Entropic Optimal Transport in Machine Learning: applications to Distributional Regression, Barycentric Estimation and Probability Matching

Giulia Luise

First Supervisor: Massimiliano Pontil

Second Supervisor: Carlo Ciliberto, John Talbot

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

I, Giulia Luise, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Regularised optimal transport theory has been gaining increasing interest in machine learning as a versatile tool to handle and compare probability measures. Entropy-based regularisations, known as Sinkhorn divergences, have proved successful in a wide range of applications: as a metric for clustering and barycenters estimation, as a tool to transfer information in domain adaptation, and as a fitting loss for generative models, to name a few. Given this success, it is crucial to investigate the statistical and optimization properties of such models. These aspects are instrumental to design new and principled paradigms that contribute to further advance the field. Nonetheless, questions on asymptotic guarantees of the estimators based on Entropic Optimal Transport have received less attention.

In this thesis we target such questions, focusing on three major settings where Entropic Optimal Transport has been used: learning histograms in supervised frameworks, barycenter estimation and probability matching. We present the first consistent estimator for learning with Sinkhorn loss in supervised settings, with explicit excess risk bounds. We propose a novel algorithm for Sinkhorn barycenters that handles arbitrary probability distributions with provable global convergence guarantees. Finally, we address generative models with Sinkhorn divergence as loss function: we analyse the role of the latent distribution and the generator from a modelling and statistical perspective. We propose a method that learns the latent distribution and the generator jointly and we characterize the generalization properties of such estimator. Overall, the tools developed in this work contribute to the understanding of the theoretical properties of Entropic Optimal Transport and their versatility in machine learning.

Impact Statement

The impact of machine learning in everyday life is becoming increasingly more significant. The amount of research towards advances of machine learning has grown immensely in the past decade and spans a broad range of topics both in theory and applications. The topic covered in this thesis is deeply theory-oriented and is of interest mainly within the academic community. The focus of this thesis is on a metric between distributions -derived from the Entropic Optimal Transport problem- that has recently gained a lot of attention in the community. The application of this metric in many machine learning tasks has proved successful. The success has motivated a new line of research that addresses the interplay between Optimal Transport and Machine Learning. The present work advances the understanding of the theoretical properties of loss functions based on Optimal Transport and designs theoretically grounded methodologies that are applied to supervised and unsupervised problems. This is necessary to pose the bases for methods which have abstract guarantees and that are not developed for specific instances only. While Optimal Transport distances have been already thoroughly studied in other applied sciences, such as economics, their application in machine learning is relatively new but very promising. This thesis benefits and contributes to the development of this active line of research. Specific contributions of the thesis are related to three core problems of machine learning, namely distributional regression, barycentric estimation and density fitting. However, the techniques developed can be deployed in other sets of problems. Finally, while the direct impact of this work is more immediate within academia, Optimal transport distances offer a versatile tool to manipulate data: with some improvement of the engineering aspects we expect that the methods developed in this thesis are of interest for practical applications.

Acknowledgements

I wish to thank my supervisors Massimiliano Pontil and Carlo Ciliberto, for their guidance and support in these years; in particular, thanks to Massi, for taking me up as a PhD student and for all the optimism and encouragements. Special thanks to Carlo for getting interested in Optimal Transport, for believing in me when I joined, for all the encouragements along the way and for bearing with my up and downs. None of our meetings ended without a laugh and these three years have truly passed by too fast.

I am deeply grateful to John Talbot, who supported my transfer from Maths to Computer Science.

I really wish to thank John Shawe-Taylor and Rémi Flamary for agreeing on being my examiners, for reading the thesis thoroughly and for their interesting and valuable feedback. I am honored that I had the chance to speak about my work in depth with you.

I am also thankful to Marco Cuturi and Gabriel Peyré, for hosting me in Paris during an internship. It was a pleasure to work with you.

Grazie ai miei amici di sempre. To Caru, you don't know how grateful I am to have you as friend, no matter where life takes us. To Tommi, for best stories on PesoSpecifico, since many years. To Maria, for our catch-ups of months condensed in phonecalls, that nonetheless allow us to be always up-to-date. To Alice and Simona, my childhood friends. I can't believe 2020 ruined our London-based reunion. Thanks to Luca, for sharing ups and downs since the linear algebra exam where we were still kiddos, and now in the asse London-NYC.

Thanks to all the amazing people I met in London. Thanks to Kate for bringing some fun in room 10.09 and for beating the laziness with some weekend runs. To KLB friends, Sean and Jessie; to Marton, for our attempts in organizing a trip to Kew Gardens that never succeeded. To Carmen, we started so well with the gym back in the days and then... And to Niki, my daily companion, for having shared every moment in the last four years and for the numerous to-do lists that we have never implemented. We will, at some point ;)

Thanks to the computer-science friends. And thanks to Anna and Adil, wonderful collaborators, for creating the most inclusive team! Working with you has really been a pleasure. And special thanks to Anna, my role model, who is always up for a chat and for great advice.
Merci à mes amis parisiens. Merci beaucoup à Francois-Pierre for being a great (and incredibly zabetta) office mate during my time in Paris and for teaching me improbable French words that I will surely use very often in the future. And to Gazza for all the Parisian beers, a noteworthy continuation of a good tradition established back in Pavia.
Thanks to my family, for all the support and all our lovely videocalls where I normally see only a wall or the corner of the sofa.
Finally, thanks to Francesco for always being by my side and for uncountable nicknames over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is about Entropic Optimal Transport in Machine Learning. Here we introduce the motivations, the context, and the contributions of this work.

A central question in machine learning is to understand the structure of data and to build models that use such data to solve a specific task. Probability distributions are an integral part in many machine learning problems, both supervised and unsupervised. For example, a classic machine learning task consists in estimating – based on a finite number of observed data $\{x_i\}_{i=1}^n$ in some space $\mathcal{X}$ – the unknown probability distribution from which the data is sampled. A common approach amounts to fitting a parametric model to a dataset, i.e. finding the parameters of a chosen model that fit the observed data in some meaningful way. In practice, to find good parameters that fit a model to the unknown distribution, one has to minimize an error (or loss) function, which quantifies the difference between the estimate and the true distribution. This means that selecting a good notion of distance to compare probabilities is a core part of the problem.

Alternatively, rather than learning the underlying distribution of the observed data, one may be interested in finding patterns, and in grouping the data points according to their similarity. Clustering is a common unsupervised task that aims to automatically divide points into groups (i.e. clusters) with similar properties. Defining what 'similar' means for complex data is challenging. In practice, many data can be represented as histograms, which are discrete probability distributions over a finite set of atoms, or encoded as densities. Examples include text documents, represented as histograms on a fixed vocabulary of n words (Kusner et al., 2015); the colour content of images, defined by the distribution of its pixels in some colour space (Rubner et al., 1997, 2000); words and graphs, that can be

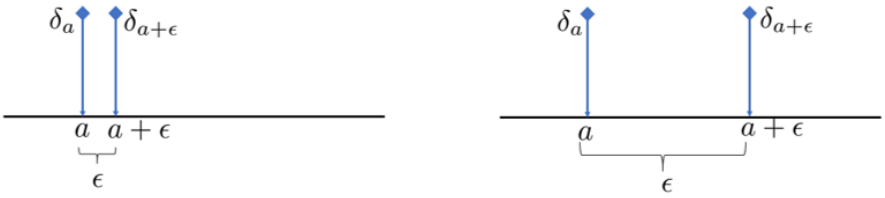


Figure 1.1: Representation of Dirac deltas centered at some $a \in \mathbb{R}$ and $a + \epsilon$ for a small (left) and a bigger (right) value of ϵ.

represented as Gaussian densities (Vilnis and McCallum, 2015; He et al., 2015). A good notion of distance when comparing probabilities is a crucial tool to define similarity for such data. Thus, quantitatively measuring the dissimilarity between two probability measures (histograms, densities or general distributions) in a meaningful way becomes essential. The metric used to quantify this dissimilarity plays a fundamental role, since different distances used in the same problem can yield very different results (Rubner et al., 2000; Cuturi and Doucet, 2014).

1.1 Discrepancies between probability measures

A widely used class of discrepancies between distributions is given by Csizar's f-divergences, which include the well-known Total Variation, Hellinger distance and Kullback-Leibler divergence. Each of them has specific properties that can be advantageous depending on the setting. For example, Total Variation is a useful upper bound of the difference between the probabilities that the two distributions can assign to the same event, Hellinger distance is convenient because of its factorization properties and Kullback-Leibler divergence arises naturally in information theory and maximum likelihood estimation (Gibbs and Su, 2002). However, all of them induce a *strong* notion of convergence, which results in not being stable with respect to some types of changes in the distributions. As illustrative example consider the following case: assume that there is a target distribution represented by a Dirac delta δ_a centered at some point $a \in \mathbb{R}$. With some procedure, one obtains as an *estimate* of such target a Delta centered at $a + \epsilon$, i.e. $\delta_{a+\epsilon}$ (Fig. 1.1). Intuitively, the mismatch in the estimation is more severe as $|\epsilon|$ grows. Using the Total Variation to measure the error results in the following behaviour: the error is 0 if $\epsilon = 0$ and is 1 for any $\epsilon > 0$, irrespective of how big or small $|\epsilon|$ is; thus, the Total Variation does not suitably capture the actual entity of the error. Ideally, a notion of discrepancy that is sensitive to changes in ϵ is favourable. Such behavior is representative of metrics that are 'weak'.

Metrics between probability measures that are labelled as *weak* and are particularly suited to capture the geometric properties of the distributions are Optimal Transport distances. [1]

Optimal Transport (OT) distances are defined starting from an Optimal Transport problem which consists in finding an 'optimal' way to move mass from a distribution α to a distribution β on some domain $\mathcal{X}$. 'Optimal' is to be interpreted with respect to a ground cost function defined on the underlying space $\mathcal{X}$, which establishes the relevant geometric structure. Normally, considering a metric space $(\mathcal{X}, \mathsf{d})$ (for example $(\mathbb{R}^d, \|\cdot\|)$) the cost function corresponds to the ground distance d or to a power of d. This is how the Optimal Transport distances incorporate the geometry of the underlying space in their definition. To clarify the terminology, we point out that when considering as cost function a distance or a power of the distance, Optimal Transport metrics are referred to as Wasserstein distance. When considering a generic cost function, the correct terminology would be Optimal Transport discrepancies. In the following we will use Optimal Transport and Wasserstein interchangeably, with some abuse of terminology. In the Dirac deltas case mentioned above, the Wasserstein distance between $\delta_{(a+\epsilon)}$ and δ_a amounts exactly to the distance between the two support points $|(a+\epsilon) - a| = |\epsilon|$: this reflects the intuition that the distance should be small if the support points are close, and should increase as the support points move away from each other. Also it showcases how the geometry of $\mathcal{X}$ (in this case $\mathbb{R}$) is lifted to impose a geometry on $\mathcal{P}(\mathcal{X})$. Wasserstein distances are then suitable to compare measures with non-overlapping support or supported on low-dimensional manifolds, to include geometric information and to metrize the weak convergence.

Therefore, on paper, Optimal Transport distances satisfy all the requirements for a good notion of distance and they seem to have a great potential in applications. However, real data is inherently partial and limited: one has access to datasets that contain a finite number of observations. When these observations are to be used as proxy for the underlying unknown distributions, it is important that the distance in use enjoys a good approximation power, namely that the distance between two distributions can be accurately estimated using their samples. Optimal Transport distances fail in this aspect, and suffer the so called 'curse of dimensionality': as the dimension of the ground space increases, the empirical measure becomes less and less representative of its continuous counterpart. In addition to this, Optimal Transport distances are affected by a second major drawback which concerns

[1] the term 'distance' will be used with some abuse of terminology to indicate any notion of discrepancy, not necessarily satisfying the axioms of a distance.

the computational complexity: computing OT distances on samples scales cubically on the number of samples (Pele and Werman, 2009), severely limiting scalability. These shortcomings originally hindered the applicability of Optimal Transport distances in large-scale data analysis. The interest for this family of metrics in machine learning community was renewed by the introduction of *regularized* versions, which benefit from computational tractability and better sample complexity, while preserving the same appealing properties as unregularized OT. The most popular regularization consists in adding an entropy penalty in the transport problem: Entropy-regularized Optimal Transport bridges the gap between the geometric flavour of Optimal Transport and the statistical and computational efficiency of other standard divergences commonly used in Machine Learning.

1.2 Entropy-regularized Optimal Transport

While entropy-regularization in transportation and linear programming had been studied since (Schrödinger, 1931; Wilson, 1969; Cominetti and Martín, 1994), it was popularised in the machine learning community by the landmark paper (Cuturi, 2013), which showed the significant computational speed-up achievable using Sinkhorn-Knopp algorithm (Sinkhorn, 1964; Sinkhorn and Knopp, 1967). The new computational solver for entropic OT problems opened the door to active research on Entropic Optimal Transport in different directions: numerical aspects, theoretical properties, and a wide range of applications.

Numerics. The computational speed-up that entropy regularisation offers was first empirically tested and highlighted in Cuturi (2013). Starting from this success, a subsequent line of works has focused on: proposing stochastic algorithms to deal with large scales settings or continuous distributions (Genevay et al., 2016); providing a refined analysis of the computational complexity (Altschuler et al., 2017); in proposing yet other variants or alternatives of Sinkhorn algorithm (Altschuler et al., 2017; Dvurechensky et al., 2018) and strategies to achieve further speed-up in specific settings (Altschuler et al., 2019); in studying stability for small regularization parameters and in developing libraries for state-of-the-art implementations that scale to millions of samples (Feydy et al., 2019).

Theory. A second line of research concerns theoretical properties of Entropic Optimal Transport, which are at the core of designing principled approaches when using entropic OT as a loss function in machine learning. Since in most frameworks distributions are

accessed only through a finite number of samples, questions on the sample complexity and the approximation power of loss functions are of particular interest. Recent results have provided the first answers to these statistical questions: (Genevay et al., 2018a) and the following improvements and extensions in (Mena and Niles-Weed, 2019) showed that the approximation rate is independent of the dimension, contrarily to unregularized Optimal Transport. When dealing with finite realizations or mini-batches, approximation rate is not the only property that matters: using $\mathsf{OT}_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)$ as an approximation of $\mathsf{OT}_\varepsilon(\alpha, \beta)$ results in a biased estimator. The bias similarly affects the estimation of the gradients, making the wide-spread mini-batch approach not theoretically justified. This problem has been addressed in a very recent work (Mensch and Peyré, 2020) that studies a method to obtain *unbiased* estimators for the gradients of Sinkhorn divergence. Besides dealing with samples, most machine learning settings also require to solve some optimization problem: when Entropic Optimal Transport is used to define a loss function, one has to minimize it over some space of probability measures. This means that studying convexity and regularity properties is useful to correctly design optimization procedures. Research on Entropic Optimal Transport is also targeting such questions.

Applications. Advances in theoretical and computational sides have paved the way to numerous applications: recent literature has explored Entropic Optimal Transport as a measure of discrepancy between histograms and probability measures in a range of problems. While this is far from being an exhaustive list, here are some examples:

- *Clustering and barycenters*: Optimal Transport and its entropic approximations have been first used for aggregation and clustering in (Cuturi and Doucet, 2014; Ye et al., 2017). Since Optimal Transport metrics faithfully incorporate the geometry of the underlying domain, they tend to preserve the geometric structure of the input measures when computing their average (Cuturi and Doucet, 2014). This means that computing the average of a set of signals with similar shape but individual noise using Optimal Transport metrics can produce a meaningful representation of the signal which faithfully maintain the profile while balancing out the noise. This advantage has motivated a rich and active line of research on computation of barycenters using Wasserstein and Entropic OT distances. When dealing with histograms, several algorithm which are both fast in practice and fully justified in theory are available (Benamou et al., 2015;

Dvurechenskii et al., 2018). On the other hand, computing barycenters of *arbitrary* distributions is way harder; while the last few years have seen progress in this direction, developing theoretically justified algorithms that cover general settings is still subject of active on-going research.

- *Generative Models:* Generative models are an established approach to learn a probability distribution in a unsupervised way. Given a dataset of examples, the goal is to fit the distribution of a parametric generative model to the unknown distribution induced by such dataset; the model then is able to ideally generate new samples from the unknown distribution. Wasserstein distance and Sinkhorn divergences have been tested as loss functions for this task (Arjovsky et al., 2017; Genevay et al., 2018b). While this is a promising direction, it is also still marginally explored. Challenges arise on multiple levels, both practical and theoretical. In particular, many questions related to asymptotic guarantees of the estimators, to the real performance of Optimal Transport distances in high dimensions and to how the intrinsic dimension of the target ρ affects the performance are still open.

- *Domain adaptation:* domain adaptation is an instance of transfer learning and aims to predict classes on a new (target) dataset that is different from the available (source) training dataset. Optimal Transport, both in its unregularised and regularised versions, has been recently used as an elegant and effective tool to transfer information between domains; the first method based on regularised optimal transport was proposed in Courty et al. (2014). The success of this work led to other extensions, such as (Courty et al., 2017; Redko et al., 2019; Bhushan Damodaran et al., 2018) and recently (Flamary et al., 2019).

Optimal Transport distances in real applications. The notion of 'transport of mass', which Wasserstein distances are based on, has also lead to remarkable applications in natural sciences. Recently, (Schiebinger et al., 2019) has deployed Optimal Transport to design tools for inferring developmental landscapes, probabilistic cellular fates and dynamic trajectories from large-scale single-cell RNA-seq data collected along a time course. Entropy-regularization is used to speed up computations and allows to handle bigger populations of cells. This is a further evidence of the versatility of Optimal Transport distances and of their potential in applied sciences, that is yet to be fully explored.

1.3 Outline of the thesis

This thesis is part of the line of research exploring the interplay between regularised optimal transport and machine learning. The goal of the work is to study Entropic Optimal Transport as a metric in distributional regression, barycenter estimation and density fitting, with a focus on designing estimators with certified theoretical guarantees. In each of the three applications, the theoretical analysis of the relevant estimators or algorithms is based on the development of new regularity properties of Entropic Optimal Transport, which are of independent interest.

As a whole, the thesis contains advances in theoretical study of Entropic Optimal Transport applied to machine learning problems. Statistics and optimization are core components in machine learning and we aim to combine the recent interest in Entropic Optimal Transport with classical questions on these aspects.

1.3.1 Contributions

We present the contributions of this thesis, dividing them into two categories: novel results on theoretical properties of Entropic Optimal Transport and novel approaches to use Entropic Optimal Transport in three standard machine learning applications. The results that fall under 'theory' are instrumental tools used to derive the asymptotic analysis of the methods designed for the targeted applications. Note that since the algorithm used to compute the Entropic Optimal Transport in practice was named after Sinkhorn (Sinkhorn, 1964; Sinkhorn and Knopp, 1967) and is known as Sinkhorn algorithm, Entropic Optimal Transport divergences are also called Sinkhorn divergences. In the following we will use both terminologies, and subtle aspects will be clarified in the background Chapter.

Contributions on Entropic Optimal Transport

On histograms.

- We prove that Entropic Regularized Optimal Transport, in its different formulations presented in the background Chapter, is a smooth function in the interior of the simplex (Chapter 3).
- We use the implicit function theorem to derive a formula for the gradient of *sharp* Sinkhorn divergence and we present an efficient algorithm to compute it in practice (Chapter 3).

On arbitrary distributions.

- We show that the gradient of Sinkhorn divergence is Lipschitz continuous with respect to Total Variation metric. Tools for the proof are based on Perron-Frobenius theory and the Hilbert metric. This result makes it possible to use Sinkhorn divergence in those optimization algorithms that require smoothness of the gradient (Chapter 4).
- We show that the gradient of Sinkhorn divergence is Lipschitz continuous with respect to MMD. The proof uses smoothness results on dual potentials of Sinkhorn divergence that were studied in Genevay et al. (2018a) (Chapter 4).
- As a corollary of the Lipschitz continuity of the gradients, we show a result on the *sample complexity* of the gradients of Sinkhorn divergence. This is useful to quantify the errors that are introduced when using samples as a proxy of the real distribution (Chapter 4).
- On the statistical side, we show the sample complexity of probability measures obtained as pushforward of oracle measures highlighting that the dimension that comes into play is the one of the oracle space (Chapter 5).

Contributions on Entropic Optimal Transport as a metric in learning problems and barycentric estimation

- We study *supervised learning* problems with histograms as outputs and Entropic Optimal Transport as a loss function. Interpreting the problem as a problem with structured outputs and relying on recent results on structured prediction, we propose the first estimator for learning in this setting which is universally consistent. We show excess risk bounds for such estimator and test it in toy image reconstruction setting (Chapter 3).
- Moving to the unsupervised world, we study Sinkhorn divergence as metric to compute the Frechet mean (i.e. barycenter) of a set of *arbitrary* input measures. Relying on Frank-Wolfe algorithm, we propose the first algorithm for Sinkhorn barycenters with free support that does not rely on an alternating procedure and that iteratively populates the barycenter, without using a fixed number of particles. We show convergence rates for both discrete and continuous input measures (Chapter 4).
- We study Sinkhorn divergence as metric for learning a distribution in a unsupervised way, namely using a latent distribution and a generator function as in GAN settings. We propose an estimator based on learning jointly the latent distribution and the generator function and we characterise the asymptotic behaviour of such estimator.

We show that the learning bound is affected by the dimension of the latent space and provide explicit upper bounds under different modelling assumptions (Chapter 5).

1.3.2 Structure of the thesis

To avoid repetitions, we keep the description of the chapters very short. The thesis is structured as follows:

Chapter 2 contains the background material which is needed for the rest of the thesis. We recall definitions and main properties of both Optimal Transport and various formulations of Entropy-regularized Optimal Transport.

Chapter 3 is about Entropic Optimal Transport as loss function in a supervised learning setting where the output set is a space of histograms. We focus on two characterisations of Entropic Optimal Transport that are available in the discrete setting and compare them in the role of loss functions on the space of histograms. We propose an estimator for learning with such losses which is universally consistent. We show excess risk bounds as statistical guarantees of the estimator.

Chapter 4 is dedicated to Sinkhorn barycenters. It presents a method for Sinkhorn barycenter with free support with convergence guarantees, based on Frank-Wolfe algorithm. In contrast to previous free-support methods, our algorithm does not perform an alternate minimization between support and weights. Instead, the Frank-Wolfe (FW) procedure allows to populate the support by incrementally adding new points and to update their weights at each iteration, similarly to kernel herding strategies (Bach et al., 2012). We prove the convergence of the proposed optimization scheme for both finitely and infinitely supported distribution settings.

Chapter 5 is concerned with Sinkhorn divergence as a metric to learn a distribution with a generative model approach. We study sample complexity of a measure obtained as a pushforward of an oracle measure supported on a lower dimensional space. We propose an estimator for jointly learning latent distribution and generator map and provide upper bounds on the generalization error that highlights the potential impact of the latent space on the statistical performance.

Chapter 6 contains the conclusions of the thesis and a glance at future directions, that are divided into two categories: i) questions which are directly motivated and tightly related to the material presented in this manuscript; ii) broader questions which are of general interest in the application of Optimal Transport tools to Machine Learning.

Publications

Published papers during my PhD that are part of the thesis

- G. Luise, A. Rudi, M. Pontil, C. Ciliberto, *Differential Properties of Sinkhorn Approximations for Learning with Wassestein Distance*, NeurIPS 2018
- G. Luise, S. Salzo, M. Pontil, C. Ciliberto, *Sinkhorn Barycenters with Free Support via Frank-Wolfe Algorithm*, NeurIPS 2019 (spotlight)

Submitted papers during my PhD that are part of the thesis

- G. Luise, M. Pontil, C. Ciliberto, *Generalization Properties of Optimal Transport GANs with Latent Distribution Learning*, arXiv:2007.14641

Further work done during the PhD that is unrelated to this thesis:

- G. Luise, D. Stamos, M. Pontil, C. Ciliberto, *Leveraging Low-Rank Relations Between Surrogate Tasks in Structured Prediction*, ICML 2019
- G. Luise, G. Savaré, *Contraction and regularizing properties of heat flows in metric measure spaces*, DCDS - S, doi: 10.3934/dcdss.2020327
- A. Salim, A. Korba, G. Luise, *The Wasserstein Proximal Gradient Algorithm*, NeurIPS 2020
- A. Korba, A. Salim, M. Arbel, G. Luise, A. Gretton, *A Non-Asymptotic Analysis for Stein Variational Gradient Descent*, NeurIPS 2020
- S. Cohen, G. Luise, A. Terenin, B. Amos, M. P. Deisenroth, *Aligning Time Series on Incomparable Spaces*, arXiv:2006.12648.

What this thesis is not about

There is a ton of extremely interesting work in Optimal Transport for Machine Learning that is not mentioned in the thesis. Three of the main fundamental areas that are not touched in this manuscript but are of deep interest for Optimal Transport in Machine Learning are the following:

Map estimation. Historically, the Optimal Transport problem introduced by Monge (Monge, 1781) was formulated as the problem of finding a map T which optimally moves mass from a distribution α to a distribution β. Details on the formulation of the Monge problem, with the related existence issues, can be found in Villani (2008). Inspired by the Monge formulation, in machine learning active research is devoted to how to estimate a good

mapping between two distributions only using their samples, which is –in most cases– all we have access to in practice. The problem has been addressed in an array of works. With no claim to give an exhaustive list, we refer to (Ferradans et al., 2014) in the case of squared Euclidean loss, (Stavropoulou and Muller, 2015) using an approximation of the barycenter mapping and more recently (Seguy et al., 2018) that can handle high scale settings for any ground cost and (Flamary et al., 2019) that is specific to Gaussian measures.

Regularizations that do not involve the entropy: Sliced Wasserstein distance. The computational burden is a major limiting factor in the application of OT distances to large-scale data analysis. However, there are a few cases where computing the Optimal Transport solution is cheap: for instance, Wasserstein distance on *one-dimensional* densities has a closed formula that can be easily computed. This fact has inspired an approximation of Wasserstein distance that exploits this property. Sliced-Wasserstein distances (Rabin et al., 2011) operate by computing (ideally) infinitely many linear projections of the high-dimensional distribution to one-dimensional subspaces and then computing the average of the Wasserstein distance between these one-dimensional projections. The sliced-Wasserstein distance has significantly lower computational cost than unregularized Wasserstein, while maintaining similar theoretical properties. Together with entropic-regularized Optimal Transport, it is an effective option to use Optimal Transport in practice, and there is active research dedicated to the application of Sliced-Wasserstein distance and some generalizations to a variety of tasks: among others, to the barycenter problem (Rabin et al., 2011), generative models (Deshpande et al., 2018; Nadjahi et al., 2019) and auto-encoders (Kolouri et al., 2019).

Optimal Transport or structured data, e.g. graphs. In this thesis we consider probability distributions on comparable domains, meaning that we can always define a cost function between points in such domains. However, many interesting applications require us to deal with distributions that do not live in comparable spaces. Hence, it is not possible to define a ground distance and consequently to use the standard definition of Optimal Transport. However, there exists an extension of Optimal Transport which is suited to this setting, named Gromov-Wasserstein (Mémoli, 2011). This has opened another line of research which targets the application of this extension of OT and variants to structured

data (Vayer et al., 2019b,a, 2020). This direction has the potential of achieving a powerful combination of geometry-aware distances that however can be applied to very complicated data.

Chapter 2

Background Material

The scope of this chapter is to present the concepts, tools and results that will be used in the rest of the thesis. We will discuss in depth definitions and properties of Entropy-regularized Optimal Transport. We will provide proofs of some of the results for completeness. For others, which require more technicalities or long arguments, we will refer to the relevant literature. Some of the concepts that are mentioned and used throughout the manuscript, but that are not core building blocks for the novel material presented in the following chapters, are deferred to Appendix A. Overall, this chapter does not contain novel results that stand as contributions of the thesis.

2.1 Notations

We set some of the notation that we will use in the rest of the thesis. In the following, $\mathcal{X}$ will denote most of the time a compact metric space or a compact subset of $\mathbb{R}^d$. When it has a different meaning, it will be specified case by case. We will use $\mathcal{C}(\mathcal{X})$ to denote the space of real valued continuous functions over $\mathcal{X}$ equipped with the supremum norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. $\mathcal{P}(\mathcal{X})$ and $\mathcal{M}(\mathcal{X})$ denote the space of Radon probability measures and finite signed measures over $\mathcal{X}$ respectively. The notation $\langle \cdot, \cdot \rangle$ is used for both scalar products and for a pairing between Banach spaces. For instance, we will use it to denote the following; the inner product in any Hilbert space $\mathcal{H}$; the inner product in the space of matrices (Frobenius product): given $A, B \in \mathbb{R}^{n\times m}$, $\langle A, B\rangle = \sum_{ij=1}^{n,m} A_{ij}B_{ij}$; the duality pairing between the space of continuous functions and the space of finite signed measures: given $f \in \mathcal{C}(\mathcal{X})$ and $\mu \in \mathcal{M}(\mathcal{X})$, $\langle f, \mu\rangle = \int_{\mathcal{X}} f(x)d\mu(x)$.

2.2 Optimal Transport Distances

In order to motivate *entropic* Optimal Transport theory, which is the main focus of this chapter, we start from classic Optimal Transport.

In this section we recall definition and properties of Optimal Transport distances. While standard Optimal Transport distances are not used in the main chapters, they serve as a preliminary material to introduce their entropic counterpart. In the following $\mathcal{X}$ and $\mathcal{Y}$ will denote domains that in full generality are complete and separable metric spaces. One can think of $\mathcal{X}$ and $\mathcal{Y}$ as subsets of Euclidean spaces for simplicity. Further or different assumptions will be specified throughout the work when needed.

2.2.1 Definition of Kantorovich problem

Definition 2.1 (Pushforward)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two separable metric spaces, $\alpha \in \mathcal{P}(\mathcal{X})$ and $T : \mathcal{X} \to \mathcal{Y}$ a measurable map. We denote by $T_{\#}\alpha$ the pushforward of α via T defined by*

$$T_{\#}\alpha(B) := \alpha\{T^{-1}(B)\} \qquad \forall\, B \in \mathcal{B}(\mathcal{Y}), \tag{2.2.1}$$

where $\mathcal{B}(\mathcal{Y})$ denotes the measurable subsets of $\mathcal{Y}$.

With this notion of pushforward measure, we can introduce the following.

Definition 2.2. *Let $\alpha \in \mathcal{P}(\mathcal{X})$, $\beta \in \mathcal{P}(\mathcal{Y})$ and $\mathsf{c} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a continuous cost function bounded from below. The Optimal Transport cost between α and β under the cost function c is defined as*

$$\mathsf{OT}(\alpha, \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} \mathsf{c}(x, y)\, d\pi(x, y), \tag{2.2.2}$$

where $\Pi(\alpha, \beta)$ is the set of admissible plans defined by $\Pi(\alpha, \beta) := \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \mathfrak{P}_{1\#}\pi = \alpha, \quad \mathfrak{P}_{2\#}\pi = \beta\}$, with $\mathfrak{P}_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and $\mathfrak{P}_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ the projections onto first and second argument; $\#$ denotes the pushforward operation defined in (2.2.1).

The product measures π in $\Pi(\alpha, \beta)$ are called transport plans. The constraints $\mathfrak{P}_{1\#}\pi = \alpha$ and $\mathfrak{P}_{2\#}\pi = \beta$ correspond to constraining the marginals of the product measure π to be α and β. Note that the set $\Pi(\alpha, \beta)$ is nonempty, since the product measure $\alpha \otimes \beta$ always belongs to $\Pi(\alpha, \beta)$. The transport plans for which the minimum in (2.2.2) is attained are called *optimal* plans. Note that the assumptions on the cost function c used in Def. 2.2 are not minimal and one can define (2.2.2) for more general cost functions; all details of the minimal assumptions that guarantee existence of minimizers can be found in (Villani, 2008).

The quantity in (2.2.2) can be seen as a sort of distance between α and β but in general it does not satisfy the axioms of a distance function. However, when c is defined in terms of a distance on the underlying space, it is possible to define a proper notion of distance as follows:

Definition 2.3. *Let $(\mathcal{X}, \mathsf{d})$ be a complete and separable metric space and let $p \in [1, \infty)$. For any probability measures $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ the Wasserstein distance of order p between α and β is defined as*

$$\mathsf{W}_p(\alpha, \beta) = \Big(\inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} \mathsf{d}(x,y)^p \, d\pi(x,y) \Big)^{\frac{1}{p}}. \tag{2.2.3}$$

Note that with the definition above, W_p may take the value $+\infty$. In order to obtain a finite quantity, one has to restrict to the space of measures

$$\mathcal{P}_p(\mathcal{X}) := \{\alpha \in \mathcal{P}(\mathcal{X}) : \quad \int_{\mathcal{X}} \mathsf{d}(x, x_0)^p \, d\alpha(x) < +\infty\},$$

where $x_0 \in \mathcal{X}$ is arbitrary. The proof that W_p satisfies the axiom of a distance can be found in (Villani, 2008, pg 106). Wasserstein distance has had a big echo in applications. In machine learning, for example, it was first 'rediscovered' and used in computer vision problems under the name of Earth-Mover-Distance (EMD) (Rubner et al., 1997, 2000). In the following we may use the 'Optimal Transport cost' and 'Wasserstein distance' interchangeably.

2.2.2 Dual formulation

The minimization problem in (2.2.2) is sometimes referred to as *primal problem*. It admits a *dual* formulation, known as Kantorovich duality. In order to present the result on Kantorovich duality, we introduce the notion of c-transform below.

Definition 2.4. *Let $\mathsf{c} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. A function $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be c-convex if it is not identically $\{+\infty\}$ and there exists a function $\zeta : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ such that*

$$\psi(x) = \sup_{y \in \mathcal{Y}} \Big(\zeta(y) - \mathsf{c}(x,y) \Big) \qquad \forall\, x \in \mathcal{X}. \tag{2.2.4}$$

Then its c-transform is the function ψ^{c} defined by

$$\psi^{\mathsf{c}}(y) = \inf_{x \in \mathcal{X}} \Big(\psi(y) + \mathsf{c}(x,y) \Big) \qquad \forall\, y \in \mathcal{Y}. \tag{2.2.5}$$

Theorem 2.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two complete and separable metric spaces. Let $\alpha \in \mathcal{P}(\mathcal{X})$*

and $\beta \in \mathcal{P}(\mathcal{Y})$ *and* $\mathsf{c} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ *a cost function satisfying the assumptions mentioned before. Then,*

$$\min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} \mathsf{c}(x,y)\, d\pi(x,y) = \sup_{(\psi,\phi)\in L^1(\alpha)\times L^1(\beta):\, \psi-\phi\leq \mathsf{c}} \Big(\int_{\mathcal{Y}} \phi(y)\, d\beta(y) - \int_{\mathcal{X}} \psi(x)\, d\alpha(x) \Big), \tag{2.2.6}$$

$$= \sup_{\psi \in L^1(\alpha)} \int_{\mathcal{Y}} \psi^{\mathsf{c}}(y)\, d\beta(y) - \int_{\mathcal{X}} \psi(x)\, d\alpha(x) \tag{2.2.7}$$

$$= \sup_{\phi \in L^1(\beta)} \int_{\mathcal{Y}} \phi(y)\, d\beta(y) - \int_{\mathcal{X}} \phi^{\mathsf{c}}(x)\, d\alpha(x) \tag{2.2.8}$$

where $\psi - \phi \leq \mathsf{c}$ *is to be interpreted as* $\psi(x) - \phi(y) \leq \mathsf{c}(x,y)$ *for any* $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

The characterization of Kantorovich duality is richer than the one reported here and it constitutes a core result in Optimal Transport theory. For a complete statement and the proof we refer to (Villani, 2008, Thm. 5.10). Note that if $\mathsf{OT}(\alpha,\beta) < +\infty$ and there exists two functions $\tilde{a} \in L^1(\alpha)$ and $\tilde{b} \in L^1(\beta)$ such that $\mathsf{c}(x,y) \leq \tilde{a}(x) + \tilde{b}(y)$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, then the supremum on the right hand side is a maximum.
When considering a metric space $(\mathcal{X}, \mathsf{d})$ and taking $\mathsf{c} = \mathsf{d}$, the dual formulation of Wasserstein-1 distance between $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ simplifies as

$$\mathsf{W}_1(\alpha,\beta) = \sup_{\psi:\, 1-\text{Lipschitz}} \Big(\int_{\mathcal{X}} \psi\, d\alpha - \int_{\mathcal{X}} \psi\, d\beta \Big). \tag{2.2.9}$$

The formula above is known as *Kantorovich-Rubenstein* and it is useful in a variety of settings (for example it is the formula used in the first paradigm of Generative Adversarial Networks using Wasserstein distance (Arjovsky et al., 2017)).

2.2.3 Convergence in Wasserstein sense

The convergence of probability measures is of crucial importance in a variety of problems. There exist different notions of convergence and they result in very different meaning of 'what is close' and 'what is far' in the space of probability measures. A very important type of convergence is the 'weak' convergence, defined below.

Definition 2.5. *Let* $(\mathcal{X}, \mathsf{d})$ *be a metric space. A sequence of probability measures* $\mu_k \in \mathcal{P}(\mathcal{X})$

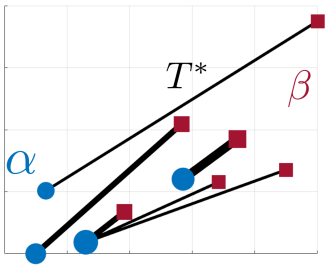


Figure 2.1: Optimal transport problem between two discrete distributions α and β represented by blue circles and red squares respectively. The area of the markers is proportional to the weight at each location. That plot displays the optimal transport plan T^* using a quadratic Euclidean cost.

weakly converges to $\mu \in \mathcal{P}(\mathcal{X})$ (and we write $\mu_k \rightharpoonup \mu$) if

$$\int_{\mathcal{X}} \psi \, d\mu_k \to \int_{\mathcal{X}} \psi \, d\mu \tag{2.2.10}$$

for any continuous and bounded function ψ. Weak convergence in $\mathcal{P}_p(\mathcal{X})$ is defined similarly, but with an extra condition: namely, μ_k weakly converges to μ in $\mathcal{P}_p(\mathcal{X})$ if

$$\mu_k \rightharpoonup \mu \quad \text{and} \quad \int_{\mathcal{X}} \mathsf{d}(x, x_o)^p \, d\mu_k(x) \to \int_{\mathcal{X}} \mathsf{d}(x, x_o)^p \, d\mu(x). \tag{2.2.11}$$

Proposition 2.2. *Let $(\mathcal{X}, \mathsf{d})$ be a metric space and $p \in [1, +\infty)$. The W_p distance metrizes the weak convergence in $\mathcal{P}_p(\mathcal{X})$. This means that if $\{\mu_k\}_k$ is a sequence of probability measures in $\mathcal{P}_p(\mathcal{X})$ and $\mu \in \mathcal{P}_p(\mathcal{X})$ is another measure, then the following are equivalent:*

- $\mathsf{W}_p(\mu_k, \mu) \to 0$
- *$\mu_k \rightharpoonup \mu$ in $\mathcal{P}_p(\mathcal{X})$.*

The proof of the result above can be found in (Villani, 2008, Thm. 6.9). An immediate consequence of the result above is the continuity of W_p: if μ_k (resp. ν_k) weakly converges to μ (resp. ν) in $\mathcal{P}_p(\mathcal{X})$ as $k \to +\infty$, then $\mathsf{W}_p(\mu_k, \nu_k) \to \mathsf{W}_p(\mu, \nu)$.

2.2.4 Discrete case and computational complexity

So far, we have introduced Optimal Transport and Wasserstein distance in full generality. In machine learning applications, many times the data is discrete: in the following we specify the definitions in this setting. Consider the probability simplex defined as $\Delta_r := \{v \in \mathbb{R}^r : v_i \geq 0, \ \sum v_i = 1\}$. Let α and β be two discrete measures defined as $\alpha = \sum_{i=1}^n \mathsf{a}_i \delta_{x_i}$

and $\beta = \sum_{j=1}^{m} \mathsf{b}_j \delta_{y_j}$ with $x_i \in \mathcal{X}$ for $i = 1, \dots, n$, $y_j \in \mathcal{Y}$ for $j = 1, \dots, m$ and $\mathsf{a} = (\mathsf{a}_i, \dots, \mathsf{a}_n) \in \Delta_n$, $\mathsf{b} = (\mathsf{b}_1, \dots, \mathsf{b}_m) \in \Delta_m$. We refer to the points x_i and y_j as *locations* and to the vectors a and b as *weights*. Optimal Transport problem in the discrete case reads as follows:

$$\mathsf{OT}(a, b) = \min_{T \in \Pi(\mathsf{a}, \mathsf{b})} \langle T, C \rangle \tag{2.2.12}$$

where $\Pi(\mathsf{a}, \mathsf{b}) := \{T \in \mathbb{R}^{n \times m} : T_{ij} \geq 0 \ , T\mathbb{1}_m = \mathsf{a}, \ \ T^\top \mathbb{1}_n = \mathsf{b}\}$ and C is the cost matrix defined by $C_{ij} = \mathsf{c}(x_i, y_j)$. The set of matrices $\Pi(\mathsf{a}, \mathsf{b})$ is bounded and defined by $n + m$ constraints (one of which is redundant). It is a convex set which is called *transport polytope*. The primal problem in (2.2.12) corresponds to a linear program (Bertsimas and Tsitsiklis, 1997). Intuitively, solutions of the primal problem prescribe a way of 'transporting' mass between the two distributions, as shown in Fig. 2.1.

Dual formulation. In the discrete case, the dual formulation corresponding to (2.2.12) is a constrained maximization problem. Given α and β as above, it reads as

$$\mathsf{OT}(\alpha, \beta) = \max_{u, v \in \mathcal{R}(C)} \langle \mathsf{a}, u \rangle + \langle \mathsf{b}, v \rangle \tag{2.2.13}$$

where $\mathcal{R}(C) := \{(u, v) \in \mathbb{R}^n \times \mathbb{R}^m : \ u_i + v_j \leq C_{ij} \text{ for all } i = 1, \dots, n, \ j = 1, \dots m\}$. The result is an application of the strong duality result for linear program, see (Bertsimas and Tsitsiklis, 1997, Thm. 4.4) and (Peyré and Cuturi, 2019, Prop. 2.4) for more details.

Several algorithms can be used to solve problem (2.2.12). Algorithmic aspects of standard Optimal Transport are unrelated to the scope of this thesis. Therefore for a detailed discussion on the algorithmic side and on OT solvers we refer to (Peyré and Cuturi, 2019, Chapter 3) and here we just briefly comment on the computational complexity aspects, which pave the way to the introduction of *regularized* variants of Optimal Transport. With no specific assumptions on the cost matrix C, the computational complexity scales as $O(n^3 \log(n))$ when computing the distance between a pair of histograms of dimension n (Pele and Werman, 2009). While active research is devoted to improve the computational complexity under some constraints on the cost matrix or when computing the distance up to some approximation error, the computational cost constitutes an issue in large scale applications. This issue, together with the statistical aspects described in the next section, is the core motivation behind the introduction of entropy-regularization.

2.2.5 Statistical Properties

When dealing with probability measures, an important property of metrics is their estimation power, meaning how quickly the empirical measure obtained from n independent samples from a given measure μ approaches μ in that distance. Also, in many practical settings, it is required to quantify how much two probability distributions μ and ν differ but μ and ν are only accessed via empirical or discretized measures μ_n and ν_n composed of n atoms. By the weak convergence result in Prop. 2.2, it follows that $\mathsf{W}_p(\mu_n, \mu) \to 0$ when $n \to +\infty$, since the empirical measure weakly converges to the population measure (Varadarajan, 1958). A key point is to *quantify* the rate, capturing how good the estimation of μ via μ_n is, in terms of the number of samples. This is not a favorable aspect for Optimal Transport distances which suffer from a *curse of dimensionality*. In the high dimensional regime, the empirical distribution μ_n becomes less and less representative as the dimension of the ambient space becomes large. An extensive review of sample complexity results on Wasserstein distances is beyond the scope of this work; here we just recall a couple of main results that clearly highlight the dependence on the dimension. Results by Dudley (1969) show that the curse of dimensionality is unavoidable and prove the lower bound $\mathsf{W}_1(\mu_n, \mu) \gtrsim n^{-\frac{1}{d}}$ for a probability measure μ on $\mathbb{R}^d$, with $d > 2$, which is absolutely continuous with respect to Lebesgue measures. A subsequent strand of works including (Boissard and Le Gouic, 2014; Weed and Bach, 2019) extended the results in Dudley (1969) to W_p. Also, a refined statement is provided in Weed and Bach (2019) for measures supported in low-dimensional manifolds, showing that the rate of approximation depends on the intrinsic dimension of the support. Overall the statistical behaviour of Wasserstein distances is not appealing when dealing with high dimensional data. This aspect, together with the computational burden, is alleviated when *regularizing* the Optimal Transport problem with entropic penalty, which is presented in the next section.

2.3 Entropic Regularized Optimal Transport

The take-home message from the previous section is that Optimal Transport distances have several good properties but also severe drawbacks. In particular, they have a rich duality formulation, they are easy to bound from above (being defined as a minimum) and they incorporate geometric information of the underlying domain, through the cost function. On the other hand, they exhibit a high computational cost and a 'curse of dimensionality' when

it comes to approximating the distance between two distributions only using finite samples. These drawbacks can be partially alleviated introducing a suitable regularization term. In this thesis we consider the entropic regularization, which operates adding to the original problem a penalty based on the entropy.

Short history of Entropic regularization. The entropic Optimal Transport problem first appeared in Schrödinger (1932), in connection with the problem of finding the most likely evolution of a particle configuration. It was later deployed for modelling purposes in transportation theory (Wilson, 1969), under the name of entropy maximizing models. Entropy was introduced in order to achieve a more accurate model when considering traffic patterns. The actual traffic patterns do not match with the ones predicted using optimal transport solutions. The former appear more diffuse than the latter, which concentrate the traffic in a few routs as a result of the sparsity of optimal plans. Entropic regularization was introduced in order to mitigate the sparsity. At a later stage, it proved useful in other applications and recently re-injected interest in Optimal Transport theory in the machine learning community (Cuturi, 2013). As a matter of terminology: in the following (and also in the literature) the discrepancy resulting from the Entropic Optimal Transport problem is referred to as Sinkhorn approximation or Sinkhorn divergence or Sinkhorn distance; the name 'Sinkhorn' is related to the algorithm used to find the solution of the entropic regularized problem, while the nomenclature 'divergence' or 'distance' is an abuse of terminology since -strictly speaking- entropic Optimal Transport as defined in (2.3.1) is neither of those. However, Sinkhorn divergence is the correct name for a variant of Entropic OT presented in Sec. 2.5, which indeed satisfies the requirement of a divergence.

2.3.1 Definition in the general setting

We introduce entropic optimal transport in its general formulation for arbitrary measures. Let $\mathcal{X}$ be a compact metric space. While we will stick to this general notion, one can think of $\mathcal{X}$ as a compact subset in some Euclidean space $\mathbb{R}^d$.

Definition 2.6 (Entropic Regularized Optimal Transport)**.** *For $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, the entropy-regularized Optimal Transport distance is defined as*

$$\mathsf{OT}_\varepsilon(\alpha,\beta) = \min_{\pi\in\Pi(\alpha,\beta)} \int \mathsf{c}(x,y)\, d\pi(x,y) + \varepsilon \mathsf{KL}(\pi \mid \alpha\otimes\beta), \tag{2.3.1}$$

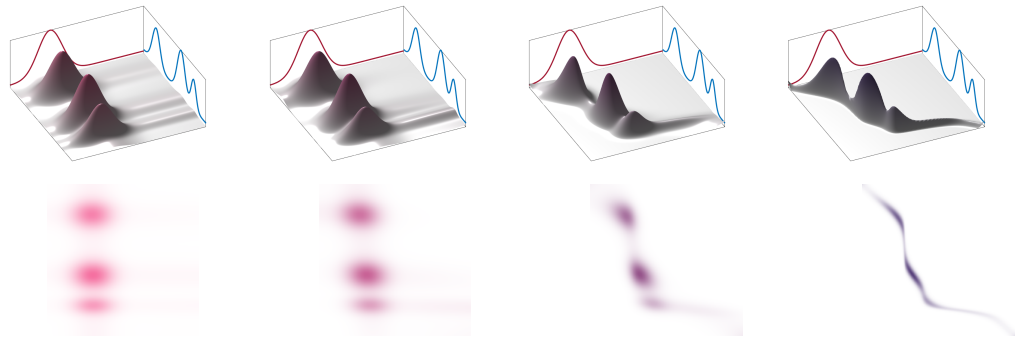


Figure 2.2: Impact of the parameter ε on the optimal transport plan between two 1-dimensional densities. From left to right $\varepsilon = 1$, $\varepsilon = 0.6$, $\varepsilon = 0.3$ and $\varepsilon = 10^{-3}$.

where $\mathsf{KL}(\pi \mid \xi)$ *denotes the Kullback-Leibler divergence between the plan* π *and the reference measure* ξ *and is defined by*

$$\mathsf{KL}(\pi \mid \xi) = \begin{cases} \displaystyle\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\xi}\right) d\pi + \int_{\mathcal{X}\times\mathcal{Y}} \big(d\xi(x,y) - d\pi(x,y)\big) & \text{if } \pi \ll \xi \\ +\infty & \text{otherwise.} \end{cases} \tag{2.3.2}$$

The measure $\alpha \otimes \beta$ in (2.3.1) is used as *reference* measure. Note that the reference measure itself plays no specific role, only its *support* does. Indeed for any $\pi \in \Pi(\alpha, \beta)$ and for any (α', β') with the same support as (α, β) (hence such that they both have densities with respect to one another) we have that $\mathsf{KL}(\pi \mid \alpha \otimes \beta) = \mathsf{KL}(\pi \mid \alpha' \otimes \beta') - \mathsf{KL}(\alpha \otimes \beta \mid \alpha' \otimes \beta')$. Intuitively, the entropic regularization encourages a 'smoothing' of the optimal coupling, as shown in Sec. 2.3.1.

2.3.2 Dual formulation

Similarly to standard Optimal Transport problem, the regularized counterpart admits both a primal and a dual formulation. The dual formulation of standard OT, recalled in (2.2.6), is a constrained problem. When adding entropic penalty, the constraint is replaced by a smooth penalty, and the dual formulation becomes *unconstrained*.

Theorem 2.3. *Let* $\varepsilon > 0$. *The dual problem of* (2.3.1)*, in the sense of Fenchel-Rockafellar, is (Chizat et al., 2018; Feydy et al., 2019)*

$$\mathsf{OT}_\varepsilon(\alpha, \beta) = \sup_{u,v \in \mathcal{C}(\mathcal{X})} \int u(x)\, d\alpha(x) + \int v(y)\, d\beta(y) - \varepsilon \int e^{\frac{u(x)+v(y)-\mathfrak{c}(x,y)}{\varepsilon}}\, d\alpha(x) d\beta(y) + \varepsilon, \tag{2.3.3}$$

where $\mathcal{C}(\mathcal{X})$ *denotes the space of real-valued continuous functions on the domain* $\mathcal{X}$*, endowed*

with the supremum norm $\|\cdot\|_\infty$.

Proof. The proof is an application of the Fenchel-Rockafellar theorem (Rockafellar, 1974, Thm. 19 - 20) (which is recalled in Appendix A). We sketch the following steps here: let $\mathcal{C}(\mathcal{X})$ be the space of continuous functions over $\mathcal{X}$, let $\mathcal{M}(\mathcal{X})$ be the set of finite signed measures over $\mathcal{X}$ and $\mathcal{M}_+(\mathcal{X})$ the set of finite positive measures over $\mathcal{X}$. The function $\mathsf{KL}(\cdot \mid \alpha \otimes \beta)$ can be extended on $\mathcal{M}(\mathcal{X})$ setting it equal to $+\infty$ outside of $\mathcal{M}_+(\mathcal{X})$. Then, problem (2.3.1) can be rewritten as

$$\min_{\pi \in \mathcal{M}(\mathcal{X})} \langle \mathsf{c}, \pi \rangle + \varepsilon \mathsf{KL}(\pi \mid \alpha \otimes \beta) + \iota_{\{(\alpha,\beta)\}}(B\pi), \tag{2.3.4}$$

where $B : \mathcal{M}(\mathcal{X} \times \mathcal{X}) \to \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is the bounded operator defined by $B\pi := (\mathfrak{P}_{1\#}\pi, \mathfrak{P}_{2\#}\pi)$ (with $\mathfrak{P}_i : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$, $i = 1, 2$ the projections onto first and second argument respectively) and ι is the indicator function defined by $\iota_K(q) = 0$ if $q \in K$ and $+\infty$ otherwise. Note that $\mathcal{M}(\mathcal{X})$ and $\mathcal{C}(\mathcal{X})$ are topologically paired under the weak* topology. Hence, we can apply Thm. A.1 choosing (in the notation of Thm. A.1) $V = \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X})$ (and hence $V^* = \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$), $W = \mathcal{C}(\mathcal{X} \times \mathcal{X})$ (and hence $W^* = \mathcal{M}(\mathcal{X} \times \mathcal{X})$), and $A^* = B$ (and therefore $A : \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X} \times \mathcal{X})$ defined by $(u, v) \mapsto ((x, y) \mapsto u(x) + v(y))$). A related proof with more details can be found in (Chizat et al., 2018, Thm. 3.2). □

First order optimality conditions. Let $\mu \in \mathcal{P}(\mathcal{X})$. We denote by $\mathsf{P}_\mu : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ the map such that, for any $w \in \mathcal{C}(\mathcal{X})$,

$$\mathsf{P}_\mu(w) : x \mapsto -\varepsilon \log \int e^{\frac{w(y) - \mathsf{c}(x,y)}{\varepsilon}} d\mu(y). \tag{2.3.5}$$

The first order optimality conditions for (2.3.3) are (see (Feydy et al., 2019) or Appendix A.4.2)

$$u = \mathsf{P}_\beta(v) \quad \alpha \text{ - a.e.} \qquad \text{and} \qquad v = \mathsf{P}_\alpha(u) \quad \beta \text{ - a.e,} \tag{2.3.6}$$

which correspond to

$$\begin{cases} e^{-\frac{u(x)}{\varepsilon}} = \int_{\mathcal{X}} e^{\frac{v(y)-\mathsf{c}(x,y)}{\varepsilon}} \, d\beta(y) & \forall\, x \in \text{supp}(\alpha) \\ e^{-\frac{v(y)}{\varepsilon}} = \int_{\mathcal{X}} e^{\frac{u(x)-\mathsf{c}(x,y)}{\varepsilon}} \, d\alpha(x) & \forall\, y \in \text{supp}(\beta). \end{cases} \tag{2.3.7}$$

Pairs (u, v) satisfying (2.3.6) exist and are referred to as *dual potentials* or *Sinkhorn potentials*. The proof of existence is nontrivial. Since the dual formulation is of crucial importance for several results presented throughout the work, we present the existence result for completeness; however, since the proof relies on technical lemmas, we defer it to Appendix A.4. A useful reference for this point is (Knopp and Sinkhorn, 1968).

The operators P_α, P_β correspond to Soft-min operators of strength ε. These are relaxations of the c-transform defined in Def. 2.4.

Remark 2.1. Extension to the whole domain. *Note that solutions of* (2.3.3) *are defined* α *and* β *almost everywhere* (*see* (2.3.6)). *However, they can be extended to be defined* everywhere *in a canonical way using the formulas* (2.3.6)*, i.e.*

$$u = \mathsf{P}_\beta(v), \qquad \textit{and} \qquad v = \mathsf{P}_\alpha(u) \quad \text{on the whole } \mathcal{X}. \tag{2.3.8}$$

This is a subtle and important point when defining the gradients of OT_ε *(see section Sec. 2.3.5).*

Assuming the existence, we now show that the potentials extended on the whole domain are unique up to a constant.

Proposition 2.4. *(Feydy et al., 2019, Prop 11) Let* (u_0, v_0) *and* (u_1, v_1) *be two pairs of functions satisfying* (2.3.8)*. Then there exists a constant* $K \in \mathbb{R}$ *such that*

$$u_0 = u_1 - K \qquad \text{and} \qquad v_0 = v_1 + K. \tag{2.3.9}$$

Proof. For $t \in [0, 1]$ let $u_t = u_0 + t(u_1 - u_0)$ and $v_t = v_0 + t(v_1 - v_0)$. Set $\phi : \mathbb{R} \to \mathbb{R}$ the function defined as

$$\phi(t) = \langle \alpha, u_t \rangle + \langle \beta, v_t \rangle - \varepsilon \left\langle \alpha \otimes \beta, \exp\left(\frac{u_t \oplus v_t - \mathsf{c}}{\varepsilon}\right) - 1 \right\rangle, \tag{2.3.10}$$

where we write $u_t \oplus v_t$ to denote the function $(x, y) \mapsto u_t(x) + v_t(y)$. Now, ϕ is concave and it is bounded from above by $\phi(0) = \phi(1) = \mathsf{OT}_\varepsilon(\alpha, \beta)$, being (u_0, v_0) and (u_1, v_1) pairs of optimal potentials. That means that ϕ is constant. Therefore,

$$0 = \phi''(t) = \frac{1}{\varepsilon}\left\langle \alpha \otimes \beta, \exp\left(\frac{u_t \oplus v_t - \mathsf{c}}{\varepsilon}\right)(u_1 - u_0 \oplus v_1 - v_0)^2 \right\rangle. \tag{2.3.11}$$

The equation above is satisfied if and only if

$$(u_1(x) - u_0(x) + v_1(y) - v_0(y))^2 = 0 \quad \alpha \otimes \beta\text{-a.e.}\,, \tag{2.3.12}$$

i.e. if and only if there exists a constant $K \in \mathbb{R}$ such that

$$u_0 = u_1 - K \quad \alpha\text{-a.e.} \qquad \text{and} \qquad v_0 = v_1 + K \qquad \beta\text{-a.e.} \tag{2.3.13}$$

Noting that T_μ defined in (2.3.5) is such that for any $\mu \in \mathcal{P}(\mathcal{X})$, $w \in \mathcal{C}(\mathcal{X})$ and $K \in \mathbb{R}$

$$\mathsf{P}_\mu(w + K) = \mathsf{P}_\mu(w) + K, \tag{2.3.14}$$

the results extends to the whole domain $\mathcal{X}$. □

Relation between primal and dual. Fenchel-Rockafellar duality also yields the following link between the solutions of primal and dual problem: let π a solution of the primal problem (2.3.1) and u, v a pair of optimal dual potentials. Then the primal-dual link is given by

$$d\pi(x, y) = e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\alpha(x) d\beta(y). \tag{2.3.15}$$

This relation leads to a neat expression of OT_ε in terms of the optimal dual potentials provided below.

Proposition 2.5. *Let (u, v) be a pair of optimal potentials. Then*

$$\mathsf{OT}_\varepsilon(\alpha, \beta) = \langle u, \alpha \rangle + \langle v, \beta \rangle\,. \tag{2.3.16}$$

Proof. By the primal dual link, we have that when (u, v) is an optimal pair, then

$$\begin{aligned}\mathsf{OT}_\varepsilon(\alpha,\beta) &= \int u(x)\,d\alpha(x) + \int v(y)\,d\beta(y) - \varepsilon \int e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}}\,d\alpha(x)d\beta(y) + \varepsilon \\ &= \int u(x)\,d\alpha(x) + \int v(y)\,d\beta(y) - \varepsilon \int d\pi(x,y) + \varepsilon,\end{aligned}$$

which gives the desired equality since $\int d\pi(x, y) = 1$. □

2.3.3 Properties of dual potentials

While the proof of existence is postponed to Appendix A.4, here we assume that pairs of dual potentials exist and we present some useful properties. In this section $\mathcal{X}$ will always be a compact metric space.

As mentioned above, pairs of optimal potentials are unique (α, β) - a.e. up to additive constant, i.e. if (u, v) is a pair of dual potentials, then $(u + K, v - K)$ is also a solution for any $K \in \mathbb{R}$.

Proposition 2.6. *Let (u, v) be a pair of potentials which are canonically extended on $\mathcal{X}$ via (2.3.6). Then,*

$$u(x) \in [-\max_y(v(y) - \mathsf{c}(x,y)), -\min_y(v(y) - \mathsf{c}(x,y))] \quad \text{for any } x \in \mathcal{X}. \tag{2.3.17}$$

Proof. The proof follows from the optimality conditions (2.3.7). □

Proposition 2.7. *Assume that the cost function $\mathsf{c} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is κ-Lipschitz. Let (u, v) be a pair of solutions of (2.3.3). Then both u and v are κ-Lipschitz on $\mathcal{X}$.*

Proof. We show the result for u only, since the argument for v is identical. Recall that $u(x) = \mathsf{P}_\beta(v)(x)$. Computing the gradient of u we obtain

$$\nabla u(x) = \int \nabla_x \mathsf{c}(x,y) e^{\frac{-\mathsf{c}(x,y)+v(y)}{\varepsilon}}\,d\beta(y) \tag{2.3.18}$$

and hence

$$\|\nabla u(x)\| \le \|\nabla_x \mathsf{c}(x,y)\| \int e^{\frac{-\mathsf{c}(x,y)+v(y)}{\varepsilon}}\,d\beta(y). \tag{2.3.19}$$

By primal constraint we have that $\int e^{\frac{-\mathsf{c}(x,y)+v(y)}{\varepsilon}}\,d\beta(y) = 1$ and from Lipschitz property of c we have that $\|\nabla_x \mathsf{c}(x, y)\| \le \kappa$. By the mean value theorem, it follows that u is κ-Lipschitz

as desired. □

Another important property of Sinkhorn potentials is their continuity with respect to the initial measure, as proved below.

Proposition 2.8. *Let $\{\alpha_n\}$ and $\{\beta_n\}$ two sequences in $\mathcal{P}(\mathcal{X})$ such that $\alpha_n \rightharpoonup \alpha$ and $\beta_n \rightharpoonup \beta$ for some $\alpha, \beta \in \mathcal{P}(\mathcal{X})$. Given some anchor point $x_o \in \mathcal{X}$, let (u_n, v_n) be the unique pair of optimal potentials for $\mathsf{OT}_\varepsilon(\alpha_n, \beta_n)$ such that $u_n(x_o) = 0$. Then $\{u_n\}$ and $\{v_n\}$ uniformly converge to the pair (u, v) of optimal potentials of $\mathsf{OT}_\varepsilon(\alpha, \beta)$ such that $u(x_o) = 0$.*

Proof. By the proposition above, u_n and v_n are κ-Lipschitz functions on the compact bounded domain $\mathcal{X}$. Setting $u_n(x_o) = 0$, we can bound $|u_n|$ by κ times the diameter $|\mathcal{X}|$ of $\mathcal{X}$. Using the fact that $v_n = \mathsf{P}_{\alpha_n}(u_n)$ and the bound on $|u_n|$, we can find a constant that bounds both $|u_n|$ and $|v_n|$ uniformly. Hence we have that $\{u_n\}$ and $\{v_n\}$ are uniformly bounded and equicontinuous (since they are Lipschitz with same constant). Applying Ascoli-Arzela theorem to the sequence $\{(u_n, v_n)\}_n$ we obtain that there exists a subsequence $\{(u_{n_k}, v_{n_k})\}_k$ indexed by k that converges uniformly to a pair (u, v) of continuous functions. Also $u(x_o) = 0$. Since $\beta_n \rightharpoonup \beta$ ($\alpha_n \rightharpoonup \alpha$ respectively) and u_{n_k} (v_{n_k} respectively) strongly converge to u (v respectively), by (Brezis, 2010, Prop. 3.5,(iv)) we have that $u = \mathsf{P}_\beta(v), v = \mathsf{P}_\alpha(u)$ and hence (u, v) is the unique pair of optimal potentials for (α, β) with $u(x_o) = 0$. Since the limit is unique, we conclude that the whole sequence $\{(u_n, v_n)\}$ uniformly converges to (u, v). □

In line with (Genevay et al., 2018a; Feydy et al., 2019) it will be useful in the following to assume (u, v) to be the Sinkhorn potentials such that: $i)$ $u(x_o) = 0$ for an arbitrary anchor point $x_o \in \mathcal{X}$ and $ii)$ (2.3.6) is satisfied pointwise on the entire domain $\mathcal{X}$. Then, u is a fixed point of the map $\mathsf{P}_{\beta\alpha} = \mathsf{P}_\beta \circ \mathsf{P}_\alpha$ (analogously for v). This suggests a fixed point iteration approach to minimize (2.3.3), yielding the well-known Sinkhorn-Knopp algorithm which has been shown to converge linearly in $\mathcal{C}(\mathcal{X})$ (Sinkhorn and Knopp, 1967; Knopp and Sinkhorn, 1968) and which is recalled in section Sec. 2.4.

2.3.4 Note on the Discrete setting

We dedicate this section to a brief discussion of entropy regularization in the discrete setting. Indeed when restricting to the discrete case, the literature presents subtle variants of Entropic-Regularized Optimal Transport.

We denote the probability simplex as

$$\Delta_r := \{a \in \mathbb{R}_+^r : \sum_{i=1}^{r} a_i = 1\}. \tag{2.3.20}$$

Let $\mathcal{X}$ denote some domain. A discrete measure α is a weighted sum of Dirac deltas, i.e.

$$\alpha = \sum_{i=1}^{n} \mathsf{a}_i \delta_{x_i}, \tag{2.3.21}$$

where $x_1, \ldots, x_n \in \mathcal{X}$ and $\mathsf{a} = (\mathsf{a}_1, \ldots, \mathsf{a}_n) \in \Delta_n$. Let $\alpha = \sum_{i=1}^{n} \mathsf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{m} \mathsf{b}_j \delta_{y_j}$, with $\mathsf{a} \in \Delta_n$ and $\mathsf{b} \in \Delta_m$ be two discrete measures. Recall that the transport polytope $\Pi(\alpha, \beta)$ in this setting amounts to

$$\Pi(\alpha, \beta) = \Pi(\mathsf{a}, \mathsf{b}) := \{T \in \mathbb{R}_+^{n \times m} \text{ such that } T\mathbb{1}_m = \mathsf{a},\ T^\top \mathbb{1}_n = \mathsf{b}\}. \tag{2.3.22}$$

Note that transport plans T in $\Pi(\mathsf{a}, \mathsf{b})$ have nm variables and $n+m$ constraints one of which is redundant.

Regularization with Relative Entropy. The discrete counterpart of the Entropic regularized Optimal Transport problem presented in (2.3.1) is the following:

$$\mathsf{OT}_\varepsilon(\alpha, \beta) = \min_{T \in \Pi(\mathsf{a}, \mathsf{b})} \langle T, C \rangle + \varepsilon \mathsf{KL}(T \mid \mathsf{a} \otimes \mathsf{b}), \tag{2.3.23}$$

where C is the cost matrix and $\mathsf{a} \otimes \mathsf{b}$ simply corresponds to $\mathsf{a}\mathsf{b}^\top$. Recall that the relative entropy (i.e. Kullback-Leibler divergence) between $P, Q \in \mathbb{R}_+^{n \times m}$ is defined by

$$\mathsf{KL}(P \mid Q) = \sum_{i,j=1}^{n,m} (P_{ij} \log\Big(\frac{P_{ij}}{Q_{ij}}\Big) - P_{ij} + Q_{ij}), \tag{2.3.24}$$

with the convention that $0 \log(0) = 0$ and $\mathsf{KL}(P \mid Q) = +\infty$ if there exists a pair (i, j) such that $P_{ij} \neq 0$ and $Q_{ij} = 0$. Note that the function $\mathsf{KL}(\cdot \mid Q)$ is strongly convex, $\mathsf{KL}(P \mid Q) \geq 0$ and $\mathsf{KL}(P \mid Q) = 0$ if and only if $P = Q$. In the discrete setting, characterizing the dual formulation and the explicit form of the solution of (2.3.23) involves less technicalities than the general case. Hence, we briefly present the related results below.

Proposition 2.9. *The solution T of (2.3.23) is unique and has the form*

$$T_{ij} = f_i K_{ij} g_j \qquad \text{with} \qquad K := e^{-\frac{C}{\varepsilon}}, \tag{2.3.25}$$

for two scaling variables $f \in \mathbb{R}_+^n$ and $g \in \mathbb{R}_+^m$.

Proof. Introducing the dual variables $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ for each constraint, the Lagrangian of (2.3.23) reads as

$$\mathcal{L}(T, u, v) = \langle T, C\rangle + \varepsilon \mathsf{KL}(T \mid \mathsf{a} \otimes \mathsf{b}) - \langle u, T\mathbb{1}_m - \mathsf{a}\rangle - \left\langle v, T^\top \mathbb{1}_n - \mathsf{b}\right\rangle. \tag{2.3.26}$$

First order conditions yield

$$\frac{\partial \mathcal{L}(T, u, v)}{\partial T_{ij}} = C_{ij} + \varepsilon \log\Big(\frac{T_{ij}}{\mathsf{a}_i \mathsf{b}_j}\Big) - u_i - v_j, \tag{2.3.27}$$

which lead to the expression $T_{ij} = \mathsf{a}_i \mathsf{b}_j e^{\frac{u_i + v_j - C_{ij}}{\varepsilon}}$. This can be rewritten in the desired form using the scaling f, g defined by $f_i := \mathsf{a}_i e^{u_i/\varepsilon}$ and $g_j := \mathsf{b}_j e^{v_j/\varepsilon}$ respectively (which are nonnegative by definition). □

The optimal scaling are linked with solutions of the dual problem, that in the discrete case reads as follows:

$$\mathsf{OT}_\varepsilon(\alpha, \beta) = \max_{u \in \mathbb{R}^n, v \in \mathbb{R}^m} \langle u, \mathsf{a}\rangle + \langle v, \mathsf{b}\rangle - \varepsilon \sum_{i,j} \exp\Big(\frac{u_i + v_j - C_{ij}}{\varepsilon}\Big) \mathsf{a}_i \mathsf{b}_j. \tag{2.3.28}$$

As one can see in the proof of Prop. 2.9, the optimal dual variables in (2.3.28) are related to the scalings f, g in Prop. 2.9 by $(f, g) = (\mathsf{a} e^{\frac{u}{\varepsilon}}, \mathsf{b} e^{\frac{v}{\varepsilon}})$.

Regularization with Discrete Entropy. Using the relative entropy as a regularizer leads to a framework that is consistent in case of continuous measures and that enjoys a well defined dual formulation. When considering *discrete* measures, the relative entropy between the plan and the product of the marginals is equivalent - as a regularizer - to the discrete entropy of the transport plan, meaning that the two formulations only differ by a constant and share the same solution. In the machine learning community the revival of entropic regularization has started with the discrete entropy and that is why we dedicate a paragraph here to discuss this formulation. Note that this is the formulation that we consider also in

Chapter 3, whose setting is restricted to the space of histograms and does not cover arbitrary measures.

Definition 2.7 (Discrete entropy)**.** *The discrete entropy of a transport plan T is defined as*

$$\mathsf{H}(T) = -\sum_{ij} T_{ij}(\log T_{ij} - 1). \tag{2.3.29}$$

The entropy regularized Optimal Transport problem that makes use of the discrete entropy is defined as

$$\mathsf{OT}_\varepsilon(\alpha, \beta) = \min_{T \in \Pi(\mathsf{a},\mathsf{b})} \langle T, C\rangle - \varepsilon \mathsf{H}(T). \tag{2.3.30}$$

Similarly to (2.3.23), the objective function is ε-strictly convex and hence (2.3.30) admits a unique solution. The characterization of the optimal plan proved in Prop. 2.9 in case of the relative entropy still holds, with the difference that going through the same proof with the discrete entropy as regularizer one obtains that the scaling f and g are of the form $f = e^u$ and $g = e^v$, where u and v are optimal solutions of the dual problem

$$\mathsf{OT}_\varepsilon(\alpha, \beta) = \max_{u \in \mathbb{R}^n, v \in \mathbb{R}^m} \langle u, \mathsf{a}\rangle + \langle v, \mathsf{b}\rangle - \varepsilon \sum_{i,j} \exp\left(\frac{u_i + v_j - C_{ij}}{\varepsilon}\right), \tag{2.3.31}$$

which differs from (2.3.28) because the sum in the right hand side does not depend on a and b.

Regularization with discrete entropy, yet another variant. The entropic regularization considered in the very first machine learning paper on the topic (Cuturi, 2013) was slightly different. Namely, in the problem regularized with discrete entropy just recalled in (2.3.30) the regularization plays a double role: in finding the optimal plan T^*_ε *and* in the associated cost, $\mathsf{OT}_\varepsilon(\alpha, \beta) = \langle T^*_\varepsilon, C\rangle + \varepsilon \mathsf{H}(T^*_\varepsilon)$. In (Cuturi, 2013, Eq. 2), the entropy penalty is used to compute the optimal plan but then disregarded in the cost: with the same notation as before,

$$\tilde{\mathsf{OT}}_\varepsilon(\alpha, \beta) = \langle T^*_\varepsilon, C\rangle \qquad \text{with} \qquad T^*_\varepsilon = \operatorname*{argmin}_{T \in \Pi(\mathsf{a},\mathsf{b})} \langle T, C\rangle + \varepsilon \mathsf{H}(T). \tag{2.3.32}$$

In Chapter 3 we will use both (2.3.30) and (2.3.32) and compare them in different tasks; we leave further details on (2.3.32) to Chapter 3 to avoid repetitions.

2.3.5 Gradients

From now on, we go back to the general setting and the formulation of OT_ε as in (2.3.1). Since they will be important concepts in the rest of the thesis, we present some results on differentiability of OT_ε. For a more in-depth analysis on this topic we refer the reader to Feydy et al. (2019) (in particular Proposition 2).

To introduce the notion of directional derivatives of OT_ε we need to define the set of feasible directions.

Definition 2.8. *Let $(\alpha, \beta) \in \mathcal{P}(\mathcal{X})^2$. The set of feasible directions of OT_ε at (α, β), denoted as $\mathcal{F}_{\mathcal{P}(\mathcal{X})^2}\big((\alpha, \beta)\big)$ is defined as*

$$\mathcal{F}_{\mathcal{P}(\mathcal{X})^2}\big((\alpha, \beta)\big) := \{(\mu, \nu) \in \mathcal{P}(\mathcal{X})^2 \text{ s.t. } \mu = \alpha' - \alpha,\ \nu = \beta' - \beta \text{ for some } \alpha', \beta' \in \mathcal{P}(\mathcal{X})\}.$$

Proposition 2.10. *Let $x_o \in \mathcal{X}$, $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ and $(u, v) \in \mathcal{C}(\mathcal{X})^2$ be the pair of corresponding Sinkhorn potentials with $u(x_o) = 0$. The function OT_ε is directionally differentiable and the directional derivative of OT_ε in (α, β) along a feasible direction $(\mu, \nu) \in \mathcal{F}_{\mathcal{P}(\mathcal{X})^2}\big((\alpha, \beta)\big)$ is*

$$\mathsf{OT}'_\varepsilon(\alpha, \beta; \mu, \nu) = \int u(x)\, d\mu(x) + \int v(y)\, d\nu(y) = \langle (u, v), (\mu, \nu) \rangle\,, \tag{2.3.33}$$

where $\langle w, \rho \rangle = \int w(x)\, d\rho(x)$ denotes the canonical pairing between the spaces $\mathcal{C}(\mathcal{X})$ and $\mathcal{M}(\mathcal{X})$. Let $\nabla\mathsf{OT}_\varepsilon \colon \mathcal{P}(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})^2$ be the operator that maps every pair of probability distributions $(\alpha, \beta) \in \mathcal{P}(\mathcal{X})^2$ to the corresponding pair of Sinkhorn potentials $(u, v) \in \mathcal{C}(\mathcal{X})^2$ with $u(x_o) = 0$. Then (2.3.33) can be written as

$$\mathsf{OT}'_\varepsilon(\alpha, \beta; \mu, \nu) = \langle \nabla\mathsf{OT}_\varepsilon(\alpha, \beta), (\mu, \nu) \rangle\,. \tag{2.3.34}$$

Proof. Let $(\delta\alpha, \delta\beta) \in \mathcal{F}_{\mathcal{P}(\mathcal{X})^2}\big((\alpha, \beta)\big)$. Denote by $\alpha_t := \alpha + t\delta\alpha$ and $\beta_t := \beta + t\delta\beta$. We define the variation ration as

$$\Delta_t = \frac{\mathsf{OT}_\varepsilon(\alpha_t, \beta_t) - \mathsf{OT}_\varepsilon(\alpha, \beta)}{t}. \tag{2.3.35}$$

The rest of the proof consists in providing an upper and lower bound for the variation ratio.

Lower bound. Let (u, v) a pair of optimal potentials for $\mathsf{OT}_\varepsilon(\alpha, \beta)$. The pair (u, v) is then suboptimal for $\mathsf{OT}_\varepsilon(\alpha_t, \beta_t)$ yielding

$$\mathsf{OT}_\varepsilon(\alpha_t, \beta_t) \geq \langle \alpha_t, u \rangle + \langle \beta_t, v \rangle - \varepsilon \left\langle \alpha_t \otimes \beta_t, \exp(\frac{1}{\varepsilon}(u \oplus v - \mathsf{c})) - 1 \right\rangle . \tag{2.3.36}$$

Since (u, v) are optimal potentials for $\mathsf{OT}_\varepsilon(\alpha, \beta)$ we also have

$$\mathsf{OT}_\varepsilon(\alpha, \beta) = \langle u, \alpha \rangle + \langle v, \beta \rangle - \varepsilon \left\langle \alpha \otimes \beta, \exp(\frac{1}{\varepsilon}(u \oplus v - \mathsf{c})) - 1 \right\rangle . \tag{2.3.37}$$

By definition of Δ_t, we have that

$$\Delta_t \geq \langle \delta\alpha, u \rangle + \langle \delta\beta, v \rangle - \varepsilon \left\langle \delta\alpha \otimes \beta + \alpha \otimes \delta\beta, \exp(\frac{1}{\varepsilon}(u \oplus v - \mathsf{c})) - 1 \right\rangle \tag{2.3.38}$$

$$\geq \langle \delta\alpha, f - \varepsilon \rangle + \langle \delta\beta, g - \varepsilon \rangle , \tag{2.3.39}$$

where the last inequality follows from the fact that $\int e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}} d\beta(y) d(\frac{\alpha_t(x)-\alpha(x)}{t}) = 0$.

Upper bound. Similarly, we can derive an upper bound for the same quantity using a pair (u_t, v_t) of optimal potentials of $\mathsf{OT}_\varepsilon(\alpha_t, \beta_t)$. Proceeding similarly as above, we obtain that

$$\Delta_t \leq \langle \delta\alpha, u_t - \varepsilon \rangle + \langle \delta\beta, v_t - \varepsilon \rangle . \tag{2.3.40}$$

Now, note that as $t \to 0$, $\alpha_t \rightharpoonup \alpha$ and $\beta_t \rightharpoonup \beta$. Therefore $u_t \to u$ and $v_t \to v$ in $\| \cdot \|_\infty$. Combining upper and lower bounds we have

$$\Delta_t \xrightarrow{t \to 0} \langle \delta\alpha, u - \varepsilon \rangle + \langle \delta\beta, v - \varepsilon \rangle = \langle \delta\alpha, u \rangle + \langle \delta\beta, v \rangle , \tag{2.3.41}$$

where the last equality follows from the fact that $\delta\alpha$ and $\delta\beta$ both have an overall mass that sums up to 0. □

Remark 2.2. In Prop. 2.10, the requirement $u(x_o) = 0$ is only a convention to remove ambiguities. Indeed, for every $K \in \mathbb{R}$, replacing the Sinkhorn potentials (u, v) with $(u + K, v - K)$ does not affect (2.3.33).

2.3.6 Convergence to unregularized Optimal Transport

In this thesis, Entropic Optimal Transport is of interest on its own, with its several elegant properties, both theoretical and computational. However, since it comes as *regularization* of standard Optimal Transport, a natural question concerns the relation between the two, in terms of convergence when the regularization parameter goes to zero. In this section we prove the result in the discrete case, since it is easier. After that we will state the result in the general case and refer to relevant literature for the proof.

Proposition 2.11 (Convergence as $\varepsilon \to 0$ in the discrete case)**.** *Let T_ε be the unique solution of (2.3.23). Then T_ε converges to the optimal solution with maximal entropy within the set of all optimal solutions of the unregularized problem, namely*

$$T_\varepsilon \xrightarrow{\varepsilon \to 0} \operatorname*{argmin}_T \{\mathsf{KL}(T \mid \mathsf{a} \otimes \mathsf{b}) : \ T \in \Pi(\mathsf{a}, \mathsf{b}) \text{ and } \langle T, C \rangle = \mathsf{OT}(\alpha, \beta).\} \tag{2.3.42}$$

In particular

$$\mathsf{OT}_\varepsilon(\alpha, \beta) \xrightarrow{\varepsilon \to 0} \mathsf{OT}(\alpha, \beta). \tag{2.3.43}$$

Proof. Consider a sequence $(\varepsilon_k)_k$ such that $\varepsilon_k \to 0$ as $k \to \infty$ and $\varepsilon_k > 0$. Let T_k be the solution of (2.3.23) with $\varepsilon = \varepsilon_k$. Since $\Pi(\mathsf{a}, \mathsf{b})$ is bounded, $\{T_k\}_k$ is a bounded sequence and hence we can extract a converging subsequence that we do not relabel, $T_k \to T^*$ for some T^*. Since $\Pi(\mathsf{a}, \mathsf{b})$ is closed, $T^* \in \Pi(\mathsf{a}, \mathsf{b})$. Now, consider any T which is a solution of the unregularized transport problem (2.2.12), i.e. $\langle T, C \rangle = \mathsf{OT}(\alpha, \beta)$. By optimality of T_k and T for their respective problems, we have

$$0 \leq \langle T_k, C \rangle - \langle T, C \rangle \leq \varepsilon_k (\mathsf{KL}(T \mid \mathsf{a} \otimes \mathsf{b}) - \mathsf{KL}(T_k \mid \mathsf{a} \otimes \mathsf{b})). \tag{2.3.44}$$

Since $\mathsf{KL}(\cdot \mid \mathsf{a} \otimes \mathsf{b})$ is continuous, taking the limit as $k \to \infty$ we obtain that $\langle T^*, C \rangle = \langle T, C \rangle$, so T^* is a feasible point in the right hand side of (2.3.42). Moreover, dividing by ε_k in (2.3.44) and taking the limit as $k \to \infty$, one obtains that $\mathsf{KL}(T^* | \mathsf{a} \otimes \mathsf{b}) \leq \mathsf{KL}(T | \mathsf{a} \otimes \mathsf{b})$ and hence T^* is a solution of (2.3.42). Now, since $\mathsf{KL}(\cdot \mid \mathsf{a} \otimes \mathsf{b})$ is strictly convex, the solution of (2.3.42) is unique. Hence the whole sequence converges. □

In the general setting, the convergence result is the following:

Theorem 2.12. *Let π_ε be the minimizer of the regularized problem in (2.3.1). Then,*

$$\mathsf{OT}_\varepsilon(\alpha,\beta) \xrightarrow{\varepsilon\to 0} \mathsf{OT}(\alpha,\beta) \qquad \text{and} \qquad \pi_\varepsilon \rightharpoonup \pi, \tag{2.3.45}$$

where π is a minimizer of the Optimal Transport problem (2.2.2).

The proof can be found in (Carlier et al., 2017, Thm. 2.7).

2.3.7 Short discussion of statistical results

We conclude the section on Entropic Optimal Transport recalling a result on the approximation power from finite samples. One of the shortcomings of standard Optimal Transport distances introduced at the beginning of the chapter is the curse of dimensionality that appears when estimating a probability measure using samples. Introducing the entropic penalty leads to advantages on this matter. Recent results showed that when considering entropic regularized Optimal Transport, the dependence on the dimension appears in the constant and not in the rate (Genevay et al., 2018a; Mena and Niles-Weed, 2019). The following holds:

Proposition 2.13. *Consider $\mathcal{X} = \mathbb{R}^d$. Let α and β be two σ^2-subgaussian measures in $\mathcal{P}(\mathcal{X})$. Let α_n and β_n be empirical measures from n iid samples of α and β. Then,*

$$\mathbb{E}\mathsf{OT}_\varepsilon(\alpha_n,\beta_n) - \mathsf{OT}_\varepsilon(\alpha,\beta) \le C_d\varepsilon\Big(1+\frac{\sigma^{\lceil 5d/2\rceil+6}}{\varepsilon^{\lceil 5d/4\rceil+3}}\Big)\frac{1}{\sqrt{n}}, \tag{2.3.46}$$

where C_d is a constant depending on the dimension d only.

This result is shown in Mena and Niles-Weed (2019) and it is specific for the cost function $\mathsf{c} = \|\cdot\|^2$. A less tight result, which however holds for more general classes of cost function, is contained in Genevay et al. (2018a). The proofs of both results are quite technical and have been the material of whole papers. Hence here we just refer to such works and we do not provide the whole proofs.

2.4 Algorithms

After presenting the Entropic Optimal Transport problem and some important theoretical results, we briefly discuss how to compute Entropic OT in practice. Algorithmic aspects are not the focus on the thesis, therefore here we keep this section brief and we just recall the basics. However, algorithmic developments are of fundamental importance and there is a lot of interesting research addressing those aspects. We refer to the book (Peyré and Cuturi, 2019, Sec. 4.2-4.6) for a thorough review. In the following we focus of the discrete setting.

For completeness, in Appendix A we also discuss the continuous counterpart.

Recall the first order conditions in (2.3.6) of the dual problem (2.3.3): for $\alpha = \sum_{i=1}^n \mathsf{a}_i \delta_{x_i}$, and $\beta = \sum_{j=1}^n \mathsf{b}_j \delta_{y_j}$, such conditions read

$$\mathsf{u}_i = -\varepsilon \log\Big(\sum_{j=1}^m e^{\frac{\mathsf{v}_j - \mathsf{c}(x_i,y_j)}{\varepsilon}} \mathsf{b}_j\Big), \qquad \mathsf{v}_j = -\varepsilon \log\Big(\sum_{i=1}^n e^{\frac{\mathsf{u}_i - \mathsf{c}(x_i,y_j)}{\varepsilon}} \mathsf{a}_i\Big), \tag{2.4.1}$$

and taking the exponential scalings of u_i and v_j, i.e. $\mathsf{f}_i = e^{\frac{\mathsf{u}_i}{\varepsilon}}$ and $\mathsf{g}_j = e^{\frac{\mathsf{v}_j}{\varepsilon}}$, we obtain the equivalent conditions

$$\mathsf{f}_i = \frac{1}{\sum_{j=1}^m \mathsf{g}_j e^{-\mathsf{c}(x_i,y_j)} \mathsf{b}_j}, \qquad \mathsf{g}_j = \frac{1}{\sum_{i=1}^n \mathsf{f}_i e^{-\mathsf{c}(x_i,y_j)} \mathsf{a}_i}. \tag{2.4.2}$$

Set K the matrix given by $\mathsf{K}_{ij} = e^{\frac{-\mathsf{c}(x_i,y_j)}{\varepsilon}}$. Then the equations above can be rewritten as

$$\mathsf{f} = \frac{1}{\mathsf{K}(\mathsf{g} \odot \mathsf{b})} = \frac{\mathsf{a}}{\mathsf{M}\mathsf{g}}, \qquad \mathsf{g} = \frac{1}{\mathsf{K}^\top(\mathsf{f} \odot \mathsf{a})} = \frac{\mathsf{b}}{\mathsf{M}^\top \mathsf{f}}, \tag{2.4.3}$$

where $\odot$ denotes the entrywise vector multiplication, the division is also to be considered entrywise, and $\mathsf{M} = \mathrm{diag}(\mathsf{a})\mathsf{K}\mathrm{diag}(\mathsf{b})$. The dual problem is concave in each variable, therefore a natural way to solve it is to iteratively optimize over each variable. The algorithm corresponding to these alternating maximizations is usually called Sinkhorn algorithm. It was pioneered in Sinkhorn (1964); Sinkhorn and Knopp (1967) and further developed in infinite dimensional settings in Nussbaum (1993). We recall Sinkhorn algorithm in Alg. 2.1. The proof of convergence of Alg. 2.1 is similar to the proof of existence of the dual potentials,

Algorithm 2.1 Sinkhorn-Knopp algorithm (finite dimensional case)

Let $\mathsf{K} \in \mathbb{R}_{++}^{n\times m}$, $\mathsf{a} \in \mathbb{R}_+^n$, with $\mathsf{a}^\top 1_n = 1$, and $\mathsf{b} \in \mathbb{R}_+^m$, with $\mathsf{b}^\top 1_m = 1$. Set $\mathsf{M} = \mathrm{diag}(\mathsf{a})\mathsf{K}\mathrm{diag}(\mathsf{b})$. Let $\mathsf{f}^{(0)} \in \mathbb{R}_{++}^n$ and define

$$\text{for } \ell = 0, 1, \dots \quad \left| \begin{array}{l} \mathsf{g}^{(\ell+1)} = \dfrac{\mathsf{b}}{\mathsf{M}^\top \mathsf{f}^{(\ell)}} \\ \mathsf{f}^{(\ell+1)} = \dfrac{\mathsf{a}}{\mathsf{M}\mathsf{g}^{(\ell+1)}}. \end{array} \right.$$

that we have deferred to Appendix A. Here we provide it for the discrete setting. As explained in Franklin and Lorenz (1989), a useful tool to show the global convergence of Alg. 2.1 is the Hilbert projective metric, introduced in Birkhoff (1957). Denoting by

$\mathbb{R}^n_+ := \{\mathsf{f} \in \mathbb{R}^n : \mathsf{f}_i > 0 \ \forall i\}$, the Hilbert projective metric on $\mathbb{R}^n_+$ is defined as

$$\text{for all } (\mathsf{f}, \mathsf{f}') \in (\mathbb{R}^n_+)^2 \qquad d_H(\mathsf{f}, \mathsf{f}') := \log \max \frac{\mathsf{f}_i \mathsf{f}'_j}{\mathsf{f}_j \mathsf{f}'_i}. \tag{2.4.4}$$

The Hilbert metric is a distance on $\mathbb{R}^n_+ / \sim$, where $\mathsf{f} \sim \mathsf{f}'$ if there exists $t > 0$ such that $\mathsf{f} = t\mathsf{f}'$, and hence $d_H(\mathsf{f}, \mathsf{f}') = 0$ if and only if $\mathsf{f} \sim \mathsf{f}'$. Also note that taking a logarithmic change of variables, the Hilbert metric is isometric to the variation seminorm which is a norm between vectors defined up to additive constant and that is closely related to the ℓ^∞ norm. Namely,

$$d_H(\mathsf{f}, \mathsf{f}') = \|\log(\mathsf{f}) - \log(\mathsf{f}')\|_{var}, \qquad \|\mathsf{u}\|_{var} := \max_i \mathsf{u}_i - (\min_i \mathsf{u}_i). \tag{2.4.5}$$

Note that $\|\mathsf{u}\|_{var} \leq 2\|\mathsf{u}\|_\infty$ and if we impose that $\mathsf{u}_i = 0$ for some fixed i, then $\|\mathsf{u}\|_\infty \leq \|\mathsf{u}\|_{var}$. The global convergence result of Sinkhorn algorithm relies on the following fundamental theorem (Birkhoff, 1957) about a contraction result under the Hilbert metric.

Theorem 2.14. *Let $\mathsf{K} \in \mathbb{R}^{n \times m}_+$. Then for any $(\mathsf{g}, \mathsf{g}') \in (\mathbb{R}^m_+)^2$,*

$$d_H(\mathsf{K}\mathsf{g}, \mathsf{K}\mathsf{g}') \leq \lambda(\mathsf{K}) d_H(\mathsf{g}, \mathsf{g}'), \qquad \text{where} \begin{cases} \lambda(\mathsf{K}) = \frac{\sqrt{\eta(\mathsf{K})}-1}{\sqrt{\eta(\mathsf{K})}+1} < 1 \\ \eta(\mathsf{K}) = \max_{i,j,k,l} \frac{\mathsf{K}_{ik}\mathsf{K}_{j\ell}}{\mathsf{K}_{jk}\mathsf{K}_{j\ell}}. \end{cases} \tag{2.4.6}$$

Using the theorem above, we can show the global convergence of Sinkhorn algorithm.

Proposition 2.15. *Let $\alpha = \sum_{i=1}^n \mathsf{a}_i \delta_{x_i}$, and $\beta = \sum_{j=1}^n \mathsf{b}_j \delta_{y_j}$ be two discrete measures and set K the matrix given by $\mathsf{K}_{ij} = e^{\frac{-\mathsf{c}(x_i, y_j)}{\varepsilon}}$. Then, the Sinkhorn iterations*

$$\mathsf{f}^{(\ell+1)} = \frac{1}{\mathsf{K}(\mathsf{g}^{(\ell)} \odot \mathsf{b})}, \qquad \mathsf{g}^{(\ell+1)} = \frac{1}{\mathsf{K}^\top(\mathsf{f}^{(\ell+1)} \odot \mathsf{a})} \tag{2.4.7}$$

converge to $(\mathsf{f}^*, \mathsf{g}^*)$ the exponential scaling of a solution of the dual problem (2.3.3) and

$$d_H(\mathsf{f}^{(\ell)}, \mathsf{f}^*) = O(\lambda(\mathsf{K})^\ell) \qquad d_H(\mathsf{g}^{(\ell)}, \mathsf{g}^*) = O(\lambda(\mathsf{K})^\ell). \tag{2.4.8}$$

Also, setting $T^{(\ell)} = \mathrm{diag}(\mathsf{f}^{(\ell)} \odot \mathsf{a})\, \mathsf{K}\, \mathrm{diag}(\mathsf{g}^{(\ell)} \odot \mathsf{b}) = \mathrm{diag}(\mathsf{f}^{(\ell)})\, \mathsf{M}\, \mathrm{diag}(\mathsf{g}^{(\ell)})$ *, we have*

$$d_H(\mathsf{f}^{(\ell)}, \mathsf{f}*) \leq \frac{d_H(T^{(\ell)}\mathbb{1}_m, \mathsf{a})}{1-\lambda(\mathsf{K})} \tag{2.4.9}$$

$$d_H(\mathsf{g}^{(\ell)}, \mathsf{g}^*) \leq \frac{d_H(T^{(\ell)\top}\mathbb{1}_n, \mathsf{b})}{1-\lambda(\mathsf{K})} \tag{2.4.10}$$

and finally

$$\|\log(T^{(\ell)}) - \log(T^*)\|_\infty \leq d_H(\mathsf{f}^{(\ell)}, \mathsf{f}^*) + d_H(\mathsf{g}^{(\ell)}, \mathsf{g}^*), \tag{2.4.11}$$

where T^* *is the solution of* (2.3.23) *.*

Proof. Note that by the defintion of the Hilbert metric, it holds that for any $(\mathsf{g}, \mathsf{g}') \in (\mathbb{R}_+^m)^2$,

$$d_H(\mathsf{g}, \mathsf{g}') = d_H(\mathsf{g}/\mathsf{g}', \mathbb{1}_m) = d_H(\mathbb{1}_m/\mathsf{g}, \mathbb{1}_m/\mathsf{g}'). \tag{2.4.12}$$

Hence,

$$d_H(\mathsf{f}^{(\ell+1)}, \mathsf{f}^*) = d_H(\frac{\mathsf{a}}{\mathsf{M}\mathsf{g}^{(\ell)}}, \frac{\mathsf{a}}{\mathsf{M}\mathsf{g}^*}) = d_H(\mathsf{M}\mathsf{g}^{(\ell)}, \mathsf{M}\mathsf{g}^*) \tag{2.4.13}$$

$$\leq \lambda(\mathsf{M}) d_H(\mathsf{g}^{(\ell)}, \mathsf{g}^*) = \lambda(\mathsf{K}) d_H(\mathsf{g}^{(\ell)}, \mathsf{g}^*) \tag{2.4.14}$$

where the equality $\lambda(\mathsf{K}) = \lambda(\mathsf{M})$ is straighforward by definition of M. This shows (2.4.8). Now, using the triangle inequality of the Hilbert metric, we obtain

$$\begin{aligned} d_H(\mathsf{f}^{(\ell)}, \mathsf{f}^*) &\leq d_H(\mathsf{f}^{(\ell+1)}, \mathsf{f}^{(\ell)}) + d_H(\mathsf{f}^{(\ell+1)}, \mathsf{f}^*) \\ &\leq d_H(\frac{\mathsf{a}}{\mathsf{M}\mathsf{g}^{(\ell)}}, \mathsf{f}^{(\ell)}) + \lambda(\mathsf{K}) d_H(\mathsf{f}^{(\ell)}, \mathsf{f}^*) \\ &\leq d_H(\mathsf{a}, \mathsf{f}^{(\ell)} \odot \mathsf{M}\mathsf{g}^{(\ell)}) + \lambda(\mathsf{K}) d_H(\mathsf{f}^{(\ell)}, \mathsf{f}^*). \end{aligned}$$

This gives the first part of (2.4.9) (the second is analogous), since $\mathsf{M}\mathsf{g}^{(\ell)} = T^{(\ell)}\mathbb{1}_m$. Finally the proof of (2.4.11) follows from (Franklin and Lorenz, 1989, Lemma 3). □

Remark 2.3. ***Computational Complexity****. The main computational bottleneck of Sinkhorn iterations is the vector-matrix multiplication against* K *and* $\mathsf{K}^\top$*: for two measures with* n *points, the complexity of this operation is* $O(n^2)$ *if implemented naively. In several cases, significant speed-ups are possible and we refer to (Peyré and Cuturi, 2019, Sec. 4.3) for the*

details on this point. In Altschuler et al. (2017) it is shown that Sinkhorn algorithm allows to find an ε-approximation of OT distance with $\tilde{O}(n^2/\varepsilon^3)$ arithmetic operations.

2.5 Sinkhorn divergence

The entropic bias. When the cost function c corresponds to a distance d or to a power d^p on the domain $\mathcal{X}$, the related related OT or $(\mathsf{OT})^{1/p}$ are -strictly speaking- distances (known as Wasserstein distances). When introducing the entropy-regularization we interfere with the metric properties: in particular, OT_ε for $\varepsilon > 0$ does not satisfy the property $0 = \mathsf{OT}_\varepsilon(\alpha, \alpha) \leq \mathsf{OT}_\varepsilon(\alpha, \beta)$, which is a desirable feature when we want to use OT_ε to compare distributions and to induce a notion of metric. While the impact is not significant when ε is small, for big ε it leads to a 'shrinking effect'; to observe this, note that for $\varepsilon \to \infty$, $\mathsf{OT}_\varepsilon(\alpha, \beta) \to \int \mathsf{c}(x, y) d\alpha(x)\, d\beta(y)$. The latter quantity is minimized when α is a Dirac delta centered at the mean (resp. median) value of β when c is $\mathsf{c}(x, y) = \|x - y\|^2$ (resp. $\mathsf{c}(x, y) = \|x - y\|$). Regardless of its impact in practise, when $\varepsilon > 0$ the entropic regularization introduces a *bias* in the optimal transport problem, since in general $\mathsf{OT}_\varepsilon(\alpha, \alpha) \neq 0$. Recently, a few works (Ramdas et al., 2017; Genevay et al., 2018b; Feydy et al., 2019) have proposed to consider a *debiased* version of the entropic problem that satisfies the properties of a divergence. The definition is below:

Definition 2.9 (Sinkhorn divergence)**.** *Sinkhorn divergence is a debiasing of OT_ε and it is defined as follows: for $\alpha, \beta \in \mathcal{P}(\mathcal{X})$*

$$\mathsf{S}_\varepsilon : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}, \qquad \mathsf{S}_\varepsilon(\alpha, \beta) = \mathsf{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\mathsf{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathsf{OT}_\varepsilon(\beta, \beta). \tag{2.5.1}$$

Nonnegativity, convexity and weak convergence. Sinkhorn divergence recovers properties which are desirable in a loss function that is often used for density fitting or matching. The theorem below summarizes those properties in a clean statement:

Theorem 2.16. *Let $\mathcal{X}$ be a compact metric space and $\mathsf{c} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Lipschitz cost function that induces for $\varepsilon > 0$ a positive universal kernel $k_\varepsilon : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as $k_\varepsilon(x, y) = e^{-\mathsf{c}(x,y)/\varepsilon}$. Then, the Sinkhorn divergence S_ε is positive definite, smooth, and convex in each of its inputs. Moreover, it metrizes the convergence in law, i.e. for any*

measures $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, it holds

$$0 = \mathsf{S}_\varepsilon(\alpha, \alpha) \leq \mathsf{S}_\varepsilon(\alpha, \beta) \tag{2.5.2}$$

$$\alpha = \beta \iff \mathsf{S}_\varepsilon(\alpha, \beta) = 0 \tag{2.5.3}$$

$$\alpha_n \rightharpoonup \alpha \iff \mathsf{S}_\varepsilon(\alpha_n, \alpha) \to 0. \tag{2.5.4}$$

Proof. The proof of this result is the subject of Feydy et al. (2019). To avoid several pages of repetitions, we refer to the paper directly. □

Limits. Sinkhorn divergence recovers the unregularized Optimal Transport as ε goes to zero, similarly to OT_ε. However, when $\varepsilon \to \infty$, the limits of OT_ε and S_ε are different. Set the following notation: for a measure $\mu \in \mathcal{P}(\mathcal{X})$ and a real-valued continuous function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the convolution $k \star \alpha$ is defined as

$$k \star \alpha \in \mathcal{C}(\mathcal{X}) \qquad x \longmapsto \int_{\mathcal{X}} k(x, y)\, d\alpha(y). \tag{2.5.5}$$

On one side we have

$$\mathsf{OT}_\varepsilon(\alpha, \beta) \xrightarrow{\varepsilon \to \infty} \langle \mathsf{c}, \alpha \otimes \beta \rangle = \langle \alpha, \mathsf{c} \star \beta \rangle \tag{2.5.6}$$

which does not have any 'norm' structure. On the other hand, Sinkhorn divergence leads to

$$\mathsf{S}_\varepsilon(\alpha, \beta) \xrightarrow{\varepsilon \to \infty} = \langle \alpha, \mathsf{c} \star \beta \rangle - \frac{1}{2} \langle \alpha, \mathsf{c} \star \alpha \rangle - \frac{1}{2} \langle \beta, \mathsf{c} \star \beta \rangle = \frac{1}{2} \langle \alpha - \beta, -\mathsf{c} \star (\alpha - \beta) \rangle . \tag{2.5.7}$$

Hence, for $k = -\mathsf{c}/2$ this is equivalent to

$$\mathsf{S}_\varepsilon(\alpha, \beta) \xrightarrow{\varepsilon \to \infty} = \mathbb{E}_{\alpha \otimes \alpha}[k(x, x')] + \mathbb{E}_{\beta \otimes \beta}[k(y, y')] - 2\mathbb{E}_{\alpha \otimes \beta}[k(x, y)], \tag{2.5.8}$$

which reminds of Maximum Mean Discrepancy MMD (recalled in Def. A.14) and exactly corresponds to MMD for some choices of c (and consequently of the kernel k).

Debiased potentials and gradients. Let u, v be optimal dual potentials. In Prop. 2.5 we

showed that $\mathsf{OT}_\varepsilon(\alpha,\beta) = \langle \alpha, u\rangle + \langle \beta, v\rangle$. Similarly,

$$\mathsf{S}_\varepsilon(\alpha,\beta) = \langle \alpha, u-p\rangle + \langle \beta, v-q\rangle\,, \tag{2.5.9}$$

where (u,v) is a pair of optimal potentials of $\mathsf{OT}_\varepsilon(\alpha,\beta)$ and $p = \mathsf{P}_\alpha(p)$, $q = \mathsf{P}_\beta(q)$ are optimal potentials for $\mathsf{OT}_\varepsilon(\alpha,\alpha)$ and $\mathsf{OT}_\varepsilon(\beta,\beta)$ respectively.

Proposition 2.17. *Let $\beta \in \mathcal{P}(\mathcal{X})$ and let $\nabla_1\mathsf{OT}_\varepsilon$ be the first component of the gradient operator defined in Prop. 2.10. Then the Sinkhorn divergence function $S_\varepsilon(\cdot,\beta)\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ in (2.5.1) is directionally differentiable and, for every $\alpha \in \mathcal{P}(\mathcal{X})$ and every $\mu \in \mathcal{F}_{\mathcal{P}(\mathcal{X})}(\alpha)$ it holds*

$$[S_\varepsilon(\cdot,\beta)]'(\alpha;\mu) = \langle \nabla S_\varepsilon(\cdot,\beta), \mu\rangle\,,$$

where

$$\nabla S_\varepsilon(\cdot,\beta)\colon \mathcal{P}(\mathcal{X}) \to \mathcal{C}(\mathcal{X}) \qquad \alpha \mapsto \nabla_1\mathsf{OT}_\varepsilon(\alpha,\beta) - \frac{1}{2}\nabla_1\mathsf{OT}_\varepsilon(\alpha,\alpha) = u - p, \tag{2.5.10}$$

with $u = \mathsf{P}_{\beta\alpha}(u)$ (where we recall that $\mathsf{P}_{\beta\alpha}$ is the composition of P_β and P_α) and $p = \mathsf{P}_\alpha(p)$ the Sinkhorn potentials of $\mathsf{OT}_\varepsilon(\alpha,\beta)$ and $\mathsf{OT}_\varepsilon(\alpha,\alpha)$ respectively.

Proof. This follows immediately from the definition of Sinkhorn divergence and Prop. 2.10. □

As for algorithms and statistical properties, it is straightforward to adapt the results recalled in the previous section on entropic optimal transport OT_ε to the Sinkhorn divergence. In the next chapters, we will explore Entropic OT and Sinkhorn divergence as discrepancy between probability measures in three machine learning problems.

Chapter 3

Learning with Sinkhorn divergence

In many settings, data can be represented as discrete probability distributions over a finite set of atoms (namely, histograms). Supervised learning problems with such data as outputs can be cast as learning problems that aim to predict a probability measure. When predicting a probability distribution, a suitable notion of closeness between measures is crucial. Optimal Transport provides a powerful tool to compare probability distributions and is a natural choice for a loss function when learning distributions.

In this chapter, we aim to use entropic optimal transport in supervised learning settings. The goal is to propose an estimator with provable consistency and explicit learning rates. As mentioned in the background chapter, slightly different variants of regularized OT are sound and available when dealing with discrete measures. In particular, in most situations the original Sinkhorn approximation (that we call sharp Sinkhorn) of the Wasserstein distance is replaced by a regularized version (that we call vanilla Sinkhorn) that is less accurate but easier to differentiate (Frogner et al., 2015; Rolet et al., 2016; Cuturi and Doucet, 2014; Benamou et al., 2015). A natural question is whether the easier tractability of this regularization is paid in terms of accuracy. Indeed, it can be shown that the sharp Sinkhorn approach provides a *tighter* approximation to the Wasserstein distance (Cominetti and Martín, 1994). We consider both vanilla and sharp Sinkhorn as viable loss functions for supervised learning with histograms as outputs and we study their differential properties, that play a key role in our analysis. We prove that sharp Sinkhorn distance enjoys the same smoothness properties as vanilla Sinkhorn and we explicitly provide an efficient algorithm to compute its gradient. We show that this result benefits both theory and applications: on one hand, we leverage high order smoothness to design an estimator for learning with Wasserstein approximations that has proved statistical guarantees. On the other hand, the gradient formula allows us

to efficiently solve learning and optimization problems in practice. Promising preliminary experiments complement our analysis. This chapter will mainly discuss the work in Luise et al. (2018).[1]

Contributions. The principal contributions of this chapter are threefold; (i) we show that both sharp and vanilla Sinkhorn distances on the simplex are smooth; (ii) we derive an explicit formula to compute the gradient of the sharp Sinkhorn efficiently. As intended, this latter result allows us to adopt this distance in applications such as approximating Wasserstein barycenters (Cuturi and Doucet, 2014), which to the best of our knowledge has not been investigated in this setting so far; (iii) we provide a novel sound approach to the challenging problem of *learning with Sinkhorn loss*, recently considered in Frogner et al. (2015). In particular, we leverage the smoothness of the Sinkhorn distance to study the generalization properties of a structured prediction estimator adapted from Ciliberto et al. (2016) to this setting, proving consistency and finite sample bounds. We provide preliminary empirical evidence of the effectiveness of the proposed approach, testing our method on a image reconstruction task.

The rest of the chapter is structured as follows: in Sec. 3.1 we briefly recall the definition of sharp and vanilla Sinkhorn and we compare their behaviors as the regularization parameter goes to zero; in Sec. 3.2 we present the results on differential properties of Sinkhorn approximation and propose a method to compute the gradient of the sharp Sinkhorn based on the implicit function theorem; Sec. 3.3 is the core section of the chapter and addresses the supervised learning framework with Sinkhorn loss. Finally, Sec. 3.4 presents experiments on the proposed estimator and Sec. 3.5 provides conluding remarks.

3.1 A comparison of variants of entropic regularization

In Chapter 2, we recalled two variants of Entropic Optimal Transport that are available when considering discrete measures. In this Chapter we will use both of them in the supervised learning setting. Here we study a comparison of the two in terms of approximation power of the Wasserstein distance. While Wassestein distance and entropic versions were presented in (2.2.12), (2.3.30), (2.3.32) in the background respectively, here we recall the definitions for convenience. Let $\mathcal{X}$ be a metric space. In the following we focus on measures with discrete support in $\mathcal{X}$. In particular, we consider distributions $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ that can be

[1] Advances in Neural Information Processing Systems (NeurIPS), Dec 2018, Montréal, Canada.

written as linear combinations $\alpha = \sum_{i=1}^n \mathsf{a}_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m \mathsf{b}_j \delta_{y_j}$ of Dirac's deltas centered on a finite number n and m of points $(x_i)_{i=1}^n$ and $(y_j)_{j=1}^m$ in $\mathcal{X}$, with the vector weights $\mathsf{a} = (\mathsf{a}_1, \dots, \mathsf{a}_n)^\top \in \Delta_n$ and $\mathsf{b} = (\mathsf{b}_1, \dots, \mathsf{b}_m)^\top \in \Delta_m$ belonging to the n and m-dimensional simplex respectively. Wasserstein distance is defined as

$$\mathsf{OT}(\alpha, \beta) = \mathsf{W}(\alpha, \beta) = \min_{T \in \Pi(\mathsf{a}, \mathsf{b})} \langle T, C \rangle, \tag{3.1.1}$$

where $\Pi(\mathsf{a}, \mathsf{b}) := \{T \in \mathbb{R}_+^{n \times m} : T \mathbb{1}_n = \mathsf{a}, \;\; T^\top \mathbb{1}_m = \mathsf{b}\}$. Now, let the discrete entropy of a matrix $T \in \mathbb{R}_+^{n \times m}$ be defined as

$$\mathsf{H}(T) \;=\; -\sum_{i,j=1}^{n,m} T_{ij}(\log T_{ij} - 1). \tag{3.1.2}$$

The two entropy-regularized optimal transport problems are defined as

$$\mathsf{OT}_\varepsilon(\alpha, \beta) \;=\; \min_{T \in \Pi(\mathsf{a}, \mathsf{b})} \;\langle T, C \rangle - \varepsilon \mathsf{H}(T) \tag{3.1.3}$$

and

$$\widetilde{\mathsf{OT}}_\varepsilon(\alpha, \beta) \;=\; \langle T_\varepsilon, C \rangle \qquad \text{with} \qquad T_\varepsilon \;=\; \operatorname*{argmin}_{T \in \Pi(\mathsf{a}, \mathsf{b})} \;\langle T, C \rangle - \varepsilon \mathsf{H}(T). \tag{3.1.4}$$

In the following we will refer to (3.1.3) as *vanilla* Sinkhorn approximation and to (3.1.4) as *sharp* Sinkhorn approximation.

Note that *sharp* Sinkhorn approximation corresponds to eliminating the contribution of the entropic regularizer $\mathsf{H}(T_\varepsilon)$ from OT_ε *after* the transport plan T_ε has been obtained. The function $\widetilde{\mathsf{OT}}_\varepsilon$ was *originally* introduced in Cuturi (2013) as the Sinkhorn distance, although recent literature on the topic has often adopted this name for the vanilla version (3.1.3). We will use both the notations $\widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a}, \mathsf{b})$ and $\widetilde{\mathsf{OT}}_\varepsilon(\alpha, \beta)$ (and also $\mathsf{OT}_\varepsilon(\alpha, \beta)$ and $\mathsf{OT}_\varepsilon(\mathsf{a}, \mathsf{b})$), although formally α and β are the discrete measures and a and b their weights. However, here we consider the support points as fixed and we are only interested in the dependence on a and b, therefore we use the two notations interchangeably.

As the intuition suggests, the absence of the entropic term $\mathsf{H}(T_\varepsilon)$ in the cost in (3.1.4) is reflected in a faster rate at approximating the Wasserstein distance, as characterized by the result below.

Proposition 3.1. *Let $\varepsilon > 0$. For any pair of discrete measures $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ with respective*

weights $\mathsf{a} \in \Delta_n$ *and* $\mathsf{b} \in \Delta_m$*, we have*

$$\left| \widetilde{\mathsf{OT}}_\varepsilon(\alpha,\beta) - \mathsf{W}(\alpha,\beta) \right| \leq c_1\, \mathrm{e}^{-\frac{1}{\varepsilon}}, \qquad c_2\varepsilon \leq \left| \mathsf{OT}_\varepsilon(\alpha,\beta) - \mathsf{W}(\alpha,\beta) \right| \leq c_3\varepsilon, \quad (3.1.5)$$

where c_1, c_2, c_3 *are constants independent of* ε*, depending on the support of* α *and* β*.*

The proof of the proposition above is contained in Appendix B.2, where we provide the explicit steps of the derivation of the two inequalities. Here we briefly summarize the main points. Regarding the right hand side in (3.1.5): the upper bound is a consequence of (Cuturi and Peyré, 2016, Prop. 2.1), that we also discussed in Prop. 2.11 in the case of the relative entropy KL as regularizer (that is equivalent to the case of the discrete entropy in this setting); the lower bound follows from the definition of $\mathsf{OT}_\varepsilon(\alpha,\beta)$.

The proof of the left hand side of Prop. 3.1 is a consequence of the result in Cominetti and Martín (1994)[Prop. 5.1], which proved exponential convergence of the approximation of linear programs with entropy penalty functions in more general cases. Specifying the result in the case of Optimal Transport with entropic regularization, Cominetti and Martín (1994)[Prop. 5.1] proves the convergence of T_ε in (3.1.4) to the optimal plan in (3.1.1) with maximum entropy, i.e.

$$T_\varepsilon \to T^* = \operatorname*{argmax}_{T \in \Pi(\mathsf{a},\mathsf{b})} \{\mathsf{H}(T) : \ \langle T, C\rangle = \mathsf{W}(\alpha,\beta)\}.$$

While the sharp Sinkhorn distance $\widetilde{\mathsf{OT}}_\varepsilon$ preserves the rate of converge of T_ε, the extra term $\varepsilon\mathsf{H}(T_\varepsilon)$ in the definition of the vanilla Sinkhorn distance OT_ε causes the slower rate. Prop. 3.1 suggests that the sharp Sinkhorn distance can offer a more accurate approximation of the Wasserstein distance *for a given* ε. This intuition is further supported in Example 3.1 where we compare the behaviour of the two approximations on the problem of finding an optimal transport barycenter of probability distributions.

Wasserstein Barycenters. Finding the barycenter of a set of discrete probability measures $\mathcal{D} = (\nu_i)_{i=1}^{\ell}$ is a challenging problem in applied optimal transport settings (Cuturi and Doucet, 2014). While this is the main content of next chapter, here we use the barycenter problem just as a tool to give an intuition of different properties of sharp and vanilla Sinkhorn.

The *Wasserstein barycenter* is defined as

$$\mu^*_{\mathsf{W}} \;=\; \operatorname*{argmin}_{\mu}\; \mathcal{B}_{\mathsf{W}}(\mu, \mathcal{D}), \qquad\qquad \mathcal{B}_{\mathsf{W}}(\mu, \mathcal{D}) \;=\; \sum_{i=1}^{\ell} q_i\, \mathsf{W}(\mu, \nu_i), \tag{3.1.6}$$

namely the point μ^*_{W} minimizing the weighted average distance between all distributions in the set $\mathcal{D}$, with q_i scalar weights. Finding the Wasserstein barycenter is computationally very expensive and the typical approach is to approximate it with the barycenter μ^*_ε, obtained by substituting the Wasserstein distance W with the regularized Sinkhorn distance OT_ε in the the objective functional of Eq. (3.1.6). However, in light of the result in Prop. 3.1, it is natural to ask whether the corresponding baricenter $\tilde{\mu}^*_\varepsilon$ of the sharp Sinkhorn distance $\widetilde{\mathsf{OT}}_\varepsilon$ could provide a better estimate of the Wasserstein one. While we defer a thorough empirical comparison of the two barycenters to Sec. 3.4, here we consider a simple scenario in which the sharp Sinkhorn can be proved to be a significantly better approximation of the Wasserstein distance.

Example 3.1 (Barycenter of two Deltas)**.** *We consider the problem of estimating the barycenter of two Dirac's deltas $\mu_1 = \delta_z, \mu_2 = \delta_y$ centered at $z = 0$ and $y = n$ with $z, y \in \mathbb{R}$ and n an even integer. Let $\mathcal{X} = \{x_0, \dots, x_n\} \subset \mathbb{R}$ be the set of all integers between 0 and n and C the cost matrix with squared Euclidean distances. It is well-known that the Wasserstein barycenter is the delta centered at the euclidean mean of z and y, $\mu^*_{\mathsf{W}} = \delta_{\frac{z+y}{2}}$. A direct calculation (see Appendix B.1) shows instead that the* vanilla *Sinkhorn barycenter $\mu^*_\varepsilon = \sum_{i=0}^n q_i \delta_{x_i}$ tends to spread the mass across all $x_i \in \mathcal{X}$, accordingly to the amount of regularization,*

$$q_i \propto \mathrm{e}^{-((z-x_i)^2 + (y-x_i)^2)/(2\varepsilon)} \qquad i = 0, \dots, n, \tag{3.1.7}$$

behaving similarly to a (discretized) Gaussian with standard deviation of the same order of the regularization ε. On the contrary, the sharp *Sinkhorn barycenter equals the Wasserstein one, namely $\tilde{\mu}^*_\varepsilon = \mu^*_{\mathsf{W}}$* for every $\varepsilon > 0$. *An example of this behavior is reported in Fig. 3.1.*

Main Challenges of the sharp Sinkhorn. The example above, together with Prop. 3.1, provides a strong argument in support of adopting the sharp Sinkhorn distance over its regularized version. However, while the gradient of the regularized Sinkhorn distance can be easily computed (see Cuturi and Doucet (2014) or Sec. 3.2) and therefore it is possible to address optimization problems such as the barycenter in (3.1.6) with first-order methods (e.g.

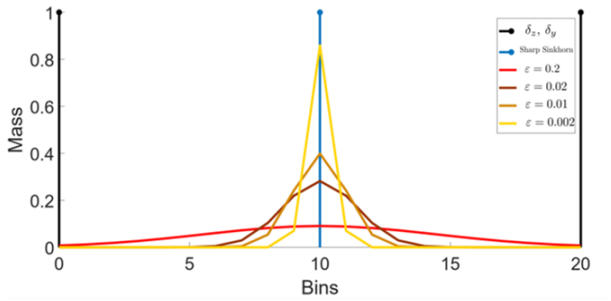


Figure 3.1: Comparison of the sharp (Blue) and regularized (Oranges) barycenters of two Dirac's deltas (Black) centered in 0 and 20 for different values of ε.

gradient descent), an explicit form for the gradient of the sharp Sinkhorn distance has not been considered. Approaches based on automatic differentiation have been recently adopted to compute the gradient of a variant of $\widetilde{\mathsf{OT}}_\varepsilon$, where the plan T_ε is the one obtained after a fixed number L of iterations (Genevay et al., 2018b; Schmitz et al., 2018; Flamary et al., 2018). These methods have been observed to be both computationally efficient and also effective in practice on a number of machine learning applications. However, in this work we are interested in investigating the analytic properties of the gradient of the sharp Sinkhorn distance, for which we provide an explicit algorithm in the following section.

3.2 Differential Properties of Sinkhorn Distances

In this section we present two main results of this chapter, namely a proof of the smoothness of the two Sinkhorn distances introduced above, and the explicit derivation of a formula for the gradient of $\widetilde{\mathsf{OT}}_\varepsilon$. These results will be key to employ the sharp Sinkhorn distance in practical applications. They are obtained leveraging the Implicit Function Theorem (Edwards, 2012) via a proof technique analogous to that in Bengio (2000) and Flamary et al. (2018) that we outline in this section and discuss in detail in the appendix.

Theorem 3.2. *For any $\varepsilon > 0$, the Sinkhorn distances OT_ε and $\widetilde{\mathsf{OT}}_\varepsilon : \Delta_n \times \Delta_n \to \mathbb{R}$ are $\mathcal{C}^\infty$ in the interior of their domain.*

Thm. 3.2 guarantees both Sinkhorn distances to be infinitely differentiable. In Sec. 3.3 this result will allow us to derive an estimator for supervised learning with Sinkhorn loss and characterize its corresponding statistical properties (i.e. universal consistency and learning rates). The proof of Thm. 3.2 is instrumental to derive a formula for the gradient of $\widetilde{\mathsf{OT}}_\varepsilon$.

Proof. Let us show the proof for $\widetilde{\mathsf{OT}}_\varepsilon$ first. We organize it in three steps:

Step 1. $\widetilde{\mathsf{OT}}_\varepsilon$ *is smooth when* T_ε *is:* when considering histograms, $\widetilde{\mathsf{OT}}_\varepsilon$ depends on its

argument a and b through the optimal coupling $T_\varepsilon(\mathsf{a},\mathsf{b})$, since the cost matrix C is fixed. Thus, since $\widetilde{\mathsf{OT}}_\varepsilon$ is a smooth function of T_ε (being the Frobenius product of T_ε with a constant matrix), showing that $\widetilde{\mathsf{OT}}_\varepsilon$ is smooth in a,b amounts to showing that T_ε is smooth in the same variables.

Step 2. T_ε is smooth when the optimal dual variables (u_*, v_*) are: By Sinkhorn's scaling theorem (Sinkhorn and Knopp, 1967), the optimal plan T_ε is characterized as follows

$$T_\varepsilon = \mathrm{diag}(\mathrm{e}^{\frac{u_*}{\varepsilon}})\mathrm{e}^{\frac{-C}{\varepsilon}}\mathrm{diag}(\mathrm{e}^{\frac{v_*}{\varepsilon}}.) \tag{3.2.1}$$

Being the exponential a smooth function, $T_\varepsilon(\mathsf{a},\mathsf{b})$ is smooth in a and b if the dual optima $u_*(\mathsf{a},\mathsf{b})$ and $v_*(\mathsf{a},\mathsf{b})$ are. Our goal is then showing the smoothness of the dual optima with respect to a and b.

Step 3. (u_*, v_*) is smooth in a,b: this is the most technical part of the proof. First of all, let us stress that one among the $n+m$ rows/columns constraints of $\Pi(\mathsf{a},\mathsf{b})$ is *redundant*: the standard dual problem recalled in (2.3.31) has an extra dual variable, and this degree of freedom is clear noticing that if (u,v) is feasible, than the pair $(u+K\mathbb{1}_n, v-K\mathbb{1}_m)$ for $K\in\mathbb{R}$ is also feasible. In the following, we get rid of the redundancy by removing one of the dual variables. Hence, let us set

$$\mathcal{L}(\mathsf{a},\mathsf{b};u,v) = -u^\top\mathsf{a} - v^\top\bar{\mathsf{b}} + \varepsilon\sum_{i,j=1}^{n,m-1}\mathrm{e}^{\frac{-(C_{ij}-u_i-v_j)}{\varepsilon}}, \tag{3.2.2}$$

where $\bar{\mathsf{b}}$ corresponds to b with the last element removed. To avoid cumbersome notation, from now on we denote $x=(\mathsf{a},\mathsf{b})$ and $\gamma=(u,v)$. The function $\mathcal{L}$ is smooth and strictly convex in γ: hence, for every fixed x in the interior of $\Delta_n\times\Delta_m$ there exist $\gamma^*(x)$ such that $\mathcal{L}(x;\gamma^*(x)) = \min_\gamma\mathcal{L}(x;\gamma)$. We now fix x_0 and show that $x\mapsto\gamma^*(x)$ is $\mathcal{C}^k$ on a neighbourhood of x_0. Set $\Psi(x;\gamma) := \nabla_\gamma\mathcal{L}(x;\gamma)$; the smoothness of $\mathcal{L}$ ensures that $\Psi\in\mathcal{C}^k$. Fix $(x_0;\gamma_0)$ such that $\Psi(x_0;\gamma_0)=0$. Since $\nabla_\gamma\Psi(x;\gamma) = \nabla^2_\gamma\mathcal{L}(x;\gamma)$ and $\mathcal{L}$ is strictly convex, $\nabla_\gamma\Psi(x_0;\gamma_0)$ is invertible. Then, by the implicit function theorem, there exist a subset $U_{x_0}\subset\Delta_n\times\Delta_m$ and a function $\phi: U_{x_0}\to\mathbb{R}^n\times\mathbb{R}^{m-1}$ such that

i) $\phi(x_0)=\gamma_0$

ii) $\Psi(x,\phi(x))=0, \qquad \forall x\in U_{x_0}$

iii) $\phi\in\mathcal{C}^k(U_{x_0})$.

For each x in U_{x_0}, since $\phi(x)$ is a stationary point for $\mathcal{L}$ and $\mathcal{L}$ is strictly convex, then

$\phi(x) = \gamma^*(x)$, which is -recalling the notation set before- (u_*, v_*). By a standard covering argument, (u_*, v_*) is $\mathcal{C}^k$ on the interior of $\Delta_n \times \Delta_m$. As this holds true for any k, the optima (u_*, v_*), and hence $\widetilde{\mathsf{OT}}_\varepsilon$, are $\mathcal{C}^\infty$ on the interior of $\Delta_n \times \Delta_m$.

Let us now focus on the smoothness of OT_ε. Note that when a, b belong to the interior of the simplex, all components are strictly positive. From the characterization of T_ε recalled in Eq. (3.2.1), we know $T_{\varepsilon ij} > 0$ for any $i, j = 1 \dots n, m$. Then, since the logarithm is a smooth function of T_ε, the term $\varepsilon\mathsf{H}(T_\varepsilon)$ is smooth in a and b. This fact combined with the first part of the proof shows the smoothness of $\mathsf{OT}_\varepsilon(\mathsf{a}, \mathsf{b}) = \langle T_\varepsilon, C\rangle - \varepsilon\mathsf{H}(T_\varepsilon)$. □

The gradient of Sinkhorn distances. After showing that Sinkhorn approximations are smooth functions with respect to the inputs a and b, we now discuss how to derive the gradient of sharp Sinkhorn with respect to one of the two variables. In both cases, the dual problem introduced in (2.3.31) plays a fundamental role. In particular, as pointed out in Cuturi and Doucet (2014), the gradient of the regularized Sinkhorn distance can be obtained directly from the dual solution as $\nabla_{\mathsf{a}}\mathsf{OT}_\varepsilon(\mathsf{a}, \mathsf{b}) = u_*(\mathsf{a}, \mathsf{b})$, for any $\mathsf{a} \in \mathbb{R}^n$ and $\mathsf{b} \in \mathbb{R}^m$. This characterization is possible because of well-known properties of primal and dual optimization problems (Bertsimas and Tsitsiklis, 1997).

The sharp Sinkhorn distance does not have a formulation in terms of a dual problem and therefore a similar argument does not apply. Nevertheless, we show here that it is still possible to obtain its gradient in closed form in terms of the dual solution.

Theorem 3.3. *Let $C \in \mathbb{R}^{n\times m}$ be a cost matrix, $\mathsf{a} \in \Delta_n$, $\mathsf{b} \in \Delta_m$ and $\varepsilon > 0$. Let $\mathcal{L}_{\mathsf{a},\mathsf{b}}(u, v)$ be defined as the argument of the maximization in the right hand side of Eq. (2.3.31), with argmax in (u_*, v_*). Let T_ε be defined as in (3.2.1). Then,*

$$\nabla_a \widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a}, \mathsf{b}) = proj_{\mathrm{T}\Delta_n}\left(A\, L\mathbb{1}_m + B\, \bar{L}^\top \mathbb{1}_n\right) \tag{3.2.3}$$

where $L = T_\varepsilon \odot C \in \mathbb{R}^{n\times m}$ is the entry-wise multiplication between T_ε and C and $\bar{L} \in \mathbb{R}^{n\times m-1}$ corresponds to L with the last column removed. The terms $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m-1}$ are

$$[A\ B] = -\frac{1}{\varepsilon}\, D\ \left[\ \nabla^2_{(u,v)}\mathcal{L}_{a,b}(u_*, v_*)\ \right]^{-1}, \tag{3.2.4}$$

with $D = [\mathrm{I}\ \mathbf{0}]$ the matrix concatenating the $n \times n$ identity matrix I and the matrix $\mathbf{0} \in \mathbb{R}^{n\times m-1}$ with all entries equal to zero. The operator $proj_{\mathrm{T}\Delta_n}$ denotes the projection onto

Algorithm 3.1 Computation of $\nabla_{\mathsf{a}}\widetilde{\mathsf{OT}}_{\varepsilon}(\mathsf{a},\mathsf{b})$

Input: $\mathsf{a} \in \Delta_n$, $\mathsf{b} \in \Delta_m$, cost matrix $C \in \mathbb{R}_+^{n,m}$, $\varepsilon > 0$.

$$
\begin{aligned}
&T = \textsc{Sinkhorn}(\mathsf{a},\mathsf{b},C,\varepsilon), \quad \bar{T} = T_{1:n,1:(m-1)} \\
&L = T \odot C, \qquad \bar{L} = L_{1:n,1:(m-1)} \\
&D_1 = \mathrm{diag}(T\mathbb{1}_m), \;\; D_2 = \mathrm{diag}(\bar{T}^\top \mathbb{1}_n)^{-1} \\
&H = D_1 - \bar{T} D_2 \bar{T}^\top, \\
&\mathrm{f} = -(L\mathbb{1}_n + D_2 \bar{T}^\top \bar{L}\mathbb{1}_{m-1}) \\
&\mathrm{g} = H^{-1}\,\mathrm{f}
\end{aligned}
$$

Return: $\mathrm{g} - \mathbb{1}_n(\mathrm{g}^\top \mathbb{1}_n)$

the tangent plane $\mathrm{T}\Delta_n = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$ *to the simplex* Δ_n.

The proof of Thm. 3.3 can be found in the supplementary material (Sec. B.3). The result is obtained by noting that the gradient of $\widetilde{\mathsf{OT}}_{\varepsilon}$ is characterized (via the chain rule) in terms of the the gradients $\nabla_{\mathsf{a}} u_*(\mathsf{a},\mathsf{b})$ and $\nabla_{\mathsf{a}} v_*(\mathsf{a},\mathsf{b})$ of the dual solutions. The main technical step of the proof is to show that these gradients correspond respectively to the terms A and B defined in (3.2.4).

To obtain the gradient of $\widetilde{\mathsf{OT}}_{\varepsilon}$ in practice, it is necessary to compute the Hessian $\nabla^2_{(u,v)}\mathcal{L}_{\mathsf{a},\mathsf{b}}(u_*,v_*)$ of the dual functional. A direct calculation shows that this corresponds to the matrix

$$
\nabla^2_{(u,v)}\mathcal{L}(u_*,v_*) = \begin{bmatrix} \mathrm{diag}(\mathsf{a}) & \bar{T}_\varepsilon \\ \bar{T}_\varepsilon{}^\top & \mathrm{diag}(\bar{\mathsf{b}}) \end{bmatrix}, \tag{3.2.5}
$$

where $\bar{T}_\varepsilon$ (respectively $\bar{\mathsf{b}}$) corresponds to T_ε (respectively b) with the last column (element) removed. See the supplementary material (Sec. B.3) for the details of this derivation.

From the discussion above, it follows that the gradient of $\widetilde{\mathsf{OT}}_{\varepsilon}$ can be obtained in closed form in terms of the transport plan T_ε. Alg. 3.1 reports an efficient approach to perform this operation. The algorithm can be derived by simple algebraic manipulation of (3.2.3), given the characterization of the Hessian in (3.2.5). We refer to Appendix B.3 for the detailed derivation of the algorithm.

Barycenters with the sharp Sinkhorn. Using Alg. 3.1 we can now apply the accelerated gradient descent approach proposed in Cuturi and Doucet (2014) to find barycenters with respect to the sharp Sinkhorn distance. Fig. 3.2 reports a qualitative experiment inspired

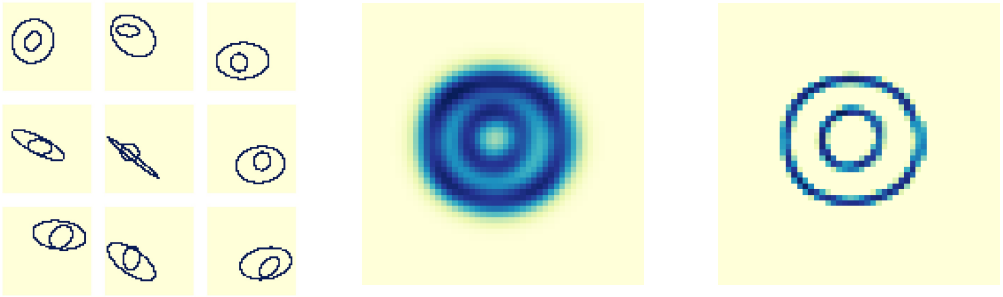


Figure 3.2: Nested Ellipses: (Left) Sample input data. (Middle) Regularized (Right) sharp Sinkhorn barycenters. We compute the barycenter of 30 nested ellipses which are represented as histograms on a 50 x 50 grid. We compare the performance of the Sharp and vanilla Sinkhorn. Sharp Sinkhorn barycenter is less impacted by the entropic regularization and suffer less blurriness, resulting more similar to the input data.

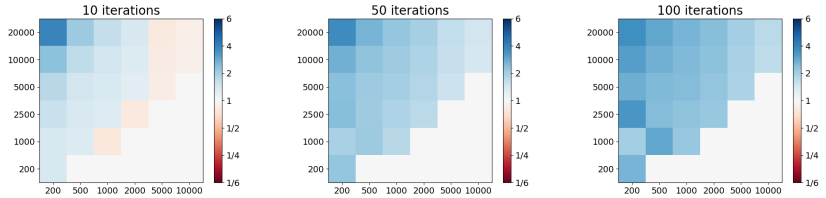


Figure 3.3: Ratio of `time`(AD) / `time`(Alg. 3.1) for 10, 50, and 100 iterations of the Sinkhorn algorithm

by the one in Cuturi and Doucet (2014), with the goal of comparing the two Sinkhorn barycenters. We considered 30 images of random nested ellipses on a 50×50 grid. We interpret each image as a distribution with support on pixels. The cost matrix is given by the squared Euclidean distances between pixels, suitably normalized. The regularization parameter in both cases is set to $\varepsilon = 0.001$. Fig. 3.2 shows some examples images in the dataset and the corresponding barycenters of the two Sinkhorn distances. While the barycenter μ_ε^* of OT_ε suffers a blurry effect, the $\widetilde{\mathsf{OT}}_\varepsilon$ barycenter $\tilde{\mu}_\varepsilon^*$ is very sharp, suggesting a better estimate of the ideal one.

We conclude this section with a computational consideration on the two methods.

Remark 3.1 (Computations). *Differentiation of sharp Sinkhorn can be efficiently carried out also via Automatic Differentiation (AD), as in Genevay et al. (2018b). Here we comment on the computational complexity of Alg. 3.1 and empirically compare the computational times of our approach and AD as dimensions and number of iterations grow. Experiments were run on a Intel(R) Xeon(R) CPU E3-1240 v3 @ 3.40GHz with 16GB RAM. The implementation of this comparison is available online*[2].

[2] https://github.com/GiulsLu/OT-gradients

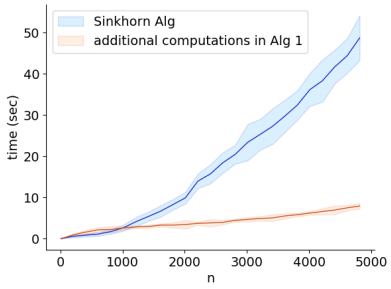


Figure 3.4: Average time (in seconds) to solve the Sinkhorn algorithm (Blue) and the remaining operations required to compute the gradient of $\widetilde{\mathsf{OT}}_\varepsilon$ in Alg. 3.1 (Orange) with respect to an increasing dimension n of the support of the distributions compared.

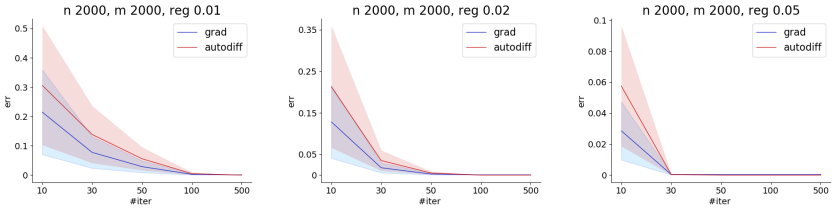


Figure 3.5: Accuracy of the Gradient obtained with Alg. 3.1 or AD with respect to the number of iterations

By leveraging the Sherman-Woodbury matrix identity, it is possible to show that the total cost of computing the gradient $\nabla_a \widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a}, \mathsf{b})$ *with* $\mathsf{a} \in \Delta_n$ *and* $\mathsf{b} \in \Delta_m$ *via Alg. 3.1 is* $O(nm \min(n, m))$. *In particular, assume* $m \leq n$. *Then, the most expensive operations are:* $O(nm^2)$ *for matrix multiplication and* $O(m^3)$ *for solving a linear system with an* $m \times m$ positive definite *matrix, where efficient off-the-shelf implementations such as Cholesky decomposition can be used.*

In Fig. 3.3 we compare the gradient obtained with Alg. 3.1 and Automatic Differentiation (AD) on random histograms with different n *(y axis),* m *(x axis), and reg.* $\lambda = 0.02$. *From left to right, we report the ratio* `time(AD) / time(Alg. 3.1)` *for* $L = 10$, $L = 50$, $L = 100$ *iterations. The results shown are averaged on* 10 *different runs. Experiments show that there exist regimes in which the gradient computed in closed form is a viable alternative to Automatic Differentiation, depending on the task. In particular, it seems that as the ratio between the supports* n *and* m *of the two distributions becomes more unbalanced, Alg. 3.1 is consistently faster than AD. Also, we note that the most expensive additional operations required to compute the gradient of sharp Sinkhorn compared to the gradient of vanilla Sinkhorn consist of matrix multiplications and the resolution of a linear system, which are very efficiently implemented on modern machines. Indeed, in our experiments the Sinkhorn*

algorithm was always the most expensive component of the computation. Fig. 3.4 reports the comparison of the time required to solve the Sinkhorn algorithm with respect to the remaining operations in Alg. 3.1 to compute the gradient of $\widetilde{\mathsf{OT}}_\varepsilon$*. It is important to notice however that in practical applications both routines can be parallelized, and several ideas can be exploited to lower the computational costs of either algorithms depending on the problem structure (see for instance the* convolutional Wasserstein distance *in Solomon et al. (2015)). Also, note that no comparison were made with a recently released library* [3] *that contains an optimized implementation of Sinkhorn algorithm. Therefore, depending on the setting, the computation of the gradient of the sharp Sinkhorn could be comparable or significantly slower than the vanilla Sinkhorn or the one obtained with automatic differentiation.*

Accuracy and approximation errors. *We conclude this discussion on computational consideration with a note on the accuracy of the method. A priori, the expression* $T_\varepsilon = \mathit{diag}(\mathrm{e}^{u_*/\varepsilon})\, \mathrm{e}^{-C/\varepsilon}\, \mathit{diag}(\mathrm{e}^{v_*/\varepsilon})$ *which is used to derive Alg. 3.1 holds 'at convergence', while in practise there is a limited budget (in terms of time and memory) for the computation of* T_ε*, i.e. limited number of iterations. In Pedregosa (2016) a similar issue is addressed. In Fig. 3.5 we empirically show that plugging an approximation* T_ε^L *obtained with a fixed number* L *of iterations in the formula for the gradient allows to reach an accuracy with respect to the 'true gradient' comparable or slightly better than automatic differentiation. Errors are measured as* ℓ^2 *norm of the difference between approximated gradient and 'true gradient', where the 'true gradient' is obtained via automatic differentiation setting* 10^5 *as maximum number of iterations.*

3.3 Learning with Sinkhorn Loss Functions

Given the characterization of smoothness for both Sinkhorn distances, in this section we focus on a specific application: supervised learning with a Sinkhorn loss function. Supervised learning with Optimal Transport loss was originally considered in Frogner et al. (2015); this work focuses on Wasserstein distance as loss function in a image-tagging application and it adopts an empirical risk minimization approach which, in practise, relies on entropic regularization for algorithmic purposes. The statistical guarantees provided for the proposed estimator are limited and provided for unregularized Wasserstein in specific settings. Motivated by this first attempt, we use Sinkhorn approximation as loss function, with the goal of providing a principled procedure that is theoretically justified and computationally

[3] https://www.kernel-operations.io/geomloss/api/pytorch-api.html

convenient.

We leverage the properties of Sinkhorn losses to study a learning algorithm with provable statistical guarantees. Differently from (Frogner et al., 2015), we interpret the output space equipped with the chosen loss function as a structured set and we rely on surrogate frameworks to design a learning procedure. This allows us to propose an estimator with strong theoretical guarantees that can be efficiently applied in practice. This section is devoted to the problem setting and the main statements, while proofs are postponed to the next section.

Problem Setting. Let $\mathcal{X}$ be an input space and $\mathcal{Y} = \Delta_n$ a set of histograms. As it is standard in supervised learning settings, the goal is to approximate a minimizer of the *expected risk*

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{S}(f(x), y)\, d\rho(x, y) \tag{3.3.1}$$

given a finite number of training points $(x_i, y_i)_{i=1}^{\ell}$ independently sampled from an unknown distribution ρ on $\mathcal{X} \times \mathcal{Y}$. The loss function $\mathcal{S} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ measures prediction errors and in our setting corresponds to either $\widetilde{\mathsf{OT}}_\varepsilon$ or OT_ε.

The idea behind the learning estimator that we will introduce below is the following: we interpret $(\mathcal{Y}, \mathcal{S})$, where $\mathcal{S}$ is either $\widetilde{\mathsf{OT}}_\varepsilon$ or OT_ε, as a *structured* output space. When dealing with structured output spaces, a viable approach consists in using a surrogate method (Bartlett et al., 2006; Mroueh et al., 2012). Intuitively, surrogate methods work as follows: using an encoding function, the original structured output space is encoded into a vector space, possibly infinite dimensional. Taking advantage of the linear structure of the space, solving the problem at this level is more amenable and a surrogate estimator can be obtained leveraging standard procedures. Once a surrogate estimator is obtained, one has to pull it back to the original problem: this is done using a decoding map which has to satisfy some assumptions in order to guarantee the effectiveness of the surrogate procedure. The formal estimator is discussed below.

Structured Prediction Estimator. Given a training set $(x_i, y_i)_{i=1}^{\ell}$, we consider $\widehat{f} : \mathcal{X} \to \mathcal{Y}$ the structured prediction estimator proposed in Ciliberto et al. (2016), defined as

$$\hat{f}(x) = \operatorname*{argmin}_{y\in\mathcal{Y}} \sum_{i=1}^{\ell} w_i(x)\, \mathcal{S}(y, y_i) \tag{3.3.2}$$

for any $x \in \mathcal{X}$. The weights $w_i(x)$ are learned from the data and can be interpreted as scores

suggesting the candidate output distribution y to be close *according to the metric* $\mathcal{S}$ to a specific output distribution y_i observed in training. While different learning strategies can be adopted to learn the score vector w, we consider the kernel-based approach in Ciliberto et al. (2016). In particular, given a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Aronszajn, 1950), we have

$$w(x) = (w_1(x), \dots, w_\ell(x))^\top = (K + \gamma \ell I)^{-1} K_x \tag{3.3.3}$$

where $\gamma > 0$ is a regularization parameter while $K \in \mathbb{R}^{\ell \times \ell}$ and $K_x \in \mathbb{R}^n$ are respectively the empirical kernel matrix with entries $K_{ij} = k(x_i, x_j)$ and the evaluation vector with entries $(K_x)_i = k(x, x_i)$, for any $i, j = 1, \dots, \ell$.

Remark 3.2 (Structured Prediction and Differentiability of Sinkhorn)**.** *The differentiability properties studied in the previous sections of this chapter play a double role. On the practical side, the gradient estimation algorithm in Alg. 3.1 allows to solve the optimization problem by adopting first order methods such as gradient descent. On the theoretical side, the smoothness guaranteed by Thm. 3.2 will allow us to characterize the generalization properties of the estimator.*

3.3.1 Theoretical Guarantees

In this section we study the theoretical guarantees of the estimator introduced in (3.3.2). We study consistency and finite samples bounds.

Universal Consistency of $\widehat{f}$. We start by showing $\widehat{f}$ is *universally consistent*, namely that it achieves minimum expected risk as the number ℓ of training points increases. To avoid technical issues on the boundary, in the following we will require $\mathcal{Y} = \Delta_n^\theta$ for some $\theta > 0$ to be the set of points $p \in \Delta_n$ with $p_i \geq \theta$ for any $i = 1, \dots, n$. The main technical step in this context is to show that for any smooth loss function on $\mathcal{Y}$, the estimator in Eq. (3.3.2) is consistent. In this sense, the characterization of smoothness in Thm. 3.2 is key to prove the following result, in combination with Thm. 4 in Ciliberto et al. (2016).

Theorem 3.4 (Universal Consistency)**.** *Let $\mathcal{Y} = \Delta_n^\theta$, $\varepsilon > 0$ and $\mathcal{S}$ be either OT_ε or $\widetilde{\mathsf{OT}}_\varepsilon$. Let k be a bounded continuous universal[4] kernel on $\mathcal{X}$. For any $\ell \in \mathbb{N}$ and any distribution ρ on $\mathcal{X} \times \mathcal{Y}$ let $\widehat{f}_\ell : \mathcal{X} \to \mathcal{Y}$ be the estimator in (3.3.2) trained with $(x_i, y_i)_{i=1}^\ell$ points*

[4]This is a standard assumptions for universal consistency (see (Steinwart and Christmann, 2008)). Example: $k(x, x') = \mathrm{e}^{-\|x - x'\|^2 / \sigma}$.

independently sampled from ρ *and* $\gamma_\ell = \ell^{-1/4}$. *Then*

$$\lim_{\ell\to\infty} \mathcal{E}(\widehat{f}_\ell) \;=\; \min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \quad \textit{with probability } 1. \tag{3.3.4}$$

A result analogous to the one above was originally proved in (Ciliberto et al., 2016, Thm. 4) for a wide family of functions referred to as *Structure Encoding Loss Function (SELF)* (Ciliberto et al., 2017) whose definition is recalled below:

Definition 3.1 (SELF)**.** *Let* $\mathcal{Y}$ *be a set. A function* $\mathcal{S} : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ *is a* Structure Encoding Loss Function *(SELF) if there exists a separable Hilbert space* $\mathcal{H}_\mathcal{Y}$ *with inner product* $\langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{Y}}$, *a continuous map* $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ *and a bounded linear operator* $V : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ *such that*

$$\mathcal{S}(y, y') = \langle \psi(y), V\psi(y')\rangle_{\mathcal{H}_\mathcal{Y}} \qquad y,\, y' \in \mathcal{H}_\mathcal{Y}. \tag{3.3.5}$$

While several classical loss functions used in structured prediction have been observed to satisfy this SELF definition, such characterization was not available for the Sinkhorn distances. The main technical step in the proof of Thm. 3.4 in this sense is to prove that any smooth function on $\mathcal{Y}$ satisfies the definition of SELF. This result is provided in the lemma below:

Theorem 3.5. *(Smooth functions are SELF) Let* $\mathcal{Y}$ *be a compact subset of* $\mathbb{R}^n$. *Any function* $\mathcal{S} : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ *such that* $\mathcal{S} \in \mathcal{C}^\infty(\mathcal{Y}\times\mathcal{Y})$ *is SELF.*

Proof. By assumption $\mathcal{S} \in \mathcal{C}^\infty(\mathcal{Y}\times\mathcal{Y})$. Since $\mathcal{Y}$ is compact,

$$\mathcal{C}^\infty(\mathcal{Y}\times\mathcal{Y}) = \mathcal{C}^\infty(\mathcal{Y}) \otimes \mathcal{C}^\infty(\mathcal{Y}) \subset H^r(\mathcal{Y}) \otimes H^r(\mathcal{Y}), \tag{3.3.6}$$

for $r > n/2$, where $H^r(\mathcal{Y})$ is the Sobolev space made of L^2 functions weakly differentiable r times (Brezis, 2010) and $\mathcal{C}^\infty(\mathcal{Y}) \otimes \mathcal{C}^\infty(\mathcal{Y})$ denotes the completion of the topological tensor product of $\mathcal{C}^\infty(\mathcal{Y})$ with itself with respect to the projective topology (see (Treves, 2016, Def 43.2 and 43.5)); the first equality follows from (Treves, 2016, Thm. 56.1). The Sobolev space $H^r(\mathcal{Y})$ is a Reproducing Kernel Hilbert Space (RKHS) (Berlinet and Thomas-Agnan, 2011) and we denote by $\mathsf{k}_y = \mathsf{k}(y,\cdot) \in H^r(\mathcal{Y})$ the reproducing kernel. The product space $H^r \otimes H^r$ is also an RKHS with reproducing kernel $\mathsf{K}((y_1,y_2),(y_1',y_2')) = \mathsf{k}(y_1,y_1')\mathsf{k}(y_2,y_2')$, i.e. in general $\mathsf{K}_{y,y'} = \mathsf{k}_y \otimes \mathsf{k}_{y'}$. Since $\mathcal{S} \in H^r \otimes H^r$, by reproducing property there exists a

function $V \in H^r \otimes H^r$ such that

$$\mathcal{S}(y, y') = \langle V, \mathsf{k}_y \otimes \mathsf{k}_{y'} \rangle_{H^r \otimes H^r}.$$

By the isometric isomorphism $H^r \otimes H^r \cong \mathrm{HS}(H^r, H^r)$ (Moretti, 2013), with $\mathrm{HS}(H^r, H^r)$ the space of Hilbert-Schmidt operators from H^r to itself, it holds

$$\mathcal{S}(y, y') = \langle V, \mathsf{k}_y \otimes \mathsf{k}_{y'} \rangle_{H^r \otimes H^r} = \langle V, \mathsf{k}_y \otimes \mathsf{k}_{y'} \rangle_{\mathrm{HS}} = \mathrm{Tr}(V^* \mathsf{k}_y \otimes \mathsf{k}_{y'}) = \langle \mathsf{k}_{y'}, V^* \mathsf{k}_y \rangle_{H^r}, \tag{3.3.7}$$

where V^* is the adjoint operator of V. To meet the conditions of Def. 3.1 it remains to show that V^* and k_y are bounded. But k_y is bounded in H^r for any $y \in \mathcal{Y}$ by definition of reproducing kernel and the operator norm $\|V^*\|$ is bounded from above by the Hilbert-Schmidt norm $\|V\|_{\mathrm{HS}}$ which is trivially bounded since $V \in \mathrm{HS}(H^r, H^r)$. □

Corollary 3.6. *The vanilla and sharp Sinkhorn losses* OT_ε *and* $\widetilde{\mathsf{OT}}_\varepsilon : \Delta_n^\theta \times \Delta_n^\theta \to \mathbb{R}$ *are SELF.*

Proof. Since $\Delta_n^\theta \subset \Delta_n$ is compact and OT_ε, $\widetilde{\mathsf{OT}}_\varepsilon$ are $\mathcal{C}^\infty$ in the interior on $\Delta_n \times \Delta_n$ by Thm. 3.2, a direct application of the result above shows that OT_ε and $\widetilde{\mathsf{OT}}_\varepsilon$ are SELF. □

Combining this result with Thm.4 in Ciliberto et al. (2016), we obtain that *for every smooth loss function* $\mathcal{S}$ *on* $\mathcal{Y}$ *the corresponding estimator* $\widehat{f}$ *in Eq. (3.3.2) is universally consistent*. Since we have shown that vanilla and sharp Sinkhorn are smooth, the proof of Thm. 3.4 easily follows:

Proof. Since OT_ε, $\widetilde{\mathsf{OT}}_\varepsilon$ are SELF functions and Δ_n^θ is compact, the result follows from Thm. 4 in Ciliberto et al. (2016). □

Thm. 3.4 guarantees $\widehat{f}$ to be a valid estimator for the learning problem. To our knowledge, this is the first result characterizing the universal consistency of an estimator for supervised learning problem with Entropic Optimal Transport losses.

Learning Rates. By imposing standard regularity conditions on the learning problem, it is possible to provide also excess risk bounds for $\widehat{f}$.

We start from the observation (see e.g. Lemma 6 in Ciliberto et al. (2016)) that the solution $f^* : \mathcal{X} \to \mathcal{Y}$ of the learning problem introduced in (3.3.1) is such that

$$f^*(x) = \operatorname*{argmin}_{z \in \mathcal{Y}} \int_{\mathcal{Y}} \mathcal{S}(z, y)\, d\rho(y|x) \tag{3.3.8}$$

almost surely on $\mathcal{X}$. In particular $f^*(x)$ corresponds to the minimizer of the *conditional expectation* $\mathbb{E}_{y|x}\mathcal{S}(z, y)$ of the loss $\mathcal{S}(z, y)$ with respect to y given $x \in \mathcal{X}$. As it is standard in statistical learning theory, in order to obtain generalization bounds for estimating f^* we will impose regularity assumptions on the conditional distribution $\rho(\cdot|x)$ or, more precisely, on its corresponding *conditional mean embedding* ((Song et al., 2009, 2013)) with respect to a suitable space of functions.

Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the kernel on $\mathcal{Y}$ associated to the reproducing kernel Hilbert spaces (RKHS) (whose definition is recalled in Appendix A.3) $\mathcal{H} = W_2^{(d+1)/2}(\mathcal{Y})$, the Sobolev space of square integrable functions with smoothness $\frac{d+1}{2}$ (see e.g. (Wendland, 2004)). We consider a function $g^* : \mathcal{X} \to \mathcal{H}$ such that

$$g^*(x) = \int_{\mathcal{Y}} h(y, \cdot) d\rho(y|x) \tag{3.3.9}$$

almost surely on $\mathcal{X}$. For any $x \in \mathcal{X}$, the quantity $g^*(x)$ is known as the *conditional mean embedding* of $\rho(\cdot|x)$ in $\mathcal{H}$, originally introduced in Song et al. (2009, 2013). In particular, in Song et al. (2009) it was shown that in order to obtain learning rates for an estimator approximating g^*, a key assumption is that g^* belongs to $\mathcal{H} \otimes \mathcal{F}$, the tensor product between the space $\mathcal{H}$ on the output and the space $\mathcal{F}$ associated to a reproducing kernel on the input. In this work we will require the same assumption.

It can be verified that $\mathcal{H} \otimes \mathcal{F}$ is a RKHS for vector-valued functions (Micchelli and Pontil, 2005; Lever et al., 2012; Alvarez et al., 2012) and that by asking $g^* \in \mathcal{H} \otimes \mathcal{F}$ we are requiring the conditional mean embedding of $\rho(\cdot|x)$ to be sufficiently regular *as a function on* $\mathcal{X}$. We are now ready to report our result on the statistical performance of $\widehat{f}$.

Theorem 3.7 (Learning Rates)**.** *Let $\mathcal{Y} = \Delta_n^\theta$, $\theta > 0$ and $\mathcal{S}$ be either OT_ε or $\widetilde{\mathsf{OT}}_\varepsilon$. Let $\mathcal{H} = W_2^{(n+1)/2}(\mathcal{Y})$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded continuous reproducing kernel on $\mathcal{X}$ with associated RKHS $\mathcal{F}$. Let $\hat{f}_\ell : \mathcal{X} \to \mathcal{Y}$ be the estimator in Eq. (3.3.2) trained with ℓ training points independently sampled from ρ and with $\gamma = \ell^{-1/2}$. If g^* defined in Eq. (3.3.9)*

is such that $g^* \in \mathcal{H} \otimes \mathcal{F}$*, then*

$$\mathcal{E}(\hat{f}_\ell) - \min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \leq c\,\tau^2 \ell^{-1/4} \tag{3.3.10}$$

holds with probability $1 - 8\mathrm{e}^{-\tau}$ *for any* $\tau > 0$*, with* c *a constant independent of* ℓ *and* τ*.*

The proof of Thm. 3.7 requires to combine our characterization of the Sinkhorn distances (or more generally smooth functions on $\mathcal{Y}$) as structure encoding loss functions (see Sec. 3.3.1) with Thm. 5 in Ciliberto et al. (2016) where a result analogous to the one above is reported for SELF loss functions.

Remark 3.3. *While the losses used in this Chapter use the original notions of Entropic regularization, the statistical guarantees of Thm. 3.4 and Thm. 3.7 still hold true when using as a loss the actual Sinkhorn divergence recalled in Sec. 2.5, where the autocorrelation terms are removed.*

Remark 3.4. *A relevant question is whether the Wasserstein distance could be similarly framed in the setting of structured prediction. However, the argument used to address Sinkhorn distances relies on their smoothness properties and cannot be extended to the Wasserstein distance, which is not smooth. A completely different approach may still be successful and it is an interesting question for future work.*

We conclude this section with a note on previous work. We recall that (Frogner et al., 2015) is the first work proposing to use Optimal Transport-related loss function in supervised learning setting. It deploys the entropic regularization in the algorithmic aspects, while sticking to unregularized Wasserstein distance in the design of the model and in the statistical analysis provided. They show a *generalization bounds* for an estimator minimizing the 1-Wasserstein distance. This chapter focuses on Entropic Optimal Transport as loss function and not as a computational tool that serves as a surrogate of Wasserstein distance. We exploit regularity properties that Entropic OT benefits from to characterize both the consistency and the *excess risk bounds* in Thm. 3.7 of the estimator in (3.3.2). The two approaches and analysis are based on different assumptions on the problem. Therefore, a comparison of the corresponding learning rates is outside the scope of this work.

3.4 Experiments

We present here some experiments that compare the two Sinkhorn approximations (sharp and vanilla) empirically. Optimization was performed with the accelerated gradient descent from

Table 3.1: Average absolute improvement in terms of the ideal Wasserstein barycenter functional $\mathcal{B}_{\mathsf{W}}$ in Eq. (3.1.6) of *sharp* vs *vanilla* Sinkhorn, for barycenters of random measures with sparse support.

	Support (% of total bins)			
	1%	2%	10%	50%
Improvement $\mathcal{B}_{\mathsf{W}}(\tilde{\mu}_\varepsilon^*) - \mathcal{B}_{\mathsf{W}}(\mu_\varepsilon^*)$	14.914 ± 0.076	12.482 ± 0.135	2.736 ± 0.569	0.258 ± 0.012

(Cuturi and Doucet, 2014) for $\widetilde{\mathsf{OT}}_\varepsilon$ and Bregman projections (Benamou et al., 2015) for OT_ε.

Barycenters with Sinkhorn Distances. We compared the quality of Sinkhorn barycenters in terms of their approximation of the (ideal) Wasserstein barycenter. We considered discrete distributions on 100 bins, corresponding to the integers from 1 to 100 and a squared Euclidean cost matrix C. We generated 4 different datasets of 10 measures each: these datasets contain histograms where only $k = 1, 2, 10, 50$ (respectively) randomly chosen consecutive bins are different from zero, with the non-zero entries sampled uniformly at random between 0 and 1 (and then normalized to sum up to 1). In Table 3.1 they correspond to the entries denoted by 1%, 2%, 10% and 50% support. We empirically chose the Sinkhorn regularization parameter ε to be the smallest value such that the output T_ε of the Sinkhorn algorithm would be within 10^{-6} from the transport polytope in 1000 iterations. Table 3.1 reports the absolute improvement of the barycenter of the sharp Sinkhorn distance with respect to the one obtained with the regularized Sinkhorn, averaged over 10 independent dataset generation for each support size k. We call absolute improvement the following quantity $\mathcal{B}_{\mathsf{W}}(\tilde{\mu}_\varepsilon^*) - \mathcal{B}_{\mathsf{W}}(\mu_\varepsilon^*)$ that compares the quality of the barycenters $\tilde{\mu}_\varepsilon^*$ and μ_ε^* in terms of the barycenter functional with Wasserstein distance $\mathcal{B}_{\mathsf{W}}$. The quantity $\mathcal{B}_{\mathsf{W}}(\tilde{\mu}_\varepsilon^*) - \mathcal{B}_{\mathsf{W}}(\mu_\varepsilon^*)$ highlights how different $\tilde{\mu}_\varepsilon^*$ and μ_ε^* are in terms of the approximation of the actual Wasserstein barycenter and it has to be interpreted as follows: the bigger this quantity is, the greater the discrepancy between using sharp and vanilla Sinkhorn (and the more favorable for the sharp). As can be noticed, the sharp Sinkhorn consistently outperforms the vanilla counterpart. Interestingly, this improvement is more evident for measures with sparse support and tends to reduce as the support increases. This is in line with the remark in example Example 3.1 and the fact that the regularization term in OT_ε tends to encourage oversmoothed solutions.

Learning with Wasserstein loss. We evaluated the Sinkhorn distances in an image reconstruction problem similar to the one considered in Weston et al. (2003) for structured

prediction. Given an image depicting a drawing, the goal is to learn how to reconstruct the lower half of the image (output) given the upper half (input). Similarly to Cuturi and Doucet (2014) we interpret each (half) image as an histogram with mass corresponding to the gray levels (normalized to sum up to 1). For all experiments, following Ciliberto et al. (2016), we evaluated the performance of the reconstruction in terms of the classification accuracy of an image recognition SVM classifier trained on a separate dataset. To train the structured prediction estimator in (3.3.2) we used a Gaussian kernel with bandwith σ and regularization parameter γ selected by cross-validation.

			Reconstruction Error (%)	
# Classes	$\widetilde{\mathsf{OT}}_\varepsilon$	OT_ε	Hell(Ciliberto et al., 2017)	KDE (Weston et al., 2003)
2	$\mathbf{3.7 \pm 0.6}$	4.9 ± 0.9	8.0 ± 2.4	12.0 ± 4.1
4	$\mathbf{22.2 \pm 0.9}$	31.8 ± 1.1	29.2 ± 0.8	40.8 ± 4.2
10	$\mathbf{38.9 \pm 0.9}$	44.9 ± 2.5	48.3 ± 2.4	64.9 ± 1.4

Table 3.2: Average reconstruction errors of the Sinkhorn (both sharp and vanilla), Hellinger, and KDE estimators on the Google QuickDraw reconstruction problem. We considered the following classes in the Google QuickDraw doodling dataset: *fish*, *apple*, (2) *mug*, *candle*, (4) *flower*, *moon*, *mushroom*, *hand*, *crown*, *broom* (10). The images have size 28 x 28. We interpret them as histograms over the grid of pixels. The input space $\mathcal{X}$ consists of the upper halves of images. The outputs space $\mathcal{Y}$ consists of the lower halves of images. We train the estimator described in the text on a dataset containing 1000 images per class and we test it on other 1000 images. The results reported are averaged on 5 independent runs. The cost matrix, containing the pairwise squared distance between pixels is normalized and the regularization parameter is set to $\varepsilon = 0.01$. The Sinkhorn losses outperform Hellinger and KDE: since they are sensitive to the geometric structure of the distributions they are better suited at capturing the shape of the images. This has a positive impact in the reconstruction.

• *Google QuickDraw.* We compared the performance of the two estimators on a challenging dataset. We selected $c = 2, 4, 10$ classes from the Google QuickDraw dataset (Google, 2017) which consists in images of size 28×28 pixels. We trained the structured prediction estimators on 1000 images per class and tested on other 1000 images. We repeated these experiments 5 times, each time randomly sampling a different training and test dataset. Table 3.2 reports the reconstruction error (i.e. the classification error of the SVM classifier) over images reconstructed by the Sinkhorn estimators, the structured prediction estimator with Hellinger loss (Ciliberto et al., 2016) and the Kernel Dependency Estimator (KDE) (Weston et al., 2003). As can be noticed, both Sinkhorn estimators perform significantly better than their competitors (except the Hellinger distance outperforming OT_ε on 4 classes). These results are not surprising as the geometric properties of optimal transport distances

make them particularly suitable in comparing probability measures with 'defined geometric structure' as it is the case here: indeed when interpreting images as probability measures over the grid of pixel, the the support of the measure has a defined shape (e.g. it is concentrated on a mug-shaped support for those in Mug class and similarly for the other classes). This is in line with the intuition that optimal transport metrics respect the way the mass is distributed on images (Cuturi, 2013; Cuturi and Doucet, 2014). Moreover, it is interesting to note that the estimator of the sharp Sinkhorn distance provides always better reconstructions than its vanilla counterpart.

3.5 Discussion

In this chapter we investigated the differential properties of Sinkhorn distances. We proved the smoothness of the two functions and derived an explicit algorithm to efficiently compute the gradient of the sharp Sinkhorn distance. Our result allows to employ the sharp Sinkhorn distance in applications that rely on first order optimization methods, such as in approximating Wasserstein barycenters and supervised learning on probability distributions. In this latter setting, our characterization of smoothness allowed to study the statistical properties of the Sinkhorn distance as loss function. In particular we considered a structured prediction estimator for which we proved universal consistency and generalization bounds.

Chapter 4

Free-support Sinkhorn barycenters

The previous chapter was devoted to Sinkhorn divergences as losses in supervised learning settings. In many applications however, data comes unlabelled and the primary interest switches from prediction to identification of clusters or aggregation of the data. A relevant example is the following: assume that multiple sensors collect data from the same environment with different noise distributions. One may be interested in assembling the gathered samples into a single signal, averaging out the noise due to individual measurements. When averaging different distributions, the choice of the metric has a deep impact. Optimal Transport distances and Sinkhorn divergences are particularly suitable in averaging problems, since they capture the geometric structure of the distributions. OT and Sinkhorn barycenters have been successfully used in many settings, including texture mixing (Rabin et al., 2011), Bayesian inference (Srivastava et al., 2018), imaging (Gramfort et al., 2015), or model ensemble (Dognin et al., 2018).

The notion of barycenter in Wasserstein space was first introduced by Agueh and Carlier (2011) and later investigated from the algorithmic perspective, in case of both the original Wasserstein distance (Staib et al., 2017; Claici et al., 2018) and its entropic regularizations (Cuturi and Doucet, 2014; Benamou et al., 2015; Dvurechenskii et al., 2018). Two main challenges in this regard are: i) how to efficiently identify the support of the candidate barycenter and ii) how to deal with continuous (or infinitely supported) probability measures. The first problem is typically addressed by either fixing the support of the barycenter a-priori (Staib et al., 2017; Dvurechenskii et al., 2018) or by adopting an alternating minimization procedure to iteratively optimize the support point locations and their weights (Cuturi and Doucet, 2014; Claici et al., 2018). While fixed-support methods enjoy better theoretical guarantees, free-support algorithms are more flexible, more memory efficient and practicable

in high dimensional settings. Free support methods are not affected by any approximation that is artificially introduced when fixing a priori the support of the target barycenter. In addition, they are more suited to address the case where input measures are continuous. The problem of dealing with continuous measures constitutes a challenge on its own; so far it has been mainly approached by adopting stochastic optimization methods to minimize the barycenter functional (Claici et al., 2018; Staib et al., 2017; Dvurechenskii et al., 2018).

In this chapter, we study this averaging, or barycenter, problem with respect to Sinkhorn divergence in *full* generality. We present a novel algorithm to estimate the barycenter of arbitrary probability distributions that does not require to fix the support beforehand. Based on a Frank-Wolfe optimization strategy, our approach proceeds by populating the support of the barycenter incrementally, without requiring any pre-allocation. In particular, the algorithm adds new points and updates their weights at each iteration, similarly to kernel herding strategies (Bach et al., 2012) and conditional gradient for sparse inverse problem (Bredies and Pikkarainen, 2013; Boyd et al., 2017). We consider discrete as well as continuous distributions, proving convergence rates of the proposed algorithm in both settings. Key elements of our analysis are a new result showing that the Sinkhorn divergence on compact domains has Lipschitz continuous gradient with respect to the Total Variation and a characterization of the sample complexity of Sinkhorn potentials. These regularity results complement recent work in Feydy et al. (2019) on theoretical properties of Sinkhorn divergence and are of independent interest. Experiments validate the effectiveness of our method in practice. This chapter will mainly discuss the work in Luise et al. (2019).[1]

Contributions. The main contributions of this chapter are the following: i) we show that the gradient of the Sinkhorn divergence is Lipschitz continuous on the space of probability measures with respect to the Total Variation. This result grants us convergence of the barycenter algorithm in finite settings. It also provides further understanding of the theoretical benefits of adding entropy penalty in terms of regularity properties of the divergence; ii) We characterize the sample complexity of Sinkhorn potentials of two empirical distributions sampled from arbitrary probability measures. While standard statistical results focus on the sample complexity of the divergence itself, here we show finite sample bounds for the gradients. This latter result allows us to iii) provide a concrete optimization scheme

[1] Advances in Neural Information Processing Systems (NeurIPS), Dec 2019, Vancouver, Canada.

to approximately solve the barycenter problem for arbitrary probability measures with convergence guarantees. iv) A byproduct of our analysis is the generalization of the Frank-Wolfe (FW) algorithm to settings where the objective functional is defined only on a set with empty interior, which is the case for Sinkhorn divergence barycenter problem.

The chapter is organized as follows: Sec. 4.1 briefly recalls the definitions that are used in this chapter. Sec. 4.2 introduces the barycenter functional, and proves the Lipschitz continuity of its gradient. Sec. 4.5 describes the implementation of our algorithm and Sec. 4.6 studies its convergence rates. Finally, Sec. 4.7 evaluates the proposed methods empirically and Sec. 4.8 provides concluding remarks.

4.1 Setting

In this section we briefly recall for convenience some definitions and properties of entropy-regularized Optimal Transport that will be repeatedly used in this chapter. We refer to Chapter 2 for more details and a thorough presentation. We consider a compact set $\mathcal{X} \subset \mathbb{R}^d$ and a symmetric cost function $\mathsf{c}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We set $\mathsf{D} := \sup_{x,y\in\mathcal{X}} \mathsf{c}(x,y)$ and denote by $\mathcal{P}(\mathcal{X})$ the space of probability measures on $\mathcal{X}$ (positive Radon measures with mass 1). For any $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, the Optimal Transport problem with entropic regularization is defined in (2.3.1). In this chapter we mostly need the *dual* formulation that is recalled here for convenience:

$$\mathsf{OT}_\varepsilon(\alpha,\beta) = \max_{u,v\in\mathcal{C}(\mathcal{X})} \int u(x)\, d\alpha(x) + \int v(y)\, d\beta(y) - \varepsilon \int e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\alpha(x) d\beta(y), \tag{4.1.1}$$

where $\mathcal{C}(\mathcal{X})$ denotes the space of real-valued continuous functions on $\mathcal{X}$, endowed with $\|\cdot\|_\infty$. Let $\mu \in \mathcal{P}(\mathcal{X})$. We denote by $\mathsf{P}_\mu\colon \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ the map such that, for any $w \in \mathcal{C}(\mathcal{X})$,

$$\mathsf{P}_\mu(w)\colon x \mapsto -\varepsilon \log \int e^{\frac{w(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\mu(y). \tag{4.1.2}$$

The first order optimality conditions for (4.1.1) are

$$u = \mathsf{P}_\beta(v) \quad \alpha\text{- a.e.} \qquad \text{and} \qquad v = \mathsf{P}_\alpha(u) \quad \beta\text{- a.e.} \tag{4.1.3}$$

Recall that pairs (u, v) satisfying (4.1.3) are referred to as *Sinkhorn* potentials. The properties of the Sinkhorn potentials have been presented in Sec. 2.3.3 in Chapter 2. In the following we assume (u, v) to be the Sinkhorn potentials such that: $i)$ $u(x_o) = 0$ for an arbitrary anchor

point $x_o \in \mathcal{X}$ and $ii)$ (4.1.3) is satisfied pointwise on the entire domain $\mathcal{X}$. The extended potentials play the role of gradients as was presented in detail in Sec. 2.3.5. In this chapter we use the unbiased version of the entropic regularization, referred to as Sinkhorn divergence and recalled in Sec. 2.5. We rewrite the definition below for convenience:

$$\mathsf{S}_\varepsilon : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}, \qquad (\alpha, \beta) \mapsto \mathsf{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\mathsf{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathsf{OT}_\varepsilon(\beta, \beta). \quad (4.1.4)$$

Remark 4.1. *The gradient of Sinkhorn divergence is a core concept in this chapter, and we refer to Prop. 2.10 and Prop. 2.17 in Chapter 2 for the details. Here we just recall that defining* $\nabla\mathsf{OT}_\varepsilon : \mathcal{P}(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})^2$ *to be the map*

$$\nabla\mathsf{OT}_\varepsilon(\alpha, \beta) = (u, v), \qquad \textit{with} \qquad u = \mathsf{P}_\beta(v), \;\; v = \mathsf{P}_\alpha(u) \;\; \textit{on } \mathcal{X}, \quad u(x_o) = 0, \tag{4.1.5}$$

then, for any $\alpha, \alpha', \beta, \beta' \in \mathcal{P}(\mathcal{X})$*, the directional derivative of* OT_ε *along* $(\mu, \nu) = (\alpha' - \alpha, \beta' - \beta)$ *is*

$$\mathsf{OT}'_\varepsilon(\alpha, \beta; \mu, \nu) = \langle \nabla\mathsf{OT}_\varepsilon(\alpha, \beta), (\mu, \nu) \rangle = \langle u, \mu \rangle + \langle v, \nu \rangle \,, \tag{4.1.6}$$

where $\langle w, \rho \rangle = \int w(x)\, d\rho(x)$ *denotes the canonical pairing between the spaces* $\mathcal{C}(\mathcal{X})$ *and* $\mathcal{M}(\mathcal{X})$*. Also, for any* $\beta \in \mathcal{P}(\mathcal{X})$ *the gradient of* $S_\varepsilon(\cdot, \beta)$ *is*

$$\nabla[S_\varepsilon(\cdot, \beta)] : \mathcal{P}(\mathcal{X}) \to \mathcal{C}(\mathcal{X}) \qquad \alpha \mapsto \nabla_1\mathsf{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\nabla_1\mathsf{OT}_\varepsilon(\alpha, \alpha) = u - p, \quad (4.1.7)$$

with $u = \mathsf{P}_{\beta\alpha}(u)$ *and* $p = \mathsf{P}_\alpha(p)$*.*

4.2 Sinkhorn barycenters with Frank-Wolfe

Given $\beta_1, \dots \beta_m \in \mathcal{P}(\mathcal{X})$ and $\omega_1, \dots, \omega_m \geq 0$ a set of weights such that $\sum_{j=1}^m \omega_j = 1$, the main goal of this chapter is to solve the following *Sinkhorn barycenter* problem

$$\min_{\alpha \in \mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha), \qquad \text{with} \qquad \mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^m \omega_j \, \mathsf{S}_\varepsilon(\alpha, \beta_j). \tag{4.2.1}$$

Although the objective functional B_ε is convex, its domain $\mathcal{P}(\mathcal{X})$ has *empty* interior in the space of finite signed measure $\mathcal{M}(\mathcal{X})$. Hence standard notions of Fréchet or Gâteaux differentiability do not apply. This causes some difficulties in devising optimization methods.

To circumvent this issue, here we adopt the Frank-Wolfe (FW) algorithm. Indeed, one key advantage of this method is that it is formulated in terms of directional derivatives along feasible directions (i.e., directions that locally remain inside the constraint set). Building upon (Demyanov and Rubinov, 1967, 1968; Dunn and Harshbarger, 1978), which study the algorithm in Banach spaces, we show that the "weak" notion of directional differentiability of S_ε (and hence of B_ε) in Remark 4.1 is sufficient to carry out the convergence analysis. While full details are provided in Appendix C.1, below we give an overview of the main result.

Frank-Wolfe in dual Banach spaces. Let $\mathcal{W}$ be a real Banach space with topological dual $\mathcal{W}^*$ and let $\mathcal{D} \subset \mathcal{W}^*$ be a nonempty, convex, closed and bounded set. For any $w \in \mathcal{W}^*$ denote by $\mathcal{F}_{\mathcal{D}}(w) = \mathbb{R}_+(\mathcal{D}-w)$ the set of feasible direction of $\mathcal{D}$ at w (namely $s = t(w'-w)$ with $w' \in \mathcal{D}$ and $t > 0$). Let $\mathsf{G}\colon \mathcal{D} \to \mathbb{R}$ be a convex function and assume that there exists a map $\nabla\mathsf{G}\colon \mathcal{D} \to \mathcal{W}$ (not necessarily unique) such that $\langle \nabla\mathsf{G}(w), s\rangle = \mathsf{G}'(w; s)$ for every $s \in \mathcal{F}_{\mathcal{D}}(w)$. In Alg. 4.1 we present a method to minimize G. The algorithm is structurally equivalent to the standard FW (Dunn and Harshbarger, 1978; Jaggi, 2013) and accounts for possible inaccuracies in solving the minimization in step (i). This will be key in Sec. 4.6 when studying the barycenter problem for β_j with infinite support. The following result (see proof in Appendix C.1) shows that under the additional assumption that $\nabla\mathsf{G}$ is Lipschitz-continuous and with sufficiently fast decay of the errors, the above procedure converges in value to the minimum of G with rate $O(1/k)$. Here $\operatorname{diam}(\mathcal{D})$ denotes the diameter of $\mathcal{D}$ with respect to the dual norm.

Theorem 4.1. *Under the assumptions above, suppose in addition that $\nabla\mathsf{G}$ is L-Lipschitz continuous with $L > 0$. Let $(w_k)_{k\in\mathbb{N}}$ be obtained according to Alg. 4.1. Then, for every integer $k \geq 1$,*

$$\mathsf{G}(w_k) - \min \mathsf{G} \leq \frac{2}{k+2} L \operatorname{diam}(\mathcal{D})^2 + \Delta_k, \tag{4.2.2}$$

where Δ_k is introduced in Alg. 4.1.

Frank-Wolfe Sinkhorn barycenters. We show that the barycenter problem (4.2.1) satisfies the setting and hypotheses of Thm. 4.1 and can be thus approached via Alg. 4.1.

Optimization domain. Let $\mathcal{W} = \mathcal{C}(\mathcal{X})$, with dual $\mathcal{W}^* = \mathcal{M}(\mathcal{X})$. The constraint set

Algorithm 4.1 FRANK-WOLFE IN DUAL BANACH SPACES

Input: initial $w_0 \in \mathcal{D}$, threshold $(\Delta_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_{++}$, such that $\Delta_k(k+2)$ is nondecreasing.

For $k = 0, 1, \dots$
Take z_{k+1} such that $\mathsf{G}'(w_k, z_{k+1} - w_k) \leq \min_{z\in\mathcal{D}} \mathsf{G}'(w_k, z - w_k) + \frac{\Delta_k}{2}$
$w_{k+1} = w_k + \frac{2}{k+2}(z_{k+1} - w_k)$

$\mathcal{D} = \mathcal{P}(\mathcal{X})$ is convex, closed, and bounded.

Objective functional. The objective functional $\mathsf{G} = \mathsf{B}_\varepsilon \colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, defined in (4.2.1), is convex since it is a convex combination of $\mathsf{S}_\varepsilon(\cdot, \beta_j)$, with $j = 1 \dots m$. The gradient $\nabla\mathsf{B}_\varepsilon \colon \mathcal{P}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is $\nabla\mathsf{B}_\varepsilon = \sum_{j=1}^m \omega_j \nabla\mathsf{S}_\varepsilon(\cdot, \beta_j)$, where $\nabla\mathsf{S}_\varepsilon(\cdot, \beta_j)$ is given in Remark 4.1.

Lipschitz continuity of the gradient. This is the most critical condition and is addressed in the following section.

4.3 Lipschitz continuity of the gradient of Sinkhorn divergence with respect to the Total Variation

In this section we show that the gradient of the Sinkhorn divergence is Lipschitz continuous with respect to the Total Variation on $\mathcal{P}(\mathcal{X})$, denoted by TV. The definition of Total Variation is provided in the brief discussion on f-divergences in Appendix A.2.1. The goal is to prove the following theorem:

Theorem 4.2. *The gradient $\nabla\mathsf{OT}_\varepsilon$ recalled in Remark 4.1 is Lipschitz continuous. In particular, the first component $\nabla_1\mathsf{OT}_\varepsilon$ is $2\varepsilon e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous, i.e., for every $\alpha, \alpha', \beta, \beta' \in \mathcal{P}(\mathcal{X})$,*

$$\|u - u'\|_\infty = \|\nabla_1\mathsf{OT}_\varepsilon(\alpha, \beta) - \nabla_1\mathsf{OT}_\varepsilon(\alpha', \beta')\|_\infty \leq 2\varepsilon e^{3\mathsf{D}/\varepsilon} \left(\|\alpha - \alpha'\|_{\mathsf{TV}} + \|\beta - \beta'\|_{\mathsf{TV}}\right), \tag{4.3.1}$$

where $\mathsf{D} = \sup_{x,y\in\mathcal{X}} \mathsf{c}(x, y)$, $u = \mathsf{P}_{\beta\alpha}(u), u' = \mathsf{P}_{\beta',\alpha'}(u')$, and $u(x_o) = u'(x_o) = 0$. Moreover, it follows from (4.1.7) that $\nabla\mathsf{S}_\varepsilon(\cdot, \beta)$ is $6\varepsilon e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous. The same holds for $\nabla\mathsf{B}_\varepsilon$.

Thm. 4.2 is one of the main contributions of this chapter. It can be rephrased by saying that the operator that maps a pair of distributions to their Sinkhorn potentials is Lipschitz

continuous. This result is significantly deeper than the one given in (Dvurechenskii et al., 2018, Lemma 1), which establishes the Lipschitz continuity of the gradient in the *semidiscrete* case. The proof relies on non-trivial tools from Perron-Frobenius theory for Hilbert's metric (Lemmens and Nussbaum, 2012), which is a well-established framework to study Sinkhorn potentials presented in Appendix A.4. This theorem provides a further understanding of the impact of the entropy regularization in Optimal Transport problems. Namely, while the benefits in terms of statistical properties and computational cost are well known, this theorem concerns the benefits of adding the entropy in terms of smoothness property of the gradients. While in this Chapter we use the theorem specifically to apply Frank-Wolfe algorithm to the barycenter problem, its applicability is not limited to this case. We provide the proof below. Most of the notions needed in the proof are presented in the introductory material in Chapter 2 and the most technical ones in the supplementary chapter Appendix A.

The proof needs some preliminary lemmas. Here we will provide the statements only and the minimum material that is necessary for the overall structure of the proof. We mostly use properties of the Hilbert metric which are presented in Appendix A.4.1 at length. Here we give the following definition of Hilbert metric:

Definition 4.1. *Set $\mathcal{C}_{++}(\mathcal{X}) := \{f \in \mathcal{C}(\mathcal{X}) \text{ such that } f > 0\}$, the set of strictly positive continuous functions. Let f, f' be two functions in $\mathcal{C}_{++}(\mathcal{X})$. The Hilbert metric d_H is defined as follows*

$$d_H(f, f') = \log \max_{x,y \in \mathcal{X}} \frac{f(x)f'(y)}{f(y)f'(x)}. \tag{4.3.2}$$

Note that normally (4.3.2) is a *characterization* of the Hilbert metric and not the definition. This is discussed in Appendix A.4.1. Here we state it directly as definition for convenience.

We start by characterizing the relation between Hilbert's metric between functions of the form $f = e^{u/\varepsilon}$ and the $\|\cdot\|_\infty$ norm between functions of the form $u = \varepsilon \log f$.

Lemma 4.3. *Let $f, f' \in \mathcal{C}_{++}(\mathcal{X})$ and set $u = \varepsilon \log f$ and $u' = \varepsilon \log f'$. Then*

$$d_H(f, f') \leq 2 \left\| \log f - \log f' \right\|_\infty \quad \textit{or, equivalently} \quad d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}) \leq \frac{2}{\varepsilon} \left\| u - u' \right\|_\infty . \tag{4.3.3}$$

Moreover, let $x_o \in \mathcal{X}$, consider the sets $\mathcal{A} = \{h \in \mathcal{C}_{++}(\mathcal{X}) \mid h(x_o) = 1\}$ and $\mathcal{B} = \{w \in \mathcal{C}(\mathcal{X}) \mid w(x_o) = 0\}$. Suppose that $f, f' \in \mathcal{A}$ (or equivalently that $u, u' \in \mathcal{B}$). Then

$$\frac{1}{2} d_H(f, f') \leq \left\| \log f - \log f' \right\|_\infty \leq d_H(f, f') \tag{4.3.4}$$

and

$$\frac{\varepsilon}{2} d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}) \leq \left\| u - u' \right\|_\infty \leq \varepsilon \, d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}). \tag{4.3.5}$$

The proof can be found in Appendix A.4.3. To proceed we recall the following two operators (introduced in Appendix A.4.1):

- $\mathsf{L}_\alpha : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ defined by

$$\mathsf{L}_\alpha(f) : y \longmapsto \int_{\mathcal{X}} e^{-\frac{\mathsf{c}(x,y)}{\varepsilon}} f(x) \, d\alpha(x) \tag{4.3.6}$$

- $\mathsf{A}_\alpha : \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ defined by

$$\mathsf{A}_\alpha(f) = 1/(\mathsf{L}_\alpha f). \tag{4.3.7}$$

In the following we will use the notation k to denote $\mathsf{k}(x,y) := e^{-\frac{\mathsf{c}(x,y)}{\varepsilon}}$. The two maps L and A are closely related to the Sinkhorn map P_α that we used to state the first order conditions of the dual problem, recalled in (4.1.2). Indeed we have

$$\mathsf{P}_\alpha(u) = -\varepsilon \log(\mathsf{A}_\alpha(e^{u/\varepsilon})). \tag{4.3.8}$$

Similarly to (4.1.3), we can express optimality conditions of the dual problem (4.1.1) using the map A_α and A_β: setting $f = e^{\frac{u}{\varepsilon}}$ and $g = e^{\frac{v}{\varepsilon}}$, then (4.1.3) is equivalent to

$$f = \mathsf{A}_\beta(g) \qquad \text{and} \qquad g = \mathsf{A}_\alpha(f), \tag{4.3.9}$$

and analogously, by setting $\mathsf{A}_{\beta\alpha} = \mathsf{A}_\beta \circ \mathsf{A}_\alpha$ and $\mathsf{A}_{\alpha\beta} = \mathsf{A}_\alpha \circ \mathsf{A}_\beta$, is equivalent to

$$f = \mathsf{A}_{\beta\alpha}(f) \qquad \text{and} \qquad g = \mathsf{A}_{\alpha\beta}(g). \tag{4.3.10}$$

As last tool before the proof, we need a property of contraction of Hilbert metric for the operator $\mathsf{A}_{\alpha\beta}$. This property with its proof is contained in Thm. A.7 in Appendix A.4.2 and the statement is recalled here for convenience:

Theorem 4.4 (Hilbert's metric contraction for $\mathsf{A}_{\beta\alpha}$)**.** *The map $\mathsf{A}_{\beta\alpha} : \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ has a unique fixed point up to positive scalar multiples. Moreover, let $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$. Then, for*

every $f, f' \in \mathcal{C}_{++}(\mathcal{X})$,

$$d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta\alpha}(f')) \leq \lambda^2 \, d_H(f, f'). \tag{4.3.11}$$

We are ready to prove the main result of the section.

Theorem 4.5 (Lipschitz continuity of the Sinkhorn potentials with respect to the total variation)**.** *Let* $\alpha, \beta, \alpha', \beta' \in \mathcal{P}(\mathcal{X})$ *and let* $x_o \in \mathcal{X}$*. Let* $(u, v), (u', v') \in \mathcal{C}(\mathcal{X})^2$ *be the two pairs of Sinkhorn potentials corresponding to the solution of the regularized OT problem in* (A.4.25) *for* (α, β) *and* (α', β') *respectively such that* $u(x_o) = u'(x_o) = 0$*. Then*

$$\left\|u - u'\right\|_\infty \leq 2\varepsilon e^{3\mathsf{D}/\varepsilon} \|(\alpha - \alpha', \beta - \beta')\|_{\mathsf{TV}}. \tag{4.3.12}$$

Hence, the map that to each pair of probability distributions $(\alpha, \beta) \in \mathcal{P}(\mathcal{X})^2$ *associates the component* u *of the corresponding Sinkhorn potentials is* $2\varepsilon e^{3\mathsf{D}/\varepsilon}$*-Lipschitz continuous* with respect to the total variation.

Proof. The functions $f = e^{u/\varepsilon}$ and $f' = e^{u'/\varepsilon}$ are fixed points of the maps $\mathsf{A}_{\beta\alpha}$ and $\mathsf{A}_{\beta'\alpha'}$ respectively. Then, it follows from Thm. 4.4 that

$$\begin{aligned} d_H(f, f') &= d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f')) \\ &\leq d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) + d_H(\mathsf{A}_{\beta'\alpha'}(f), \mathsf{A}_{\beta'\alpha'}(f')) \\ &\leq d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) + \lambda^2 d_H(f, f'), \end{aligned}$$

hence,

$$d_H(f, f') \leq \frac{1}{1 - \lambda^2} \, d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)). \tag{4.3.13}$$

Moreover, using (4.3.3), we have

$$\begin{aligned} d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) &\leq d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha}(f)) + d_H(\mathsf{A}_{\beta'\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) \\ &\leq d_H(\mathsf{A}_{\beta}(g), \mathsf{A}_{\beta'}(g)) + \lambda d_H(\mathsf{A}_{\alpha}(f), \mathsf{A}_{\alpha'}(f)) \\ &\leq 2 \left\| \log \frac{\mathsf{A}_\beta(g)}{\mathsf{A}_{\beta'}(g)} \right\|_\infty + 2\lambda \left\| \log \frac{\mathsf{A}_\alpha(f)}{\mathsf{A}_{\alpha'}(f)} \right\|_\infty. \end{aligned} \tag{4.3.14}$$

Now, note that by Lemma A.9

$$\left|\log \frac{\mathsf{A}_\beta(g)}{\mathsf{A}_{\beta'}(g)}\right| = \left|\log \frac{\mathsf{L}_{\beta'} g}{\mathsf{L}_\beta g}\right| \leq \max\{1/\mathsf{L}_\beta g, 1/\mathsf{L}_{\beta'} g\}|(\mathsf{L}_{\beta'} - \mathsf{L}_\beta)g| \tag{4.3.15}$$

and that, for every $x \in \mathcal{X}$,

$$\begin{aligned}[(\mathsf{L}_{\beta'} - \mathsf{L}_\beta)g](x) &= \int \mathsf{k}(x,z)g(z)\, d(\beta - \beta')(z) \\ &= \left\langle \mathsf{k}(x,\cdot)g, \beta - \beta' \right\rangle \leq \|g\|_\infty \|\beta - \beta'\|_{\mathsf{TV}},\end{aligned} \tag{4.3.16}$$

where k denotes the function $\mathsf{k}(x,z) = e^{-\frac{\mathsf{c}(x,z)}{\varepsilon}}$. Similarly, $[(\mathsf{L}_\beta - \mathsf{L}_{\beta'})g](x) \leq \|g\|_\infty \|\beta - \beta'\|_{\mathsf{TV}}$. Therefore, since $1/(\mathsf{L}_\beta g) = \mathsf{A}_\beta(g) = f$ and $\mathsf{L}_{\beta'} g \geq e^{-\mathsf{D}/\varepsilon} \min g$, it follows from Lemma A.8Item (v) and (A.4.34) (applied to g) that

$$\left\|\log \frac{\mathsf{A}_\beta(g)}{\mathsf{A}_{\beta'}(g)}\right\|_\infty \leq \max\left\{\|f\|_\infty, \frac{e^{\mathsf{D}/\varepsilon}}{\min g}\right\} \|g\|_\infty \|\beta - \beta'\|_{\mathsf{TV}} \leq e^{2\mathsf{D}/\varepsilon} \|\beta - \beta'\|_{\mathsf{TV}}. \tag{4.3.17}$$

Analogously, it holds

$$\left\|\log \frac{\mathsf{A}_\alpha(f)}{\mathsf{A}_{\alpha'}(f)}\right\|_\infty \leq e^{2\mathsf{D}/\varepsilon} \|\alpha - \alpha'\|_{\mathsf{TV}}. \tag{4.3.18}$$

Putting (4.3.13), (4.3.14), (4.3.17), and (4.3.18) together, we have

$$d_H(f, f') \leq \frac{2e^{2\mathsf{D}/\varepsilon}}{1 - \lambda^2} \left(\lambda \|\alpha - \alpha'\|_{\mathsf{TV}} + \|\beta - \beta'\|_{\mathsf{TV}}\right). \tag{4.3.19}$$

Now, note that since $e^{\mathsf{D}/\varepsilon} \geq 1$

$$\frac{1}{1 - \lambda^2} = \frac{(e^{\mathsf{D}/\varepsilon} + 1)^2}{4e^{\mathsf{D}/\varepsilon}} \leq e^{\mathsf{D}/\varepsilon}. \tag{4.3.20}$$

Finally, recalling (4.3.5), we have

$$\left\|u - u'\right\|_\infty \leq 2\varepsilon e^{3\mathsf{D}/\varepsilon} \|(\alpha - \alpha', \beta - \beta')\|_{\mathsf{TV}}, \tag{4.3.21}$$

where $\|(\alpha - \alpha', \beta - \beta')\|_{\mathsf{TV}} = \|\alpha - \alpha'\|_{\mathsf{TV}} + \|\beta - \beta'\|_{\mathsf{TV}}$ is the total variation norm on $\mathcal{M}(\mathcal{X})^2$. □

We are now ready to prove the theorem Thm. 4.2, stated at the beginning of the section,

which is now a easy consequence of the results above:

Proof. The first part is just a consequence of Thm. 4.5 and (4.1.6). The second part follows from the first part and Remark 4.1. □

4.4 Lipschitz continuity with respect to the MMD and sample complexity of Sinkhorn gradients

In the previous section, we proved Lipschitz continuity of the gradient of Sinkhorn divergence with respect to Total Variation. Total Variation is a strong metric on the space of measures which makes $\mathcal{M}(\mathcal{X})$ a Banach space. This will be key in showing that Frank-Wolfe algorithm can be applied to the barycenter problem and inherits the guarantees in the optimization process. However, when we deal with continuous measures another ingredient comes into play: in practise we access the absolutely continuous measures via samples. Therefore this statistical aspects plays a role in the kind of theoretical guarantees that we can achieve. In order to take the sampling procedure into account, we have to quantify the approximation resulting by using a gradient computed on a sample rather than the actual gradient. In statistical terms, we need a result on sample complexity of the gradients of Sinkhorn. This section is devoted to this and the main tool is -again- a Lipschitz continuity result. However, differently from the previous section where we proved a Lipschitz continuity result with respect to a *strong* metric, here we need a Lipschitz continuity result with respect to a *weak* metric, as specified in the statements below.

The main result that we prove in this section is the following, that *quantifies* the approximation error between $\nabla_1 \mathsf{OT}_\varepsilon(\cdot, \beta)$ and $\nabla_1 \mathsf{OT}_\varepsilon(\cdot, \hat{\beta})$ in terms of the sample size of $\hat{\beta}$.

Theorem 4.6 (Sample Complexity of Sinkhorn Potentials)**.** *Suppose that* $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$ *with* $s > d/2$*. Then, there exists a constant* $\bar{\mathsf{r}} = \bar{\mathsf{r}}(\mathcal{X}, \mathsf{c}, d)$ *such that for any* $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ *and any empirical measure* $\hat{\beta}$ *of a set of* n *points independently sampled from* β*, we have, for every* $\tau \in (0, 1]$

$$\|u - u_n\|_\infty = \|\nabla_1 \mathsf{OT}_\varepsilon(\alpha, \beta) - \nabla_1 \mathsf{OT}_\varepsilon(\alpha, \hat{\beta})\|_\infty \leq \frac{8\varepsilon\, \bar{\mathsf{r}} e^{3\mathsf{D}/\varepsilon} \log \frac{3}{\tau}}{\sqrt{n}} \tag{4.4.1}$$

with probability at least $1 - \tau$*, where* $u = \mathsf{P}_{\beta\alpha}(u), u_n = \mathsf{P}_{\hat{\beta}\alpha}(u_n)$ *and* $u(x_o) = u_n(x_o) = 0$.

We point out that it *cannot* be obtained by means of the Lipschitz continuity of $\nabla_1 \mathsf{OT}_\varepsilon$ in Thm. 4.2, since empirical measures do not converge in $\|\cdot\|_{\mathsf{TV}}$ to their target distribution

(Devroye et al., 1990). Instead, the proof consists in considering the weaker *Maximum Mean Discrepancy (*MMD*, recalled in Def. A.14)* metric associated to a universal kernel (Song, 2008), which metrizes the topology of the convergence in law of $\mathcal{P}(\mathcal{X})$ (Sriperumbudur et al., 2011). Empirical measures converge in MMD metric to their target distribution (Song, 2008). Therefore, in order to show that (4.4.1) holds, it is sufficient to prove the Lipschitz continuity of $\nabla_1 \mathsf{OT}_\varepsilon$ with *respect to* MMD. We show this result below.

In general, recall that a well-established approach to approximate a distribution $\beta \in \mathcal{P}(\mathcal{X})$ is to independently sample a set of points $x_1, \dots, x_n \in \mathcal{X}$ from β and consider the empirical distribution $\beta_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. As an intermediate step, the following lemma shows that β_n converges to β in MMD with high probability. The original version of this result can be found in Song (2008), we report an independent proof in Appendix C.4. for completeness.

Lemma 4.7. *Let $\beta \in \mathcal{P}(\mathcal{X})$. Let $x_1, \dots, x_n \in \mathcal{X}$ be independently sampled according to β and denote by $\beta_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Then, for any $\tau \in (0, 1]$, we have*

$$\mathsf{MMD}(\beta_n, \beta) \leq \frac{4 \log \frac{3}{\tau}}{\sqrt{n}} \tag{4.4.2}$$

with probability at least $1 - \tau$.

We now proceed to the main result on Lipschitz continuity of the potentials with respect to MMD.

Proposition 4.8 (Lipschitz continuity of the Sinkhorn Potentials with respect to the MMD)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact Lipschitz domain and $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$, with $s > d/2$. Let $\alpha, \beta, \alpha', \beta' \in \mathcal{P}(\mathcal{X})$. Let $x_o \in \mathcal{X}$ and let $(u, v), (u', v') \in \mathcal{C}(\mathcal{X})^2$ be the two Sinkhorn potentials corresponding to the solution of the regularized OT problem in (4.1.1) for (α, β) and (α', β') respectively such that $u(x_o) = u'(x_o) = 0$. Then*

$$\left\|u - u'\right\|_\infty \leq 2\varepsilon \bar{\mathsf{r}} e^{3\mathsf{D}/\varepsilon} \left(\mathsf{MMD}(\alpha, \alpha') + \mathsf{MMD}(\beta, \beta')\right), \tag{4.4.3}$$

with $\bar{\mathsf{r}}$ from Lemma C.10. In other words, the operator $\nabla_1 \mathsf{OT}_\varepsilon \colon \mathcal{P}(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})$ is $2\varepsilon\bar{\mathsf{r}}e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous with respect to the MMD.

Proof. Let $f = e^{u/\varepsilon}$ and $g = e^{v/\varepsilon}$. A quite technical result, which is postponed to Lemma C.10 in the appendix, shows a uniform bound on the norm of $\mathsf{k}(x, \cdot)e^{u/\varepsilon}$ and $\mathsf{k}(x, \cdot)e^{v/\varepsilon}$ in the Sobolev space $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$. Using those, we can now refine the analysis

in Thm. 4.5. More precisely, we observe that in (4.3.16) we have

$$\begin{aligned}
[(\mathsf{L}_{\beta'} - \mathsf{L}_{\beta})g](x) &= \int \mathsf{k}(x,z)g(z)\, d(\beta - \beta')(z) \\
&= \int \langle \mathsf{k}(x,\cdot)g, \mathsf{h}(z,\cdot)\rangle_{\mathcal{H}}\, d(\beta - \beta')(z) \\
&= \left\langle \mathsf{k}(x,\cdot)g, \mathsf{h}_{\beta} - \mathsf{h}_{\beta'} \right\rangle_{\mathcal{H}} \\
&\leq \|\mathsf{k}(x,\cdot)g\|_{\mathcal{H}}\, \|\mathsf{h}_{\beta} - \mathsf{h}_{\beta'}\|_{\mathcal{H}} \\
&\leq \bar{\mathsf{r}}\, \mathsf{MMD}(\beta, \beta'),
\end{aligned}$$

where h_{β} ($\mathsf{h}_{\beta'}$) is the kernel mean embedding of β (β'), where in the first equality, with some abuse of notation, we have implicitly considered the extension of $\mathsf{k}(x,\cdot)g$ to $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$ as discussed in Lemma C.10. The rest of the analysis in Thm. 4.5 remains unvaried, eventually leading to (4.4.3). □

We now have all the tools to prove Thm. 4.6, which easily follows from the results discussed so far.

Proof. The theorem is just a consequence of Lemma 4.7 and Prop. 4.8. □

Remark 4.2. *This latter result relies on higher regularity properties of Sinkhorn potentials, which have been recently shown (Genevay et al., 2018a, Thm.2) to be uniformly bounded in Sobolev spaces under the additional assumption* $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$. *For sufficiently large* s, *the Sobolev norm is in duality with the* MMD *(Muandet et al., 2017) and allows us to derive the required Lipschitz continuity. We conclude noting that while Genevay et al. (2018a) studied the sample complexity of the Sinkhorn* divergence, *Thm. 4.6 is a sample complexity result for Sinkhorn* potentials. *In this sense, we observe that the constants appearing in the bound are tightly related to those in (Genevay et al., 2018a, Thm.3) and have similar behavior with respect to* ε. *However, in the subsequent work (Mena and Niles-Weed, 2019), the exponential dependence on* ε^{-1} *in the sample complexity result has been removed, in favour of polynomial dependence. This change cannot be done here with the present proof, which deeply exploits the structure of DAD problems. It would be interesting to see whether with a completely different approach it is possible to improve the constants in our results as well. In a very recent paper (Shen et al., 2020) our result is generalized to cost functions with weaker assumptions. However, the dependence on the regularization parameter remains the same and there is no improvement on that regard.*

4.5 Algorithm: practical Sinkhorn barycenters

According to Sec. 4.2, FW is a valid approach to tackle the barycenter problem (4.2.1). Here we describe how to implement in practice the abstract procedure of Alg. 4.1 to obtain a sequence of distributions $(\alpha_k)_{k\in\mathbb{N}}$ minimizing B_ε. A main challenge in this sense resides in finding a minimizing feasible direction for $\mathsf{B}'_\varepsilon(\alpha_k; \mu-\alpha_k) = \langle \nabla \mathsf{B}_\varepsilon(\alpha_k), \mu - \alpha_k\rangle$. According to Remark 4.1, this amounts to solve

$$\mu_{k+1} \in \operatorname*{argmin}_{\mu\in\mathcal{P}(\mathcal{X})} \sum_{j=1}^m \omega_j \ \langle u_{jk} - p_k, \mu\rangle \qquad \text{where} \qquad u_{jk} - p_k = \nabla \mathsf{S}_\varepsilon[(\cdot,\beta_j)](\alpha_k), \tag{4.5.1}$$

with $p_k = \nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \alpha_k)$ not depending on j. In general (4.5.1) would entail a minimization over the set of all probability distributions on $\mathcal{X}$. However, since the objective functional is linear in μ and $\mathcal{P}(\mathcal{X})$ is a weakly-$*$ compact convex set, we can apply Bauer maximum principle (see e.g., (Aliprantis, 2006, Thm. 7.69)). Hence, solutions are achieved at the extreme points of the optimization domain: in the case of $\mathcal{P}(\mathcal{X})$, the extreme points correspond to Dirac's deltas (Chouquet, 1969, p. 108). Now, denote by $\delta_x \in \mathcal{P}(\mathcal{X})$ the Dirac's delta centered at $x \in \mathcal{X}$. We have $\langle w, \delta_x\rangle = w(x)$ for every $w \in \mathcal{C}(\mathcal{X})$. Hence (4.5.1) is equivalent to

$$\mu_{k+1} = \delta_{x_{k+1}} \qquad \text{with} \qquad x_{k+1} \in \operatorname*{argmin}_{x\in\mathcal{X}} \sum_{j=1}^m \omega_j \ \big(u_{jk}(x) - p_k(x)\big). \tag{4.5.2}$$

Once the new support point x_{k+1} has been obtained, the update in Alg. 4.1 corresponds to

$$\alpha_{k+1} = \alpha_k + \frac{2}{k+2}(\delta_{x_{k+1}} - \alpha_k) = \frac{k}{k+2}\alpha_k + \frac{2}{k+2}\delta_{x_{k+1}}. \tag{4.5.3}$$

In particular, if FW is initialized with a distribition with finite support, say $\alpha_0 = \delta_{x_0}$ for some $x_0 \in \mathcal{X}$, then also every further iterate α_k will have at most $k+1$ support points. According to (4.5.2), the inner optimization for FW consists in minimizing the functional $x \mapsto \sum_{j=1}^m \omega_j \ \big(u_{jk}(x) - p_k(x)\big)$ over $\mathcal{X}$. In practice, having access to such functional poses already a challenge, since it requires computing the Sinkhorn potentials u_{jk} and p_k, which are continuous functions on the domain $\mathcal{X}$. Below we discuss how to estimate these potentials when the β_j have finite support. We then address the general setting.

Algorithm 4.2 SINKHORN BARYCENTERS

Input: $\mathbf{Y}_j \in \mathbb{R}^{d\times m_j}, \mathsf{b}_j \in \mathbb{R}^{m_k}$ for $j = 1,\dots,J$, initial point $\mathbf{X}_0 = x_0 \in \mathbb{R}^d$, $\varepsilon > 0$.

Initialize: $\beta_j = (\mathbf{Y}_j, \mathsf{b}_j)$ and $\alpha_0 = (\mathbf{X}_0, \mathsf{a}_0)$ probability distributions with $\mathsf{a}_0 = 1$.

For $k = 0, 1, \dots, K-1$

$\mathsf{p} = \text{SINKHORNKNOPP}(\alpha_k, \alpha_k, \varepsilon)$

For $j = 1, \dots J$

$\_, \mathsf{v}_j = \text{SINKHORNKNOPP}(\alpha_k, \beta_j, \varepsilon)$

Let $\varphi_k : x \mapsto -\sum_{q=1}^{Q} \log \sum_{j=1}^{m_q} e^{(\mathsf{v}_{qj} - \mathsf{c}(x, \mathbf{Y}_q^j))/\varepsilon}\, \mathsf{b}_{qj} + \log \sum_{i=0}^{k} e^{(\mathsf{p}_i - \mathsf{c}(x, \mathbf{X}^i))/\varepsilon} \mathsf{a}_i$

$x_{k+1} = \text{MINIMIZE}(\varphi_k)$

$\mathbf{X}_{k+1} = [\mathbf{X}_k, x_{k+1}]$ and $\mathsf{a}_{k+1} = \frac{1}{k+1}[k\ \mathsf{a}_k, 1]$

$\alpha_{k+1} = (\mathbf{X}_{k+1}, \mathsf{a}_{k+1})$

Return: α_K

Computing $\nabla_1 \mathsf{OT}_\varepsilon$ for probability distributions with finite support. Let $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, with $\beta = \sum_{i=1}^n \mathsf{b}_i \delta_{y_i}$ a probability measure with finite support, with $\mathsf{b} = (\mathsf{b}_i)_{i=1}^n$ nonnegative weights summing up to 1. It is useful to identify β with the pair $(\mathbf{Y}, \mathsf{b})$, where $\mathbf{Y} \in \mathbb{R}^{d\times n}$ is the matrix with i-th column equal to y_i. Let now $(u, v) \in \mathcal{C}(\mathcal{X})^2$ be the pair of Sinkhorn potentials associated to α and β in Prop. 2.10, recall that $u = \mathsf{P}_\beta(v)$. Denote by $\mathsf{v} \in \mathbb{R}^n$ the *evaluation vector* of the Sinkhorn potential v, with i-th entry $\mathsf{v}_i = v(y_i)$. According to the definition of P_β in (4.1.2), for any $x \in \mathcal{X}$

$$[\nabla_1 \mathsf{OT}_\varepsilon(\alpha, \beta)](x) = u(x) = [\mathsf{P}_\beta(v)](x) = -\varepsilon \log \sum_{i=1}^{n} e^{(\mathsf{v}_i - \mathsf{c}(x, y_i))/\varepsilon}\, \mathsf{b}_i, \tag{4.5.4}$$

since the integral $\mathsf{P}_\beta(v)$ reduces to a sum over the support of β. Hence, the gradient of OT_ε (i.e. the potential u), *is uniquely characterized in terms of the finite dimensional vector* v *collecting the values of the potential* v *on the support of* β . We refer as SINKHORNGRADIENT to the routine which associates to each triplet $(\mathbf{Y}, \mathbf{b}, \mathbf{v})$ the map $x \mapsto -\varepsilon \log \sum_{i=1}^n e^{(\mathsf{v}_i - \mathsf{c}(x, y_i))/\varepsilon}\, \mathsf{b}_i$.

Sinkhorn barycenters: finite case. Alg. 4.2 summarizes FW applied to the barycenter problem (4.2.1) when the β_j's have finite support. Starting from a Dirac's delta $\alpha_0 = \delta_{x_0}$, at each iteration $k \in \mathbb{N}$ the algorithm proceeds by: i) finding the corresponding evaluation vectors v_j's and p of the Sinkhorn potentials for $\mathsf{OT}_\varepsilon(\alpha_k, \beta_j)$ and $\mathsf{OT}_\varepsilon(\alpha_k, \alpha_k)$ respectively, via the routine SINKHORNKNOPP (see (Cuturi, 2013; Feydy et al., 2019) or Alg. 2.1).

This is possible since both β_j and α_k have finite support and therefore the problem of approximating the evaluation vectors v_j and p reduces to an optimization problem over finite vector spaces that can be efficiently solved (Cuturi, 2013); $ii)$ obtain the gradients $u_j = \nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \beta_j)$ and $p = \nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \alpha_k)$ via SINKHORNGRADIENT; $iii)$ minimize $\varphi : x \mapsto \sum_{j=1}^n \omega_j \, u_j(x) - p(x)$ over $\mathcal{X}$ to find a new point x_{k+1} (we comment on this meta-routine Minimize below); $iv)$ finally update the support and weights of α_k according to (4.5.3) to obtain the new iterate α_{k+1}.

A key feature of Alg. 4.2 is that the support of the candidate barycenter is updated *incrementally* by adding at most one point at each iteration, a procedure similar in flavor to the kernel herding strategy in Bach et al. (2012) and Lacoste-Julien et al. (2015). This contrasts with previous methods for barycenter estimation (Cuturi and Doucet, 2014; Benamou et al., 2015; Staib et al., 2017; Dvurechenskii et al., 2018), which require the support set, or at least its cardinality, to be fixed beforehand. However, identifying the new support point requires solving the nonconvex problem (4.5.2), a task addressed by the meta-routine MINIMIZE. This problem is typically smooth (e.g., a linear combination of Gaussians when $\mathsf{c}(x, y) = \|x - y\|^2$) and first or second order nonlinear optimization methods can be adopted to find stationary points. We note that all free-support methods in the literature for barycenter estimation are also affected by nonconvexity since they typically require solving a bi-convex problem (alternating minimization between support points and weights) which is not jointly convex (Cuturi and Doucet, 2014; Claici et al., 2018). We conclude by observing that if we restrict to the setting of (Staib et al., 2017; Dvurechenskii et al., 2018) with fixed support set, then MINIMIZE can be solved exactly by evaluating the functional in (4.5.2) on each candidate support point.

Sinkhorn barycenters: general case. When the β_j's have infinite support, it is not possible to apply Sinkhorn-Knopp in practice. In line with (Genevay et al., 2018a; Staib et al., 2017), we can randomly sample empirical distributions $\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^n \delta_{x_{ij}}$ from each β_j and apply Sinkhorn-Knopp to $(\alpha_k, \hat{\beta}_j)$ in Alg. 4.1 rather than to the ideal pair (α_k, β_j). This strategy is motivated by Prop. 2.8, where it was shown that Sinkhorn potentials vary continuously with the input measures. However, it opens two questions: $i)$ whether this approach is theoretically justified (consistency) and $ii)$ how many points should we sample from each β_j to ensure convergence (rates). We answer these questions in Thm. 4.10 in the

next section.

4.6 Convergence analysis

We finally address the convergence of FW applied to both the finite and infinite settings discussed in Sec. 4.5. We begin by considering the finite setting.

Theorem 4.9. *Suppose that $\beta_1, \dots \beta_m \in \mathcal{P}(\mathcal{X})$ have finite support and let α_k be the k-th iterate of Alg. 4.2 applied to (4.2.1). Then,*

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{48\, \varepsilon\, e^{3\mathsf{D}/\varepsilon}}{k+2}. \tag{4.6.1}$$

The result follows by the convergence result of FW in Thm. 4.1 applied with the Lipschitz constant computed in Thm. 4.2, and recalling that $\operatorname{diam}(\mathcal{P}(\mathcal{X})) = 2$ with respect to the Total Variation. We note that Thm. 4.9 assumes SINKHORNKNOPP and MINIMIZE in Alg. 4.2 to yield exact solutions. In Appendix C.3 we comment how approximation errors in this context affect the bound in (4.6.1).

We can now study the convergence of FW in continuous settings.

Theorem 4.10. *Suppose that $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$ with $s > d/2$. Let $n \in \mathbb{N}$ and $\hat{\beta}_1, \dots, \hat{\beta}_m$ be empirical distributions with n support points, each independently sampled from $\beta_1, \dots, \beta_m$. Let α_k be the k-th iterate of Alg. 4.2 applied to $\hat{\beta}_1, \dots, \hat{\beta}_m$. Then for any $\tau \in (0, 1]$, the following holds with probability larger than $1 - \tau$*

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{64 \bar{\mathsf{r}} \varepsilon e^{3\mathsf{D}/\varepsilon} \log \frac{3}{\tau}}{\min(k, \sqrt{n})}. \tag{4.6.2}$$

Proof. Let $\widehat{\mathsf{B}_\varepsilon}(\alpha) = \sum_{j=1}^m \omega_j \mathsf{S}_\varepsilon(\alpha, \hat{\beta}_j)$. Then, it follows from the definition of B_ε and Thm. 4.6 that, for every $k \in \mathbb{N}$, and with probability larger than $1 - \tau$, we have

$$\begin{aligned}
\|\nabla \widehat{\mathsf{B}_\varepsilon}(\alpha_k) - \nabla \mathsf{B}_\varepsilon(\alpha_k)\|_\infty &\leq \sum_{j=1}^m \omega_j \|\nabla[\mathsf{S}_\varepsilon(\cdot, \hat{\beta}_j)](\alpha_k) - \mathsf{S}_\varepsilon(\cdot, \beta_j)](\alpha_k)\|_\infty \\
&= \sum_{j=1}^m \omega_j \|\nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \hat{\beta}_j) - \nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \beta_j)\|_\infty \\
&\leq \frac{8\varepsilon\, \bar{\mathsf{r}} e^{3\mathsf{D}/\varepsilon} \log \frac{3}{\tau}}{\sqrt{n}} \\
&= \frac{\Delta_1}{4},
\end{aligned}$$

where

$$\Delta_1 := \frac{32\varepsilon\,\bar{\mathsf{r}}e^{3\mathsf{D}/\varepsilon}\log\frac{3}{\tau}}{\sqrt{n}}.$$

Now, let $\gamma_k = 2/(k+2)$. Since Alg. 4.2 is applied to $\hat{\beta}_1, \dots \hat{\beta}_m$, we have

$$\delta_{x_{k+1}} \in \operatorname*{argmin}_{\mathcal{P}(\mathcal{X})} \langle \nabla \widehat{\mathsf{B}_\varepsilon}(\alpha_k), \cdot \rangle \quad \text{and} \quad \alpha_{k+1} = (1-\gamma_k)\alpha_k + \gamma_k \delta_{x_{k+1}}.$$

Therefore, it follows from Thm. C.1, Prop. C.3, and Thm. 4.2 that, with probability larger than $1-\tau$, we have

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon \leq 6\varepsilon\bar{\mathsf{r}}e^{3\mathsf{D}/\varepsilon}\mathrm{diam}(\mathcal{P}(\mathcal{X}))^2\gamma_k + \Delta_1 \mathrm{diam}(\mathcal{P}(\mathcal{X})).$$

The statement follows by noting that $\mathrm{diam}(\mathcal{P}(\mathcal{X})) = 2$. □

A consequence of Thm. 4.10 is that the accuracy of FW depends simultaneously on the number of iterations and the sample size used in the approximation of the gradients: by choosing $n = k^2$ we recover the $O(1/k)$ rate of the finite setting, while for $n = k$ we have a rate of $O(k^{-1/2})$, which is reminiscent of typical sample complexity results, highlighting the statistical nature of the problem.

Remark 4.3 (Incremental Sampling)**.** *The above strategy requires sampling the empirical distributions for $\beta_1, \dots, \beta_m$ beforehand. A natural question is whether it would be possible to do this* incrementally*, sampling new points and updating $\hat{\beta}_j$ accordingly, as the number of* FW *iterations increase. To this end, one can perform an intersection bound and see that this strategy is still consistent, but the bound in Thm. 4.10 worsens the logarithmic term, which becomes* $\log(3mk/\tau)$.

4.7 Experiments

In this section we show the performance of our method in a range of experiments [2].

Discrete measures: barycenter of nested ellipses. We compute the barycenter of 30 randomly generated nested ellipses on a 50×50 grid similarly to Cuturi and Doucet (2014). We interpret each image as a probability distribution in 2D. The cost matrix is given by the squared Euclidean distances between pixels. Fig. 4.1 reports 8 samples of the input ellipses

[2] https://github.com/GiulsLu/Sinkhorn-Barycenters

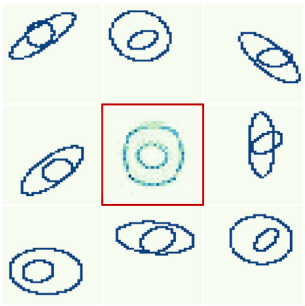


Figure 4.1: We compute the barycenter of 30 pairs of nested ellipses, randomly generated on a 50×50 grid. We use the Alg. 4.2 to compute the barycenter of the 30 input measures. A sample of the inputs is provided in the outer figures, while the barycenter retrieved with Alg. 4.2 is displayed in the center, with a red frame.

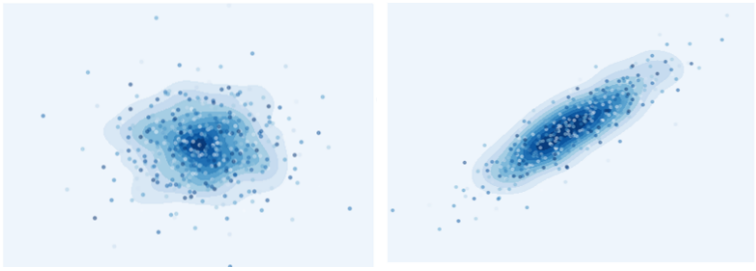


Figure 4.2: Barycenters of Gaussians: Alg. 4.2 is tested in the computation of the barycenter of 5 Gaussian distributions $\mathcal{N}(m_i, C_i)$ $i = 1, \dots, 5$ in $\mathbb{R}^2$, with mean $m_i \in \mathbb{R}^2$ and covariance C_i randomly generated. Scatter plot: output of our method; Density level sets: the true Wasserstein barycenter.

and the barycenter obtained with Alg. 4.2. It shows qualitatively that our approach captures key geometric properties of the input measures.

Continuous measures: barycenter of Gaussians. We compute the barycenter of 5 Gaussian distributions $\mathcal{N}(m_i, C_i)$ $i = 1, \dots, 5$ in $\mathbb{R}^2$, with mean $m_i \in \mathbb{R}^2$ and covariance C_i randomly generated. We apply Alg. 4.2 to empirical measures obtained by sampling $n = 500$ points from each $\mathcal{N}(m_i, C_i)$, $i = 1, \dots, 5$. Since the (Wasserstein) barycenter of Gaussian distributions can be estimated accurately (see (Agueh and Carlier, 2011)), in Fig. 4.2 we report both the output of our method (as a scatter plot) and the true Wasserstein barycenter (as level sets of its density). We observe that the barycenter found by our algorithm recovers both the mean and covariance of the target barycenter. In Appendix C.5 we provide additional experiments also in the case of mixtures of Gaussians.

Image "compression" via distribution matching. Similarly to Claici et al. (2018), we test Alg. 4.2 in the special case of computing the "barycenter" of a single measure

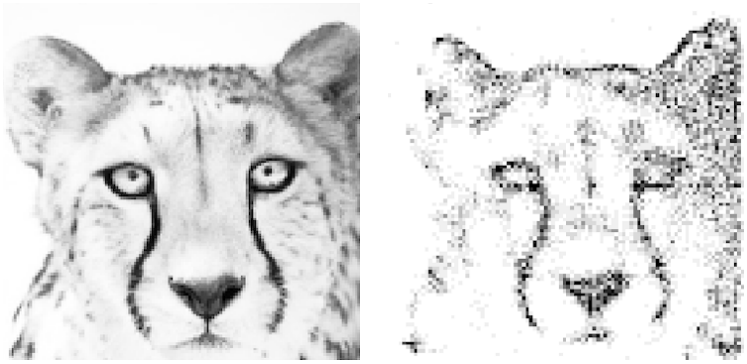


Figure 4.3: Image compression: original image 140x140 pixels (left), sample (right). Alg. 4.2 is used to match a single probability measure supported on 140^2 points. On the right, a sample of the barycenter retrived by the algorithm after around 3900 iterations.

Figure 4.4: k-means clustering experiment: 20 centroids obtained by performing k-mean with Alg. 4.2. The experiment is run on a subset of 500 random images from the MNIST dataset. Each image is suitably normalized to be interpreted as a probability distribution on the grid of 28×28 pixels with values scaled between 0 and 1. The initialization consists of 20 centroids according to the k-means++ strategy (Arthur and Vassilvitskii, 2007).

$\beta \in \mathcal{P}(\mathcal{X})$. While the solution of this problem is the distribution β itself, we can interpret the intermediate iterates α_k of Alg. 4.2 as compressed version of the original measure. In this sense k would represent the level of compression since α_k is supported on *at most* k points. Fig. 4.3 (Right) reports iteration $k = 5000$ of Alg. 4.2 applied to the 140×140 image in Fig. 4.3 (Left) interpreted as a probability measure β in 2D. We note that the number of points in the support is ~ 3900: indeed, Alg. 4.2 selects the most relevant support points multiple times to accumulate the right amount of mass on each of them (darker color = higher weight). This shows that FW tends to greedily search for the most relevant support points, prioritizing those with higher weight, with an echo of quantization in image processing.

k-means on MNIST digits. We tested our algorithm on a k-means clustering experiment.

We consider a subset of 500 random images from the MNIST dataset. Each image is suitably normalized to be interpreted as a probability distribution on the grid of 28×28 pixels with values scaled between 0 and 1. The $n = 500$ random images constitute our set of observations $\{\beta_1, \dots, \beta_{500}\}$ and the goal of k-means clustering is to find a partition of the n observations into $k(\leq n)$ sets $\mathcal{S} = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster 'variance'. In our setting, the objective is

$$\underset{\mathcal{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\beta \in S_i} \mathsf{S}_\varepsilon(\beta, \alpha_i) \tag{4.7.1}$$

where α_i is the barycenter of the points in S_i. The naive procedure to find a solution consists in the following: given an initial set of k means $\alpha_1^{(1)}, \dots \alpha_k^{(1)}$ the algorithm proceeds by alternating between two steps, the assignment step and the update step. The *assignment step* assigns each observation to the cluster with the nearest mean, w.r.t. Sinkhorn divergence:

$$S_i^{(t)} := \{\beta_q : \ \mathsf{S}_\varepsilon(\beta_q, \alpha_i^{(t)}) \leq \mathsf{S}_\varepsilon(\beta_q, \alpha_j^{(t)}) \quad \forall\, j,\, 1 \leq j \leq k\}. \tag{4.7.2}$$

The *update steps* recalculates the means (i.e. centroids) for the observations assigned to each cluster

$$\alpha_i^{(t)} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{|S_i^{(t)}|} \sum_{j=1}^{|S_i^{(t)}|} \mathsf{S}_\varepsilon(\alpha, \beta_j). \tag{4.7.3}$$

This step corresponds to computing a barycenter w.r.t Sinkhorn divergence and we use Alg. 4.2. We initialize 20 centroids according to the k-means++ strategy (Arthur and Vassilvitskii, 2007). Fig. 4.4 depicts the 20 centroids obtained by performing k-means with Alg. 4.2. We see that the structure of the digits is successfully detected, recovering also minor details (e.g. note the difference between the centroids related to the digit 2).

Real data: Sinkhorn propagation of weather data. We consider the problem of Sinkhorn *propagation* similar to the one in Solomon et al. (2014). The goal is to predict the distribution of missing measurements for weather stations in the state of Texas, US by "propagating" measurements from neighboring stations in the network. The problem can be formulated as minimizing the functional $\sum_{(v,u)\in\mathcal{E}} r_{uv} \mathsf{S}_\varepsilon(\rho_v, \rho_u)$ over the set $\{\rho_v \in \mathcal{P}(\mathbb{R}^2) | v \in \mathcal{V}_0\}$ with: $\mathcal{V}_0 \subset \mathcal{V}$ the subset of stations with missing measurements, $G = (\mathcal{V}, \mathcal{E})$ the whole graph of the stations network, r_{uv} a weight inversely proportional to the

geographical distance between two vertices/stations $u, v \in \mathcal{V}$. Edges $\mathcal{E}$ are selected as follows: we created a matrix D such that D_{uv} contains the distance between station at vertex u and station at vertex j, computed using the geographical coordinates of the stations. Each node v in $\mathcal{V}$, is connected to those nodes $u \in \mathcal{V}$ such that $D_{vu} \leq 3$. If the number of nodes u that meet this condition is *less* than 5, we connect v with its 5 nearest nodes. If the number of nodes u that meet this condition is *more* than 10, we connect v with its 10 nearest nodes. Each edge e_{uv} is weighted with $\omega_{uv} := D_{uv}$. Since intuitively we may expect that nearer nodes should have more influence in the construction of the histograms of unknown nodes, in the propagation functional we weight $\mathsf{S}_\varepsilon(\rho_v, \rho_u)$ with $r_{uv} = \exp(-\omega_{uv}/\sigma)$ or $r_{uv} = 1/\omega_{vu}$ suitably normalized.

The variable $\rho_v \in \mathcal{P}(\mathbb{R}^2)$ denotes the distribution of measurements at station v of daily *temperature* and *atmospheric pressure* over one year. This is a generalization of the barycenter problem (4.2.1) (see also (Peyré and Cuturi, 2019)). From the total $|\mathcal{V}| = 115$, we randomly select $10\%, 20\%$ or 30% to be *available* stations, and use Alg. 4.2 to propagate their measurements to the remaining "missing" ones. As in Solomon et al. (2014), we use as baseline the Dirichlet (DR) approach, that we briefly recall here. The Dirichlet (DR) approach is a classic method for label propagation proposed in Zhu et al. (2003) that works as follows: assume a label function f is unknown on a subset of vertices $\mathcal{V}_0 \subset \mathcal{V}$ and we wish to infer f on $\mathcal{V}_0$ based on the known values on $\mathcal{V} \setminus \mathcal{V}_0$. The DR method minimizes the Dirichlet energy $\mathcal{E}_D(f) := \sum_{(u,v)\in\mathcal{E}} r_{uv}(f_u - f_v)^2$ over the set of functions with prescribed values on $\mathcal{V} \setminus \mathcal{V}_0$. As in Solomon et al. (2014), when considering probability distributions ρ_v, the Dirichlet method can be naively extended as follows: for each $x \in \mathbb{R}$, $\rho_v(x)$ is interpreted as label f_v for v and the DR method recalled before can be applied. We compare our approach (FW) with the Dirichlet baseline (DR) in terms of the error $d(C_T, \hat{C})$ between the covariance matrix C_T of the groundtruth distribution and that of the predicted one. Here $d(A, B) = \|\log(A^{-1/2}BA^{-1/2})\|$ is the geodesic distance on the cone of positive definite matrices. The average prediction errors are: 2.07 (FW), 2.24 (DR) for 10%, 1.47 (FW), 1.89 (DR) for 20% and 1.3 (FW), 1.6 (DR) for 30%. Standard deviations across 5 runs with randomly selected known vertices are ~ 0.2 for the 10% case and ~ 0.05 for 20% and 30% cases. Fig. 4.5 qualitatively reports the improvement $\Delta = d(C_T, C_{DR}) - d(C_T, C_{FW})$ of our method on individual stations: a higher color intensity corresponds to a wider gap in our favor between prediction errors, from light green ($\Delta \sim 0$) to red ($\Delta \sim 2$). Our

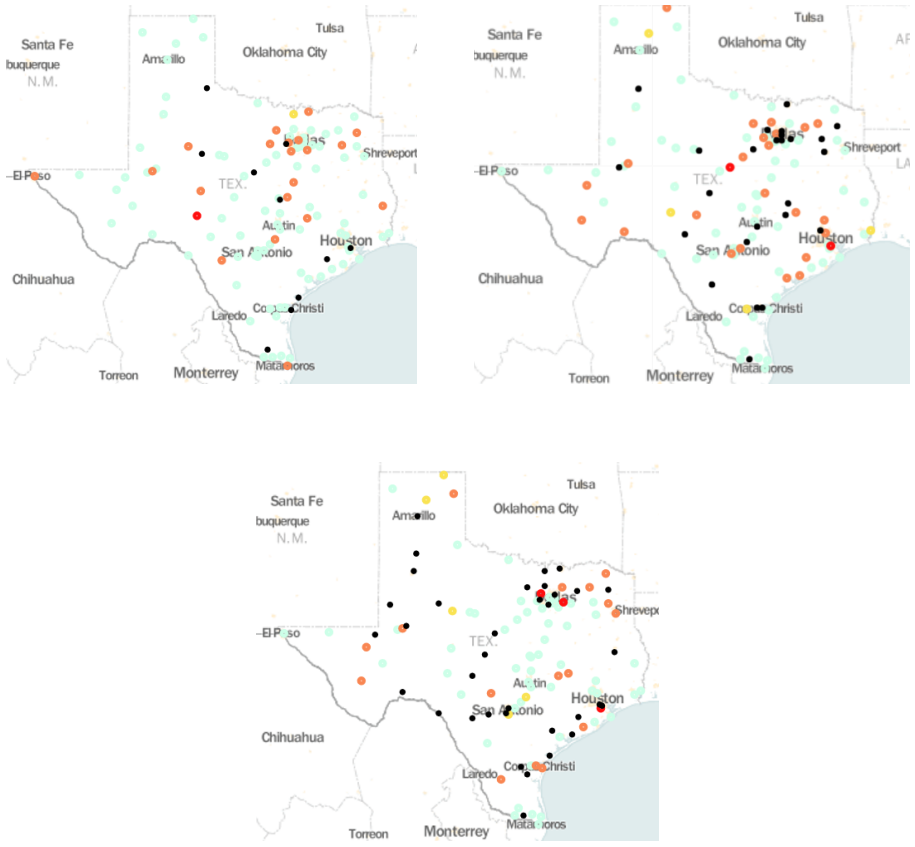


Figure 4.5: From Left to Right: propagation of weather data with $10\%, 20\%$ and 30% stations with available measurements (represented by the black markers). The propagation problem can be interpreted as a generalization of the barycenter problem: given a graph with measurements available in some vertices, the goal is to predict the missing measurements. Alg. 4.2 is tested against the Dirichlet baseline (see text). The quality of the predictions are measured comparing the covariance matrices of the groundtruth distribution C_T and the predicted ones C_{DR} for the Dirichlet method and C_{FW} for Alg. 4.2. The figure displays the improvement $\Delta = d(C_T, C_{DR}) - d(C_T, C_{FW})$: higher color intensity (in the scale light green, yellow, orange, red) corresponds to a bigger gap in favour of Alg. 4.2, from light green $\Delta \sim 0$ to red $\Delta \sim 2$.

approach tends to propagate the distributions to missing locations with higher accuracy. This is in line with the fact that barycenters with Sinkhorn divergence successfully captures the overall geometric structure of the input distributions (intuitively, the distributions of e.g. temperatures over a year registered in close stations have a similar shape) and that our algorithm successfully computes such barycenters.

4.8 Discussion

This chapter dealt with Sinkhorn divergence as a metric for barycenter estimation. We proposed a Frank-Wolfe-based algorithm to find the Sinkhorn barycenter of probability distributions with either finitely or infinitely many support points. Our algorithm belongs

to the family of barycenter methods with free support since it adaptively identifies support points rather than fixing them a-priori. In the finite settings, we were able to guarantee convergence of the proposed algorithm by proving the Lipschitz continuity of gradient of the barycenter functional in the Total Variation sense. Then, by studying the sample complexity of Sinkhorn potential estimation, we proved the convergence of our algorithm also in the infinite case. We empirically assessed our method on a number of synthetic and real datasets, showing that it exhibits good qualitative and quantitative performance. While in this work we have considered FW iterates that are a convex combination of Dirac's delta, models with higher regularity (e.g. mixture of Gaussians) might be more suited to approximate the barycenter of distributions with smooth density. Hence, future work will investigate how the perspective adopted in this work could be extended also to other barycenter estimators.

Chapter 5

Probability matching with Sinkhorn Divergence

In this chapter, we study Sinkhorn divergence as a metric to learn a distribution within the generative model frameworks. Learning a parametric model that fits a set of observation is a fundamental statistical problem. When the target distributions admit a density, the classic approach to estimate such density is the maximum-likelihood estimation. In machine learning problems however, the target distribution is often supported on a low-dimensional domain and does not admit a density. In this case, an established approach to learn such target invokes generative models. Generative models are obtained learning a parametric mapping that given samples from a reference can generate samples that follow the target distribution. The advantage of this approach is that the ability to easily generate samples is often more desirable than knowing the numerical value of the density. The Generative Adversarial Networks (GAN) framework is a well-established paradigm for generative models (Goodfellow et al., 2014). In its original form, it deals with two models simultaneously: a generator g that captures the data distribution, and a discriminator (or critic) D that estimates the probability that a sample came from the training data rather than g. Algorithms in this class aim to reproduce the sampling behavior of the target distribution, rather than explicitly fitting a density function. This is done by modeling the target probability as the *pushforward* via the generator map g of a probability measure in a latent space. Since their introduction, GANs have achieved remarkable progress. From a practical perspective, a large number of model architectures have been explored, leading to impressive results in image generation (Vondrick et al., 2016; Isola et al., 2017; Ledig et al., 2017).

The original paradigm with generator and discriminator models has been subsequently

generalized in a line of work devoted to identify rich metrics for generator training that serve the role of discriminator: namely, the discriminator is played by a distance between probability measures that quantify the discrepancy between the model distribution and the data distribution; the generative model can optimize such distance to produce data that more closely resembles the training data. Rich metrics that have been proposed as discriminator are f-divergences (Nowozin et al., 2016), integral probability metrics (IPM) (Dziugaite et al., 2015) or optimal transport distances (Arjovsky et al., 2017). While recent attention has been devoted to studying the theoretical properties of such models (Liu et al., 2017; Bai et al., 2018; Zhang et al., 2018), a full theoretical understanding of the main building blocks is still missing.

We focus on generative models with regularized Optimal Transport metrics as discriminators (Salimans et al., 2016; Genevay et al., 2018b). This metrics are known to have a bad dependence, in terms of constants, on the dimension of the underlying space when it comes to estimation from samples. Motivated by this insight, our goal is to understand the role of the ambient space dimension and the latent dimension. In particular, we study how the interplay between the latent distribution and the complexity of the pushforward map (generator) affects the overall generalization performance, from both statistical and modelling perspectives. We prove an upper bound on the learning rates of such generative models in terms of a notion of complexity for: $i)$ the ideal generator network and $ii)$ the latent space and distribution.

Our analysis leads us to advocate learning the latent distribution as well as the pushforward map within the generative model paradigm. The approach is in line with previous work on multi-modal GANs (Ben-Yosef and Weinshall, 2018; Pandeva and Schubert, 2019), which models the latent distribution as a Gaussian mixture whose parameters are inferred during training. In fact, our results potentially provide a theoretical justification to the empirical analysis of Ben-Yosef and Weinshall (2018) and Pandeva and Schubert (2019). In contrast to the methods above, our estimator is not limited to Gaussian mixtures but can asymptotically learn any sub-Gaussian latent distribution. Additionally, we characterize the learning rates of our joint estimator and discuss the theoretical advantages over performing standard GANs training, namely fixing the latent distribution a-priori. This chapter is based on (Luise et al., 2020).

Contributions. The main contributions of this chapter include: $i)$ showing how the regu-

larity of a generator network (e.g. in terms of its smoothness) potentially affects the sample complexity of pushforward measures and consequently of the generative models estimator; ii) introducing a novel algorithm for joint training of generator and latent distributions; iii) study the statistical properties (i.e. learning rates) of the resulting estimator in both settings where the generative model is exact as well as the case where the target distribution is only approximately supported on a low dimensional domain.

The rest of the chapter is organized as follows: Sec. 5.1 formally introduces the GANs framework in terms of pushforward measures. Sec. 5.2 discusses the limitations of fixing the latent distribution a-priori and motivates the proposed approach. Sec. 5.3 constitutes the core of the chapter, proposing the joint estimator and studying its generalization properties. Training and sampling strategies are discussed in Sec. 5.4. Finally Sec. 5.5 presents preliminary experiments highlighting the effectiveness of the proposed estimator, while Sec. 5.6 provides a brief discussion of the results and potential future directions.

5.1 Background

The goal of probability matching is to find a good approximation of a distribution ρ given only a finite number of points sampled from it. Typically, one is interested in finding a distribution $\hat{\mu}$ in a class $\mathcal{M}$ of probability measures that best approximates ρ, ideally minimizing

$$\inf_{\mu \in \mathcal{M}} \mathsf{d}(\mu, \rho). \tag{5.1.1}$$

Here d is a discrepancy measure between probability distributions. A wide range of hypotheses spaces $\mathcal{M}$ have been considered in the literature, such as space of distributions parametrized via mixture models (Dempster et al., 1977; Bishop, 2006; Sugiyama et al., 2010; Sriperumbudur et al., 2017), deep belief networks (Hinton et al., 2006; Van den Oord et al., 2016; Van Oord et al., 2016) and variational autencoders (Kingma and Welling, 2014) among the most well known approaches. In this work we focus on generative models that resemble 'in spirit' generative adversarial networks (GAN) (Goodfellow et al., 2014) and that can be formulated as in (5.1.1), leveraging the notions of adversarial divergences and pushforward measures. Below, we introduce these two notions and the notation used in this work.

Adversarial divergences. Let $\mathcal{P}(\mathcal{X})$ be the space of probability measures over a set

$\mathcal{X} \subset \mathbb{R}^d$. Given a space $\mathcal{F}$ of functions $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we define the adversarial divergence between $\mu, \rho \in \mathcal{P}(\mathcal{X})$

$$\mathsf{d}_{\mathcal{F}}(\mu, \rho) \;=\; \sup_{F \in \mathcal{F}} \int F(x', x)\, d(\mu \otimes \rho)(x', x), \tag{5.1.2}$$

namely the supremum over $\mathcal{F}$ of all expectations $\mathbb{E}\left[F(x, x')\right]$ with respect to the joint distribution $\mu \otimes \rho$ (see (Liu et al., 2017)). A well-established family of adversarial divergences are *integral probability metrics (IPM)* (whose definition is recalled in Appendix A.2.2), where $F(x', x) = f(x') - f(x)$ for $f : \mathcal{X} \to \mathbb{R}$ in a suitable space (e.g. a ball in a Sobolev space). Here $\mathsf{d}_{\mathcal{F}}$ measures the largest gap $\mathbb{E}_\mu f(x) - \mathbb{E}_\rho f(x)$ between the expectations of μ and ρ. Examples of IPM used in the GAN paradigm include Maximum Mean Discrepancy (Dziugaite et al., 2015) and Sobolev-IPM (Mroueh et al., 2018). Other adversarial divergences are f-divergences (Nowozin et al., 2016; Goodfellow et al., 2014). Recently, Optimal Transport-based adversarial divergences have attracted significant attention from the GANs literature, such as the Wasserstein distance (Arjovsky et al., 2017), the Sliced-Wasserestein distance (Liutkus et al., 2019; Nadjahi et al., 2019; Deshpande et al., 2018; Wu et al., 2019) or the Sinkhorn divergence (Genevay et al., 2018b; Sanjabi et al., 2018). For completeness, in Appendix D.1 (see also (Liu et al., 2017)) we review how to formulate the adversarial divergences mentioned above within the form of (5.1.2).

Pushforward measures. Pushforward measures are a central component of the GAN paradigm. The notion of pushforward was introduced in Def. 2.1 and we recall it here for convenience.

Definition 5.1 (Pushforward). *Let $\mathcal{Z}$ and $\mathcal{X}$ be two measurable spaces and $T : \mathcal{Z} \to \mathcal{X}$ a measurable map. Let $\eta \in \mathcal{P}(\mathcal{Z})$ be a probability measure over $\mathcal{Z}$. The **pushforward** of η via T is defined to be the measure $T_\# \eta$ in $\mathcal{P}(\mathcal{X})$ such that for any Borel subset B of $\mathcal{X}$,*

$$(T_\# \eta)(B) \;=\; \eta(T^{-1}(B)). \tag{5.1.3}$$

To clarify the notation, in the rest of the paper we will refer to a measure $T_\# \eta$ as pushforward *measure*, and to the corresponding T as pushforward *map*. A key property of pushforward measures is the *Transfer lemma* (Ambrosio et al., 2008, Sec 5.2), which states that for any

measurable $f : \mathcal{X} \to \mathbb{R}$,

$$\int_{\mathcal{X}} f(x)\, d(T_{\#}\eta)(x) \;=\; \int_{\mathcal{Z}} f(T(z))\, d\eta(z). \tag{5.1.4}$$

This property is particularly useful within GAN settings as we discuss in the following.

Generative models with adversarial divergences. The generative adversarial network (GAN) paradigm consists in parametrizing the space $\mathcal{M}$ of candidate models in (5.1.1) as *a set of pushforwards measures of a latent distribution*. From now on, $\mathcal{Z}$ and $\mathcal{X}$ will denote latent and target spaces and we will assume $\mathcal{Z} \subset \mathbb{R}^k$ and $\mathcal{X} \subset \mathbb{R}^d$. Given a set $\mathcal{T}$ of functions $T : \mathcal{Z} \to \mathcal{X}$ and given a (latent) probability distribution $\eta \in \mathcal{P}(\mathcal{Z})$, we consider the space

$$\mathcal{M}(\mathcal{T}, \eta) \;=\; \{\, \mu = T_{\#}\eta \mid T \in \mathcal{T} \,\}.$$

While this choice allows to parameterize the target distribution only implicitly, it offers a significant advantage at sampling time: sampling x from $\mu = T_{\#}\eta$ corresponds to sampling a z from η and then taking $x = T(z)$. By leveraging the Transfer lemma (5.1.4) and using an adversarial divergence $\mathsf{d}_{\mathcal{F}}$, the probability matching problem in (5.1.1) recovers the original minimax game formulation in Goodfellow et al. (2014)

$$\inf_{\mu \in \mathcal{M}(\mathcal{T},\eta)} \mathsf{d}_{\mathcal{F}}(\mu, \rho) \;=\; \inf_{T \in \mathcal{T}} \mathsf{d}_{\mathcal{F}}(T_{\#}\eta, \rho) \;=\; \inf_{T \in \mathcal{T}} \sup_{F \in \mathcal{F}} \int F(T(z), x)\, d(\eta \otimes \rho)(z, x). \tag{5.1.5}$$

Within the GAN literature, the pushforward T is referred to as the *generator* and optimization is performed over a suitable $\mathcal{T}$ (e.g. a set of neural networks (Goodfellow et al., 2014)) for a fixed η (e.g. a Gaussian or uniform distribution). The term F is called *discriminator* since, when for instance $\mathsf{d}_{\mathcal{F}}$ is an IPM, $F(x', x) = f(x') - f(x)$ aims at maximally separating (discriminating) the expectations of μ and ρ. While this is not the original GANs formulation and it is a variant of generative model based on adversarial divergences, in the following with some abuse of terminology we will use the term GAN for the model above.

5.2 The Complexity of Modeling the Generator

In this section we discuss a main limitation of choosing the latent distribution a-priori within the GANs framework. This will motivate our analysis in Sec. 5.3 to learn the latent distribution jointly with the generator. Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ be the (unknown) target distribution

and $\rho_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ an empirical distribution of n Dirac's deltas δ_{x_i} centered on i.i.d. points $(x_i)_{i=1}^n$ sampled from ρ. Given a latent distribution $\eta \in \mathcal{P}(\mathbb{R}^k)$ (such as Gaussian or uniform distribution), GAN training consists in learning a map $\hat{T}$ such that

$$\hat{T} = \operatorname*{argmin}_{T\in\mathcal{T}} \mathsf{d}_{\mathcal{F}}(T_{\#}\eta, \rho_n). \tag{5.2.1}$$

The potential downside of this strategy is that it offloads all the complexity of modeling the target ρ onto the generator T. Therefore, for a given η, it might happen that the equality $T_{\#}\eta = \rho$ is satisfied only by very complicated – hence hard to learn – pushforward maps. To illustrate when this might be the case and its effects on modeling and learning, we discuss some examples below (see Appendix D.2 for technical details). We first recall a characterization of pushforward measures that will be instrumental in building this intuition.

Proposition 5.1 (Simplified version of (Ambrosio et al., 2008, Lemma 5.5.3))**.** *Let ρ and $\eta \in \mathcal{P}(\mathbb{R}^d)$ admit density functions $f, g : \mathbb{R}^d \to \mathbb{R}$ with respect to the Lebesgue measure, denoted $\eta = f\mathcal{L}^d$ and $\rho = g\mathcal{L}^d$. Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be injective a.e. and differentiable, then $\rho = T_{\#}\eta$ if and only if*

$$g(T(x))|\det\nabla T| = f(x). \tag{5.2.2}$$

Using Prop. 5.1 we can interpret the GAN problem as akin to *solving the differential equation* (5.2.2). Therefore, choosing η a-priori might implicitly require a very complex model space $\mathcal{T}$ to find such a solution. In contrast, the following example describes the case where $\mathcal{T}$ contains only simple models.

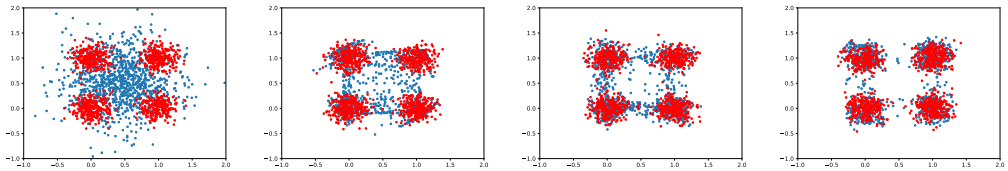
Example 5.1 (Affine Pushforward Maps)**.** *Let $\eta \in \mathcal{P}(\mathbb{R}^d)$ admit a density $f : \mathbb{R}^d \to \mathbb{R}$ and let*

$$\mathcal{T} = \{\, T_{A,b} : \mathbb{R}^d \to \mathbb{R}^d \mid T_{A,b}(z) = Az + b,\ A \in \mathbb{R}^{d\times d},\ b \in \mathbb{R}^d,\ \det(A) \neq 0 \,\}. \tag{5.2.3}$$

Then, for $T = T_{A,b} \in \mathcal{T}$, the measure $\rho = T_{\#}\eta \in \mathcal{P}(\mathbb{R}^d)$ admits a density $g : \mathbb{R}^d \to \mathbb{R}$ such that

$$g(z) = f(A^{-1}(z-b)) \cdot |\det(A^{-1})|. \tag{5.2.4}$$

The set $\mathcal{T}$ of affine generators is able to parametrize only a limited family of distributions (essentially translations and re-scaling of the latent η). This prevents significant changes to



(a) 1 layer **(b)** 2 layers **(c)** 5 layers **(d)** 7 layers

Figure 5.1: Sinkhorn GAN estimation between a 2D Gaussian and a mixture of four 2D Gaussians with generator space $\mathcal{T}$ of increasing complexity (depth of the network). Real (Red) vs generated (Blue) samples.

the shape of the latent distribution to match the target; for example, a uniform (alternatively a Gaussian) measures can match only uniform (Gaussian) measures. Below, we illustrate two examples where the pushforward map can indeed be quite complex and therefore require a larger space $\mathcal{T}$ when solving (5.2.1).

Example 5.2 (Uniform to Gaussian)**.** *Let η be the uniform distribution on the interval $[-1, 1]$ and ρ the Gaussian distribution on $\mathbb{R}$, with zero mean and unit variance. Then $T_{\#}\eta = \rho$, with $T(x) = \sqrt{2}\text{erf}^{-1}(x)$ the inverse to the standard error function $\text{erf}(x) = \rho\big((-\infty, x]\big)$.*

The map erf^{-1} is highly nonlinear and with steep derivatives. Therefore, learning a GAN from a uniform to a Gaussian distribution would require choosing a significantly large space $\mathcal{T}$ to approximate erf^{-1}. We further illustrate the effect of a similar behaviour with an additional empirical example.

Empirical example (Multi-modal Target). We consider the case where ρ is multimodal (a mixture of four Gaussian distributions in 2D), while η is unimodal (a Gaussian in 2D). Fig. 5.1 qualitatively compares samples from the real distribution ρ against samples from $T_{\#}\eta$, with T learned via GAN training in (5.2.1) (with Sinkhorn loss $\mathsf{d}_{\mathcal{F}}$, see Sec. 5.3) for spaces $\mathcal{T}$ of increasing complexity (neural networks with increasing depth). Linear generators are clearly unsuited for this task, and only highly non-linear models yield reasonable estimates. See Appendix D.2 for details on the experimental setup.

5.3 Learning the Latent Distribution

The arguments above suggest that choosing the latent distribution a-priori can be limiting in several settings. Therefore, in this work we propose to learn the latent distribution jointly with the generator. Given a family $\mathcal{H}$ of latent distributions, we aim to solve

$$(\hat{T}, \hat{\eta}) \;=\; \underset{T\in\mathcal{T},\eta\in\mathcal{H}}{\text{argmin}}\; \mathsf{d}_{\mathcal{F}}(T_{\#}\eta, \rho_n). \tag{5.3.1}$$

A natural question is how the learning rates of $(\hat{T}, \hat{\eta})$ are affected by the choice of $\mathcal{T}$ and $\mathcal{H}$. In this work we address this question for the case where $\mathsf{d}_{\mathcal{F}}$ is the Sinkhorn divergence (Cuturi, 2013). Indeed, Optimal Transport is particularly suited to capture the geometric properties of distribution supported on low-dimensional manifolds (Weed and Bach, 2019) (e.g. pushforward measures from a low-dimensional latent space). Moreover, for the Sinkhorn divergence, discriminator training, i.e. finding F in (5.1.5), can be efficiently solved to arbitrary precision via the Sinkhorn-Knopp algorithm (see (Cuturi, 2013) and Sec. 5.4). Below, we introduce our choices for $d_{\mathcal{F}}, \mathcal{H}$ and $\mathcal{T}$ and then proceed to characterize the learning rates of $\hat{T}_{\#}\hat{\eta}$.

Choosing $\mathsf{d}_{\mathcal{F}}$: Sinkhorn divergence. Sinkhorn divergence and its properties were presented in Sec. 2.5. Here just recall the definition: for any $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ Sinkhorn divergence is defined as

$$\mathsf{S}_\varepsilon(\alpha, \beta) \;=\; \mathsf{OT}_\varepsilon(\alpha, \beta) \;-\; \frac{1}{2}\mathsf{OT}_\varepsilon(\alpha, \alpha) \;-\; \frac{1}{2}\mathsf{OT}_\varepsilon(\beta, \beta), \tag{5.3.2}$$

where OT_ε is the Entropic Optimal Transport cost defined in (2.3.1).

Choosing $\mathcal{H}$: sub-Gaussian distributions. For the purpose of our analysis, in the following we will restrict to a class $\mathcal{H}$ of distribution that are not too spread out on the entire latent domain $\mathcal{Z} \subseteq \mathbb{R}^k$. In particular, we will parametrize $\mathcal{H} \subset \mathcal{G}_\sigma(\mathcal{Z})$ the space of σ-*sub-Gaussian* distributions on $\mathcal{Z}$, namely distributions η such that $\int e^{\|z\|^2/2k\sigma^2}\, d\eta(z) \leq 2$. Gaussian distributions and probabilities supported on a compact set belong to this family. Thus, $\mathcal{H}$ recovers the case of standard GANs. Note that the parameter σ allows us to upper bound all moments of a distribution and can therefore be interpreted as a quantity that controls the complexity of η.

Choosing $\mathcal{T}$: balls in $C^s(\mathcal{Z}, \mathcal{X})$. In the following we will restrict our analysis to spaces of functions that satisfy specific regularity conditions. In particular, we will consider $\mathcal{T} \subset C^s_{\tau,L}(\mathcal{Z}, \mathcal{X})$ to be contained in the set of L-Lipscthitz functions in the ball of radius τ in the space of continuous functions from $\mathcal{Z} \subset \mathbb{R}^k$ to $\mathcal{X} \subset \mathbb{R}^d$ equipped with the uniform norm $\|\cdot\|_{\infty,s}$ on all partial derivatives up to order s. Intuitively, the norm $\|T\|_{\infty,s}$ quantifies the complexity of the generator T, hence reflecting how easy (or hard) it is to learn it in

practice. In our analysis we will require $s \geq \lceil k/2 \rceil + 1$. To simplify our analysis in the following, we add the additional requirements that $T(0) = 0$ for any $T \in \mathcal{T}$ (in case of an offset, one can factor out a translation first (see (Peyré and Cuturi, 2019, Remark 2.19)). This choice of the space $\mathcal{T}$ allows to formally model a wide range of smooth generators T (e.g. pushforward maps parametrized by neural networks with smooth activations or via smooth reproducing kernels).

We are now ready to state our main result, which characterizes the learning rates of the estimator in (5.3.1) in terms of the complexity parameters associated to the spaces $\mathcal{T}$ and $\mathcal{H}$ introduced above.

Theorem 5.2. *Let $\mathcal{Z} \subset \mathbb{R}^k$, $\mathcal{X} \subset \mathbb{R}^d$ and $\rho = T^*_\# \eta^*$ with $T^* \in \mathcal{T} \subset C^{\lceil k/2 \rceil + 1}_{\tau, L}(\mathcal{Z}, \mathcal{X})$ and $\eta^* \in \mathcal{H} \subset \mathcal{G}_\sigma(\mathcal{Z})$. Let $(\hat{T}, \hat{\eta})$ satisfy (5.3.1) with $\mathsf{d}_\mathcal{F} = \mathsf{S}_\varepsilon$ and ρ_n a sample of n i.i.d. points from ρ. Then,*

$$\mathbb{E}\, \mathsf{S}_\varepsilon(\hat{T}_\# \hat{\eta}, \rho) \;\leq\; \frac{\mathsf{b}(\tau, L, \sigma, k)}{\sqrt{n}}$$

where $\mathsf{b}(\tau, L, \sigma, k) \;=\; \mathsf{C}_k \, (1 + \tau^k L^{\lceil 3k/2 \rceil + 1} \, \sigma^{\lceil 5k/2 \rceil + 6} \, \varepsilon^{-\lceil 5k/4 \rceil - 3})$ with C_k a constant depending only on the latent space dimension k.

Thm. 5.2 quantifies the tradeoff between the complexity terms τ, L of the pushforward map T^* and σ of the latent distribution η^*. In particular, we note that: $i)$ we pay a polynomial cost in terms of the sub-Gaussian parameter σ of the latent distribution η^*; $ii)$ we pay a cost proportional to the complexity of the target generator, including its $\|T^*\|_{s,\infty}$ norm and Lipschitz constant L; $iii)$ all terms depend on the dimension k of the latent space and *not* on the target space dimension d. This result suggests that the GAN paradigm is particularly suited to settings where the target distribution can be modeled in terms of a low-dimensional latent distribution *and* a regular pushforward map. Extending the result to larger families of pushforward maps (e.g. with weaker regularity assumptions) would be interesting but would require a different strategy than the one proposed here. Thus, it will be the subject of future work.

Sketch of the proof. The proof of Thm. 5.2 is quite technical so we moved it to Appendix D.4. Here we present the main steps and key ideas. We begin by observing that the matching error

$\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta}, \rho) - \mathsf{S}_\varepsilon(T^*_\#\eta^*, \rho)$ (which is equal to $\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta}, \rho)$ by assumption) is controlled by

$$\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta}, \rho) - \mathsf{S}_\varepsilon(T^*_\#\eta^*, \rho) = A_1 + A_2 + A_3, \tag{5.3.3}$$

where

$$A_1 = \mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta}, \rho) - \mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta}, \rho_n) \tag{5.3.4}$$

$$A_2 = \mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta}, \rho_n) - \mathsf{S}_\varepsilon(T^*_\#\eta^*, \rho_n) \tag{5.3.5}$$

$$A_3 = \mathsf{S}_\varepsilon(T^*_\#\eta^*, \rho_n) - \mathsf{S}_\varepsilon(T^*_\#\eta^*, \rho). \tag{5.3.6}$$

By leveraging the optimality of the estimator $\hat{T}_\#\hat{\eta}$ in minimizing (5.3.1), we have that $A_2 \leq 0$. Note that

$$A_1 + A_3 \leq 2 \sup_{T\in\mathcal{T},\eta\in\mathcal{H}} \Big[\mathsf{S}_\varepsilon(T_\#\eta, \rho_n) - \mathsf{S}_\varepsilon(T_\#\eta, \rho)\Big]. \tag{5.3.7}$$

So combining the two equations above, we have that the matching error is upper bounded by

$$\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta}, \rho) \;\leq\; 2 \sup_{T\in\mathcal{T},\eta\in\mathcal{H}} |\mathsf{S}_\varepsilon(T_\#\eta, \rho) - \mathsf{S}_\varepsilon(T_\#\eta, \rho_n)|. \tag{5.3.8}$$

The right hand side corresponds to the largest generalization error of estimators in $(\mathcal{T}, \mathcal{H})$. This quantity is related to the sample complexity of ρ_n with respect to the Sinkhorn divergence. The latter is a topic recently studied in Genevay et al. (2018a); Mena and Niles-Weed (2019), with bounds available for controlling $|\mathsf{S}_\varepsilon(\mu, \rho) - \mathsf{S}_\varepsilon(\mu, \rho_n)|$ for μ a fixed distribution. However, to control (5.3.8) we need to provide a uniform upper bound for the sample complexity of the Sinkhorn divergence over the class $(\mathcal{T}, \mathcal{H})$. To do so, we use the following.

Lemma 5.3 (Informal)**.** *Let $\eta, \nu_1, \nu_2 \in \mathcal{G}_\sigma(\mathcal{Z})$ and $T, T' \in \mathcal{T}$ with $\mathcal{T}$ as in Thm. 5.2. Then,*

$$|\mathsf{S}_\varepsilon(T_\#\eta, T'_\#\nu_1) - \mathsf{S}_\varepsilon(T_\#\eta, T'_\#\nu_2)| \;\leq\; \sup_{u\in\mathcal{F}_{\sigma,\tau,L}} |\int u(z)\, d\nu_1 - \int u(z)\, d\nu_2(z)| \tag{5.3.9}$$

with $\mathcal{F}_{\sigma,\tau,L}$ a suitable space of functions $u : \mathcal{Z} \to \mathbb{R}$ that does not depend on η, T and T' but only on the complexity parameters σ, τ and L (see Appendix D.3 for the characterization of $\mathcal{F}_{\sigma,\tau,L}$).

The result implies that we can upper bound the generalization error of $T_{\#}\eta$ in terms of the integral probability metric $\mathsf{d}_{\mathcal{F}_{\sigma,\tau,L}}(\eta^*, \eta_n^*)$ (i.e. (right hand side of (5.3.9)) between the true latent η^* and its empirical sample η_n^*. Note that this quantity is uniform with respect to the sub-Gaussian parameter of distributions in $\mathcal{H}$ and the regularity of the class $\mathcal{T}$. Following (Mena and Niles-Weed, 2019, Thm. 2), we can control $\mathsf{d}_{\mathcal{F}_{\sigma,\tau,L}}(\eta^*, \eta_n^*)$ in expectation by estimating the covering numbers of a rescaling of $\mathcal{F}_{\sigma,\tau,L}$. □

Thm. 5.2 studies the learning rates of the estimator in (5.3.1) when the GAN model is exact. A natural question is whether similar results hold when this is only an approximation. Below, we consider the case where the target distribution is 'almost' a low-dimensional pushforward (e.g. it is concentrated around a low-dimensional manifold) but is supported on a larger domain (e.g. due to noise).

Remark 5.1. *The interest in this case is inspired by some results in Weed and Bach (2019) on approximation rates of standard Wasserstein distance. As briefly discussed in Chapter 2, given a probability measure ρ on a $d-$dimensional domain, its empirical version ρ_n approaches ρ in Wasserstein distance at the rate $n^{-1/d}$. In Weed and Bach (2019), it was shown that if ρ is supported on a k dimensional domain embedded in a d-dimensional space, with $k < d$, then the rate of approximation depends on k only. It is natural to ask what happens for measures which are 'approximately' low-dimensional, meaning that they are supported on a low-dimensional manifold up to some noise, for example. The analysis in Weed and Bach (2019) covers this case and proves that the advantage resides is non-asymptotic faster rates, i.e. a better approximation for n up to a certain threshold. The result that we obtain here is different in nature but follows the same motivation: we are interested in understanding whether the 'almost' low-dimensional support of the target distribution can still leads to any benefit in the statistical analysis.*

Approximation error for noisy models. Let the target distribution ρ be obtained by convolving $T_{\#}^*\eta^*$ with a distribution Φ_δ with sub-Gaussian parameter $\delta > 0$. Recall that the convolution $\Phi_\delta * \mu$ is defined as the distribution such that, for any measurable $f : \mathcal{X} \to \mathbb{R}$

$$\int f(x)\, d(\Phi_\delta * \mu)(x) \;=\; \int f(w+y)\, d\mu(y) d\Phi_\delta(w).$$

Therefore, $\rho = \Phi_\delta * T_{\#}\eta$ can be interpreted as the process of 'perturbing' the distribution $T_{\#}\eta$

by means of a probability Φ_δ. Standard examples are the cases where Gaussian or uniform noise is added to samples from the pushforward. This perturbation affects the generalization of the proposed estimators by a term proportional to the sub-Gaussian parameter δ of the noise Φ_δ, as follows.

Corollary 5.4. *Under the same assumption of Thm. 5.2, let $\Phi_\delta \in \mathcal{G}_\delta(\mathcal{X})$ and $\rho = \Phi_\delta * T^*_\# \eta^*$. Then,*

$$\mathbb{E}\, \mathsf{S}_\varepsilon(\hat{T}_\# \hat{\eta}, \rho) \;\leq\; \frac{2\,\mathsf{b}(\tau, L, \sigma, k)}{\sqrt{n}} \;+\; 3\mathsf{b}_1(L, \sigma, d, k)\;\delta,$$

where $\mathsf{b}(\tau, L, \sigma, k)$ is the same constant as in Thm. 5.2 and $\mathsf{b}_1(L, \sigma, d, k)$ is another constant that is reported in the formal version of the result in Cor. D.15.

Cor. 5.4 characterizes the approximation behavior of the GAN paradigm (see Appendix D.4.1 for a proof). It shows that when the target distribution is essentially low-dimensional, we can recover it up to a quantity that depends on the intensity of the noise Φ_δ. This is reminiscent of the irreducible error in supervised learning settings when approximating a function lying outside the hypothesis space (Shalev-Shwartz and Ben-David, 2014).

5.4 Optimization

In this section we discuss how to parametrize the spaces $\mathcal{T}$ and $\mathcal{H}$ and tackle the joint GAN problem in practice. We note that minimizing $\mathsf{S}_\varepsilon(T_\# \eta, \rho_n)$ with respect to either T or η critically hinges on the dual formulation of entropic Optimal Transport. Therefore, we first briefly recall its formulation and main properties for convenience. Then, we use such notion to address (5.3.1).

Dual Formulation of Entropic OT. Recall that the dual formulation of $\mathsf{OT}_\varepsilon(\alpha, \beta)$ (presented in (2.3.3)) is

$$\max_{u,v \in \mathcal{C}(\mathcal{X})} \int u(x)\, d\alpha(x) \;+\; \int v(y)\, d\beta(y) \;-\; \varepsilon \int e^{\frac{u(x)+v(y)-\|x-y\|^2}{\varepsilon}}\, d\alpha(x) d\beta(y).$$

This problem always admits a pair of minimizers (u^*, v^*), also known as *Sinkhorn potentials* (Sinkhorn, 1964). When α and β are probability distributions with finite support, the well-established SINKHORNKNOPP algorithm can be applied to efficiently obtain the values of u^* and v^* on the support points of α and β respectively (Sinkhorn, 1964; Cuturi, 2013). Then,

u^* and v^* can be evaluated on any point of $\mathcal{X}$ by means of the following characterization of the Sinkhorn potentials (presented in (2.3.8))

$$u^*(x) \;=\; -\,\varepsilon\,\log\int e^{\frac{v^*(y)-\|x-y\|^2}{\varepsilon}}\,d\beta(y). \tag{5.4.1}$$

This characterization will be of particular interest in the following. Indeed both optimization of the generator and latent distribution will make use of the explicit calculation of the gradients of u^*.

Learning the Generator. Minimizing (5.3.1) with respect to T for a fixed latent distribution η corresponds to training a standard GAN. The case of Sinkhorn GANs was originally studied in Salimans et al. (2018); Genevay et al. (2018b). In practice, one considers a parametric family of generators T_θ with θ in some parameters space Θ. The gradients of Sinkhorn divergence $\mathsf{S}_\varepsilon(T_{\theta\#}\eta, \rho_n)$ with respect to θ can be obtained via automatic differentiation. Here, we also provide an analytic formula which highlights the dependence on the gradient of the potential and the gradient of the generator map. We will assume the parametrization to be differentiable a.e., and denote by $\nabla_\theta T_\theta$ the gradient of T_θ with respect to θ. By leveraging the characterization of dual potential in (5.4.1), we have the following.

Proposition 5.5. *Let $\eta \in \mathcal{P}(\mathcal{Z})$ and $\rho \in \mathcal{P}(\mathcal{X})$. Let (u^*, v^*) be a pair of minimizers of (4.1.1) with $\alpha = T_{\theta\#}\eta$ and $\beta = \rho$. Then, the gradient of $\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho)$ in θ_0 is*

$$\big[\nabla_\theta \mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho)\big]|_{\theta=\theta_0} = \int \big[\nabla_x u^*(\cdot)\big]|_{x=T_{\theta_0}(z)}\,\big[\nabla_\theta T_\theta(z)\big]|_{\theta=\theta_0}\,d\eta(z). \tag{5.4.2}$$

Remark 5.2 (Mini-batches)**.** *When η has dense support or the number of points in the support is very large, computing the gradient of OT_ε with either (5.4.2) or automatic differentiation can become prohibitive. A commonly used approach (Genevay et al., 2018b) is to sample m points from η and ρ and compute $\nabla_\theta \mathsf{OT}_\varepsilon(T_{\theta\#}\eta_m, \rho_m)$ as a proxy of the target gradient. We care to point out that such gradient is not an unbiased estimator of the real gradient. This means that stochastic gradient descent approaches are not theoretically justified and may fail in practice. Recent research (Mensch and Peyré, 2020; Fatras et al., 2019) is exploring this aspect that is key to enable effective large scale applications.*

Learning the Latent Distribution. To guarantee $\hat{\eta}$ to be sub-Gaussian, we consider

Algorithm 5.1 LATENT DISTRIBUTION LEARNING GANS

Input: Target ρ_n, latent dimension k, initial network params θ, Sinkhorn param $\varepsilon > 0$, step sizes $\alpha_1, \alpha_2 > 0$, perturbation Φ_δ (e.g. Gaussian $\mathcal{N}(0, \delta I_k)$ in $\mathbb{R}^k$), starting particles $(z_i)_{i=1}^m$, sampling size ℓ.

Until convergence do:

Sample $(i_j, w_j)_{j=1}^\ell$ with $i_j \sim \text{Unif.}\{1, \dots, m\}$ and $w_j \sim \Phi_\delta$

Let $\mu = \frac{1}{\ell}\sum_{j=1}^\ell \delta_{x_j}$ with $x_j = T_\theta(z_{i_j} + w_j)$

$(u^*, v^*) = \text{SINKHORNKNOPP}(\mu, \rho_n, \varepsilon)$

$\theta \leftarrow \theta - \alpha_1 \sum_{j=1}^\ell \nabla_x u^*(x_j)\, \nabla_\theta T_\theta(z_{i_j} + w_j)$

$z_i \leftarrow z_i - \alpha_2 \sum_{j \mid i_j = i} \nabla_x u^*(x_j)\, \nabla_z T_\theta(z_i + w_j)$

Return: $\hat{\mu} = T_{\theta\#}\hat{\eta}$ with $\hat{\eta} = \Phi_\delta * \hat{\nu}$ and $\hat{\nu} = \frac{1}{m}\sum_{i=1}^m z_i$.

Sampling: $x \sim \hat{\mu}$ obtained as $x = T(z_i + w)$ with $i \sim \text{Unif.}\{1, \dots, m\}$ and $w \sim \Phi_\delta$.

$\mathcal{H} = \mathcal{P}(\mathcal{Z})$ the set of probability measures over a compact subset $\mathcal{Z}$ of the latent space $\mathbb{R}^k$. Optimization over a space of measures $\mathcal{P}(\mathcal{Z})$ is itself an active research topic. Possible strategies include Conditional Gradient (Bredies and Pikkarainen, 2013; Boyd et al., 2017; Mensch et al., 2019; Luise et al., 2019), Mirror Descent (Hsieh et al., 2019) or the particle based approaches discussed below (Feydy et al., 2019; Chizat, 2019).

Flow-based methods approximate the target distribution with a set of m particles $\eta = \sum_{i=1}^m \omega_i \delta_{z_i}$ whose position is then optimized to minimize $\mathsf{S}_\varepsilon(T_\#\eta, \rho_n)$ (for simplicity, here we do not learn the ω_i but fix them to $1/m$). This problem can be solved by a gradient descent-based algorithm in the direction minimizing the associated Sinkhorn potentials (Feydy et al., 2019). More precisely, given (u^*, v^*) a minimizer of (4.1.1), we update the position of each particle via a gradient step of size $\alpha > 0$

$$z_+ = z - \alpha \nabla_z u^*(T(z)). \tag{5.4.3}$$

We refer to (Chizat, 2019) for more details and a comprehensive analysis of convergence and approximation guarantees for particle-based methods with respect to m the number of particles.

Sampling & **Training.** Both conditional gradient and flow-based methods approximate the ideal $\hat{\eta}$ via a discrete distribution. While these strategies are guaranteed to approximate to arbitrary precision the ideal solution, they cannot be directly used for sampling new points. To this end here we propose to model $\eta = \Phi_\delta * \nu$ as the convolution of a discrete $\nu = \frac{1}{m}\sum_{i=1}^m \delta_{z_i}$ with a δ-variance Gaussian distribution Φ_δ. We can then address the

following variant to the joint problem (5.3.1)

$$\min_{\theta\in\Theta,\nu\in\mathcal{H}} \mathsf{S}_\varepsilon(T_{\theta\#}(\Phi_\delta * \nu), \rho_n), \tag{5.4.4}$$

where ν is learned by means of the flow-based approaches introduced above (by sampling a new set of points from Φ_δ at each iteration). This strategy effectively renders the estimated η to be a mixture of m Gaussian distributions, whose position on the latent space is optimized iteratively.

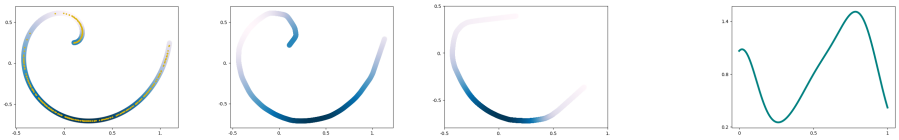
Alg. 5.1 summarizes the process of jointly learning $\eta = \Phi_\delta * \nu$ and T_θ, according to (5.4.4). For simplicity, we consider the case where we optimize simultaneously both network parameters and support points of the latent distribution. However other options are viable, such as block coordinate descent or alternating minimization, where each term is optimized while keeping the other fixed to the previous step. The algorithm proceeds iteratively by: i) sampling points from the current estimate of $\hat{\eta}$ in terms of the discrete $\hat{\nu}$ and the perturbation Φ_δ; ii) computing the Sinkhorn potential u^* via the SINKHORNKNOPP[1] algorithm; iii) update the network parameters θ and latent $\hat{\nu}$ according to the gradient steps (5.4.2) and (5.4.3) respectively.

5.5 Experiments

We tested the proposed strategy of jointly learning the latent distribution and generator on two synthetic experiments. We do so by comparing the performance of the joint GAN estimator from Alg. 5.1 and of the standard GAN estimator, with fixed latent distribution. We report both the qualitative sampling behavior of the two methods as well as their quantitative performance in terms of *generalization gap*, namely the value $\mathsf{S}_\varepsilon(T_{\#}\eta, \rho)$ attained at convergence (using the Sinkhorn distance between generated and *new* real samples as a proxy). Details on the setup, data generation, networks specifications and training are reported in Appendix D.6.

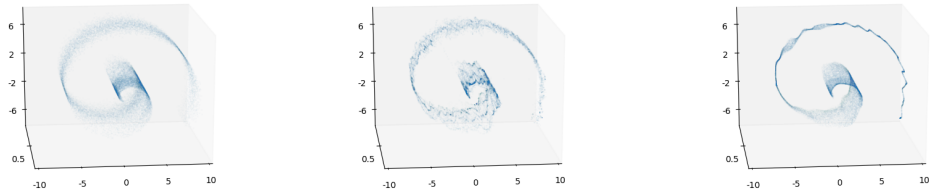
Spiral. We chose the target ρ to be a multimodal probability measure in $\mathbb{R}^2$ supported on a spiral-shaped 1D manifold (Fig. 5.2a, where the color intensity is proportional to higher density). Given the low-dimensionality of the target, we consider a GAN model with latent space $\mathcal{Z}$ in $\mathbb{R}$. We compare Alg. 5.1 against a GAN trained with fixed latent distribution

[1] We used the implementation available at https://www.kernel-operations.io/geomloss/.



(a) Target density **(b)** Alg. 5.1 **(c)** Fixed latent **(d)** Learned $\hat{\eta}$ by Alg. 5.1

Figure 5.2: Left: Multimodal distribution supported on the spiral: the target ρ is a multimodal probability measure in $\mathbb{R}^2$ supported on a 1-dimensional spiral-shaped manifold; middle: estimator $\hat{T}_{\#}\hat{\eta}$ trained on a sample ρ_n with $n = 1000$ points iid from ρ using Alg. 5.1; Right: estimator $\hat{T'}_{\#}\mathcal{N}(0,1)$ trained on a sample ρ_n with $n = 1000$ points iid from ρ. The latent distribution is fixed.

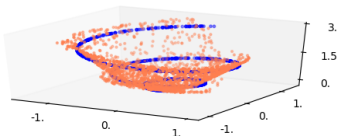


(a) Samples from the target ρ **(b)** Samples from $\hat{T}_{\#}\hat{\eta}$ from Alg. 5.1 **(c)** Samples from GAN with fixed η_0

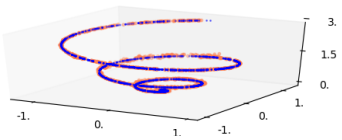
Figure 5.3: Left: Multimodal distribution supported on the swiss-roll: the target ρ is a multimodal probability measure in $\mathbb{R}^3$ supported on a 2-dimensional swiss-roll manifold; middle: estimator $\hat{T}_{\#}\hat{\eta}$ trained on a sample ρ_n with $n = 1000$ points iid from ρ using Alg. 5.1; Right: estimator $\hat{T'}_{\#}\mathcal{N}(0, Id)$ trained on a sample ρ_n with $n = 1000$ points iid from ρ. The latent distribution is fixed.

$\eta_0 = \mathcal{N}(0,1)$ (the univariate Gaussian measure on $\mathbb{R}$) by training them on a sample ρ_n of $n = 1000$ i.i.d. points sampled from ρ. Fig. 5.2 reports the density and a sample of the ground-truth target ρ (Fig. 5.2a), our $\hat{T}_{\#}\hat{\eta}$ estimated via Alg. 5.1 (Fig. 5.2b) and $\hat{T}'_{\#}\eta_0$ trained with standard Sinkhorn GAN (Fig. 5.2c). We see that the generator $\hat{T}'$ is unable to apply enough distortion to η_0 to match ρ (see our discussion in Sec. 5.2). In contrast, our method recovers the target distribution with high accuracy. This is quantitatively reflected by the generalization gaps: $\mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho) < 10^{-5}$ for Alg. 5.1 and $\mathsf{S}_\varepsilon(\hat{T}'_{\#}\eta_0, \rho) > 0.1$ for the fixed latent. Fig. 5.2d reports the density of the latent $\hat{\eta}$ on $\mathbb{R}$ to show how our method captures the bi-modality of ρ.

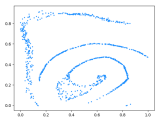
Swiss Roll. Similarly to the previous setting, we consider ρ a multimodal distribution in $\mathbb{R}^3$ supported on the 2D swiss roll manifold. Latent space was set as $\mathcal{Z} \subset \mathbb{R}^2$. Fig. 5.3 (Left to right) shows samples from the ground truth, our joint $\hat{T}_{\#}\hat{\eta}$ and the standard GAN $\hat{T}_{\#}\eta_0$ with $\eta_0 = \mathcal{N}(0, I)$. Also in this case, the latter generator was not able to fully recover the geometry of ρ, as indicated by the different generalization gaps $\mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho) < 0.5$ and $\mathsf{S}_\varepsilon(\hat{T}'_{\#}\eta_0, \rho) > 1.8$.



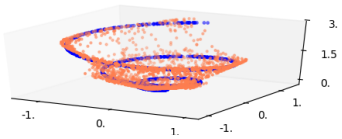
(a) Samples standard GAN



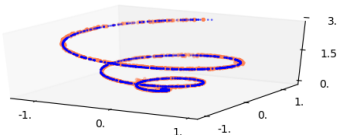
(b) Samples Alg. 5.1



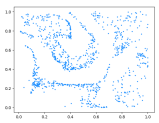
(c) Samples from $\hat{\eta}$ by Alg. 5.1



(d) Samples standard GAN



(e) Samples Alg. 5.1



(f) Samples from $\hat{\eta}$ by Alg. 5.1

Figure 5.4: results for GAN training with $\mathcal{T}$ space of generators of increasing complexity. (Top row) $\mathcal{T}$ space of 2-*layers generators*, (Bottom row) $\mathcal{T}$ space of 2-*layers generators*. (First two columns) Samples from the target distribution ρ (blue) and generated samples (orange) for respectively the standard GANs with fixed latent Gaussian distribution (left column) and $\hat{T}_{\#}\hat{\eta}$ learned via Alg. 5.1 (central column). (Right column) samples from the latent distribution $\hat{\eta}$ learned via Alg. 5.1.

From 2D to 3D: matching of a 1-dimensional helix. We considered the task of matching a probability measure $\rho \in \mathcal{P}(\mathbb{R}^3)$ supported on a 1-dimensional helix-shaped manifold. While the target distribution could be modeled in terms of a latent distribution on the real line and a suitable pushforward, here we consider a model where $\eta \in \mathcal{P}(\mathbb{R}^2)$ is a probability in $2D$ and $T : \mathbb{R}^2 \to \mathbb{R}^3$. The goal of this experiment is to qualitatively assess the impact of learning the latent distribution. Analogously to the previous experiments, we compare our algorithm against the standard GAN approach, with latent distribution fixed and equal to the Gaussian measure $\eta_0 = \mathcal{N}([0.0, 0.0], I)$. We consider two options for the family $\mathcal{T}$ of candidate generators from $\mathbb{R}^2$ to $\mathbb{R}^3$ with increasing complexity. We aim to show that the joint GAN estimator can efficiently learn the target distribution while the standard GAN algorithm requires a significantly larger space of generators. We show results corresponding to two architectures for learning the generator.

2-*layers generator* : we considered $\mathcal{T}$ the space of neural networks from $\mathbb{R}^2$ to $\mathbb{R}^3$ with 2 hidden layers of dimensions 128 and respectively ReLu and Tanh activation functions. Fig. 5.4 (top row) reports the results for GAN training in this setting for the standard GAN (Fig. 5.4a) and the proposed estimator (Fig. 5.4b). When keeping the latent distribution fixed, the generator is not able to match the target. In contrast, by applying Alg. 5.1 to learn the latent distribution, part of the complexity of modeling ρ is offloaded to η and therefore the final estimator is able to approximately match the target. This is visually reported in

Fig. 5.4c, which shows a sample from the learned latent distribution $\hat{\eta}$. It can be noticed that such distribution is significantly different from the Gaussian measure used for standard GAN training. As a result, the generalization gaps (see section Sec. 5.5) are respectively 0.0003 for Alg. 5.1 and 0.0311 when keeping the latent distribution fixed.

4-*layers generator*: we considered $\mathcal{T}$ the space of neural networks from $\mathbb{R}^2$ to $\mathbb{R}^3$ with 4 hidden layers with dimensions $128, 512, 512, 128$ and ReLu activation functions for layers 1,3 and Tanh for layers 2,4. Fig. 5.4 (bottom row) reports the results of GAN training for the standard estimator with fixed latent distribution η_0 (Fig. 5.4d) and the estimator from Alg. 5.1 (Fig. 5.4e). We note that the standard GAN method is still not able to correctly match the target distribution but is incurring in a smaller error. The two methods achieve generalization gaps respectively 0.0003 for Alg. 5.1 and 0.0209 when keeping the latent distribution fixed. We note that estimator from Alg. 5.1 is achieving similar qualitative and quantitative performance to the one obtained using a simpler space of generators $\mathcal{T}$. However we observe a key difference in Fig. 5.4f, which reports a sample from the estimated latent distribution $\hat{\eta}$. Since the generator class is larger, it allows to apply more distortion to latent distributions. As a consequence, the latent distribution can have a less sharp support shape and still realize a good matching.

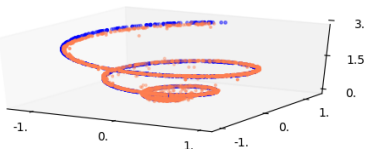


Figure 5.5: result with standard GAN with $\mathcal{T}$ a space of generators with 6 layers with 512 dimensions.

Deeper models. In our experiments, the standard GAN was able to match the target distribution only when $\mathcal{T}$ allowed for deeper networks (Fig. 5.5 shows this for $\mathcal{T}$ a space of networks up to 6 layers with 512 neurons each and ReLU activation functions). This is in line with the intuition in Sec. 5.2 and the theoretical analysis in this chapter: while choosing a fixed latent distribution still allows to recover the target probability, doing so might impose tight requirements on the complexity of the space of generators that needs to be considered.

Dependence on the dimensions of the latent and ambient space. Finally, we conclude with an example that highlights the dependence on the ambient space and latent space in terms of generalization error in practice. When testing the dependence on the

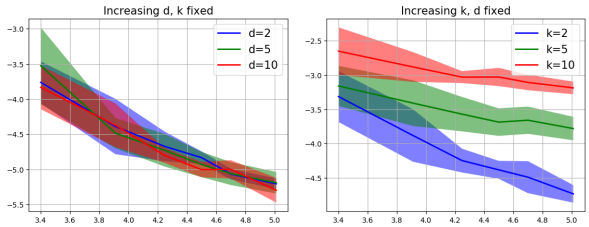


Figure 5.6: Impact of the latent dimension and the ambient dimension on the statistical performance of an estimator of the form $T_{\#}\eta$. The plots display on a log-log scale how the generalization error (y-axis) decreases when the number of training points (x-axis) increases.

ambient space d: we considered as target ρ a distribution a 2-dimensional spiral-shaped manifold embedded in $\mathbb{R}^d$. We consider $n \in \{30, 50, 70, 90, 110, 150, 200, 500\}$ training points and we aim to display how the behaviour of the generalization error as n increases is affected by the latent and ambient dimensions. For $d = 10$ fixed, we show that the generalization error deteriorates as k increases, namely for $k = 2, 5, 10$ (see Fig. 5.6 (right)). On the other hand, we keep $k = 2$ fixed and consider $d = 2, 5, 10$, i.e. we increase the dimension of the ambient space. In this setting, increasing the dimension does not impact the performance (see Fig. 5.6 (left)).

5.6 Discussion

In this chapter we studied the role of pushforward maps (generators) and latent distributions in a generative modelling framework with Sinkhorn divergence, from a theoretical perspective. We characterized the learning rates of the estimator in terms of the complexity (i.e. smoothness) of the class of generators. We introduced a novel 'GAN' estimator that jointly learns both latent distribution generator, studied its generalization properties and proposed a practical algorithm to train it. We performed some toy experiments that were conceived as a proof-of-concept of the ideas and intuition underlying the proposed method. Future work will focus on two main directions. First, we plan to investigate more empirically oriented questions related to our framework. In particular, we plan to evaluate our approach on large scale real data, to test the limits and benefits of the proposed strategy in practice. Secondly, on a more theoretical direction, we plan to extend our analysis to a larger family of adversarial divergences and generator networks.

Chapter 6

Conclusion and future directions

We end the thesis with some concluding remarks and potential future directions. Optimal Transport is an elegant theory at the intersection of probability, analysis and geometry and has proved to be a useful tool in a variety of applications. The Entropy-regularized variant was first popularized in the machine learning community as a computational tool to solve Optimal Transport problem efficiently up to some approximation. However, the computational efficiency has paved the way to numerous applications and has motivated further research on theoretical properties and advantages of Entropic Optimal Transport. The thesis fits in this line of research. The scope of the thesis was to study the regularity induced by the entropy penalty at different levels and to leverage this regularity when using Entropic Optimal Transport in different machine learning problems, to design estimators and algorithms with provable theoretical guarantees. We focused on supervised learning, barycenter estimation and density matching.

Supervised Learning. We used Entropic Optimal Transport as a loss in supervised learning settings with the simplex as output space. We proved high order differentiability of Entropic OT which extends standard results on *first* order differentiability. We leveraged such smoothness to design an estimator for learning with Sinkhorn loss that is consistent and with provable learning rates.

Barycenter estimation. We used Sinkhorn divergence as metric in the barycenter problem and proposed the first algorithm that computes both the weights and the atoms, with provable convergence guarantees. The algorithm is based on Frank-Wolfe procedure. The algorithm does not invoke an alternating minimization scheme as in previous literature, but proceeds by

adding a new atom at each iteration and by suitably rescaling the weights. We proved new regularity properties of the Sinkhorn divergence that were necessary to apply the Frank-Wolfe paradigm, such as the Lipschitz continuity of the gradient. We showed convergence rates for both discrete input measures and continuous input measures.

Density matching. We used the Sinkhorn divergence as a metric when learning a probability measure in a unsupervised fashion. Our goal was to understand how properties of the generator and the latent space potentially impact the statistical performance of the estimator parametrized as pushforward of such latent distribution and generator. We showed that within the modelling range that is able to capture the target distribution, using a lower latent dimension is beneficial. Based on our analysis, we proposed an estimator that learns both the generator and the latent distribution and we studied the statistical performance of such estimator.

6.1 Future directions

We first discuss potential further work which is tightly related to the material presented in this thesis and then discuss broader future questions.

Future directions inspired by Chapter 3. *Supervised learning in infinite dimensional settings.* The distributional regression setting considered in this work deals with probability measures over a finite set. This is the framework originally proposed in the first work considering Optimal Transport distances as losses in supervised learning setting and captures various applications. However, being able to deal with general measures rather than histograms would greatly improve the flexibility of the paradigm. An open question is then how to design an estimator with provable theoretical guarantees that is suitable to infinite dimensional settings. This would allow to go beyond histograms and could provide a more flexible and realistic approach suitable to a wider range or applications.

This is a hard problem on multiple levels; bridging the gap from the finite dimensional case to the space of arbitrary probability distribution is ambitious in terms of both the evaluation of the losses and the optimization pipelines. We believe that designing a framework for supervised learning with Sinkhorn loss that successfully supports infinite dimensional spaces both in the algorithmic modelling and in the statistical analysis constitutes an interesting and important direction for future research.

Future directions inspired by Chapter 4. *Potential improvement of the constants in the Lipschitz smoothness of Sinkhorn gradients.* In Chapter 4 we proved Lipschitz continuity of Sinkhorn gradients with respect to MMD and Total Variation. The Lipschitz constant $e^{\mathsf{D}/\varepsilon}$ where D is the diameter of the domain, has a bad dependence on ε. At the time of publication of (Luise et al., 2019), which Chapter 4 is based on, the constant was in line with the ones in Genevay et al. (2018a) regarding the sample complexity of Sinkhorn divergence. However, for sample complexity results the dependence on ε was improved from exponential to polynomial in the subsequent work (Mena and Niles-Weed, 2019). It is not clear whether such improvement could hold in this case too. Our current analysis based on Perron-Frobenius theory does not allow to eliminate the exponential dependence and an interesting question is whether exploring other approaches can lead to better constants. The very recent work (Shen et al., 2020) improves minor details of our results but does not remove the exponential dependence. A relevant question would be to determine whether the exponential dependence is tight in this case or to understand how to remove it.

Explore variants of Frank-Wolfe algorithm that could achieve better rates. In Chapter 4, we considered the vanilla version of Frank-Wolfe algorithm, with oblivious step-size $2/(2+k)$ where k is the present iteration. However, there exists an extensive literature on Frank-Wolfe variants (Lacoste-Julien and Jaggi, 2015) that can achieve better rates under specific assumptions. Studying those variants both in theory and practice for the Sinkhorn barycenter problem would be a natural extension of the work on barycenters presented in this thesis.

Extend the results to unbalanced Sinkhorn divergences. In this thesis we considered Entropic Optimal Transport between probability measures, i.e. between measures with the same mass. However, the formulations of Optimal Transport, its entropic version and the Sinkhorn divergence can be extended to the unbalanced setting of arbitrary positive finite measures, i.e. measures are not required to have the same mass (equal to one). Unbalanced Optimal Transport is an extension of standard Optimal transport that is particularly suited for settings where creation/ destruction of mass is natural, such as frameworks involving birth/death dynamics or populations of cells (Schiebinger et al., 2019). Future work will be devoted to exploring whether the results in Chapter 4 on the smoothness of the gradients hold for the unbalance case as well. This would allow to study the barycenter problem and the proposed

algorithm in the unbalanced setting.

Future directions inspired by Chapter 5. Chapter 5 is the most exploratory among the chapters in the thesis. While it frames the main ideas and motivations, it leaves several open questions. *Weaken the regularity assumptions in the main results.* In theorem Thm. 5.2 we assume high order differentiability of the generator T. This happens because in the proof presented we need to differentiate and to estimate the norm of those derivatives. It would be interesting to see whether it is possible to relax this assumption and to ask for *weak* derivatives only. Allowing T to be in a Sobolev space would increase the generality of the result and to cover standard activation functions in GANs generators such as ReLu.

Better statistical bounds for smaller classes of distributions. In our analysis, we leveraged state-of-the-art results on sample complexity of Sinkhorn divergences, recently proved in Mena and Niles-Weed (2019). The results are quite general, holding for any subgaussian measure. However, this generality is paid in terms of constants in the bounds: in particular, the leading term has the form $(\frac{\sigma}{\varepsilon})^{5k/2}$ where k is the dimension of the domain and σ is the subGaussianity. This quantity rapidly grows as k increases. Adding stronger requirements on the distributions (e.g. restricting the attention to measures with smooth densities as in Weed and Berthet (2019)) may lead to refined and more informative bounds.

6.2 Broader questions

Optimal Transport has received increasing attention in machine learning in the past decade and there are countless aspects that have not been discussed in this thesis. The flexibility of Optimal Transport metrics makes them a great tool in many settings. In particular, while it is not to be considered as the panacea, Optimal Transport distances can be a powerful option in a variety of tasks: as a loss function for distributional learning and as a metric for barycenters, as considered in this thesis; but also as a notion to transfer information across different domains, exploited in domain adaptation, and as geometric structure to equip the space of probability measures with. This latter direction is relevant when optimizing on the space of probability distributions, e.g. for sampling purposes. If on one side Optimal Transport has great potential, on the other there are key fundamental aspects that still have to be understood. We conclude the work with open questions that are not directly related to the material presented but we believe are crucial for developments of Optimal Transport for

Machine Learning.

Learning the cost function. In this thesis, we either developed theoretical results in full generality (e.g. assuming the domain $\mathcal{X}$ to be a metric space with basic assumptions and the cost function c to satisfy some mild requirements) without focusing on any specific application, or we considered subsets of the Euclidean space equipped with standard Euclidean norm as a show-case. In most applications -not covered or mentioned in this work, but deeply relevant to the developments of the interaction between Optimal Transport and Machine Learning- the cost function plays a fundamental role and simply sticking to a norm-related quantity cannot lead to acceptable results in practice. One example among all concerns images: there is no standard and effective way to define a distance between pairs of images. Considering images as sufficiently high-dimensional vectors and using standard Euclidean norm between them does not capture any notion of shape and texture which are however crucial. Thus, the geometric flavour of Optimal Transport and its regularized variants is a double-edged sword. Optimal Transport metric faithful incorporates the ground geometry of the underlying domain, but what if the choice of the ground cost itself is not trivial? In some cases, the ground geometry needs to be properly engineered for each specific problem and becomes part of the problem itself. Relying on OT metrics without having access to a sensible cost function can lead to poor results. Therefore, a very important and broad question concerns how to design a ground metric which is meaningful for a given task. While an option is to learn adversarially a deep embedding of the data, many aspects on this question are still open and unexplored.

Large scale settings. While the entropic regularization improves the computational complexity of Optimal Transport distances, in many large scale setting the computational cost is still a burden. The most immediate solution is to rely on mini-batches. However, gradients computed on batches are biased estimator of the actual gradients and therefore no theoretical justification can back-up the mini-batch trick. A crucial question in this regard consists in understanding how to make the use of batches theoretically sound. Simultaneously, exploring computational methods that can efficiently deal with large scale settings and small regularization parameter is also of fundamental importance.

Dealing with high dimensions. The curse of dimensionality of Wasserstein distance is alleviated by the entropy penalty: while for unregularized Optimal Transport the dependence on the dimension appears in the rate of convergence, Sinkhorn divergences display a better rate $(n^{-1/2})$ and suffer the dependence on dimensionality in terms of constants. However, the constants degrades as the regularization decreases and this poses challenges on how to choose the parameter; how to optimally select the regularization is still an open question and is application dependent. A large regularization is more efficient on the computational side and probably on statistical sides as well; however, the geometry on the space of probability measures induced by highly regularized Sinkhorn divergences becomes 'flatter' and less discriminative, undermining some of the reasons why Optimal Transport metrics may be favourable in the first place. Alternatives to Entropy regularization have been recently proposed, for example in the line of works on Sliced Optimal Transport distances. Overall, it seems that being able to combine good geometric properties with efficient computational costs and effective statistical features is still an open challenge. Entropic Optimal Transport is a first elegant answer but in its original definition, probability not the final one. More research in this direction will be precious for future developments.

Appendix A

Appendix of Background material

In this Appendix we cover some material -useful to understand the results in the thesis- that was not included in Chapter 2. In particular: in Appendix A.1 we recall some basic concepts on duality of convex functions and introduce the definitions that are necessary to state the Fenchel-Rockafellar theorem. In Appendix A.2 we briefly introduce two important families of divergences between probability measures, namely f-divergences and Integral Probability Metrics, which includes the Maximum Mean Discrepancy (MMD). Related to MMD, in Appendix A.3 we introduce some definitions and concepts on reproducing kernels and Reproducing Kernel Hilbert Spaces, that are used in a few results in this thesis. In Appendix A.4 we introduce Hilbert metric, its properties and all the technical lemmas and details needed to prove the existence of solutions of the dual formulation of Entropic Optimal Transport (2.3.3).

A.1 Useful concepts and Fenchel-Rockafellar theorem

In this section we introduce the tools that are necessary to state the Fenchel-Rockafellar theorem. Let V be a Banach space.

Definition A.1. *The set of functions $f : V \to \bar{\mathbb{R}}$ that are convex, proper and lower semicontinuous is denoted by $\Gamma_0(V)$.*

Definition A.2. *Let V be a Banach space. The topological dual of V, denoted by V^*, is the space of all continuous linear functionals on V.*

To state the Fenchel-Rockafeller theorem we need the notion of pairing (or duality), which is recalled below:

Definition A.3. *Two Banach spaces V and W are said* topologically paired *if all continuous linear functionals on one space can be identified with the elements of the other, and vice-versa.*

Formally, V and W are topologically paired if there is a bilinear form $\langle \cdot, \cdot \rangle : V \times W \to \mathbb{R}$ that separates points, i.e. (i.e. $\forall v \neq 0,\ \exists w \in W$ such that $\langle v, w \rangle \neq 0$ and $\forall w \neq 0\ \exists v \in V$ such that $\langle v, w \rangle \neq 0$) .

Note that a Banach space V (equipped with the strong or weak topology) and its topological dual V^* equipped with the weak* topology are topologically paired. As an example, relevant to our setting, consider the following: let $\mathcal{X}$ be a compact metric space and denote by $\mathcal{C}(\mathcal{X})$ the Banach space of continuous functions on $\mathcal{X}$ endowed with the norm of uniform convergence. We denote by $\mathcal{M}(\mathcal{X})$ the space of (signed) Borel measures on $\mathcal{X}$ endowed with the norm of total variation (see (A.2.7)), which is the dual of $\mathcal{C}(\mathcal{X})$. Then $\mathcal{C}(\mathcal{X})$ and $\mathcal{M}(\mathcal{X})$ are topologically paired, with the following duality:

$$\langle f, \mu \rangle = \int_{\mathcal{X}} f \, d\mu, \qquad \text{for any } f \in \mathcal{C}(\mathcal{X})\ \mu \in \mathcal{M}(\mathcal{X}). \tag{A.1.1}$$

Two other notions that we need are the Fenchel-Legendre conjugate and the definition of adjoin operator which are recalled below.

Definition A.4. *(Fenchel-Legendre conjugate) Le $f \in \Gamma_0(V)$. The conjugate $f^* : V^* \to \mathbb{R} \cup \{+\infty\}$ is defined as follows: for any $v^* \in V^*$*

$$f^*(v^*) := \sup_{v \in V} \langle v, v^* \rangle - f(v). \tag{A.1.2}$$

Definition A.5 (Adjoint operator)**.** *Let (V, V^*) and (W, W^*) be two pairs of topologically paired spaces and $A : V \to W$ be a continuous linear operator. For all $w^* \in W^*$, the application $V \ni v \mapsto \langle Av, w^* \rangle$ is a continuous linear form on V and hence admits a representer in V^* that we denote $A^* w^*$. This uniquely defines an operator $A^* : W^* \to V^*$ that is called* adjoint *of A.*

With these tools we can now state the Fenchel-Rockafellar theorem. We refer to (Rockafellar, 1974) for the proof and a more general statement.

Theorem A.1. *[Fenchel-Rockafellar] Let (V, V^*) and (W, W^*) be two pairs of topologically paired spaces. Let $f \in \Gamma_0(V)$ and $g \in \Gamma_0(W)$ and $A : V \to W$ be a continuous linear operator with adjoint $A^* : W^* \to V^*$. The so-called qualification constraint is the property: there exists $v \in \mathrm{dom}(f)$ such that g is continuous at Av . If the qualification constraint holds, then*

$$\sup_{v \in V} -f(-v) - g(Av) = \min_{w^* \in W^*} f^*(A^* w^*) + g^*(w^*), \tag{A.1.3}$$

and the min is attained. Moreover, if the minimum is finite, $(v, w^*) \in V \times W^*$ *is a couple of optimizers if and only if* $Av \in \partial g^*(w^*)$ *and* $A^*w^* \in \partial f(-v)$.

A.2 Comparing probability measures

In this section we review common discrepancies between probability measures. Chapter 2 is entirely dedicated to Optimal Transport and Sinkhorn divergences, while here we introduce f-divergences and Maximum Mean Discrepancy, which are widely used in machine learning and data science. We recall the definition of weak convergence of probability measures, that will be often mentioned in the following section.

Definition A.6. *Let* $\mathcal{X}$ *be a measurable space. A sequence* $(\alpha_n)_n$ *with* $\alpha_n \in \mathcal{P}(\mathcal{X})$ *for all* n *weak-converges to* $\alpha \in \mathcal{P}(\mathcal{X})$ *(denoted by* $\alpha_n \rightharpoonup \alpha$*) if*

$$\int_{\mathcal{X}} f(x)\, d\alpha_n(x) \to \int_X f(x)\, d\alpha(x), \qquad \text{for all} f \in \mathcal{C}_b(\mathcal{X}) \tag{A.2.1}$$

where $\mathcal{C}_b(\mathcal{X})$ *is the space of continuous and bounded functions over* $\mathcal{X}$.

A divergence D *on* $\mathcal{P}(\mathcal{X})$ *is said to* metrize *the weak convergence if*

$$D(\alpha_n, \alpha) \to 0 \qquad \Longleftrightarrow \qquad \alpha_n \rightharpoonup \alpha. \tag{A.2.2}$$

A.2.1 f-divergences

f-divergences are a class of divergences routinely used in many applications. They compare two input measures by comparing their mass pointwise, without using any notion of mass transportation. The pointwise comparison is based on the pointwise ratio between two measures, that gives a sense of how close they are. The definition of f-divergences builds upon the entropy functionals which are recalled below.

Definition A.7. *(Entropy functional) A function* $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *is an entropy functional if it is convex, lower-semicontinuous and* $f(1) = 0$*. The speed of growth of* f *at infinity is given by*

$$f'_\infty := \lim_{x \to \infty} \frac{f(x)}{x} \in \mathbb{R} \cup \{\infty\}. \tag{A.2.3}$$

An entropy functional f induces an f-divergence as follows:

Definition A.8. *Let* f *be an entropy functional. For* $\alpha, \beta \in \mathcal{M}(\mathcal{X})$*, let* $\frac{d\alpha}{d\beta}\beta + \alpha^\perp$ *be the Lebesgue decomposition of* α*. Then, the* f*-divergence* D_f *associated to the entropy*

functional f is defined as

$$D_f(\alpha, \beta) = \int_X f\Big(\frac{d\alpha}{d\beta}\Big) d\beta + \alpha^{\perp} f'_{\infty}, \tag{A.2.4}$$

if α and β are nonnegative and $+\infty$ otherwise.

If f has a superlinear growth, i.e. $f'_{\infty} = \infty$, then $D_f(\alpha, \beta) = \infty$ when α does not admit a density with respect to β. We provide two well-known examples of f-divergences below.

Kullback-Leibler divergence. One of the best-known examples of f-divergences is the Kullback-Leibler divergence, defined as

$$D_{\mathsf{KL}}(\alpha, \beta) = \mathsf{KL}(\alpha \mid \beta) = \int_{\mathcal{X}} \log\Big(\frac{d\alpha}{d\beta}\Big) d\alpha - \int_{\mathcal{X}} d\alpha + \int_{\mathcal{X}} d\beta, \tag{A.2.5}$$

which is associated to the following function

$$f_{\mathsf{KL}}(x) = \begin{cases} x \log x - x + 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{A.2.6}$$

The entropy function f_{KL} has superlinear growth, meaning that $f'_{\infty} = \infty$. This means that when α and β do not share the same support, i.e. α does not admit a density with respect to β, then $\mathsf{KL}(\alpha \mid \beta)$ is equal to ∞. For instance, consider $\alpha_n = \delta_{1/n}$ and $\beta = \delta_0$. Since α_n and β have disjoint support, $D_{\mathsf{KL}}(\alpha_n, \beta) = \infty$ for any n, although the support of α_n gets closer to the support of β as n goes to infinity. This is illustrative of the fact that the Kullback-Leibler divergence does not metrize the weak convergence.

Total Variation. Another well-known f-divergence is the Total Variation, defined as

$$\mathsf{TV}(\alpha, \beta) = D_{\mathsf{TV}}(\alpha, \beta) = 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\alpha(A) - \beta(A)|, \tag{A.2.7}$$

where $\mathcal{B}(\mathcal{X})$ denotes the set of measurable subsets of $\mathcal{X}$. D_{TV} is associated to the entropy functional

$$f_{\mathsf{TV}}(x) = \begin{cases} |x - 1| & \text{if } x \geq 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{A.2.8}$$

Total Variation defines a norm on the space of measures $\mathcal{M}(\mathcal{X})$ as follows

$$\mathsf{TV}(\alpha, \beta) = \|\alpha - \beta\|_{\mathsf{TV}}, \qquad \|\alpha\|_{\mathsf{TV}} = |\alpha|(\mathcal{X}) = \int_{\mathcal{X}} d|\alpha|, \tag{A.2.9}$$

where $|\alpha| = \alpha^+ + \alpha^-$, with α^+ and α^- the two nonnegative measures that constitute the Hahn-Jordan decomposition (Bogachev, 2007, Def. 3.1.4) of α, i.e. $\alpha = \alpha^+ - \alpha^-$. Note that f_{TV} has linear growth at infinity and $f'_\infty = 1$. Hence, unlike KL, D_{TV} is not infinity when α is not absolutely continuous with respect to β. However, in the examples with $\alpha_n = \delta_n$ considered before, we have that $D_{\mathsf{TV}}(\alpha_n, \beta)$ is constant; this means that Total Variation also fails in metrizing the weak convergence. Total Variation also belongs to another class of metrics called Integral Probability Metrics (IPM), which are another family of divergences between probability measures. Under some specific assumptions IPM can metrize weak convergence, as briefly discussed below.

A.2.2 Integral Probability Metrics

Integral Probability Metrics (Müller, 1997) are another class of widely-used discrepancies between probability distributions. We recall the definition below.

Definition A.9 (Integral Probability Metric)**.** *Let $\mathcal{X}$ be a measurable space. For $\alpha, \beta \in \mathcal{P}(\mathcal{X})$, and Integral Probability Metric IPM is defined as*

$$\gamma_{\mathcal{F}}(\alpha, \beta) = \sup_{f \in \mathcal{F}} \Big[\int_{\mathcal{X}} f(x)\, d\alpha(x) - \int_{\mathcal{X}} f(x)\, d\beta(x) \Big], \tag{A.2.10}$$

where $\mathcal{F}$ is a space of real-valued bounded measurable functions on $\mathcal{X}$.

The function class $\mathcal{F}$ fully characterizes the IPM $\gamma_{\mathcal{F}}(\alpha, \beta)$. When choosing $\mathcal{F}$, there are two important aspects: on one hand, the function class must be rich enough so that $\gamma_{\mathcal{F}}(\alpha, \beta)$ vanishes if and only if $\alpha = \beta$, since this is a feature that is natural when we want to define a notion of similarity. On the other hand, the larger the function class $\mathcal{F}$, the harder it is to estimate $\gamma_{\mathcal{F}}(\alpha, \beta)$. For example, when $\mathcal{F} = \mathcal{C}_b(\mathcal{X})$, then the corresponding IPM is a metric on the space of probability measures (Muandet et al., 2017, Thm. 3.5), but handling $\mathcal{C}_b(\mathcal{X})$ is very difficult in practice. Necessary and sufficient conditions on the class $\mathcal{F}$ to ensure that $\gamma_{\mathcal{F}}$ metrizes the weak convergence are discussed in (Müller, 1997).

Examples. One example of IPM is the Total variation, which is the only metric that is both an f-divergence and an Integral Probability Metric (Sriperumbudur et al., 2009). The

corresponding space $\mathcal{F}$ is given by

$$\mathcal{F} = \{f \in \mathcal{C}_b(\mathcal{X}) : \ \|f\|_\infty \leq 1\}. \tag{A.2.11}$$

Another well-known example is Wasserstein-1 distance, with

$$\mathcal{F} = \{f \in \mathcal{C}_b(\mathcal{X}) : \ \|f\|_{\text{Lip}} \leq 1\}, \tag{A.2.12}$$

that is the space of Lipchitz functions with Lipschitz constant smaller than 1.

Besides the examples above, a fundamental class of IPMs is given by choosing $\mathcal{F}$ as a unit ball in a Reproducing Kernel Hilbert Space. The resulting IPM is called Maximum Mean Discrepancy (MMD) (Gretton et al., 2007). MMDs are among the most important IPMs in machine learning. We recall the definition in the next section, where we also provide some background on kernels and RKHS that it is useful to read the manuscript.

A.3 Reminders on Kernels and MMD

This is an introductory section where we list basic definitions of concepts that are mentioned throughout the thesis. It is to be interpreted as a minimal glossary, by no means exhaustive.

Definition A.10 (Kernel)**.** *Let $\mathcal{X}$ be a nonempty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exist a Hilbert space $\mathcal{H}$ and a function $\varphi : \mathcal{X} \to \mathcal{H}$ such that*

$$k(x, x') = \big\langle \varphi(x), \varphi(x') \big\rangle_{\mathcal{H}} \qquad \text{for any } x, x' \in \mathcal{X}. \tag{A.3.1}$$

The function φ is referred to as feature map.

Definition A.11 (Reproducing kernel Hilbert Space)**.** *Let $\mathcal{X}$ a nonempty set and $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is called a Reproducing kernel Hilbert Space if all the evaluation functionals F_x defined by $\mathsf{F}_x(f) = f(x)$ are bounded, i.e. for all $x \in \mathcal{X}$ there exists some $C > 0$ such that*

$$|\mathsf{F}_x(f)| = |f(x)| \leq C\|f\|_{\mathcal{H}} \qquad \text{for all } f \in \mathcal{H}. \tag{A.3.2}$$

Proposition A.2 (Reproducing property)**.** *Let $\mathcal{H}$ be an RKHS. For each $x \in \mathcal{X}$ there exists a function $k_x \in \mathcal{H}$ such that $\mathsf{F}_x(f) = f(x) = \langle f, k_x \rangle$.*

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a two-variables function defined by $k(x, y) := k_y(x)$. Then, by

the reproducing property it follows that $k(x,y) = k_y(x) = \langle k_x, k_y \rangle = \langle \varphi(x), \varphi(y) \rangle$, where $\varphi(x) := k_x$ is the feature map $\varphi : \mathcal{X} \to \mathcal{H}$.

Definition A.12 (Universal kernel)**.** *Let $\mathcal{C}_b(\mathcal{X})$ be the space of continuous bounded functions on a compact metric space $\mathcal{X}$. A continuous positive definite kernel k on $\mathcal{X}$ is said to be universal if the corresponding RKHS $\mathcal{H}$ is dense in $\mathcal{C}_b(\mathcal{X})$, i.e., for any $f \in \mathcal{C}_b(\mathcal{X})$ and $\varepsilon > 0$, there exists a function $h \in \mathcal{H}$ such that $\|f - h\|_\infty \leq \varepsilon$.*

Definition A.13 (Kernel Mean Embedding)**.** *Let $\mathcal{X}$ be a measurable space and $\mathcal{P}(\mathcal{X})$ denote the set of probability measures over $\mathcal{X}$. The kernel mean embedding of probability measures in $\mathcal{P}(\mathcal{X})$ into an RKHS $\mathcal{H}$ associated to a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as*

$$\mathsf{M} : \mathcal{P}(\mathcal{X}) \to \mathcal{H}, \qquad \alpha \longmapsto \int_{\mathcal{X}} k(\cdot, x)\, d\alpha(x). \tag{A.3.3}$$

Definition A.14 (Maximum Mean Discrepancy)**.** *Let $\mathcal{H}$ be a RKHS. Then, for $\alpha, \beta \in \mathcal{P}(\mathcal{X})$,*

$$\mathsf{MMD}(\mathcal{H}; \alpha, \beta) = \sup_{\{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}} \Big[\int_{\mathcal{X}} f(x)\, d\alpha(x) - \int_{\mathcal{X}} f(x)\, d\beta(x) \Big]. \tag{A.3.4}$$

Note that MMD can be expressed in terms of mean embeddings, namely

$$\mathsf{MMD}(\mathcal{H}; \alpha, \beta) = \|\mathsf{M}_\alpha - \mathsf{M}_\beta\|_{\mathcal{H}}, \tag{A.3.5}$$

or equivalently using the correspondent kernel k

$$\mathsf{MMD}^2(\mathcal{H}; \alpha, \beta) = \mathbb{E}_{\alpha \otimes \alpha}(k(x, x')) + \mathbb{E}_{\beta \otimes \beta}(k(y, y')) - 2\mathbb{E}_{\alpha \otimes \beta}(k(x, y)). \tag{A.3.6}$$

A.4 Hilbert metric and existence of dual potentials

In this section we present the proof of existence of dual potentials, which are solutions of the dual formulation of the Entropic Optimal Transport problem presented in Thm. 2.3. The proof needs some preliminary material, including definition and properties of the Hilbert metric, that are also necessary for other proofs in the thesis, mainly in Chapter 4. Some results are quite technical and we present the proofs for completeness.

A.4.1 Hilbert's metric and the Birkhoff-Hopf theorem

We first review the basic concepts of the nonlinear Perron-Frobenius theory (Lemmens and Nussbaum, 2012) which provides tools to deal with DAD problems and ultimately to study existence of dual potentials.

We consider $\mathcal{X} \subset \mathbb{R}^d$ to be a compact set. We denote by $\mathcal{C}(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$ endowed with the sup norm, namely $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. Let $\mathcal{C}_+(\mathcal{X})$ be the cone of non-negative continuous functions, that is, $\mathcal{C}_+(\mathcal{X}) := \{f \in \mathcal{C}(\mathcal{X})$ such that $f(x) \geq 0,$ for every $x \in \mathcal{X}\}$. Also, we denote by $\mathcal{C}_{++}(\mathcal{X})$ the set of continuous and (strictly) positive functions on $\mathcal{X}$, $\mathcal{C}_{++}(\mathcal{X}) := \{f \in \mathcal{C}(\mathcal{X})$ such that $f(x) > 0,$ for every $x \in \mathcal{X}\}$, which turns out to be the interior of $\mathcal{C}_+(\mathcal{X})$.

Recall that $\mathsf{c} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a positive, symmetric, and continuous function and define $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{++}$ as

$$\forall x, y \in \mathcal{X}, \qquad \mathsf{k}(x, y) = e^{-\frac{\mathsf{c}(x,y)}{\varepsilon}}. \tag{A.4.1}$$

Set $\mathsf{D} = \sup_{x,y\in\mathcal{X}} \mathsf{c}(x, y)$. Then, we have $\mathsf{k}(x, y) \in [e^{-\mathsf{D}/\varepsilon}, 1]$ for all $x, y \in \mathcal{X}$. Let $\alpha \in \mathcal{P}(\mathcal{X})$. Define the operator $\mathsf{L}_\alpha \colon \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ as

$$\forall f \in \mathcal{C}(\mathcal{X}), \qquad \mathsf{L}_\alpha f \colon x \mapsto \int \mathsf{k}(x, z) f(z)\, d\alpha(z). \tag{A.4.2}$$

Note that L_α is linear and continuous. In particular, since $\mathsf{k}(x, y) \in [0, 1]$ for all $x, y \in \mathcal{X}$, we have

$$\forall\, f \in \mathcal{C}_+(\mathcal{X}), \qquad \mathsf{L}_\alpha f \geq 0 \tag{A.4.3}$$

and

$$\forall\, f \in \mathcal{C}(\mathcal{X}), \qquad \|\mathsf{L}_\alpha f\|_\infty \leq \|f\|_\infty. \tag{A.4.4}$$

Hilbert's Metric. The cone $\mathcal{C}_+(\mathcal{X})$ induces a partial ordering $\leq$ on $\mathcal{C}(\mathcal{X})$, such that

$$\forall\, f, g \in \mathcal{C}(\mathcal{X}), \qquad f \leq g \Leftrightarrow\ g - f \in \mathcal{C}_+(\mathcal{X}). \tag{A.4.5}$$

According to (Lemmens and Nussbaum, 2012), we say that a function $g \in \mathcal{C}_+(\mathcal{X})$ *dominates* $f \in \mathcal{C}(\mathcal{X})$ if there exist $t, s \in \mathbb{R}$ such that

$$tg \leq f \leq sg. \tag{A.4.6}$$

This notion induces an equivalence relation on $\mathcal{C}_+(\mathcal{X})$, denoted $f \sim g$, meaning that f dominates g and g dominates f. The corresponding equivalence classes are called *parts* of $\mathcal{C}_+(\mathcal{X})$. Let $f, g \in \mathcal{C}_+(\mathcal{X})$ be such that $f \sim g$. We define

$$M(f/g) = \inf\{s \in \mathbb{R} \,|\, f \leq sg\} \qquad \text{and} \qquad m(f/g) = \sup\{t \in \mathbb{R} \,|\, tg \leq f\}. \quad \text{(A.4.7)}$$

Note that $m(f/g) \leq M(f/g)$. Moreover, for every $f, g \in \mathcal{C}_+(\mathcal{X})$ such that $f \sim g$, we have that $\text{supp}(f) = \text{supp}(g)$ and if $g \neq 0$ (hence $f \neq 0$), then

$$M(f/g) = \max_{x \in \text{supp}(g)} \frac{f(x)}{g(x)} > 0 \quad \text{and} \quad m(f/g) = \min_{x \in \text{supp}(g)} \frac{f(x)}{g(x)} > 0. \quad \text{(A.4.8)}$$

Definition A.15. *With the notation introduced above, the Hilbert's metric is defined as*

$$d_H(f, g) = \log \frac{M(f/g)}{m(f/g)}, \quad \text{(A.4.9)}$$

for all $f \sim g$ with $f \neq 0$ and $g \neq 0$, $d_H(0, 0) = 0$ and $d_H(f, g) = +\infty$ otherwise.

Direct calculation shows that (Lemmens and Nussbaum, 2013, Proposition 2.1.1)

(i) $d_H(f, g) \geq 0$ and $d_H(f, g) = d_H(g, f)$, for every $f, g \in \mathcal{C}_+(\mathcal{X})$;

(ii) $d_H(f, h) \leq d_H(f, g) + d_H(g, h)$, for every $f, g, h \in \mathcal{C}_+(\mathcal{X})$ with $f \sim g$ and $g \sim h$;

(iii) $d_H(sf, tg) = d_H(f, g)$, for every $f, g \in \mathcal{C}_+(\mathcal{X})$ and $s, t > 0$.

Note that d_H induces a metric on the rays of the parts of $\mathcal{C}_+(\mathcal{X})$ (Lemmens and Nussbaum, 2013, Lemma 2.1).

We now focus on $\mathcal{C}_{++}(\mathcal{X})$. A direct consequence of Hilbert's metric properties is the following.

Lemma A.3 (Hilbert's Metric on $\mathcal{C}_{++}(\mathcal{X})$)**.** *The interior of $\mathcal{C}_+(\mathcal{X})$ corresponds to the set of (strictly) positive functions $\mathcal{C}_{++}(\mathcal{X})$ and is a part of $\mathcal{C}_+(\mathcal{X})$ with respect to the equivalence relation induced by dominance. For every $f, g \in \mathcal{C}_{++}(\mathcal{X})$,*

$$M(f/g) = \max_{x \in \mathcal{X}} \frac{f(x)}{g(x)} \qquad m(f/g) = \min_{x \in \mathcal{X}} \frac{f(x)}{g(x)}, \quad \text{(A.4.10)}$$

and $M(f/g) \geq m(f/g) > 0$. Therefore

$$d_H(f, g) = \log \max_{x,y \in \mathcal{X}} \frac{f(x)\, g(y)}{f(y)\, g(x)}. \quad \text{(A.4.11)}$$

Proof. Since $\mathcal{X}$ is compact it is straightforward to see that $\mathcal{C}_{++}(\mathcal{X})$ is the interior of $\mathcal{C}_{+}(\mathcal{X})$. By applying (Lemmens and Nussbaum, 2012, Lemma 1.2.2) we have that $\mathcal{C}_{++}(\mathcal{X})$ is a part of $\mathcal{C}_{+}(\mathcal{X})$. The characterization of $M(f/g)$ and $m(f/g)$ follows by direct calculation from the definition using the fact that $\inf_{\mathcal{X}} h = \min_{\mathcal{X}} h > 0$ for any $h \in \mathcal{C}_{++}(\mathcal{X})$ since $\mathcal{X}$ is compact. Finally, the characterization of Hilbert's metric on $\mathcal{C}_{++}(\mathcal{X})$ is obtained by recalling that $(\min_{x\in\mathcal{X}} h(x))^{-1} = \max_{x\in\mathcal{X}} h(x)^{-1}$ for every $h \in \mathcal{C}_{++}(\mathcal{X})$. □

Lemma A.4 (Ordering properties of L_α)**.** *Let $\alpha \in \mathcal{P}(\mathcal{X})$. Then the following holds:*

(i) the operator L_α is order-preserving *(with respect to the cone $\mathcal{C}_{+}(\mathcal{X})$), that is,*

$$(\forall\, f, g \in \mathcal{C}(\mathcal{X})) \qquad f \leq g \;\Rightarrow\; \mathsf{L}_\alpha f \leq \mathsf{L}_\alpha g; \tag{A.4.12}$$

(ii) L_α maps parts of $\mathcal{C}_{+}(\mathcal{X})$ to parts of $\mathcal{C}_{+}(\mathcal{X})$, that is,

$$(\forall\, f, g \in \mathcal{C}(\mathcal{X})) \qquad f \sim g \;\Rightarrow\; \mathsf{L}_\alpha f \sim \mathsf{L}_\alpha g; \tag{A.4.13}$$

(iii) $\mathsf{L}_\alpha(\mathcal{C}_{+}(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$ and $\mathsf{L}_\alpha(\mathcal{C}_{++}(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X})$.

Proof. (i): Let $f, g \in \mathcal{C}(\mathcal{X})$ with $f \leq g$. Then $g - f \in \mathcal{C}_{+}(\mathcal{X})$ and by linearity of L_α combined with (A.4.3), we have $\mathsf{L}_\alpha g - \mathsf{L}_\alpha f = \mathsf{L}_\alpha(g - f) \geq 0$.

(ii): Let $f, g \in \mathcal{C}_{+}(\mathcal{X})$ with $f \sim g$. Then there exist $t, s \in \mathbb{R}$ and $s', t' \in \mathbb{R}$ such that $tg \leq f \leq sg$ and $t'f \leq g \leq s'f$. Since L_α is linear and order-preserving, we have $\mathsf{L}_\alpha f \sim \mathsf{L}_\alpha g$.

(iii): Let $f \in \mathcal{C}_{+}(\mathcal{X})$. By (A.4.3) and (A.4.4), for any $x \in \mathcal{X}$

$$0 \leq (\mathsf{L}_\alpha f)(x) \leq \|\mathsf{L}_\alpha f\|_\infty \leq \int f(x)\, d\alpha(x) = \|f\|_{L^1(\mathcal{X},\alpha)}. \tag{A.4.14}$$

Moreover,

$$\mathsf{L}_\alpha f(x) = \int k(y, x) f(y)\, d\alpha(y) \geq e^{-\mathsf{D}/\varepsilon}\, \|f\|_{L^1(\mathcal{X},\alpha)}. \tag{A.4.15}$$

Therefore, if $\|f\|_{L^1(\mathcal{X},\alpha)} = 0$ then by (A.4.14) $\mathsf{L}_\alpha f = 0$; if $\|f\|_{L^1(\mathcal{X},\alpha)} > 0$ then by (A.4.15) $\mathsf{L}_\alpha f \in \mathcal{C}_{++}(\mathcal{X})$. We conclude that the operator L_α maps $\mathcal{C}_{+}(\mathcal{X})$ in $\mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$. Moreover, $\mathsf{L}_\alpha(\mathcal{C}_{++}(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X})$, since for every $f \in \mathcal{C}_{++}(\mathcal{X})$ we have $\|f\|_{L^1(\mathcal{X},\alpha)} \geq \min_{\mathcal{X}} f > 0$. □

Following (Lemmens and Nussbaum, 2012, Section A.4) we now introduce a quantity which plays a role of central importance.

Definition A.16 (Projective Diameter of L_α)**.** *Let* $\alpha \in \mathcal{P}(\mathcal{X})$. *The* projective diameter of L_α *is*

$$\Delta(\mathsf{L}_\alpha) = \sup\{d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha g) \mid f, g \in \mathcal{C}_+(\mathcal{X}),\ \mathsf{L}_\alpha f \sim \mathsf{L}_\alpha g\}. \tag{A.4.16}$$

The following result shows that it is possible to find a finite upper bound on $\Delta(\mathsf{L}_\alpha)$ that is independent on α.

Proposition A.5 (Upper bound on the Projective Diameter of L_α)**.** *Let* $\alpha \in \mathcal{P}(\mathcal{X})$. *Then*

$$\Delta(\mathsf{L}_\alpha) \leq 2\mathsf{D}/\varepsilon. \tag{A.4.17}$$

Proof. Let $f, g \in \mathcal{C}_+(\mathcal{X})$. Recall that L_α maps $\mathcal{C}_+(\mathcal{X})$ into $\mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$ (see Lemma A.4Item (iii)) and that $\{0\}$ and $\mathcal{C}_{++}(\mathcal{X})$ are two parts of $\mathcal{C}_+(\mathcal{X})$ with respect to the relation $\sim$ (see (Lemmens and Nussbaum, 2012, Lemma 1.2.2)). Now, if $\mathsf{L}_\alpha f = \mathsf{L}_\alpha g = 0$, then we have $d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha g) = d_H(0, 0) = 0$. Therefore it is sufficient to study the case that $\mathsf{L}_\alpha f, \mathsf{L}_\alpha g \in \mathcal{C}_{++}(\mathcal{X})$. Following the characterization of Hilbert's metric on $\mathcal{C}_{++}(\mathcal{X})$ given in Lemma A.3, we have

$$\begin{aligned}
d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') &= \log \max_{x,y\in\mathcal{X}} \frac{(\mathsf{L}_\alpha f)(x)\ (\mathsf{L}_\alpha f')(y)}{(\mathsf{L}_\alpha f)(y)\ (\mathsf{L}_\alpha f')(x)} \\
&= \log \max_{x,y\in\mathcal{X}} \frac{\int \mathsf{k}(x,z) f(z)\ d\alpha(z)\ \int \mathsf{k}(y,w) f'(w)\ d\alpha(w)}{\int \mathsf{k}(y,z) f(z)\ d\alpha(z)\ \int \mathsf{k}(x,w) f'(w)\ d\alpha(w)} \\
&= \log \max_{x,y\in\mathcal{X}} \frac{\int \mathsf{k}(x,z)\mathsf{k}(y,w)\ f(z) f'(w)\ d\alpha(z) d\alpha(w)}{\int \mathsf{k}(y,z)\mathsf{k}(x,w)\ f(z) f'(w)\ d\alpha(z) d\alpha(w)} \\
&= \log \max_{x,y\in\mathcal{X}} \frac{\int \frac{\mathsf{k}(x,z)\mathsf{k}(y,w)}{\mathsf{k}(y,z)\mathsf{k}(x,w)}\ \mathsf{k}(y,z)\mathsf{k}(x,w)\ f(z) f'(w)\ d\alpha(z) d\alpha(w)}{\int \mathsf{k}(y,z)\mathsf{k}(x,w)\ f(z) f'(w)\ d\alpha(z) d\alpha(w)} \\
&\leq \log \max_{x,y,z,w\in\mathcal{X}} \frac{\mathsf{k}(x,z)\mathsf{k}(y,w)}{\mathsf{k}(y,z)\mathsf{k}(x,w)}.
\end{aligned}$$

Since for every $x, y \in \mathcal{X}$ $\mathsf{c}(x,y) \in [0, \mathsf{D}]$, we have $k(x,y) \in [e^{-\mathsf{D}/\varepsilon}, 1]$ and hence

$$d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \leq 2\mathsf{D}/\varepsilon. \qquad \square$$

A consequence of Prop. A.5 is a special case of Birkhoff-Hopf theorem.

Theorem A.6 (Birkhoff-Hopf Theorem)**.** *Let* $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$ *and* $\alpha \in \mathcal{P}(\mathcal{X})$. *Then, for every*

$f, f' \in \mathcal{C}_+(\mathcal{X})$ such that $f \sim f'$, we have

$$d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \leq \lambda \, d_H(f, f'). \tag{A.4.18}$$

Proof. The statement is a direct application of the Birkhoff-Hopf theory (Lemmens and Nussbaum, 2012, Sections A.4 and A.7). The *Birkhoff contraction ratio* of L_α is defined as

$$\kappa(\mathsf{L}_\alpha) = \inf \left\{ \hat{\lambda} \in \mathbb{R}_+ \;\middle|\; d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \leq \hat{\lambda} d_H(f, f') \;\; \forall f, f' \in \mathcal{C}_+(\mathcal{X}), \;\; f \sim f' \right\}.$$

Then it follows from Birkhoff-Hopf theorem (Lemmens and Nussbaum, 2012, Theorem A.4.1) that

$$\kappa(\mathsf{L}_\alpha) = \tanh\left(\frac{1}{4}\Delta(\mathsf{L}_\alpha)\right). \tag{A.4.19}$$

Recalling the upper bound on the projective diameter of L_α given in Prop. A.5, we have

$$\kappa(\mathsf{L}_\alpha) \leq \tanh\left(\frac{\mathsf{D}}{2\varepsilon}\right) = \frac{e^{\mathsf{D}/\varepsilon} - 1}{e^{\mathsf{D}/\varepsilon} + 1} = \lambda,$$

and Eq. (A.4.18) follows. □

A.4.2 DAD problems

We introduce the formalism of DAD problems. The dual formulation of Entropic OT falls into this framework and hence the results reported in this section will be instrumental for the proof of existence of dual potentials. Recall that the operator L_α introduced in the previous section is defined as follows: for $\alpha \in \mathcal{P}(\mathcal{X})$, $\mathsf{L}_\alpha \colon \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is given by

$$(\forall f \in \mathcal{C}(\mathcal{X})) \qquad \mathsf{L}_\alpha f \colon x \mapsto \int \mathsf{k}(x, z) f(z) \, d\alpha(z). \tag{A.4.20}$$

We now introduce another operator, closely related to L_α.

The map A_α. Let $\alpha \in \mathcal{P}(\mathcal{X})$. We define the map $\mathsf{A}_\alpha \colon \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$, such that

$$\forall \, f \in \mathcal{C}_{++}(\mathcal{X}), \qquad \mathsf{A}_\alpha(f) = \mathsf{R} \circ \mathsf{L}_\alpha(f) = 1/(\mathsf{L}_\alpha f), \tag{A.4.21}$$

where $\mathsf{R}\colon \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ is defined by $\mathsf{R}(f) = 1/f$ with

$$(1/f)\colon x \mapsto \frac{1}{f(x)}. \tag{A.4.22}$$

Note that A_α is well defined since, by Lemma A.4Item (iii), $\mathsf{L}_\alpha(\mathcal{C}_{++}(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X})$ and, for every $f \in \mathcal{C}_{++}(\mathcal{X})$, $\min_{\mathcal{X}} f > 0$, because $\mathcal{X}$ compact. Moreover, it follows from (A.4.11) in Lemma A.3, that, for any two $f, f' \in \mathcal{C}_{++}(\mathcal{X})$

$$d_H(1/f, 1/f') = \log \max_{x,y\in\mathcal{X}} \frac{f(y)f'(x)}{f(x)f'(y)} = d_H(f, f'). \tag{A.4.23}$$

We highlight here the connection between P_α introduced in (2.3.5) and A_α, namely for any $\alpha \in \mathcal{P}(\mathcal{X})$ and $u \in \mathcal{C}(\mathcal{X})$

$$\mathsf{P}_\alpha(u) = -\varepsilon \log(\mathsf{A}_\alpha(e^{u/\varepsilon})). \tag{A.4.24}$$

Dual OT_ε Problem. Recall the dual problem introduced (2.3.3) of the optimal transport problem with entropic regularization: for $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ and $\varepsilon > 0$

$$\max_{u,v\in\mathcal{C}(\mathcal{X})} \int u(x)\, d\alpha + \int v(y)\, d\beta(y) - \varepsilon \int e^{\frac{u(x)+v(y)-\mathsf{d}(x,y)}{\varepsilon}}\, d\alpha(x) d\beta(y). \tag{A.4.25}$$

The optimality conditions for the problem above are

$$\begin{cases} e^{-\frac{u(x)}{\varepsilon}} = \displaystyle\int_{\mathcal{X}} e^{\frac{v(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\beta(y) \quad (\forall\, x \in \mathrm{supp}(\alpha)) \\ e^{-\frac{v(y)}{\varepsilon}} = \displaystyle\int_{\mathcal{X}} e^{\frac{u(x)-\mathsf{c}(x,y)}{\varepsilon}}\, d\alpha(x) \quad (\forall\, y \in \mathrm{supp}(\beta)), \end{cases} \tag{A.4.26}$$

which are equivalent to

$$\begin{cases} g(y)^{-1} = \displaystyle\int_{\mathcal{X}} e^{\frac{-\mathsf{c}(x,y)}{\varepsilon}} f(x)\, d\alpha(x) \quad (\forall\, y \in \mathrm{supp}(\beta)) \\ f(x)^{-1} = \displaystyle\int_{\mathcal{X}} e^{\frac{-\mathsf{c}(x,y)}{\varepsilon}} g(y)\, d\beta(y) \quad (\forall\, x \in \mathrm{supp}(\alpha)), \end{cases} \tag{A.4.27}$$

where $f = e^{u/\varepsilon} \in \mathcal{C}_{++}(\mathcal{X})$ and $g = e^{v/\varepsilon} \in \mathcal{C}_{++}(\mathcal{X})$. In the rest of the section we will

consider the following *DAD problem*: (Lemmens and Nussbaum, 2012; Nussbaum, 1993)

$$\forall\, y \in \mathcal{X} \ \int_{\mathcal{X}} f(x)\mathsf{k}(x,y)g(y)\, d\alpha(x) = 1 \ \text{ and } \ \forall\, x \in \mathcal{X} \ \int_{\mathcal{X}} f(x)\mathsf{k}(x,y)g(y)\, d\beta(y) = 1. \tag{A.4.28}$$

It is clear that a solution of (A.4.28) is also a solution of (A.4.27). The vice versa is in general not true; however, there is a canonical way to build solutions of (A.4.28) starting from solutions of (A.4.27): indeed if (f, g) is a solution of (A.4.27), then the functions $\bar{f}, \bar{g} \colon \mathcal{X} \to \mathbb{R}$ defined through $\bar{f}(x)^{-1} = \int_{\mathcal{X}} \mathsf{k}(x,y)g(y)\, d\beta(y)$ and $\bar{g}(y)^{-1} = \int_{\mathcal{X}} \mathsf{k}(x,y)g(y)\, d\beta(y)$ provide a solution of (A.4.28). So, the dual OT_ε problem (A.4.25) admits a solution if and only if the corresponding DAD problem (A.4.28) admits a solution. Recalling the definition of A_α in (A.4.21), problem (A.4.28) can be more compactly written as

$$f = \mathsf{A}_\beta(g) \qquad \text{and} \qquad g = \mathsf{A}_\alpha(f), \tag{A.4.29}$$

or equivalently, by setting $\mathsf{A}_{\beta\alpha} = \mathsf{A}_\beta \circ \mathsf{A}_\alpha$ and $\mathsf{A}_{\alpha\beta} = \mathsf{A}_\alpha \circ \mathsf{A}_\beta$,

$$f = \mathsf{A}_{\beta\alpha}(f) \qquad \text{and} \qquad g = \mathsf{A}_{\alpha\beta}(g). \tag{A.4.30}$$

This shows that the solutions of the DAD problem (A.4.28) are the fixed points of $\mathsf{A}_{\alpha\beta}$ and $\mathsf{A}_{\beta\alpha}$ respectively. Note that the operators $\mathsf{A}_{\beta\alpha}$ and $\mathsf{A}_{\alpha\beta}$ are positively homogeneous, that is, for every $t \in \mathbb{R}_{++}$ and $f \in \mathcal{C}_{++}(\mathcal{X})$, $\mathsf{A}_{\beta\alpha}(tf) = t\mathsf{A}_{\beta\alpha}(f)$ and $\mathsf{A}_{\alpha\beta}(tf) = t\mathsf{A}_{\alpha\beta}(f)$. Thus, if f is a fixed point of $\mathsf{A}_{\beta\alpha}$, then tf is also a fixed point of $\mathsf{A}_{\beta\alpha}$, for every $t > 0$. If (f, g) is a solution of the DAD problem (A.4.28), then the pair (u, v), with $u = \varepsilon \log f$ and $v = \varepsilon \log g$ is a solution of (A.4.25). We refer to these solutions as *Sinkhorn potentials* of the pair (α, β). Finally, note that, it follows from (A.4.26) that solutions of (A.4.25) are determined (α, β)-a.e. on $\mathcal{X}$ and up to a translation of the form $(u + t, v - t)$, for some $t \in \mathbb{R}$.

The following result is essentially the specialization of (Lemmens and Nussbaum, 2012, Thm. 7.1.4) to the case of the map $\mathsf{A}_{\beta\alpha}$.

Theorem A.7 (Hilbert's metric contraction for $\mathsf{A}_{\beta\alpha}$)**.** *The map $\mathsf{A}_{\beta\alpha} : \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ has a unique fixed point up to positive scalar multiples. Moreover, let $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$. Then, for*

every $f, f' \in \mathcal{C}_{++}(\mathcal{X})$,

$$d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta\alpha}(f')) \leq \lambda^2 \, d_H(f, f'). \tag{A.4.31}$$

Proof. To show that $\mathsf{A}_{\beta\alpha} : \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ has a unique fixed point up to positive scalar we refer to the proof of (Lemmens and Nussbaum, 2012, Thm. 7.1.4). As for (A.4.31), this can be easily seen with the tools that we have introduced so far. By combining (A.4.23) with Thm. A.6 we obtain that, for any $f, f' \in \mathcal{C}_{++}(\mathcal{X})$

$$d_H(\mathsf{A}_\alpha(f), \mathsf{A}_\alpha(f')) \;=\; d_H(1/(\mathsf{L}_\alpha f), 1/(\mathsf{L}_\alpha f')) \;=\; d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \;\leq\; \lambda \, d_H(f, f'). \tag{A.4.32}$$

Since the same holds for A_β then (A.4.31) is satisfied. □

We conclude this part gathering a few easy results that are useful in the rest of the appendix.

Lemma A.8. *(Auxiliary Cone) Consider the set*

$$K = \{f \in \mathcal{C}_+(\mathcal{X}) \mid f(x) \leq f(y)\, e^{\mathsf{D}/\varepsilon} \;\; \forall x, y \in \mathcal{X}\}. \tag{A.4.33}$$

Let $\alpha \in \mathcal{P}(\mathcal{X})$. *Then the following holds.*

(i) K *is a closed convex cone and* $K \subset \mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$;

(ii) $\mathsf{L}_\alpha(\mathcal{C}_+(\mathcal{X})) \subset K$;

(iii) $\mathsf{R}(K) \subset K$, *where recall that* $\mathsf{R}\colon \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ *is defined by* $\mathsf{R}(f) = 1/f$;

(iv) $Ran(\mathsf{A}_\alpha) \subset K$;

(v) If $f \in K$ *and* $g = \mathsf{A}_\alpha f$, *then* $g \in K$ *and* $1 \leq (\min_{\mathcal{X}} g)\, \|f\|_\infty \leq \|g\|_\infty \, \|f\|_\infty \leq e^{2\mathsf{D}/\varepsilon}$.

(vi) If $f \in K$ *is such that* $f(x_o) = 1$ *for some* $x_o \in \mathcal{X}$, *then* $\|\varepsilon \log f\|_\infty \leq \mathsf{D}$.

Proof. (i): We see that for any $f \in K$,

$$\max_{\mathcal{X}} f \leq (\min_{\mathcal{X}} f)\, e^{\mathsf{D}/\varepsilon}, \tag{A.4.34}$$

so, if $f(x) = 0$ for some $x \in \mathcal{X}$, then $f(x) = 0$ on all $\mathcal{X}$. Hence $K \subseteq \mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$. It is straightforward to verify that K is a convex cone. Moreover K is also closed. Indeed if $(f_n)_{n\in\mathbb{N}}$ is a sequence in K which converges uniformly to $f \in \mathcal{C}(\mathcal{X})$, then, for every

$x, y \in \mathcal{X}$ and every $n \in \mathbb{N}$, $f_n(x) \leq f_n(y)e^{\mathsf{D}/\varepsilon}$ and hence, letting $n \to +\infty$, we have $f(x) \leq f(y)e^{\mathsf{D}/\varepsilon}$.

(ii): For every $f \in \mathcal{C}_+(\mathcal{X})$ and $x, y \in \mathcal{X}$, we have

$$(\mathsf{L}_\alpha f)(x) = \int \mathsf{k}(x,z) f(z)\, d\alpha(z) = \int \frac{\mathsf{k}(x,z)}{\mathsf{k}(y,z)}\, \mathsf{k}(y,z) f(z)\, d\alpha(z)$$
$$\leq e^{\mathsf{D}/\varepsilon} \int \mathsf{k}(y,z) f(z)\, d\alpha(z) = e^{\mathsf{D}/\varepsilon} (\mathsf{L}_\alpha f)(y).$$

(iii): For every $f \in K$,

$$(\forall\, x, y \in \mathcal{X}) \qquad f(x) \leq f(y)\, e^{\mathsf{D}/\varepsilon} \;\Leftrightarrow\; \frac{1}{f(y)} \leq \frac{1}{f(x)}\, e^{\mathsf{D}/\varepsilon}.$$

(iv) It follows from Item (ii) and Item (iii) and the definition of A_α.

(v): It follows from (iv), Eq. (A.4.38), and Eq. (A.4.34).

(vi): Let $f \in K$ be such that $f(x_o) = 1$. Then $\min_{\mathcal{X}} f \leq 1 \leq \max_{\mathcal{X}} f$. Thus, it follows from Eq. (A.4.34) that

$$\max_{\mathcal{X}} f \leq e^{\mathsf{D}/\varepsilon} \quad \text{and} \quad \min_{\mathcal{X}} f \geq e^{-\mathsf{D}/\varepsilon} \tag{A.4.35}$$

and hence, for every $x \in \mathcal{X}$, $-\mathsf{D} \leq \varepsilon \log f(x) \leq \mathsf{D}$. □

A.4.3 Hilbert metric and relation with supremum norm

While the Hilbert metric is useful for DAD problems and to handle the dual potentials, ultimately one is more interested in dealing with more classic norms, such as the supremum norm. We conclude this section with two results on Hilbert metric and its relation with L_∞ norm.

Lemma 4.3. *Let $f, f' \in \mathcal{C}_{++}(\mathcal{X})$ and set $u = \varepsilon \log f$ and $u' = \varepsilon \log f'$. Then*

$$d_H(f, f') \leq 2 \left\| \log f - \log f' \right\|_\infty \quad \textit{or, equivalently} \quad d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}) \leq \frac{2}{\varepsilon} \left\| u - u' \right\|_\infty . \tag{4.3.3}$$

Moreover, let $x_o \in \mathcal{X}$, consider the sets $\mathcal{A} = \{h \in \mathcal{C}_{++}(\mathcal{X}) \mid h(x_o) = 1\}$ and $\mathcal{B} = \{w \in \mathcal{C}(\mathcal{X}) \mid w(x_o) = 0\}$. Suppose that $f, f' \in \mathcal{A}$ (or equivalently that $u, u' \in \mathcal{B}$). Then

$$\frac{1}{2} d_H(f, f') \leq \left\| \log f - \log f' \right\|_\infty \leq d_H(f, f') \tag{4.3.4}$$

and

$$\frac{\varepsilon}{2} d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}) \leq \left\|u - u'\right\|_\infty \leq \varepsilon \, d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}). \tag{4.3.5}$$

Proof. We have

$$\begin{aligned} d_H(f, f') &= \log \max_{x,y \in \mathcal{X}} \frac{f(x) f'(y)}{f(y) f'(x)} \\ &= \log \max_{x \in \mathcal{X}} \frac{f(x)}{f'(x)} + \log \max_{y \in \mathcal{X}} \frac{f'(y)}{f(y)} \\ &= \max_{x \in \mathcal{X}} \log \frac{f(x)}{f'(x)} + \max_{y \in \mathcal{X}} \log \frac{f'(y)}{f(y)} \\ &\leq 2 \max_{x \in \mathcal{X}} \left| \log \frac{f(x)}{f'(x)} \right| = 2 \left\|\log(f/f')\right\|_\infty = 2 \left\|\log f - \log f'\right\|_\infty \end{aligned}$$

and (4.3.3) follows. Suppose that $f, f' \in \mathcal{A}$. Then

$$\begin{aligned} \left\|\log f - \log f'\right\|_\infty &= \max \left\{ \log \max_{x \in \mathcal{X}} \frac{f(x)}{f'(x)}, \log \max_{x \in \mathcal{X}} \frac{f'(x)}{f(x)} \right\} \\ &= \max \left\{ \log \max_{x \in \mathcal{X}} \frac{f(x) f'(x_o)}{f(x_o) f'(x)}, \log \max_{x \in \mathcal{X}} \frac{f(x_o) f'(x)}{f(x) f'(x_o)} \right\} \\ &\leq \max \left\{ \log \max_{x,y \in \mathcal{X}} \frac{f(x) f'(y)}{f(y) f'(x)}, \log \max_{x,y \in \mathcal{X}} \frac{f(y) f'(x)}{f(x) f'(y)} \right\} \\ &= d_H(f, f'), \end{aligned}$$

since $f(x_o)/f'(x_o) = f'(x_o)/f(x_o) = 1$. Therefore, (4.3.4) follows. □

Lemma A.9. *For every $x, y \in \mathbb{R}_{++}$ we have*

$$|\log x - \log y| \leq \max\left\{x^{-1}, y^{-1}\right\} |x - y|. \tag{A.4.36}$$

The following result allows to extend the previous observations on a pair f, f' to the corresponding $g = \mathsf{A}_\alpha f$ and $g' = \mathsf{A}_\alpha f'$.

Lemma A.10. *Let $x_o \in \mathcal{X}$ and $K \subset \mathcal{C}_+(\mathcal{X})$ the cone from Lemma A.8. Let $f, f' \in K$ be such that $f(x_o) = f'(x_o) = 1$, and set $g = \mathsf{A}_\alpha f$ and $g' = \mathsf{A}_\alpha f'$. Then,*

$$\left\|\log g - \log g'\right\|_\infty \leq e^{3\mathsf{D}/\varepsilon} \left\|\log f - \log f'\right\|_\infty. \tag{A.4.37}$$

Proof. It follows from (A.4.21) and Lemma A.9 that

$$|\log g - \log g'| = \left|\log \frac{g}{g'}\right| = \left|\log \frac{\mathsf{L}_\alpha f'}{\mathsf{L}_\alpha f}\right| \leq \max\{g', g\}|\mathsf{L}_\alpha f - \mathsf{L}_\alpha f'|.$$

Therefore, since $1 \leq \|f\|_\infty, \|f'\|_\infty$, and recalling Lemma A.8Item (v) and (A.4.4), we have

$$\begin{aligned}
\left\|\log g - \log g'\right\|_\infty &\leq \max\{\|g\|_\infty, \left\|g'\right\|_\infty\} \left\|\mathsf{L}_\alpha f - \mathsf{L}_\alpha f'\right\|_\infty \\
&\leq \max\{\|f\|_\infty \|g\|_\infty, \left\|f'\right\|_\infty \left\|g'\right\|_\infty\} \left\|\mathsf{L}_\alpha f - \mathsf{L}_\alpha f'\right\|_\infty \\
&\leq e^{2\mathsf{D}/\varepsilon} \left\|f - f'\right\|_\infty \\
&= e^{2\mathsf{D}/\varepsilon} \|e^{\log f} - e^{\log f'}\|_\infty.
\end{aligned}$$

Now, since $f, f' \leq e^{\mathsf{D}/\varepsilon}$, we have $\log f, \log f' \leq \mathsf{D}/\varepsilon$. Thus, the statement follows by noting that the exponential function is Lipschitz continuous on $]-\infty, \mathsf{D}/\varepsilon]$ with constant $e^{\mathsf{D}/\varepsilon}$. □

A.4.4 Existence of potentials and properties

In the previous subsections we have presented all the tools that are needed to show existence of the potentials, which is now a corollary of previous results.

Corollary A.11 (Existence and uniqueness of Sinkhorn potentials)**.** *Let $\alpha, \beta \in \mathcal{P}(\mathcal{X})$. Then, the DAD problem (A.4.28) admits a solution (f, g) and every other solution is of type $(tf, t^{-1}g)$, for some $t > 0$. Moreover, there exists a pair $(u, v) \in \mathcal{C}(\mathcal{X})^2$ of Sinkhorn potentials and every other pair of Sinkhorn potentials is of type $(u + s, v - s)$, for some $s \in \mathbb{R}$. In particular, for every $x_o \in \mathcal{X}$, there exist a unique pair (u, v) of Sinkhorn potentials such that $u(x_0) = 0$.*

Proof. It follows from Thm. A.7 and the discussion after (A.4.30). □

Bounding (f, g) point-wise. We conclude this section by providing additional properties of the solutions (f, g) of the DAD problem (A.4.29). In particular, we show that there exists one such solution for which it is possible to provide a point-wise upper and lower bound independent on α and β.

Remark A.1. *Let $f \in \mathcal{C}_{++}(\mathcal{X})$ and set $g = \mathsf{A}_\alpha(f)$. Then, recalling (A.4.21) and Eq. (A.4.4), we have that, for every $x \in \mathcal{X}$,*

$$1 = g(x)(\mathsf{L}_\alpha\, f)(x) \leq g(x) \|\mathsf{L}_\alpha f\|_\infty \leq g(x) \|f\|_\infty$$

and

$$1 = g(x)(\mathsf{L}_\alpha\, f)(x) \geq g(x)(\min_{\mathcal{X}} f) \int \mathsf{k}(x,z)\, d\alpha(z) \geq g(x)(\min_{\mathcal{X}} f)e^{-\mathsf{D}/\varepsilon}.$$

Therefore,

$$\min_{\mathcal{X}} g \geq \frac{1}{\|f\|_\infty} \quad \textit{and} \quad \|g\|_\infty \leq \frac{e^{\mathsf{D}/\varepsilon}}{\min_{\mathcal{X}} f}. \tag{A.4.38}$$

As a direct consequence of Lemma A.8 we can establish a uniform point-wise upper and lower bound for the value of DAD solutions.

Corollary A.12. *Let $\alpha, \beta \in \mathcal{P}(\mathcal{X})$. Let $x_o \in \mathcal{X}$ and let (f, g) be the solution of* (A.4.29) *such that $f(x_o) = 1$. Then $\|f\|_\infty \leq e^{\mathsf{D}/\varepsilon}$ and $\|g\|_\infty \leq e^{2\mathsf{D}/\varepsilon}$. Moreover, the corrisponding pair (u, v) of Sinkhorn potentials satifies $\|u\|_\infty \leq \mathsf{D}$ and $\|v\|_\infty \leq 2\mathsf{D}$.*

Proof. Since f and g are fixed points of $\mathsf{A}_{\beta\alpha}$ and $\mathsf{A}_{\alpha\beta}$ respectively, it follows from Lemma A.8Item (iv) that $f, g \in K$. Then, Lemma A.8Item (vi) yields $\|f\|_\infty \leq e^{\mathsf{D}/\varepsilon}$, whereas by the second of (A.4.38) and (A.4.35) we derive that $\|g\|_\infty \leq e^{2\mathsf{D}/\varepsilon}$. □

Appendix B

Appendix of Chapter 3

This appendix is structured as follows:

Appendix B.1. This section presents more details on the Example 3.1 in the main body of the manuscript.

Appendix B.2. This section contains the proof of Prop. 3.1.

Appendix B.3. This section presents the proof of Thm. 3.3 on the differential properties on sharp and vanilla Sinkhorn and the formula of the gradient of sharp Sinkhorn.

B.1 Example: Barycenter of Dirac Deltas

Wasserstein barycenter problems can be divided into two main classes: problems in which the support is free (and must be computed, generating a nonconvex problem (Cuturi and Doucet, 2014)) and problems where the support is fixed. In some cases, the latter is the only valid choice: for instance, when the geometric domain is a space of symbols and the cost matrix M contains the symbol-to-symbol dissimilarities, no extra information of the symbol space is available and the support of the barycenter will have to lie on a pre-determined set in order to be meaningful. A concrete example is the following: when dealing with histograms on words, the barycenter will optimize how to spread the mass among a set of known words that are used to build the matrix C, through a word2vec operation. In the following we carry out the computation of the barycenter of two Dirac deltas with regularized Sinkhorn and Sinkhorn distances, in order to prove what stated in Example 3.1.

Barycenter with OT_ε. Let $\mu = \delta_z$ be the Dirac delta centered at $z \in \mathbb{R}^d$ and $\nu = \delta_y$ the Dirac delta centered at $y \in \mathbb{R}^d$. We fix the set of admissible support of the barycenter $\mathcal{X} = \{x_1, \dots, x_n\}$, where $x_i \in \mathbb{R}^d$ for any i. For the sake of simplicity let us assume that $\mathcal{X}$

contains the point $(y+z)/2$. The cost matrices with mutual distances between z and $\mathcal{X}$ and y and $\mathcal{X}$ will be

$$C^z = \{\mathsf{d}(z, x_i)\}_{i=1}^n \in \mathbb{R}^n, \qquad C^y = \{\mathsf{d}(y, x_i)\}_{i=1}^n.$$

Since the support is fixed, only the weights $\mathsf{a} = (\mathsf{a}_1, \ldots, \mathsf{a}_n)$ of the barycenter $\mu_\varepsilon = \sum_{i=1}^n \mathsf{a}_i \delta_{x_i}$ are to be computed. Vector a is the minimizer of the following functional

$$\Delta_n \ni \mathsf{a} \longrightarrow \mathcal{B}_{\mathsf{OT}_\varepsilon}(\mathsf{a}) = \frac{1}{2}\mathsf{OT}_\varepsilon(\mathsf{a}, \delta_z) + \frac{1}{2}\mathsf{OT}_\varepsilon(\mathsf{a}, \delta_y).$$

Note that since Dirac delta has mass 1 concentrated at a point, the transport polytope corresponding to a and a Dirac delta is $\Pi(\mathsf{a}, 1)$. The elements in $\Pi(\mathsf{a}, 1)$ are those matrices $T \in \mathbb{R}^{n\times 1}$ such that $T\mathbb{1}_1 = \mathsf{a}$ and $T^\top \mathbb{1}_n = 1$. Thus,

$$\begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{pmatrix} \begin{pmatrix} 1 \end{pmatrix} = \begin{pmatrix} \mathsf{a}_1 \\ \mathsf{a}_2 \\ \vdots \\ \mathsf{a}_n \end{pmatrix} \tag{B.1.1}$$

which implies $T_1 = \mathsf{a}_1, \ldots, T_n = \mathsf{a}_n$. In this case, $\Pi(\mathsf{a}, 1)$ contains only one matrix, which coincides with $\mathsf{a}^\top$. The distance $\mathsf{OT}_\varepsilon(\mathsf{a}, \delta_z)$ is given by $\langle \mathsf{a}^\top, C^z\rangle - \varepsilon \mathsf{H}(\mathsf{a})$ and, similarly, $\mathsf{OT}_\varepsilon(\mathsf{a}, \delta_y) = \langle \mathsf{a}^\top, C^y\rangle - \varepsilon \mathsf{H}(\mathsf{a})$. Then, the goal is to minimize

$$\mathsf{a} \longrightarrow \frac{1}{2}\langle \mathsf{a}, C^z\rangle + \frac{1}{2}\langle \mathsf{a}, C^y\rangle + \varepsilon \sum_{i=1}^n \mathsf{a}_i(\log \mathsf{a}_i - 1)$$

with the constraint that $\mathsf{a} \in \Delta_n$. The partial derivative with respect to a_i is given by

$$\frac{\partial \mathcal{B}_{\mathsf{OT}_\varepsilon}}{\partial \mathsf{a}_i} = \frac{1}{2}(C_i^z + C_i^y) + \varepsilon \log \mathsf{a}_i$$

Setting it equal to zero, it yields $\mathsf{a}_i = \mathrm{e}^{\frac{-(C_i^z + C_i^y)}{2\varepsilon}}$. The constraint $\mathsf{a} \in \Delta_n$ leads to

$$\mathsf{a}_i = \frac{\mathrm{e}^{\frac{-(C_i^z + C_i^y)}{2\varepsilon}}}{\sum_{j=1}^n \mathrm{e}^{\frac{-(C_j^z + C_j^y)}{2\varepsilon}}}.$$

Then the barycenter μ_ε^* has weights $(\mathsf{a}_1, \dots, \mathsf{a}_n)$ where each a_i is strictly positive, with maximum at the entry corresponding to the point x_i which realizes the minimum distance from z and y, i.e. $(z+y)/2$. The sparsity of the initial deltas is lost.

Barycenter with $\widetilde{\mathsf{OT}}_\varepsilon$. On the other hand, let us compute the barycenter between μ and ν with respect to the Sinkhorn distance recalled in Eq. (3.1.4). The very same considerations on $\Pi(\mathsf{a}, 1)$ still hold, so $\Pi(\mathsf{a}, 1)$ contains $T = \mathsf{a}^\top$ only. Hence, in this case the Sinkhorn barycenter functional $\mathcal{B}_{\widetilde{\mathsf{OT}}_\varepsilon}$ coincides with the Wasserstein barycenter functional $\mathcal{B}_{\mathsf{W}}$, since $\widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a}, \delta_j) = \langle \mathsf{a}^\top, C^j \rangle = \mathsf{W}(\mathsf{a}, \delta_j)$, for $j = z, y$. This trivially implies that $\tilde{\mu}_\varepsilon^* = \mu_{\mathsf{W}}^*$.

B.2 Proof of Proposition 3.1 in Section 3.1

Proposition 3.1. *Let $\varepsilon > 0$. For any pair of discrete measures $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ with respective weights $\mathsf{a} \in \Delta_n$ and $\mathsf{b} \in \Delta_m$, we have*

$$\left| \widetilde{\mathsf{OT}}_\varepsilon(\alpha, \beta) - \mathsf{W}(\alpha, \beta) \right| \leq c_1 \, \mathrm{e}^{-\frac{1}{\varepsilon}}, \qquad c_2 \varepsilon \leq \left| \mathsf{OT}_\varepsilon(\alpha, \beta) - \mathsf{W}(\alpha, \beta) \right| \leq c_3 \varepsilon, \quad (3.1.5)$$

where c_1, c_2, c_3 are constants independent of ε, depending on the support of α and β.

Proof. As shown in (Cominetti and Martín, 1994)(Prop.5.1), the sequence T_ε converges to an optimal plan of W as ε goes to zero. More precisely,

$$T_\varepsilon \to T^* = \mathrm{argmax}_{T \in \Pi(\mathsf{a}, \mathsf{b})} \{ \mathsf{H}(T); \;\; \langle T, C \rangle = \mathsf{W}(\alpha, \beta) \}$$

exponentially fast, that is $\|T_\varepsilon - T^*\|_{\mathbb{R}^{nm}} \leq c \, \mathrm{e}^{-\frac{1}{\varepsilon}}$. Thus,

$$|\widetilde{\mathsf{OT}}_\varepsilon(\alpha, \beta) - \mathsf{W}(\alpha, \beta)| = |\langle T_\varepsilon, C \rangle - \langle T^*, C \rangle| \leq \|T_\varepsilon - T^*\| \|C\| \leq c \, \mathrm{e}^{-\frac{1}{\varepsilon}} \|C\| =: c_1 \mathrm{e}^{-\frac{1}{\varepsilon}}.$$

As for the second part, note that if T^* is a solution of $\min_{T \in \Pi(\mathsf{a}, \mathsf{b})} \langle T, C \rangle$, then $\langle T^*, C \rangle \leq \langle T, C \rangle$, for any other $T \in \Pi(\mathsf{a}, \mathsf{b})$, including any solution T_ε of $\min_{T \in \Pi(\mathsf{a}, \mathsf{b})} \langle T, C \rangle - \varepsilon \mathsf{H}(T)$. Also, by definition, $\mathsf{H}(T) \geq 0$ for any T in the transport polytope. Hence, $-\varepsilon \mathsf{H}(T) \leq 0$ and therefore $\mathsf{OT}_\varepsilon(\alpha, \beta) \leq \mathsf{W}(\alpha, \beta)$. Using this fact and the definition of entropy, we obtain

$$0 \leq \langle T^*, C \rangle - (\langle T_\varepsilon, C \rangle - \varepsilon \mathsf{H}(T_\varepsilon)) = \langle T^*, C \rangle - \langle T_\varepsilon, C \rangle + \varepsilon \mathsf{H}(T_\varepsilon) \leq \varepsilon \mathsf{H}(T_\varepsilon), \quad (B.2.1)$$

since $\langle T^*, C\rangle - \langle T_\varepsilon, C\rangle \leq 0$. This yields,

$$|\mathsf{OT}_\varepsilon(\alpha,\beta) - \mathsf{W}(\alpha,\beta)| \leq \varepsilon \mathsf{H}(T_\varepsilon) = \varepsilon\big(1 - \sum_{i,j} T_{\varepsilon,ij}\log(T_{\varepsilon,ij})\big) \leq c_3\varepsilon, \tag{B.2.2}$$

for some constant c_3, proving the desired upper bound. As for the lower bound, note that since T_ε is the optimum, it attains the minimum and hence

$$\langle T_\varepsilon, C\rangle - \varepsilon\mathsf{H}(T_\varepsilon) \leq \langle T, C\rangle - \varepsilon\mathsf{H}(T)$$

for any other T, including T^*. Therefore,

$$\mathsf{W}(\alpha,\beta) - \mathsf{OT}_\varepsilon(\alpha,\beta) = \langle T^*, C\rangle - \langle T_\varepsilon, C\rangle + \varepsilon\mathsf{H}(T_\varepsilon) \geq \varepsilon\mathsf{H}(T^*) \geq 0, \tag{B.2.3}$$

leading to

$$|\mathsf{W}(\alpha,\beta) - \mathsf{OT}_\varepsilon(\alpha,\beta)| \geq c_2\varepsilon, \tag{B.2.4}$$

for some constant c_2.

□

B.3 Proof of the formula of the gradient

With a similar procedure, the implicit function theorem provides a formula for the gradient of sharp Sinkhorn distance.

Theorem 3.3. *Let $C \in \mathbb{R}^{n\times m}$ be a cost matrix, $\mathsf{a} \in \Delta_n$, $\mathsf{b} \in \Delta_m$ and $\varepsilon > 0$. Let $\mathcal{L}_{\mathsf{a},\mathsf{b}}(u,v)$ be defined as the argument of the maximization in the right hand side of Eq. (2.3.31), with argmax in (u_*, v_*). Let T_ε be defined as in (3.2.1). Then,*

$$\nabla_a \widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a},\mathsf{b}) = proj_{\mathrm{T}\Delta_n}\big(A\, L\mathbb{1}_m + B\,\bar{L}^\top \mathbb{1}_n\big) \tag{3.2.3}$$

where $L = T_\varepsilon \odot C \in \mathbb{R}^{n\times m}$ is the entry-wise multiplication between T_ε and C and $\bar{L} \in \mathbb{R}^{n\times m-1}$ corresponds to L with the last column removed. The terms $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m-1}$ are

$$[A\ B] = -\frac{1}{\varepsilon}\, D\ \big[\ \nabla^2_{(u,v)}\mathcal{L}_{a,b}(u_*,v_*)\ \big]^{-1}, \tag{3.2.4}$$

with $D = [\mathrm{I}\ \mathbf{0}]$ the matrix concatenating the $n \times n$ identity matrix I and the matrix $\mathbf{0} \in \mathbb{R}^{n\times m-1}$ with all entries equal to zero. The operator $proj_{\mathrm{T}\Delta_n}$ denotes the projection onto

the tangent plane $\mathrm{T}\Delta_n = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 0\}$ *to the simplex* Δ_n.

Proof. Let us adopt the same notation as in the proof of Thm. 3.2. Recall that $\mathcal{L}$ is defined by

$$\mathcal{L}(\mathsf{a}, \mathsf{b}, u, v) = \langle \mathsf{a}, u \rangle + \langle \mathsf{b}, v \rangle - \varepsilon \sum_{i,j=1}^{n,m-1} e^{\frac{u_i + v_j - C_{ij}}{\varepsilon}}. \tag{B.3.1}$$

Since $\Psi = \nabla_{(u,v)}\mathcal{L}$, by a direct computation, Ψ can be written as

$$\Psi(\mathsf{a}, \mathsf{b}; u, v) = \begin{pmatrix} \mathsf{a} - T\mathbb{1} \\ \mathsf{b} - T^\top \mathbb{1} \end{pmatrix},$$

where T is the $n \times m-1$ matrix given by $\mathrm{diag}(\mathrm{e}^{\frac{u_*}{\varepsilon}})\mathrm{e}^{\frac{\bar{C}}{\varepsilon}}\mathrm{diag}(\mathrm{e}^{\frac{v_*}{\varepsilon}})$ and $\bar{C}$ is the matrix C with m^{th} column removed. In the following, we keep track of the dependence on a only. Being Ψ the gradient of $\mathcal{L}$, and $\gamma^*(\mathsf{a}) = (u_*(\mathsf{a}), v_*(\mathsf{a}))$ a stationary point, we have

$$\Psi(\mathsf{a}; \gamma^*(\mathsf{a})) = 0. \tag{B.3.2}$$

For the sake of clarity, notice that:

i) $\mathsf{a} \in \mathbb{R}^n$;

ii) $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{m-1} \longrightarrow \mathbb{R}$, as we are considering it is a function of a, u , v;

iii) $\Psi(\mathsf{a}, \gamma(\mathsf{a})) = \nabla_{u,v}\mathcal{L}(\mathsf{a}, \gamma(\mathsf{a})) \in \mathbb{R}^{n+m-1 \times 1}$;

iv) $u_* : \mathbb{R}^n \to \mathbb{R}^n$, $v_* : \mathbb{R}^n \to \mathbb{R}^{m-1}$, thus $\gamma^* : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^{m-1}$.

Our goal is to derive $\nabla_{\mathsf{a}}\gamma^*(\mathsf{a})$: by matrix differentiation rules (Kollo and von Rosen, 2006) and Eq. (B.3.2),

$$\nabla_{\mathsf{a}}\Psi(\mathsf{a}, \gamma^*(\mathsf{a})) = \nabla_1\Psi(\mathsf{a}, \gamma^*(\mathsf{a})) + \nabla_{\mathsf{a}}\gamma^*(\mathsf{a})\nabla_2\Psi(\mathsf{a}, \gamma^*(\mathsf{a})) = 0. \tag{B.3.3}$$

Let us analyse each term: $\nabla_1\Psi(\mathsf{a}, \gamma^*(\mathsf{a})) = [\mathbf{I}_n, \mathbf{0}_{n,m-1}]$ is $n \times n+m-1$ matrix with identity and zeros block, and $\nabla_2\Psi(\mathsf{a}, \gamma^*(\mathsf{a})) = \nabla^2_{u,v}\mathcal{L}(\mathsf{a}, \gamma^*(\mathsf{a})) =: H$ is the Hessian of $\mathcal{L}$ evaluated at $(\mathsf{a}, \gamma^*(\mathsf{a}))$, which is a $n+m-1 \times n+m-1$ matrix. Together with Eq. (B.3.3), this yields

$$\nabla_{\mathsf{a}}\gamma^*(\mathsf{a}) = [\nabla_{\mathsf{a}}u_*(\mathsf{a}), \nabla_{\mathsf{a}}v_*(\mathsf{a})] = -DH^{-1}.$$

For the sake of clarity, note that $\nabla_{\mathsf{a}}u_*(\mathsf{a})$ and $\nabla_{\mathsf{a}}v_*(\mathsf{a})$ contains the gradients of the compo-

nents as columns, i.e.

$$\nabla_{\mathsf{a}} u_* = \begin{pmatrix} \nabla_{\mathsf{a}} u_{*1}, & \nabla_{\mathsf{a}} u_{*2}, & \dots, & \nabla_{\mathsf{a}} u_{*n} \end{pmatrix}$$
$$\nabla_{\mathsf{a}} v_* = \begin{pmatrix} \nabla_{\mathsf{a}} v_{*1}, & \nabla_{\mathsf{a}} v_{*2}, & \dots, & \nabla_{\mathsf{a}} v_{*m-1} \end{pmatrix}.$$

Now, since $\widetilde{\mathsf{OT}}_\varepsilon(a, b) = \langle T_\varepsilon, C \rangle$ and T_ε corresponds to Eq. (3.2.1) a straightforward computation shows that

$$\nabla_{\mathsf{a}} \widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a}, \mathsf{b}) = \sum_{i,j=1}^{n,m} \nabla_a T_{\varepsilon ij} C_{ij} = \frac{1}{\varepsilon} \sum_{i,j=1}^{n,m} T_{\varepsilon ij} C_{ij} \nabla_{\mathsf{a}} u_{*i} + \frac{1}{\varepsilon} \sum_{i,j=1}^{n,m-1} T_{\varepsilon ij} C_{ij} \nabla_{\mathsf{a}} v_{*j}.$$

Setting $L := T_\varepsilon \odot C$, then the formula above can be rewritten as

$$\nabla_{\mathsf{a}} \widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a}, \mathsf{b}) = \frac{1}{\varepsilon} \sum_{i}^{n} \nabla_{\mathsf{a}} u_{*i} \sum_{j=1}^{m} L_{ij} + \frac{1}{\varepsilon} \sum_{j=1}^{m-1} \nabla_{\mathsf{a}} v_{*j} \sum_{i=1}^{n} L_{ij},$$

which is exactly

$$\nabla_{\mathsf{a}} \widetilde{\mathsf{OT}}_\varepsilon(\mathsf{a}, \mathsf{b}) = \frac{1}{\varepsilon} (\nabla_{\mathsf{a}} u_* L \mathbb{1}_m + \nabla_{\mathsf{a}} v_* \bar{L}^\top \mathbb{1}_n).$$

Since by definition, the gradient belongs to the tangent space of the domain, and $\mathsf{a} \in \Delta_n$, we project on the tangent space of the simplex, recovering $\mathrm{proj}_{\mathrm{T}\Delta_n} \frac{1}{\varepsilon} (\nabla_{\mathsf{a}} u_* L \mathbb{1}_m + \nabla_{\mathsf{a}} v_* \bar{L}^\top \mathbb{1}_n)$. □

B.3.1 Massaging the gradient to get an algorithmic-friendly form

In the proof of theorem 3.3 in Chapter 3 we have derived a formula for the gradient of Sinkhorn distance. In this section we further manipulate it in order to obtain an algorithmic friendly expression that also points out some interesting bits that were hidden in the formula above. All the notation has already been introduced: from now on, we will drop the ε and denote the optimal plan by T to make the notation neater.

An explicit computation of the second derivatives of $\mathcal{L}$ with respect to u_i and v_j for $i = 1, \dots, n$ and $j = 1, \dots, m-1$ leads to the following identity

$$H = \frac{1}{\varepsilon} \begin{pmatrix} \mathrm{diag}(T\mathbb{1}) & \bar{T} \\ \bar{T}^\top & \mathrm{diag}(\bar{T}^\top \mathbb{1}) \end{pmatrix} = \frac{1}{\varepsilon} \begin{pmatrix} \mathrm{diag}(\mathsf{a}) & \bar{T} \\ \bar{T}^\top & \mathrm{diag}(\bar{\mathsf{b}}) \end{pmatrix}.$$

That is, H is a block matrix and each block can be expressed in terms of the plan T. The block structure can be exploited when it comes to compute the inverse: we have shown that the gradient of the dual potentials is given by

$$[\nabla_{\mathsf{a}} u_*, \nabla_{\mathsf{a}} v_*] = -DH^{-1}, \qquad D = [\mathrm{I}_n, \mathbf{0}_{n,m-1}].$$

Now, the inverse of a block matrix is again a block matrix, say

$$H^{-1} = \varepsilon \begin{pmatrix} K_1 & K_2 \\ K_3 & K_4 \end{pmatrix}.$$

Then, $[\nabla_{\mathsf{a}} u_*, \nabla_{\mathsf{a}} v_*] = -\varepsilon[K_1, K_2]$. By the formula of the block inverse, setting

$$\mathcal{K} = \mathrm{diag}(T\mathbb{1}) - \bar{T}\mathrm{diag}(\bar{T}^\top \mathbb{1})^{-1}\bar{T}^\top,$$

the blocks K_1 and K_2 are given by

$$K_1 = \mathcal{K}^{-1}, \qquad K_2 = -\mathcal{K}^{-1}\bar{T}\mathrm{diag}(\bar{T}^\top \mathbb{1})^{-1}.$$

Note that $\mathcal{K}$ is symmetric and so its inverse. Now, we can rewrite $\frac{1}{\varepsilon}(\nabla_{\mathsf{a}} u_* L\mathbb{1}_m + \nabla_{\mathsf{a}} v_* \bar{L}^\top \mathbb{1}_n)$, with $L = T \odot C$, as

$$\frac{1}{\varepsilon}\varepsilon\big(- \mathcal{K}^{-1} L\mathbb{1}_m + \mathcal{K}^{-1}\bar{T}\mathrm{diag}(\bar{T}^\top \mathbb{1})^{-1}\bar{L}^\top \mathbb{1}_n\big)$$

and, since ε cancels out, we conclude that

$$\nabla_{\mathsf{a}} \widetilde{\mathsf{OT}}_\varepsilon(a, b) = \cdot\mathrm{solve}(\mathcal{K}, -L\mathbb{1}_m + \bar{T}\mathrm{diag}(\bar{T}^\top \mathbb{1})^{-1}\bar{L}^\top \mathbb{1}_n).$$

Comment on Remark 3.1. In the recent work (Altschuler et al., 2017), it has been shown that Sinkhorn-Knopp algorithm outputs a matrix T_ε whose distance $\|T_\varepsilon \mathbb{1} - \mathsf{a}\|_1 + \|T_\varepsilon^\top \mathbb{1} - \mathsf{b}\|_1$ from the transport polytope $\Pi(\mathsf{a}, \mathsf{b})$ is smaller than σ in $O(\sigma^{-2}\log(s/\ell))$ iterations, where $s = \sum_{ij} \mathrm{e}^{-\frac{C_{ij}}{\varepsilon}}$ and $\ell = \min_{ij} \mathrm{e}^{-\frac{C_{ij}}{\varepsilon}}$. Let us denote by $C_{\max}$ and $C_{\min}$ the maximum and

minimum elements of C respectively. Then,

$$\frac{s}{\ell} = \sum_{ij} \mathrm{e}^{-\frac{(C_{ij}-C_{\max})}{\varepsilon}} \geq \mathrm{e}^{-\frac{(C_{\min}-C_{\max})}{\varepsilon}} \geq 1.$$

This yields the lower bound

$$\log\left(\frac{s}{\ell}\right) \geq c\varepsilon^{-1}$$

where c is a constant independent of ε. We can then conclude that Sinkhorn-Knopp algorithm returns a matrix T_ε such that

$$\langle T_\varepsilon, C\rangle \leq \mathsf{W}(\mathsf{a},\mathsf{b}) + \sigma$$

in $O(n^2\sigma^{-2}C_{\max}^2\varepsilon^{-1})$.

Appendix C

Appendix of Chapter 4

Below we give an overview of the structure of the supplementary material and highlight the main novel results.

Appendix C.1: abstract Frank-Wolfe algorithm in dual Banach spaces. This section contains full details on Frank-Wolfe algorithm. The novelty stands in the relaxation of the differentiability assumptions.

Appendix C.2 Sinkhorn algorithm in infinite dimensions. This section contains the generalization of Sinkhorn algorithm, presented in Chapter 2, in the infinite dimensional setting.

Appendix C.3: Frank-Wolfe algorithm for Sinkhorn barycenters. This section contains the complete analysis of FW algorithm for Sinkhorn barycenters, which takes into account the error in the computation of Sinkhorn potentials and the error in their minimization. The main result is the convergence of the Frank-Wolfe scheme for finitely supported distributions in Thm. C.7.

Appendix C.4: Sample complexity of Sinkhorn potential and convergence of Alg. 4.2 in case of continuous measures. This section contains the discussion and the proofs of two of main results of the work Thm. 4.6, Thm. 4.10.

Appendix C.5: additional experiments. This section contains additional experiment on barycenters of mixture of Gaussian and the barycenter of meshes in 3D (dinosaur).

C.1 The Frank-Wolfe algorithm in dual Banach spaces

In this section we detail the convergence analysis of the Frank-Wolfe algorithm in abstract dual Banach spaces and under mild directional differentiablility assumptions so to cover the setting of Sinkhorn barycenters described in Sec. 4.2 in Chapter 4.

Let $\mathcal{W}$ be a real Banach space and let be $\mathcal{W}^*$ its topological dual. Let $\mathcal{D} \subset \mathcal{W}^*$ be a nonempty, closed, convex, and bounded set and let $\mathsf{G}\colon \mathcal{D} \to \mathbb{R}$ be a convex function. We address the following optimization problem

$$\min_{w \in \mathcal{D}} \mathsf{G}(w), \tag{C.1.1}$$

assuming that the set of solutions is nonemtpy.

We recall the concept of the tangent cone of feasible directions.

Definition C.1. *Let $w \in \mathcal{D}$. Then* the cone of feasible directions of $\mathcal{D}$ at w *is* $\mathcal{F}_{\mathcal{D}}(w) = \mathbb{R}_+(\mathcal{D} - w)$ *and the* tangent cone of $\mathcal{D}$ at w *is*

$$\mathcal{T}_{\mathcal{D}}(w) = \overline{\mathcal{F}_{\mathcal{D}}(w)} = \big\{v \in \mathcal{W}^* \,|\, \exists (t_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_{++} \text{ such that } t_k \to 0 \\ \text{and } \exists (w_k)_{k\in\mathbb{N}} \in \mathcal{D}^{\mathbb{N}} \text{ such that } t_k^{-1}(w_k - w) \to v\big\}.$$

Remark C.1. $\mathcal{F}_{\mathcal{D}}(w)$ is the cone generated by $\mathcal{D} - w$, and it is a convex cone. Indeed, if $t > 0$ and $v \in \mathcal{F}_{\mathcal{D}}(w)$, then $tv \in \mathcal{F}_{\mathcal{D}}(w)$. Moreover, if $v_1, v_2 \in \mathcal{F}_{\mathcal{D}}(w)$, then there exists $t_1, t_2 > 0$ and $w_1, w_2 \in \mathcal{D}$ such that $v_i = t_i(w_i - w)$, $i = 1, 2$. Thus,

$$v_1 + v_2 = (t_1 + t_2)\Big(\frac{t_1}{t_1 + t_2} w_1 + \frac{t_2}{t_1 + t_2} w_2 - w\Big) \in \mathbb{R}_+(\mathcal{D} - w).$$

So, $\mathcal{T}_{\mathcal{D}}(w)$ is a closed convex cone too.

Definition C.2. *Let $w \in \mathcal{D}$ and $v \in \mathcal{F}_{\mathcal{D}}(w)$. Then,* the directional derivative of G at w in the direction v *is*

$$\mathsf{G}'(w; v) = \lim_{t \to 0^+} \frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t} \in [-\infty, +\infty[\,.$$

Remark C.2. The above definition is well-posed. Indeed, since v is a feasible direction of

$\mathcal{D}$ at w, there exists $t_1 > 0$ and $w_1 \in \mathcal{D}$ such that $v = t_1(w_1 - w)$; hence

$$\forall\, t \in]0, 1/t_1]\,, \quad w + tv = w + t\,t_1(w_1 - w) = (1 - t\,t_1)w + t\,t_1 w_1 \in \mathcal{D}.$$

Moreover, since G is convex, the function $t \in]0, 1/t_1] \mapsto (\mathsf{G}(w + tv) - \mathsf{G}(w))/t$ is increasing, hence

$$\lim_{t \to 0^+} \frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t} = \inf_{t \in]0, 1/t_1]} \frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t}. \tag{C.1.2}$$

It is easy to prove that the function

$$v \in \mathcal{F}_{\mathcal{D}}(w) \mapsto \mathsf{G}'(w; v) \in [-\infty, +\infty[$$

is positively homogeneous and sublinear (hence convex), that is,

(i) $\forall\, v \in \mathcal{F}_{\mathcal{D}}(w)$ and $\forall\, t \in \mathbb{R}_+$, $\mathsf{G}'(w; tv) = t\mathsf{G}'(w; v)$;

(ii) $\forall\, v_1, v_2 \in \mathcal{F}_{\mathcal{D}}(w)$, $\mathsf{G}'(w; v_1 + v_2) \leq \mathsf{G}'(w; v_1) + \mathsf{G}'(w; v_2)$.

We make the following assumptions on G:

H1) $\forall\, w \in \mathcal{D}$, the function $v \mapsto \mathsf{G}'(w; v)$ is finite, that is, $\mathsf{G}'(w; v) \in \mathbb{R}$.

H2) The *curvature of* G is finite, that is,

$$C_{\mathsf{G}} = \sup_{\substack{w, z \in \mathcal{D} \\ \gamma \in [0,1]}} \frac{2}{\gamma^2} \big(\mathsf{G}(w + \gamma(z - w)) - \mathsf{G}(w) - \gamma \mathsf{G}'(w; z - w)\big) < +\infty. \tag{C.1.3}$$

Remark C.3. For every $w, z \in \mathcal{D}$, we have

$$\mathsf{G}(z) - \mathsf{G}(w) \geq \mathsf{G}'(w; z - w). \tag{C.1.4}$$

This follows from Eq. (C.1.2) with $w_1 = z$ and $t = 1$ ($t_1 = 1$).

The (inexact) Frank-Wolfe algorithm is detailed in Alg. C.1.

Remark C.4.

(i) Alg. C.1 does not require the sub-problem $\min_{z \in \mathcal{D}} \mathsf{G}'(w_k, z - w_k)$ to have solutions. Indeed it only requires computing a Δ_k-minimizer of $\mathsf{G}'(w_k; \cdot - w_k)$ on $\mathcal{D}$, which always exists.

(ii) Since $\mathcal{D}$ is weakly-$*$ compact (by Banach-Alaoglu theorem), if $\mathsf{G}'(w_k, \cdot - w_k)$ is weakly-$*$ continuous on $\mathcal{D}$, then the sub-problem $\min_{z \in \mathcal{D}} \mathsf{G}'(w_k, z - w_k)$ admits

Algorithm C.1 Frank-Wolfe in Dual Banach Spaces

Let $(\gamma_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_{++}$ be such that $\gamma_0 = 1$ and, for every $k \in \mathbb{N}$, $1/\gamma_k \leq 1/\gamma_{k+1} \leq 1/2+1/\gamma_k$ (i.e., $\gamma_k = 2/(k+2)$). Let $w_0 \in \mathcal{D}$ and $(\Delta_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_{+}$ be such that $(\Delta_k/\gamma_k)_{k\in\mathbb{N}}$ is nondecreasing. Then

for $k = 0, 1, \ldots$

find $z_{k+1} \in \mathcal{D}$ is such that $\mathsf{G}'(w_k; z_{k+1} - w_k) \leq \inf_{z\in\mathcal{D}} \mathsf{G}'(w_k; z - w_k) + \frac{1}{2}\Delta_k$

$w_{k+1} = w_k + \gamma_k(z_{k+1} - w_k)$

solutions. Note that this occurs when the directional derivative $\mathsf{G}'(w;\cdot)$ is linear and can be represented in $\mathcal{W}$. This case is addressed in the subsequent Prop. C.3.

Theorem C.1. *Let $(w_k)_{k\in\mathbb{N}}$ be defined according to Alg. C.1. Then, for every integer $k \geq 1$,*

$$\mathsf{G}(w_k) - \min \mathsf{G} \leq C_{\mathsf{G}}\gamma_k + \Delta_k. \tag{C.1.5}$$

Proof. Let $w_* \in \mathcal{D}$ be a solution of problem Eq. (C.1.1). It follows from Item H2) and the definition of w_{k+1} in Alg. C.1, that

$$\mathsf{G}(w_{k+1}) \leq \mathsf{G}(w_k) + \gamma_k \mathsf{G}'(w_k; z_{k+1} - w_k) + \frac{\gamma_k^2}{2} C_{\mathsf{G}}.$$

Moreover, it follows from the definition of z_{k+1} in Alg. C.1 and Eq. (C.1.4) that

$$\begin{aligned}
\mathsf{G}'(w_k; z_{k+1} - w_k) &\leq \inf_{z\in\mathcal{D}} \mathsf{G}'(w_k; z - w_k) + \frac{1}{2}\Delta_k \\
&\leq \mathsf{G}'(w_k; w_* - w_k) + \frac{1}{2}\Delta_k \\
&\leq -(\mathsf{G}(w_k) - \mathsf{G}(w_*)) + \frac{1}{2}\Delta_k.
\end{aligned}$$

Then,

$$\mathsf{G}(w_{k+1}) - \mathsf{G}(w_*) \leq (1-\gamma_k)(\mathsf{G}(w_k) - \mathsf{G}(w_*)) + \frac{\gamma_k^2}{2}\Big(C_{\mathsf{G}} + \frac{\Delta_k}{\gamma_k}\Big). \tag{C.1.6}$$

Now, similarly to (Jaggi, 2013, Theorem 2), we can prove (C.1.5) by induction. Since $\gamma_0 = 1$, $1/\gamma_1 \leq 1/2 + 1/\gamma_0$, and $\Delta_0/\gamma_0 \leq \Delta_1/\gamma_1$, it follows from Eq. (C.1.6) that

$$\mathsf{G}(w_1) - \mathsf{G}(w_*) \leq \frac{1}{2}\Big(C_{\mathsf{G}} + \frac{\Delta_0}{\gamma_0}\Big) \leq \gamma_1\Big(C_{\mathsf{G}} + \frac{\Delta_1}{\gamma_1}\Big), \tag{C.1.7}$$

hence (C.1.5) is true for $k = 1$. Set, for the sake of brevity, $C_k = C_{\mathsf{G}} + \Delta_k/\gamma_k$ and suppose that Eq. (C.1.5) holds for $k \in \mathbb{N}$, $k \geq 1$. Then, it follows from Eq. (C.1.6) and the properties of $(\gamma_k)_{k\in\mathbb{N}}$ that

$$\begin{aligned}
\mathsf{G}(w_{k+1}) - \mathsf{G}(w_*) &\leq (1-\gamma_k)\gamma_k C_k + \frac{\gamma_k^2}{2} C_k \\
&= C_k \gamma_k \Big(1 - \frac{\gamma_k}{2}\Big) \\
&\leq C_k \gamma_k \Big(1 - \frac{\gamma_{k+1}}{2}\Big) \\
&\leq C_k \frac{1}{1/\gamma_{k+1} - 1/2} \Big(1 - \frac{\gamma_{k+1}}{2}\Big) \\
&= C_k \gamma_{k+1} \\
&\leq C_{k+1}\gamma_{k+1}. \qquad \square
\end{aligned}$$

Corollary C.2. *Under the assumptions of Thm. C.1, suppose in addition that $\Delta_k = \Delta\gamma_k^\zeta$, for some $\zeta \in [0,1]$ and $\Delta \geq 0$. Then we have*

$$\mathsf{G}(w_k) - \min \mathsf{G} \leq C_{\mathsf{G}}\gamma_k + \Delta\gamma_k^\zeta. \tag{C.1.8}$$

Proof. It follows from Thm. C.1 by noting that the sequence $\Delta_k/\gamma_k = \Delta/\gamma_k^{1-\zeta}$ is nondecreasing. $\square$

Proposition C.3. *Suppose that there exists a mapping $\nabla\mathsf{G}\colon \mathcal{D} \to \mathcal{W}$ such that*[1]*,*

$$\forall\, w \in \mathcal{D},\ \forall\, z \in \mathcal{D} \quad \langle \nabla\mathsf{G}(w), z - w\rangle = \mathsf{G}'(w; z - w). \tag{C.1.9}$$

Then the following holds.

(i) Let $k \in \mathbb{N}$ and suppose that there exists $u_k \in \mathcal{W}$ such that $\|u_k - \nabla\mathsf{G}(w_k)\| \leq \Delta_{1,k}/4$ and that $z_{k+1} \in \mathcal{D}$ satisfies

$$\langle u_k, z_{k+1}\rangle \leq \min_{z\in\mathcal{D}} \langle u_k, z\rangle + \frac{\Delta_{2,k}}{2},$$

for some $\Delta_{1,k}, \Delta_{2,k} > 0$. Then

$$\mathsf{G}'(w_k; z_{k+1} - w_k) \leq \min_{z\in\mathcal{D}} \mathsf{G}'(w_k; z - w_k) + \frac{1}{2}(\Delta_{1,k}\mathrm{diam}(\mathcal{D}) + \Delta_{2,k}). \tag{C.1.10}$$

[1] This mapping does not need to be unique.

(ii) Suppose that $\nabla \mathsf{G}\colon \mathcal{D} \to \mathcal{W}$ is L-Lipschitz continuous for some $L > 0$. Then, for every $w, z \in \mathcal{D}$ and $\gamma \in [0,1]$,

$$\mathsf{G}(w + \gamma(z-w)) - \mathsf{G}(w) - \gamma \langle z - w, \nabla \mathsf{G}(w) \rangle \leq \frac{L}{2}\gamma^2 \|z-w\|^2$$

and hence $C_{\mathsf{G}} \leq L\mathrm{diam}(\mathcal{D})^2$.

Proof. Item (i): we have

$$\begin{aligned}\langle \nabla \mathsf{G}(w_k), z_{k+1} - w_k \rangle &= \langle u_k, z_{k+1} - w_k \rangle + \langle \nabla \mathsf{G}(w_k) - u_k, z_{k+1} - w_k \rangle \\ &\leq \min_{z \in \mathcal{D}} \langle u_k, z - w_k \rangle + \frac{\Delta_{2,k}}{2} + \frac{\Delta_{1,k}}{4}\mathrm{diam}(\mathcal{D}). \end{aligned} \tag{C.1.11}$$

Moreover,

$$\begin{aligned}\forall\, z \in \mathcal{D}, \quad \langle u_k, z - w_k \rangle &= \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \langle u_k - \nabla \mathsf{G}(w_k), z - w_k \rangle \\ &\leq \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_{1,k}}{4}\mathrm{diam}(\mathcal{D}),\end{aligned}$$

hence

$$\min_{z \in \mathcal{D}} \langle u_k, z - w_k \rangle \leq \min_{z \in \mathcal{D}} \langle \nabla \mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_{1,k}}{4}\mathrm{diam}(\mathcal{D}). \tag{C.1.12}$$

Thus, Eq. (C.1.10) follows from Eq. (C.1.11), Eq. (C.1.12), and Eq. (C.1.9).

Item (ii): let $w, z \in \mathcal{D}$, and define $\psi\colon [0,1] \to \mathcal{W}^*$ such that $\psi(\gamma) = \mathsf{G}(w + \gamma(z-w))$ $\forall\, \gamma \in [0,1]$. Then, it is easy to see that for every $\gamma \in]0,1[$, ψ is differentiable at γ and $\psi'(\gamma) = \mathsf{G}'(w + \gamma(z-w); z-w) = \langle \nabla \mathsf{G}(w + \gamma(z-w)), z - w \rangle$. Moreover, ψ is continuous on $[0,1]$. Therefore, the fundamental theorem of calculus yields

$$\psi(\gamma) - \psi(0) = \int_0^\gamma \psi'(t)dt$$

and hence

$$\begin{aligned}\mathsf{G}(w+\gamma(z-w))-\mathsf{G}(w)-\gamma\langle\nabla\mathsf{G}(w),z-w\rangle &= \int_0^\gamma \langle\nabla\mathsf{G}(w+t(z-w))-\nabla\mathsf{G}(w),z-w\rangle\,dt\\ &\le \int_0^\gamma \|\nabla\mathsf{G}(w+t(z-w))-\nabla\mathsf{G}(w)\|\|z-w\|dt\\ &\le \int_0^\gamma Lt\|z-w\|^2dt\\ &= L\frac{\gamma^2}{2}\|z-w\|^2, \qquad \square\end{aligned}$$

where we used Cauchy-Schwarz inequality and the Lipschiz continuity of $\nabla\mathsf{G}$.

C.2 Sinkhorn algorithm in infinite dimesional setting

In the context of optimal transport, Sinkhorn-Knopp algorithm is often presented and studied in finite dimension (Cuturi, 2013; Peyré and Cuturi, 2019). The algorithm originates from the so called *matrix scaling problems*, also called *DAD problems*, which consists in finding, for a given matrix A with nonnegative entries, two diagonal matrices D_1, D_2 such that D_1AD_2 is doubly stochastic (Sinkhorn and Knopp, 1967). In our setting it is crucial to analyze the algorithm in infinite dimension.

Thm. A.7 shows that $\mathsf{A}_{\beta\alpha}$ is a contraction with respect to the Hilbert's metric. This suggests a direct approach to find the solutions of the DAD problem by adopting a fixed-point strategy, which amounts to applying the operators A_α and A_β alternatively, starting from some $f^{(0)} \in \mathcal{C}_{++}(\mathcal{X})$. This is exactly the approach to the Sinkhorn algorithm pioneered by Menon (1967) and Franklin and Lorenz (1989) and further developed in an infinite dimensional setting in Nussbaum (1993). In this section we review the algorithm and give the convergence properties for the special kernel k in (A.4.1). In particular we provide rate of convergence in the sup norm $\|\cdot\|_\infty$.

Algorithm C.2 Sinkhorn-Knopp algorithm (infinite dimensional case)

Let $\alpha, \beta \in \mathcal{P}(\mathcal{X})$. Let $f^{(0)} \in \mathcal{C}_{++}(\mathcal{X})$ and define,

$$\begin{aligned}&\text{for } \ell = 0, 1, \dots\\ &\quad g^{(\ell+1)} = \mathsf{A}_\alpha(f^{(\ell)})\\ &\quad f^{(\ell+1)} = \mathsf{A}_\beta(g^{(\ell+1)})\end{aligned}$$

Theorem C.4 (Convergence of Sinkhorn-Knopp algorithm)**.** *Let* D *be the diameter of the domain* $\mathcal{X}$ *and recall that* $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$. *Let* $(f^{(\ell)})_{\ell\in\mathbb{N}}$ *be defined according to Alg. C.2. Let* $x_o \in \mathcal{X}$ *and let* (f, g) *be the solution of the DAD problem Eq. (A.4.27) such that* $f(x_o) = 1$. *Then, defining for every* $\ell \in \mathbb{N}$, $\tilde{f}^{(\ell)} = f^{(\ell)}/f^{(\ell)}(x_o)$ *and* $\tilde{g}^{(\ell+1)} = g^{(\ell+1)}f^{(\ell)}(x_o)$, *we have*

$$\begin{cases} \|\log \tilde{f}^{(\ell)} - \log f\|_\infty \le \lambda^{2\ell}\left(\dfrac{\mathsf{D}}{\varepsilon} + \log \dfrac{\|f^{(0)}\|_\infty}{\min_{\mathcal{X}} f^{(0)}}\right) \\ \|\log \tilde{g}^{(\ell+1)} - \log g\|_\infty \le e^{3\mathsf{D}/\varepsilon}\|\log \tilde{f}^{(\ell)} - \log f\|_\infty. \end{cases} \tag{C.2.1}$$

Moreover, let the potentials $(u, v) = (\varepsilon \log f, \varepsilon \log g)$ *and set* $(\tilde{u}^{(\ell)}, \tilde{v}^{(\ell)}) = (\varepsilon \log \tilde{f}^{(\ell)}, \varepsilon \log \tilde{g}^{(\ell)})$ *for every* $\ell \in \mathbb{N}$. *Then we have*

$$\|\tilde{u}^{(\ell)} - u\|_\infty \le \lambda^{2\ell}\left(\frac{\mathsf{D} + \max_{\mathcal{X}} u^{(0)} - \min_{\mathcal{X}} u^{(0)}}{\varepsilon}\right). \tag{C.2.2}$$

Proof. Let $\mathcal{A}$ be the set defined in Lemma 4.3. For every $\ell \in \mathbb{N}$, we have $f^{(\ell+1)} = \mathsf{A}_{\beta\alpha}(f^{(\ell)})$ and $\tilde{f}^{\ell} \in \mathcal{A}$. Thus, it follows from Thm. A.7 and (4.3.4) in Lemma 4.3 that, for every $\ell \in \mathbb{N}$,

$$\|\log \tilde{f}^{(\ell)} - \log f\|_\infty \le d_H(\tilde{f}^{\ell}, f) = d_H(\mathsf{A}^{(\ell)}_{\beta\alpha}(f^{(0)}), f) \le \lambda^{2\ell} d_H(f^{(0)}, f).$$

Moreover, recalling (A.4.11), we have

$$d_H(f^{(0)}, f) = d_H(1/f^{(0)}, \mathsf{L}_\beta g) = \log \max_{x,y\in\mathcal{X}} \frac{f^{(0)}(y)\mathsf{L}_\beta g(y)}{f^{(0)}(x)\mathsf{L}_\beta g(x)} \le \log\left[e^{\mathsf{D}/\varepsilon} \max_{x,y\in\mathcal{X}} \frac{f^{(0)}(y)}{f^{(0)}(x)}\right]$$

where we used the fact that $\mathsf{L}_\beta(\mathcal{C}_{++}(\mathcal{X})) \subset K$, with K defined in Lemma A.8 and the definition (A.4.33). Thus, the first inequality in (C.2.1) follows. The second inequality in (C.2.1) and (C.2.2) follow directly from Lemma A.10 and the fact that $u^{(0)} = \varepsilon \log f^{(0)}$. □

Proposition C.5. *Suppose that* α *and* β *are probability measures with finite support. Then Alg. C.2 can be reduced to the finite dimensional Alg. 2.1. More specifically, suppose that* $\alpha = \sum_{i=1}^{n_1} a_i\delta_{x_i}$, *and* $\beta = \sum_{i=1}^{n_2} b_i\delta_{y_i}$, *where* $\mathsf{a} = (a_i)_{1\le i\le n_1} \in \mathbb{R}^{n_1}_+$, $\sum_{i=1}^{n} a_i = 1$ *and* $\mathsf{b} = (b_i)_{1\le i\le n_2} \in \mathbb{R}^{n_2}_+$, $\sum_{i=1}^{n} b_i = 1$. *Let* $\mathsf{K} \in \mathbb{R}^{n_1\times n_2}$ *be such that* $\mathsf{K}_{i_1,i_2} = \mathsf{k}(x_{i_1}, y_{i_2})$ *and let* $\mathsf{M} = \mathrm{diag}(\mathsf{a})\mathsf{K}\mathrm{diag}(\mathsf{b}) \in \mathbb{R}^{n_1\times n_2}$. *Let* $(\mathsf{f}^{(\ell)})_{\ell\in\mathbb{N}}$ *and* $(f^{(\ell)})_{\ell\in\mathbb{N}}$ *be defined according to*

Alg. 2.1 and Alg. C.2 respectively, with $\mathsf{f}^{(0)} = (f^{(0)}(x_i))_{1\le i\le n_1}$. *Then, for every* $\ell \in \mathbb{N}$,

$$g^{(\ell+1)}(y)^{-1} = \sum_{i_1=1}^{n_1} k(x_{i_1}, y) a_{i_1} \mathsf{f}_{i_1}^{(\ell)} \qquad \forall\, y \in \mathcal{X}$$
$$f^{(\ell+1)}(x)^{-1} = \sum_{i_2=1}^{n_2} k(x, y_{i_2}) b_{i_2} \mathsf{g}_{i_2}^{(\ell+1)} \qquad \forall\, x \in \mathcal{X}.$$

Moreover, setting $u^{(\ell)} = \varepsilon \log f^{(\ell)}$, $v^{(\ell)} = \varepsilon \log g^{(\ell)}$, $\mathsf{u}^{(\ell)} = \varepsilon \log \mathsf{f}^{(\ell)}$, *and* $\mathsf{v}^{(\ell)} = \varepsilon \log \mathsf{g}^{(\ell)}$, *we have*

$$\begin{cases} v^{(\ell+1)}(y) = -\varepsilon \log \sum_{i_1=1}^{n_1} \exp(\mathsf{u}_{i_1}^{(\ell)} - \mathsf{c}(x_{i_1}, y)) a_{i_1} & \forall\, y \in \mathcal{X} \\ u^{(\ell+1)}(x) = -\varepsilon \log \sum_{i_2=1}^{n_2} \exp(\mathsf{v}_{i_2}^{(\ell+1)} - \mathsf{c}(x, y_{i_2})) b_{i_2} & \forall\, x \in \mathcal{X}. \end{cases} \tag{C.2.3}$$

Proof. Since α and β have finite support, we derive from the definitions of $f^{(\ell+1)}$ and $g^{(\ell+1)}$ in Alg. C.2 and those of A_α and A_β that

$$\begin{cases} g^{(\ell+1)}(y)^{-1} = (\mathsf{L}_\alpha f^{(\ell)}))(y) = \sum_{i_1=1}^{n_1} a_{i_1} \mathsf{k}(x_{i_1}, y) f^{(\ell)}(x_{i_1}) & \forall\, x \in \mathcal{X} \\ f^{(\ell+1)}(x)^{-1} = (\mathsf{L}_\beta g^{(\ell+1)}))(x) = \sum_{i_2=1}^{n_2} \mathsf{k}(x, y_{i_2}) b_{i_2} g^{(\ell+1)}(y_{i_2}) & \forall\, y \in \mathcal{X}. \end{cases}$$

Now, multiplying the above equations by b_{i_2} and a_{i_1} respectively, and recalling that $\mathsf{M}_{i_1,i_2} = a_{i_1} \mathsf{k}(x_{i_1}, y_{i_2}) b_{i_2}$, we have

$$\begin{bmatrix} b_1 g^{(\ell+1)}(y_1)^{-1} \\ \vdots \\ b_{n_2} g^{(\ell+1)}(y_{n_2})^{-1} \end{bmatrix} = \mathsf{M}^\top \begin{bmatrix} f^{(\ell)}(x_1) \\ \vdots \\ f^{(\ell)}(x_{n_1}) \end{bmatrix}, \quad \begin{bmatrix} a_1 f^{(\ell+1)}(x_1)^{-1} \\ \vdots \\ a_{n_1} f^{(\ell+1)}(x_{n_1})^{-1} \end{bmatrix} = \mathsf{M} \begin{bmatrix} g^{(\ell+1)}(y_1) \\ \vdots \\ g^{(\ell+1)}(y_{n_2}) \end{bmatrix},$$

and hence

$$\begin{bmatrix} g^{(\ell+1)}(y_1) \\ \vdots \\ g^{(\ell+1)}(y_{n_2}) \end{bmatrix} = \mathsf{b} \Big/ \mathsf{M}^\top \begin{bmatrix} f^{(\ell)}(x_1) \\ \vdots \\ f^{(\ell)}(x_{n_1}) \end{bmatrix}, \quad \begin{bmatrix} f^{(\ell+1)}(x_1) \\ \vdots \\ f^{(\ell+1)}(x_{n_1}) \end{bmatrix} = \mathsf{a} \Big/ \mathsf{M} \begin{bmatrix} g^{(\ell+1)}(y_1) \\ \vdots \\ g^{(\ell+1)}(y_{n_2}) \end{bmatrix}.$$

Therefore, since $\mathsf{f}^{(0)} = (f^{(0)}(x_i))_{1\le i\le n_1}$, recalling Alg. 2.1, it follows by induction that, for

every $\ell \in \mathbb{N}$, $\mathsf{f}^{(\ell)} = (f^{(\ell)}(x_i))_{1\le i\le n_1}$ and $\mathsf{g}^{(\ell)} = (g^{(\ell)}(x_i))_{1\le i\le n_1}$. Thus, the first part of the statement follows. The second part follows directly from the definitions of $u^{(\ell)}$, $v^{(\ell)}$, $\mathsf{u}^{(\ell)}$, and $\mathsf{v}^{(\ell)}$. □

C.3 Frank-Wolfe algorithm for Sinkhorn barycenters

In this section we finally analyze the Frank-Wolfe algorithm for the Sinkhorn barycenters and give convergence results. The following result is a direct consequence of the convergence of Sinkhorn algorithm Thm. C.4 and the definition of gradient of OT_ε, presented in the background material in (2.3.33) and recalled in Chapter 4 in Remark 4.1.

Theorem C.6. *Let D be the diameter of the domain $\mathcal{X}$ and let $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$. Let $(\tilde{u}^{(\ell)})_{\ell\in\mathbb{N}}$ be generated with Sinkhorn algorithm in the continuous setting, recalled in Alg. C.2. Then,*

$$\forall\, \ell \in \mathbb{N}, \quad \|\tilde{u}^{(\ell)} - \nabla_1 \mathsf{OT}_\varepsilon(\alpha,\beta)\|_\infty \le \lambda^{2\ell}\left(\frac{\mathsf{D} + \max_{\mathcal{X}} u^{(0)} - \min_{\mathcal{X}} u^{(0)}}{\varepsilon}\right), \tag{C.3.1}$$

where $u^{(\ell)} = \varepsilon \log f^{(\ell)}$ and $\tilde{u}^{(\ell)} = u^{(\ell)} - u^{(\ell)}(x_o)$.

In view of Prop. 2.17, Thm. C.6, and Prop. C.3, we can address the problem of the Sinkhorn barycenter (4.2.1) via the Frank-Wolfe Alg. C.1. Note that, according to Prop. C.3(ii), since the diameter of $\mathcal{P}(\mathcal{X})$ with respect to $\|\cdot\|_{\mathsf{TV}}$ is 2, we have that the curvature of B_ε is upper bounded by

$$C_{\mathsf{B}_\varepsilon} \le 24\varepsilon e^{3\mathsf{D}/\varepsilon}. \tag{C.3.2}$$

Let $k \in \mathbb{N}$ and α_k be the current iteration. For every $j \in \{1,\dots,m\}$, we can compute $\nabla_1 \mathsf{OT}_\varepsilon(\alpha_k,\beta_j)$ and $\nabla_1 \mathsf{OT}_\varepsilon(\alpha_k,\alpha_k)$ by the Sinkhorn-Knopp algorithm. By (C.3.1), we can find $\ell \in \mathbb{N}$ large enough so that $\|\tilde{u}_j^{(\ell)} - \nabla_1 \mathsf{OT}_\varepsilon(\alpha_k,\beta_j)\|_\infty \le \Delta_{1,k}/8$ and $\|\tilde{p}^{(\ell)} - \nabla_1 \mathsf{OT}_\varepsilon(\alpha_k,\alpha_k)\|_\infty \le \Delta_{1,k}/8$ and we set

$$\tilde{u}^{(\ell)} := \sum_{j=1}^m \omega_j \tilde{u}_j^{(\ell)} - \tilde{p}^{(\ell)}. \tag{C.3.3}$$

Then,

$$\|\tilde{u}^{(\ell)} - \nabla \mathsf{B}_\varepsilon(\alpha_k)\|_\infty \le \frac{\Delta_{1,k}}{4}. \tag{C.3.4}$$

Now, Frank-Wolf Alg. C.1 (in the version considered in Prop. C.3(i)) requires finding

$$\eta_{k+1} \in \operatorname*{argmin}_{\eta \in \mathcal{P}(\mathcal{X})} \langle \tilde{u}^{(\ell)}, \eta - \alpha_k \rangle \tag{C.3.5}$$

and making the update

$$\alpha_{k+1} = (1 - \gamma_k)\alpha_k + \gamma_k \eta_{k+1}. \tag{C.3.6}$$

Since the solution of (C.3.5) is a Dirac measure (see Sec. 4.5 in Chapter 4), the algorithm reduces to

$$\begin{cases} \text{find } x_{k+1} \in \mathcal{X} \text{ such that } \tilde{u}^{(\ell)}(x_{k+1}) \leq \min_{x \in \mathcal{X}} \tilde{u}^{(\ell)}(x) + \dfrac{\Delta_{2,k}}{2} \\ \alpha_{k+1} = (1 - \gamma_k)\alpha_k + \gamma_k \delta_{x_{k+1}}. \end{cases} \tag{C.3.7}$$

So, if we initialize the algorithm with $\alpha_0 = \delta_{x_0}$, then any α_k will be a discrete probability measure with support contained in $\{x_0, \dots, x_k\}$. This implies that if all the β_j's are probability measures with finite support, the computation of $\nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \beta_j)$ by the Sinkhorn algorithm can be reduced to a fully discrete algorithm, as showed in Prop. C.5. More precisely, assume that

$$\beta_j = \sum_{i_2=0}^{n} b_{j,i_2} \delta_{y_{j,i_2}} \qquad \forall\, j = 1, \dots, m, \tag{C.3.8}$$

and that at iteration k we have

$$\alpha_k = \sum_{i_1=0}^{k} a_{k,i_1} \delta_{x_{i_1}}. \tag{C.3.9}$$

Set $\mathsf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be defined as $\mathsf{k}(x, y) = e^{\frac{-\mathsf{c}(x,y)}{\varepsilon}}$, and

$$\mathsf{a}_k = \begin{bmatrix} a_{k,0} \\ \vdots \\ a_{k,k} \end{bmatrix} \in \mathbb{R}^{k+1}, \quad \mathsf{M}_0 = \begin{bmatrix} a_{k,0}\mathsf{k}(x_0, x_0)a_{k,0} & \dots & a_{k,0}\mathsf{k}(x_0, x_k)a_{k,k} \\ \vdots & \ddots & \vdots \\ a_{k,k}\mathsf{k}(x_k, x_0)a_{k,0} & \dots & a_{k,k}\mathsf{k}(x_k, x_k)a_{k,k} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)} \tag{C.3.10}$$

and, for every $j = 1 \ldots, m$,

$$\mathsf{b}_j = \begin{bmatrix} b_{j,0} \\ \vdots \\ b_{j,n} \end{bmatrix} \in \mathbb{R}^{n+1}, \ \mathsf{M}_j = \begin{bmatrix} a_{k,0}\mathsf{k}(x_0, y_{j,0})b_{j,0} & \ldots & a_{k,0}\mathsf{k}(x_0, y_{j,n})b_{j,n} \\ \vdots & \ddots & \vdots \\ a_{k,k}\mathsf{k}(x_k, y_{j,0})b_{j,0} & \ldots & a_{k,n}\mathsf{k}(x_k, y_{j,n})b_{j,n} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(n+1)}. \tag{C.3.11}$$

Then, run Alg. 2.1, with input a_k, and M_0 to get $(\mathsf{e}^{(\ell)}, \mathsf{h}^{(\ell)})$, and, for every $j = 1, \ldots, m$, with input a_k, b_j, and M_j to get $(\mathsf{f}_j^{(\ell)}, \mathsf{g}_j^{(\ell)})$. So, we have,

$$\forall\, \ell \in \mathbb{N} \quad \begin{cases} \mathsf{h}^{(\ell+1)} = \dfrac{\mathsf{a}_k}{\mathsf{M}_0^\top \mathsf{e}^{(\ell)}}, & \mathsf{e}^{(\ell+1)} = \dfrac{\mathsf{a}_k}{\mathsf{M}_0 \mathsf{h}^{(\ell+1)}} \\ \mathsf{g}_j^{(\ell+1)} = \dfrac{\mathsf{b}_j}{\mathsf{M}_j^\top \mathsf{f}_j^{(\ell)}}, & \mathsf{f}_j^{(\ell+1)} = \dfrac{\mathsf{a}_k}{\mathsf{M}_j \mathsf{g}_j^{(\ell+1)}} \quad \forall\, j = 1, \ldots, m. \end{cases} \tag{C.3.12}$$

According to Prop. C.5, for every $\ell \in \mathbb{N}$, we have

$$\forall\, x \in \mathcal{X} \quad \begin{cases} e^{(\ell)}(x)^{-1} = \displaystyle\sum_{i_2=0}^{k} \mathsf{k}(x, x_{i_2}) \mathsf{h}_{i_2}^{(\ell-1)} a_{k,i_2}, \\ p^{(\ell)}(x) = \varepsilon \log e^{(\ell)}(x) = -\varepsilon \log \displaystyle\sum_{i_2=0}^{k} \mathsf{k}(x, x_{i_2}) \mathsf{h}_{i_2}^{(\ell-1)} a_{k,i_2} \\ \tilde{p}^{(\ell)}(x) = p^{(\ell)}(x) - p^{(\ell)}(x_o), \end{cases} \tag{C.3.13}$$

and, for every $j = 1, \ldots, m$,

$$(\forall\, x \in \mathcal{X}) \quad \begin{cases} f_j^{(\ell)}(x)^{-1} = \displaystyle\sum_{i_2=0}^{n} \mathsf{k}(x, y_{i_2}) \mathsf{g}_{j,i_2}^{(\ell-1)} b_{j,i_2}, \\ u_j^{(\ell)}(x) = \varepsilon \log f_j^{(\ell)}(x) = -\varepsilon \log \displaystyle\sum_{i_2=0}^{n} \mathsf{k}(x, y_{i_2}) \mathsf{g}_{j,i_2}^{(\ell-1)} b_{j,i_2} \\ \tilde{u}_j^{(\ell)}(x) = u_j^{(\ell)}(x) - u_j^{(\ell)}(x_o). \end{cases} \tag{C.3.14}$$

Since the $\tilde{u}_j^{(\ell)}$'s and $u_j^{(\ell)}$'s, and $\tilde{p}^{(\ell)}$ and $p^{(\ell)}$, differ by a constant only, the final algorithm can be written as in Alg. C.3. We stress that this algorithm is more theoretically accurate than Alg. 4.2 since, in the computation of the Sinkhorn potentials and in their minimization, errors have been taken into account.

We now give a final convergence theorem, of which Thm. 4.9 in Chapter 4 is a special case.

Theorem C.7. *Suppose that $\beta_1, \ldots, \beta_m \in \mathcal{P}(\mathcal{X})$ are probability measures with finite*

Algorithm C.3 Frank-Wolfe algorithm for Sinkhorn barycenter

Let $\alpha_0 = \delta_{x_0}$ for some $x_0 \in \mathcal{X}$. Let $(\Delta_k)_{k\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ be such that Δ_k/γ_k is nondecreasing. Define

for $k = 0, 1, \dots$
- run Alg. 2.1 with input $\mathsf{a}_k, \mathsf{a}_k, \mathsf{M}_0$ till $\lambda^{2\ell}\mathsf{D}/\varepsilon \le \frac{\Delta_{1,k}}{8} \to \mathsf{h} \in \mathbb{R}^{k+1}$
- compute p via (C.3.13) with h
- for $j = 1, \dots m$
  - run Alg. 2.1 with input $\mathsf{a}_k, \mathsf{b}_j, \mathsf{M}_j$ till $\lambda^{2\ell}\mathsf{D}/\varepsilon \le \frac{\Delta_{1,k}}{8} \to \mathsf{g}_j \in \mathbb{R}^{n+1}$
  - compute u_j via (C.3.14) with g_j
- set $u = \sum_{j=1}^m \omega_j u_j - p$
- find $x_{k+1} \in \mathcal{X}$ such that $u(x_{k+1}) \le \min_{x\in\mathcal{X}} u(x) + \dfrac{\Delta_{2,k}}{2}$
- $\alpha_{k+1} = (1-\gamma_k)\alpha_k + \gamma_k\delta_{x_{k+1}}$.

support, each of cardinality $n \in \mathbb{N}$. Let $(\alpha_k)_{k\in\mathbb{N}}$ be generated by Alg. C.3. Then, for every $k \in \mathbb{N}$,

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha\in\mathcal{P}(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \le \gamma_k 24\varepsilon e^{3\mathsf{D}/\varepsilon} + 2\Delta_{1,k} + \Delta_{2,k} \tag{C.3.15}$$

Proof. It follows from Thm. C.1, Prop. C.3, and (C.3.2), recalling that $\operatorname{diam}(\mathcal{P}(\mathcal{X})) = 2$. □

C.4 Sample complexity of Sinkhorn potential

In the following we will denote by $\mathcal{C}^s(\mathcal{X})$ the space of s-differentiable functions with continuous derivatives and by $W^{s,p}(\mathcal{X})$ the Sobolev space of functions $f\colon \mathcal{X} \to \mathbb{R}$ with p-summable weak derivatives up to order s (Adams and Fournier, 2003). We denote by $\|\cdot\|_{s,p}$ the corresponding norm.

The following result shows that under suitable smoothness assumptions on the cost function c, the Sinkhorn potentials are uniformly bounded as functions in a suitable Sobolev space of corresponding smoothness. This fact will play a key role in approximating the Sinkhorn potentials of general distributions in practice. We recall the result on this Sobolev regularity of the potentials below.

Theorem C.8 (Proposition 2 in (Genevay et al., 2018a))**.** *Let $\mathcal{X}$ be a closed bounded domain with Lipschitz boundary in $\mathbb{R}^d$ (Adams and Fournier, 2003, Definition 4.9) and let $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X}\times\mathcal{X})$. Then for every $(\alpha,\beta) \in \mathcal{P}(\mathcal{X})^2$, the associated Sinkhorn potentials*

$(u, v) \in \mathcal{C}(\mathcal{X})^2$ are functions in $W^{s,\infty}(\mathcal{X})$. Moreover, let $x_o \in \mathcal{X}$. Then there exists a constant $\mathsf{r} > 0$, depending only on ε, s and $\mathcal{X}$, such that for every $(\alpha, \beta) \in \mathcal{P}(\mathcal{X})^2$ the associated Sinkhorn potentials $(u, v) \in \mathcal{C}(\mathcal{X})^2$ with $u(x_o) = 0$ satisfy $\|u\|_{s,\infty}, \|v\|_{s,\infty} \leq \mathsf{r}$.

In the original statement of (Genevay et al., 2018a, Proposition 2) the above result is formulated for $\mathsf{c} \in \mathcal{C}^\infty(\mathcal{X} \times \mathcal{X})$ for simplicity. However, as clarified by the authors, it holds also for the more general case $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X}'/)$.

Lemma C.9. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a closed bounded domain with Lipschitz boundary and let $u, u' \in W^{s,\infty}(\mathcal{X})$. Then the following holds*

(i) $\|uu'\|_{s,\infty} \leq \mathsf{m}_1 \|u\|_{s,\infty} \|u'\|_{s,\infty}$,

(ii) $\|e^u\|_{s,\infty} \leq \|e^u\|_\infty (1 + \mathsf{m}_2 \|u\|_{s,\infty})$,

where $\mathsf{m}_1 = \mathsf{m}_1(s, d)$ and $\mathsf{m}_2 = \mathsf{m}_2(s, d) > 0$ depend only on the dimension d and the order of differentiability s but not on u and u'.

Proof. Item (i) follows directly from Leibniz formula. To see Item (ii), let $\mathbf{i} = (i_1, \dots, i_d) \in \mathbb{N}^d$ be a multi-index with $|\mathbf{i}| = \sum_{\ell=1}^d i_\ell \leq s$ and note that by chain rule the derivatives of e^u can be written as

$$D^{\mathbf{i}}\, e^u = e^u\, P_{\mathbf{i}}\Big((D^{\mathbf{j}} u)_{\mathbf{j} \leq \mathbf{i}}\Big),$$

where $P_{\mathbf{i}}$ is a polynomial of degree $|\mathbf{i}|$ and $\mathbf{j} \leq \mathbf{i}$ is the ordering associated to the cone of non-negative vectors in $\mathbb{R}^d$. Note that $P_0 = 1$, while for $|\mathbf{i}| > 0$, the associated polyomial $P_{\mathbf{i}}$ has a root in zero (i.e. it does not have constant term). Hence

$$\|e^u\|_{s,\infty} \leq \|e^u\|_\infty \left(1 + |P|\Big((\|D^{\mathbf{i}} u\|_\infty)_{|\mathbf{i}| \leq s}\Big)\right),$$

where we have denoted by $P = \sum_{0 < |\mathbf{i}| \leq s} P_{\mathbf{i}}$ and by $|P|$ the polynomial with coefficients corresponding to the absolute value of the coefficients of P. Therefore, since $\|D^{\mathbf{i}} u\|_\infty \leq \|u\|_{s,\infty}$ for any $|\mathbf{i}| \leq s$, by taking

$$\mathsf{m}_2 = |P|\Big((1)_{|\mathbf{i}| \leq s}\Big),$$

(namely the sum of all the coefficients of $|P|$), we obtain the desired result. Indeed note that the coefficients of P do not depend on u but only on the smoothness s and dimension d. □

Lemma C.10. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a closed bounded domain with Lipschitz boundary and let $x_o \in \mathcal{X}$. Let $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$, for some $s \in \mathbb{N}$. Then for any $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ and corresponding pair of Sinkhorn potentials $(u, v) \in \mathcal{C}(\mathcal{X})^2$ with $u(x_o) = 0$, the functions $\mathsf{k}(x, \cdot)e^{u/\varepsilon}$ and $\mathsf{k}(x, \cdot)e^{v/\varepsilon}$ belong to $W^{s,2}(\mathcal{X})$ for every $x \in \mathcal{X}$. Moreover, they admit an extension to $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$ and there exists a constant $\bar{\mathsf{r}}$ independent on α and β, such that for every $x \in \mathcal{X}$*

$$\left\|\mathsf{k}(x, \cdot)e^{u/\varepsilon}\right\|_{\mathcal{H}}, \ \left\|\mathsf{k}(x, \cdot)e^{v/\varepsilon}\right\|_{\mathcal{H}} \leq \bar{\mathsf{r}} \tag{C.4.1}$$

(with some abuse of notation, we have identified $\mathsf{k}(x, \cdot)e^{u/\varepsilon}$ and $\mathsf{k}(x, \cdot)e^{v/\varepsilon}$ with their extensions to $\mathbb{R}^d$).

Proof. In the following we denote by $\|\cdot\|_{s,2} = \|\cdot\|_{s,2,\mathcal{X}}$ the norm of $W^{s,2}(\mathcal{X})$ and by $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_{s,2,\mathbb{R}^d}$ the norm of $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$. Let $x \in \mathcal{X}$. Then, since $u - \mathsf{c}(x, \cdot) \in W^{s,\infty}(\mathcal{X})$ and $\|u\|_{s,\infty} \leq \mathsf{r}$, it follows from Lemma C.9 that

$$\begin{aligned}
\left\|\mathsf{k}(x, \cdot)e^{u/\varepsilon}\right\|_{s,\infty} &= \left\|e^{(u-\mathsf{c}(x,\dot{)})/\varepsilon}\right\|_{s,\infty} \\
&\leq \left\|e^{(u-\mathsf{c}(x,\dot{)})/\varepsilon}\right\|_{\infty}(1 + \mathsf{m}_2\|u - \mathsf{c}(x, \cdot)\|_{s,\infty}) \\
&= \left\|\mathsf{k}(x, \cdot)e^{u/\varepsilon}\right\|_{\infty}(1 + \mathsf{m}_2\|u - \mathsf{c}(x, \cdot)\|_{s,\infty}) \\
&\leq \left\|e^{u/\varepsilon}\right\|_{\infty}(1 + \mathsf{m}_2(\mathsf{r} + \|\mathsf{c}\|_{s,\infty})) \\
&\leq e^{\mathsf{D}/\varepsilon}(1 + \mathsf{m}_2(\mathsf{r} + \|\mathsf{c}\|_{s,\infty})),
\end{aligned}$$

where we used the fact that $D^{\mathsf{i}}[\mathsf{c}(x, \cdot)] = (D^{\mathsf{i}}\mathsf{c})(x, \cdot)$. This implies

$$\left\|\mathsf{k}(x, \cdot)e^{u/\varepsilon}\right\|_{s,2} \leq |\mathcal{X}|^{1/2}e^{\mathsf{D}/\varepsilon}(1 + \mathsf{m}_2(\mathsf{r} + \|\mathsf{c}\|_{s,\infty}))$$

where $|\mathcal{X}|$ is the Lebesgue measure of $\mathcal{X}$. Proceeding analogously to (Genevay et al., 2018a, Proposition 2) and using Stein's Extension Theorem (Adams and Fournier, 2003, Theorem 5.24),(Stein, 2016, Chapter 6), we can guarantee the existence of a *total extension operator* (Adams and Fournier, 2003, Definition 5.17). In particular, there exists a constant $\mathsf{m}_3 = \mathsf{m}_3(s, 2, \mathcal{X})$ such that for any $\varphi \in W^{s,2}(\mathcal{X})$ there exists $\tilde{\varphi} \in W^{s,2}(\mathbb{R}^d)$ such that

$$\|\tilde{\varphi}\|_{\mathcal{H}} = \|\tilde{\varphi}\|_{s,2,\mathbb{R}^d} \leq \mathsf{m}_3\|\varphi\|_{s,2,\mathcal{X}} = \mathsf{m}_3\|\varphi\|_{s,2}. \tag{C.4.2}$$

Therefore, we conclude

$$\left\|\mathsf{k}(x,\cdot)e^{u/\varepsilon}\right\|_{\mathcal{H}} \leq \mathsf{m}_3|\mathcal{X}|^{1/2}e^{\mathsf{D}/\varepsilon}(1+\mathsf{m}_2(\mathsf{r}+\|\mathsf{c}\|_{s,\infty})) =: \bar{\mathsf{r}}. \tag{C.4.3}$$

The same argument applies to $\mathsf{k}(x,\cdot)e^{v/\varepsilon}$ with the only exception that now, in virtue of Cor. A.12, we have $\|e^{v/\varepsilon}\|_\infty \leq e^{2\mathsf{D}/\varepsilon}$. Note that $\bar{\mathsf{r}}$ is a constant depending only on $\mathcal{X}$, c, s and d but it is independent on the probability distributions α and β. □

Sobolev spaces and reproducing kernel Hilbert spaces. Recall that for $s > d/2$ the space $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$, is a reproducing kernel Hilbert space (RKHS) (Wendland, 2004, Chapter 10). In this setting we denote by $\mathsf{h} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ the associated reproducing kernel, which is continuous and bounded and satisfies the reproducing property

$$\forall\, x \in \mathcal{X},\ \forall\, f \in \mathcal{H} \qquad \langle f, \mathsf{h}(x,\cdot)\rangle_{\mathcal{H}} = f(x). \tag{C.4.4}$$

We can also assume that h is *normalized*, namely, $\|\mathsf{h}(x,\cdot)\|_{\mathcal{H}} = 1$ for all $x \in \mathcal{X}$ (Wendland, 2004, Chapter 10).

Kernel mean embeddings. For every $\beta \in \mathcal{P}(\mathcal{X})$, we denote by $\mathsf{h}_\beta \in \mathcal{H}$ the *Kernel Mean Embedding* of β in $\mathcal{H}$ (Smola et al., 2007; Muandet et al., 2017), that is, the vector

$$\mathsf{h}_\beta = \int \mathsf{h}(x,\cdot)\, d\beta(x). \tag{C.4.5}$$

In other words, the kernel mean embedding of a distribution β corresponds to the expectation of $\mathsf{h}(x,\cdot)$ with respect to β. By the linearity of the inner product and the integral, for every $f \in \mathcal{H}$, the inner product

$$\langle f, \mathsf{h}_\beta\rangle_{\mathcal{H}} = \int \langle f, \mathsf{h}(x,\cdot)\rangle\ d\beta(x) = \int f(x)\, d\beta(x), \tag{C.4.6}$$

corresponds to the expectation of $f(x)$ with respect to β. The *Maximum Mean Discrepancy (MMD)* (Song, 2008; Sriperumbudur et al., 2011; Muandet et al., 2017) between two probability distributions $\beta, \beta' \in \mathcal{P}(\mathcal{X})$, recalled inDef. A.14, corresponds to

$$\mathsf{MMD}(\beta,\beta') = \|\mathsf{h}_\beta - \mathsf{h}_{\beta'}\|_{\mathcal{H}}. \tag{C.4.7}$$

In the case of the Sobolev space $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$, the MMD metrizes the weak-$*$ topology of $\mathcal{P}(\mathcal{X})$ (Sriperumbudur et al., 2010, 2011). Finally, we show the sample complexity of MMD:

Lemma 4.7. *Let $\beta \in \mathcal{P}(\mathcal{X})$. Let $x_1, \dots, x_n \in \mathcal{X}$ be independently sampled according to β and denote by $\beta_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Then, for any $\tau \in (0, 1]$, we have*

$$\mathsf{MMD}(\beta_n, \beta) \leq \frac{4 \log \frac{3}{\tau}}{\sqrt{n}} \tag{4.4.2}$$

with probability at least $1 - \tau$.

Proof. The proof follows by applying Pinelis' inequality (Yurinskii, 1976; Pinelis, 1994; Smale and Zhou, 2007) for random vectors in Hilbert spaces. More precisely, for $i = 1, \dots, n$, denote by $\zeta_i = \mathsf{h}(x_i, \cdot) \in \mathcal{H}$ and recall that $\|\zeta_i\| = \|\mathsf{h}(x, \cdot)\| = 1$ for all $x \in \mathcal{X}$. We can therefore apply (Smale and Zhou, 2007, Lemma 2) with constants $\widetilde{M} = 1$ and $\sigma^2 = \sup_i \mathbb{E}\|\zeta_i\|^2 \leq 1$, which guarantees that for every $\tau \in (0, 1]$

$$\|\frac{1}{n}\sum_{i=1}^n \Big[\zeta_i - \mathbb{E}\,\zeta_i\Big]\|_{\mathcal{H}} \leq \frac{2 \log \frac{2}{\tau}}{n} + \sqrt{\frac{2 \log \frac{2}{\tau}}{n}} \leq \frac{4 \log \frac{3}{\tau}}{\sqrt{n}}, \tag{C.4.8}$$

holds with probability at least $1 - \tau$. Here, for the second inequality we have used the fact that $\log \frac{2}{\tau} \leq \log \frac{3}{\tau}$ and $\log \frac{3}{\tau} \geq 1$ for every $\tau \in (0, 1]$. The desired result follows by observing that

$$\mathsf{h}_\beta = \int \mathsf{h}(x, \cdot)\, d\beta(x) = \mathbb{E}\,\zeta_i \tag{C.4.9}$$

for all $i = 1, \dots, n$, and

$$\mathsf{h}_\beta = \frac{1}{m}\sum_{i=1}^m \mathsf{h}(x_i, \cdot) = \frac{1}{m}\sum_{i=1}^m \zeta_i. \tag{C.4.10}$$

Therefore,

$$\mathsf{MMD}(\beta_k, \beta) = \|\mathsf{h}_{\beta_k} - \mathsf{h}_\beta\|_{\mathcal{H}} = \|\frac{1}{n}\sum_{i=1}^n \Big[\zeta_i - \mathbb{E}\,\zeta_i\Big]\|_{\mathcal{H}}, \tag{C.4.11}$$

which combined with (C.4.8) leads to the conclusion. □

C.5 Additional experiments

Sampling of continuous measures: mixture of Gaussians. We compute the barycen-

ter of 5 mixtures of two Gaussians μ_j, centered at $(j/2, 1/2)$ and $(j/2, 3/2)$ for $j - 0, \ldots, 4$ respectively. Samples are provided in Fig. C.1. We use different relative weights pairs in the mixture of Gaussians, namely $(1/10, 9/10), (1/4, 3/4), (1/2, 1/2)$. At each iteration, a sample of $n = 500$ points is drawn from μ_j, $j = 0 \ldots, 4$. Results are reported in Fig. C.2.

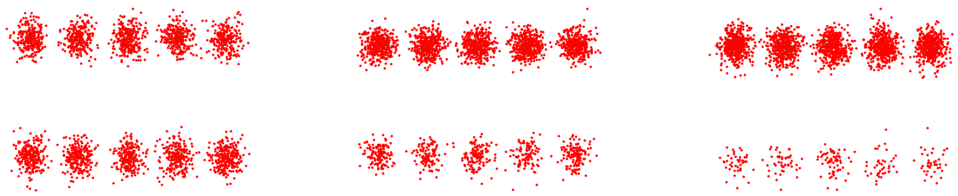


Figure C.1: Samples of input measures

Figure C.2: Barycenters of Mixture of Gaussians

Large scale discrete measures: meshes. We compute the barycenter of two discrete measures with support in $\mathbb{R}^3$. Meshes of the dinosaur are taken from (Solomon et al., 2015) and rescaled by a 0.5 factor. The internal problem in Frank-Wolfe algorithm is solved using L-BFGS-B SciPy optimizer. Formula of the Jacobian is passed to the method. The barycenter is displayed in Fig. C.3 together with an example of the input.

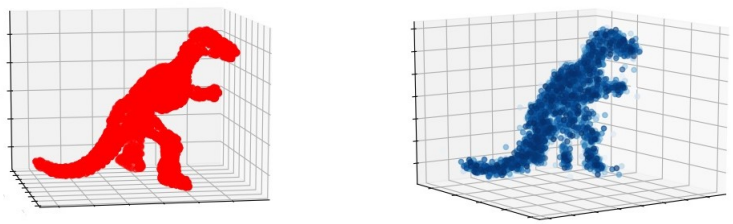


Figure C.3: 3D dinosaur mesh (left), barycenter of 3D meshes (right)

Appendix D

Appendix of Chapter 5

This chapter is organized as follows:

Appendix D.1. This section recalls how common loss functions used for GAN training can be formulated as adversarial divergences.

Appendix D.2. This section provides details on the examples made in Sec. 5.2.

Appendix D.3. In this section we derive technical results that will be used to prove the main results of this work.

Appendix D.4. This section proves the learning rates of the joint GAN estimator proposed in Sec. 5.3.

Appendix D.5. This section proves the formula of the gradient of Sinkhorn divergence with respect to network parameters.

Appendix D.6. This section describes the experimental setup and network specification for the empirical evaluation reported in Sec. 5.5.

D.1 Adversarial Divergences

The notion of *adversarial divergence* was originally introduced in Liu et al. (2017). For completeness, we review how most loss functions used for probability matching within the GANs literature can be formulated as an adversarial divergence in (5.1.2). We recall here the definition in (5.1.2) of adversarial divergence over a space $\mathcal{F}$ of functions $F : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

bewteen two distributions $\mu, \rho \in \mathcal{P}(\mathcal{X})$ as

$$\mathsf{d}_{\mathcal{F}}(\mu, \rho) \;=\; \sup_{F \in \mathcal{F}} \int F(x', x)\, d\mu(x') d\rho(x). \tag{D.1.1}$$

Depending on the choice of $\mathcal{F}$ we recover different choices of adversarial divergences as discussed below.

Integral Probability Metrics. One of the most notable examples of adversarial divergences are integral probability metrics (IPM), briefly discussed in Appendix A.2.2. Recall that the IPM over a space $\mathcal{F}_0$ of functions $f : \mathcal{X} \to \mathbb{R}$, between two distribution μ, ρ is defined as

$$\mathsf{IPM}_{\mathcal{F}_0}(\mu, \rho) \;=\; \sup_{f \in \mathcal{F}_0} \Big| \int f(x)\, d\mu(x) - \int f(x) d\rho(x) \Big|. \tag{D.1.2}$$

The IPM is an adversarial divergence with $\mathcal{F}$ in (D.1.1) defined as

$$\mathcal{F} \;=\; \{\, F \mid F : (x', x) \mapsto f(x') - f(x), \quad f \in \mathcal{F}_0 \,\}.$$

Examples include:

- *Maximum Mean Discrepancy (*MMD*).* Here $\mathcal{F}_0$ is the ball of radius 1 in a Reproducing Kernel Hilbert Space (Dziugaite et al., 2015).
- *μ-Sobolev IPM.* Proposed in (Mroueh et al., 2018). Given a reference measure $\mu \in \mathcal{P}(\mathcal{X})$, Sobolev-IPM consider $\mathcal{F}_0$ to be

  $$\mathcal{F}_0 \;=\; \{\, f \mid \mathbb{E}_\mu \|\nabla f(\cdot)\|^2 \leq 1, \quad f \in W^{1,2}(\mathcal{X}, \mu) \,\}$$

  with $W^{1,2}(\mathcal{X}, \mu)$ denoting the space of μ-square integrable functions on $\mathcal{X}$ with μ-square integrable first weak derivative. Similarly, *μ-Fisher-IPM* (Mroueh and Sercu, 2017) are IPM defined over $\mathcal{F}_0$ the ball of radius 1 in $L^2(\mathcal{X}, \mu)$.
- 1-*Wasserstein.* The 1-Wasserstein distance is defined as in (2.2.3) with $p = 1$ and with dual formulation (2.2.9) in Chapter 2. The dual formulation is in an IPM loss, with $\mathcal{F}_0$ the ball of radius 1 in the space of Lipschitz functions (namely all functions with Lipschitz constant less or equal than 1).

f-Divergences. f-Divergences, briefly introduced in Appendix A.2.1, are discrepancy measures between two distributions of the form

$$\mathsf{d}_f(\mu,\rho) \;=\; \int f\left(\frac{d\mu}{d\rho}(x)\right) d\rho(x) \tag{D.1.3}$$

where $d\mu/d\rho$ denotes the Radon-Nykodin derivative of μ with respect to ρ and $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is suitable a convex function. By leveraging the notion of Fenchel dual $f^*(y) = \sup_x \langle x, y\rangle - f(x)$ and the fact that $f^{**} = f$ for convex functios, in (Nowozin et al., 2016; Liu et al., 2017) it was observed that f-divergences of the form in (D.1.3) can be written as adversarial divergences with

$$\mathcal{F}_f \;=\; \{\, F \mid F : (x', x) \mapsto g(x') - f^*(g(x)), \quad g \in C_b(\mathcal{X}),\ \mathrm{dom}(g) \subset \mathrm{dom}(f) \,\}$$

with $C_b(\mathcal{X})$ the set of continuous bounded functions on $\mathcal{X}$.

Entropic Optimal Transport. We conclude this section by reviewing how entropic optimal transport functions can be formulated as adversarial divergences. As observed in Sec. 5.4, the dual probem associated to the definition of OT_ε corresponds to

$$\sup_{u,v \in \mathcal{C}(\mathcal{X})} \int u(x)\, d\mu(x) \;+\; \int v(y)\, d\rho(y) \;-\; \varepsilon \int e^{\frac{u(x)+v(y)-\|x-y\|^2}{\varepsilon}}\, d\mu(x) d\rho(y).$$

This problem can be written in the form of adversarial divergence in (5.1.2) by taking

$$\mathcal{F} \;=\; \{\, F \mid F : (x, y) \mapsto u(x) + v(y) - \varepsilon e^{\frac{u(x)+v(y)-\|x-y\|^2}{\varepsilon}}, \quad u, v \in C(\mathcal{X}) \,\}.$$

D.2 The Complexity of Pushforward Maps/Generators

We provide here some examples and remarks on pushforward maps. Assume that $\mathcal{Z} = \mathcal{X} = \mathbb{R}^d$ and consider μ to be a Gaussian measure. Consider $\rho \in \mathcal{P}(\mathcal{X})$ a target distribution. Since μ is absolutely continuous with respect to Lebesgue measure $\mathcal{L}^d$, there always exists a measurable map T such that $T_\# \mu = \rho$ (see for example (Ambrosio and Gigli, 2013, Thm. 1.33). However, existence of a pushforward map T does not imply anything on its regularity, unless further assumptions hold on the measures μ and ρ. For example, consider the case where the support of ρ is disconnected: in this case, any pushforward will exhibit

discontinuities (since the image through a continuous map of a connected set is always connected). In the following we provide a few examples of pushforward map between distributions.

1. Book-shifting. Let $\mu = \chi_{[0,1]}$ and $\rho = \chi_{[1,2]}$. The maps $T_1, T_2 : [0,1] \to [1,2]$ defined by $T_1(x) = 2 - x$, $T_2(x) = x + 1$ are pushforwards from μ to ρ, i.e. $T_{i\#}\mu = \rho$ for $i = 1, 2$.
2. Gaussians. Let $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$ and $\rho = \mathcal{N}(m_\rho, \Sigma_\rho)$ be two Gaussians on $\mathbb{R}^d$. The following map

$$T : x \to m_\rho + A(x - m_\mu), \qquad A = \Sigma_\mu^{-\frac{1}{2}} \big(\Sigma_\mu^{\frac{1}{2}} \Sigma_\rho \Sigma_\mu^{\frac{1}{2}}\big)^{\frac{1}{2}} \Sigma_\mu^{-\frac{1}{2}} \tag{D.2.1}$$

is such that $T_\# \mu = \rho$.

Intuitively, when considering measures with the same 'structure' (i.e. in the examples above, both uniform or both Gaussian), the 'distortion' needed to match the two distribution is mild and this results in very regular pushforward maps, namely linear in the case above. Viceversa, given a measure μ, pushforward via linear maps can target measures with the same structure as μ only.

Example: class of functions and correspondent pushforward measures. Consider $\mathcal{T}$ the class of affine maps on the real line, i.e.

$$\mathcal{T} := \{T^{m,q} : \mathbb{R} \to \mathbb{R} : \ T^{m,q}(x) = mx + q, \ \ m, q \in \mathbb{R}, \ \ m \neq 0.\} \tag{D.2.2}$$

Let $\mu = r\mathcal{L}^1$ with $r = \mathbb{1}_{[0,1]}$. Then, all the pushforward measures $T_\# \mu$ with $T \in \mathcal{T}$ are of the form

$$T^{m,q}_\# \mu = \frac{\mathbb{1}_{[0,1]}\big(\frac{x-q}{m}\big)}{m} \mathcal{L}^1. \tag{D.2.3}$$

Hence, all the measures that can be written as pushforward of μ via maps in $\mathcal{T}$ are uniform measures on intervals in $\mathbb{R}$, namely $T^{m,q}_\# \mu = \frac{1}{m} \mathbb{1}_{[q,q+m]}$.

When μ and the target measure ρ are very different, pushforward maps will be more complex:

3. From uniform to troncated Gaussian. Let $\rho = f\mathbb{1}_{[-1,1]}$ and $\mu = g\mathbb{1}_{[-1.1]}$ with $f(x) = \sqrt{2/(\pi c)} e^{-\frac{x^2}{2}}$, $c = \text{erf}(1/\sqrt{2})$ and $g(x) = \frac{1}{2}$. Using (5.2.2), one can show that the map T such that $T_\# \mu = \rho$ is of the form

$$T(x) = \sqrt{2}\text{erf}^{-1}\big(cx/2\big). \tag{D.2.4}$$

Since ρ is not of the form (D.2.3), there exists no $T \in \mathcal{T}$ with $\mathcal{T}$ defined in (D.2.2) such that $T_{\#}\mu = \rho$. In order to be able to map μ into ρ, one has to consider a class of function $\mathcal{T}$ large enough to include the function erf^{-1}.

From the examples above, it is clear that when given a fixed μ and a target ρ, the regularity (in terms of upper bounds of derivatives) varies significantly, depending on the properties of μ and ρ. In particular, a measure μ with a specific structure, may require a pushforward T to have big derivatives, in order to satisfy $T_{\#}\mu = \rho$.

Example: pushforward from one Gaussian to a mixture of three Gaussians in 1d. We computed the pushforward from a Gaussian distribution to a mixture of the Gaussian distributions in 1D with variance 0.05 and 0.01 (see Fig. D.1a and Fig. D.1c). We computed the pushforward map using a neural network with 5 layers, alternating ReLu and tanh as activation functions. Fig. D.1b and Fig. D.1d display the graphs of the computed pushforward maps. One can notice that the maps alternate regions with steep derivatives to regions with flat derivatives, needed to distort the mass of the Gaussian in order to match the multimodal shape of the target. The steepness is significantly higher in the case where the target has smaller variance (i.e. the three Gaussians are more concentrated leading to areas with a very small amount of mass).

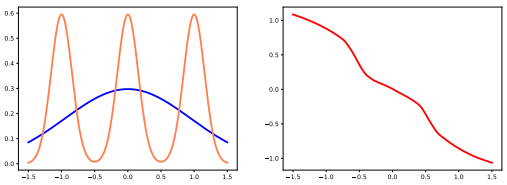
Experimental setup for Fig. 5.1. We used neural network with 1 layer (linear network), 2 layers (with ReLu activation), 5 (up to 256 dimensions) and 7 layers (up to 512 dimensions), alternating ReLu to Tanh activation functions.

D.3 Technical results

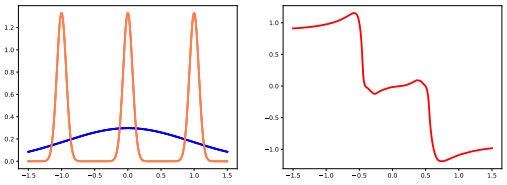
We introduce some notation first; in the following, the map $\mathsf{c} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denotes the Euclidean squared norm, i.e. $\mathsf{c}(x, y) = \frac{1}{2}\|x - y\|^2$. Given two maps $V, T : \mathcal{Z} \to \mathcal{X}$, we used the symbol c decorated with subscripts T and V to denote the following: $\mathsf{c}_{T,V} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is the function defined by

$$\mathsf{c}_{T,V}(z, w) = \mathsf{c}(T(z), V(w)), \qquad \text{for all } z, w \in \mathcal{Z}.$$

Since we need to highlight the dependence on the cost, we will incorporate it in the notation used for Sinkhorn divergence, namely $\mathsf{S}_{\varepsilon,c}$ with c denoting the cost function used. We first recall a straightforward result which links Sinkhorn divergence of pushforward measures



(a) Distributions **(b)** Pushforward map



(c) Distributions **(d)** Pushforward map

Figure D.1: Pushforward functions that map 1 Gaussian to a mixture of three Gaussians. Distributions displayed on the left. Graphs of pushforward maps on the right.

with Sinkhorn divergence with modified cost function.

Lemma D.1. *Let $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ and $T, V : \mathcal{Z} \to \mathcal{X}$ be continuous maps. Let $\mathsf{c} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\mathsf{c}_{T,V} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be as defined above. Then,*

$$\mathsf{S}_{\varepsilon,\mathsf{c}}(T_{\#}\mu, V_{\#}\nu) = \mathsf{S}_{\varepsilon,\mathsf{c}_{T,V}}(\mu, \nu).$$

Proof. Similarly to $\mathsf{S}_{\varepsilon,\mathsf{c}}$, let $\mathsf{OT}_{\varepsilon,\mathsf{c}}$ be the biased entropic OT problem with cost function c. Let $F(\mu, \nu, u, v, \mathsf{c})$ be defined as

$$F(\alpha, \beta, u, v, \mathsf{c}) = \int_{\mathcal{X}} u(x)\, d\alpha(x) + \int_{\mathcal{X}} v(y)\, d\beta(y) - \varepsilon \int e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\alpha(x) d\beta(y).$$

By the dual definition of $\mathsf{OT}_{\varepsilon,\mathsf{c}}$, we have

$$\begin{aligned}
\mathsf{OT}_{\varepsilon,\mathsf{c}}(T_{\#}\mu, V_{\#}\nu) &= \sup_{(u,v)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{X})} F(T_{\#}\mu, V_{\#}\nu, u, v, \mathsf{c}) && \text{(D.3.1)}\\
&= \sup_{(u,v)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{X})} F(\mu, \nu, u \circ T, v \circ V, \mathsf{c}_{T,V}) && \text{(D.3.2)}\\
&= \sup_{(\tilde{u},\tilde{v})\in(\mathcal{C}(\mathcal{X})\circ T)\times(\mathcal{C}(\mathcal{X})\circ V)} F(\mu, \nu, \tilde{u}, \tilde{v}, \mathsf{c}_{T,V}), && \text{(D.3.3)}
\end{aligned}$$

by the property of the pushforward and where $\mathcal{C}(\mathcal{X}) \circ T := \{f \circ T : \ f \in \mathcal{C}(\mathcal{X})\}$ and

similarly for $\mathcal{C}(\mathcal{X}) \circ V$. Now, consider

$$\mathsf{OT}_{\varepsilon,\mathsf{c}_{T,V}}(\mu,\nu) = \sup_{(\tilde{u},\tilde{v})\in\mathcal{C}(\mathcal{Z})\times\mathcal{C}(\mathcal{Z})} \int_{\mathcal{Z}} \tilde{u}(z)\, d\mu(z) + \int_{\mathcal{Z}} \tilde{v}(w)\, d\nu(w) - \varepsilon \int e^{\frac{\tilde{u}(z)+\tilde{v}(w)-\mathsf{c}_{T,V}(z,w)}{\varepsilon}}\, d\mu(z)d\nu(w).$$

We note that the optimal potentials $\tilde{u}, \tilde{v}$ of $\mathsf{OT}_{\varepsilon,\mathsf{c}_{T,V}}$ satisfy

$$\tilde{u}(z) = -\log \int_{\mathcal{Z}} e^{\tilde{v}(w)-\mathsf{c}_{T,V}(z,w)}\, d\nu(w).$$

Recalling that $\mathsf{c}_{T,V}(z,w) = \mathsf{c}(T(z),V(w))$, we note that $\tilde{u}$ and $\tilde{v}$ are functions of the form $u \circ T$ and $v \circ V$. Hence the supremum in (D.3.4) can be restricted to be on the sets $\mathcal{C}(\mathcal{X}) \circ T$ and $\mathcal{C}(\mathcal{X}) \circ V$. Thus, the quantity in (D.3.3) equals $\mathsf{OT}_{\varepsilon,\mathsf{c}_{T,V}}$. Extending the same argument to the autocorrelation terms, we conclude that $\mathsf{S}_{\varepsilon,\mathsf{c}}(T_{\#}\mu, V_{\#}\nu) = \mathsf{S}_{\varepsilon,\mathsf{c}_{T,V}}(\mu,\nu)$ as desired. □

Before proceeding with the results bounding the potentials of Sinkhorn divergence with cost function $\mathsf{c}_{T,V}$, we provide some technical results the will be needed in the rest.

Lemma D.2 (Lemma 1 in (Mena and Niles-Weed, 2019))**.** *If $\mu \in \mathcal{P}(\mathcal{Z})$ with $\mathcal{Z} \subset \mathbb{R}^k$ is σ^2-subgaussian, then*

$$\mathbb{E}_\mu \|X\|^{2r} \leq (2k\sigma^2)^r r!,$$

for all nonnegative integers r. Also,

$$\mathbb{E}_\mu e^{v\cdot X} \leq \mathbb{E}_\mu e^{\|v\|\|X\|} \leq 2e^{\frac{k\sigma^2}{2}\|v\|^2}$$

for any $v \in \mathbb{R}^k$.

Lemma D.3. *Let $P = T_{\#}\mu$ with $T : \mathbb{R}^k \to \mathbb{R}^d$ such that $\|T(z)\| \leq L\|z\|$ and μ σ^2-sub-Gaussian. Then*

$$\mathbb{E}_P\|X\|^2 \leq 2k(L\sigma)^2 \qquad \text{and} \qquad \mathbb{E}_P\|X\| \leq L\sigma\sqrt{2k}.$$

Proof.

$$\mathbb{E}_P\|X\|^2 = \int_{\mathcal{X}} \|x\|^2\, dP(x) = \int_{\mathcal{X}} \|x\|^2\, d(T_{\#}\mu)(x) \tag{D.3.4}$$

$$= \int_{\mathcal{Z}} \|T(z)\|^2\, d\mu(z) \leq L^2 \int_{\mathcal{Z}} \|z\|^2\, d\mu(z). \tag{D.3.5}$$

Since μ is σ^2-sub-Gaussian, we have that

$$\mathbb{E}_\mu \frac{\|z\|^{2r}}{(2k\sigma^2)^r r!} \leq \mathbb{E}_\mu e^{\frac{\|z\|^2}{2k\sigma^2}} - 1 \leq 1.$$

Thus, $L^2\mathbb{E}_\mu\|z\|^2 \leq 2L^2k\sigma^2$ and combining this with (D.3.5) we obtain

$$\mathbb{E}_P\|X\|^2 \leq L^2 2k\sigma^2.$$

An easy application of Jensen inequality yields the bound for $\mathbb{E}_P\|X\|$. □

The lemma below contains bound of Sinkhorn potentials for a cost function of the form $\mathsf{c}_{T,V}$. The proof is quite technical, mostly based on long but basic computations, and it follows the same steps and ideas as the proofs in (Mena and Niles-Weed, 2019).

Lemma D.4 (Bounds on potentials with changed cost function). *Let $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ be σ^2-sub-Gaussian measures and consider Sinkhorn divergence with $\varepsilon = 1$ and $\mathsf{c}(x,y) = \frac{1}{2}\|T(x) - V(y)\|$ as cost function, where $T, V : \mathcal{Z} \to \mathcal{X}$ are such that $\|K(z)\| \leq L\|z\|$ for $K = T, V$. Let (u, v) denote a pair of optimal potentials. Then,*

$$-1 - k(L\sigma)^2\big(1 + \frac{1}{2}(\|T(x)\| + \sqrt{2k}L\sigma)\big)^2 \leq u(x) \leq \frac{1}{2}\big(\|T(x)\| + \sqrt{2k}L\sigma\big)^2$$
$$-1 - k(L\sigma)^2\big(1 + \frac{1}{2}(\|V(y)\| + \sqrt{2k}L\sigma)\big)^2 \leq v(y) \leq \frac{1}{2}\big(\|V(y)\| + \sqrt{2k}L\sigma\big)^2.$$

Proof. Let (u_0, v_0) any pair of optimal potentials. Since potentials are defined up to constant, we assume as in (Mena and Niles-Weed, 2019) that $\mathbb{E}_\mu u_0 = \mathbb{E}_\nu v_0 = \frac{1}{2}\mathsf{S}(\mu, \nu)$. We define

$$u(x) = -\log \int_{\mathcal{X}} e^{v_0(y) - \frac{1}{2}\|T(x) - V(y)\|^2}\, d\nu(y)$$
$$v(y) = -\log \int_{\mathcal{X}} e^{u(x) - \frac{1}{2}\|T(x) - V(y)\|^2}\, d\mu(x),$$

for any $x, y \in \mathcal{X}$. Once we have proved that they are well defined and we have shown the desired lower and upper bound, the proof that they are optimal potentials is exactly the same as in (Mena and Niles-Weed, 2019, Prop 6). By Jensen inequality

$$v_0(y) = -\log \int_{\mathcal{X}} e^{u_0(x) - \frac{1}{2}\|T(x) - V(y)\|^2}\, d\mu(x)$$
$$\leq -\mathbb{E}_\mu u_0(x) + \frac{1}{2}\mathbb{E}_\mu\|T(X) - V(y)\|^2 \leq \frac{1}{2}\mathbb{E}_\mu\|T(X) - V(y)\|^2.$$

Therefore

$$e^{v_0(y)-\frac{1}{2}\|T(x)-V(y)\|^2} \leq e^{\frac{1}{2}\mathbb{E}_\mu\|T(X)-V(y)\|^2-\frac{1}{2}\|T(x)-V(y)\|}.$$

Expanding the squares we have

$$\begin{aligned}&\frac{1}{2}\mathbb{E}_\mu\|T(X)-V(y)\|^2-\frac{1}{2}\|T(x)-V(y)\|^2\\&=\frac{1}{2}\mathbb{E}_\mu\|T(X)\|^2+\frac{1}{2}\|V(y)\|^2-\mathbb{E}_\mu\langle T(X),V(y)\rangle-\frac{1}{2}\|T(x)\|^2-\frac{1}{2}\|V(y)\|^2-\langle T(x),V(y)\rangle.\end{aligned}$$

Using Lemma D.3, we obtain

$$\begin{aligned}\frac{1}{2}\mathbb{E}_\mu\|T(X)-V(y)\|^2-\frac{1}{2}\|T(x)-V(y)\|^2&\leq L^2k\sigma^2+\mathbb{E}_\mu\|T(X)\|\|V(y)\|+\|T(x)\|\|V(y)\|\\&\leq k(L\sigma)^2+\|V(y)\|(\|T(x)\|+\sqrt{2k}L\sigma).\end{aligned}$$

With elementary computations and using σ^2-sub-Gaussianity of ν, we get

$$\int_{\mathcal{Z}} e^{k(L\sigma)^2+\|V(y)\|(\|T(x)\|+\sqrt{2k}L\sigma)}\,d\nu(y)\leq 2e^{k(L\sigma)^2\left(1+\frac{1}{2}(\|T(x)\|+\sqrt{2k}L\sigma)\right)^2}.$$

So combining the last steps, we have shown that

$$\int e^{v_0(y)-\frac{1}{2}\|T(x)-V(y)\|^2}\,d\nu(y)\leq 2e^{k(L\sigma)^2\left(1+\frac{1}{2}(\|T(x)\|+\sqrt{2k}L\sigma)\right)^2}.$$

Now,

$$\begin{aligned}u(x)&=-\log\int e^{v_0(y)-\frac{1}{2}\|T(x)-V(y)\|^2}\,d\nu(y)\\&\geq-\log(2e^{k(L\sigma)^2\left(1+\frac{1}{2}(\|T(x)\|+\sqrt{2k}L\sigma)\right)^2})\\&\geq-1-k(L\sigma)^2\big(1+\frac{1}{2}(\|T(x)\|+\sqrt{2k}L\sigma)\big)^2,\end{aligned}$$

proving the desired lower bound. We now study the upper bound for u:

$$\begin{aligned} u(x) &= -\log \int e^{v_0(y) - \frac{1}{2}\|T(x)-V(y)\|^2} \, d\nu(y) \\ &\leq \int -\log e^{v_0(y) - \frac{1}{2}\|T(x)-V(y)\|^2} \, d\nu(y) \\ &= -\int v_0(y) \, d\nu(y) + \int \frac{1}{2}\|T(x) - V(y)\|^2 \, d\nu(y) \\ &\leq \int \frac{1}{2}\|T(x) - V(y)\|^2 \, d\nu(y). \end{aligned}$$

Developing the square and bounding $\mathbb{E}_\nu\|V(Y)\|^2$ and $\mathbb{E}_\nu\|V(Y)\|$ with Lemma D.3, we have

$$u(x) \leq \frac{1}{2}\big(\|T(x)\| + \sqrt{2k}L\sigma\big)^2.$$

With the exact same reasoning one can derive the analogous bound for v. □

Note that in terms of $\|x\|$ (and not $\|T(x)\|$) the derived bounds become

$$-1 - k(L\sigma)^2\big(1 + \frac{1}{2}L^2(\|x\| + \sqrt{2k}\sigma)\big)^2 \leq u(x) \leq \frac{1}{2}L^2\big(\|x\| + \sqrt{2k}\sigma\big)^2 \tag{D.3.6}$$

$$-1 - k(L\sigma)^2\big(1 + \frac{1}{2}L^2(\|y\| + \sqrt{2k}\sigma)\big)^2 \leq v(y) \leq \frac{1}{2}L^2\big(\|y\| + \sqrt{2k}\sigma\big)^2. \tag{D.3.7}$$

Therefore we have the following result:

Lemma D.5. *In the assumptions of Lemma D.4 we have*

$$|u(z)| \leq C_k \begin{cases} 1 + (L\sigma)^4 & \text{if } \|z\| \leq \sqrt{k}\sigma \\ 1 + (1 + (L\sigma)^2)L^2\|z\|^2 & \text{if } \|z\| > \sqrt{k}\sigma. \end{cases} \tag{D.3.8}$$

where C_k is a constant depending only on k.

Proof. This is an immediate consequence of (D.3.6) and (D.3.7). □

While the previous results provided bounds on the potentials, the following will focus on their *derivatives*. As before, the proof is technical (with even more computations) but mainly follows its analogous in (Mena and Niles-Weed, 2019) (which is done for the squared norm as a cost and does not highlights the dependence on T and V).

Lemma D.6 (Bounds on derivatives of potentials with changed cost function). *Let $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ be σ^2-sub-Gaussian measures and consider Sinkhorn divergence with $\varepsilon = 1$*

and $\mathsf{c}(x,y) = \frac{1}{2}\|T(x) - V(y)\|$ *as cost function, where* $T, V : \mathcal{Z} \to \mathcal{X}$ *are such that* $\|K(z)\| \leq L\|z\|$ *for* $K = T, V$*. Also, assume that for any multi-index* α *with length at most* $|\alpha| \leq \lfloor k/2 \rfloor + 1$*,* $\|D^\alpha T(x)\|_\infty \leq \tau$*. Let* (u, v) *denote a pair of optimal potentials. Then,*

$$|D^\alpha\big(u(\cdot) - \frac{1}{2}\|T(\cdot)\|^2\big)(z)| \leq C_{k,|\alpha|} \begin{cases} (\tau L\sigma)^{|\alpha|}(1 + ((L\sigma)^2 + L\sigma)^{|\alpha|}) & \text{if } \|z\| \leq \sqrt{k\sigma} \\ (\tau L\sigma)^{|\alpha|}(1 + (\sqrt{(L\sigma)\|z\|} + (L^2\sigma)\|z\|)^{|\alpha|}) & \text{if } \|z\| > \sqrt{k\sigma}. \end{cases} \tag{D.3.9}$$

Proof. Potentials (u, v) are chosen as in Lemma D.4. For convenience, set $\bar{u}$ the function defined by $\bar{u}(x) = u(x) - \frac{1}{2}\|T(x)\|^2$. The goal is now to bound the derivatives of $\bar{u}$, namely $|D^\alpha \bar{u}(x)|$. Note that

$$\begin{aligned} D^\alpha \bar{u}(x) = -D^\alpha \log(e^{-\bar{u}(x)}) &= -D^\alpha(\log(\int e^{v(y) - \frac{1}{2}\|T(x) - V(y)\|^2 + \frac{1}{2}\|T(x)\|^2}\, d\nu(y)) \\ &= -D^\alpha(\log(\int e^{v(y) - \frac{1}{2}\|V(y)\|^2 + \langle V(y), T(x)\rangle}\, d\nu(y)) \\ &= \frac{D^\alpha \int e^{v(y) - \frac{1}{2}\|V(y)\|^2 + \langle V(y), T(x)\rangle}\, d\nu(y))}{\int e^{v(y) - \frac{1}{2}\|V(y)\|^2 + \langle V(y), T(x)\rangle}\, d\nu(y))}. \end{aligned}$$

Using Faa' di Bruno formula, we have

$$D^\alpha \int e^{v(y) - \frac{1}{2}\|V(y)\|^2 + \langle V(y), T(x)\rangle}\, d\nu(y)) = \int \mathbf{P}([\langle D^j T(x), V(y)\rangle]_{j \leq |\alpha|}) e^{v(y) - \frac{1}{2}\|V(y)\|^2 - \langle T(x), V(y)\rangle}\, d\nu(y),$$

where $\mathbf{P}$ is a polynomial of degree $|\alpha|$. In order to bound $|D^\alpha \bar{u}(x)|$, we have to bound the quantity

$$A(x) := \frac{\int \mathbf{P}([\langle D^j T(x), V(y)\rangle]_{j \leq |\alpha|}) e^{v(y) - \frac{1}{2}\|V(y)\|^2 - \langle T(x), V(y)\rangle}\, d\nu(y)}{\int e^{v(y) - \frac{1}{2}\|V(y)\|^2 + \langle V(y), T(x)\rangle}\, d\nu(y))}.$$

To simplify the notation, set

$$E(v, V)(y) := e^{v(y) - \frac{1}{2}\|V(y)\|^2 - \langle T(x), V(y)\rangle} \quad \text{and} \quad \mathsf{B} := \int e^{v(y) - \frac{1}{2}\|V(y)\|^2 + \langle V(y), T(x)\rangle}\, d\nu(y)./$$

Now, set $\mathcal{D} := \{y : \|y\| \leq \mathsf{h}\}$ with h to be chosen later, and $\mathcal{D}^c$ the complementary set. We

split the quantity $A(x)$ as follows:

$$A(x) = A_1(x) + A_2(x)$$

with

$$A_1(x) = \int \mathbb{1}_{\mathcal{D}} \mathbf{P}([\langle D^j T(x), V(y)\rangle]_{j\le|\alpha|}) E(v,V)(y)\, d\nu(y)/\mathsf{B},$$
$$A_2(x) = \int \mathbb{1}_{\mathcal{D}^c} \mathbf{P}([\langle D^j T(x), V(y)\rangle]_{j\le|\alpha|}) E(v,V)(y)\, d\nu(y)/\mathsf{B}.$$

We bound the two terms separately: note that on $\mathcal{D}$ we have $\|V(y)\| \le L\|y\| \le L\mathsf{h}$ and hence

$$A_1(x) \le \sup_x \mathbf{P}([\langle \|D^j T(x)\|, L\mathsf{h}\rangle]_{j\le|\alpha|}) \le C_{|\alpha|}\tau^{|\alpha|}(L\mathsf{h})^{|\alpha|},$$

since we can assume without loss of generality that $\tau \ge 1$ and $L \ge 1$. As for A_2, we proceed as follows. First, applying Lemma D.4 we have that

$$\frac{1}{\mathsf{B}} = \Big(\int E(v,V)(y)\, d\nu(y)\Big)^{-1} = e^{\bar{u}(x)} \le e^{-\frac{1}{2}\|T(x)\|^2 + u(x)} \le e^{k(L\sigma)^2 + \|T(x)\|\sqrt{2k}L\sigma}$$

and

$$e^{v(y) - \frac{1}{2}\|V(y)\|^2} \le e^{k(L\sigma)^2 + \|V(y)\|\sqrt{2k}L\sigma}.$$

Using these inequalities, we obtain

$$\begin{aligned}
A_2 &\le c_1 \int \mathbb{1}_{\mathcal{D}^c} \mathbf{P}([\langle D^j T(x), V(y)\rangle]_{j\le|\alpha|}) e^{\|V(y)\|\sqrt{2k}L\sigma + \langle T(x), V(y)\rangle}\, d\nu(y) \\
&\le c_1 \int \mathbb{1}_{\mathcal{D}^c} \mathbf{P}([\langle D^j T(x), V(y)\rangle]_{j\le|\alpha|}) e^{\|V(y)\|(\sqrt{2k}L\sigma + \|T(x)\|)}\, d\nu(y) \\
&\le C_{|\alpha|} c_1 \tau^{|\alpha|} L^{|\alpha|} \Big(\int \mathbb{1}_{\mathcal{D}^c} \|y\|^{2|\alpha|}\, d\nu(y)\Big)^{1/2} \Big(\int \mathbb{1}_{\mathcal{D}^c} e^{2\|V(y)\|(\sqrt{2k}L\sigma + \|T(x)\|)}\, d\nu(y)\Big)^{1/2},
\end{aligned}$$

with $c_1 = e^{2k(L\sigma)^2 + \|T(x)\|\sqrt{2k}\sigma L}$. Now,

$$\Big(\int \mathbb{1}_{\mathcal{D}^c} \|y\|^{2|\alpha|}\, d\nu(y)\Big)^{1/2} \le e^{\frac{-\mathsf{h}^2}{8k\sigma^2}} \Big(\int \mathbb{1}_{\mathcal{D}^c} e^{\frac{\|y\|^2}{4k\sigma^2}} \|y\|^{2|\alpha|}\, d\nu(y)\Big)^{1/2},$$

and applying Young inequality, the subgaussianity of ν and Lemma D.2, we have

$$\Big(\int \mathbb{1}_{\mathcal{D}^c}\|y\|^{2|\alpha|}\,d\nu(y)\Big)^{1/2} \le e^{\frac{-\mathsf{h}^2}{8k\sigma^2}}\sqrt{2}(2|\alpha|)!^{1/4}(\sqrt{2k}\sigma)^{|\alpha|}.$$

Also,

$$\Big(\int \mathbb{1}_{\mathcal{D}^c}e^{2\|V(y)\|(\sqrt{2k}L\sigma+\|T(x)\|)}\,d\nu(y)\Big)^{1/2} \le 2e^{2L^2(\sqrt{2k}L\sigma+\|T(x)\|)^2k\sigma^2}.$$

Choosing $\mathsf{h}^2 \ge C_{k,|\alpha|}\sigma^2((L\sigma)^2+(L\sigma)^4)$ if $\|x\| \le \sqrt{k\sigma}$ and $\mathsf{h}^2 \ge C_{k,|\alpha|}\sigma^2(\sigma L\|x\| + \sigma^2L^4\|x\|^2)$ if $\|x\| > \sqrt{k\sigma}$ for a sufficiently large constant $C_{k,|\alpha|}$, then we have that

$$A_2 \le C_{k,|\alpha|}(\tau\sigma L)^{|\alpha|}.$$

Combining this with the bound on A_1, we obtain:

$$A(x) \le C_{k,|\alpha|}(\tau L\sigma)^{|\alpha|}(1+((L\sigma)^2+L\sigma)^{|\alpha|}) \qquad \text{if } \|x\| \le \sqrt{k\sigma},$$

and

$$A(x) \le C_{k,|\alpha|}(\tau L\sigma)^{|\alpha|}(1+(\sqrt{(L\sigma)\|x\|}+(L^2\sigma)\|x\|)^{|\alpha|}) \qquad \text{if } \|x\| > \sqrt{k\sigma}.$$

□

Lemma D.7. *In the assumptions of Lemma D.6 we have, for any multi-index α,*

$$|D^\alpha u(z)| \le C_{k,|\alpha|}\begin{cases}\tau L\|z\| + (\tau L\sigma)^{|\alpha|}(1+(L\sigma)^{2|\alpha|}) & \text{if } \|z\| \le \sqrt{k}\sigma \\ \tau L\|z\| + (\tau L\sigma)^{|\alpha|}(1+(L^2\sigma\|z\|)^{|\alpha|}) & \text{if } \|z\| > \sqrt{k}\sigma.\end{cases} \tag{D.3.10}$$

where C_k is a constant depending only on k.

Proof. The proof follows by easy manipulation of the terms in (D.3.9). □

Finally, we present the formal version of Lemma 5.3.

Lemma D.8. *Let $\mathcal{F}_{\sigma,\tau,L}$ be the space of functions satisfying inequalities (D.3.8) and (D.3.10). Let $\eta,\nu_1,\nu_2 \in \mathcal{G}_\sigma(\mathcal{Z})$ and $T,T' \in \mathcal{T}$ with $\mathcal{T}$ as in Thm. 5.2. Let S denote Sinkhorn*

divergence with $\varepsilon = 1$. *Then,*

$$|\mathsf{S}(T_{\#}\eta, T'_{\#}\nu_1) - \mathsf{S}(T_{\#}\eta, T'_{\#}\nu_2)| \;\leq\; \sup_{u\in\mathcal{F}_{\sigma,\tau,L}} |\int u(z)\, d\nu_1 - \int u(z)\, d\nu_2(z)|. \quad \text{(D.3.11)}$$

Proof. The proof follows exactly the same lines as the proof of (Mena and Niles-Weed, 2019, Cor 2) with this variant: the set $\mathcal{F}_\sigma$ is replaced by the set $\mathcal{F}_{\sigma,\tau,L}$, thanks to our estimates on the potentials in the bounds Lemma D.5 and Lemma D.7. □

Remark D.1. *Define the set* $\mathcal{F}^s$ *to be the set of functions satisfying*

$$|u(x)| \leq C_{s,k}(1 + \|x\|^2)$$
$$|D^\alpha u(x)| \leq C_{s,k}(1 + \|x\|^s) \quad |\alpha| \leq s.$$

Note that for a sufficiently big constant $C_{s,k}$, *for any* $u \in \mathcal{F}_{\sigma,\tau,L}$ *the function* $\frac{1}{1+(\tau L)^s+(\sigma L)^{3s}\tau^s} u$ *belongs to* $\mathcal{F}^s$.

Theorem D.9. *With the same notation as above, the following holds*

$$\mathbb{E} \sup_{T\in\mathcal{T},\eta\in\mathcal{H}} |\mathsf{S}_\varepsilon(T_{\#}\eta, \rho_n) - \mathsf{S}_\varepsilon(T_{\#}\eta, \rho)| \;\leq\; \frac{\mathsf{b}(\tau, L, \sigma, k)}{\sqrt{n}} \quad \text{(D.3.12)}$$

where $\mathsf{b}(\tau, L, \sigma, k) \;=\; \mathsf{C}_k\,(\tau L)^{\lceil \frac{k}{2}\rceil+1}\big(1+L^{k+2}(1+\sigma^{\lceil\frac{5k}{2}\rceil+6})\,\varepsilon^{-\lceil\frac{5k}{4}\rceil-3}\big)$ *with* C_k *a constant depending only on the latent space dimension* k.

Proof. We first set $\varepsilon = 1$ and consider S, and then obtain the bound for the general case. For a given $T \in \mathcal{T}$ and $\eta \in \mathcal{H}$, by (Mena and Niles-Weed, 2019, Prop 2), and Lemma D.1 we have

$$|\mathsf{S}(T_{\#}\eta, \rho_n) - \mathsf{S}(T_{\#}\eta, \rho)| = |\mathsf{S}(T_{\#}\eta, T^*_{\#}\eta^*_n) - \mathsf{S}(T_{\#}\eta, T^*_{\#}\eta^*)|$$
$$= |\mathsf{S}_{\mathsf{c}_{T,T^*}}(\eta, \eta^*_n) - \mathsf{S}_{\mathsf{c}_{T,T^*}}(\eta, \eta^*)| \leq \sup_{f\in\mathcal{F}_{\sigma,L,\tau}} |\int_{\mathcal{Z}} f\, d(\eta^*_n - \eta^*)|.$$

Note that the set $\mathcal{F}_{\sigma,L,\tau}$ is independent of the specific η, η^* and T, T^* and depends only on the properties of the classes that we consider, i.e. σ^2-sub-gaussianity and boundedness L and smoothness τ for functions in $\mathcal{T}$. Thus, we can take the supremum in the left hand side

over $\eta \in \mathcal{H}$ and $T \in \mathcal{T}$:

$$\sup_{T\in\mathcal{T},\eta\in\mathcal{H}} |\mathsf{S}(T_{\#}\eta, \rho_n) - \mathsf{S}(T_{\#}\eta, \rho)| \leq \sup_{f\in\mathcal{F}_{\sigma,L,\tau}} |\int_{\mathcal{Z}} f\, d(\eta_n^* - \eta^*)|.$$

From now on, recalling that for any $f \in \mathcal{F}_{\sigma,\tau,L}$ the function $\frac{1}{(\tau L)^s+(\sigma L)^{3s}\tau^s} f$ belongs to $\mathcal{F}^s$, the proof is identical to the proof of (Mena and Niles-Weed, 2019, Thm. 2) and it leads to the following bound for any ε:

$$\mathbb{E} \sup_{T\in\mathcal{T},\eta\in\mathcal{H}} |\mathsf{S}_\varepsilon(T_{\#}\eta, \rho_n) - \mathsf{S}_\varepsilon(T_{\#}\eta, \rho)| \;\leq\; \frac{\mathsf{b}(\tau, L, \sigma, k)}{\sqrt{n}}$$

where $\mathsf{b}(\tau, L, \sigma, k) \;=\; \mathsf{C}_k\,(\tau L)^{\lceil \frac{k}{2}\rceil+1}\big(1+L^{k+2}(1+\sigma^{\lceil\frac{5k}{2}\rceil+6})\,\varepsilon^{-\lceil\frac{5k}{4}\rceil-3}\big)$ with C_k a constant depending only on the latent space dimension k. □

D.4 Learning Rates

We provide here a formal statement of Thm. 5.2.

Theorem D.10. *Let $\mathcal{Z} \subset \mathbb{R}^k$, $\mathcal{X} \subset \mathbb{R}^d$ and $\rho = T^*_{\#}\eta^*$ with $T^* \in \mathcal{T} \subset C^{\lceil k/2\rceil+1}_{\tau,L}(\mathcal{Z}, \mathcal{X})$ and $\eta^* \in \mathcal{H} \subset \mathcal{G}_\sigma(\mathcal{Z})$. Let $(\hat{T}, \hat{\eta})$ satisfy (5.3.1) with $\mathsf{d}_{\mathcal{F}} = \mathsf{S}_\varepsilon$ and ρ_n a sample of n i.i.d. points from ρ. Then,*

$$\mathbb{E}\, \mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho) \;\leq\; \frac{\mathsf{b}(\tau, L, \sigma, k)}{\sqrt{n}}$$

where $\mathsf{b}(\tau, L, \sigma, k) \;=\; \mathsf{C}_k\,(\tau L)^{\lceil \frac{k}{2}\rceil+1}\big(1+L^{k+2}(1+\sigma^{\lceil\frac{5k}{2}\rceil+6})\,\varepsilon^{-\lceil\frac{5k}{4}\rceil-3}\big)$ with C_k a constant depending only on the latent space dimension k and where the expectation is taken with respect to ρ_n.

Proof. We decompose the error as follows

$$\mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho) - \mathsf{S}_\varepsilon(T^*_{\#}\eta^*, \rho) = A_1 + A_2 + A_3 \tag{D.4.1}$$

where

$$A_1 = \mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho) - \mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho_n) \tag{D.4.2}$$

$$A_2 = \mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho_n) - \mathsf{S}_\varepsilon(T^*_{\#}\eta^*, \rho_n) \tag{D.4.3}$$

$$A_3 = \mathsf{S}_\varepsilon(T^*_{\#}\eta^*, \rho_n) - \mathsf{S}_\varepsilon(T^*_{\#}\eta^*, \rho). \tag{D.4.4}$$

Note that by optimality of $\hat{T}$ and $\hat{\eta}$, $A_2 \leq 0$. Now,

$$A_1 + A_3 \leq 2 \sup_{T\in\mathcal{T},\eta\in\mathcal{H}} \Big[\mathsf{S}_\varepsilon(T_\#\eta, \rho_n) - \mathsf{S}_\varepsilon(T_\#\eta, \rho)\Big]. \tag{D.4.5}$$

Applying Thm. D.9 and combining it with (D.4.1) and (D.4.5) yields the desired result. □

D.4.1 Perturbation case

We conclude this section extending Thm. 5.2 to the case where the model is accurate up to a perturbation of the pushforward measure in terms of a subgaussian distribution.

Lemma D.11 (Pushforward of a sub-Gaussian measure)**.** *Let $T : \mathcal{Z} \to \mathcal{X}$ be a Lipschitz continuous map from $\mathcal{Z} \subset \mathbb{R}^k$ to $\mathcal{X} \subset \mathbb{R}^d$ with Lipschitz constant L and such that $T(0) = 0$. Let $\eta \in \mathcal{G}_\sigma(\mathcal{Z})$. Then $T_\#\eta \in \mathcal{G}_{\sigma_L}(\mathcal{X})$ with $\sigma_L = \sigma L\sqrt{k/d}$.*

Proof. The result follows by observing that for any σ_0 we have

$$\int_{\mathcal{X}} e^{\frac{\|x\|^2}{2d\sigma_0^2}}\, d(T_\#\eta)(x) = \int_{\mathcal{Z}} e^{\frac{\|T(z)\|^2}{2d\sigma_0^2}}\, d\eta(z) \leq \int_{\mathcal{Z}} e^{\frac{L^2\|z\|^2}{2d\sigma_0^2}}\, d\eta(z).$$

Since $\eta \in \mathcal{G}_\sigma(\mathcal{Z})$, we have that

$$\int_{\mathcal{Z}} e^{\frac{\|z\|^2}{2k\sigma^2}} \leq 2.$$

Choosing $\sigma_0 = \sigma L\sqrt{k/d}$ yields the we obtain that

$$\int_{\mathcal{X}} e^{\frac{\|x\|^2}{2d\sigma_0^2}}\, d(T_\#\eta)(x) \leq \int_{\mathcal{Z}} e^{\frac{\|z\|^2}{2k\sigma^2}} \leq 2$$

and hence $T_\#\eta$ belongs to $\mathcal{G}_{\sigma_L}(\mathcal{X})$ as required. □

Lemma D.12 (Convolution of two sub-Gaussian measures)**.** *Let $\sigma_1, \sigma_2 > 0$, $\mu \in \mathcal{G}_{\sigma_1}(\mathcal{X})$ and $\rho \in \mathcal{G}_{\sigma_2}(\mathcal{X})$. Then $\mu * \rho \in \mathcal{G}_{2\bar{\sigma}}$ with $\bar{\sigma} = \max(\sigma_1, \sigma_2)$.*

Proof. For any $\sigma > 0$ we have

$$\begin{aligned}
\int e^{\frac{\|x\|^2}{2d\sigma^2}}\, d(\mu * \rho)(x) &= \int e^{\frac{\|y+w\|^2}{2d\sigma^2}}\, d\mu(y)d\rho(w) \\
&\leq \int e^{\frac{\|y\|^2+\|w\|^2}{d\sigma^2}}\, d\mu(y)d\rho(w) \\
&\leq \left(\int e^{\frac{2\|y\|^2}{d\sigma^2}}\, d\mu(y)\right)^{1/2} \left(\int e^{\frac{2\|w\|^2}{d\sigma^2}}\, d\rho(w)\right)^{1/2}.
\end{aligned}$$

Now, if $\sigma \geq 2\sigma_1$, we have

$$\int e^{\frac{2\|y\|^2}{d\sigma^2}} \, d\mu(y) \leq \int e^{\frac{\|y\|^2}{2d\sigma_1^2}} \, d\mu(y) \leq 2.$$

Analogously for $\sigma \geq 2\sigma_2$. Therefore by taking $\sigma = 2\bar{\sigma}$ with $\bar{\sigma} = \max(\sigma_1, \sigma_2)$, we have

$$\int e^{\frac{\|x\|^2}{2d\sigma^2}} \, d(\mu * \rho)(x) \leq 2,$$

and hence $\mu * \rho \in \mathcal{G}_{2\bar{\sigma}}$ as required. □

Lemma D.13 (Perturbation)**.** *Let $\mu \in \mathcal{G}_\sigma$ with $\sigma \geq 1$. Let $\Phi_\delta \in \mathcal{G}_\delta$ for $0 \leq \delta \leq \sigma$. Then, for any $\nu \in \mathcal{G}_{2\sigma}$, we have*

$$|\mathsf{S}_\varepsilon(\nu, \Phi_\delta * \mu) - \mathsf{S}_\varepsilon(\nu, \mu)| \leq \mathsf{b}_1(\sigma, d) \, \delta$$

with

$$\mathsf{b}_1(\sigma, d) = 6d^{3/2}\sigma(1 + C_{1,d}4\sigma^2(1 + 2\sigma))$$

and $C_{1,d}$ a constant depending only on the ambient dimension d.

Proof. Since $\delta \leq \sigma$, by applying Lemma D.12 we have $\Phi_\delta * \mu \in \mathcal{G}_{2\sigma}$ and $\mu \in \mathcal{G}_\sigma \subset \mathcal{G}_{2\sigma}$. Therefore, we apply Lemma D.8 to control

$$|\mathsf{S}_\varepsilon(\nu, \Phi_\delta * \mu) - \mathsf{S}_\varepsilon(\nu, \mu)| \leq \sup_{u \in \mathcal{F}_{2\sigma}} \left| \int u(x + w) \, d\mu(x) d\Phi_\delta(w) - \int u(x) \, d\mu(x) \right| \tag{D.4.6}$$

$$= \sup_{u \in \mathcal{F}_{2\sigma}} \left| \int \big(u(x + w) - u(x)\big) \, d\Phi_\delta(w) d\mu(x) \right|. \tag{D.4.7}$$

Note that for any $x, w \in \mathcal{X}$ we can define $H : [0, 1] \to \mathbb{R}$ such that for any $t \in [0, 1]$

$$H(t) = u(x + tw).$$

Then, by the fundamental theorem of calculus we have

$$\begin{aligned}\int_0^1 H'(t)\, dt &= H(1) - H(0) \\ &= u(x+w) - u(x).\end{aligned}$$

Now

$$H'(t) = \langle \nabla u(x+tw), w \rangle ,$$

which implies

$$\begin{aligned}|u(x+w) - u(x)| &\leq \int_0^1 \|\nabla u(x+tw)\| \|w\|\, dt \\ &\leq \sqrt{d}\|w\| \int_0^1 \|\nabla u(x+tw)\|_\infty .\end{aligned}$$

Now, by direct application of (Mena and Niles-Weed, 2019, Prop. 1) for the functions in $\mathcal{F}_{2\sigma}$, we have

$$|D^1 u(x)| \leq \begin{cases} \|x\| + C_{1,d} 4\sigma^2 (1+2\sigma) & \text{if} \quad \|x\| \leq \sqrt{d} 2\sigma \\ \|x\|(1 + C_{1,d} 2^{3/2} \sigma^{3/2} (1 + (2\sigma)^{1/2})) & \text{otherwise.} \end{cases}$$

Therefore, since $\sigma > 1$ we can upper bound $\sigma^{2/3}$ with σ^2 and $\sigma^{1/2}$ with σ, to obtain the neater formula below

$$\|\nabla f(x)\|_\infty \leq \mathsf{b}_0(\sigma, d)(1 + \|x\|) \qquad \text{with} \qquad \mathsf{b}_0(\sigma, d) = 1 + C_{1,d} 4\sigma^2 (1+2\sigma).$$

We can therefore bound

$$\begin{aligned}\int_0^1 \|\nabla u(x+tw)\|\, dt &\leq \mathsf{b}_0(\sigma, d) \int_0^1 (1 + \|x+tw\|)\, dt \\ &\leq \mathsf{b}_0(\sigma, d)(1 + \|x\| + \|w\|).\end{aligned}$$

Combining the steps above, for any $x, w \in \mathcal{X}$

$$|u(x+w) - u(x)| \leq \sqrt{d} \mathsf{b}_0 s(\sigma, d)(\|x\|\|w\| + \|w\|^2 + \|w\|).$$

Plugging this inequality in (D.4.7), we have

$$\sup_{u\in\mathcal{F}_\sigma} \int u(x+w)-u(x)\,d\mu(x) \le \sqrt{d}\mathsf{b}_0(\sigma,d)\int(\|x\|\|w\|+\|w\|^2+\|w\|)\,d\mu(x)d\Phi_\delta(w)$$
$$\le 2d^{3/2}\mathsf{b}_0(\sigma,d)\,\delta(\sigma+\delta+1),$$

where we have used Lemma D.2 in the last inequality. Since $\delta \le \sigma$ and $\sigma \ge 1$, we have that $1+\delta+\sigma \le 3\sigma$ and the inequality above yields the required result. □

Lemma D.14 (Perturbation on samples)**.** *Let $\mu \in \mathcal{G}_\sigma$ with $\sigma \ge 1$. Let $\Phi_\delta \in \mathcal{G}_\delta$ for $0 \le \delta \le \sigma$. Consider $\rho = \Phi_\delta * \mu$ and denote by ρ_n and μ_n the empirical measures $\rho_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i+w_i}$ and $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ where x_i are sampled i.i.d. from μ and w_i are samples i.i.d. from Φ_δ. Then, for any $\nu \in \mathcal{G}_{2\sigma}$, we have*

$$\mathbb{E}|\mathsf{S}_\varepsilon(\nu,\rho_n)-\mathsf{S}_\varepsilon(\nu,\mu_n)| \le \mathsf{b}_1(\sigma,d)\,\delta$$

with

$$\mathsf{b}_1(\sigma,d) = 6d^{3/2}\sigma(1+C_{1,d}4\sigma^2(1+2\sigma))$$

and $C_{1,d}$ a constant depending only on the ambient dimension d.

Proof. Similarly to the proof of Lemma D.13, we have

$$|\mathsf{S}_\varepsilon(\nu,\rho_n)-\mathsf{S}_\varepsilon(\nu,\mu_n)| \le \sup_{u\in\mathcal{F}_{2\sigma}} |\int u(y)\,d\rho_n(y) - \int u(y)\,d\mu_n(y)| \tag{D.4.8}$$
$$= \sup_{u\in\mathcal{F}_{2\sigma}} |\frac{1}{n}\sum_{i=1}\big(u(x_i+w_i)-u(x_i)\big)|. \tag{D.4.9}$$

Using the upper bound on the gradient of u proved in Lemma D.13, we obtain

$$|\mathsf{S}_\varepsilon(\nu,\rho_n)-\mathsf{S}_\varepsilon(\nu,\mu_n)| \le \sup_{u\in\mathcal{F}_{2\sigma}} |\frac{1}{n}\sum_{i=1}\big(u(x_i+w_i)-u(x_i)\big)| \tag{D.4.10}$$
$$\le \sqrt{d}\mathsf{b}_0(\sigma,d)\frac{1}{n}\sum_{i=1}^n(\|x_i\|\|w_i\|+\|w_i\|^2+\|w_i\|). \tag{D.4.11}$$

Taking the expectation, we get

$$\mathbb{E}|\mathsf{S}_\varepsilon(\nu,\rho_n)-\mathsf{S}_\varepsilon(\nu,\mu_n)|\le \mathbb{E}_{x\sim\mu,\,w\sim\Phi_\delta}\sqrt{d}\mathsf{b}_0(\sigma,d)\frac{1}{n}\sum_{i=1}^{n}(\|x_i\|\|w_i\|+\|w_i\|^2+\|w_i\|), \tag{D.4.12}$$

and using Lemma D.2 we conclude

$$\mathbb{E}|\mathsf{S}_\varepsilon(\nu,\rho_n)-\mathsf{S}_\varepsilon(\nu,\mu_n)|\le 2d^{3/2}\mathsf{b}_0(\sigma,d)\;\delta(\sigma+\delta+1). \tag{D.4.13}$$

Since $\delta\le\sigma$ and $\sigma\ge 1$, we have that $1+\delta+\sigma\le 3\sigma$ and the inequality above yields the required result. □

We are finally ready to prove our result on perturbed GAN models.

Corollary D.15 (Formal version of Cor. 5.4)**.** *Under the same assumption of Thm. 5.2, let $\Phi_\delta\in\mathcal{G}_\delta(\mathcal{X})$ and $\rho=\Phi_\delta * T^*_\#\eta^*$. Let c be the same constant of Thm. 5.2. Then,*

$$\mathbb{E}\;\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},\rho)\;\le\;\frac{2\,\mathsf{b}(\tau,L,\sigma,k)}{\sqrt{n}}\;+\;3\mathsf{b}_1(L\sigma\sqrt{k/d},d)\;\delta.$$

with $\mathsf{b}_1(L\sigma\sqrt{k/d},d)$ the constant defined in Lemma D.13.

Proof. Let ρ_n be the empirical sample used to obtain $(\hat{T},\hat{\eta})$. By definition of ρ, we have that ρ_n corresponds to an empirical sample of points $(T^*(z_i)+w_i)_{i=1}^n$ with z_i i.i.d. points sampled from η^* and w_i i.i.d. points sampled from Φ_δ. We denote $\eta_n^*=\frac{1}{n}\sum_{i=1}^n\delta_{z_i}$. Consider the following decomposition of the error

$$\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},\rho)=A_1+A_2+A_3+A_4+A_5+A_6$$

with

$$\begin{aligned}
A_1&=\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},\rho)-\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},T^*_\#\eta^*)\\
A_2&=\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},T^*_\#\eta^*)-\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},T^*_\#\eta_n^*)\\
A_3&=\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},T^*_\#\eta_n^*)-\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},\rho_n)\\
A_4&=\mathsf{S}_\varepsilon(\hat{T}_\#\hat{\eta},\rho_n)-\mathsf{S}_\varepsilon(T^*_\#\eta^*,\rho_n)\\
A_5&=\mathsf{S}_\varepsilon(T^*_\#\eta^*,\rho_n)-\mathsf{S}_\varepsilon(T^*_\#\eta^*,T^*_\#\eta_n^*)\\
A_6&=\mathsf{S}_\varepsilon(T^*_\#\eta^*,T^*_\#\eta_n^*)
\end{aligned}$$

We start by controlling the term A_1. First we note that according to Lemma D.11, both distributions $\hat{T}_{\#}\hat{\eta}$ and $T^*_{\#}\eta^*$ are sub-Gaussian with parameter $L\sigma\sqrt{k/d}$. Therefore, by applying Lemma D.13 we obtain

$$A_1 \leq \mathsf{b}_1(L\sigma\sqrt{k/d}, d)\ \delta$$

where b_1 is the constant introduced in Lemma D.13. As for A_3 and A_5, we bound them using Lemma D.14 and we obtain

$$\mathbb{E}[A_3] \leq \mathsf{b}_1(L\sigma\sqrt{k/d}, d)\ \delta$$
$$\mathbb{E}[A_5] \leq \mathsf{b}_1(L\sigma\sqrt{k/d}, d)\ \delta.$$

The term A_2 corresponds to the sample complexity of $T^*_{\#}\eta^*$. Therefore we can apply Thm. D.9 to obtain

$$\mathbb{E}[A_2] = \mathbb{E}\left[\mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, T^*_{\#}\eta^*) - \mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, T^*_{\#}\eta^*_n)\right] \leq \frac{\mathsf{b}(\sigma,\tau,L)}{\sqrt{n}}.$$

The same holds for A_6, namely

$$\begin{aligned}\mathbb{E}[A_6] &= \mathbb{E}\left[\mathsf{S}_\varepsilon(T^*_{\#}\eta^*, T^*_{\#}\eta^*_n)\right] \\ &= \mathbb{E}\left[\mathsf{S}_\varepsilon(T^*_{\#}\eta^*, T^*_{\#}\eta^*_n) - \mathsf{S}_\varepsilon(T^*_{\#}\eta^*, T^*_{\#}\eta^*)\right] \\ &\leq \frac{\mathsf{b}(\sigma,\tau,L)}{\sqrt{n}}.\end{aligned}$$

Finally, we note that since $(\hat{T}, \hat{\eta})$ is the minimizer of $\mathsf{S}_\varepsilon(T_{\#}\eta, \rho_n)$,

$$A_4 = \mathsf{S}_\varepsilon(\hat{T}_{\#}\hat{\eta}, \rho_n) - \mathsf{S}_\varepsilon(T^*_{\#}\eta^*, \rho_n) \leq 0.$$

Combining all the bounds above yields the required result. □

D.5 Optimization

D.5.1 Computing the gradient with respect to the network parameters

In this section we provide the analytic formula for the gradient of the Sinkhorn divergence with respect to the generator network parameters. We recall here the statement.

Proposition 5.5. *Let $\eta \in \mathcal{P}(\mathcal{Z})$ and $\rho \in \mathcal{P}(\mathcal{X})$. Let (u^*, v^*) be a pair of minimizers of (4.1.1) with $\alpha = T_{\theta\#}\eta$ and $\beta = \rho$. Then, the gradient of $\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho)$ in θ_0 is*

$$\left[\nabla_\theta \mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho)\right]|_{\theta=\theta_0} = \int \left[\nabla_x u^*(\cdot)\right]|_{x=T_{\theta_0}(z)} \left[\nabla_\theta T_\theta(z)\right]|_{\theta=\theta_0} \, d\eta(z). \tag{5.4.2}$$

The proof of the result hinges on the following characterization of the directional derivative of functionals that admit a variational form. Note that here we consider a generic smooth cost function $\mathsf{c} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ in the definition of OT_ε, even though most of the material in the chapter was presented with the squared Euclidean distance.

Theorem D.16 (Thm. 4.13 in Bonnans and Shapiro (2013))**.** *Let $f : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is a continuous function such that for all $x \in \mathcal{X}$ the function $f(x, \cdot)$ is (Gateaux) differentiable, that $f(x, u)$ and $\nabla_u f(x, u)$ are continuous on $\mathcal{X} \times \mathcal{U}$, and that the inf-compactness condition holds. Then the optimal value function*

$$v : u_0 \mapsto \inf_{x \in \mathcal{X}} f(x, u_0),$$

is Fréchet directionally differentiable at u_0 with directional derivative

$$v'(u_0; \bar{u}) = \inf_{x \in \mathcal{S}(u_0)} \nabla_u f(x, u_0)\bar{u},$$

for any $\bar{u} \in U$, with $\mathcal{S}(u_0)$ the set of minimizer of $f(\cdot, u_0)$.

We recall that inf-compactness is the condition requiring the existence of a neighborhood of u_0 and a constant such that the level sets of $f(\cdot, u)$ associated to such constant are compact for any u in such neighborhood. We note that the same result holds when considering the supremum of a joint function $f(x, u_0)$, which is the case of the Sinkhorn divergence considered in the following.

Let now $\eta \in \mathcal{P}(\mathcal{Z}), \rho \in \mathcal{P}(\mathcal{X})$ and Θ a space of parameters for the pushforward maps $T_\theta : \mathcal{Z} \to \mathcal{X}$. We will apply Thm. D.16 to the functional

$$F(\theta) = \mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho) = \sup_{u,v \in \mathcal{C}(\mathcal{X})} G(T_{\theta\#}\eta, \rho, u, v),$$

where we have denoted

$$G(T_{\theta\#}\eta, \rho, u, v) = \int u(x)\, d(T_{\theta\#}\eta)(x) + \int v(y)\, d\rho(y) \\ - \varepsilon \int e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d(T_{\theta\#}\eta)(x) d\rho(y).$$

We recall that the solution (u^*, v^*) to the Sinkhorn dual problem is unique up to a constant shift $(u^* + r, v^* - r)$ for any $r \in \mathbb{R}$ (Feydy et al., 2019). We can therefore restrict the above optimization problem to

$$F(\theta) = \sup_{(a,b)\in\mathcal{D}} G(T_{\theta\#}\eta, \rho, u, v), \tag{D.5.1}$$

to the domain

$$\mathcal{D} = \left\{ (u, v) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X}) \;\middle|\; \int u(x)\, d(T_{\theta\#}\eta)(x) = \int v(y)\, d\rho(y) \right\}.$$

Therefore, over this linear subspace of $\mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X})$, the functional $\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho, \cdot, \cdot)$ admits a unique minimizer and is actually strictly concave, which guarantees inf-compactness (actually sup-compactness in this case) to hold.

We can therefore apply Thm. D.16 with the following substitutions in our setting: $x \leftarrow (u, v)$, $\mathcal{X} \leftarrow \mathcal{D}$, $u \leftarrow \theta$ and $U \leftarrow \Theta$. Let (u^*, v^*) be the minimizer of (D.5.1). We have

$$[\nabla_\theta F(\cdot)]|_{\theta=\theta_0} = \nabla_\theta [G(T_{\theta\#}\eta, \rho, u^*, v^*)|_{\theta=\theta_0}.$$

Now, by applying the Transfer lemma, we have

$$G(T_{\theta\#}\eta, \rho, u^*, v^*) = G(\eta, \rho, u^* \circ T_\theta, v^*)$$

and therefore,

$$\begin{aligned}
[\nabla_\theta F(\cdot)]|_{\theta=\theta_0} &= \nabla_\theta [G(\eta, \rho, u^* \circ T_\theta, v^*)|_{\theta=\theta_0} \\
&= \int \left[\nabla_\theta u^*(T_\theta(z))\right]|_{\theta=\theta_0}\, d\eta(z) \\
&\quad - \varepsilon \int \left[\nabla_\theta e^{\frac{u^*(T_\theta(z))+v^*(y)-\mathsf{c}(T(z),y)}{\varepsilon}}\right]|_{\theta=\theta_0}\, d\eta(z) d\rho(y).
\end{aligned}$$

Recall that u^* is differentiable (actually C^∞, see e.g. (Genevay et al., 2018a, Thm. 2) characterizing the regularity of Sinkhorn potential when using a smooth cost). The chain rule yields

$$\int \left[\nabla_\theta u^*(T_\theta(z))\right]|_{\theta=\theta_0}\, d\eta(z) = \int \left[\nabla_x u^*(\cdot)|_{x=T_{\theta_0}(z)}\right]\left[\nabla_\theta T_\theta(z)\right]|_{\theta=\theta_0}\, d\eta(z). \quad \text{(D.5.2)}$$

By computing the gradient of the exponential term, the second term in the gradient can be split in two parts

$$\varepsilon \int \left[\nabla_\theta\, e^{\frac{u^*(T_\theta(z))+v^*(y)-c(T_\theta(z),y)}{\varepsilon}}\right]|_{\theta=\theta_0}\, d\eta(z)d\rho(y) = A_1 - A_2$$

with

$$\begin{aligned} A_1 &= \int \left[\nabla_\theta u^*(T_\theta(z))\right]|_{\theta=\theta_0}\, e^{\frac{u^*(T_{\theta_0}(z))+v^*(y)-c(T_{\theta_0}(z),y)}{\varepsilon}}\, d\eta(x)d\rho(y) \\ A_2 &= \int \left[\nabla_\theta c(T_\theta(z),y)\right]|_{\theta=\theta_0}\, e^{\frac{u^*(T_{\theta_0}(z))+v^*(y)-c(T_{\theta_0}(z),y)}{\varepsilon}}\, d\eta(x)d\rho(y). \end{aligned}$$

Recall that since (u^*, v^*) is a pair of minimizers, the characterization of u^* from (5.4.1) holds, implying that for any $z \in \mathcal{Z}$,

$$\int e^{\frac{u^*(T_{\theta_0}(z))+v^*(y)-c(T_{\theta_0}(z),y)}{\varepsilon}}\, d\rho(y) = 1.$$

Therefore

$$\begin{aligned} A_1 &= \int \left[\nabla_\theta u^*(T_\theta(z))\right]|_{\theta=\theta_0} \left(\int e^{\frac{u^*(T_{\theta_0}(z))+v^*(y)-c(T_{\theta_0}(z),y)}{\varepsilon}}\, d\rho(y) \right) d\eta(z) \\ &= \int \left[\nabla_\theta u^*(T_\theta(z))\right]|_{\theta=\theta_0}\, d\eta(z). \end{aligned}$$

Hence, analogously to (D.5.2) we have

$$A_1 = \int \left[\nabla_x u^*(\cdot)|_{x=T_{\theta_0}(z)}\right]\left[\nabla_\theta T_\theta(z)\right]|_{\theta=\theta_0}\, d\eta(z).$$

Regarding the term A_2, we apply the chain rule to the cost term, obtaining

$$A_2 = \int \left[\nabla_x c(\cdot,y)|_{x=T_{\theta_0}(z)}\right]\left[\nabla_\theta T_\theta(z)\right]|_{\theta=\theta_0}\, e^{\frac{u^*(T_{\theta_0}(z))+v^*(y)-c(T_{\theta_0}(z),y)}{\varepsilon}}\, d\eta(z)d\rho(y).$$

Since the term in (D.5.2) and A_1 eliminate each other, we have

$$[\nabla_\theta F(\cdot)]|_{\theta=\theta_0} = \int \left[\nabla_x c(\cdot, y)|_{x=T_{\theta_0}(z)}\right] \left[\nabla_\theta T_\theta(z)\right]|_{\theta=\theta_0} e^{\frac{u^*(T_\theta(z))+v^*(y)-c(T_\theta(z),y)}{\varepsilon}} d\eta(z) d\rho(y).$$

Now, by the characterization of Sinkhorn potential in (5.4.1), we have that for any $x_0 \in \mathcal{X}$,

$$\nabla_x u^*(\cdot)|_{x=x_0} = \int \nabla_x c(\cdot, y)|_{x=x_0} e^{\frac{u^*(x_0)+v(y)-c(x_0,y))}{\varepsilon}} d\rho(y).$$

Replacing the equality above in the formula of $\nabla_\theta F$, we have

$$\left[\nabla_\theta F(\cdot)\right]|_{\theta=\theta_0} = \int \left[\nabla_z u^*(\cdot)\right]|_{x=T_{\theta_0}(z)} \left[\nabla_\theta T_\theta(z)\right]|_{\theta=\theta_0} d\eta(z),$$

as required.

Gradient of the Sinkhorn Divergence. Prop. 5.5 characterizes the gradient of $\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho)$ with respect to the network parameters θ. However, the Sinkhorn divergence, defined in (2.5.1) depends also on the so-called *autocorrelation* term $-\frac{1}{2}\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, T_{\theta\#}\eta)$. By following the same reasoning in the proof of Prop. 5.5, we have that

$$\left[\nabla_\theta \mathsf{OT}_\varepsilon(T_{\theta\#}\eta, T_{\theta\#}\eta)\right]|_{\theta=\theta_0} = 2 \int \left[\nabla_z u^*(\cdot)\right]|_{x=T_{\theta_0}(z)} \left[\nabla_\theta T_\theta(z)\right]|_{\theta=\theta_0} d\eta(z)$$

with u^* the Sinkhorn potential minimizing

$$\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, T_{\theta\#}\eta) = \sup_{u \in \mathcal{C}(\mathcal{X})} G(T_{\theta\#}\eta, T_{\theta\#}\eta, u, u).$$

Thanks to the linearity of the gradient, we can therefore compute the gradient of the Sinkhorn divergence by combining the gradient of $\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, \rho)$ and $\mathsf{OT}_\varepsilon(T_{\theta\#}\eta, T_{\theta\#}\eta)$.

D.6 Experiments

In this section we provide the details on the experimental setup of section Sec. 5.5.

Spiral. We describe the setting reported in Fig. 5.2, where the target $\rho \in \mathcal{P}(\mathbb{R}^2)$ is a multi-modal distribution on a spiral-shaped 1D manifold in $\mathbb{R}^2$. In particular we modeled

$\rho^* = T_\# \eta^*$ with η a mixture of three Gaussian distributions on $\mathbb{R}$,

$$\eta^* = \frac{1}{3}\sum_{j=1}^{3} \mathcal{N}(m_j, \sigma)$$

with means respectively in $m_1 = 0.1$, $m_2 = 0.7$ and $m_3 = 0.9$ and same variance $\sigma^2 = 0.1$. To map η to $\mathbb{R}^2$ we considered the pushforward map $T^* : \mathbb{R} \to \mathbb{R}^2$ such that

$$T^*(x) = (x\sin(2\pi x),\ x\cos(2\pi x)).$$

To approximate T we considered a fully connected neural network with 4 hidden layers with dimensions $256, 1024, 256, 256$, two ReLUs activation functions for the first and second layers and one sigmoid ($\tanh$) for the third layer. To minimize $\mathsf{S}_\varepsilon(T_\#\eta, \rho_n)$ in T for η fixed, we used ADAM as optimizer, with learning rate of $\alpha_1 = 10^{-4}$. To learn η for T fixed we used step size $\alpha_2 = 10^{-3}$. We run Alg. 5.1 with $m = 1000$ particles and sampling size $\ell = 100$ (the number of "perturbation" points sampled around the particles at each iteration). We set the regularization parameter of the Sinkhorn divergence equal to $\varepsilon = 0.005$. When keeping η fixed, we chose η to be $\mathcal{N}(0.5, 1)$.

Swiss Roll. Similary to the previous setting we considered $\rho = T^*_\# \eta^*$ the pushforward measure of a multimodal latent distribution. Here $\rho \in \mathcal{P}(\mathbb{R}^3)$, with $\eta^* \in \mathcal{P}(\mathbb{R}^2)$ the "restriction" of a Gaussian mixture on $[0,1]^2$. More formally, let $g : \mathbb{R}^2 \to \mathbb{R}$ be the density of the mixture of 3 isotropic Gaussian measures

$$\rho = \frac{1}{3}\sum_{j=1}^{3} \mathcal{N}(m_j, \Sigma)$$

with means $m_1 = (0.4, 0.4)$, $m_2 = (0.2, 0.8)$ and $m_3 = (0.8, 0.5)$, and same covariance $\Sigma = \sigma I$ with $\sigma^2 = 0.15$. Then we consider η^* the distribution with density $h : \mathbb{R}^2 \to \mathbb{R}$, proportional to

$$h \propto g \cdot \mathbb{1}_{[0,1]^2},$$

where $\mathbb{1}_{[0,1]^2}$ is the indicator function of the interval $[0,1]^2$. We used the pushforward map

of the swiss roll $T^* : \mathbb{R}^2 \to \mathbb{R}^3$

$$T(x, y) \;=\; (\, x\cos(2\pi x), y, x\sin(2\pi x)\,).$$

To approximate T we considered the same structure used in the spiral setting: a fully connected neural network with 4 hidden layers with dimensions $256, 1024, 256, 256$, two ReLUs activation functions for the first and second layers and one sigmoid ($\tanh$) for the third layer. To minimize $\mathsf{S}_\varepsilon(T_\#\eta, \rho_n)$ in T for η fixed, we used ADAM as optimizer, with learning rate of $\alpha_1 = 5 \cdot 10^{-5}$. To learn η for T fixed we used step size $\alpha_2 = 10^{-4}$. We run Alg. 5.1 with $m = 1000$ particles and sampling size $\ell = 100$ performing block-coordinate descent, alternating 50 iterations when learning T with η fixed and 20 iterations when learning η for a fixed T. We set the regularization parameter of the Sinkhorn divergence starting from $\varepsilon = 2$ and decreasing every 50 iterations of the generator training (both for Alg. 5.1 and the standard Sinkhorn GAN) by a factor $\varepsilon \leftarrow 0.9 \cdot \varepsilon$. We did not allow ε to drop below 10^{-3} When keeping the latent fixed, we chose η to be $\mathcal{N}([0.5, 0.5], I)$.

Bibliography

Adams, R. A. and Fournier, J. J. (2003). *Sobolev spaces*. Elsevier.

Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Analysis*, 43(2):904–924.

Aliprantis, C. D. and Border, K. (2006). *Infinite Dimensional Analysis: a Hitchhiker's guide*. Springer Science & Business Media.

Altschuler, J., Bach, F., Rudi, A., and Niles-Weed, J. (2019). Massively scalable Sinkhorn Distances via the Nyström Method. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4427–4437.

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1961–1971.

Alvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

Ambrosio, L. and Gigli, N. (2013). A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, pages 214–223.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035.

Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012). On the Equivalence between Herding and Conditional Gradient Algorithms. In *International Conference on Machine Learning (ICML)*, pages 1355–1362.

Bai, Y., Ma, T., and Risteski, A. (2018). Approximability of Discriminators Implies Diversity in GANs. In *International Conference on Learning Representations (ICLR)*.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Ben-Yosef, M. and Weinshall, D. (2018). Gaussian Mixture Generative Adversarial Networks for Diverse Datasets, and the Unsupervised Clustering of Images. *arXiv preprint arXiv:1808.10356*.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 2(37):A1111–A1138.

Bengio, Y. (2000). Gradient-based Optimization of Hyperparameters. *Neural computation*, 12(8):1889–1900.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bertsimas, D. and Tsitsiklis, J. (1997). *Introduction to Linear Optimization*. Athena Scientific, 1st edition.

Bhushan Damodaran, B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In *European Conference on Computer Vision (ECCV)*, pages 447–463.

Birkhoff, G. (1957). Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bogachev, V. I. (2007). *Measure theory*, volume 1. Springer Science & Business Media.

Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563.

Bonnans, J. F. and Shapiro, A. (2013). *Perturbation analysis of optimization problems*. Springer Science & Business Media.

Boyd, N., Schiebinger, G., and Recht, B. (2017). The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems. *SIAM Journal on Optimization*, 27(2):616–639.

Bredies, K. and Pikkarainen, H. K. (2013). Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218.

Brezis, H. (2010). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York.

Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. (2017). Convergence of Entropic Schemes for Optimal Transport and Gradient Flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418.

Chizat, L. (2019). Sparse Optimization on Measures with Over-parameterized Gradient Descent. *arXiv preprint arXiv:1907.10300*.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). Scaling Algorithms for Unbalanced Optimal Transport problems. *Mathematics of Computation*, 87(314):2563–2609.

Chouquet, G. (1969). *Lectures on Analysis, Vol. II*. W. A. Bejamin, Inc., Reading, MA, USA.

Ciliberto, C., Rosasco, L., and Rudi, A. (2016). A Consistent Regularization Approach for Structured Prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4412–4420.

Ciliberto, C., Rudi, A., Rosasco, L., and Pontil, M. (2017). Consistent Multitask Learning with Nonlinear Output Relations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1983–1993.

Claici, S., Chien, E., and Solomon, J. (2018). Stochastic Wasserstein Barycenters. In *International Conference on Machine Learning (ICML)*, pages 999–1008.

Cominetti, R. and Martín, J. S. (1994). Asymptotic Analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1):169–187.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3730–3739.

Courty, N., Flamary, R., and Tuia, D. (2014). Domain adaptation with regularized optimal transport. In *ECML/PKDD 2014*, LNCS, pages 1–16, Nancy, France.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2292–2300.

Cuturi, M. and Doucet, A. (2014). Fast Computation of Wasserstein Barycenters. In *International Conference on Machine Learning (ICML)*, pages 685–693.

Cuturi, M. and Peyré, G. (2016). A Smoothed Dual Approach for Variational Wasserstein Problems. *SIAM J. Imaging Sciences*, 9(1):320–343.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Demyanov, V. F. and Rubinov, A. M. (1967). The Minimization of Smooth Convex Functional on a Convex Set. *J. SIAM Control.*, 5(2):280–294.

Demyanov, V. F. and Rubinov, A. M. (1968). Minimization of Functionals in Normed Spaces. *J. SIAM Control.*, 6(1):73–88.

Deshpande, I., Zhang, Z., and Schwing, A. G. (2018). Generative Modeling using the Sliced Wasserstein Distance. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3483–3491.

Devroye, L., Gyorfi, L., et al. (1990). No empirical probability measure can converge in the total variation sense for all distributions. *The Annals of Statistics*, 18(3):1496–1499.

Dognin, P., Melnyk, I., Mroueh, Y., Ross, J., Dos Santos, C., and Sercu, T. (2018). Wasserstein Barycenter Model Ensembling. In *International Conference on Learning Representations (ICLR)*.

Dudley, R. M. (1969). The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50.

Dunn, J. C. and Harshbarger, S. (1978). Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444.

Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., and Nedich, A. (2018). Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10760–10770.

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. (2018). Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm. In *International Conference on Machine Learning (ICML)*, pages 2196–2220.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence (UAI)*, pages 258–267.

Edwards, C. (2012). *Advanced Calculus of Several Variables*. Dover Publications.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. (2019). Learning with minibatch wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. (2014). Regularized Discrete Optimal Transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-I., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics (AIStats)*.

Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. (2018). Wasserstein Discriminant Analysis. *Machine Learning*.

Flamary, R., Lounici, K., and Ferrari, A. (2019). Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*.

Franklin, J. and Lorenz, J. (1989). On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2053–2061.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2018a). Sample complexity of Sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics (AIStats)*.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for Large-scale Optimal Transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3440–3448.

Genevay, A., Peyré, G., and Cuturi, M. (2018b). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics (AIStats)*, pages 1608–1617.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680.

Google, I. (2017). *QuickDraw Dataset*. Available on line.

Gramfort, A., Peyré, G., and Cuturi, M. (2015). Fast Optimal Transport Averaging of Neuroimaging Data. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 261–272.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel Method for the Two-Sample Problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 513–520.

He, S., Liu, K., Ji, G., and Zhao, J. (2015). Learning to Represent Knowledge Graphs with Gaussian Embedding. In *International on Conference on Information and Knowledge Management (CIKM)*, pages 623–632. ACM.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural computation*, 18(7):1527–1554.

Hsieh, Y.-P., Liu, C., and Cevher, V. (2019). Finding Mixed Nash Equilibria of Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, pages 2810–2819.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134.

Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *International Conference on Machine Learning (ICML)*, pages 427–435.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. *International Conference on Learning Representation (ICLR)*.

Knopp, P. and Sinkhorn, R. (1968). A note concerning simultaneous integral equations. *Canadian Journal of Mathematics*, 20:855–861.

Kollo, T. and von Rosen, D. (2006). *Advanced multivariate statistics with matrices*, volume 579. Springer Science & Business Media.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). Generalized Sliced Wasserstein Distances. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 261–272.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From Word Embeddings to Document Distances. In *International Conference on Machine Learning (ICML)*, pages 957–966.

Lacoste-Julien, S. and Jaggi, M. (2015). On the Global Linear Convergence of Frank-Wolfe Optimization Variants. In *Advances in Neural Information Processing Systems (NIPS)*, pages 496–504.

Lacoste-Julien, S., Lindsten, F., and Bach, F. (2015). Sequential Kernel Herding: Frank-Wolfe Optimization for Particle Filtering. In *International Conference on Artificial Intelligence and Statistics (AIStats)*, pages 544–552.

Ledig, C., Theis, L., Huszár, F., Caballero, et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690.

Lemmens, B. and Nussbaum, R. (2012). *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press.

Lemmens, B. and Nussbaum, R. (2013). Birkhoff's version of Hilbert's metric and its applications in analysis. *arXiv preprint arXiv:1304.7921*.

Lever, G., Baldassarre, L., Patterson, S., Gretton, A., Pontil, M., and Grünewälder, S. (2012). Conditional Mean Embeddings as Regressors. In *International Conference on Machine Learing (ICML)*, volume 5.

Liu, S., Bousquet, O., and Chaudhuri, K. (2017). Approximation and Convergence Properties of Generative Adversarial Learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5545–5553.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. (2019). Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions. In *International Conference on Machine Learning (ICML)*, pages 4104–4113.

Luise, G., Pontil, M., and Ciliberto, C. (2020). Generalization Properties of Optimal Transport GANs with Latent Distribution Learning. *arXiv preprint arXiv:2007.14641*.

Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. (2018). Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5859–5870.

Luise, G., Salzo, S., Pontil, M., and Ciliberto, C. (2019). Sinkhorn Barycenters with Free Support via Frank-Wolfe Algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9322–9333.

Mémoli, F. (2011). Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487.

Mena, G. and Niles-Weed, J. (2019). Statistical Bounds for Entropic Optimal Transport: Sample Complexity and the Central Limit Theorem. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4541–4551.

Menon, M. V. (1967). Reduction of a matrix with positive elements to a doubly stochastic matrix. *Proc. Amer. Math. Soc.*, 18:244–247.

Mensch, A., Blondel, M., and Peyré, G. (2019). Geometric Losses for Distributional Learning. In *International Conference on Machine Learning (ICML)*, pages 4516–4525.

Mensch, A. and Peyré, G. (2020). Online Sinkhorn: Optimal Transportation Distances from Sample Streams. *arXiv preprint arXiv:2003.01415*.

Micchelli, C. A. and Pontil, M. (2005). On learning vector-valued functions. *Neural computation*, 17(1):177–204.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Moretti, V. (2013). *Spectral Theory and Quantum Mechanics: With an Introduction to the Algebraic Formulation*. UNITEXT. Springer Milan.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. (2018). Sobolev GAN. In *International Conference on Learning Representations (ICLR)*.

Mroueh, Y., Poggio, T., Rosasco, L., and Slotine, J.-J. (2012). Multiclass learning with simplex coding. In *Advances in Neural Information Processing Systems (NIPS) 25*, pages 2798–2806.

Mroueh, Y. and Sercu, T. (2017). Fisher GAN. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2513–2523.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. (2019). Asymptotic guarantees for Learning Generative Models with the Sliced-Wasserstein Distance. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 250–260.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in neural information processing systems (NIPS)*, pages 271–279.

Nussbaum, R. (1993). Entropy minimization, Hilbert's projective metric and scaling integral kernels. *Journal of Functional Analysis*, 115:45–99.

Pandeva, T. and Schubert, M. (2019). MMGAN: Generative Adversarial Networks for Multi-Modal Distributions. *arXiv preprint arXiv:1911.06663*.

Pedregosa, F. (2016). Hyperparameter optimization with approximate gradient. *arXiv preprint arXiv:1602.02355*.

Pele, O. and Werman, M. (2009). Fast and robust Earth Mover's Distances. In *International Conference on Computer Vision (ICCV)*, pages 460–467.

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pages 1679–1706.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein Barycenter and its Application to Texture Mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*, pages 435–446. Springer.

Ramdas, A., Trillos, N. G., and Cuturi, M. (2017). On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47.

Redko, I., Courty, N., Flamary, R., and Tuia, D. (2019). Optimal Transport for Multi-source Domain Adaptation under Target Shift. In *International Conference on Artificial Intelligence and Statistics (AIStats)*, pages 849–858.

Rockafellar, R. T. (1974). *Conjugate duality and optimization*. SIAM.

Rolet, A., Cuturi, M., and Peyré, G. (2016). Fast Dictionary Learning with a Smoothed Wasserstein Loss. In *International Conference on Artificial Intelligence and Statistics (AIStats)*, pages 630–638.

Rubner, Y., Guibas, L. J., and Tomasi, C. (1997). The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval. In *Proceedings of the ARPA Image Understanding Workshop*, volume 661, page 668.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International journal of computer vision*, 40(2):99–121.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, et al. (2016). Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2234–2242.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. (2018). Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. (2018). On the Convergence and Robustness of Training GANs with Regularized Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7091–7101.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4):928–943.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. (2018). Wasserstein Dictionary Learning: Optimal Transport-Based Unsupervised Nonlinear Dictionary Learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678.

Schrödinger, E. (1931). Über die Umkehrung der Naturgesetze. *Verlag der Akademie der Wissenschaften, in Kommission bei Walter de Gruyter*.

Schrödinger, E. (1932). Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2018). Large Scale Optimal Transport and Mapping Estimation. In *International Conference on Learning Representations (ICLR)*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shen, Z., Wang, Z., Ribeiro, A., and Hassani, H. (2020). Sinkhorn Barycenter via Functional Gradient Descent. *arXiv preprint arXiv:2007.10449*.

Sinkhorn, R. (1964). A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *Ann. Math. Statist.*, 35(2):876–879.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348.

Smale, S. and Zhou, D.-X. (2007). Learning Theory Estimates via Integral Operators and their Approximations. *Constructive approximation*, 26(2):153–172.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains. *ACM Transactions on Graphics (TOG)*, 34(4):66.

Solomon, J., Rustamov, R. M., Guibas, L., and Butscher, A. (2014). Wasserstein Propagation for Semi-supervised Learning. In *International Conference on Machine Learning (ICML)*.

Song, L. (2008). Learning via Hilbert space embedding of distributions. *Citeseer*.

Song, L., Fukumizu, K., and Gretton, A. (2013). Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models. *IEEE Signal Process. Mag.*, 30(4):98–111.

Song, L., Huang, J., Smola, A., and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning (ICML)*, pages 961–968.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density Estimation in Infinite Dimensional Exponential Families. *Journal of Machine Learning Research*, 18(57):1–59.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2009). On integral probability metrics, ϕ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(70):2389–2410.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11(50):1517–1561.

Srivastava, S., Li, C., and Dunson, D. B. (2018). Scalable Bayes via Barycenter in Wasserstein Space. *Journal of Machine Learning Research*, 19(8):1–35.

Staib, M., Claici, S., Solomon, J. M., and Jegelka, S. (2017). Parallel streaming Wasserstein barycenters. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2647–2658.

Stavropoulou, F. and Muller, J. (2015). Parametrization of Random Vectors in Polynomial Chaos Expansions via Optimal Transportation. *SIAM Journal on Scientific Computing*, 37(6):A2535–A2557.

Stein, E. M. (2016). *Singular integrals and differentiability properties of functions (PMS-30)*, volume 30. Princeton university press.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., and Okanohara, D. (2010). Least-Squares Conditional Density Estimation. *IEICE Transactions on Information and Systems*, 93(3):583–594.

Treves, F. (2016). *Topological Vector Spaces, Distributions and Kernels: Pure and Applied Mathematics*, volume 25. Elsevier.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. (2016). Conditional Image Generation with PixelCNN Decoders. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4790–4798.

Van Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel Recurrent Neural Networks. In *International Conference on Machine Learning (ICML)*, pages 1747–1756.

Varadarajan, V. S. (1958). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2019a). Optimal Transport for structured data with application on graphs. In *International Conference on Machine Learning (ICML)*.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. (2020). Fused Gromov-Wasserstein Distance for Structured Objects. *Algorithms*, 13(9):212.

Vayer, T., Flamary, R., Tavenard, R., Chapel, L., and Courty, N. (2019b). Sliced gromov-wasserstein. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Vilnis, L. and McCallum, A. (2015). Word Representations via Gaussian Embedding. In *International Conference on Learning Representations (ICLR)*.

Vondrick, C., Pirsiavash, H., and Torralba, A. (2016). Generating Videos with Scene Dynamics. In *Advances in Neural Information Processing Systems (NIPS)*, pages 613–621.

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

Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory (COLT)*, pages 3118–3119.

Wendland, H. (2004). *Scattered Data Approximation*, volume 17. Cambridge University press.

Weston, J., Chapelle, O., Elisseeff, A., Schölkopf, B., and Vapnik, V. (2003). Kernel Dependency Estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 873–880.

Wilson, A. G. (1969). The Use of Entropy Maximising Models, in the Theory of Trip Distribution, Mode Split and Route Split. *Journal of Transport Economics and Policy*, 3(1):108–126.

Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P., and Gool, L. V. (2019). Sliced Wasserstein Generative Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722.

Ye, J., Wu, P., Wang, J. Z., and Li, J. (2017). Fast Discrete Distribution Clustering using Wasserstein Barycenter with Sparse Support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332.

Yurinskii, V. (1976). Exponential Inequalities for Sums of Random Vectors. *Journal of multivariate analysis*, 6(4):473–499.

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. (2018). On the Discrimination-Generalization Tradeoff in GANs. In *International Conference on Learning Representations (ICLR)*.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *International Conference on Machine Learning (ICML)*, pages 912–919.