rho(k)\big)^{2/5} \Big\} \,. \tag{6.56}$$

Proof of Proposition 6.8. We proceed as in the proof of Proposition 6.7, *i.e.* by bounding $\mathtt{m}_2(\mu_d)$, $\alpha(\mu_d)$ and $\beta_2(\mu_d)$.

Since $(X_j)_{j\in\mathbb{N}^*}$ is assumed to be fourth-order weakly dependent, then by Definition 6.4, there exist some constant $K \geq 0$ and a nonincreasing sequence of real coefficients $\{\rho(n)\}_{n\in\mathbb{N}}$ such that, for any $1 \leq i \leq j \leq d$,

$$|\mathrm{Cov}(X_i^2, X_j^2)| \leq K\rho(j-i), \quad |\mathrm{Cov}(X_i, X_j)| \leq K\rho(j-i)\ . \tag{6.57}$$

First, using the same arguments as in (6.47), we have $\mathtt{m}_2(\mu_d) = \sum_{j=1}^d \mathrm{Var}[X_j]$. We then use the second inequality in (6.57) to bound $\mathtt{m}_2(\mu_d)$ as follows.

$$\mathtt{m}_2(\mu_d) = \sum_{j=1}^d \mathrm{Cov}(X_j, X_j) \leq dK\rho(0) \tag{6.58}$$

Regarding $\alpha(\mu_d)$, we use the Cauchy–Schwarz inequality again (6.48) but in this setting, the right-hand side features non-zero covariance terms:

$$\int_{\mathbb{R}^d} \big(\, \|x_{1:d}\|^2 - \mathtt{m}_2(\mu_d)\big)^2 \mathrm{d}\mu_d(x_{1:d}) = \mathrm{Var}\big[\|X_{1:d}\|^2\big] \tag{6.59}$$

$$= \sum_{j=1}^d \mathrm{Var}\big[X_j^2\big] + 2\ \sum_{i<j} \mathrm{Cov}\big(X_i^2, X_j^2\big)\ . \tag{6.60}$$

By using the first inequality in (6.57), we get for any $d \in \mathbb{N}^*$,

$$\sum_{j=1}^d \mathrm{Var}\big[X_j^2\big] = \sum_{j=1}^d \mathrm{Cov}(X_j^2, X_j^2) \leq Kd\rho(0)\ , \tag{6.61}$$

$$\sum_{i<j} \mathrm{Cov}\big(X_i^2, X_j^2\big) \leq \sum_{i<j} |\mathrm{Cov}\big(X_i^2, X_j^2\big)| \leq K \sum_{i<j} \rho(j-i) \tag{6.62}$$

$$\leq K \sum_{n=1}^{d-1} (d-n)\rho(n) \tag{6.63}$$

$$\leq Kd \sum_{n=1}^{d-1} \rho(n) \leq Kd\rho_\infty \tag{6.64}$$

where (6.63) results from the change of variable $n = j - i$. Besides, by Definition 6.4, $\{\rho(n)\}_{n\in\mathbb{N}}$ is a nonincreasing sequence satisfying $\sum_{n=0}^{+\infty} \rho(n) \leq \rho_\infty < +\infty$, hence (6.64). We conclude that for any $d \in \mathbb{N}^*$,

$$\alpha(\mu_d) \leq d^{1/2} K^{1/2} \big(\rho(0) + 2\rho_\infty\big)^{1/2}\ . \tag{6.65}$$

Let us now bound $\beta_2(\mu_d)$. First, for any $d \in \mathbb{N}^*$,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} (\langle x_{1:d}, x'_{1:d} \rangle)^2 \mathrm{d}(\mu_d \otimes \mu_d)(x_{1:d}, x'_{1:d}) \tag{6.66}$$

$$= \sum_{j=1}^{d} \mathbb{E}\big[{X_j}^2\big]\mathbb{E}\big[{X'_j}^2\big] + 2\sum_{i<j} \mathbb{E}\big[X_i X_j\big]\mathbb{E}\big[X'_i X'_j\big] \tag{6.67}$$

$$= \sum_{j=1}^{d} \mathbb{E}\big[{X_j}^2\big]^2 + 2\sum_{i<j} \mathbb{E}\big[X_i X_j\big]^2 \tag{6.68}$$

$$= \sum_{j=1}^{d} \mathrm{Var}\big[X_j\big]^2 + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j)^2 \,, \tag{6.69}$$

where we used $\mathbb{E}[X_i] = 0$ for any $i \geq 1$. To bound (6.69), we apply the second inequality in (6.57), and adapt the arguments used to prove (6.61) and (6.63), .

$$\sum_{j=1}^{d} \mathrm{Var}\big[X_j\big]^2 \leq K^2 d \rho(0)^2 \tag{6.70}$$

$$\sum_{i<j} \mathrm{Cov}(X_i, X_j)^2 \leq K^2 d \sum_{n=1}^{d-1} \rho(n)^2 \leq K^2 d \rho_\infty \max_{1 \leq n \leq d-1} \rho(n) \tag{6.71}$$

Since $\sum_{n=0}^{+\infty} \rho(n) \leq \rho_\infty < \infty$, $\{\rho(n)\}_{n \in \mathbb{N}}$ converges to 0 as $n \to +\infty$ and is thus bounded, so $\max_{1 \leq n \leq d-1} \rho(n) < \infty$. We then use (6.70) and (6.71) in the definition of $\beta_2(\mu_d)$, and $\beta_1(\mu_d) \leq \beta_2(\mu_d)$, to derive the upper-bound below for any $d \in \mathbb{N}^*$.

$$\beta_1(\mu_d), \beta_2(\mu_d) \leq d^{1/2} K \big\{\rho(0)^2 + 2\rho_\infty \max_{1 \leq n \leq d-1} \rho(n)\big\}^{1/2} \tag{6.72}$$

□

6.6.7 Setup for synthetic experiments

We explain in more details the setup for the synthetic experiments discussed in Section 6.4, specifically the procedure to generate data. For $d \in \mathbb{N}^*$, we generate $n = 10^4$ i.i.d. realizations of two random variables in $\mathbb{R}^d$, denoted by $X_{1:d} = \{X_j\}_{j=1}^d$ and $Y_{1:d} = \{Y_j\}_{j=1}^d$ and respectively distributed from μ_d, $\nu_d \in \mathcal{P}_2(\mathbb{R}^d)$. The n generated samples of $X_{1:d}$ and $Y_{1:d}$ are respectively denoted by $\{x^{(j)}\}_{j=1}^n$, $\{y^{(j)}\}_{j=1}^n$, with $x^{(j)}, y^{(j)} \in \mathbb{R}^d$ for $j \in \{1, \ldots, n\}$. We approximate SW of order 2 between the empirical distributions of $\{x^{(j)}\}_{j=1}^n$ and $\{y^{(j)}\}_{j=1}^n$, given by $\hat{\mu}_{d,n} = n^{-1} \sum_{j=1}^n \delta_{x^{(j)}}$ and $\hat{\nu}_{d,n} = n^{-1} \sum_{j=1}^n \delta_{y^{(j)}}$ respectively. Note that in Section 6.4, these two distributions were denoted by μ_d, ν_d instead of $\hat{\mu}_{d,n}$, $\hat{\nu}_{d,n}$, to simplify the notation.

We first consider the setting described in Section 6.6.5, where $\mu_d = \mu^{(1)} \otimes \cdots \otimes \mu^{(d)}$ and $\nu_d = \nu^{(1)} \otimes \cdots \otimes \nu^{(d)}$ with $\mu^{(j)}, \nu^{(j)} \in \mathcal{P}_4(\mathbb{R})$ for $j \in \{1, \ldots, d\}$. This means that $\{X_j\}_{j=1}^d$ and $\{Y_j\}_{j=1}^d$ are two sequences of d independent random variables. For each $j \in \{1, \ldots, d\}$, $\mu^{(j)}$ (or $\nu^{(j)}$) refers to a Gaussian or a Gamma distribution, centered or not, as we explain hereafter.

Gaussian distributions (Figure 6.1a). Here, for $j \in \{1, \dots, d\}$, we have $\mu^{(j)} = \mathcal{N}(m_1^{(j)}, \sigma_1^2)$ and $\nu^{(j)} = \mathcal{N}(m_2^{(j)}, \sigma_2^2)$, where $m_1^{(j)}$, $m_2^{(j)}$ are two i.i.d. samples from $\mathcal{N}(1,1)$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 10$. Therefore, $\mu_d = \mathcal{N}(\mathbf{m}_1, \mathbf{I}_d)$ and $\nu_d = \mathcal{N}(\mathbf{m}_2, 10\,\mathbf{I}_d)$, where $\mathbf{I}_d$ denotes the identity matrix of size d, and $\mathbf{m}_1 = \{m_1^{(j)}\}_{j=1}^d$, $\mathbf{m}_2 = \{m_2^{(j)}\}_{j=1}^d \in \mathbb{R}^d$.

We prove that the SW of order 2 between such Gaussian distributions admits a closed-form expression: for any $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^d$ and $\sigma_1^2, \sigma_2^2 > 0$,

$$\mathbf{SW}_2^2\{\mathcal{N}(\mathbf{m}_1, \sigma_1^2\,\mathbf{I}_d), \mathcal{N}(\mathbf{m}_2, \sigma_2^2\,\mathbf{I}_d)\} = \frac{1}{d}\|\mathbf{m}_1 - \mathbf{m}_2\|^2 + (\sigma_1 - \sigma_2)^2 \tag{6.73}$$

Proof of (6.73). First, given the properties of affine transformations of Gaussian random variables, we know that for any $\theta \in \mathbb{S}^{d-1}$, $\mathbf{m} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ symmetric positive-definite, $\theta_\sharp^\star \mathcal{N}(\mathbf{m}, \Sigma)$ is the univariate Gaussian distribution $\mathcal{N}(\langle \theta, \mathbf{m}\rangle, \theta^\intercal \Sigma \theta)$. Using this property in the definition of SW (2.20) and the fact that $\|\theta\| = 1$ for $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
&\mathbf{SW}_2^2\{\mathcal{N}(\mathbf{m}_1, \sigma_1^2\,\mathbf{I}_d), \mathcal{N}(\mathbf{m}_2, \sigma_2^2\,\mathbf{I}_d)\} \\
&= \int_{\mathbb{S}^{d-1}} \mathbf{W}_2^2\{\mathcal{N}(\langle \theta, \mathbf{m}_1\rangle, \sigma_1^2), \mathcal{N}(\langle \theta, \mathbf{m}_2\rangle, \sigma_2^2)\} \mathrm{d}\boldsymbol{\sigma}(\theta) &(6.74)\\
&= \int_{\mathbb{S}^{d-1}} \big\{ \langle \theta, \mathbf{m}_1 - \mathbf{m}_2\rangle^2 + (\sigma_1 - \sigma_2)^2 \big\} \mathrm{d}\boldsymbol{\sigma}(\theta)\,, &(6.75)
\end{aligned}$$

where (6.75) results from the closed-form solution of the Wasserstein distance of order 2 between Gaussian distributions (2.10). Besides, by definition of the Euclidean inner-product, for any $\theta \in \mathbb{S}^{d-1}$,

$$\langle \theta, \mathbf{m}_1 - \mathbf{m}_2\rangle^2 = \big(\theta^\intercal(\mathbf{m}_1 - \mathbf{m}_2)\big)^2 = (\mathbf{m}_1 - \mathbf{m}_2)^\intercal \theta\theta^\intercal (\mathbf{m}_1 - \mathbf{m}_2)\,. \tag{6.76}$$

We can thus rewrite (6.75) to obtain

$$\begin{aligned}
&\mathbf{SW}_2^2\{\mathcal{N}(\mathbf{m}_1, \sigma_1^2\,\mathbf{I}_d), \mathcal{N}(\mathbf{m}_2, \sigma_2^2\,\mathbf{I}_d)\} \\
&= (\mathbf{m}_1 - \mathbf{m}_2)^\intercal \Big\{ \int_{\mathbb{S}^{d-1}} \theta\theta^\intercal \mathrm{d}\boldsymbol{\sigma}(\theta) \Big\} (\mathbf{m}_1 - \mathbf{m}_2) + (\sigma_1 - \sigma_2)^2\,. &(6.77)
\end{aligned}$$

We conclude by using the fact that $\int_{\mathbb{S}^{d-1}} \theta\theta^\intercal \mathrm{d}\boldsymbol{\sigma}(\theta) = (1/d)\mathbf{I}_d$.

□

Gamma distributions (Figure 6.1a). Denote by $\Gamma(k, s)$ the Gamma distribution with shape parameter $k > 0$ and scale $s > 0$. For $j \in \{1, \dots, d\}$, $\mu^{(j)} = \Gamma(k_1^{(j)}, s_1)$ and $\nu^{(j)} = \Gamma(k_2^{(j)}, s_2)$, where $k_1^{(j)}$ (respectively, $k_2^{(j)}$) is drawn from the uniform distribution over $[1, 5)$ (respectively, over $[5, 10)$), $s_1 = 2$ and $s_2 = 3$.

Centered (Gaussian or Gamma) distributions (Figures 6.1b and 6.2). We first generate $\{x^{(j)}\}_{j=1}^n$, $\{y^{(j)}\}_{j=1}^n$ using the Gaussian (or Gamma) distributions described in the two paragraphs above. Then, we center the data: for $j \in \{1, \dots, n\}$, $\bar{x}^{(j)} = x^{(j)} - n^{-1}\sum_{i=1}^n x^{(i)}$ and $\bar{y}^{(j)} = y^{(j)} - n^{-1}\sum_{i=1}^n y^{(i)}$. The two distributions that we compare with SW, referred to as $\bar{\mu}_d, \bar{\nu}_d$ in Section 6.4, correspond to the empirical distributions of the centered datasets $\{\bar{x}^{(j)}\}_{j=1}^n$, $\{\bar{y}^{(j)}\}_{j=1}^n$, which can be denoted by $\bar{\mu}_{d,n}$ and $\bar{\nu}_{d,n}$.

We prove in the next proposition that our theoretical bounds derived in Section 6.6.5 can be improved for centered Gaussian distributions: in this setting, the expected approximation error is upper-bounded by a term in $d^{-1/2}$, which is consistent with the slope observed in Figure 6.1b.

Proposition 6.9. *For $d \in \mathbb{N}^*$, let $\mu_d = \mathcal{N}(\mathbf{m}_1, \sigma_1^2 \mathbf{I}_d)$ and $\nu_d = \mathcal{N}(\mathbf{m}_2, \sigma_2^2 \mathbf{I}_d)$, and denote by $\bar{\mu}_d$, $\bar{\nu}_d$ their centered versions, i.e. $\bar{\mu}_d = \mathcal{N}(\mathbf{0}, \sigma_1^2 \mathbf{I}_d)$ and $\bar{\nu}_d = \mathcal{N}(\mathbf{0}, \sigma_2^2 \mathbf{I}_d)$. Consider the empirical distributions $\bar{\mu}_{d,n}$, $\bar{\nu}_{d,n}$ given by*

$$\bar{\mu}_{d,n} = (1/n) \sum_{j=1}^{n} \delta_{(X_{1:d}^{(j)} - \bar{X}_{1:d})} , \quad \bar{\nu}_{d,n} = (1/n) \sum_{j=1}^{n} \delta_{(Y_{1:d}^{(j)} - \bar{Y}_{1:d})} , \tag{6.78}$$

where $\{X_{1:d}^{(j)}\}_{j=1}^n$ (respectively, $\{Y_{1:d}^{(j)}\}_{j=1}^n$) is a sequence of n random variables i.i.d. from μ_d (respectively, from ν_d), $\bar{X}_{1:d} = n^{-1} \sum_{j=1}^n X_{1:d}^{(j)}$, and $\bar{Y}_{1:d} = n^{-1} \sum_{j=1}^n Y_{1:d}^{(j)}$. Then,

$$\mathbb{E}\left|\mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_{d,n}))\}\right| \leq \frac{\sigma_1 + \sigma_2}{(2dn)^{1/2}} + \mathcal{O}\left(\frac{1}{dn}\right) ,$$

where $\mathbb{E}$ is the expectation with respect to $\{X_{1:d}^{(j)}\}_{j=1}^n$ and $\{Y_{1:d}^{(j)}\}_{j=1}^n$, and $\mathtt{m}_2(\bar{\mu}_{d,n}), \mathtt{m}_2(\bar{\nu}_{d,n})$ are defined in (6.3): $\mathtt{m}_2(\bar{\mu}_{d,n}) = n^{-1} \sum_{j=1}^n \|X_{1:d}^{(j)} - \bar{X}_{1:d}\|^2$, $\mathtt{m}_2(\bar{\nu}_{d,n}) = n^{-1} \sum_{j=1}^n \|Y_{1:d}^{(j)} - \bar{Y}_{1:d}\|^2$.

Proof of Proposition 6.9. Given the closed-form expressions in (6.73) and (2.10), we have

$$\begin{aligned}
&\mathbb{E}\left|\mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d) - \mathbf{W}_2\{\mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})), \mathcal{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_{d,n}))\}\right| \\
&= \mathbb{E}\left||\sigma_1 - \sigma_2| - |d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2} - d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}|\right| \\
&\leq \mathbb{E}\left|\sigma_1 - \sigma_2 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2} + d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}\right| \qquad (6.79) \\
&\leq \mathbb{E}\left|\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}\right| + \mathbb{E}\left|\sigma_2 - d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}\right| . \qquad (6.80)
\end{aligned}$$

where (6.79) results from applying the reverse triangle inequality, and (6.80) follows from the triangle inequality and the linearity of the expectation.

The final result follows from bounding from above the two terms in (6.80), *i.e.* $\mathbb{E}\left|\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}\right|$ and $\mathbb{E}\left|\sigma_2 - d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}\right|$. First, by the Cauchy–Schwarz inequality,

$$\mathbb{E}\left|\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}\right| \leq \left\{\mathbb{E}\left[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2})^2\right]\right\}^{1/2} , \tag{6.81}$$

with

$$\mathbb{E}\left[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2})^2\right] = \sigma_1^2 - 2\sigma_1 d^{-1/2}\mathbb{E}[\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}] + \mathbb{E}[d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})] . \tag{6.82}$$

Consider the random variable defined as $Z = \sqrt{\sum_{i=1}^{dn} \{(X_i - \bar{X})^2/\sigma_1^2\}}$, where $\{X_i\}_{i=1}^{dn}$ are i.i.d. from $\mathcal{N}(0, \sigma_1^2)$ and $\bar{X} = (dn)^{-1} \sum_{i=1}^{dn} X_i$. Then, by Cochran's theorem, Z is distributed from the chi distribution with $dn - 1$ degrees of freedom. This implies that,

$$\begin{aligned}
&\mathbb{E}[d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})] = \sigma_1^2 \, \frac{dn-1}{dn} , \\
&\mathbb{E}[Z] = \sqrt{2} \, \frac{\Gamma(dn/2)}{\Gamma((dn-1)/2)} = \sqrt{dn-1} \left[1 - \frac{1}{4dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right)\right] .
\end{aligned}$$

Hence, (6.82) boils down to

$$\mathbb{E}\left[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\hat{\mu}_{d,n})^{1/2})^2\right] = \sigma_1^2\left[2 - \frac{1}{dn} - 2\left(1-\frac{1}{dn}\right)^{1/2}\left\{1 - \frac{1}{4dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right)\right\}\right] . \tag{6.83}$$

Besides, we know that

$$\left(1-\frac{1}{dn}\right)^{1/2} = 1 - \frac{1}{2dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right) , \tag{6.84}$$

so we can write (6.83) as

$$\mathbb{E}\left[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\hat{\mu}_{d,n})^{1/2})^2\right] = \frac{\sigma_1^2}{2dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right) . \tag{6.85}$$

By plugging (6.85) in (6.81), we conclude that

$$\mathbb{E}\left|\sigma_1 - d^{-1/2}\mathtt{m}_2(\hat{\mu}_{d,n})^{1/2}\right| \leq \frac{\sigma_1}{(2dn)^{1/2}} + \mathcal{O}\left(\frac{1}{dn}\right) . \tag{6.86}$$

We can use the same reasoning to prove that

$$\mathbb{E}\left|\sigma_2 - d^{-1/2}\mathtt{m}_2(\hat{\nu}_{d,n})^{1/2}\right| \leq \frac{\sigma_2}{(2dn)^{1/2}} + \mathcal{O}\left(\frac{1}{dn}\right) , \tag{6.87}$$

and we use (6.86) and (6.87) to bound (6.80), which concludes the proof.

□

We move on to the explanation of autoregressive processes: let $(X_j)_{j\in\mathbb{N}^*}$ be an autoregressive process of order 1 defined as $X_1 = \varepsilon_1$ and for $t \in \mathbb{N}^*$, $t > 1$, $X_t = \alpha X_{t-1} + \varepsilon_t$, where $\alpha \in [0,1)$ and $(\varepsilon_j)_{j\in\mathbb{N}^*}$ is a sequence of i.i.d. real random variables such that $\mathbb{E}[\varepsilon_1] = 0$ and $\mathbb{E}[\varepsilon_1^2] < \infty$.

For $d \in \mathbb{N}^*$ and $B = 10^4$, we generate n realizations of $\{X_j\}_{j=B+1}^{B+d} \in \mathbb{R}^d$ using the aforementionned recursion. This gives us our first dataset $\{x^{(j)}\}_{j=1}^n$, where $x^{(j)} \in \mathbb{R}^d$ for $j \in \{1,\dots,n\}$. Note that the first B steps of the process are discarded in order to reach its stationary regime (which exists since $|\alpha| < 1$), and thus meet the weak dependence condition [Doukhan and Neumann, 2008]. We repeat the same procedure to obtain the second dataset, $\{y^{(j)}\}_{j=1}^n$. Since the two datasets are generated using the same AR(1) model, μ_d and ν_d are the same distribution, so the exact value of SW is zero.

We conducted our experiments on two types of AR(1) processes, which differ from the distribution used to draw n i.i.d. samples of $\{\varepsilon_j\}_{j=1}^{B+d}$. The two settings are specified below.

Gaussian noise (Figure 6.1c). For $j \in \{1,\dots,B+d\}$, $\varepsilon_j \sim \mathcal{N}(0,1)$.

Student's t noise (Figure 6.1d). Denote by $t(r)$ the Student's t distribution with $r > 0$ degrees of freedom. For $j \in \{1,\dots,B+d\}$, $\varepsilon_j \sim t(10)$.

Finally, we specify that the experiment comparing the computation time of our methodology against Monte Carlo estimation (Figure 6.2) was conducted on a daily-use laptop equipped with 8 × Intel Core i7-8650U CPU @ 1.90GHz, 16GB of RAM.

6.6.8 Experimental details for image generation

Architecture. For each model (SWG, reg-SWG or reg-det-SWG), we used the architectures described in [Deshpande et al., 2018]: the "Conv & Deconv" generator and discriminator in [Deshpande et al., 2018, Section D] for MNIST, and DCGAN [Radford et al., 2016] with layernorm for both the generator and discriminator for CelebA.

Data preprocessing. For MNIST, we do not apply any specific preprocessing. For CelebA, each image is cropped at the center and resized to 140×140 (using the notation width $\times$ height, both in pixels), then resized to 64×64.

Optimization. For each model, we used the same optimization routine as in [Deshpande et al., 2018]: one training iteration consists in performing one update for the generator then one update for the discriminator, both with the default setting of Adam [Kingma and Ba, 2015] (*i.e.* $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$). The values of other important hyperparameters are given in Table 6.2.

Dataset	Batch size	Learning rate	Total number of epochs
MNIST	512	5×10^{-4}	200
CelebA	64	1×10^{-4}	20

Table 6.2: Hyperparameters used when training each model.

Regularization parameters. For reg-SWG and reg-det-SWG, we tuned the regularization coefficients (λ_1, λ_2) via cross-validation: we trained the models for $\lambda_1 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and $\lambda_2 \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and selected the tuple that minimizes the average FID over 5 runs.

Computing infrastructure. The FID and computation times on GPU reported in Table 6.1 (columns "FID", "T_{SW}, GPU" and "T_{tot}, GPU") were obtained by training each model on a computer cluster equipped with 3 GPUs (NVIDIA Tesla V100-PCIE-32GB and 2 $\times$ NVIDIA Tesla V100-PCIE-16GB) for CelebA, and with 1 GPU (NVIDIA GP100GL, Tesla P100 PCIe 16GB) for MNIST.

To obtain the computation times on CPU (Table 6.1, columns "T_{SW}, CPU" and "T_{tot}, CPU"), we used a workstation equipped with 24 $\times$ Intel Xeon CPU E5-2620 v3 @ 2.40GHz.

Chapter 7

Statistical and Topological Properties of Sliced Probability Divergences

This chapter is based on [Nadjahi et al., 2020b].

The idea of slicing divergences has been proven to be successful when comparing two probability measures in various machine learning applications including generative modeling, and consists in computing the expected value of a "base divergence" between *one-dimensional random projections* of the two measures. However, the topological, statistical, and computational consequences of this technique have not yet been well-established.

In this chapter, we aim at bridging this gap and derive various theoretical properties of sliced probability divergences. First, we show that slicing preserves the metric axioms and the weak continuity of the divergence, implying that the sliced divergence will share similar topological properties. We then precise the results in the case where the base divergence belongs to the class of integral probability metrics. On the other hand, we establish that, under mild conditions, the sample complexity of a sliced divergence does not depend on the problem dimension. We finally apply our general results to several base divergences, and illustrate our theory on both synthetic and real data experiments.

7.1 Introduction

Most inference methods in implicit generative modeling rely on the use of a particular divergence in order to be able to discriminate probability distributions. Recent advances in this field have illustrated that the choice of this divergence is of crucial importance since it can lead to very different practical and theoretical properties (Chapter 1). In this context, "sliced" probability divergences, such as Sliced-Wasserstein or Sliced-Cramér [Kolouri et al., 2020a], have become increasingly popular.

This slicing strategy has been essentially motivated by two main purposes. The first purpose is that some probability divergences are only defined to compare measures supported on one-dimensional spaces (e.g., Cramér distance, Cramér [1928]); hence, the slicing operation allows the use of such divergences to multivariate distributions [Knop et al., 2020, Kolouri et al., 2020a]. The second purpose arises when the computational complexity of a divergence becomes excessive when comparing measures on

high-dimensional spaces, but can efficiently be computed in the univariate case (e.g., the Wasserstein distance (2.12)). The slicing operation then leverages these advantages originally available in one dimension to define divergences achieving computational efficiency on multivariate settings [Rabin et al., 2012, Deshpande et al., 2019, Paty and Cuturi, 2019, Kolouri et al., 2019b, Vayer et al., 2019].

Even though various sliced divergences have successfully been deployed in practical applications, their theoretical properties have not yet been well understood, and existing results are largely restricted to the specific case of the Sliced-Wasserstein distance. Besides, some properties of SW have only been characterized for specific settings, in particular its statistical benefits observed in practice [Deshpande et al., 2018, 2019]. In this chapter, we aim to bridge this gap by investigating the theoretical properties of sliced probability divergences from a general point of view: since such divergences are all characterized via the same slicing operation, we explore in depth the topological and statistical implications of this operation. Specifically, we consider a generic base divergence $\boldsymbol{\Delta}$ between one-dimensional probability measures, and define its sliced version, denoted by $\mathbf{S}\boldsymbol{\Delta}$, which operates on multivariate settings.

We first establish several topological properties of $\mathbf{S}\boldsymbol{\Delta}$. Thanks to our general approach, our findings can directly be applied to any instance of sliced divergence, including those motivated by the two aforementioned purposes. Specifically, we show that slicing preserves the metric properties: if $\boldsymbol{\Delta}$ is a metric, so is $\mathbf{S}\boldsymbol{\Delta}$ (Proposition 7.1). We then focus on finer topological properties of $\mathbf{S}\boldsymbol{\Delta}$ and show in Theorem 7.2 that, if the convergence in $\boldsymbol{\Delta}$ implies the weak convergence of measures (or conversely), then slicing preserves this property, *i.e.* the convergence in $\mathbf{S}\boldsymbol{\Delta}$ implies the weak convergence of measures (or conversely). We also consider the case when $\boldsymbol{\Delta}$ is an integral probability metric (Definition 2.3) and identify sufficient conditions for $\mathbf{S}\boldsymbol{\Delta}$ to be upper-bounded by $\boldsymbol{\Delta}$, which implies that $\mathbf{S}\boldsymbol{\Delta}$ induces a weaker topology (Theorem 7.4). Similarly, we identify sufficient conditions such that $\boldsymbol{\Delta}$ and $\mathbf{S}\boldsymbol{\Delta}$ are strongly equivalent (Corollary 7.6), meaning that $\boldsymbol{\Delta}$ is upper- and lower-bounded by $\mathbf{S}\boldsymbol{\Delta}$.

Then, we derive the following statistical properties of $\mathbf{S}\boldsymbol{\Delta}$: we prove that the "sample complexity" of $\mathbf{S}\boldsymbol{\Delta}$ is proportional to the sample complexity of $\boldsymbol{\Delta}$ for one-dimensional measures, and does not depend on the dimension d (Theorems 7.7, 7.8). This property explains why *any* $\mathbf{S}\boldsymbol{\Delta}$ motivated by the second purpose offers statistical benefits when the original divergence suffers from the curse of dimensionality. However, this comes with a caveat: we show that, if one approximates the expectation over the random projections that appears in $\mathbf{S}\boldsymbol{\Delta}$ with a Monte Carlo average, which is the most common practice, then an additional variance term appears in the sample complexity and can limit the performance of $\mathbf{S}\boldsymbol{\Delta}$ in high dimensions (Theorem 7.9). Our results agree with the recent empirical observations reported in [Deshpande et al., 2019], which motivated Chapters 5 and 6, and provide a better understanding for them.

We illustrate all our theoretical findings on various examples, which demonstrate their applicability. In particular, our general topological analysis allows us to establish a novel result for the Sliced-Cramér distance. We also derive a sample complexity result for SW which has never been shown before, under different assumptions on the measures to be compared. We then consider Sinkhorn divergences (Definition 2.7), whose sample complexity is known to have an exponential dependence on the dimension d and regularization parameter ε (Theorem 2.8), and introduce its sliced version, referred to as the *Sliced-Sinkhorn divergence*. We prove that this new divergence has several merits: we derive its sample complexity by combining our general results with recent work [Genevay et al., 2019, Mena and Niles-Weed, 2019], and obtain rates that do not depend

on d nor on ε. We also show that this divergence improves the worst-case computational complexity bounds of Sinkhorn divergences in $\mathbb{R}^d$. Finally, we support our theory with numerical experiments on synthetic and real data.

7.2 Sliced Probability Divergences

In this section, we define the family of *Sliced Probability Divergences*, then we present our theoretical contributions regarding their topological and statistical properties. We provide all the proofs in Sections 7.6.1 and 7.6.3.

7.2.1 Definition

Consider a "*base divergence*" $\mathbf{\Delta} : \mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R}) \to \mathbb{R}_+ \cup \{\infty\}$ which measures the dissimilarity between two probability measures on $\mathbb{R}$, and let $p \in [1, \infty)$. We define the *Sliced Probability Divergence of order p* associated to $\mathbf{\Delta}$, denoted by $\mathbf{S\Delta}_p$, for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ as

$$\mathbf{S\Delta}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{\Delta}^p(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) \mathrm{d}\boldsymbol{\sigma}(\theta) \ . \tag{7.1}$$

We assume that $\theta \mapsto \mathbf{\Delta}^p(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu)$ is measurable so that (7.1) is well-defined. This can easily be checked if $(\mu', \nu') \mapsto \mathbf{\Delta}(\mu', \nu')$ is continuous for the weak topology on $\mathcal{P}(\mathbb{R})$, since this implies $\theta \mapsto \mathbf{\Delta}^p(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu)$ is continuous.

In practice, since the integration over $\mathbb{S}^{d-1}$ in (7.1) does not admit an analytical form in general, it is approximated with a simple Monte Carlo scheme (e.g., Section 2.6, Vayer et al. [2019], Kolouri et al. [2020a]). The Monte Carlo estimate of $\mathbf{S\Delta}_p$ obtained with L random projection directions is defined as

$$\widehat{\mathbf{S\Delta}}_{p,L}^p(\mu, \nu) = \frac{1}{L} \sum\nolimits_{l=1}^{L} \mathbf{\Delta}^p\big(\theta^\star_{l\sharp} \mu, \theta^\star_{l\sharp} \nu\big) \ , \tag{7.2}$$

with $\{\theta_l\}_{l=1}^L$ i.i.d. from $\boldsymbol{\sigma}$ and $\theta^\star_l(x) = \langle \theta_l, x \rangle$. Since each term of the sum in (7.2) can be computed independently from each other, the approximation of SPDs can be carried out in parallel, which constitutes a nice practical feature. Recent work [Paty and Cuturi, 2019, Deshpande et al., 2019] has shown that sampling many projection directions uniformly on the sphere might not be the best strategy, in the sense that some directions can be more helpful than others to discriminate the two distributions at hand. However, the Monte Carlo estimate based on uniform sampling (7.2) is the most common method used in practice to approximate sliced divergences, hence we focus on this approximation throughout the rest of the chapter.

7.2.2 Topological properties

We provide several results to describe the topology induced by SPDs, given the properties of base divergences. We first relate in Proposition 7.1 the metric properties of $\mathbf{\Delta}$ and $\mathbf{S\Delta}_p$, $p \in [1, \infty)$.

Proposition 7.1. *(i) If $\mathbf{\Delta}$ is non-negative (or symmetric), then $\mathbf{S\Delta}_p$ is non-negative (symmetric resp.).*

(ii) If $\mathbf{\Delta}$ satisfies for $\mu', \nu' \in \mathcal{P}(\mathbb{R})$, $\mathbf{\Delta}(\mu', \nu') = 0$ if and only if $\mu' = \nu'$, then for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\mathbf{S\Delta}_p(\mu, \nu) = 0$ if and only if $\mu = \nu$.

(iii) If $\boldsymbol{\Delta}$ is a metric, then $\mathbf{S}\boldsymbol{\Delta}_p$ is a metric.

Next, we extend our result stated in Theorem 3.1, which showed that the convergence under SW implies the weak convergence of probability measures: we prove that this property holds for the general class of SPDs, but also that the converse implication is true, provided that $\boldsymbol{\Delta}$ is weakly continuous. We refer to Section 2.1 for the definitions of convergence under a probability divergence, weak convergence of probability measures, and weak continuity.

Theorem 7.2. *Let $p \in [1, \infty)$ and $\boldsymbol{\Delta}$ be a non-negative base divergence.*

(i) If the convergence under $\boldsymbol{\Delta}$ implies the weak convergence in $\mathcal{P}(\mathbb{R})$, then the convergence under $\mathbf{S}\boldsymbol{\Delta}_p$ implies the weak convergence in $\mathcal{P}(\mathbb{R}^d)$.

(ii) If $\boldsymbol{\Delta}$ is bounded and the weak convergence in $\mathcal{P}(\mathbb{R})$ implies the convergence under $\boldsymbol{\Delta}$, then the weak convergence in $\mathcal{P}(\mathbb{R}^d)$ implies the convergence under $\mathbf{S}\boldsymbol{\Delta}_p$.

We now focus on IPMs and formally define Sliced-IPMs, before providing finer topological results. We introduce the following notations. Let X be a closed and measurable subset of $\mathbb{R}^d$. $\mathbb{M}(\mathsf{X})$ denotes the set of real-valued measurable functions on X, $\mathbb{M}_b(\mathsf{X})$ is the set of bounded functions of $\mathbb{M}(\mathsf{X})$, and $\mathrm{B}_d(\mathbf{0}, R) = \left\{x \in \mathbb{R}^d \,:\, \|x\| < R\right\}$ is the open ball in $\mathbb{R}^d$ of radius $R > 0$ centered around $\mathbf{0} \in \mathbb{R}^d$.

Definition 7.3. *Let $\widetilde{\mathsf{F}} \subset \mathbb{M}_b(\mathbb{R})$ and $p \in [1, \infty)$. The Sliced Integral Probability Metric of order p associated with $\widetilde{\mathsf{F}}$, denoted by $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}$, is defined for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ as*

$$(\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p})^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \boldsymbol{\gamma}^p_{\widetilde{\mathsf{F}}}(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) \mathrm{d}\boldsymbol{\sigma}(\theta) \; .$$

Since $\gamma_{\widetilde{\mathsf{F}}}$ is a pseudo-metric, $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}$ is a pseudo-metric as well by Proposition 7.1. We now identify some regularity conditions on the function classes F and $\widetilde{\mathsf{F}}$ such that we are able to show that Sliced-IPMs can be bounded above and below by IPMs. Note that for the next results, we will assume that the supremum in (2.1) is attained. This property is for example verified for $\mathbf{W}_1$ and MMD, by [Villani, 2008] and [Gretton et al., 2012] respectively.

Theorem 7.4. *Let $\widetilde{\mathsf{F}} \subset \mathbb{M}_b(\mathbb{R})$, $\mathsf{F} \subset \mathbb{M}_b(\mathbb{R}^d)$, and assume that*

$$\left\{f : \mathbb{R}^d \to \mathbb{R} \,:\, f = \tilde{f} \circ \theta^\star, \text{ with } \tilde{f} \in \widetilde{\mathsf{F}}, \theta \in \mathbb{S}^{d-1}\right\} \subset \mathsf{F} \; .$$

Then, for any $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}(\mu, \nu) \leq \boldsymbol{\gamma}_{\mathsf{F}}(\mu, \nu)$.

Theorem 7.4 states that $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}$ induces a weaker topology, which is computationally beneficial as argued in Arjovsky et al. [2017], but also indicates that $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}$ comes with less discriminative power, which can be restrictive for hypothesis testing applications such as in [Gretton et al., 2012].

We now derive a lower-bound on compact domains.

Theorem 7.5. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, with support included in $\mathrm{B}_d(\mathbf{0}, R)$. Let $\mathsf{G} \subset \mathbb{M}_b(\mathbb{R}^d)$ and suppose that there exists $\mathsf{L} \geq 0$ such that for any $g \in \mathsf{G}$, g is L-Lipschitz continuous. Consider a class of functions $\widetilde{\mathsf{G}}$ satisfying*

$$\begin{aligned}\widetilde{\mathsf{G}} \quad \supset \quad \{\tilde{g} : \mathbb{R} \to \mathbb{R} \,:\, &\text{there exist } x \in \mathbb{R}^d,\ \theta \in \mathbb{S}^{d-1},\ g \in \mathsf{G} \\ &\text{such that } \tilde{g}(t) = g(x - \theta t) \text{ for any } t \in \mathbb{R}\} \; .\end{aligned}$$

Furthermore, suppose that $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}$ *is bounded. Then, for any* $p \in [1, +\infty)$*, there exists* $C_p > 0$ *such that*

$$\boldsymbol{\gamma}_{\mathsf{G}}(\mu, \nu) \leq C_p \, \mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu, \nu)^{1/(d+1)} \, . \tag{7.3}$$

One can show that the exponent $1/(d+1)$ in (7.3) is intrinsic to slicing, hence cannot be avoided. By combining Theorems 7.4 and 7.5, we finally establish a strong equivalence result below, which implies that the convergence of probability measures in $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}$ is equivalent to the convergence in $\boldsymbol{\gamma}_{\mathsf{G}}$.

Corollary 7.6. *Let* $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$*, with support included in* $\mathrm{B}_d(\mathbf{0}, R)$*, and let* $\mathsf{G} \subset \mathbb{M}_b(\mathbb{R}^d)$*. Assume that the conditions of Theorems 7.4 and 7.5 are satisfied. Then, for any* $p \in [1, +\infty)$*, there exists* $C_p \geq 0$ *independent of* μ, ν *such that*

$$\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu, \nu) \leq \boldsymbol{\gamma}_{\mathsf{G}}(\mu, \nu) \leq C_p \, \mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu, \nu)^{1/(d+1)} \, .$$

Our analysis on IPMs builds on [Bonnotte, 2013, Chapter 5.1], which contains analogous results for the Sliced-Wasserstein distance only. The novelty of Theorems 7.4 and 7.5 is the identification of the relationships between the function classes $\widetilde{\mathsf{F}}$,F and $\widetilde{\mathsf{G}}$,G, which might provide a useful guideline for practitioners interested in slicing any IPM, and cannot be directly obtained from [Bonnotte, 2013]. We further illustrate these relations in Section 7.6.2 for classical instances of IPMs.

7.2.3 Statistical properties

In most practical applications, we have at hand finite sets of samples drawn from unknown underlying distributions. An important question is then the bound of the error made when approximating a divergence with finitely many samples: given $\mathbf{S}\boldsymbol{\Delta}_p$ and any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, our goal is to quantify the *sample complexity* of $\mathbf{S}\boldsymbol{\Delta}_p$, *i.e.* the convergence rate of $\mathbf{S}\boldsymbol{\Delta}_p(\hat{\mu}_n, \hat{\nu}_n)$ to $\mathbf{S}\boldsymbol{\Delta}_p(\mu, \nu)$ according to n. We show in Theorem 7.7 that the sample complexity of any SPD is proportional to the sample complexity of the base divergence, and more importantly, does not depend on d.

Theorem 7.7. *Let* $p \in [1, \infty)$*. Suppose that* $\boldsymbol{\Delta}^p$ *admits the following sample complexity: for any* $\mu', \nu' \in \mathcal{P}(\mathbb{R})$ *with respective empirical measures* $\hat{\mu}'_n, \hat{\nu}'_n$*,*

$$\mathbb{E}\left|\boldsymbol{\Delta}^p(\mu', \nu') - \boldsymbol{\Delta}^p(\hat{\mu}'_n, \hat{\nu}'_n)\right| \leq \beta(p, n) \, .$$

Then, for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ *with respective empirical measures* $\hat{\mu}_n, \hat{\nu}_n$*, the sample complexity of* $\mathbf{S}\boldsymbol{\Delta}_p$ *is given by*

$$\mathbb{E}\left|\mathbf{S}\boldsymbol{\Delta}_p^p(\mu, \nu) - \mathbf{S}\boldsymbol{\Delta}_p^p(\hat{\mu}_n, \hat{\nu}_n)\right| \leq \beta(p, n) \, .$$

If $\boldsymbol{\Delta}$ is a bounded pseudo-metric and we have a direct control over the convergence rate of empirical measures in $\boldsymbol{\Delta}$, we can further derive the following result.

Theorem 7.8. *Let* $p \in [1, \infty)$*. Assume that for any* $\mu' \in \mathcal{P}(\mathbb{R})$ *with empirical measure* $\hat{\mu}'_n$*,*

$$\mathbb{E}|\boldsymbol{\Delta}^p(\hat{\mu}'_n, \mu')| \leq \alpha(p, n) \, .$$

Then, for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ *with empirical measure* $\hat{\mu}_n$*, we have*

$$\mathbb{E}\left|\mathbf{S}\boldsymbol{\Delta}_p^p(\hat{\mu}_n, \mu)\right| \leq \alpha(p, n) \, .$$

Besides, if $\boldsymbol{\Delta}$ *is non-negative, symmetric, and satisfies the triangle inequality, then*

$$\mathbb{E}\left|\mathbf{S}\boldsymbol{\Delta}_p(\mu, \nu) - \mathbf{S}\boldsymbol{\Delta}_p(\hat{\mu}_n, \hat{\nu}_n)\right| \leq 2 \, \alpha(p, n)^{1/p} \, .$$

So far, our results show that slicing preserves some useful topological properties of the base divergence, and additionally offers a dimension-free sample complexity. On the other hand, slicing results in less discriminant divergences, as we mentioned for IPMs (Theorem 7.4), and in such a case, the improvement in the sample complexity might be less significant. More analysis is required to understand the potential reduction in the discriminative power, and we leave it out of scope of this study.

In practice, SPDs also induce an approximation error due to the Monte Carlo estimate (7.2). We use the term *projection complexity* to refer to the convergence rate of $\widehat{\mathbf{S}\boldsymbol{\Delta}}_{p,L}$ to $\mathbf{S}\boldsymbol{\Delta}_p$ as a function of the number of projections L.

Theorem 7.9. *Let $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Then, the error made with the Monte Carlo estimation of $\mathbf{S}\boldsymbol{\Delta}_p$ can be bounded as follows*

$$\Big\{\mathbb{E}\big|\widehat{\mathbf{S}\boldsymbol{\Delta}}^p_{p,L}(\mu,\nu) - \mathbf{S}\boldsymbol{\Delta}^p_p(\mu,\nu)\big|\Big\}^2 \leq L^{-1}\int_{\mathbb{S}^{d-1}} \big\{\boldsymbol{\Delta}^p(\theta^\star_\sharp\mu, \theta^\star_\sharp\nu) - \mathbf{S}\boldsymbol{\Delta}^p_p(\mu,\nu)\big\}^2 \mathrm{d}\boldsymbol{\sigma}(\theta)\ . \tag{7.4}$$

By definition of $\mathbf{S}\boldsymbol{\Delta}^p_p(\mu,\nu)$, Theorem 7.9 illustrates that the quality of the Monte Carlo estimates is impacted by the number of projections as well as the variance of the evaluations of the base divergence. This behavior has previously been empirically observed in different scenarios [Deshpande et al., 2019, Paty and Cuturi, 2019], and paved the way for the max-sliced distances and our methodologies in Chapters 5 and 6.

We now leverage Theorems 7.7 and 7.9 to derive the *overall complexity* of sliced divergences, *i.e.* the convergence rate of $\widehat{\mathbf{S}\boldsymbol{\Delta}}_p(\hat{\mu}_n, \hat{\nu}_n)$ to $\mathbf{S}\boldsymbol{\Delta}_p(\mu,\nu)$.

Corollary 7.10. *Let $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. Denote by $\hat{\mu}_n$ (respectively, $\hat{\nu}_n$) the empirical distribution computed over a sequence of i.i.d. random variables $X_{1:n} = \{X_k\}_{k=1}^n$ from μ (resp., $Y_{1:n} = \{Y_k\}_{k=1}^n$ from ν). Assume $\boldsymbol{\Delta}^p$ admits the following sample complexity: for any $\mu', \nu' \in \mathcal{P}(\mathbb{R})$ and empirical instantiations $\hat{\mu}'_n, \hat{\nu}'_n$,*

$$\mathbb{E}[\big|\boldsymbol{\Delta}^p(\mu',\nu') - \boldsymbol{\Delta}^p(\hat{\mu}'_n, \hat{\nu}'_n)\big|] \leq \beta(p,n)\ .$$

Then,

$$\mathbb{E}\big[|\widehat{\mathbf{S}\boldsymbol{\Delta}}^p_{p,L}(\hat{\mu}_n,\hat{\nu}_n) - \mathbf{S}\boldsymbol{\Delta}^p_p(\mu,\nu)|\big] \leq \beta(p,n) + L^{-1/2}\ \left[\int_{\mathbb{S}^{d-1}} \mathbb{E}\big[\ \big(\boldsymbol{\Delta}^p(\theta^\star_\sharp\hat{\mu}_n, \theta^\star_\sharp\hat{\nu}_n) - \mathbf{S}\boldsymbol{\Delta}^p_p(\hat{\mu}_n,\hat{\nu}_n)\big)^2\ \big]\mathrm{d}\boldsymbol{\sigma}(\theta)\right]^{1/2}\ ,$$

where $\widehat{\mathbf{S}\boldsymbol{\Delta}}^p_{p,L}(\hat{\mu}_n,\hat{\nu}_n)$ is defined by (7.2), and $\mathbb{E}$ is the expectation w.r.t. $X_{1:n}$, $Y_{1:n}$ and $\{\theta_l\}_{l=1}^L$ i.i.d. from the uniform distribution on $\mathbb{S}^{d-1}$.

Hence, the overall complexity $\big|\widehat{\mathbf{S}\boldsymbol{\Delta}}_{p,L}(\hat{\mu}_n,\hat{\nu}_n) - \mathbf{S}\boldsymbol{\Delta}_p(\mu,\nu)\big|$ is bounded by the sum of the sample and the projection complexities. This result is useful as it helps understanding the behavior of sliced divergences in most practical applications, where $\mathbf{S}\boldsymbol{\Delta}_p(\mu,\nu)$ is approximated using finite sets of samples drawn from μ and ν along with Monte Carlo estimates.

7.3 Applications

We already referred to Section 7.6.2, which contains applications of Theorems 7.4 and 7.5 to classical instances of IPMs in order to clarify the assumptions of these theorems. In this section, we apply the rest of our topological and statistical results to specific sliced divergences and present the interesting properties that we obtained. Our goal is to further illustrate the significance of the general theoretical analysis that we conducted in the previous section. In particular, we will introduce a novel divergence based on Sinkhorn divergences, and provide theoretical results that emphasize its statistical and computational advantages. The associated proofs are given in Sections 7.6.4 to 7.6.6.

7.3.1 Topology induced by the Sliced-Cramér distance

First, Theorem 7.2 can be applied to various base divergences (e.g., see those listed in [Gibbs and Su, 2002, Theorem 6]) and foster interesting applications. In particular, we focus on the Cramér distance [Cramér, 1928] and its sliced version [Knop et al., 2020, Kolouri et al., 2020a], whose definitions are recalled in Definitions 7.11 and 7.12 respectively.

Definition 7.11 (Cramér distance)**.** *Let $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Denote by F_μ, F_ν the cumulative distribution functions of μ, ν respectively. The Cramér distance of order p between μ and ν is defined by*

$$\mathbf{C}_p^p(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)|^p \, \mathrm{d}t \, . \tag{7.5}$$

By [Dedecker and Merlevède, 2007, Lemma 1], the Cramér distance can be written as an IPM.

Definition 7.12 (Sliced-Cramér distance)**.** *Let $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. The Sliced-Cramér distance of order p between μ and ν is defined by*

$$\mathbf{SC}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{C}_p^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu) \mathrm{d}\boldsymbol{\sigma}(\theta) \, . \tag{7.6}$$

We establish theoretical guarantees which, to the best of our knowledge, have not been proved before: we show that convergence under the Sliced-Cramér distance implies weak convergence in $\mathcal{P}(\mathbb{R}^d)$, and the converse is true for measures supported on a compact space.

Corollary 7.13. *Let $p \in [1, \infty)$. For any sequence $(\mu_k)_{k \in \mathbb{N}}$ in $\mathcal{P}(\mathbb{R}^d)$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\lim_{k \to \infty} \mathbf{SC}_p(\mu_k, \mu) = 0$ implies $(\mu_k)_{k \in \mathbb{N}}$ converges weakly to μ.*

Besides, if $(\mu_k)_{k \in \mathbb{N}}$ and μ are supported on a compact space $\mathsf{K} \subset \mathbb{R}^d$, then the converse implication holds, meaning that the convergence under $\mathbf{SC}_p$ is equivalent to the weak convergence in $\mathcal{P}(\mathsf{K})$.

Theorem 7.2 also applies to the broader class of Sliced-IPMs, assuming a density property for the space of functions associated with the base IPM. We provide the formal statements and proofs of these results in Section 7.6.4.

7.3.2 Sample complexity of the Sliced-Wasserstein distance

Then, we derive the sample complexity of $\mathbf{SW}_p$ under different moment conditions. While previous works have illustrated the statistical benefits of SW, our next corollary establishes a novel result.

Corollary 7.14. *Let $p \in [1, \infty)$, $q > p$, and $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$ with corresponding empirical measures $\hat{\mu}_n, \hat{\nu}_n$. We use the notation*

$$M_q^{1/q}(\mu, \nu) = M_q^{1/q}(\mu) + M_q^{1/q}(\nu) \ ,$$

where $M_q(\zeta)$ refers to the moment of order q of $\zeta \in \mathcal{P}_q(\mathbb{R}^d)$. Then, there exists a constant $C_{p,q}$ depending on p, q such that

$$\mathbb{E}\big|\mathbf{SW}_p(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SW}_p(\mu, \nu)\big| \leq C_{p,q}^{1/p} M_q^{1/q}(\mu, \nu) \begin{cases} n^{-1/(2p)} & \text{if } q > 2p, \\ n^{-1/(2p)} \log(n)^{1/p} & \text{if } q = 2p, \\ n^{-(q-p)/(pq)} & \text{if } q \in (p, 2p). \end{cases} \tag{7.7}$$

Corollary 7.14 completes the literature on the sample complexity of SW: [Deshpande et al., 2019] derived the sample complexity for Gaussian distributions only and [Manole et al., 2019] provided confidence intervals which partially cover our result.

7.3.3 Sliced-Sinkhorn divergences

We now introduce a new family of probability divergences obtained by slicing the regularized OT cost and Sinkhorn divergences, and called Sliced-Sinkhorn divergences (SSD): for $p \in [1, \infty)$, $\varepsilon \geq 0$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$\begin{aligned} \mathbf{SW}_{p,\varepsilon}(\mu, \nu) &= \int_{\mathbb{S}^{d-1}} \mathbf{W}_{p,\varepsilon}(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) \, \mathrm{d}\boldsymbol{\sigma}(\theta), \\ \overline{\mathbf{SW}}_{p,\varepsilon}(\mu, \nu) &= \int_{\mathbb{S}^{d-1}} \overline{\mathbf{W}}_{p,\varepsilon}(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) \, \mathrm{d}\boldsymbol{\sigma}(\theta) \end{aligned} \tag{7.8}$$

We show that these divergences enjoy interesting statistical and computational properties. For clarity purposes, our results are only presented for $\mathbf{SW}_{p,\varepsilon}$, but also apply for $\overline{\mathbf{SW}}_{p,\varepsilon}$. Since $\mathbf{W}_{p,\varepsilon}$ is not an IPM, we first derive a topological property analogous to Theorem 7.4.

Theorem 7.15. *Let $p \in [1, \infty)$ and $\varepsilon \geq 0$. For any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,*

$$\mathbf{SW}_{p,\varepsilon}(\mu, \nu) \leq \mathbf{W}_{p,\varepsilon}(\mu, \nu) \ .$$

Next, we show that on compact domains, while the sample complexity of regularized OT exponentially worsens as ε decreases [Genevay et al., 2019, Theorem 3], the sample complexity of SSD does not depend on ε.

Theorem 7.16. *Let X be a compact subset of $\mathbb{R}^d$, $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathsf{X})$, with respective empirical instanciations $\hat{\mu}_n, \hat{\nu}_n$. Then, there exists a constant $C(\mu, \nu)$ that depends on the moments of μ and ν, such that*

$$\mathbb{E}\big|\mathbf{SW}_{p,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SW}_{p,\varepsilon}(\mu, \nu)\big| \leq \mathrm{diam}(\mathsf{X}) C(\mu, \nu) n^{-1/2} \ .$$

In practice, we approximate SSD by using (7.2). The estimator corresponds to randomly picking a finite set of directions and solving, for each direction, a regularized OT problem in $\mathbb{R}$. To obtain solutions associated to the regularized Wasserstein cost, a method which is now standard is the Sinkhorn's algorithm: more details are given in Section 2.5 and at the end of Section 7.6.6. In particular, if we use the squared Euclidean ground cost and consider the empirical measures $\hat{\mu}_n, \hat{\nu}_n$ on $\mathbb{R}^d$ associated to the observations $(x_i)_{i=1}^n, (y_j)_{j=1}^n$ respectively, computing $\mathbf{W}_{p,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n)$ has a worst-case convergence rate that depends on

$$C(\hat{\mu}_n, \hat{\nu}_n) = \max_{i,j \in \{1,\dots,n\}} \frac{\|x_i - y_j\|^2}{\varepsilon} .$$

See also [Altschuler et al., 2017] for a sublinear rate with a better constant, still depending on this quantity. The rate for $\mathbf{W}_{p,\varepsilon}(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n)$, with $\theta \in \mathbb{S}^{d-1}$, then depends on $C(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n) = \max_{i,j \in \{1,\dots,n\}} \| \langle \theta, x_i - y_j \rangle \|^2 / \varepsilon$.

We show in Proposition 7.17 that with high probability, $C(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n)$ is smaller than $C(\hat{\mu}_n, \hat{\nu}_n)$ by a factor of d at least, unless n grows super-polynomially with d. Our result, combined with the parallel computation of (7.2), implies that slicing the regularized OT may lead to significant computational benefits.

Proposition 7.17. *Let $(x_i)_{i=1}^n$ be a set of vectors in $\mathbb{R}^d$ such that*

$$\max_{i,j} \|x_i - x_j\|_2^2 \leq R^2 ,$$

and θ chosen uniformly at random on $\mathbb{S}^{d-1}$. Then for $\delta \in (0, 1]$, it holds with probability $1 - \delta$,

$$\max_{i,j} |\langle \theta, x_i - x_j \rangle|^2 \leq \frac{2R^2}{d} \log(\sqrt{2\pi} n^2 / \delta) .$$

Finally, we note that an advantage of the Sinkhorn divergence over the Wasserstein distance is that the former is always differentiable [Feydy et al., 2019, Proposition 2] while the latter is not. This property, which is crucial in differential programming pipelines, suggests that SSD is potentially better-behaved than SW in tasks such as generative modeling. We leave its analysis to future work.

7.4 Experiments

We present the numerical experiments that we conducted to illustrate our theoretical findings, and we provide the code to reproduce them[1]. We also provide additional empirical results in Section 7.6.7.

We first verify that IPMs and Sinkhorn divergences are bounded below by their sliced versions, as demonstrated in Theorems 7.4 and 7.15 respectively. Consider $n = 1000$ observations i.i.d. from $\mathcal{N}(\mathbf{0}, \sigma_\star^2 \mathbf{I}_d)$, with $\sigma_\star^2 = 4$. We generate n i.i.d. samples from $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ for σ^2 varying between 0.1 and 9. We compute MMD between the empirical distributions of the observations and the generated datasets, as well as the Wasserstein distance of order 1 and normalized Sinkhorn divergence (7.8) with order 1 and $\varepsilon = 1$. We used a Gaussian kernel for MMD combined with the heuristic proposed in [Gretton et al., 2012], which sets the kernel width to be the median distance over the aggregated data, and we approximated this discrepancy with the biased estimator in [Gretton et al.,

[1] See our GitHub repository: https://github.com/kimiandj/sliced_div

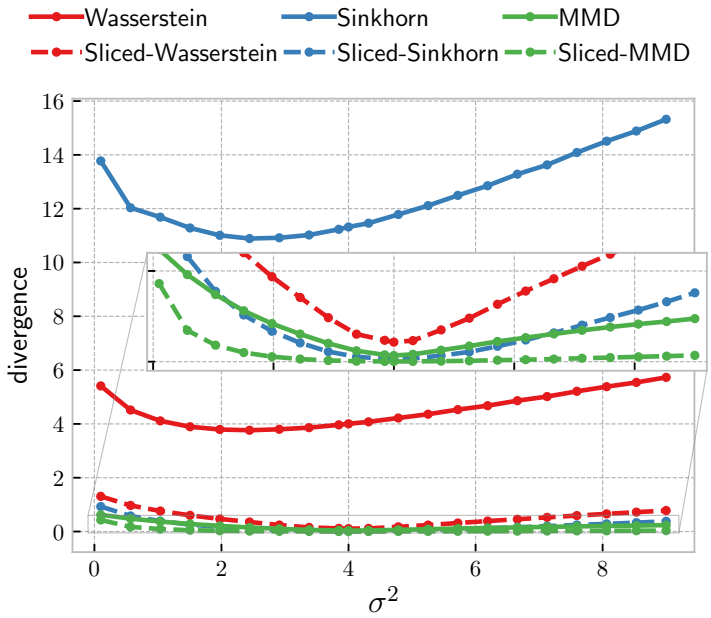


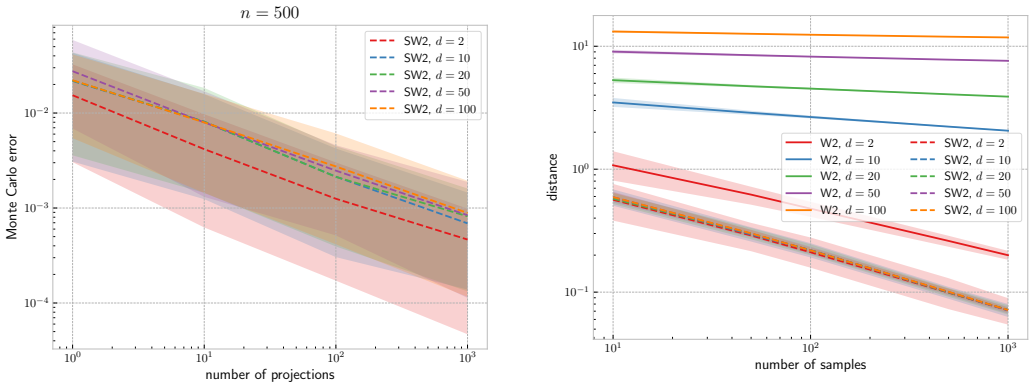
Figure 7.1: (Sliced-)Divergences between two sets of 1000 samples in $\mathbb{R}^{10}$ i.i.d. from $\mathcal{N}(\mathbf{0}, 4\mathbf{I})$ and $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, for varying σ^2.

2012, Equation 5]. Then, we compute Sliced-Wasserstein, Sliced-Sinkhorn and Sliced-MMD. Each of these sliced divergences was approximated with a Monte Carlo estimate based on 50 randomly picked projections. Figure 7.1 reports the divergences against σ^2 for $d = 10$. Results are averaged over 10 runs, and for clarity reasons, we do not plot the error bands (based on the 10th-90th percentiles) as these were very tight.

The curves for Wasserstein, Sinkhorn and MMD are above their respective sliced version's ones, as predicted by our theoretical bounds. This figure also illustrates the statistical benefits induced by slicing: all sliced divergences attain their minimum at $\sigma_\star$, while Wasserstein and Sinkhorn fail at this. This observation is in line with [Bellemare et al., 2017], where the authors showed that both the minimum point and gradients of the Wasserstein distance have a bias, which can be prominent unless n is large enough. MMD performs well in this task, and this can be explained by its dimension-free sample complexity. In that sense, Sliced-MMD acts more as a sanity-check of our theory, rather than a practical proposal.

The next experiments aim at illustrating our statistical properties. We first analyze the convergence rate of the Monte Carlo estimates (Theorem 7.9) in a synthetical setting. We consider two sets of 500 samples i.i.d. from the d-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and we approximate $\mathbf{SW}_2$ between the empirical distributions with a Monte Carlo scheme that uses a high number of projections $L_\star = 10\,000$. Then, we compute the Monte Carlo estimate $\widehat{\mathbf{SW}}_{2,L}$ obtained with $L < L_\star$ random projections. Figure 7.2a shows the absolute difference of $\widehat{\mathbf{SW}}_{2,L}$ and $\widehat{\mathbf{SW}}_{2,L_\star}$ against L, for different values of dimension d. We observe that the Monte Carlo error indeed shrinks to zero when we increase the number of projections, with a convergence rate of order $L^{-1/2}$.

Then, we illustrate the sample complexity of Sliced-Wasserstein and Sliced-Sinkhorn (Corollary 7.14 and Theorem 7.16, respectively). We consider two sets of n samples i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and we compute $\mathbf{W}_2$ and $\overline{\mathbf{W}}_{2,\varepsilon}$ and their sliced versions approxi-



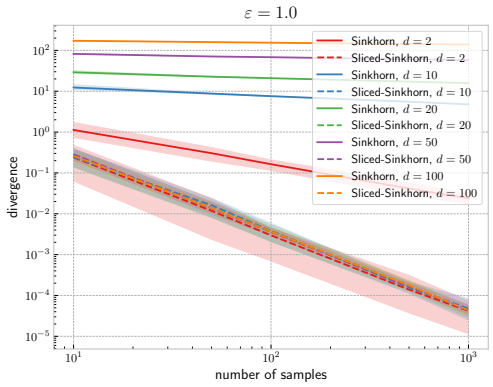
(a) Projection complexity

(b) Sample complexity

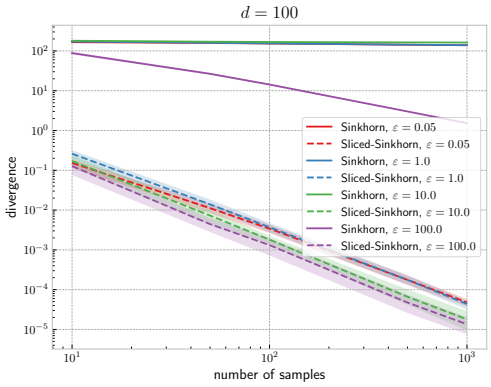
Figure 7.2: (Sliced-)Wasserstein distances of order 2 between two sets of n samples generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for different d, on log-log scale. Results are averaged over 100 runs, and the shaded areas correspond to the 10th-90th percentiles.

mated with 100 random projections. We analyze the convergence rate for different n and dimensions d. We also study the influence of the regularization parameter ε for Sinkhorn divergences. Figure 7.2b reports the Wasserstein and Sliced-Wasserstein distances vs. n, for d between 2 and 100. We observe that, as opposed to $\mathbf{W}_2$, the convergence rate of $\mathbf{SW}_2$ does not depend on the dimension, therefore $\mathbf{SW}_2$ converges faster than $\mathbf{W}_2$ when the dimension increases. Figures 7.3a and 7.3b show Sinkhorn and Sliced-Sinkhorn divergences vs. n, and respectively study the influence of d and ε on the convergence rate. As predicted by the theory, Sliced-Sinkhorn offers more "robustness" than Sinkhorn: its convergence rate does not depend on the dimension nor on the regularization coefficient. To illustrate Proposition 7.17, we plot on Figure 7.3c the number of iterations when the convergence of Sinkhorn's algorithm is reached, as a function of d. For Sliced-Sinkhorn, this number is an average over the number of projections used in the approximation. Our experiment emphasizes the computational advantages of Sliced-Sinkhorn, since its number of iterations remains the same with the increasing dimension, while it grows exponentially for Sinkhorn.

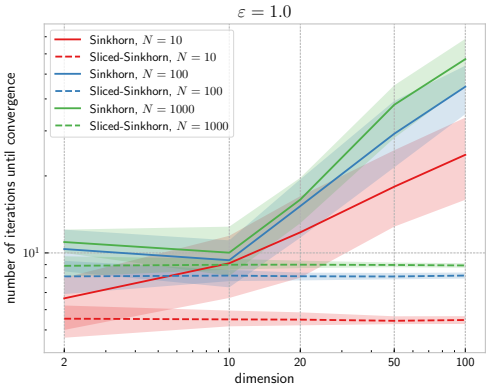
Our last experiment operates on real data and is motivated by the two-sample testing problem [Gretton et al., 2012], whose goal is to determine whether two sets of samples were generated from the same distribution or not. This is useful for various applications, including data integration, where we wish to understand that two datasets were drawn from the same distribution in order to merge them. In this context, we run the following experiment: for different values of n, we randomly select two subsets of n samples from the same dataset, and we compute the Wasserstein and Sliced-Wasserstein distances (of order 2) between the empirical distributions, as well as the Sinkhorn and Sliced-Sinkhorn divergences ($\varepsilon = 1$). The sliced divergences are approximated with 10 random projections. We use the MNIST [LeCun and Cortes, 2010] and CIFAR-10 [Krizhevsky, 2009, Chapter 3] datasets, and we report the divergences against n, and the mean execution time for the computation of Sinkhorn and Sliced-Sinkhorn, on Figure 7.4. The sliced divergences perform the best, in the sense that they need less samples to converge to zero. Besides, Sliced-Sinkhorn is faster than Sinkhorn in terms of execution time (which was expected, given our discussion above Proposition 7.17), and the difference is even more visible for a high number of samples. For example, for $n = 2500$ on MNIST or



(a) Influence of dimension



(b) Influence of regularization



(c) Study of number of iterations

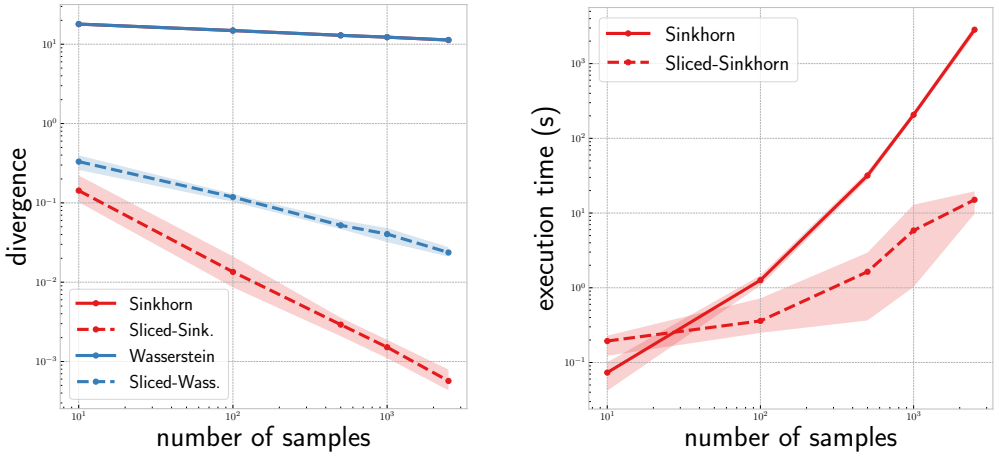
Figure 7.3: (Sliced-)Sinkhorn divergences between two sets of n samples generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for different values of n, dimension d, and regularization coefficient ε. Results are averaged over 100 runs, and the shaded areas correspond to the 10th-90th percentiles. All plots have a log-log scale.

$n = 1000$ on CIFAR-10, Sliced-Sinkhorn is almost 130 times faster than for Sinkhorn on average.

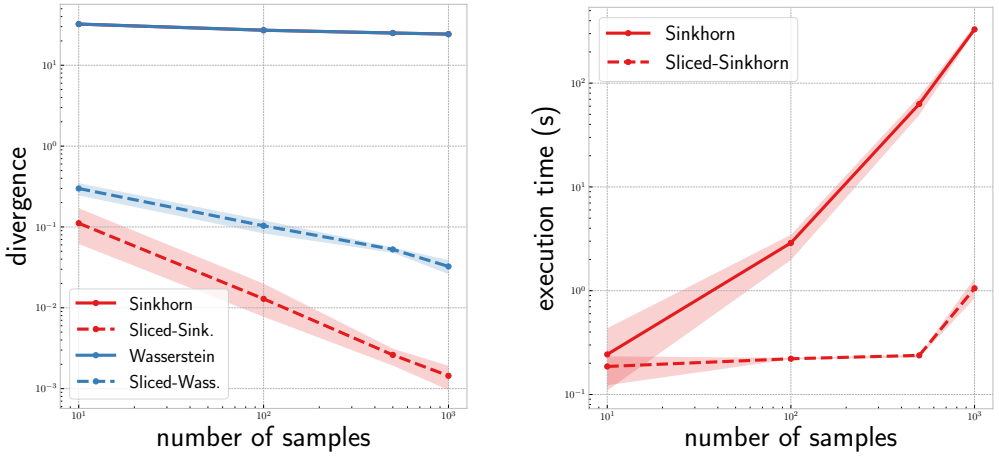
7.5 Conclusion

In this study, we considered sliced probability divergences, which have been increasingly popular in machine learning applications. We derived theoretical results about their induced topology as well as their statistical efficiency in terms of number of samples and projections, and we empirically illustrated our findings on different setups. Specifically, we proved that the preserved topology and dimension-free sample complexity are intrinsic to slicing. Since this was unclear in the previous literature, which combined slicing with a specific distance, our unified treatment of these results brings insight to the properties of particular instances used in practice.

The gains in statistical efficiency could be explained by an ability of slicing to overlook irrelevant characteristics of the distributions. An interesting question for future work is then to understand precisely what geometrical features are well preserved by the slicing operation.



Results on MNIST



Results on CIFAR-10

Figure 7.4: (Sliced-)Wasserstein and (Sliced-)Sinkhorn ($\varepsilon = 1$) between two random subsets of n samples of real datasets (top: MNIST, bottom: CIFAR-10), for different values of n. Results are averaged over 10 runs, and the shaded areas correspond to the 10th-90th percentiles. All plots have a log-log scale.

7.6 Appendix: Postponed proofs and Additional Empirical Results

7.6.1 Proofs for Section 7.2.2

Proof of Proposition 7.1. (i) The fact that $\mathbf{S}\boldsymbol{\Delta}_p$ is non-negative (or symmetric) if $\boldsymbol{\Delta}$ is, immediately follows from the definition of $\mathbf{S}\boldsymbol{\Delta}_p$ (7.1).

(ii) Assume that $\boldsymbol{\Delta}$ satisfies the identity of indiscernibles, *i.e.* for $\mu', \nu' \in \mathcal{P}(\mathbb{R})$, $\boldsymbol{\Delta}(\mu', \nu') = 0$ if and only if $\mu' = \nu'$. For any $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$, $\boldsymbol{\Delta}(\theta^\star_\sharp \mu, \theta^\star_\sharp \mu) = 0$, therefore $\mathbf{S}\boldsymbol{\Delta}_p(\mu, \mu) = 0$ by its definition (7.1).

Now, consider μ, $\nu \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathbf{S}\boldsymbol{\Delta}_p(\mu, \nu) = 0$. Then, by the definition of $\mathbf{S}\boldsymbol{\Delta}_p$ (7.1), we have $\boldsymbol{\Delta}(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) = 0$ for $\boldsymbol{\sigma}$-almost every ($\boldsymbol{\sigma}$-a.e.) $\theta \in \mathbb{S}^{d-1}$, therefore $\theta^\star_\sharp \mu = \theta^\star_\sharp \nu$ for $\boldsymbol{\sigma}$-a.e. $\theta \in \mathbb{S}^{d-1}$. Next, we use the same technique as in [Bonnotte,

2013, Proposition 5.1.2]: for any measure $\xi \in \mathcal{P}(\mathbb{R}^s)$ $(s \geq 1)$, $\mathcal{F}[\xi]$ denotes the Fourier transform of ξ and is defined as, for any $w \in \mathbb{R}^s$,

$$\mathcal{F}[\xi](w) = \int_{\mathbb{R}^s} e^{-\mathrm{i}\langle w,x\rangle} \mathrm{d}\xi(x) \; . \tag{7.9}$$

Then, by using (7.10) and the property of push-forward measures, we have for any $t \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$,

$$\mathcal{F}[\theta^\star_\sharp \mu](t) = \int_{\mathbb{R}} e^{-\mathrm{i}tu} \mathrm{d}\theta^\star_\sharp \mu(u) = \int_{\mathbb{R}^d} e^{-\mathrm{i}t\langle \theta,x\rangle} \mathrm{d}\mu(x) = \mathcal{F}[\mu](t\theta) \; . \tag{7.10}$$

Since for $\boldsymbol{\sigma}$-a.e. $\theta \in \mathbb{S}^{d-1}$, $\theta^\star_\sharp \mu = \theta^\star_\sharp \nu$ thus $\mathcal{F}[\theta^\star_\sharp \mu] = \mathcal{F}[\theta^\star_\sharp \nu]$, we obtain $\mathcal{F}[\mu] = \mathcal{F}[\nu]$. By the injectivity of the Fourier transform, we conclude that $\mu = \nu$.

(iii) Suppose $\boldsymbol{\Delta}$ is a metric. Based on the previous results, to show that $\mathbf{S}\boldsymbol{\Delta}_p$ is a metric, all we need to prove here is that it verifies the triangle inequality. Let $\mu, \nu, \xi \in \mathcal{P}(\mathbb{R}^d)$. Using that $\boldsymbol{\Delta}$ satisfies the triangle inequality and the Minkowski inequality in $\mathcal{L}^p(\mathbb{S}^{d-1}, \boldsymbol{\sigma})$, we get

$$\begin{aligned}
\mathbf{S}\boldsymbol{\Delta}_p(\mu,\nu) &= \left\{ \int_{\mathbb{S}^{d-1}} \boldsymbol{\Delta}^p\big(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu\big) \mathrm{d}\boldsymbol{\sigma}(\theta) \right\}^{1/p} \\
&\leq \left\{ \int_{\mathbb{S}^{d-1}} \Big[\boldsymbol{\Delta}\big(\theta^\star_\sharp \mu, \theta^\star_\sharp \xi\big) + \boldsymbol{\Delta}\big(\theta^\star_\sharp \xi, \theta^\star_\sharp \nu\big) \Big]^p \mathrm{d}\boldsymbol{\sigma}(\theta) \right\}^{1/p} \\
&\leq \left\{ \int_{\mathbb{S}^{d-1}} \boldsymbol{\Delta}^p\big(\theta^\star_\sharp \mu, \theta^\star_\sharp \xi\big) \mathrm{d}\boldsymbol{\sigma}(\theta) \right\}^{1/p} + \left\{ \int_{\mathbb{S}^{d-1}} \boldsymbol{\Delta}^p\big(\theta^\star_\sharp \xi, \theta^\star_\sharp \nu\big) \mathrm{d}\boldsymbol{\sigma}(\theta) \right\}^{1/p} \\
&\leq \mathbf{S}\boldsymbol{\Delta}_p(\mu,\xi) + \mathbf{S}\boldsymbol{\Delta}_p(\xi,\nu) \; .
\end{aligned} \tag{7.11}$$

□

For the proof of Theorem 7.2, we start by proving Lemma 7.18 below, which extends Lemma 3.16 to the more general class of Sliced Probability Divergences.

Lemma 7.18. *Consider $(\mu_k)_{k\in\mathbb{N}}$ a sequence in $\mathcal{P}(\mathbb{R}^d)$ satisfying*

$$\lim_{k\to\infty} \mathbf{S}\boldsymbol{\Delta}_1(\mu_k, \mu) = 0 \; ,$$

with $\mu \in \mathcal{P}(\mathbb{R}^d)$, and assume that the convergence in $\boldsymbol{\Delta}$ implies the weak convergence in $\mathcal{P}(\mathbb{R})$. Then, there exists an increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that the subsequence $(\mu_{\phi(k)})_{k\in\mathbb{N}}$ converges weakly to μ.

Proof. We assume that $\lim_{k\to\infty} \mathbf{S}\boldsymbol{\Delta}_1(\mu_k, \mu) = 0$, *i.e.*:

$$\lim_{k\to\infty} \int_{\mathbb{S}^{d-1}} \boldsymbol{\Delta}(\theta^\star_\sharp \mu_k, \theta^\star_\sharp \mu) \mathrm{d}\boldsymbol{\sigma}(\theta) = 0 \tag{7.12}$$

By [Bogachev, 2007, Theorem 2.2.5], (7.12) implies that, there exists an increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that for $\boldsymbol{\sigma}$-a.e. $\theta \in \mathbb{S}^{d-1}$,

$$\lim_{k\to\infty} \boldsymbol{\Delta}(\theta^\star_\sharp \mu_{\phi(k)}, \theta^\star_\sharp \mu) = 0 \; .$$

Since $\boldsymbol{\Delta}$ is assumed to imply weak convergence in $\mathcal{P}(\mathbb{R})$, then, for $\boldsymbol{\sigma}$-a.e. $\theta \in \mathbb{S}^{d-1}$, $(\theta^\star_\sharp \mu_{\phi(k)})_{k\in\mathbb{N}}$ converges weakly to $\theta^\star_\sharp \mu$. By Lévy's characterization [Kallenberg, 1997, Theorem 4.3], we have for $\boldsymbol{\sigma}$-a.e. $\theta \in \mathbb{S}^{d-1}$ and any $s \in \mathbb{R}$,

$$\lim_{k\to\infty} \Phi_{\theta^\star_\sharp \mu_{\phi(k)}}(s) = \Phi_{\theta^\star_\sharp \mu}(s) \;, \tag{7.13}$$

where Φ_ν is the characteristic function of $\nu \in \mathcal{P}(\mathbb{R}^s)$ $(s \geq 1)$ and is defined for any $v \in \mathbb{R}^s$ as, $\Phi_\nu(v) = \int_{\mathbb{R}^s} \mathrm{e}^{\mathrm{i}\langle v,w\rangle} \mathrm{d}\nu(w)$. Therefore, for Lebesgue (Leb)-almost every $z \in \mathbb{R}^d$,

$$\lim_{k\to\infty} \Phi_{\mu_{\phi(k)}}(z) = \Phi_\mu(z) \;. \tag{7.14}$$

We now use (7.14) to show that $(\mu_{\phi(k)})_{k\in\mathbb{N}}$ converges weakly to μ. By [Billingsley, 1999, Problem 1.11, Chapter 1], this boils down to proving that, for any $f : \mathbb{R}^d \to \mathbb{R}$ continuous with compact support,

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu_{\phi(k)}(z) = \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu(z) \;. \tag{7.15}$$

Consider $\sigma > 0$ and a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ with compact support. We introduce the function f_σ defined as: for any $x \in \mathbb{R}^d$,

$$f_\sigma(x) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} f(x-z) \exp\left(-\|z\|^2/(2\sigma^2)\right) \mathrm{d}z = f * g_\sigma(x) \;, \tag{7.16}$$

where $*$ is the convolution product, and g_σ is the density of the d-dimensional Gaussian with zero mean and covariance matrix $\sigma^2 \mathbf{I}_d$. First, we prove that (7.15) holds with f_σ in place of f. The characteristic function associated to a d-dimensional Gaussian random variable G with zero mean and covariance matrix $(1/\sigma^2)\mathbf{I}_d$ is given by: for any $z \in \mathbb{R}^d$, $\mathbb{E}\left[\mathrm{e}^{\mathrm{i}\langle z,G\rangle}\right] = \mathrm{e}^{-\|z\|^2/(2\sigma^2)}$. By plugging this in the definition of f_σ and using Fubini's theorem, we obtain for any $k \in \mathbb{N}$,

$$\begin{aligned}
\int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu_{\phi(k)}(z) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(w) g_\sigma(z-w) \mathrm{d}w \mathrm{d}\mu_{\phi(k)}(z) \\
&= (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(w) \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle z-w,x\rangle} g_{1/\sigma}(x) \mathrm{d}x \mathrm{d}w \mathrm{d}\mu_{\phi(k)}(z) \\
&= (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(w) \mathrm{e}^{-\mathrm{i}\langle w,x\rangle} g_{1/\sigma}(x) \Phi_{\mu_{\phi(k)}}(x) \mathrm{d}x \mathrm{d}w \\
&= (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_{\mu_{\phi(k)}}(x) \mathrm{d}x \;,
\end{aligned} \tag{7.17}$$

where $\mathcal{F}[f](x) = \int_{\mathbb{R}^d} f(w) \mathrm{e}^{-\mathrm{i}\langle w,x\rangle} \mathrm{d}w$ is the Fourier transform of f. Since the support of f is assumed to be compact, $\mathcal{F}[f]$ exists and is bounded by $\int_{\mathbb{R}^d} |f(w)| \, \mathrm{d}w < +\infty$, therefore, for any $k \in \mathbb{N}$ and $x \in \mathbb{R}^d$,

$$\left| \mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_{\mu_{\phi(k)}}(x) \right| \leq g_{1/\sigma}(x) \int_{\mathbb{R}^d} |f(w)| \mathrm{d}w \;. \tag{7.18}$$

We can prove with similar techniques that (7.17) holds with μ in place of $\mu_{\phi(k)}$, *i.e.*

$$\int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu(z) = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_\mu(x) \mathrm{d}x \;. \tag{7.19}$$

Using (7.14), (7.17), (7.19) and Lebesgue's Dominated Convergence Theorem, we obtain:

$$\lim_{k\to\infty} (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_{\mu_{\phi(k)}}(x) \mathrm{d}x \\ = (2\pi\sigma^2)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}[f](x) g_{1/\sigma}(x) \Phi_{\mu}(x) \mathrm{d}x \ ,$$

$$\textit{i.e.,} \ \lim_{k\to\infty} \int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu_{\phi(k)}(z) = \int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu(z) \ . \tag{7.20}$$

We can now prove (7.15): for any $\sigma > 0$,

$$\left| \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu_{\phi(k)}(z) - \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu(z) \right| \tag{7.21}$$

$$\leq 2 \sup_{z\in\mathbb{R}^d} |f(z) - f_\sigma(z)| + \left| \int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu_{\phi(k)}(z) - \int_{\mathbb{R}^d} f_\sigma(z) \mathrm{d}\mu(z) \right| \ . \tag{7.22}$$

By (7.20), we deduce that for any $\sigma > 0$,

$$\limsup_{k\to+\infty} \left| \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu_{\phi(k)}(z) - \int_{\mathbb{R}^d} f(z) \mathrm{d}\mu(z) \right| \leq 2 \sup_{z\in\mathbb{R}^d} |f(z) - f_\sigma(z)| \ , \tag{7.23}$$

and since $\lim_{\sigma\to 0} \sup_{z\in\mathbb{R}^d} |f(z) - f_\sigma(z)| = 0$ [Folland, 1999, Theorem 8.14-b], we conclude that $(\mu_{\phi(k)})_{k\in\mathbb{N}}$ converges weakly to μ.

□

We can now prove Theorem 7.2.

Proof of Theorem 7.2. Let $p \in [1, \infty)$ and $(\mu_k)_{k\in\mathbb{N}}$ be a sequence of probability measures in $\mathcal{P}(\mathbb{R}^d)$.

First, suppose $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to $\mu \in \mathcal{P}(\mathbb{R}^d)$. By the continuous mapping theorem, since for any $\theta \in \mathbb{S}^{d-1}$, $\theta^\star$ is a bounded linear form thus continuous, then $(\theta^\star_\sharp \mu_k)_{k\in\mathbb{N}}$ converges weakly to $\theta^\star_\sharp \mu$. Therefore, according to our assumption on $\mathbf{\Delta}$, for any $\theta \in \mathbb{S}^{d-1}$,

$$\lim_{k\to\infty} \mathbf{\Delta}(\theta^\star_\sharp \mu_k, \theta^\star_\sharp \mu) = 0 \ . \tag{7.24}$$

Besides, $\mathbf{\Delta}$ is assumed to be non-negative and bounded. Hence, there exists $M > 0$ such that, for any $k \in \mathbb{N}$,

$$\mathbf{\Delta}^p(\theta^\star_\sharp \mu_k, \theta^\star_\sharp \mu) \leq M \ . \tag{7.25}$$

Using (7.24), (7.25) and the bounded convergence theorem, we obtain

$$\lim_{k\to\infty} \mathbf{S}\mathbf{\Delta}_p^p(\mu_k, \mu) = \lim_{k\to\infty} \int_{\mathbb{S}^{d-1}} \mathbf{\Delta}^p(\theta^\star_\sharp \mu_k, \theta^\star_\sharp \mu) \mathrm{d}\boldsymbol{\sigma}(\theta) = \int_{\mathbb{S}^{d-1}} 0^p \ \mathrm{d}\boldsymbol{\sigma}(\theta) = 0 \ . \tag{7.26}$$

Since the mapping $t \mapsto t^{1/p}$ is continuous on $\mathbb{R}+$ (and can be applied to $\mathbf{S}\mathbf{\Delta}_p^p$, which is non-negative by the non-negativity of $\mathbf{\Delta}$ and Proposition 7.1), then (7.26) implies $\lim_{k\to\infty} \mathbf{S}\mathbf{\Delta}_p(\mu_k, \mu) = 0$.

Now, let us prove the other implication, *i.e.* $\lim_{k\to\infty} \mathbf{S}\mathbf{\Delta}_p\big(\mu_k, \mu\big) = 0$ implies the weak convergence of $(\mu_k)_{k\in\mathbb{N}}$ to μ, given the assumptions on $\mathbf{\Delta}$. This result is a generalization

of Theorem 3.1, and is proved analogously, using Lemma 7.18: consider $(\mu_k)_{k\in\mathbb{N}}$ and μ in $\mathcal{P}(\mathbb{R}^d)$ such that

$$\lim_{k\to\infty} \mathbf{S}\boldsymbol{\Delta}_p(\mu_k, \mu) = 0 \,, \tag{7.27}$$

and suppose $(\mu_k)_{k\in\mathbb{N}}$ does not converge weakly to μ. Then, $\lim_{k\to\infty} \mathbf{d}_{\mathcal{P}}(\mu_k, \mu) \neq 0$, where $\mathbf{d}_{\mathcal{P}}$ is the Lévy-Prokhorov metric, *i.e.* there exists $\epsilon > 0$ and a subsequence $(\mu_{\psi(k)})_{k\in\mathbb{N}}$ with $\psi : \mathbb{N} \to \mathbb{N}$ increasing, such that for any $k \in \mathbb{N}$,

$$\mathbf{d}_{\mathcal{P}}(\mu_{\psi(k)}, \mu) > \epsilon \,. \tag{7.28}$$

On the other hand, an application of Hölder's inequality on $\mathbb{S}^{d-1}$ gives for any μ, ν in $\mathcal{P}(\mathbb{R}^d)$,

$$\mathbf{S}\boldsymbol{\Delta}_1(\mu, \nu) \leq \mathbf{S}\boldsymbol{\Delta}_p(\mu, \nu) \,. \tag{7.29}$$

Then, by (7.27), $\lim_{k\to\infty} \mathbf{S}\boldsymbol{\Delta}_1(\mu_{\psi(k)}, \mu) = 0$. Since we assume the convergence in $\boldsymbol{\Delta}$ implies the weak convergence in $\mathcal{P}(\mathbb{R})$, Lemma 7.18 gives us: there exists a subsequence $(\mu_{\phi(\psi(k))})_{k\in\mathbb{N}}$ with $\phi : \mathbb{N} \to \mathbb{N}$ increasing such that $(\mu_{\phi(\psi(k))})_{k\in\mathbb{N}}$ converges weakly to μ. This is equivalent to

$$\lim_{k\to\infty} \mathbf{d}_{\mathcal{P}}(\mu_{\phi(\psi(k))}, \mu) = 0 \,,$$

which contradicts (7.28). We conclude that (7.27) implies the weak convergence of $(\mu_k)_{k\in\mathbb{N}}$ to μ.

□

Proof of Theorem 7.4. Let $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.

$$(\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p})^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \boldsymbol{\gamma}^p_{\widetilde{\mathsf{F}}}(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.30}$$

$$= \int_{\mathbb{S}^{d-1}} \left\{ \sup_{\tilde{f}\in\widetilde{\mathsf{F}}} \left| \int_{\mathbb{R}} \tilde{f}(t) \, \mathrm{d}(\theta^\star_\sharp \mu - \theta^\star_\sharp \nu)(t) \right| \right\}^p \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.31}$$

$$= \int_{\mathbb{S}^{d-1}} \left| \int_{\mathbb{R}} \tilde{f}^*(t) \mathrm{d}(\theta^\star_\sharp \mu - \theta^\star_\sharp \nu)(t) \right|^p \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.32}$$

$$= \int_{\mathbb{S}^{d-1}} \left| \int_{\mathbb{R}^d} \tilde{f}^*\big(\theta^\star(x)\big) \mathrm{d}(\mu - \nu)(x) \right|^p \mathrm{d}\boldsymbol{\sigma}(\theta) \,, \tag{7.33}$$

with $\tilde{f}^* = \operatorname{argmax}_{\tilde{f}\in\widetilde{\mathsf{F}}} \left| \int_{\mathbb{R}} \tilde{f}(t) \mathrm{d}\theta^\star_\sharp \mu(t) - \int_{\mathbb{R}} \tilde{f}(t) \mathrm{d}\theta^\star_\sharp \nu(t) \right|$, which is assumed to exist. Note that (7.33) results from applying the property of push-forward measures.

By definition of F, for any $\theta \in \mathbb{S}^{d-1}$, there exists $f^*_\theta \in \mathsf{F}$ such that $f^*_\theta = \tilde{f}^* \circ \theta^\star$. Therefore, we obtain

$$(\mathbf{S}\boldsymbol{\gamma}_{\mathsf{F},p})^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \left| \int_{\mathbb{R}^d} f^*_\theta(x) \mathrm{d}(\mu - \nu)(x) \right|^p \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.34}$$

$$\leq \int_{\mathbb{S}^{d-1}} \left\{ \sup_{f\in\mathsf{F}} \left| \int_{\mathbb{R}^d} f(x) \mathrm{d}(\mu - \nu)(x) \right| \right\}^p \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.35}$$

$$= \boldsymbol{\gamma}^p_{\mathsf{F}}(\mu, \nu) \int_{\mathbb{S}^{d-1}} \mathrm{d}\boldsymbol{\sigma}(\theta) = \boldsymbol{\gamma}^p_{\mathsf{F}}(\mu, \nu) \,, \tag{7.36}$$

which completes the proof.

□

Proof of Theorem 7.5. We start by upper bounding the distance between two regularized measures. Denote by $\mathrm{supp}(\zeta)$ the support of the function ζ. Let $\varphi : \mathbb{R} \to \mathbb{R}_+^*$ be a smooth and even function verifying $\mathrm{supp}(\varphi) \subset [-1, 1]$ and $\int_{\mathbb{R}} \varphi(t)\mathrm{dLeb}_1(t) = 1$. Define

$$\varphi_\lambda(x) = \lambda^{-d}\varphi(\|x\|/\lambda)/\mathcal{A}(\mathbb{S}^{d-1}) ,$$

with $\mathcal{A}(\mathbb{S}^{d-1})$ denoting the surface area of the d-dimensional unit sphere,*i.e.*

$$\mathcal{A}(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma(d/2) ,$$

where Γ is the gamma function. Denote by $\mathcal{F}[f]$ the Fourier transform of any function f defined on $\mathbb{R}^s$ ($s \geq 1$), given for any $x \in \mathbb{R}^s$ by, $\mathcal{F}[f](x) = \int_{\mathbb{R}^s} f(w)e^{-\mathrm{i}\langle w,x\rangle}\mathrm{d}w$. Let $g \in \mathsf{G}$. By the isometry properties of the Fourier transform and the definition of φ_λ, we have

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu_\lambda - \nu_\lambda)(x) = \int_{\mathbb{R}^d} \mathcal{F}[g](w)\left\{\mathcal{F}[\mu](w) - \mathcal{F}[\nu](w)\right\}\mathcal{F}[\varphi](\lambda w)\mathrm{d}w , \tag{7.37}$$

where $\mu_\lambda = \mu * \varphi_\lambda$ and $\nu_\lambda = \nu * \varphi_\lambda$. By representing w with its polar coordinates $(r, \theta) \in [0, \infty) \times \mathbb{S}^{d-1}$, we obtain

$$\begin{aligned}&\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu_\lambda - \nu_\lambda)(x)\\ &= \int_{\mathbb{S}^{d-1}}\int_0^\infty \mathcal{F}[g](r\theta)\left\{\mathcal{F}[\mu](r\theta) - \mathcal{F}[\nu](r\theta)\right\}\mathcal{F}[\varphi](\lambda r)r^{d-1}\mathrm{d}r\mathrm{d}\boldsymbol{\sigma}(\theta) .\end{aligned}$$

Since g is a real function, $\mathcal{F}[g]$ is an even function, hence

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu_\lambda - \nu_\lambda)(x) \tag{7.38}$$

$$= \frac{1}{2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} \mathcal{F}[g](r\theta)\left\{\mathcal{F}[\mu](r\theta) - \mathcal{F}[\nu](r\theta)\right\}\mathcal{F}[\varphi](\lambda r)\,|r|^{d-1}\,\mathrm{d}r\mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.39}$$

$$= \frac{1}{2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}} \mathcal{F}[g](r\theta)\left\{\mathcal{F}[\theta_\sharp^\star\mu](r) - \mathcal{F}[\theta_\sharp^\star\nu](r)\right\}\mathcal{F}[\varphi](\lambda r)\,|r|^{d-1}\,\mathrm{d}r\mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.40}$$

$$= \frac{1}{2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\int_{-R}^{R} \mathcal{F}[g](r\theta)e^{-\mathrm{i}ru}\mathrm{d}(\theta_\sharp^\star\mu - \theta_\sharp^\star\nu)(u)\mathcal{F}[\varphi](\lambda r)\,|r|^{d-1}\,\mathrm{d}r\mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.41}$$

$$= \frac{1}{2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\int_{\mathbb{R}^d}\int_{-R}^{R} g(x)e^{-\mathrm{i}r(u+\langle\theta,x\rangle)}\left\{\mathrm{d}(\theta_\sharp^\star\mu - \theta_\sharp^\star\nu)(u)\right\}\mathcal{F}[\varphi](\lambda r)\,|r|^{d-1}\,\mathrm{d}x\mathrm{d}r\mathrm{d}\boldsymbol{\sigma}(\theta) , \tag{7.42}$$

where (7.40) follows from (7.10), (7.41) results from the definition of the Fourier transform and the fact that $u \in [-R, R]$, and in the last line, we used the definition of the Fourier transform and Fubini's theorem. By making the change of variables $x \to x - u\theta$, we obtain

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu_\lambda - \nu_\lambda)(x) \tag{7.43}$$

$$= \frac{1}{2}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\int_{\mathbb{R}^d}\int_{-R}^{R} g(x - u\theta)e^{-\mathrm{i}r\langle\theta,x\rangle}\mathrm{d}(\theta_\sharp^\star\mu - \theta_\sharp^\star\nu)(u)\mathcal{F}[\varphi](\lambda r)\,|r|^{d-1}\,\mathrm{d}x\mathrm{d}r\mathrm{d}\boldsymbol{\sigma}(\theta) . \tag{7.44}$$

Since we assumed $\mathrm{supp}(\mu)$, $\mathrm{supp}(\nu)$ are included in $B_d(\mathbf{0}, R)$, then $\mathrm{supp}(\mu_\lambda)$, $\mathrm{supp}(\mu_\lambda)$ are in $B_d(\mathbf{0}, R+\lambda)$, and the domain of $x \mapsto g(x-u\theta)$ must be contained in $B_d(\mathbf{0}, 2R+\lambda)$. By Fubini's theorem and the definition of $\widetilde{\mathsf{G}}$, we have

$$\left|\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu_\lambda - \nu_\lambda)(x)\right| \tag{7.45}$$

$$\leq \frac{1}{2}\int_{\mathbb{R}}\int_{B_d(\mathbf{0},2R+\lambda)}\int_{\mathbb{S}^{d-1}}\left|\int_{-R}^{R} g(x-u\theta)\mathrm{d}(\theta^\star_\sharp\mu - \theta^\star_\sharp\nu)(u)e^{-\mathrm{i}r\langle\theta,x\rangle}\mathcal{F}[\varphi](\lambda r)\,|r|^{d-1}\right|\mathrm{d}\boldsymbol{\sigma}(\theta)\mathrm{d}x\mathrm{d}r \tag{7.46}$$

$$\leq \frac{1}{2}\int_{\mathbb{R}}\int_{B_d(\mathbf{0},2R+\lambda)}\int_{\mathbb{S}^{d-1}}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}}}(\theta^\star_\sharp\mu, \theta^\star_\sharp\nu)\left|e^{-\mathrm{i}r\langle\theta,x\rangle}\mathcal{F}[\varphi](\lambda r)\,|r|^{d-1}\right|\mathrm{d}\boldsymbol{\sigma}(\theta)\mathrm{d}x\mathrm{d}r \tag{7.47}$$

$$\leq C(2R+\lambda)^d\int_{\mathbb{S}^{d-1}}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}}}(\theta^\star_\sharp\mu, \theta^\star_\sharp\nu)\mathrm{d}\boldsymbol{\sigma}(\theta)\int_{\mathbb{R}}\lambda^{-d}\left|\mathcal{F}[\varphi](r)\,|r|^{d-1}\right|\mathrm{d}r \tag{7.48}$$

$$\leq C(2R+\lambda)^d\lambda^{-d}\left(\int_{\mathbb{S}^{d-1}}\boldsymbol{\gamma}^p_{\widetilde{\mathsf{G}}}(\theta^\star_\sharp\mu, \theta^\star_\sharp\nu)\mathrm{d}\boldsymbol{\sigma}(\theta)\right)^{1/p}\int_{\mathbb{R}}\left|\mathcal{F}[\varphi](r)|r|^{d-1}\right|\mathrm{d}r \tag{7.49}$$

$$\leq C_1(2R+\lambda)^d\lambda^{-d}\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu,\nu)\,, \tag{7.50}$$

where in (7.48), $C > 0$ and does not depend on μ and ν, (7.49) results from applying Hölder's inequality on $\mathbb{S}^{d-1}$ if $p > 1$, and in (7.50), $C_1 = C\int_{\mathbb{R}}\left|\mathcal{F}[\varphi](r)|r|^{d-1}\right|\mathrm{d}r$.

By using the definition of $\boldsymbol{\gamma}_{\mathsf{G}}$ and (7.50), we obtain

$$\boldsymbol{\gamma}_{\mathsf{G}}(\mu_\lambda,\nu_\lambda) = \sup_{g\in\mathsf{G}}\left|\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu_\lambda-\nu_\lambda)(x)\right| \leq C_1(2R+\lambda)^d\lambda^{-d}\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu,\nu)\,. \tag{7.51}$$

We now relate $\boldsymbol{\gamma}_{\mathsf{G}}(\mu_\lambda,\nu_\lambda)$ with $\boldsymbol{\gamma}_{\mathsf{G}}(\mu,\nu)$. We start with the following estimate

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu-\nu)(x) - \boldsymbol{\gamma}_{\mathsf{G}}(\mu_\lambda,\nu_\lambda) \tag{7.52}$$

$$\leq \int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu-\nu)(x) - \int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu_\lambda-\nu_\lambda)(x) \tag{7.53}$$

$$\leq \int_{\mathbb{R}^d}\left|g(x)-(\varphi_\lambda * g)(x)\right|\mathrm{d}\mu(x) + \int_{\mathbb{R}^d}\left|g(x)-(\varphi_\lambda * g)(x)\right|\mathrm{d}\nu(x) \tag{7.54}$$

$$\tag{7.55}$$

Since we assumed any $g \in \mathsf{G}$ is $\mathtt{L}$-Lipschitz continuous, we can bound the integrand in (7.54) as follows: for $x \in \mathbb{R}^d$,

$$\left|g(x)-(\varphi_\lambda * g)(x)\right| = \left|\lambda^{-d}\int_{\mathbb{R}^d}\big(g(x)-g(y)\big)\varphi\big((x-y)/\lambda\big)\mathrm{d}y\right| \tag{7.56}$$

$$\leq \lambda^{-d}\int_{\mathbb{R}^d}\left|g(x)-g(y)\right|\varphi\big((x-y)/\lambda\big)\mathrm{d}y \tag{7.57}$$

$$\leq \mathtt{L}\lambda^{-d+1}\int_{\mathbb{R}^d}\|x-y\|\lambda^{-1}\varphi\big((x-y)/\lambda\big)\mathrm{d}y \tag{7.58}$$

$$\leq \mathtt{L}\lambda^{-d+1}\int_{\mathbb{R}^d}\|u\|\lambda^{-1}\varphi\big(u/\lambda\big)\mathrm{d}u \leq \mathtt{L}\lambda\int\|z\|\varphi(z)\mathrm{d}z\,. \tag{7.59}$$

Hence, by denoting by $M_1(\varphi)$ the moment of order 1 of φ, (7.54) is bounded by

$$\int_{\mathbb{R}^d} g(x)\mathrm{d}(\mu-\nu)(x) - \boldsymbol{\gamma}_{\mathsf{G}}(\mu_\lambda,\nu_\lambda) \leq 2\mathtt{L}M_1(\varphi)\lambda\,. \tag{7.60}$$

Taking the supremum of both sides over G gives us

$$\boldsymbol{\gamma}_{\mathsf{G}}(\mu,\nu) - \boldsymbol{\gamma}_{\mathsf{G}}(\mu_\lambda,\nu_\lambda) \leq 2\mathrm{L}M_1(\varphi)\lambda\,. \tag{7.61}$$

By combining the above inequality with (7.51), we get

$$\begin{aligned}\boldsymbol{\gamma}_{\mathsf{G}}(\mu,\nu) &\leq C_1(2R+\lambda)^d\lambda^{-d}\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu,\nu) + 2\mathrm{L}M_1(\varphi)\lambda &(7.62)\\ &\leq C_2\lambda\Big((2R+\lambda)^d\lambda^{-(d+1)}\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu,\nu)+1\Big)\,, &(7.63)\end{aligned}$$

with C_2 satisfying $C_2 \geq C_1$ and $C_2 \geq 2\mathrm{L}M_1(\varphi)$. Finally, by choosing

$$\lambda = R^{d/(d+1)}\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu,\nu)^{1/(d+1)}\,,$$

and using the hypothesis that $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}$ is bounded, we obtain

$$\begin{aligned}\boldsymbol{\gamma}_{\mathsf{G}}(\mu,\nu) &\leq C_2R^{d/(d+1)}\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu,\nu)^{1/(d+1)}\Big((2R+\lambda)^dR^{-d}+1\Big) &(7.64)\\ &\leq C_p\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{G}},p}(\mu,\nu)^{1/(d+1)}, &(7.65)\end{aligned}$$

for some $C_p > 0$, as desired. This concludes the proof.

□

Proof of Corollary 7.6. The desired result is obtained as a direct application of Theorems 7.4 and 7.5.

□

7.6.2 Application of Theorems 7.4 and 7.5

In order to illustrate their assumptions, we apply Theorems 7.4 and 7.5 to well-known instances of IPMs given below. Note that some of these IPMs were already presented in Section 2.2: we recall their definitions in this chapter for completeness.

(1) *Wasserstein distance of order 1.* By the Monge Kantorovich duality theorem [Villani, 2008, Theorem 5.10], when $\mathsf{F} = \{f : \mathsf{Y} \to \mathbb{R} : \|f\|_{\mathrm{Lip}} \leq 1\}$, where $\|f\|_{\mathrm{Lip}} = \sup_{x,y\in\mathsf{Y},x\neq y}\{|f(x)-f(y)|\,/\,\|x-y\|\}$, γ_{F} is the Wasserstein distance of order 1, denoted by $\mathbf{W}_1$.

(2) *Maximum mean discrepancy.* Let H be a reproducing kernel Hilbert space (RKHS) for real-valued functions on Y, and F be the unit ball in H. Then, γ_{F} defines the MMD in RKHS [Gretton et al., 2012, Section 2].

(3) *Total variation distance.* (TV) By choosing $\mathsf{F} = \{f : \mathsf{Y} \to \mathbb{R} : \|f\|_\infty \leq 1\}$, with $\|f\|_\infty = \sup_{x\in\mathsf{Y}}|f(x)|$, γ_{F} corresponds to TV [Douc et al., 2018, Proposition D.2.4].

Informally, the condition on the function classes in Theorem 7.4 requires that F and $\widetilde{\mathsf{F}}$ should be linked to each other in the way that F should be large enough to contain the composition of *all* elements of $\widetilde{\mathsf{F}}$ with *all* possible linear forms $\theta^\star$ for $\theta \in \mathbb{S}^{d-1}$.

Let us illustrate this condition by considering the Wasserstein distance of order 1. In this case, F is the set of 1-Lipschitz functions from $\mathbb{R}^d$ to $\mathbb{R}$, and $\widetilde{\mathsf{F}}$ is the set of 1-Lipschitz functions from $\mathbb{R}$ to $\mathbb{R}$. Then, the condition on F boils down to showing that the composition of any $\tilde{f} \in \widetilde{\mathsf{F}}$ with any linear projection $\theta^\star$ results in a 1-Lipschitz function in $\mathbb{R}^d$, which is simply true since $\tilde{f}$ is 1-Lipschitz and $\|\theta\| = 1$ for all $\theta \in \mathbb{S}^{d-1}$.

In the next three corollaries, we formally prove that Theorem 7.4 holds for the Wasserstein distance of order 1 $\mathbf{W}_1$, total variation distance $\mathbf{TV}$ and maximum mean discrepancy $\mathbf{MMD}$. We denote by $\mathbf{SW}_1$, $\mathbf{STV}_p$ and $\mathbf{SMMD}_p$ the respective sliced versions of these IPMs with order $p \in [1, \infty)$.

Corollary 7.19. *Let $p \in [1, \infty)$. For any $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, $\mathbf{SW}_1(\mu, \nu) \leq \mathbf{W}_1(\mu, \nu)$.*

Proof. Choose $\widetilde{\mathsf{F}} = \{\tilde{f} : \mathbb{R} \to \mathbb{R} \,:\, \|\tilde{f}\|_{\mathrm{Lip}} \leq 1\}$, where

$$\|\tilde{f}\|_{\mathrm{Lip}} = \sup_{x,y \in \mathbb{R}^d, x \neq y} \left\{ \left|\tilde{f}(x) - \tilde{f}(y)\right| / \|x - y\| \right\} .$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ such that $f = \tilde{f} \circ \theta^\star$ with $\tilde{f} \in \widetilde{\mathsf{F}}, \theta \in \mathbb{S}^{d-1}$. Then, by using the Cauchy-Schwarz inequality and the definition of $\widetilde{\mathsf{F}}$, we have for any $x, y \in \mathbb{R}^d$,

$$|f(x) - f(y)| = \left|\tilde{f}\big(\theta^\star(x)\big) - \tilde{f}\big(\theta^\star(y)\big)\right| \leq \left| \langle \theta, x - y \rangle \right| \leq \|\theta^\star\| \, \|x - y\| \leq \|x - y\| \, . \tag{7.66}$$

Therefore, $f \in \mathsf{F} = \{f : \mathbb{R}^d \to \mathbb{R} \,:\, \|f\|_{\mathrm{Lip}} \leq 1\}$. Corollary 7.19 follows from the application of Theorem 7.4 along with the definition of $\mathbf{W}_1$.

□

Note that Corollary 7.19 is not a new result: the fact that $\mathbf{SW}_p$ is bounded above by $\mathbf{W}_p$ for $p \in [1, \infty)$ was established in [Bonnotte, 2013, Proposition 5.1.3]. While their result is proved using the primal formulation of the OT problem, we used the dual formulation available for $p = 1$ to illustrate the applicability of Theorem 7.4. Our result is thus consistent with the existing results in the literature.

Corollary 7.20. *Let $p \in [1, \infty)$. For any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\mathbf{STV}_p(\mu, \nu) \leq \mathbf{TV}(\mu, \nu) \, . \tag{7.67}$$

Proof. Choose $\widetilde{\mathsf{F}} = \left\{\tilde{f} : \mathbb{R} \to \mathbb{R}, \, \|\tilde{f}\|_\infty \leq 1\right\}$, and let $f : \mathbb{R}^d \to \mathbb{R}$ such that $f = \tilde{f} \circ \theta^\star$ with $\tilde{f} \in \widetilde{\mathsf{F}}, \theta \in \mathbb{S}^{d-1}$. Then,

$$\|f\|_\infty = \|\tilde{f} \circ \theta^\star\|_\infty = \sup_{x \in \mathbb{R}^d} \left|\tilde{f}\big(\theta^\star(x)\big)\right| \leq \sup_{t \in \mathbb{R}} \left|\tilde{f}(t)\right| = \|\tilde{f}\|_\infty \leq 1 \, , \tag{7.68}$$

hence, $f \in \mathsf{F} = \left\{f : \mathbb{R}^d \to \mathbb{R} \,:\, \|f\|_\infty \leq 1\right\}$. We obtain the final result by using Theorem 7.4 and the definition of TV.

□

Corollary 7.21. *Let $\widetilde{\mathsf{F}} \subset \mathbb{M}_b(\mathbb{R})$ be the unit ball of the RKHS with reproducing kernel $\tilde{k}$, and k be the positive definite kernel such that for any $x_i, x_j \in \mathbb{R}^d$,*

$$k(x_i, x_j) = \int_{\mathbb{S}^{d-1}} \tilde{k}\big(\theta^\star(x_i), \theta^\star(x_j)\big) \mathrm{d}\boldsymbol{\sigma}(\theta) \, . \tag{7.69}$$

Define $\mathsf{F} \subset \mathbb{M}_b(\mathbb{R}^d)$ as the unit ball of the RKHS whose reproducing kernel $\hat{k}$ satisfies $k - \hat{k}$ is positive definite. Then, for any $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,

$$\mathbf{SMMD}_p(\mu, \nu; \widetilde{\mathsf{F}}) \leq \mathbf{MMD}(\mu, \nu; \mathsf{F}) \, , \tag{7.70}$$

where $\mathbf{MMD}(\cdot, \cdot \,;\, \mathsf{F}')$ and $\mathbf{SMMD}_p(\cdot, \cdot \,;\, \mathsf{F}')$ respectively denote the MMD and the Sliced-MMD of order p in the RKHS whose unit ball is F'.

In particular, this property holds for,

(i) Linear kernels: $\tilde{k}(t_i, t_j) = t_i t_j$ *for* $t_i, t_j \in \mathbb{R}$, *and* $\hat{k}(x_i, x_j) = x_i^\top x_j / d'$ *for* $x_i, x_j \in \mathbb{R}$ *and* $d' \geq d$.

(ii) Radial basis function (RBF) kernels: let $h \geq 0$, $\tilde{k}(t_i, t_j) = e^{-|t_i - t_j|^2/h}$ *for* $t_i, t_j \in \mathbb{R}$, *and* $\hat{k}(x_i, x_j) = e^{-\|x_i - x_j\|^2/h}$ *for* $x_i, x_j \in \mathbb{R}^d$.

Proof. Define $\widetilde{\mathsf{F}}$ as the unit ball of an RKHS whose reproducing kernel is denoted by $\tilde{k}$. Then, any $\tilde{f} \in \widetilde{\mathsf{F}}$ satisfies

$$\|\tilde{f}\|_{\widetilde{\mathsf{F}}}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \tilde{k}(t_i, t_j) \leq 1, \tag{7.71}$$

where $n \in \mathbb{N}^*$, $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $t_1, \ldots, t_n \in \mathbb{R}$.

Consider $f : \mathbb{R}^d \to \mathbb{R}$ such that $f = \tilde{f} \circ \theta^*$ with $\tilde{f} \in \widetilde{\mathsf{F}}$ and $\theta \in \mathbb{S}^{d-1}$. By (7.71), we have

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \tilde{k}\big(\theta^\star(x_i), \theta^\star(x_j)\big) \leq 1 \tag{7.72}$$

The integration of (7.72) over $\mathbb{S}^{d-1}$ gives us

$$\int_{\mathbb{S}^{d-1}} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \tilde{k}\big(\theta^\star(x_i), \theta^\star(x_j)\big) \mathrm{d}\boldsymbol{\sigma}(\theta) \leq \int_{\mathbb{S}^{d-1}} 1 \, \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.73}$$

$$\textit{i.e.,} \quad \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \int_{\mathbb{S}^{d-1}} \tilde{k}\big(\theta^\star(x_i), \theta^\star(x_j)\big) \mathrm{d}\boldsymbol{\sigma}(\theta) \leq 1 \, . \tag{7.74}$$

Define $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $k(x_i, x_j) = \int_{\mathbb{S}^{d-1}} \tilde{k}\big(\theta^\star(x_i), \theta^\star(x_j)\big) \mathrm{d}\boldsymbol{\sigma}(\theta)$ for $x_i, x_j \in \mathbb{R}^d$. Since $\tilde{k}$ is positive definite, so is k. By the Moore-Aronszajn theorem, there exists a unique RKHS with reproducing kernel k. Therefore, (7.74) means that f is in the unit ball of the RKHS associated with k.

Additionally, consider a positive definite kernel $\hat{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $k - \hat{k}$ is positive definite on $\mathbb{R}^d$. In other words, the following holds for any $n \in \mathbb{N}$, $v_1, \ldots, v_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$\sum_{i=1}^n \sum_{j=1}^n v_i v_j \{k(x_i, x_j) - \hat{k}(x_i, x_j)\} \geq 0 \, . \tag{7.75}$$

Then, by (7.74), we obtain $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \hat{k}(x_i, x_j) \leq 1$.

Therefore, any f defined as $f = \tilde{f} \circ \theta$ with $\tilde{f} \in \widetilde{\mathsf{F}}$ and $\theta \in \mathbb{S}^{d-1}$ is in the unit ball of the RKHS associated with $\hat{k}$, which we denote by F. By using Theorem 7.4 and the definition of MMD, we obtain the desired result: for any $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,

$$\mathbf{SMMD}_p(\mu, \nu; \widetilde{\mathsf{F}}) \leq \mathbf{MMD}(\mu, \nu; \mathsf{F}) \, . \tag{7.76}$$

Next, we show that this result holds for two popular choices of kernels. First, we choose $\tilde{k}$ as the linear kernel: $\tilde{k}(t_i, t_j) = t_i t_j$ for $t_i, t_j \in \mathbb{R}$. Define $\hat{k}$ as a rescaled version

of the linear kernel in $\mathbb{R}^d$: $\hat{k}(x_i, x_j) = x_i^\top x_j / d'$ for $x_i, x_j \in \mathbb{R}^d$ and $d' \geq d$. Then, for any $n \in \mathbb{N}$, $v_1, \ldots, v_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathbb{R}^d$,

$$\sum_{i=1}^n \sum_{j=1}^n v_i v_j \{k(x_i, x_j) - \hat{k}(x_i, x_j)\} \tag{7.77}$$

$$= \sum_{i=1}^n \sum_{j=1}^n v_i v_j \Big\{ \int_{\mathbb{S}^{d-1}} \theta(x_i)\theta(x_j) \mathrm{d}\boldsymbol{\sigma}(\theta) - x_i^\top x_j / d' \Big\} \tag{7.78}$$

$$= \sum_{i=1}^n \sum_{j=1}^n v_i v_j \Big\{ x_i^\top \Big(\int_{\mathbb{S}^{d-1}} \theta\theta^\top \mathrm{d}\boldsymbol{\sigma}(\theta) \Big) x_j - x_i^\top x_j / d' \Big\} \tag{7.79}$$

$$= \sum_{i=1}^n \sum_{j=1}^n v_i v_j x_i^\top x_j \Big(1/d - 1/d' \Big) \geq 0 \,, \tag{7.80}$$

where (7.80) results from $\sum_{i=1}^n \sum_{j=1}^n v_i v_j x_i^\top x_j \geq 0$ (the linear kernel is positive definite) and $d' \geq d$. We conclude that (7.76) holds with $\widetilde{\mathsf{F}}$ defined as the unit ball of the RKHS associated with the linear kernel $\tilde{k}(t_i, t_j) = t_i t_j$ for $t_i, t_j \in \mathbb{R}$, and F being the unit ball of the RKHS associated with the rescaled linear kernel $\hat{k}(x_i, x_j) = x_i^\top x_j / d'$ for $x_i, x_j \in \mathbb{R}^d$ and $d' \geq d$.

We focus now on RBF kernels: let $h \geq 0$ and choose $\tilde{k}(t_i, t_j) = e^{-|t_i - t_j|^2/h}$ for $t_i, t_j \in \mathbb{R}$, and $\hat{k}(x_i, x_j) = e^{-\|x_i - x_j\|^2/h}$ for $x_i, x_j \in \mathbb{R}^d$. We have for any $x_i, x_j \in \mathbb{R}^d$,

$$k(x_i, x_j) = \int_{\mathbb{S}^{d-1}} \tilde{k}\big(\theta(x_i), \theta(x_j)\big) \mathrm{d}\boldsymbol{\sigma}(\theta) = \int_{\mathbb{S}^{d-1}} e^{-|\theta^\top x_i - \theta^\top x_j|^2/h} \, \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.81}$$

$$= \int_{\mathbb{S}^{d-1}} e^{-|\theta^\top (x_i - x_j)|^2/h} \, \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.82}$$

$$= \int_{\mathbb{S}^{d-1}} e^{(-\|x_i - x_j\|^2/h)(\theta^\top (x_i - x_j)/\|x_i - x_j\|)^2} \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.83}$$

$$= M\left(\frac{1}{2}, \frac{d}{2}, -\frac{\|x_i - x_j\|^2}{h} \right) \,, \tag{7.84}$$

where $M(a, c, \kappa)$ stands for the confluent hypergeometric function evaluated at $a, c, \kappa \in \mathbb{R}$, and appears in the normalizing constant of the multivariate Watson distribution: see [Sra, 2016, Section 2.3] for more details.

M satisfies the following property,

$$M\left(\frac{1}{2}, \frac{d}{2}, -\frac{\|x_i - x_j\|^2}{h} \right) = e^{-\|x_i - x_j\|^2/h} \, M\left(\frac{d-1}{2}, \frac{d}{2}, \frac{\|x_i - x_j\|^2}{h} \right) \,. \tag{7.85}$$

Since $\|x_i - x_j\|^2/h \geq 0$ and $\kappa \mapsto M(\cdot, \cdot, \kappa)$ is increasing, we have

$$M\left(\frac{d-1}{2}, \frac{d}{2}, \frac{\|x_i - x_j\|^2}{h} \right) \geq M\left(\frac{d-1}{2}, \frac{d}{2}, 0 \right) = M\left(\frac{1}{2}, \frac{d}{2}, 0 \right) = 1 \,. \tag{7.86}$$

Finally, by using (7.84) and (7.85), we obtain: for any $n \in \mathbb{N}$, $v_1, \ldots, v_n \in \mathbb{R}$ and

$x_1, \dots, x_n \in \mathbb{R}^d$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \{k(x_i, x_j) - \hat{k}(x_i, x_j)\} \tag{7.87}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \left[M\left(\frac{1}{2}, \frac{d}{2}, -\frac{\|x_i - x_j\|^2}{h}\right) - e^{-\|x_i - x_j\|^2/h} \right] \tag{7.88}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j e^{-\|x_i - x_j\|^2/h} \left[M\left(\frac{d-1}{2}, \frac{d}{2}, \frac{\|x_i - x_j\|^2}{h}\right) - 1 \right] \tag{7.89}$$

$$\geq 0\,, \tag{7.90}$$

where the last line follows from (7.86) and $\sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j e^{-\|x_i - x_j\|^2/h} \geq 0$ (RBF kernels are positive definite). We conclude that $k - \hat{k}$ is positive definite, hence (7.76) holds for RBF kernels.

□

As with Theorem 7.4, Theorem 7.5 assumes that the function classes G and $\widetilde{\mathsf{G}}$ are linked to each other and sufficiently regular. The condition on G is verified with $\mathbf{W}_1$ (simply by definition) and MMD (provided that the reproducing kernel is Lipschitz-continuous, which holds on compact spaces for classical choices of kernels), but not with TV.

On the other hand, the second condition requires $\widetilde{\mathsf{G}}$ to be large enough to contain *any* possible slice $g(x - u\theta)$ for any $g \in \mathsf{G}$.

7.6.3 Proofs for Section 7.2.3

Proof of Theorem 7.7. Let $p \in [1, \infty)$ and μ, ν in $\mathcal{P}(\mathbb{R}^d)$ with respective empirical measures $\hat{\mu}_n, \hat{\nu}_n$. By using the definition of $\mathbf{S\Delta}_p$, the triangle inequality and the assumption on the sample complexity of $\mathbf{\Delta}^p$, we have

$$\mathbb{E}\left|\mathbf{S\Delta}_p^p(\mu, \nu) - \mathbf{S\Delta}_p^p(\hat{\mu}_n, \hat{\nu}_n)\right|$$

$$= \mathbb{E}\left| \int_{\mathbb{S}^{d-1}} \left\{ \mathbf{\Delta}^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu) - \mathbf{\Delta}^p(\theta_\sharp^\star \hat{\mu}_n, \theta_\sharp^\star \hat{\nu}_n) \right\} \mathrm{d}\boldsymbol{\sigma}(\theta) \right| \tag{7.91}$$

$$\leq \mathbb{E}\left\{ \int_{\mathbb{S}^{d-1}} \left| \mathbf{\Delta}^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu) - \mathbf{\Delta}^p(\theta_\sharp^\star \hat{\mu}_n, \theta_\sharp^\star \hat{\nu}_n) \right| \mathrm{d}\boldsymbol{\sigma}(\theta) \right\} \tag{7.92}$$

$$\leq \int_{\mathbb{S}^{d-1}} \mathbb{E}\left| \mathbf{\Delta}^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu) - \mathbf{\Delta}^p(\theta_\sharp^\star \hat{\mu}_n, \theta_\sharp^\star \hat{\nu}_n) \right| \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.93}$$

$$\leq \int_{\mathbb{S}^{d-1}} \beta(p, n) \mathrm{d}\boldsymbol{\sigma}(\theta) = \beta(p, n)\,, \tag{7.94}$$

which completes the proof.

□

Proof of Theorem 7.8. Let $p \in [1, \infty)$ and $\mu \in \mathcal{P}(\mathbb{R}^d)$ with corresponding empirical measure $\hat{\mu}_n$. By using the definition of $\mathbf{S\Delta}_p$, the triangle inequality and the assumed convergence rate of empirical measures in $\mathbf{\Delta}^p$, we obtain the convergence rate in $\mathbf{S\Delta}_p$

as follows

$$\mathbb{E}\left|\mathbf{S}\mathbf{\Delta}_p^p(\hat{\mu}_n,\mu)\right| = \mathbb{E}\left|\int_{\mathbb{S}^{d-1}} \mathbf{\Delta}^p(\theta_\sharp^\star\hat{\mu}_n,\theta_\sharp^\star\mu)\mathrm{d}\boldsymbol{\sigma}(\theta)\right| \tag{7.95}$$

$$\leq \mathbb{E}\left\{\int_{\mathbb{S}^{d-1}} \left|\mathbf{\Delta}^p(\theta_\sharp^\star\hat{\mu}_n,\theta_\sharp^\star\mu)\right| \mathrm{d}\boldsymbol{\sigma}(\theta)\right\} \tag{7.96}$$

$$\leq \int_{\mathbb{S}^{d-1}} \mathbb{E}\left|\mathbf{\Delta}^p(\theta_\sharp^\star\hat{\mu}_n,\theta_\sharp^\star\mu)\right| \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.97}$$

$$\leq \int_{\mathbb{S}^{d-1}} \alpha(p,n)\mathrm{d}\boldsymbol{\sigma}(\theta) = \alpha(p,n)\,. \tag{7.98}$$

Additionally, if we assume that $\mathbf{\Delta}$ satisfies non-negativity, symmetry and the triangle inequality, then $\mathbf{S}\mathbf{\Delta}_p$ also verifies these three properties by Proposition 7.1, and we can derive its sample complexity: for any μ,ν in $\mathcal{P}(\mathbb{R}^d)$ with respective empirical measures $\hat{\mu}_n,\hat{\nu}_n$, the triangle inequality give us

$$|\mathbf{S}\mathbf{\Delta}_p(\mu,\nu) - \mathbf{S}\mathbf{\Delta}_p(\hat{\mu}_n,\hat{\nu}_n)| \leq \mathbf{S}\mathbf{\Delta}_p(\hat{\mu}_n,\mu) + \mathbf{S}\mathbf{\Delta}_p(\hat{\nu}_n,\nu) \tag{7.99}$$

By taking the expectation of (7.99) with respect to $\hat{\mu}_n,\hat{\nu}_n$, we obtain

$$\begin{aligned}
&\mathbb{E}\,|\mathbf{S}\mathbf{\Delta}_p(\mu,\nu) - \mathbf{S}\mathbf{\Delta}_p(\hat{\mu}_n,\hat{\nu}_n)| \\
&\leq \mathbb{E}\,|\mathbf{S}\mathbf{\Delta}_p(\hat{\mu}_n,\mu)| + \mathbb{E}\,|\mathbf{S}\mathbf{\Delta}_p(\hat{\nu}_n,\nu)| && (7.100)\\
&\leq \left\{\mathbb{E}\left|\mathbf{S}\mathbf{\Delta}_p^p(\hat{\mu}_n,\mu)\right|\right\}^{1/p} + \left\{\mathbb{E}\left|\mathbf{S}\mathbf{\Delta}_p^p(\hat{\nu}_n,\nu)\right|\right\}^{1/p} && (7.101)\\
&\leq \alpha(p,n)^{1/p} + \alpha(p,n)^{1/p} = 2\alpha(p,n)^{1/p}\,, && (7.102)
\end{aligned}$$

where (7.101) results from applying Hölder's inequality on $\mathbb{S}^{d-1}$ if $p>1$, and (7.102) follows from the convergence rate result in (7.98).

□

Proof of Theorem 7.9. Let $p\in[1,\infty)$ and $\mu,\nu\in\mathcal{P}(\mathbb{R}^d)$. We recall that $\widehat{\mathbf{S}\mathbf{\Delta}}_{p,L}(\mu,\nu)$ denotes the approximation of $\mathbf{S}\mathbf{\Delta}_p(\mu,\nu)$ obtained with a Monte Carlo scheme that uniformly picks L projection directions on $\mathbb{S}^{d-1}$ (cf. Equation (7.2)).

By using Hölder's inequality and the results on the moments of the Monte Carlo estimation error, we obtain

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\boldsymbol{\sigma}}\left|\widehat{\mathbf{S}\mathbf{\Delta}}_{p,L}^p(\mu,\nu) - \mathbf{S}\mathbf{\Delta}_p^p(\mu,\nu)\right| \\
&\leq \left\{\mathbb{E}_{\theta\sim\boldsymbol{\sigma}}\left|\widehat{\mathbf{S}\mathbf{\Delta}}_{p,L}^p(\mu,\nu) - \mathbf{S}\mathbf{\Delta}_p^p(\mu,\nu)\right|^2\right\}^{1/2} && (7.103)\\
&\leq L^{-1/2}\left\{\int_{\mathbb{S}^{d-1}} \left\{\mathbf{\Delta}^p(\theta_\sharp^\star\mu,\theta_\sharp^\star\nu) - \mathbf{S}\mathbf{\Delta}_p^p(\mu,\nu)\right\}^2 \mathrm{d}\boldsymbol{\sigma}(\theta)\right\}^{1/2}\,, && (7.104)
\end{aligned}$$

Since $\mathbf{S}\mathbf{\Delta}_p^p(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathbf{\Delta}^p(\theta_\sharp^\star\mu,\theta_\sharp^\star\nu)\mathrm{d}\boldsymbol{\sigma}(\theta)$ by definition, the quantity given by,

$$\int_{\mathbb{S}^{d-1}} \left\{\mathbf{\Delta}^p(\theta_\sharp^\star\mu,\theta_\sharp^\star\nu) - \mathbf{S}\mathbf{\Delta}_p^p(\mu,\nu)\right\}^2 \mathrm{d}\boldsymbol{\sigma}(\theta)$$

is the variance of $\mathbf{\Delta}^p(\theta_\sharp^\star\mu,\theta_\sharp^\star\nu)$ with respect to $\theta\sim\boldsymbol{\sigma}$.

□

Proof of Corollary 7.10. Let $p \in [1, \infty)$, $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and the respective empirical distributions $\hat{\mu}_n, \hat{\nu}_n$. By the triangle inequality,

$$\begin{aligned}
&|\widehat{\mathbf{SA}}^p_{p,L}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\mu, \nu)| \\
&\quad \leq |\widehat{\mathbf{SA}}^p_{p,L}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n)| + |\mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\mu, \nu)| \, .
\end{aligned}$$

Therefore, by linearity of expectation, we have

$$\mathbb{E}\big[|\widehat{\mathbf{SA}}^p_{p,L}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\mu, \nu)|\big] \tag{7.105}$$

$$\leq \mathbb{E}\left[\mathbb{E}[|\widehat{\mathbf{SA}}^p_{p,L}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n)|\big| \, X_{1:n}, Y_{1:n}]\right] \tag{7.106}$$

$$+ \mathbb{E}\left[|\mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\mu, \nu)|\right] \, . \tag{7.107}$$

We bound (7.106): by Theorem 7.9, we have

$$\mathbb{E}\big[|\widehat{\mathbf{SA}}^p_{p,L}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n)| \;\big|\; X_{1:n}, Y_{1:n}\big] \tag{7.108}$$

$$\leq L^{-1/2} \left\{ \int_{\mathbb{S}^{d-1}} \left\{ \boldsymbol{\Delta}^p(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n) \right\}^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \right\}^{1/2} \, . \tag{7.109}$$

By taking the expectation then using Jensen's inequality, we get

$$\mathbb{E}\left[\mathbb{E}\big[|\widehat{\mathbf{SA}}^p_{p,L}(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n)| \;\big|\; X_{1:n}, Y_{1:n}\big]\right] \tag{7.110}$$

$$\leq L^{-1/2} \, \mathbb{E}\left[\left\{ \int_{\mathbb{S}^{d-1}} \left\{ \boldsymbol{\Delta}^p(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n) \right\}^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \right\}^{1/2} \right] \tag{7.111}$$

$$\leq L^{-1/2} \, \mathbb{E}^{1/2}\left[\int_{\mathbb{S}^{d-1}} \left\{ \boldsymbol{\Delta}^p(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n) \right\}^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \right] \, . \tag{7.112}$$

Next, we bound (7.107): by the sample complexity assumption for $\boldsymbol{\Delta}^p$ and Theorem 7.7, we have

$$\mathbb{E}\left[|\mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n) - \mathbf{SA}^p_p(\mu, \nu)|\right] \leq \beta(p, n) \, . \tag{7.113}$$

Combining (7.112) and (7.113) in (7.106) and (7.107) completes the proof.

□

Remark 7.22. *Note that by Fubini's theorem,*

$$\int_{\mathbb{S}^{d-1}} \mathbb{E}[(\boldsymbol{\Delta}^p(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n) - \mathbf{SA}^p_p(\hat{\mu}_n, \hat{\nu}_n))^2] \mathrm{d}\boldsymbol{\sigma}(\theta) \, ,$$

which appears in Corollary 7.10, is equal to $\mathbb{E}[\mathrm{Var}\{\boldsymbol{\Delta}^p(\theta^\star_\sharp \hat{\mu}_n, \theta^\star_\sharp \hat{\nu}_n)|X_{1:n}, Y_{1:n}\}]$, *where* Var *is the variance w.r.t.* $X_{1:n}$, $Y_{1:n}$ *and* θ *(which is distributed according to the uniform distribution on* $\mathbb{S}^{d-1}$ *and independent of* $X_{1:n}, Y_{1:n}$*).*

7.6.4 Proofs for Section 7.3.1

As discussed in Section 7.3.1, we can use the general result in Theorem 7.2 to establish novel topological properties for specific sliced probability divergences, for example the Sliced-Cramér distance (whose definition is recalled in Definition 7.12) and the broader class of Sliced-IPMs. We present our results and proofs for these examples below.

Proof of Corollary 7.13. Let $p \in [1, \infty)$. By Hölder's inequality, for any $\mu', \nu' \in \mathcal{P}(\mathbb{R})$, we have

$$\mathbf{C}_1(\mu', \nu') \leq \mathbf{C}_p(\mu', \nu') \,. \tag{7.114}$$

Consider a sequence $(\mu'_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(\mathbb{R})$ and $\mu' \in \mathcal{P}(\mathbb{R})$ such that

$$\lim_{k\to\infty} \mathbf{C}_p(\mu'_k, \mu') = 0 \,.$$

By (7.114), this implies $\lim_{k\to\infty} \mathbf{C}_1(\mu'_k, \mu') = 0$. Since the Cramér distance of order 1 is equivalent to the Wasserstein distance of order 1, then by [Villani, 2008, Theorem 6.8], the convergence of $(\mu'_k)_{k\in\mathbb{N}}$ to μ' under $\mathbf{C}_p$ implies $(\mu'_k)_{k\in\mathbb{N}}$ converges weakly to μ' in $\mathcal{P}(\mathbb{R})$. By Theorem 7.2, we conclude that the convergence under $\mathbf{SC}_p$ implies the weak convergence in $\mathcal{P}(\mathbb{R}^d)$.

We now show the second part of the statement. This result partly follows from slight modifications of the techniques we used in the proof of Theorem 7.2.

Consider a compact space $\mathsf{K}' \subset \mathbb{R}$ and a sequence $(\mu'_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(\mathsf{K}')$. Suppose that $(\mu'_k)_{k\in\mathbb{N}}$ converges weakly to $\mu' \in \mathcal{P}(\mathsf{K}')$. Since $F_{\mu'}$ is non-decreasing, it is almost everywhere continuous w.r.t. to the Lebesgue convergence, and using the Portmanteau theorem, we get that for Leb-almost every $t \in \mathbb{R}$, $\lim_{k\to\infty} F_{\mu'_k}(t) = F_{\mu'}(t)$. Besides, for any $k \in \mathbb{N}$ and $t \in \mathsf{K}'$, $|F_{\mu'_k}(t)| \leq 1$, and since K' is compact, $\left(\int_{\mathsf{K}'} 1^p \mathrm{d}t\right)^{1/p} < \infty$. By the dominated convergence theorem in L^p-spaces, we conclude that

$$\lim_{k\to\infty} \left\{ \int_{\mathsf{K}'} |F_{\mu'_k}(t) - F_{\mu'}(t)|^p \mathrm{d}t \right\}^{1/p} = 0 \,, \tag{7.115}$$

in other words, the weak convergence of measures in $\mathcal{P}(\mathsf{K}')$, where K' is a compact subspace of $\mathbb{R}$, implies the convergence under $\mathbf{C}_p$.

Now, consider a compact space $\mathsf{K} \subset \mathbb{R}^d$ and a sequence $(\mu_k)_{k\in\mathbb{N}}$ in $\mathcal{P}(\mathsf{K})$ which converges weakly to $\mu \in \mathcal{P}(\mathsf{K})$. For any $\theta \in \mathbb{S}^{d-1}$, define

$$\mathsf{K}_\theta = \{\langle \theta, x\rangle \,:\, x \in \mathsf{K}\} \,,$$

which is a compact subset of $\mathbb{R}$ (since it is the image of K by a continuous function) with $\mathrm{diam}(\mathsf{K}_\theta) \leq \mathrm{diam}(\mathsf{K})$ (by the Cauchy-Schwarz inequality). The sequence of push-forward measures $(\theta^\star_\sharp \mu_k)_{k\in\mathbb{N}}$ is in $\mathcal{P}(\mathsf{K}_\theta)$ and, by the continuous mapping theorem, converges weakly to $\theta^\star_\sharp \mu \in \mathcal{P}(\mathsf{K}_\theta)$. Therefore, by (7.115), for any $\theta \in \mathbb{S}^{d-1}$,

$$\lim_{k\to\infty} \mathbf{C}_p(\theta^\star_\sharp \mu_k, \theta^\star_\sharp \mu) = 0 \,. \tag{7.116}$$

Besides, for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with support in K, and $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned} \mathbf{C}_p(\theta^\star_\sharp \nu, \theta^\star_\sharp \mu) = \int_{\mathbb{R}} |F_\nu(t) - F_\mu(t)|^p \,\mathrm{d}t = \int_{\mathsf{K}_\theta} |F_\nu(t) - F_\mu(t)|^p \,\mathrm{d}t \\ \leq 2^p \mathrm{diam}(\mathsf{K}_\theta) \leq 2^p \mathrm{diam}(\mathsf{K}) \,. \end{aligned} \tag{7.117}$$

By (7.116) and the dominated convergence theorem, $\lim_{k\to\infty} \mathbf{SC}_p(\mu_k, \mu) = 0$.

□

Corollary 7.23. *Let $p \in [1, \infty)$ and $\widetilde{\mathsf{F}} \subset \mathbb{M}_b(\mathbb{R})$. Suppose that the space spanned by $\widetilde{\mathsf{F}}$ is dense in the space of continous functions for $\|\cdot\|_\infty$. Then, the convergence under the Sliced Integral Probability Metric of order p associated with $\widetilde{\mathsf{F}}$, $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}$, implies the weak convergence in $\mathcal{P}(\mathbb{R}^d)$. Besides, if $\boldsymbol{\gamma}_{\widetilde{\mathsf{F}}}$ is bounded, the converge implication holds, i.e. the weak convergence in $\mathcal{P}(\mathbb{R}^d)$ implies the convergence under $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}$.*

Proof. By construction of $\widetilde{\mathsf{F}}$ and [Ambrosio et al., 2005, Section 5.1], $\boldsymbol{\gamma}_{\widetilde{\mathsf{F}}}$ metrizes the weak convergence in $\mathcal{P}(\mathbb{R})$, *i.e.* the weak convergence in $\mathcal{P}(\mathbb{R})$ is equivalent to the convergence of measures under $\boldsymbol{\gamma}_{\widetilde{\mathsf{F}}}$. The properties of $\mathbf{S}\boldsymbol{\gamma}_{\widetilde{\mathsf{F}},p}$, $p \in [1,\infty)$ result from the application of Theorem 7.2.

□

Remark 7.24. *The boundedness assumption for $\boldsymbol{\gamma}_{\widetilde{\mathsf{F}}}$ is achieved if we additionally suppose that $\widetilde{\mathsf{F}}$ is a uniformly bounded family of functions in $\mathbb{M}(\mathbb{R})$, which is a mild assumption.*

7.6.5 Proof of Corollary 7.14

Lemma 7.25. *Let $p \in [1,\infty)$ and $\mu' \in \mathcal{P}(\mathbb{R})$ with empirical distribution $\hat{\mu}'_n$. Suppose there exists $q > p$ such that the moment of order q of μ', defined as $M_q(\mu') = \int_{\mathbb{R}} |t|^q \, \mathrm{d}\mu'(t)$, is bounded above by $K < \infty$. Then, there exists a constant $C_{p,q}$ depending on p, q such that*

$$\mathbb{E}\left[\mathbf{W}_p^p(\hat{\mu}'_n, \mu')\right] \leq C_{p,q} K \begin{cases} n^{-1/2} & \text{if } q > 2p, \\ n^{-1/2}\log(n) & \text{if } q = 2p, \\ n^{-(q-p)/q} & \text{if } q \in (p, 2p). \end{cases} \tag{7.118}$$

Proof. This immediately results from [Fournier and Guillin, 2015, Theorem 1]. □

Proof of Corollary 7.14. We first recall that, for any $\xi \in \mathcal{P}(\mathbb{R}^s)$ $(s \geq 1)$ and $\theta \in \mathbb{S}^{d-1}$, the moment of order $k > 0$ of $\theta^\star_\sharp \xi$ is lower than the one associated with ξ. Indeed, by using the property of push-forward measures, the Cauchy-Schwarz inequality, and $\|\theta\| \leq 1$, we have

$$M_k(\theta^\star_\sharp \xi) = \int_{\mathbb{R}} |t|^k \, \mathrm{d}\theta^\star_\sharp \xi(t) = \int_{\mathbb{R}^d} |\langle \theta, x \rangle|^k \, \mathrm{d}\xi(x) \leq \int_{\mathbb{R}^d} \|x\|^k \, \mathrm{d}\xi(x) = M_k(\xi) \, . \tag{7.119}$$

Now, let $p \in [1,\infty)$ and $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ $(q > p)$ with empirical distribution $\hat{\mu}_n$. Then, by (7.119), for any $\theta \in \mathbb{S}^{d-1}$, $M_q(\theta^\star_\sharp \mu) \leq M_q(\mu) < \infty$, and we can apply Lemma 7.25 and Theorem 7.8 to derive the convergence rate under $\mathbf{SW}_p$: there exists a constant $C_{p,q}$ such that,

$$\mathbb{E}\left[\mathbf{SW}_p^p(\hat{\mu}_n, \mu)\right] \leq C_{p,q} M_q^{p/q}(\mu) \begin{cases} n^{-1/2} & \text{if } q > 2p, \\ n^{-1/2}\log(n) & \text{if } q = 2p, \\ n^{-(q-p)/q} & \text{if } q \in (p, 2p). \end{cases} \tag{7.120}$$

Besides, since $\mathbf{W}_p$ is a metric, we can apply Theorem 7.8 to derive the sample complexity of $\mathbf{SW}_p$. Consider $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$ with $q > p$, with respective empirical measures $\hat{\mu}_n, \hat{\nu}_n$. Then, starting from (7.101) and using the convergence rate derived in (7.120), we obtain the desired result as follows

$$\mathbb{E}\left[|\mathbf{SW}_p(\mu,\nu) - \mathbf{SW}_p(\hat{\mu}_n, \hat{\nu}_n)|\right] \tag{7.121}$$

$$\leq \left\{\mathbb{E}\left|\mathbf{SW}_p^p(\hat{\mu}_n, \mu)\right|\right\}^{1/p} + \left\{\mathbb{E}\left|\mathbf{SW}_p^p(\hat{\nu}_n, \nu)\right|\right\}^{1/p} \tag{7.122}$$

$$\leq C_{p,q}^{1/p}\left(M_q^{1/q}(\mu) + M_q^{1/q}(\nu)\right) \begin{cases} n^{-1/(2p)} & \text{if } q > 2p, \\ n^{-1/(2p)}\log(n)^{1/p} & \text{if } q = 2p, \\ n^{-(q-p)/(pq)} & \text{if } q \in (p, 2p). \end{cases} \tag{7.123}$$

□

7.6.6 Proofs for Section 7.3.3

Proof of Theorem 7.15. Let $p \in [1, \infty)$ and $\varepsilon \geq 0$. We use the reformulation of $\mathbf{W}_{p,\varepsilon}$ as the maximum of an expectation, as given in [Genevay et al., 2016, Proposition 2.1],

$$\mathbf{SW}^p_{p,\varepsilon}(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{W}^p_{p,\varepsilon}(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.124}$$

$$= \int_{\mathbb{S}^{d-1}} \left\{ \max_{\tilde{u}, \tilde{v} \in \mathrm{C}(\mathbb{R})} \mathbb{E}_{\theta^\star_\sharp \mu \otimes \theta^\star_\sharp \nu} \left[\phi_\varepsilon\big(\tilde{u}(\tilde{X}), \tilde{v}(\tilde{Y}), \tilde{X}, \tilde{Y}\big) \right] \right\}^p \mathrm{d}\boldsymbol{\sigma}(\theta) \,, \tag{7.125}$$

where $\mathrm{C}(\mathbb{R})$ denotes the set of continuous real functions, and $\phi_\varepsilon(t, s, x, y) = t + s - \varepsilon e^{(t+s-\|x-y\|^p)/\varepsilon}$.

Consider for any $\theta \in \mathbb{S}^{d-1}$, $\tilde{u}^\star_\theta$, $\tilde{v}^\star_\theta$ as the functions attaining the maximum in (7.125), which exist by [Genevay et al., 2019, Theorem 4 in the supplementary document]. We obtain

$$\mathbf{SW}^p_{p,\varepsilon}(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \left\{ \mathbb{E}_{\theta^\star_\sharp \mu \otimes \theta^\star_\sharp \nu} \left[\phi_\varepsilon\big(\tilde{u}^\star_\theta(\tilde{X}), \tilde{v}^\star_\theta(\tilde{Y}), \tilde{X}, \tilde{Y}\big) \right] \right\}^p \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.126}$$

$$= \int_{\mathbb{S}^{d-1}} \left\{ \mathbb{E}_{\mu \otimes \nu} \left[\phi_\varepsilon\big(\tilde{u}^\star_\theta \circ \theta^\star(X), \tilde{v}^\star_\theta \circ \theta^\star(Y), X, Y\big) \right] \right\}^p \mathrm{d}\boldsymbol{\sigma}(\theta) \,. \tag{7.127}$$

Since for all $\tilde{w} \in \mathrm{C}(\mathbb{R})$ and $\theta \in \mathbb{S}^{d-1}$, $\tilde{w} \circ \theta^\star \in \mathrm{C}(\mathbb{R}^d)$, we can bound (7.127) as follows

$$\mathbf{SW}^p_{p,\varepsilon}(\mu, \nu) \leq \int_{\mathbb{S}^{d-1}} \left\{ \max_{u, v \in \mathrm{C}(\mathbb{R}^d)} \mathbb{E}_{\mu \otimes \nu} \left[\phi_\varepsilon\big(u(X), v(Y), X, Y\big) \right] \right\}^p \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{7.128}$$

$$= \mathbf{W}^p_{p,\varepsilon}(\mu, \nu) \,. \tag{7.129}$$

By Proposition 7.1, since $\mathbf{W}_{p,\varepsilon}$ is non-negative, so is $\mathbf{SW}_{p,\varepsilon}$, and we can apply $t \mapsto t^{1/p}$ on both sides of (7.129) to obtain the final result.

□

We move on to the proof of Theorem 7.16, which requires preliminary technical results.

Proposition 7.26. *Let $\tilde{\mathsf{X}}$ be a compact subset of $\mathbb{R}$, and $\mu', \nu' \in \mathcal{P}(\tilde{\mathsf{X}})$ with respective empirical instantiations $\hat{\mu}'_n, \hat{\nu}'_n$. Let $p \in [1, \infty)$ and $\varepsilon \geq 0$. Then,*

$$\left| \mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n, \hat{\nu}'_n) - \mathbf{W}_{p,\varepsilon}(\mu', \nu') \right| \leq 2 \operatorname{diam}(\tilde{\mathsf{X}}) \left\{ \mathbf{W}_1(\mu', \hat{\mu}'_n) + \mathbf{W}_1(\nu', \hat{\nu}'_n) \right\} \,. \tag{7.130}$$

Proof. Let $p \in [1, \infty)$, $\varepsilon \geq 0$ and $\tilde{\mathsf{X}} \subset \mathbb{R}$ compact. Consider $\mu', \nu' \in \mathcal{P}(\tilde{\mathsf{X}})$ with respective empirical distributions $\hat{\mu}'_n, \hat{\nu}'_n$. We first express the regularized OT cost as the maximum of an expectation [Genevay et al., 2016, Proposition 2.1]

$$\mathbf{W}_{p,\varepsilon}(\mu', \nu') = \max_{\tilde{u}, \tilde{v} \in \mathrm{C}(\mathbb{R})} \mathbb{E}_{\mu' \otimes \nu'} \left[\phi_\varepsilon\big(\tilde{u}(\tilde{X}), \tilde{v}(\tilde{Y}), \tilde{X}, \tilde{Y}\big) \right] \tag{7.131}$$

$$\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n, \nu') = \max_{\tilde{u}, \tilde{v} \in \mathrm{C}(\mathbb{R})} \mathbb{E}_{\hat{\mu}'_n \otimes \nu'} \left[\phi_\varepsilon\big(\tilde{u}(\tilde{X}), \tilde{v}(\tilde{Y}), \tilde{X}, \tilde{Y}\big) \right] \,, \tag{7.132}$$

where $\phi_\varepsilon(t, s, x, y) = t + s - \varepsilon e^{(t+s-\|x-y\|^2/2)/\varepsilon}$. By [Genevay et al., 2019, Proposition 1], the Sinkhorn potentials $(\tilde{u}, \tilde{v})$ are Lipschitz continuous with Lipschitz constant $\operatorname{diam}(\tilde{\mathsf{X}}) < \infty$. Therefore, by denoting by $\mathrm{Lip}_{\operatorname{diam}(\tilde{\mathsf{X}})}(\mathbb{R})$ the space of $\operatorname{diam}(\tilde{\mathsf{X}})$-Lipschitz

continuous functions defined on $\mathbb{R}$, (7.131) and (7.132) can be rewritten with the maximization over $\mathrm{Lip}_{\mathrm{diam}(\tilde{\mathsf{X}})}(\mathbb{R})$.

We can now use [Mena and Niles-Weed, 2019, Proposition 2] to bound the absolute difference of $\mathbf{W}_{p,\varepsilon}(\mu',\nu')$ and $\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\nu')$. We provide the detailed proof below for completeness. By [Mena and Niles-Weed, 2019, Proposition 6, Appendix A], there exist smooth potentials $(\tilde{u}^\star,\tilde{v}^\star)$ attaining the maximum in (7.131) such that, for all $\tilde{x},\tilde{y}\in\mathbb{R}$,

$$\int_{\mathbb{R}} \phi_\varepsilon(\tilde{u}^\star(\tilde{x}),\tilde{v}^\star(\tilde{y}),\tilde{x},\tilde{y})\mathrm{d}\nu'(\tilde{y}) = 1 \quad \mu'\text{-almost surely}, \tag{7.133}$$

$$\int_{\mathbb{R}} \phi_\varepsilon(\tilde{u}^\star(\tilde{x}),\tilde{v}^\star(\tilde{y}),\tilde{x},\tilde{y})\mathrm{d}\mu'(\tilde{x}) = 1 \quad \nu'\text{-almost surely}\,. \tag{7.134}$$

Analogously, there exist smooth optimal potentials $(\tilde{u}^\star_n,\tilde{v}^\star_n)$ for (7.132) satisfying (7.133) and (7.134) where $\tilde{u}^\star$, $\tilde{v}^\star$ and μ' are replaced by $\tilde{u}^\star_n$, $\tilde{v}^\star_n$ and $\hat{\mu}'_n$ respectively.

The optimality of these potentials give us

$$\mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star_n(\tilde{X}),\tilde{v}^\star_n(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star_n(\tilde{X}),\tilde{v}^\star_n(\tilde{Y}),\tilde{X},\tilde{Y})\right] \tag{7.135}$$

$$\leq \mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star_n(\tilde{X}),\tilde{v}^\star_n(\tilde{Y}),\tilde{X},\tilde{Y})\right] \tag{7.136}$$

$$\leq \mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right]\,. \tag{7.137}$$

Therefore,

$$\left|\mathbf{W}_{p,\varepsilon}(\mu',\nu') - \mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\nu')\right| \tag{7.138}$$

$$= \left|\mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star_n(\tilde{X}),\tilde{v}^\star_n(\tilde{Y}),\tilde{X},\tilde{Y})\right]\right| \tag{7.139}$$

$$\leq \left|\mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right]\right| \tag{7.140}$$

$$+ \left|\mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star_n(\tilde{X}),\tilde{v}^\star_n(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star_n(\tilde{X}),\tilde{v}^\star_n(\tilde{Y}),\tilde{X},\tilde{Y})\right]\right|\,. \tag{7.141}$$

We bound each term of the sum in (7.141) as follows

$$\left|\mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right]\right| \tag{7.142}$$

$$= \left|\int_{\mathbb{R}} \tilde{u}^\star(\tilde{x})\mathrm{d}(\mu'-\hat{\mu}'_n)(\tilde{x}) - \varepsilon\int_{\mathbb{R}}\int_{\mathbb{R}} e^{(\tilde{u}^\star(\tilde{x})+\tilde{v}^\star(\tilde{y})-|\tilde{x}-\tilde{y}|^2/2)/\varepsilon}\mathrm{d}\nu'(\tilde{y})\mathrm{d}(\mu'-\hat{\mu}'_n)(\tilde{x})\right| \tag{7.143}$$

$$= \left|\int_{\mathbb{R}} \tilde{u}^\star(\tilde{x})\mathrm{d}(\mu'-\hat{\mu}'_n)(\tilde{x})\right| \leq \sup_{\tilde{u}\in\mathrm{Lip}_{\mathrm{diam}(\tilde{\mathsf{X}})}(\mathbb{R})}\left|\int_{\mathbb{R}} \tilde{u}(\tilde{x})\mathrm{d}(\mu'-\hat{\mu}'_n)(\tilde{x})\right|\,, \tag{7.144}$$

where (7.144) results from (7.133). Since for any $f\in\mathrm{Lip}_{\mathtt{L}}(\mathbb{R})$ with $\mathtt{L}>0$, $f/\mathtt{L}\in\mathrm{Lip}_1(\mathbb{R})$, (7.144) can be bounded as follows

$$\left|\mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right] - \mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}^\star(\tilde{X}),\tilde{v}^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right]\right| \tag{7.145}$$

$$\leq \mathrm{diam}(\tilde{\mathsf{X}})\sup_{\tilde{u}\in\mathrm{Lip}_1(\mathbb{R})}\left|\int_{\mathbb{R}} \tilde{u}(\tilde{x})\mathrm{d}(\theta^\star_\sharp\mu - \theta^\star_\sharp\hat{\mu}_n)(\tilde{x})\right| = \mathrm{diam}(\tilde{\mathsf{X}})\mathbf{W}_1(\mu',\hat{\mu}'_n)\,, \tag{7.146}$$

where (7.146) follows from the dual formulation of the Wasserstein distance of order 1 [Villani, 2008, Theorem 5.10].

We show with an analogous proof that

$$\left|\mathbb{E}_{\mu'\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}_n^\star(\tilde{X}),\tilde{v}_n^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right]-\mathbb{E}_{\hat{\mu}'_n\otimes\nu'}\left[\phi_\varepsilon(\tilde{u}_n^\star(\tilde{X}),\tilde{v}_n^\star(\tilde{Y}),\tilde{X},\tilde{Y})\right]\right| \tag{7.147}$$

$$\leq \operatorname{diam}(\tilde{\mathsf{X}})\mathbf{W}_1(\mu',\hat{\mu}'_n)\;, \tag{7.148}$$

which leads to the conclusion that

$$\left|\mathbf{W}_{p,\varepsilon}(\mu',\nu')-\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\nu')\right|\leq 2\operatorname{diam}(\tilde{\mathsf{X}})\mathbf{W}_1(\mu',\hat{\mu}'_n)\;. \tag{7.149}$$

By using the triangle inequality and (7.149), we obtain the final result

$$\left|\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\hat{\nu}'_n)-\mathbf{W}_{p,\varepsilon}(\mu',\nu')\right| \tag{7.150}$$

$$\leq \left|\mathbf{W}_{p,\varepsilon}(\mu',\nu')-\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\nu')\right|+\left|\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\nu')-\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\hat{\nu}'_n)\right| \tag{7.151}$$

$$\leq 2\operatorname{diam}(\tilde{\mathsf{X}})\left\{\mathbf{W}_1(\mu',\hat{\mu}'_n)+\mathbf{W}_1(\nu',\hat{\nu}'_n)\right\}\;. \tag{7.152}$$

□

Corollary 7.27. *Let $\tilde{\mathsf{X}}$ be a compact subset of $\mathbb{R}$, and $\mu',\nu'\in\mathcal{P}(\tilde{\mathsf{X}})$. Denote by $\hat{\mu}'_n,\hat{\nu}'_n$ their respective empirical instantiations. Let $p\in[1,\infty)$ and $\varepsilon\geq 0$. Then,*

$$\mathbb{E}\left|\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\hat{\nu}'_n)-\mathbf{W}_{p,\varepsilon}(\mu',\nu')\right|\leq 2\operatorname{diam}(\tilde{\mathsf{X}})C_q\left[M_q^{1/q}(\mu')+M_q^{1/q}(\nu')\right]n^{-1/2}\;, \tag{7.153}$$

where $q>2$, $C_q<\infty$ is a constant that depends on q, and $M_q(\mu'),M_q(\nu')$ are the moments of order q of μ',ν' respectively.

Proof. We apply Proposition 7.26 and take the expectation of (7.130) with respect to $\tilde{X}_{1:n}\sim\hat{\mu}'_n$ and $\tilde{Y}_{1:n}\sim\hat{\nu}'_n$

$$\mathbb{E}\left|\mathbf{W}_{p,\varepsilon}(\hat{\mu}'_n,\hat{\nu}'_n)-\mathbf{W}_{p,\varepsilon}(\mu',\nu')\right|\leq 2\operatorname{diam}(\tilde{\mathsf{X}})\mathbb{E}\left\{\mathbf{W}_1(\mu',\hat{\mu}'_n)+\mathbf{W}_1(\nu',\hat{\nu}'_n)\right\}\;. \tag{7.154}$$

Since μ' and ν' are both supported on a compact space, they have infinitely many finite moments. We can then bound (7.154) using the convergence rate of empirical measures in $\mathbf{W}_1$, recalled in Lemma 7.25. This concludes the proof.

□

Proof of Theorem 7.16. Let $p\in[1,\infty)$ and $\varepsilon\geq 0$. Consider $\mu,\nu\in\mathcal{P}(\mathsf{X})$ with $\mathsf{X}\subset\mathbb{R}^d$ compact, and denote by $\hat{\mu}_n,\hat{\nu}_n$ their respective empirical distributions.

Let $\theta\in\mathbb{S}^{d-1}$ and define $\mathsf{X}_\theta=\{\langle\theta,x\rangle\,:\,x\in\mathsf{X}\}$. X_θ is compact (since X is compact and $\theta^\star$ is continuous) and verifies $\operatorname{diam}(\mathsf{X}_\theta)\leq\operatorname{diam}(\mathsf{X})$ (by the Cauchy-Schwarz inequality). Besides, by (7.119), for any $k>0$, $M_k(\theta^\star_\sharp\mu)\leq M_k(\mu)$ and $M_k(\theta^\star_\sharp\nu)\leq M_k(\nu)$. By Corollary 7.27, there exists $C_q<\infty$ which depends on $q>2$ such that,

$$\mathbb{E}\left|\mathbf{W}_{p,\varepsilon}(\theta^\star_\sharp\hat{\mu}_n,\theta^\star_\sharp\hat{\nu}_n)-\mathbf{W}_{p,\varepsilon}(\theta^\star_\sharp\mu,\theta^\star_\sharp\nu)\right| \tag{7.155}$$

$$\leq 2\operatorname{diam}(\mathsf{X})C_q\left[M_q^{1/q}(\mu)+M_q^{1/q}(\nu)\right]n^{-1/2}\;. \tag{7.156}$$

The sample complexity of $\mathbf{SW}_{p,\varepsilon}$ is finally obtained by applying Theorem 7.7.

□

Finally, in addition to the background elements given in Section 2.5, we recall an important result regarding the convergence of Sinkhorn's algorithm, which will be useful for the proof of Proposition 7.17.

Sinkhorn's algorithm refers to an iterative procedure which operates on empirical distributions as follows: consider a cost matrix C between two sets of n samples, and define the matrix K such that, for $1 \leq i, j \leq n$,

$$K_{i,j} = \exp\left(-\frac{C_{i,j}}{\varepsilon}\right) ,$$

and initialize $b^{(0)} = 1 \in \mathbb{R}^n$; then, compute for $\ell > 1$,

$$a^{(\ell)} = 1./n(Kb^{(\ell-1)}) ,$$
$$b^{(\ell)} = 1./n(Ka^{(\ell)}) ,$$

where ./ stands for the entry-wise division. This defines a sequence $\gamma_{i,j}^{(\ell)} = a_i^{(\ell)} K_{i,j} b_j^{(\ell)}$, which converges to a solution of the regularized OT problem (2.16) at a linear rate. The convergence rate of Sinkhorn's algorithm is recalled in Theorem 7.28. For an extended discussion on this result, we refer to [Peyré and Cuturi, 2019, Section 4.2].

Theorem 7.28 (Franklin and Lorenz [1989]). *The iterates $a^{(\ell)}$ and $b^{(\ell)}$ of Sinkhorn's algorithm converge linearly for the Hilbert metric at a rate $1-\tanh(\tau(K)/4)$, with $\tau(K) = \log \max_{i,j,i',j'} \frac{K_{ij}K_{i'j'}}{K_{ij'}K_{i'j}}$. In particular, for the squared-norm cost, i.e. $K_{ij} = \exp(-\|x_i - x_j\|^2/\varepsilon)$, it holds*

$$\tau(K) \leq 2 \max_{i,j} \|x_i - x_j\|^2/\varepsilon. \tag{7.157}$$

Proof of Proposition 7.17. For $i, j \in \{1, \ldots, n\}$, $f_{i,j} : \theta \in \mathbb{S}^{d-1} \mapsto \frac{1}{R}\langle \theta, x_i - x_j \rangle$ is 1-Lipschitz and has median 0 for θ uniformly distributed on the unit sphere. Thus, by concentration of measure on the sphere [Wainwright, 2019, Example 3.12], it holds for $\varepsilon > 0$,

$$\mathbb{P}\left(|f_{i,j}(\theta)| \geq \varepsilon\right) \leq \sqrt{2\pi} \exp(-d\varepsilon^2/2) .$$

Taking a union bound over the $n(n-1)$ pairs of indices and setting $\tau = (R\varepsilon)^2$, it follows

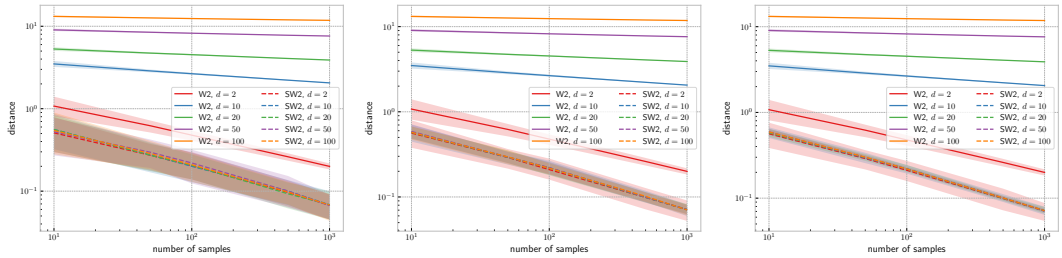
$$\mathbb{P}\left(\max_{i,j} |\langle \theta, x_i - x_j \rangle|^2 \geq \tau\right) \leq \sqrt{2\pi} n^2 \exp(-d\tau/2R^2) .$$

Hence, for any $\delta > 0$, it holds with probability $1 - \delta$ that $\max_{i,j} |\langle \theta, x_i - x_j \rangle|^2 \leq \frac{2R^2}{d} \log(\sqrt{2\pi} n^2/\delta)$. This argument was suggested to us by an anonymous reviewer at the NeurIPS 2020 conference.

□

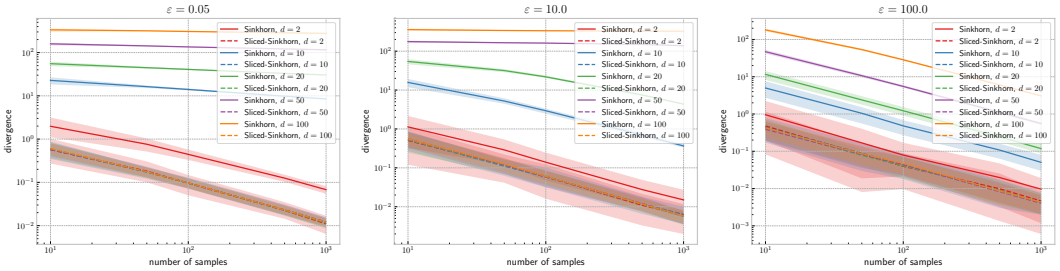
7.6.7 Additional empirical results

In this section, we provide additional results obtained for the synthetical experiments illustrating the sample complexity of Sliced-Wasserstein and Sliced-Sinkhorn divergences: we produce figures analogously to Figures 7.2b, 7.3a and 7.3b, with different hyperparameter values.

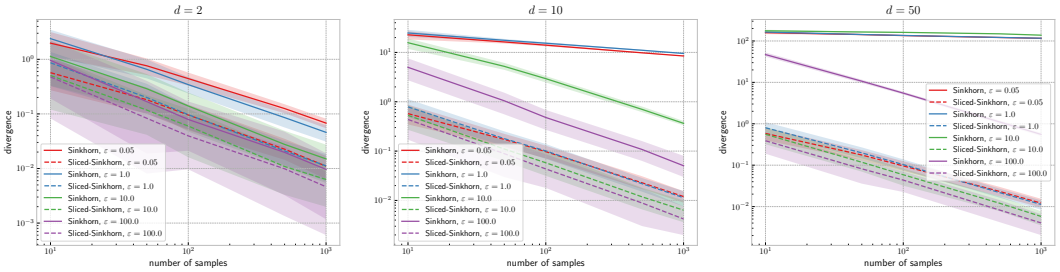


$L = 1$ $L = 10$ $L = 1000$

Figure 7.5: Illustration of Corollary 7.14: Wasserstein and Sliced-Wasserstein distances of order 2 between two sets of n samples generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ vs. n, for different d, on log-log scale. $\mathbf{SW}_2$ is approximated with L random projections, $L \in \{1, 10, 1000\}$. Results are averaged over 100 runs, and the shaded areas correspond to the 10th-90th percentiles. Figure 7.2b shows the results for $L = 100$.



Influence of the data dimension for $\varepsilon \in \{0.05, 10, 100\}$



Influence of the regularization coefficient for $d \in \{2, 10, 50\}$

Figure 7.6: Illustration of Theorem 7.16: Sinkhorn and Sliced-Sinkhorn divergences between two sets of n samples generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for different values of n, dimension d, and regularization coefficient ε. Sliced-Sinkhorn is approximated with 10 random projections. Results are averaged over 100 runs, and the shaded areas correspond to the 10th-90th percentiles. All plots have a log-log scale. Figure 7.3a shows the influence of the dimension for $\varepsilon = 1$, and Figure 7.3b shows the influence of the regularization for $d = 100$.

Chapter 8

Conclusion

8.1 Summary

This thesis builds on the recent success of optimal transport tools for machine learning models, and is specifically focused on the increasingly popular Sliced-Wasserstein distance. Our general purpose was to provide a more comprehensive understanding of this metric by investigating its theoretical and practical implications, especially when applied for parameter inference in generative models. Some of our results enabled us to identify the limitations caused by the estimation of SW and overcome them by designing novel methodologies. We summarize our main contributions below, accordingly to the objectives listed in Section 1.4 (page 22).

Theoretical properties of SW. *This paragraph provides answers to objective 1.* We established several theoretical results that aim at shedding light on the empirical behavior of SW as reported in prior work. Specifically, in Chapter 3, we studied the estimators obtained by minimizing SW over a parametric space and proved some of their asymptotic properties. Besides ensuring theoretical consistency to existing SW-based generative models, our results allowed us to derive new topological properties, namely the lower semi-continuity of SW and the fact that convergence under SW implies weak convergence.

Then, we demonstrate that SW is able to mitigate the statistical limitations of the Wasserstein distance in high-dimensional settings. The central limit theorem proved in Chapter 3 and characterizing the asymptotic distribution of minimum SW estimators exhibits a convergence rate of $\sqrt{n}$, where n is the number of observations. This dimension-free rate as well as the empirical comparison in Chapter 4 (Figure 4.1, page 74) are first evidence that SW offers important statistical benefits over the Wasserstein distance. This is further confirmed by the sample complexity of SW derived in Chapter 7, which does not depend on the dimension, as opposed to the sample complexity of the Wasserstein distance which can grow exponentially in dimension.

Finally, our theoretical results in Chapter 7 also explain why the statistical efficiency is actually balanced with the fact that SW is defined as an integral over $\mathbb{S}^{d-1}$. We elaborate on this aspect later, as an answer to objective 3.

New methodology for approximate inference. *This paragraph provides answers to objective 2.* We expanded the applicability of SW by developing a likelihood-free approximation inference technique based on this metric and called SW-ABC (Chapter 4). We showed that SW-ABC comes with convergence guarantees under different asymptotic regimes and offers a superior empirical performance as compared to existing ABC

techniques which rely on different divergences. Besides, we leveraged our methodology to design a novel image denoising algorithm, which demonstrates the flexibility of SW-ABC.

Limitations of SW and solutions. *This paragraph provides answers to objective 3.* As illustrated in previous empirical studies, the common practice that consists in approximating SW with a standard Monte Carlo estimate (over $L \in \mathbb{N}^*$ random projections) can be inefficient in high-dimensional settings and might thus degrade the performance of SW-based generative models. We support this observation with our theoretical findings in Chapter 7: we showed that the overall complexity of computing SW (and more generally, sliced divergences, as explained in objective 4) depends on the sample complexity but also on an error due to the Monte Carlo approximation, which depends on L and an additional variance term.

We then proposed two different approaches to overcome the limitations induced by the Monte Carlo estimates of SW. On the one hand, the literature on concentration of random projections, combined with the analytical expression of the Wasserstein distance between Gaussian distributions, helped us formulate a novel technique to compute SW with simple deterministic operations (Chapter 6). The returned SW estimates are guaranteed to converge to the exact value of SW, provided that the compared distributions satisfy a weak dependence condition. Besides, our approximation method can be used to improve the performance of generative models that are based on traditional Monte Carlo estimates.

On the other hand, we defined the class of Generalized Sliced-Wasserstein distances in Chapter 5 and illustrated their ability to outperform SW on generative modeling applications. These novel metrics bring an interesting perspective on adversarial generative modeling, showing that such algorithms contain an implicit stage for learning projections with different cost functions than ours.

Theoretical analysis of "slicing". *This paragraph provides answers to objective 4.* In order to understand what the slicing operation itself is bringing, we introduced in Chapter 7 the first general definition for sliced divergences, and derived their topological and statistical properties. Our results show that slicing leads to a dimension-free sample complexity, while carrying out useful topological properties of the "base divergence", e.g., metric axioms and metrizing weak convergence. If the focus is on sustaining such topological properties, then the improvement in the convergence rate is meaningful and circumvents the curse of dimensionality – but in practice, this rate is impacted by the Monte Carlo approximation.

8.2 Future Research Directions

Eventually, we hope that our contributions provide useful insights for practitioners in terms of designing new methodologies based on the Sliced-Wasserstein distance and its variants, as well as obtaining a better understanding of these tools. This thesis also opens up new prospects which motivate future research directions, notably on the approximation of sliced divergences and the analysis of GSW, as we describe below.

Improve the estimation of sliced divergences. As we argued in Chapters 5 to 7, the Monte Carlo strategy, widely used to compute sliced probability divergences in practice, is not ideal given the induced approximation error. Our alternative SW estimate,

which we developed in Chapter 6 to address this aspect, provides important computational advantages over Monte Carlo, but still leaves room for improvement. First, the convergence rate of our approximate to the true SW seems slow according to our nonasymptotical guarantees in Section 6.3.3 (page 102), while it is reasonably fast in our experiments. To bridge this gap between theory and practice, we can study if our proofs and the ones in [Reeves, 2017] can be refined when assuming additional structure on the distributions, e.g. sub-Gaussian and sub-exponential: this will help identify the settings under which our current bounds are tight or can be improved.

Furthermore, since our methodology in Chapter 6 applies to the Sliced-Wasserstein distance only, another crucial question is how to effectively approximate the larger class of sliced divergences introduced in Chapter 7. One idea would be to resort to alternative Monte Carlo algorithms, such as sequential Monte Carlo samplers [Chopin, 2002, Del Moral et al., 2006], in order to develop a technique that improves the convergence rate or the variance of the traditional Monte Carlo estimates.

In-depth analysis of GSW. Our work on Generalized Sliced-Wasserstein distances in Chapter 5 have since then inspired other follow-up studies, including [Nguyen et al., 2021, Chen et al., 2021, Naderializadeh et al., 2021], and pave the way for a deeper theoretical investigation. Indeed, our experiments showed that the choice of the defining function g is data-dependent and highly impacts on the performance of the associated GSW, and the question "when and why do certain choices of g perform well in practice?" has not yet been elucidated. To address this matter, we can study to what extent the theoretical analysis in [Nadjahi et al., 2019, 2020b] can be generalized to GSW: the topological and statistical properties of GSW might highly depend on the types of one-dimensional representations used. This requires an analysis of the generalized Radon transform and would help identifying the advantages of some representations against others.

On the other hand, we can explore whether a connection can be established between kernel functions and GSW (see [Kolouri et al., 2020b] as our preliminary study), and between SW for manifolds and GSW. One hypothesis would be that, with appropriate representations, GSW might be equivalent to defining SW on manifolds. To verify this assumption, a starting point would be to formulate a definition of SW on compact Riemannian manifolds, for example by leveraging existing characterizations of the Wasserstein distance on such spaces [Rabin et al., 2011]. Apart from understanding better GSW, such a distance would be very useful in practice: the Sliced-Wasserstein distance has only been defined for data living on $\mathbb{R}^d$, although there are various fields where the training data are supported on Riemannian manifolds, e.g., in bioinformatics [Mardia et al., 2018], neurology [Kaufman et al., 2005], life sciences [Ameijeiras-Alonso et al., 2018] and text mining [Banerjee et al., 2005].

Bibliography

B. K. Abid and R. M. Gower. Greedy stochastic algorithms for entropy-regularized optimal transport problems. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, Lanzarote, Spain, Apr. 2018.

J. Adler and S. Lunz. Banach Wasserstein GAN. In *Advances in Neural Information Processing Systems*, pages 6754–6763, 2018.

S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966. ISSN 00359246.

C. Aliprantis, K. Border, and K. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Studies in economic theory. Springer, 1999. ISBN 9783540658542.

J. Altschuler, J. Weed, and P. Rigollet. Near-Linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 1961–1971, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2005.

J. Ameijeiras-Alonso, R. M. Crujeiras, and A. Rodríguez-Casal. Directional Statistics for Wildfires. *Applied Directional Statistics*, page 203–226, Sep 2018. doi: 10.1201/9781315228570-17.

M. Anttila, K. Ball, and I. Perissinaki. The Central Limit Problem for Convex Bodies. *Transactions of the American Mathematical Society*, 355(12):4723–4735, 2003. ISSN 00029947.

M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum Mean Discrepancy Gradient Flow. In *Advances in Neural Information Processing Systems*, 2019.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

A. Banerjee, J. Ghosh, S. Sra, S. Dhillon, and Banerjee. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statistics & Probability Letters*, 76(12):1298 – 1302, 2006. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2006.02.001.

A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2011. ISBN 9781420099669.

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

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035, 2002. ISSN 0016-6731.

M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, Jan 2019. doi: 10.1093/imaiai/drn000.

G. Beylkin. The inversion problem and applications of the generalized Radon transform. *Communications on pure and applied mathematics*, 37(5):579–599, 1984.

P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9. A Wiley-Interscience Publication.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

S. G. Bobkov. On concentration of distributions of random weighted sums. *The Annals of Probability*, 31(1):195 – 215, 2003. doi: 10.1214/aop/1046294309.

V. Bogachev. *Measure Theory*. Number vol. 1 in Measure Theory. Springer Berlin Heidelberg, 2007. ISBN 9783540345145.

E. Boissard and T. L. Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539 – 563, 2014. doi: 10.1214/12-AIHP517.

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

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

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

L. D. Brown and R. Purves. Measurable Selections of Extrema. *Ann. Statist.*, 1(5): 902–912, 09 1973. doi: 10.1214/aos/1176342510.

A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005.

M. Carrière, M. Cuturi, and S. Oudot. Sliced Wasserstein Kernel for Persistence Diagrams. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 664–673, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

X. Chen, Y. Yang, and Y. Li. Augmented Sliced Wasserstein Distances, 2021.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3): 539–552, 08 2002. ISSN 0006-3444. doi: 10.1093/biomet/89.3.539.

I. Ciszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

S. Cohen, K. S. S. Kumar, and M. P. Deisenroth. Sliced Multi-Marginal Optimal Transport, 2021.

J. Cornebise, E. Moulines, and J. Olsson. Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing*, 18(4): 461–480, Aug 2008. doi: 10.1007/s11222-008-9089-4.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2017.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

H. Cramér. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):141–180, 1928. doi: 10.1080/03461238.1928.10416872.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

B. Dai and U. Seljak. Sliced Iterative Normalizing Flows. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2352–2364. PMLR, 18–24 Jul 2021.

S. Deans. *The Radon Transform and Some of Its Applications*. Dover Books on Mathematics Series. Dover Publications, 2007. ISBN 9780486462417.

S. Dede. An empirical central limit theorem in L1 for stationary sequences. *Stochastic Processes and their Applications*, 119(10):3494–3515, 2009.

J. Dedecker and F. Merlevède. The empirical distribution function for dependent variables: asymptotic and nonasymptotic results in $\mathbb{L}^p$. *ESAIM: Probability and Statistics*, 11:102–114, 2007. doi: 10.1051/ps:2007009.

E. del Barrio, E. Giné, and C. Matrán. Central Limit Theorems for the Wasserstein Distance Between the Empirical and the True Distributions. *Ann. Probab.*, 27(2): 1009–1071, 04 1999. doi: 10.1214/aop/1022677394.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(3):411–436, 2006.

A. Denisyuk. Inversion of the generalized Radon transform. *Translations of the American Mathematical Society-Series 2*, 162:19–32, 1994.

S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 49(4):1183 – 1203, 2013. doi: 10.1214/12-AIHP489.

I. Deshpande, Z. Zhang, and A. G. Schwing. Generative modeling using the sliced Wasserstein distance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. Schwing. Max-Sliced Wasserstein distance and its use for GANs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

P. Diaconis and D. Freedman. Asymptotics of Graphical Projection Pursuit. *The Annals of Statistics*, 12(3):793 – 815, 1984. doi: 10.1214/aos/1176346703.

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018. ISBN 978-3-319-97703-4.

P. Doukhan and S. Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84(2):313–342, 1999. ISSN 0304-4149. doi: https://doi.org/10.1016/S0304-4149(99)00055-1.

P. Doukhan and M. H. Neumann. Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117(7):878–903, 2007. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2006.10.011.

P. Doukhan and M. H. Neumann. The notion of Ψ-weak dependence and its applications to bootstrapping time series. *Probability Surveys*, 5(none):146 – 168, 2008. doi: 10.1214/06-PS086.

D. Dowson and B. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(82)90077-X.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Statist.*, 40(1):40–50, 02 1969. doi: 10.1214/aoms/1177697802.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376. PMLR, 10–15 Jul 2018.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training Generative Neural Networks via Maximum Mean Discrepancy Optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267, Arlington, Virginia, USA, 2015. ISBN 9780996643108.

L. Dümbgen and P. Del Conte-Zerial. On low-dimensional projections of high-dimensional distributions. *Institute of Mathematical Statistics Collections*, pages 91 – 104, 2013. doi: 10.1214/12-imscoll908.

L. Ehrenpreis. *The universality of the Radon transform.* Oxford University Press on Demand, 2003.

L. Fan, F. Zhang, H. Fan, and C. Zhang. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*, 2(1):7, 2019.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012. doi: 10.1111/j.1467-9868.2011.01010.x.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 16–18 Apr 2019.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

G. B. Folland. *Real analysis.* Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999. ISBN 0-471-31716-0. Modern techniques and their applications, A Wiley-Interscience Publication.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, Aug. 2015.

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

D. T. Frazier, G. M. Martin, C. Robert, and J. Rousseau. Asymptotic properties of approximate Bayesian computation. *Biometrika*, 105(3):593–607, 06 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy027.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

C. Frogner, F. Mirzazadeh, and J. Solomon. Learning Embeddings into Entropic Wasserstein Spaces. In *International Conference on Learning Representations (ICLR)*, 2019.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

I. M. Gel'fand, M. I. Graev, and Z. Y. Shapiro. Differential forms and integral geometry. *Functional Analysis and its Applications*, 3(2):101–114, 1969.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic Optimization for Large-scale Optimal Transport. In *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

A. Genevay, G. Peyré, and M. Cuturi. GAN and VAE from an optimal transport point of view. *arXiv preprint arXiv:1706.01807*, 2017.

A. Genevay, G. Peyre, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 09–11 Apr 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample Complexity of Sinkhorn Divergences. In *Proceedings of Machine Learning Research*, pages 1574–1583, 2019.

A. L. Gibbs and F. E. Su. On Choosing and Bounding Probability Metrics. *International Statistical Review*, 70(3):419–435, 2002. doi: 10.1111/j.1751-5823.2002.tb00178.x.

S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. *Proceedings of Machine Learning Research*, 145:1 – 46, 2021.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A Kernel Statistical Test of Independence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *J. Mach. Learn. Res.*, 13(null):723–773, Mar. 2012. ISSN 1532-4435.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

P. Hall and K.-C. Li. On almost Linearity of Low Dimensional Projections from High Dimensional Data. *The Annals of Statistics*, 21(2):867 – 889, 1993. doi: 10.1214/aos/1176349155.

S. Helgason. The Radon transform on Rn. In *Integral Geometry and Radon Transforms*, pages 1–62. Springer, 2011.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

A. Homan and H. Zhou. Injectivity and stability for a generic class of generalized Radon transforms. *The Journal of Geometric Analysis*, 27(2):1515–1529, 2017.

G. Huber. Gamma Function Derivation of n-Sphere Volumes. *The American Mathematical Monthly*, 89(5):301–302, 1982. doi: 10.1080/00029890.1982.11995438.

B. Jiang, T.-Y. Wu, and W.-H. Wong. Approximate Bayesian Computation with Kullback-Leibler Divergence as Data Discrepancy. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1711–1721, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

O. Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, 1997. ISBN 0-387-94957-7.

L. V. Kantorovich. On the Translocation of Masses (title translated from Russian). *Dokl. Akad. Nauk SSSR*, 37(7–8):227–229, 1942.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

C. G. Kaufman, V. Ventura, and R. E. Kass. Spline-based non-parametric regression for periodic functions and its application to directional tuning of neurons. *Statistics in Medicine*, 24(14):2255–2265, 2005. doi: 10.1002/sim.2104.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

B. Klartag. A central limit theorem for convex sets. *Inventiones mathematicae*, 168(1): 91–131, Jan 2007. doi: 10.1007/s00222-006-0028-8.

E. Klinger, D. Rickert, and J. Hasenauer. pyABC: distributed, likelihood-free inference. *Bioinformatics*, 34(20):3591–3593, 05 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty361.

S. Knop, P. Spurek, J. Tabor, I. Podolak, M. Mazur, and S. Jastrzębski. Cramer-Wold Auto-Encoder. *Journal of Machine Learning Research*, 21(164):1–28, 2020.

S. Kolouri, Y. Zou, and G. K. Rohde. Sliced-Wasserstein Kernels for Probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2016.

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

S. Kolouri, G. K. Rohde, and H. Hoffmann. Sliced Wasserstein distance for Learning Gaussian Mixture Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436, 2018.

S. Kolouri, K. Nadjahi, U. Şimşekli, R. Badeau, and G. Rohde. Generalized Sliced Wasserstein Distances. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019a.

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

S. Kolouri, N. A. Ketz, A. Soltoggio, and P. K. Pilly. Sliced Cramer Synaptic Consolidation for Preserving Deeply Learned Representations. In *International Conference on Learning Representations*, 2020a.

S. Kolouri, K. Nadjahi, U. Şimşekli, and S. Shahrampour. Generalized Sliced Distances for Probability Distributions, 2020b.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

P. Kuchment. Generalized transforms of Radon type and their applications. In *Proceedings of Symposia in Applied Mathematics*, volume 63, page 67, 2006.

Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

H. Leeb. On the conditional distributions of low-dimensional projections from high-dimensional data. *The Annals of Statistics*, 41(2):464 – 483, 2013. doi: 10.1214/12-AOS1081.

S. Leglaive, U. Şimşekli, A. Liutkus, R. Badeau, and G. Richard. Alpha-stable multichannel audio source separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 576–580. IEEE, 2017.

Y. Li, K. Swersky, and R. Zemel. Generative Moment Matching Networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, page 1718–1727, 2015.

T. Lin, N. Ho, and M. Jordan. On Efficient Optimal Transport: An Analysis of Greedy and Accelerated Mirror Descent Algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3982–3991. PMLR, 09–15 Jun 2019.

S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

A. Liutkus, U. Şimşekli, S. Majewski, A. Durmus, and F.-R. Stöter. Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4104–4113, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

M. Loève. *Probability Theory I.* Springer-Verlag New York, 1977. doi: 10.1007/978-1-4684-9464-8.

B. B. Mandelbrot. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E.* Springer Science & Business Media, 2013.

T. Manole, S. Balakrishnan, and L. Wasserman. Minimax Confidence Intervals for the Sliced Wasserstein Distance, 2019.

K. V. Mardia, J. Illemann Foldager, and J. Frellsen. Directional Statistics in Protein Bioinformatics. *Applied Directional Statistics*, page 17–40, Sep 2018. doi: 10.1201/9781315228570-9.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, 2001. doi: 10.1109/ICCV.2001.937655.

E. Meckes. Approximation of Projections of Random Vectors. *Journal of Theoretical Probability*, 25(2):333–352, Jun 2010. doi: 10.1007/s10959-010-0299-2.

G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems 32*, pages 4543–4553. Curran Associates, Inc., 2019.

C. Meng, Y. Ke, J. Zhang, M. Zhang, W. Zhong, and P. Ma. Large-scale optimal transport map estimation using projection pursuit. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

A. Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429–443, 1997. doi: 10.2307/1428011.

N. Naderializadeh, S. Kolouri, J. F. Comer, R. W. Andrews, and H. Hoffmann. Set Representation Learning with Generalized Sliced-Wasserstein Embeddings. *CoRR*, abs/2103.03892, 2021.

K. Nadjahi, A. Durmus, U. Şimşekli, and R. Badeau. Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

K. Nadjahi, V. De Bortoli, A. Durmus, R. Badeau, and U. Şimşekli. Approximate Bayesian Computation with the Sliced-Wasserstein Distance. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020a.

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Şimşekli. Statistical and Topological Properties of Sliced Probability Divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020b.

K. Nadjahi, A. Durmus, P. E. Jacob, R. Badeau, and U. Şimşekli. Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

F. Natterer. *The mathematics of computerized tomography*, volume 32. SIAM, 1986.

J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308.

K. Nguyen, N. Ho, T. Pham, and H. Bui. Distributional Sliced-Wasserstein and Applications to Generative Modeling. In *International Conference on Learning Representations*, 2021.

J. P. Nolan. Multivariate elliptically contoured stable distributions: theory and estimation. *Computational Statistics*, 28(5):2067–2089, Oct 2013. ISSN 1613-9658. doi: 10.1007/s00180-013-0396-7.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian Computation with Kernel Embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 398–407, Cadiz, Spain, 09–11 May 2016. PMLR.

G. Patrini, M. Carioni, P. Forre, S. Bhargav, M. Welling, R. v. d. Berg, T. Genewein, and F. Nielsen. Sinkhorn autoencoders. *arXiv preprint arXiv:1810.01118*, 2018.

F.-P. Paty and M. Cuturi. Subspace Robust Wasserstein distances. In *International Conference on Machine Learning*, 2019.

G. Peters and S. Sisson. Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk*, 1, 09 2006. doi: 10.21314/JOP.2006.014.

G. Peyré and M. Cuturi. 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.

D. Pollard. The minimum distance method of testing. *Metrika*, 27(1):43–70, Dec 1980. ISSN 1435-926X. doi: 10.1007/BF01893576.

J. Rabin, J. Delon, and Y. Gousseau. Transportation Distances on the Circle. *Journal of Mathematical Imaging and Vision*, 41(1–2):147–167, May 2011. doi: 10.1007/s10851-011-0284-0.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein Barycenter and Its Application to Texture Mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-24785-9.

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

J. Radon. Uber die bestimmug von funktionen durch ihre integralwerte laengs geweisser mannigfaltigkeiten. *Berichte Saechsishe Acad. Wissenschaft. Math. Phys., Klass*, 69: 262, 1917.

A. Ramdas, N. G. Trillos, and M. Cuturi. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2), 2017. ISSN 1099-4300. doi: 10.3390/e19020047.

I. Redko, N. Courty, R. Flamary, and D. Tuia. Optimal Transport for Multi-source Domain Adaptation under Target Shift. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 849–858. PMLR, 16–18 Apr 2019.

G. Reeves. Conditional central limit theorems for Gaussian projections. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3045–3049, 2017. doi: 10.1109/ISIT.2017.8007089.

R. Rockafellar, M. Wets, and R. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2009. ISBN 9783642024313.

F. Rouvière. Nonlinear Radon and Fourier Transforms, 2015. URL https://math.unice.fr/~frou/recherche/Nonlinear%20RadonW.pdf.

M. Rowland, J. Hron, Y. Tang, K. Choromanski, T. Sarlos, and A. Weller. Orthogonal Estimation of Wasserstein Distances. In *Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 186–195, 16–18 Apr 2019.

P. Rubenstein, O. Bousquet, J. Djolonga, C. Riquelme, and I. O. Tolstikhin. Practical and Consistent Estimation of f-Divergences. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A. Rényi. On Measures of Entropy and Information. In *Berkeley Symposium on Mathematical Statistics and Probability*, page 547–561, 1961.

G. Samorodnitsky and M. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Stochastic Modeling Series. Taylor & Francis, 1994. ISBN 9780412051715.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer International Publishing, 2015. ISBN 9783319208275. doi: 10.1007/978-3-319-20828-2.

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

V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large-Scale Optimal Transport and Mapping Estimation. In *International Conference on Learning Representations (ICLR)*, 2018.

U. Şimşekli, A. Liutkus, and A. T. Cemgil. Alpha-stable matrix factorization. *IEEE Signal Processing Letters*, 22(12):2289–2293, 2015.

S. A. Sisson, Y. Fan, and M. A. Beaumont. Overview of Approximate Bayesian Computation. *arXiv e-prints*, art. arXiv:1802.09720, Feb 2018.

J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains. *ACM Trans. Graph.*, 34(4), July 2015. ISSN 0730-0301. doi: 10.1145/2766963.

S. Sra. Directional Statistics in Machine Learning: a Brief Review, 2016.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On Integral Probability Metrics, ϕ-Divergences and Binary Classification, 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010. ISSN 1532-4435.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none):1550 – 1599, 2012. doi: 10.1214/12-EJS722.

V. N. Sudakov. Typical distributions of linear functionals in finite dimensional spaces of high dimension. *Soviet Math. Dokl.*, 19(6):1578 – 1582, 1978.

D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. In *International Conference on Learning Representations*, 2017.

M. Tanaka, A. Francis, F. Luciani, and S. Sisson. Using Approximate Bayesian Computation to Estimate Tuberculosis Transmission Parameters From Genotype Data. *Genetics*, 173:1511–20, 08 2006. doi: 10.1534/genetics.106.055574.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145(2):505–518, 1997. ISSN 0016-6731.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein Auto-Encoders. In *6th International Conference on Learning Representations (ICLR)*, May 2018.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society, Interface / the Royal Society*, 6:187–202, 03 2009. doi: 10.1098/rsif.2008.0172.

G. Uhlmann. *Inside out: inverse problems and applications*, volume 47. Cambridge University Press, 2003.

T. Vayer, R. Flamary, R. Tavenard, L. Chapel, and N. Courty. Sliced Gromov-Wasserstein. In *NeurIPS 2019 - Thirty-third Conference on Neural Information Processing Systems*, volume 32, Vancouver, Canada, Dec. 2019.

C. Villani. *Optimal Transport: Old and New.* Grundlehren der mathematischen Wissenschaften. Springer, 2009 edition, Sept. 2008. ISBN 3540710493.

H. von Weizsäcker. Sudakov's typical marginals, random linear functionals and a conditional central limit theorem. *Probability Theory and Related Fields*, 107(3):313 – 324, Mar 1997.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

W. Wang, Y. Mo, J. A. Ozolek, and G. K. Rohde. Penalized Fisher discriminant analysis and its application to image-based morphometry. *Pattern recognition letters*, 32(15): 2128–2135, 2011.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 11 2019. doi: 10.3150/18-BEJ1065.

J. Wolfowitz. The Minimum Distance Method. *Ann. Math. Statist.*, 28(1):75–88, 03 1957. doi: 10.1214/aoms/1177707038.

S. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 8 2010. ISSN 0028-0836. doi: 10.1038/nature09319.

J. Wu, Z. Huang, W. Li, J. Thoma, and L. Van Gool. Sliced Wasserstein Generative Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

ECOLE DOCTORALE

Titre : La distance de *Sliced-Wasserstein* pour l'Apprentissage Automatique à Grande Echelle : Théorie, Méthodologie et Extensions

Mots clés : Apprentissage Automatique, Transport Optimal, Modélisation Générative

Résumé : De nombreuses méthodes d'inférence statistique et de modélisation générative ont recours à une *divergence* pour pouvoir comparer de façon pertinente deux distributions de probabilité. La *distance de Wasserstein*, qui découle du *transport optimal*, est un choix intéressant, mais souffre de limites computationnelle et statistique à grande échelle. Plusieurs alternatives ont alors été proposées, notamment la *distance de Sliced-Wasserstein* (SW), une métrique de plus en plus utilisée en pratique en raison de ses avantages computationnels. Cependant, peu de travaux ont analysé ses propriétés théoriques. Cette thèse examine plus en profondeur l'utilisation de SW pour des problèmes modernes de statistique et d'apprentissage automatique, avec un double objectif : 1) apporter de nouvelles connaissances théoriques permettant une compréhension approfondie des algorithmes basés sur SW, et 2) concevoir de nouveaux outils inspirés de SW afin d'améliorer son application et sa scalabilité. Nous prouvons d'abord un ensemble de propriétés asymptotiques sur les estimateurs obtenus en minimisant SW, ainsi qu'un théorème central limite dont le taux de convergence est indépendant de la dimension. Nous développons également une nouvelle technique d'inférence basée sur SW qui n'utilise pas la vraisemblance, offre des garanties théoriques et s'adapte bien à la taille et à la dimension des données. Etant donné que SW est couramment estimée par une simple méthode de Monte Carlo, nous proposons ensuite deux approches pour atténuer les inefficacités dues à l'erreur d'approximation : d'une part, nous étendons la définition de SW pour introduire les *distances de Sliced-Wasserstein généralisées*, et illustrons leurs avantages sur des applications de modélisation générative ; d'autre part, nous tirons parti des résultats de *concentration de la mesure* pour formuler une nouvelle approximation déterministe de SW, qui est plus efficace à calculer que la technique de Monte Carlo et présente des garanties non asymptotiques sous une condition de dépendance faible. Enfin, nous définissons la classe générale de *divergences "sliced"* et étudions leurs propriétés topologiques et statistiques ; en particulier, nous prouvons que l'erreur d'approximation de toute divergence *sliced* par des échantillons ne dépend pas de la dimension du problème.

Title : Sliced-Wasserstein Distance for Large-Scale Machine Learning: Theory, Methodology and Extensions

Keywords : Machine Learning, Optimal Transport, Generative Modeling

Abstract : Many methods for statistical inference and generative modeling rely on a *probability divergence* to effectively compare two probability distributions. The *Wasserstein distance*, which emerges from *optimal transport*, has been an interesting choice, but suffers from computational and statistical limitations on large-scale settings. Several alternatives have then been proposed, including the *Sliced-Wasserstein distance* (SW), a metric that has been increasingly used in practice due to its computational benefits. However, there is little work regarding its theoretical properties. This thesis further explores the use of SW in modern statistical and machine learning problems, with a twofold objective: 1) provide new theoretical insights to understand in depth SW-based algorithms, and 2) design novel tools inspired by SW to improve its applicability and scalability. We first prove a set of asymptotic properties on the estimators obtained by minimizing SW, as well as a central limit theorem whose convergence rate is dimension-free. We also design a novel likelihood-free approximate inference method based on SW, which is theoretically grounded and scales well with the data size and dimension. Given that SW is commonly estimated with a simple Monte Carlo scheme, we then propose two approaches to alleviate the inefficiencies caused by the induced approximation error: on the one hand, we extend the definition of SW to introduce the *Generalized Sliced-Wasserstein distances* and illustrate their advantages on generative modeling applications; on the other hand, we leverage *concentration of measure* results to formulate a new deterministic approximation for SW, which is computationally more efficient than the usual Monte Carlo technique and has nonasymptotical guarantees under a weak dependence condition. Finally, we define the general class of *sliced probability divergences* and investigate their topological and statistical properties; in particular, we establish that the sample complexity of any sliced divergence does not depend on the problem dimension.

Institut Polytechnique de Paris
91120 Palaiseau, France