Title
An Assortment of Analyses of Optimal Transport Inspired by Domain Adaptation

Author
Pitcan, Yannik Kashif

Publication Date
2021

Peer reviewed|Thesis/dissertation

An Assortment of Analyses of Optimal Transport Inspired by Domain Adaptation

by

Yannik Pitcan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter Bartlett, Chair
Professor Steven Evans
Associate Professor Avi Feller

Summer 2021

An Assortment of Analyses of Optimal Transport Inspired by Domain Adaptation

Abstract

An Assortment of Analyses of Optimal Transport Inspired by Domain Adaptation

by

Yannik Pitcan

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter Bartlett, Chair

This dissertation consists of several papers. First, we start off introducing domain adaptation theory and briefly introduce optimal transport. Such an introduction allows the reader to understand why studying problems in optimal transport theory is so valuable.

Our first key result establishes bounds between regularized and unregularized optimal transport. Instead of using an entropic regularization, which is used in the Sinkhorn divergence, we regularize using dual potentials in a reproducing kernel Hilbert space. After this, we derive sample complexity bounds for the regularized optimal transport problem, and we show this is a substantial improvement over unregularized optimal transport. With these two results, one can approximate the theoretical optimal transport distance.

Next, we prove the first and second moments of the source and target distributions are enough to determine explicitly the optimal transport map and also that this is a linear mapping. Furthermore, we propose an alternative regularization for the transport map between two distributions.

After this, we briefly diverge from optimal transport theory and introduce work on prior elicitation. In particular, we extend a result from [39] on non-asymptotic bounds for maximum likelihood estimators to that for M-estimators. Crucially, we show sufficient assumptions for these to hold and use these to theoretically justify our prior elicitation objective.

Last, we return to optimal transport and introduce a variant to compare multiple probability measures, which we call sliced multi-marginal optimal transport. There, we propose a paradigm based on random one-dimensional projections.

This is dedicated to my parents.

Contents

List of Figures

Acknowledgments

Someone once said "failure is one's own doing, but success takes a village." I forget who said that, but it could not be more apt. There are multiple people in my life to whom I am incredibly appreciative.

First and foremost, I would like to thank Dr. Peter Bartlett. He has been an excellent advisor who suggested this topic and kept me motivated even when I thought all was lost. Without him, I do not know where I would be today. I not only consider him an advisor but now as a friend.

My other committee members, Steven Evans and Avi Feller, were also invaluable when I was seeking guidance and needed someone to discuss my work with. I can never repay them enough.

I would also like to thank my colleagues Soren Kuenzel, Devin Salle, Andre Waschka, and Alexander Tsigler. All three gave me many suggestions and tips that helped immensely. Their support when I was talking about research problems was invaluable.

Acknowledgements also go to Professor Sarah Brown for providing extra feedback on my research and Alexander Terenin and Samuel Cohen for being excellent collaborators.

From a personal standpoint, I'm grateful for the extra moral support from Jada Golden Sherman, whose pep talks were most appreciated when I was writing this dissertation. I also would like to acknowledge Shwin Ricci, Devin Saxon, Julia Sullivan, and Ashia Wilson for boosting my spirits at times.

I am thankful for my parents, Grace and Clyde Pitcan, for always believing in me.

And lastly, I am thankful for me for believing in myself.

Chapter 1

Introduction

First we give a primer on domain adaptation and then an introduction to optimal transport theory.

1.1 Motivation

In statistical learning theory, many results study the problem of estimating when a hypothesis from a select hypothesis class produces a low true risk. This is often expressed as a generalization bound on the true risk. The typical generalization problem assumes that the training and test distributions are identical.

One example of a case where training and test data differ is facial recognition, where an image classification model is learned on a community and is then used to classify those in another community who may have different facial features. The image recognition performance will deteriorate when the classification model does not account for the disparity between training and test distributions [41].

Another instance in which this assumption is violated is the spam filtering problem. A given user will be targeted with spam messages depending on his browsing history. If a working professional sets up his corporate mailbox on his home computer and transfers his settings, many personal emails he may want could be perceived as spam by an algorithm that learned preferences from professional communications. A classifier distinguishing spam from non-spam may not perform as well on another user if it does not adapt to different circumstances [41].

Such examples motivate the domain adaptation problem and extend traditional learning paradigms. For the rest of this dissertation, we investigate the scenario where a model may be learned on one distribution but evaluated on another.

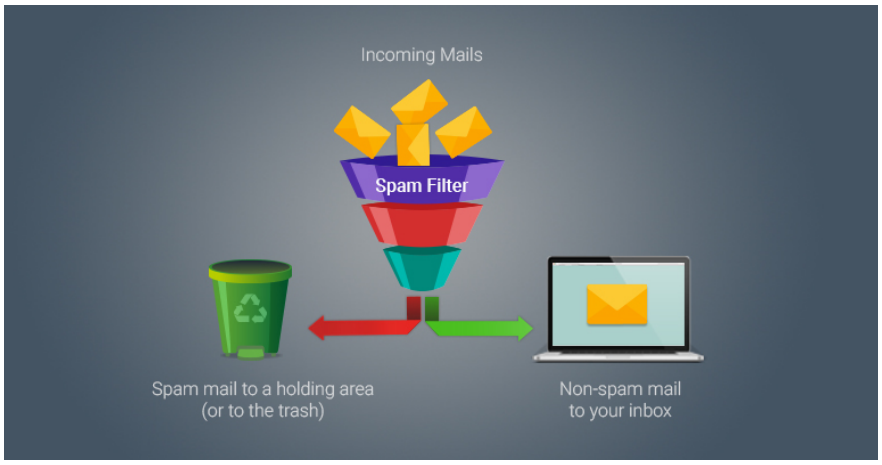


Figure 1.1: One application of transfer learning: spam filtering [41].

1.2 Background

For the applications considered above, the goal is to find a model that remains robust under changes in the environment. In other words, if a model is learned from the source, we want to measure how well it performs on the target domain. Formally, we describe this as follows.

Definition 1 (Transfer learning)**.** *Let S be a source data distribution called the source domain and T be a target data distribution called the target domain. Consider $X_S \times Y_S$ as the source input and output spaces and $X_T \times Y_T$ as target input and output spaces. Denote S_X and T_X to be the marginal distributions of X_S and X_T and by t_S and t_T the source and target learning tasks depending on Y_S and Y_T respectively. Tasks are defined as the combination of an input and an output we want to predict, and, for now, we will focus on classification tasks. We seek to improve the performance of $f_T : X_T \to Y_T$ for t_T using information gained from S where $S \neq T$.*

Transfer learning scenarios

Furthermore, we have the following types of learning:

- Inductive transfer learning. $X_S \overset{d}{=} X_T$ but $t_S \neq t_T$. Here, one can imagine t_S and t_T to be the tasks of detecting spam and hoaxes respectively for the same user.
- Transductive transfer learning. $X_S \neq X_T$ but $t_S = t_T$. An example of a task here is spam detection for two different users.

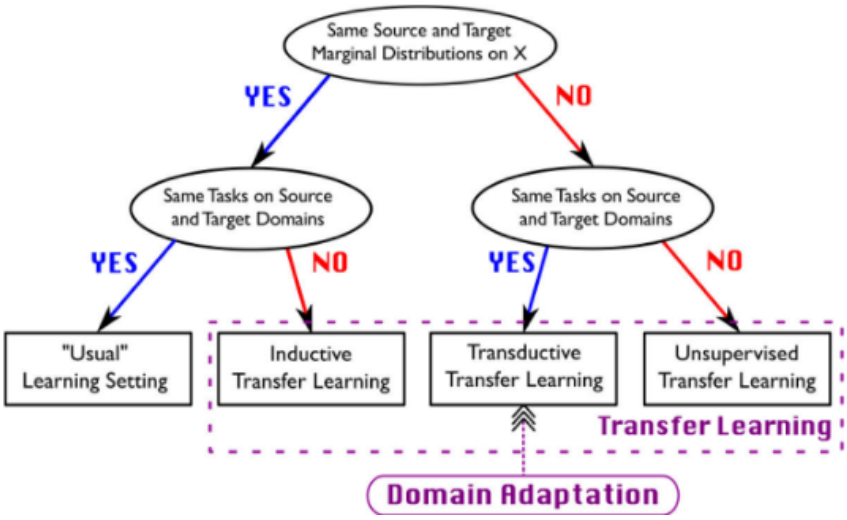


Figure 1.2: Positioning of Domain Adaptation compared to other learning techniques [41]

- Unsupervised transfer learning. $t_S \neq t_T$ and $X_S \neq X_T$. To make this more intuitive, one can think of collecting web data for a user and pages taken from the web. For the first user, one detects spam, and for the second, one detects hoaxes.

The category we focus on is transductive transfer learning, which we hereafter call domain adaptation.

From a probabilistic point of view, we can categorize our problem via the causal link between labels and instances.

- $X \to Y$ problems where the class label is causally determined by instance values. This labeling comes up in image classification where the object description determines the label. The joint distribution can be decomposed into $P(X, Y) = P(X)P(Y|X)$.
- $Y \to X$ where this is the reverse. Class labels causally determine instance values. A good example here is in medical applications where we observe disease symptoms but want to predict the disease [41]. The joint decomposition here is $P(X, Y) = P(Y)P(X|Y)$.

It follows that we can categorize different types of transfer learning scenarios based on the probabilistic point of view. The following are some such scenarios:

- Covariate-shift $P(X_S) \neq P(X_T)$ but $P(Y_T|X_T) = P(Y_S|X_S)$

  This is a case of the $X \to Y$ problem where $X_S \not\equiv X_T$ while $Y_S|X_S \equiv Y_T|X_T$. Here, the marginal distributions between the source and target are different while the predictive

behavior stays the same. One example of this is the Office/Caltech dataset [27] with domains:

1. Amazon images from online merchants.
2. Low-quality webcam images.
3. High-quality images taken with a DSLR.
4. Images from Caltech dataset for object recognition.

Solving the covariate shift problem involves a reweighting as seen by the following:

$$\begin{aligned} R_T^\ell(h) &= \mathbf{E}_{(x,y)\sim T}(\ell(h(x),y)) \\ &= \mathbf{E}_{x\sim T(x)}\mathbf{E}_{T(y|x)}(\ell(h(x),y)) \\ &= \int \frac{dT(x)}{dS(x)} dS(x)\mathbf{E}_{T(y|x)}\ell(h(x),y) \\ &= \int \frac{dT(x)}{dS(x)} dS(x)\mathbf{E}_{S(y|x)}\ell(h(x),y) \\ &= \mathbf{E}_{x\sim S(x)}\mathbf{E}_{S(y|x)}[\beta(x)\ell(h(x),y)] \\ &= \mathbf{E}_{(x,y)\sim S}[\beta(x)\ell(h(x),y)] \end{aligned}$$

where $\beta(x) := \frac{dT(x)}{dS(x)}$, the Radon-Nikodym derivative of the target distribution with respect to the source distribution.

- Target-shift $P(X_T|Y_T) \neq P(X_S|Y_S)$

  These occur in $Y \to X$ problems. In this case, $Y_S \not\equiv Y_T$–the target distributions are different. Generally, this occurs when different sampling methods are used for the source and target datasets.

- Concept shift $P(X_T, Y_T) \neq P(X_S, Y_S)$ This occurs both in $X \to Y$ and $Y \to X$ problems when $P(Y_S|X_S) \neq P(Y_T|X_T)$ and $P(X_S|Y_S) \neq P(X_T|Y_T)$ respectively.

- Sample-selection bias

  Here, the source and target distributions differ because of a latent variable that excludes some sample observations conditional on their labeling or nature. For example, if we are classifying images of people, we may discard images that are unclear. This exclusion leads to a sample-selection bias, since some devices may take less clear pictures by default.

- Ideal joint error. We may claim the existence of a low-error hypothesis for both the source and target domain. Usually, this is characterized by

$$\lambda_{\mathcal{H}} = \min_{h\in\mathcal{H}} R_S(h) + R_T(h)$$

where $R_S(h)$ and $R_T(h)$ are the true risks over the source and target domains S and T.

Definition 2 (True Risk)**.** *Given a loss function $l : Y \times Y \to [0, 1]$, the true risk or generalization error $R^l_D(h)$ for a hypothesis $h \in \mathcal{H}$ on a distribution D over $X \times Y$ is defined as*

$$R^l_D(h) = \mathrm{E}_{(x,y)\sim D} l(h(x), y).$$

and for a given pair of hypotheses $(h, h') \in \mathcal{H}^2$,

$$R^l_D(h, h') = \mathrm{E}_{(x,y)\sim D} l(h(x), h'(x)).$$

As a side-note, there are three predominant algorithmic techniques used for domain adaptation. They are

1. Reweighting the source labeled examples to be more similar to the target examples. This is done in cases such as covariate shift.

2. Iteratively “auto-labeling" target examples. Here, a model is learned from source labeled examples and then automatically labels some target examples. We then learn a new model from the new labeled examples.

3. Finding a common representation space. In this situation, we find a space where the source and target domains are close while maintaining a good performance on the source domain task.

Divergence between domains

In domain adaptation, we must quantify the difference between source and target domains. Unlike classical supervised learning, transfer learning involves a discrepancy between the two domains. There are many metrics, such as Hellinger distance, total-variation distance, Renyi divergence, or Wasserstein metric, that exist to measure such a discrepancy [41].

Often, one wants to prove that a divergence measure can relate errors between source and target domains. This relation means we can establish error guarantees by minimizing the divergence between the source and target distributions.

Along with analyzing existing divergence measures, one may also design a new divergence measure suitable for domain adaptation. Additionally, we investigate a new specific divergence measure in Chapter 3.

In the subsequent paragraphs, we discuss seminal work in this area of research. We do this to better demonstrate what we mean by relating errors between domains with respect to a divergence measure.

A First Theoretical Analysis

From a theory perspective, the seminal work was done by Ben-David et al. In their work, they considered a binary loss function in a binary classification setting and the L^1-distance [6].

First, we provide some definitions.

Definition 3 (Rademacher complexity)**.** *Denote $\sigma_1, \sigma_2, \ldots, \sigma_m$ as independent random variables drawn from the Rademacher distribution i.e. $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$ for $i = 1, 2, \ldots, m$.*

Given a sample $S = (z_1, z_2, \ldots, z_m) \in Z^m$, and a class F of real-valued functions defined on a domain space Z,

$$\mathrm{Rad}_S(F) = \frac{1}{m} \mathrm{E}_\sigma \left[\sup_{f \in F} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

Definition 4. *Shattering*

A family H shatters a set $S \subseteq \mathcal{X}$ if for every subset $T \subseteq S$ there exists a function $h \in H$ such that $h(s) = 1_{s \in T}$ for all $s \in S$, that is, $h(s) = 1$ if $s \in T$ and $h(s) = 0$ if $s \in S \setminus T$.

Intuitively, we say that H shatters some set $S \subseteq \mathcal{X}$ if we can realize any labelings on S using functions from H.

Definition 5. *VC Dimension*

The VC dimension of a set of hypothesis functions H is the cardinality of the largest set which H can shatter.

Definition 6. *$\mathcal{H}$-divergence*

Denote $\mathcal{A}$ the set of measurable subsets under two probability distributions $\mathcal{D}$ and $\mathcal{D}'$. Then the $\mathcal{H}$-divergence is defined as

$$d_1(\mathcal{D}, \mathcal{D}') = 2 \sup_{A \in \mathcal{A}} \left| P_D(A) - P_{D'}(A) \right|.$$

This one compares how two classifiers disagree on both domains. Here, it finds the pair of classifiers with the largest disparity in disagreements between the source and target domains.

Using this notion of distance, Ben-David et al. derived the first generalization bounds.

Definition 7 (Labeling function)**.** *A labeling function $f : X \to Y$ is a mapping of feature input X to a class label Y. In the next theorem, $Y = \{0, 1\}$.*

Theorem 1.2.1. *Generalization bound with respect to $\mathcal{H}$-divergence [6]*

Let l represent the $0 - 1$ loss function and f_S, f_T the source and target true labeling functions respectively.

$$R_T^l(h) \leq R_S^l(h) + d_1(X_S, X_T) + \min \left\{ \mathrm{E}_{x \sim X_S}[\|f_S(x) - f_T(x)\|], \mathrm{E}_{x \sim X_T}[\|f_S(x) - f_T(x)\|] \right\}$$

This was the first theoretical generalization bound, and it had some flaws. In practice, one may want to obtain finite-sample estimates, but that isn't possible with $\mathcal{H}$-divergence. Also, the $\mathcal{H}$-divergence does not incorporate the hypothesis class. Both of these issues are resolved with the introduction of another type of divergence: the symmetric difference hypothesis divergence.

Definition 8 (Symmetric Difference Hypothesis Divergence)**.**

$$D_{\mathcal{H}\Delta\mathcal{H}}(S,T) = 2 \sup_{h,h' \in \mathcal{H}} |P_S[h(x) \neq h'(x)] - P_T[h(x) \neq h'(x)]|$$

Definition 9 (Empirical Symmetric Difference Hypothesis Divergence)**.**

$$\hat{D}_{\mathcal{H}\Delta\mathcal{H}}(\hat{S},\hat{T}) = 2 \sup_{h,h' \in \mathcal{H}} \left| \sum_{i=1}^{m} I[h(x_i^s) \neq h'(x_i^s)] - \sum_{i=1}^{m} I[h(x_i^t) \neq h'(x_i^t)] \right|.$$

Theorem 1.2.2. *Here, $\hat{S}, \hat{T}$ are independent size-m samples drawn from S and T respectively. For $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$D_{\mathcal{H}\Delta\mathcal{H}}(S,T) \leq \hat{D}_{\mathcal{H}\Delta\mathcal{H}}(\hat{S},\hat{T}) + 4\sqrt{\frac{2VC(\mathcal{H})\log(2m) + \log(2/\delta)}{m}}$$

The above tells us that, for a finite VC dimension class $\mathcal{H}$, the empirical $\mathcal{H}\Delta\mathcal{H}$ divergence is a good estimate for its true variant.

Furthermore, one can compute the empirical divergence. Ben-David et al then obtained a bound for risk on the target domain that involved the empirical divergence.

We first define empirical risk. Unlike true risk, which defines the theoretical error over a distribution, empirical risk represents the error of a hypothesis on an observed training sample.

Definition 10 (Empirical Risk)**.** *Given a loss function $l : Y \times Y \to [0,1]$ and a training sample $S = (x_i, y_i)_{i=1}^m$ where each sample is drawn i.i.d. from D, the empirical risk $R^l_{\hat{D}}(h)$ for a hypothesis $h \in \mathcal{H}$ is*

$$R^l_{\hat{D}}(h) = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i),$$

where $\hat{D}$ is the empirical distribution associated with S.

Theorem 1.2.3. *Let $\lambda^* = \min_{h \in \mathcal{H}} R_S(h) + R_T(h)$ be the minimum joint risk. With probability at least $1-\delta$:*

$$R^l_T(h) \leq \hat{R}^l_S(h) + \frac{1}{2}D_{\mathcal{H}\Delta\mathcal{H}}(\hat{S},\hat{T}) + \lambda^* + O\left(\sqrt{\frac{VC(\mathcal{H})\log(m) + \log(2/\delta)}{m}}\right)$$

Of note is that the risk bound presented is only relevant if the optimal joint risk is controlled.

Critique of $\mathcal{H}\Delta\mathcal{H}$-divergence

A flaw of the $\mathcal{H}\Delta\mathcal{H}$-divergence is that it relies on a specific loss function (0-1 loss). Contrarily, one may want to work with a more general loss function. This desire motivated other work [29] to use Renyi and $\mathcal{Y}$-discrepancy distances.

Definition 11 (Renyi Divergence)**.** *Let p and q be two probability densities.*

$$D_\alpha(p,q) = \frac{1}{\alpha-1}\log_2 \int_{\mathcal{X}} p^\alpha(x)/q^{\alpha-1}(x)\,dx,$$

where α denotes its order.

Definition 12 (Expected Loss)**.** *Let $f, g : X \to Y$ be labeling functions and $L : Y \times Y \to \mathbb{R}$ be a loss function. Then for any distribution D, our expected loss is*

$$\mathcal{L}_D(f,g) = \mathrm{E}_{x\sim D}[L(f(x), g(x))].$$

Definition 13. *$\mathcal{Y}$-Discrepancy*

Let f_P and f_Q be the labeling functions on P and Q. Then the $\mathcal{Y}$-discrepancy between domains (P, f_P) and (Q, f_Q) is

$$disc_{\mathcal{Y}}(P,Q) = \sup_{h\in H}|\mathcal{L}_Q(h, f_Q) - \mathcal{L}_P(h, f_P)|$$

In the majority of this dissertation, we study divergences inspired by optimal transport theory. This brings us to the next section, which introduces some of the foundational material on Wasserstein spaces.

1.3 Brief Introduction to Optimal Transport

Monge Problem

In 1781, Gaspard Monge asked how one can transport a pile of sand into a pit when both have equal volumes.

Intuitively, the goal is to minimize the expected “cost" of moving the sand, and it turns out this has a mathematical formulation as follows:

Let X be the space of sand, Y be the space for the pit, and define a cost function $c : X \times Y \to \mathbb{R}$ that demonstrates the cost of moving a unit of sand $x \in X$ to a pit location $y \in Y$. And denote μ to be the distribution on X

The choice of where to place a unit of sand can be represented as the function $T : X \to Y$, which has a total transport cost of

$$\int_X c(x, T(x))\,d\mu(x).$$

Moreover, the function T must satisfy a mass-preservation requirement: the volume $\nu(B)$ of any region in the pit $B \subseteq Y$ must be the same as the volume of the sand moved into B.

Formally, we can write this as

$$\mu(T^{-1}(B)) = \nu(B) \text{ for all } B \subseteq Y$$

which we denote $T\#\mu = \nu$. We say that ν is the push-forward measure of μ under T.

If c and T are measurable, and $\mu(T^{-1}(B)) = \nu(B)$ for all measurable subsets B of Y, then T is called a transport map. Normalizing μ and ν to be probability measures, the Monge problem finds the optimal transport map minimizing transport costs [34].

Definition 14. *Monge Problem*

Let $T : X \to Y$ be a transport map with an associated total cost

$$C(T) = \int_X c(x, T(x))\, d\mu(x),$$

where μ and ν are again the probability measures assigned to X and Y.

The Monge problem finds

$$\inf_{T:T\#\mu=\nu} C(T).$$

The Monge problem is very hard because the set of transport maps $\{T : T\#\mu = \nu\}$ is intractable to work with. If $\mu = \delta\{x_0\}$ is a Dirac measure and ν is not, then no transport maps exist.

But what if we can split the mass of sand particles? That is to say, what if we don't have the strict conditions as above. This brings us to the Kantorovich relaxation.

Kantorovich Relaxation

For each point $x \in X$, a probability measure μ_x defines how the mass at x is split. If $\mu_x = \delta\{y\}$ for $y \in Y$, then all the mass at x is sent to y.

Represent π to be the joint probability measure on $X \times Y$, where $\pi(A \times B)$ is the amount of sand moved from $A \subseteq X$ to $B \subseteq Y$. The total mass sent from A is $\pi(A \times Y)$ and the total moved into B is $\pi(X \times B)$. Such a measure π is called a transference plan when

$$\pi(A \times Y) = \mu(A), \quad A \subseteq X$$
$$\pi(X \times B) = \nu(B), \quad B \subseteq Y$$

where A and B are Borel sets. The set of transference plans is denoted $\Pi(\mu, \nu)$.

Definition 15. *Kantorovich Problem*

Let $\pi \in \Pi(\mu, \nu)$ be a transference plan with an associated total cost

$$C(\pi) = \int_{X \times Y} c(x, y)\, d\pi(x, y).$$

The Kantorovich problem solves for the optimal plan given by

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi).$$

Probabilistic Interpretations of Monge and Kantorovich Problems

We can view the above optimization problems from a probabilistic perspective. The Monge solution minimizes $\mathrm{E}_X[c(X, T(X))]$ over T (measurable) whereas the Kantorovich solution minimizes $\mathrm{E}_{\pi \in \Pi(\mu,\nu)}[c(X, Y)]$. We call $\pi \in \Pi(\mu, \nu)$ a coupling between μ and ν.

A Divergence Measure Inspired by Optimal Transport

If $X = Y$, then we can define a distance between measures μ and ν using a special cost function c.

Let $c(x_1, x_2) = [d(x_1, x_2)]^p$, where $d(x_1, x_2)$ denotes the distance between x_1 and x_2 and p is a real-valued constant.

Definition 16. *Wasserstein Distance of Order p*

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x_1, x_2)^p \, d\pi(x_1, x_2) \right)^{1/p} = \left(\inf_{\pi \in \Pi(\mu,\nu)} \mathrm{E}_\pi[d(x_1, x_2)^p] \right)^{1/p}.$$

With this being said, let's begin.

Chapter 2

Theoretical Analysis of Domain Adaptation with Optimal Transport

Previously, we gave a glimpse at domain adaptation and briefly discussed optimal transport (OT), without tying the two concepts together. Thus, we give a more in-depth introduction here to how optimal transport and domain adaptation are linked from a theoretical viewpoint. Specifically, we explicitly show how generalization bounds for domain adaptation may involve optimal transport quantities. This subsequently inspires our analysis of OT.

Why Use Optimal Transport in Domain Adaptation?

Optimal transport is capable of taking into consideration the geometry of the data. In domain adaptation problems, this is helpful, especially since when dealing with a source and target distribution, a natural idea is to look for a nonlinear transformation between the two distributions. This makes optimal transport distances (e.g. Wasserstein) highly promising. Another concern is that the source and target distributions lack a shared support. Using a distance that does not require a shared support makes sense, and the Wasserstein is one such distance. This property distinguishes it from other divergences, such as Maximum Mean Discrepancy or Kullback-Leibler, which usually require a common support.

2.1 Notation and Preliminaries

Definition 17. *Reproducing Kernel Hilbert Space*

Let X be an arbitrary set and H a Hilbert space of real-valued functions on X. The evaluation functional over the Hilbert space of functions H is a linear functional that evaluates each function at a point x,

$$L_x : f \mapsto f(x) \; \forall f \in H.$$

We say that H is a reproducing kernel Hilbert space if, for all $x \in X$, L_x is continuous at any $f \in H$ or, equivalently, if L_x is a bounded operator on H, i.e. there exists some $M > 0$ such that

$$|L_x(f)| := |f(x)| \leq M\|f\|_H \quad \forall f \in H.$$

Definition 18. *Kullback-Leibler Divergence If P and Q are probability measures on a set $\mathcal{X}$, and P is absolutely continuous with respect to Q, then the Kullback-Liebler divergence from Q to P is*

$$D_{KL}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP.$$

Definition 19. *Maximum mean discrepancy*

MMD represents distances between distributions as distances between mean embeddings of features. If we have distributions p and q over a set $\mathcal{X}$, the MMD is defined by a feature map $\varphi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a reproducing kernel Hilbert space with φ the reproducing kernel.

$$MMD(p, q) = \|\mathrm{E}_{X\sim p}[\varphi(X)] - \mathrm{E}_{Y\sim q}[\varphi(Y)]\|_{\mathcal{H}}.$$

We can alternatively characterize the MMD as follows:

$$\begin{aligned} MMD(p, q) &= \|\mathrm{E}_{X\sim p}[\varphi(X)] - \mathrm{E}_{Y\sim q}[\varphi(Y)]\|_{\mathcal{H}} \\ &= \sup_{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\leq 1} \langle f, \mathrm{E}_{X\sim p}[\varphi(X)] - \mathrm{E}_{Y\sim q}[\varphi(Y)]\rangle_{\mathcal{H}} \\ &= \sup_{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\leq 1} \left[\langle f, \mathrm{E}_{X\sim p}[\varphi(X)]\rangle_{\mathcal{H}} - \langle f, \mathrm{E}_{Y\sim q}[\varphi(Y)]\rangle_{\mathcal{H}}\right] \\ &= \sup_{f\in\mathcal{H}:\|f\|_{\mathcal{H}}\leq 1} [\mathrm{E}_{X\sim p}[f(X)] - \mathrm{E}_{Y\sim q}[f(Y)]] \end{aligned}$$

The alternative characterization holds because of the reproducing property: $\langle f, \varphi(x)\rangle_{\mathcal{H}} = f(x)$ for any $f \in \mathcal{H}$. The second line holds since $\sup_{f:\|f\|\leq 1}\langle f, g\rangle_{\mathcal{H}} = \|g\|$ is attained when $f = g/\|g\|$. The fourth relies on Bochner integrability, but assuming our kernel or distributional support is bounded, this is true.

The following extension of Wasserstein to empirical measures will be used when contrasting empirical to theoretical distances.

Definition 20. *Discrete Wasserstein [41]*

If we deal with empirical measures $\hat{\mu}_S = \frac{1}{N_S}\sum_{i=1}^{N_S} \delta_{x_s^i}$ and $\hat{\mu}_T = \frac{1}{N_T}\sum_{i=1}^{N_T} \delta_{x_T^i}$ represented by the uniformly weighted sums of N_S and N_T Diracs with mass at locations x_S^i and x_T^i respectively, then the Kantorovich problem is defined in terms of the inner product between the coupling matrix γ and the cost matrix C:

$$W_1(\hat{\mu}_S, \hat{\mu}_T) = \min_{\gamma\in\Pi(\hat{\mu}_s,\hat{\mu}_T)} \langle C,\ \gamma\rangle_F$$

where $\langle\ ,\ \rangle_F$ *denotes the Frobenius inner product,* $\Pi(\hat{\mu}_s, \hat{\mu}_T) = \{\gamma \in \mathbb{R}_+^{N_S \times N_T} | \gamma 1 = \hat{\mu}_S, \gamma^T 1 = \hat{\mu}_T\}$ *is a set of doubly stochastic matrices and* C *is a dissimilarity matrix, i.e.,* $C_{ij} = c(x_S^i,\ x_T^j)$*, defining the cost needed to move a probability mass from* x_S^i *to* x_T^j*.*

Definition 21 (Expected Loss)**.** *Let* l *be a convex loss-function. Given a distribution* μ_D*, a hypothesis* $h \in H$ *and a labeling function* f_D *(which may be a hypothesis in* H*), the expected loss is defined as*

$$R_D^l(h, f_D) = \mathrm{E}_{X \sim \mu_D}[l(h(x), f_D(x))].$$

Our source and target spaces are denoted by S and T respectively. S has a distribution μ_S and T has as its underlying distribution, μ_T. Our loss function is denoted by $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.

2.2 Prior Work

First, we introduce some past results pertaining to risk bounds with respect to the Wasserstein distance. The purpose of this is to show the explicit connections between optimal transport work and domain adaptation theory and also serve as inspiration for the research presented in the next several chapters.

First Theoretical Bounds

Lemma 2.2.1. *This lemma and theorem are from [41].*

Let μ_S, $\mu_T \in \mathcal{P}(\mathbf{X})$ *be two probability measures on* $\mathbb{R}^d$*. And assume the following hold.*

- *The cost function* $c(\mathbf{x}, \mathbf{x}') = \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_{\mathcal{H}_{k_\ell}}$*, where* $\mathcal{H}$ *is a reproducing kernel hilbert space (RKHS) 17 equipped with kernel* $k_\ell : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ *induced by* $\phi : \mathbf{X} \to \mathcal{H}_{k_\ell}$ *and* $k_\ell(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}_{k_\ell}}$*.*

  And let us briefly discuss the use of the aforementioned cost function. It can be seen that:

$$\begin{aligned} c(x, x') &= \|\phi(x) - \phi(x')\|_{\mathcal{H}_{k_l}} \\ &= \sqrt{\langle \phi(x) - \phi(x'), \phi(x) - \phi(x') \rangle_{\mathcal{H}_{k_l}}} \\ &= \sqrt{k_l(x, x) - 2k_l(x, x') + k_l(x', x')}. \end{aligned}$$

  There also exists a one-to-one relationship between the choice of a positive-definite kernel k *and the cost function* c *as proven in [44].*

- *The loss function* $\ell_{h,f} : \mathbf{x} \longrightarrow \ell(h(\mathbf{x}), f(\mathbf{x}))$ *is convex, symmetric, bounded, obeys triangle equality, and has the parametric form* $|h(\mathbf{x}) - f(\mathbf{x})|^q$ *for some* $q > 0$*. The loss function* ℓ *is a nonlinear mapping of* $\mathcal{H}_k$ *for the family of* ℓ_q *losses.*

  $\ell_{h,f}$ *also belongs to the RKHS* $\mathcal{H}_{k_\ell}$ *admitting the reproducing kernel* k_ℓ *and that its norm obeys the following inequality:*

$$\|\ell_{h,f}\|^2_{\mathcal{H}_{k_\ell}} \leq \|h - f\|^{2q}_{\mathcal{H}_k}.$$

- *The kernel* k_ℓ *in the RKHS* $\mathcal{H}_{k_\ell}$ *is square-root integrable w.r.t. both* μ_S, μ_T *for all* μ_S, $\mu_T \in \mathcal{P}(\mathbf{X})$ *where* $\mathbf{X}$ *is separable and* $0 \leq k_\ell(\mathbf{x}, \mathbf{x}') \leq K, \forall \mathbf{x}, \mathbf{x}' \in \mathbf{X}$.

- $\|\ell\|_{\mathcal{H}_{k_\ell}} \leq 1$.

Then the following bound holds

$$\forall (h, f) \in \mathcal{H}^2_{k_\ell}, \quad \mathrm{R}^{\ell_q}_{\mathcal{T}}(h, f) \leq \mathrm{R}^{\ell_q}_{S}(h, f) + W_1(\mu_S, \mu_T).$$

With the use of a concentration inequality on Wasserstein distances [10], empirical risk bounds are obtained. The theorem is provided below for convenience.

Theorem 2.2.2. *Let* μ *be a probability measure on* $\mathbb{R}^d$ *such that*

$$\exists \alpha > 0, \int_{\mathbb{R}^d} e^{\alpha \|x\|^2} d\mu(x) < \infty$$

and let $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}$ *be the empirical measure on* $\{x_i\}$*. The* $\{x_i\}$ *are sampled from* μ*.*

Then for all $d' > d$ *and all* $\xi < \sqrt{2}$*, there exists* $N_0(d')$ *and* $\alpha > 0$ *with*

$$\int e^{\alpha c(x, x')} d\mu(x) < \infty$$

for a fixed x' *such that for all* $\epsilon > 0$ *and all* $N \geq N_0 \max\{\epsilon^{-(d'+2)}, 1\}$*,*

$$\Pr[W_1(\mu, \hat{\mu}) > \epsilon] \leq e^{-\frac{\xi}{2} N \epsilon^2}.$$

Theorem 2.2.3. *[41] Under the assumptions of 2.2.1, let* S_u *and* T_u *be two samples of size* N_S *and* N_T *drawn i.i.d. from* μ_S *and* μ_T*, respectively. Let* $\hat{\mu_S} = \frac{1}{N_S} \sum_{i=1}^{N_S} \delta_{\mathbf{x}_i^s}$ *and* $\hat{\mu_T} = \frac{1}{N_T} \sum_{i=1}^{N_T} \delta_{\mathbf{x}_i^T}$ *be the associated empirical measures. Then for any* $d' > d$ *and* $\varsigma' < \sqrt{2}$*, there exists some constant* N_0 *depending on* d'*, such that for any* $\delta > 0$ *and* $\min(N_S, N_T) \geq N_0 \max\left(\delta^{-(d'+2)}, 1\right)$ *with probability of at least* $1 - \delta$ *for all* h*, we have*

$$\mathrm{R}^{\ell_q}_{\mathcal{T}}(h) \leq \mathrm{R}^{\ell_q}_{\mathcal{S}}(h) + W_1(\hat{\mu_S}, \hat{\mu_T}) + \sqrt{2 \log\left(\frac{1}{\delta}\right) / \varsigma'} \left(\sqrt{\frac{1}{N_S}} + \sqrt{\frac{1}{N_T}}\right) + \lambda$$

where λ *is the combined error of the ideal hypothesis* h^* *that minimizes the combined error of* $\mathrm{R}^{\ell_q}_{\mathcal{S}}(h) + \mathrm{R}^{\ell_q}_{\mathcal{T}}(h)$*.*

In this chapter, we introduced optimal transport in the context of domain adaptation and discussed how these calculations appear in theoretical and empirical risk bounds. This serves to show the significance of studying the statistical properties of optimal transportation, which is the motivation for the next three chapters in this dissertation.

Chapter 3

Regularized Optimal Transport for Domain Adaptation Problems

3.1 Introduction

Previously, we looked at unregularized optimal transport for domain adaptation and some statistical properties – in particular, how it may be used in theoretical generalization bounds. However, a fundamental limitation of the optimal transportation methods for domain adaptation is that they turn out to be computationally difficult. For example, in the discrete optimal transport problem, assuming P and Q were of size n, algorithms such as the simplex algorithm and Hungarian algorithm have a complexity of at most $O(n^3 \log(n))$.

In this chapter, we examine regularization methods for optimal transport that statistically outperform that of unregularized optimal transport. This then leads us to introduce an alternative concept of regularization involving dual potentials that may be more suitable for domain adaptation problems. In particular, we study briefly the behavior of joint couplings in a simpler setting and then introduce our main result of this chapter, which are sample complexity bounds for this alternative regularization. And lastly, we demonstrate that, under certain assumptions, it also has statistical benefits over unregularized OT.

3.2 Entropic Regularization

One popular method of addressing the aforementioned limitation of OT is by using an entropic regularization as seen in [16].

Definition 22 (Entropic Regularization of Wasserstein)**.**

$$W_\epsilon(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon H(\pi | P \otimes Q)$$

where

$$H(\pi|P \otimes Q) = \int_{\mathcal{X}\times\mathcal{Y}} \log(\frac{d\pi(x,y)}{dP(x)dQ(y)})d\pi(x,y).$$

If we use the relative entropy as a regularizer, then we can formulate the dual of regularized OT as the maximization of an expectation problem [23].

$$\begin{aligned} W_\epsilon(P,Q) = & \sup_{u\in L_1(P), v\in L_1(Q)} \int_{\mathcal{X}} u(x)\mathrm{d}P(x) + \int_y v(y)\mathrm{d}Q(y) \\ & - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\epsilon}} \mathrm{d}P(x)\mathrm{d}Q(y) + \varepsilon \\ = & \sup_{u\in L_1(P), v\in L_1(Q)} \mathrm{E}_{P\otimes Q}[f_\epsilon^{XY}(u,\ v)] + \epsilon \end{aligned}$$

where $f_\epsilon^{XY}(u,\ v) = u(x) + v(y) - \epsilon e^{\frac{u(x)+v(y)-\mathrm{c}(x,y)}{\epsilon}}$.

Results in [23] demonstrated the superiority of entropic regularized OT over unregularized OT in terms of sample complexity. In the case where $c(x,y) = \frac{1}{2}\|x-y\|^2$, they showed the following:

Theorem 3.2.1. *Let P and Q be two probability measures on a bounded domain in $\mathbb{R}^d$ of diameter D. Then*

$$\sup_{P,Q} E_{P,Q}\left|W_\epsilon(P,Q) - W_\epsilon\left(P_n,Q_n\right)\right| \leq K_{D,d}\left(1 + \frac{1}{\epsilon^{\lfloor d/2\rfloor}}\right)\frac{e^{D^2/\epsilon}}{\sqrt{n}}$$

where $K_{D,d}$ is a constant depending on D and d.

3.3 Downside of Entropic Regularization

Although entropic regularized OT has benefits over unregularized OT as demonstrated in 3.2.1, it may not be the best regularization for domain adaptation work. When we prescribe a divergence measure for domain adaptation, we seek to penalize when the source and target distributions are not identical. However, with entropic regularization, we penalize when the source and target distributions are not independent. But if one can look at a "mapping" between the two distributions, it makes sense to regularize with respect to the deviation between our "mapping" and an identity map. What we eventually demonstrate is that we can perform such a regularization by looking at the dual potentials.

In order to discuss this more rigorously, first we give a more precise description of a mapping between two probability measures.

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 $\mathcal{X}$,

$$\left(T_\#\eta\right)(B) = \eta\left(T^{-1}(B)\right).$$

From here on, we will refer to a measure $T_{\#}\eta$ as pushforward measure, and to the corresponding T as pushforward map.

Lemma 3.3.1. *For any measurable $f : \mathcal{X} \to \mathbb{R}$*

$$\int_{\mathcal{X}} f(x) d\left(T_{\#}\eta\right)(x) = \int_{\mathcal{Z}} f(T(z)) d\eta(z)$$

3.4 Existence and Uniqueness of an Optimal Transport Map

The intuition behind our proposed regularization stems from Brenier's theorem, which concerns the existence and uniqueness of optimal maps.

Theorem 3.4.1 (Brenier's Theorem)**.** *Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ with $X, Y \subset \mathbb{R}^n$ and assume that μ, ν both have finite second moments and that μ is absolutely continuous. Then there is a unique solution $\pi^\dagger \in \Pi(\mu, \nu)$ to the Kantorovich optimal transport problem with cost $c(x, y) = \frac{1}{2}|x - y|^2$ which is given by*

$$\pi^\dagger = (\mathrm{Id} \times \nabla\varphi)_{\#}\mu \quad \text{or equivalently} \quad \mathrm{d}\pi^\dagger(x, y) = \mathrm{d}\mu(x)\delta(y = \nabla\varphi(x))$$

where $\nabla\varphi$ is the gradient of a convex function (defined μ -almost everywhere) that pushes μ forward to ν, i.e. $(\nabla\varphi)_{\#}\mu = \nu$.

Any convex function is locally Lipschitz on the interior of its domain and u is almost everywhere differentiable on the interior of its domain.

When $c(\boldsymbol{x}, \boldsymbol{y}) = 1/2\|\boldsymbol{x} - \boldsymbol{y}\|^2$, we have an explicit representation for T, the transport map.

$$T(\boldsymbol{x}) = \boldsymbol{x} - \nabla u(\boldsymbol{x}) = \nabla\left[\frac{1}{2}\|\boldsymbol{x}\|^2 - u(\boldsymbol{x})\right] = \nabla\varphi(\boldsymbol{x}).$$

In this case, the Brenier's potential φ and the Kantorovich's potential u are related by the following: $\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2 - u(\boldsymbol{x})$

Corollary 3.4.1.1. *Under the assumptions of Theorem 3.4.1, ∇u is the unique solution to the Monge transportation problem:*

$$\frac{1}{2}\int_X |x - \nabla\varphi(x)|^2 \, \mathrm{d}\mu(x) = \frac{1}{2}\inf_{T_{\#}\mu=\nu} \int_X |x - T(x)|^2 \, \mathrm{d}\mu(x).$$

This motivates our proposed regularization as now we have a clear relationship between potentials and the optimal transport map. Since we established a relationship between those and the Kantorovich potentials and the latter potentials are easier to work with as they come up in the dual form of the optimal transport problem, we may try to regularize on the Kantorovich potentials.

Proposed Regularization Problem

The first optimization problem we examine is the following, where H is a reproducing kernel Hilbert space and $c(x, y)$ denotes an arbitrary cost function in x and y. The functions u and v here are dual potentials.

$$\inf_{u,v} \left[\int u \, ds + \int v \, dt + \lambda(\|u\|_H^2 + \|v\|_H^2) \right], \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - u(x) - v(y) \leq c(x, y) \, \forall x, y.$$

In particular, we can derive a general form of the primal for this and then investigate behavior of the optimal coupling in this setting.

We will circle back to this after the next section.

3.5 Convex Analysis Prerequisites

Before we continue, we introduce some concepts from convex analysis that will be needed going forward.

Definition 23. *Let $X \subset \mathbb{R}^n$ and $f : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function and*

$$X^* = \left\{ x^* \in \mathbb{R} : \sup_{x \in X}(\langle x^*, x \rangle - f(x)) < \infty \right\}.$$

The Fenchel-Legendre transform $f^* : X^* \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$f^*(x^*) = \sup_{x \in X}[\langle x^*, x \rangle - f(x)]$$

This f^* is always convex.

Definition 24. *The subdifferential of a lower semi-continuous convex function ϕ at $x \in$ dom ϕ is defined by*

$$\partial\phi(x) = \{x^* \in X^* : \forall y \in X, \, \phi(y) - \phi(x) \geq \langle x^*, y - x \rangle\}$$

Theorem 3.5.1. *Let $f : X \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Then for any $x \in$ int dom f,*

$$\partial f(x) \neq \emptyset.$$

Theorem 3.5.2 (Fenchel-Young inequality)**.** *For any function f,*

$$f(x) + f^*(x^*) \geq \langle x^*, x \rangle$$

Theorem 3.5.3 (Fenchel-Young equality)**.**

$$f(x) + f^*(x^*) = \langle x^*, x \rangle$$

iff

$$x^* \in \partial f(x).$$

Theorem 3.5.4. *Let* $v(y) = \inf_x[f(x) + g(x+y)]$. *The Fenchel primal problem is*

$$p = v(0) = \inf_x[f(x) + g(x)].$$

The dual problem is

$$d = v^{**}(0) = \sup_{y^*}[-f^*(y^*) - g^*(-y^*)].$$

Proof. Calculate $v^*(-y^*) = \sup_{x,y}[\langle -y^*, y\rangle - f(x) - g(x+y)]$.

Let $u = x + y$, so we have

$$\begin{aligned} v^*(-y^*) &= \sup_{x,u}\langle -y^*, u - x\rangle - f(x) - g(u) \\ &= \sup_x[\langle y^*, x\rangle - f(x)] + \sup_u[\langle -y^*, u\rangle - g(u)] \\ &= f^*(y^*) + g^*(-y^*). \end{aligned}$$

Thus,

$$d = v^{**}(0) = \sup_{-y^*}[0 - v^*(-y^*)] = \sup_{-y^*}[-f^*(y^*) - g^*(-y^*)].$$

□

3.6 Derivation of Primal Formulation

Our goal is to find the primal formulation for the following optimization problem:

$$\inf_{u,v}\left[\int u\,ds + \int v\,dt + \lambda(\|u\|_H^2 + \|v\|_H^2)\right],\ \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - u(x) - v(y) \le c(x,y) \quad \forall x, y.$$

Ultimately, this is done in order to understand how to relate the optimal joint couplings with the source and target measures. Such a coupling then determines a "transport plan" between the source and target distributions via a computation of its barycentric projections [43]. The arguments that follow are similar to those used to prove Kantorovich duality.

Let

$$\phi_c = \{(u,v) \in C(X) \times C(Y) : \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - u(x) - v(y) \le c(x,y) \quad \forall x, y\}$$

.

In the next few lines, $z \in E = C(X \times Y)$ is the set of continuous functions on $X \times Y$ with sup. norm and $\pi \in E^* = M(X \times Y)$ is the set of regular Radon measures with total variation norm.

$$f(z) = \begin{cases} \lambda(\|u\|_H^2 + \|v\|_H^2) + \int u\, ds + \int v\, dt, & \text{if } z(x,y) = u(x) + v(y) \quad \forall x, y \\ +\infty, & \text{otherwise.} \end{cases}$$

$$g(z) = \begin{cases} 0, & \text{if } \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - z(x,y) \leq c(x,y) \quad \forall x, y \\ +\infty, & \text{otherwise.} \end{cases}$$

$$f(z)+g(z) = \begin{cases} \int_X u\, ds + \int_Y v\, dt + \lambda(\|u\|_H^2 + \|v\|_H^2), & \text{if } \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - u(x) - v(y) \leq c(x,y) \\ \infty, & \text{otherwise.} \end{cases}$$

$$\inf_{z \in E}[f(z) + g(z)] = \inf_{(u,v) \in \Phi_c} \left\{ \int_X u\, ds + \int_Y v\, dt + \lambda(\|u\|_H^2 + \|v\|_H^2) \right\}$$

$$\begin{aligned} g^*(-\pi) &= \sup_{z \in E} \left[-\int_{X \times Y} z\, d\pi - g(z) \right] \\ &= \sup_{z \in E} \left[-\int_{X \times Y} z\, d\pi : \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - c(x,y) \leq z(x,y) \right] \\ &= \begin{cases} \int_{X \times Y} \left[c(x,y) - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 \right] d\pi, & \text{if } \pi \in M_+(X \times Y) \\ +\infty, \text{ otherwise} \end{cases} \end{aligned}$$

$$\begin{aligned} f^*(\pi) &= \sup_{z \in E} \left[\int_{X \times Y} z\, d\pi - f(z) \right] \\ &= \sup_{z \in E} \left[\int_{X \times Y} z(x,y)\, d\pi(x,y) - \int_X u\, ds - \int_Y v\, dt - \lambda(\|u\|_H^2 + \|v\|_H^2) : z = u \oplus v \right] \\ &= \begin{cases} -\lambda(\|u\|_H^2 + \|v\|_H^2), & \text{if } \pi \in \Pi(s,t) \\ -\inf_u(\int u\, ds - \int u\, d\pi + \lambda\|u\|_H^2) - \inf_v(\int v\, dt - \int v\, d\pi + \lambda\|v\|_H^2), & \text{otherwise} \end{cases} \end{aligned}$$

Let π_1 and π_2 be the projections of the measure π onto the spaces X and Y respectively and let

$$Q(s, \pi_1, \lambda) := \inf_u \left(\int u\, ds - \int u\, d\pi + \lambda \|u\|_H^2 \right)$$

and

$$Q(t, \pi_2, \lambda) := \inf_v \left(\int v\, dt - \int v\, d\pi + \lambda \|v\|_H^2 \right).$$

Then our dual problem is

$$\sup_{\pi\in E^*}(-f^*(\pi)-g^*(-\pi)) = \sup_{\pi\in M_+}\left[Q(s,\pi_1,\lambda)+Q(t,\pi_2,\lambda)-\int\left(c(x,y)-\frac{1}{2}\|x\|^2-\frac{1}{2}\|y\|^2\right)d\pi\right].$$

If $c(x,y)=\frac{1}{2}\|x-y\|^2$, this becomes

$$\sup_{\pi\in E^*}(-f^*(\pi)-g^*(-\pi)) = \sup_{\pi\in M_+}\left[Q(s,\pi_1,\lambda)+Q(t,\pi_2,\lambda)+\int\langle x,y\rangle\, d\pi\right].$$

If we have additional knowledge about H, we can explicitly calculate closed form expressions for $Q(s,\pi_1)$ and $Q(t,\pi_2)$ that are dependent only on s, t, λ, and π.

3.7 An Alternative Optimization Problem

If we are working with finite sets S and T, we can represent the joint distribution on $S\times T$ by a doubly stochastic $|S|\times|T|$ matrix Π. We can write the marginal distributions on S and T as $\pi_1=\Pi 1$ and $\pi_2^T=1^T\Pi$, i.e. taking the row and column marginals of Π.

Here, we propose an alternative optimization inspired by our primal derivation in the previous section.

One possible optimization problem is

$$\sup_{\Pi}\frac{1}{4\lambda}(-\|s-\Pi 1\|^2-\|t-\Pi^T 1\|^2)+tr(\Pi)=\inf_{\Pi}\frac{1}{4\lambda}(\|s-\Pi 1\|^2+\|t-\Pi^T 1\|^2)-tr(\Pi).$$

We analyze the behavior of solutions when $|S|=|T|=2$.

What we will see is that the mass is concentrated along the diagonal of the joint probability matrix.

We look at the Karush-Kuhn-Tucker conditions and take derivatives with respect to Π,f$(\Pi,\lambda)=\frac{(\Pi\vec{1}-s)^T(\Pi\vec{1}-s)+(\Pi^T\vec{1}-t)^T(\Pi^T\vec{1}-t)}{4\lambda}-Tr[\Pi]+\tau(1-\vec{1}^T\Pi 1)-v_1E_1^T\Pi E_1-v_2E_2^T\Pi E_1-v_3E_1^T\Pi E_2-v_4E_2^T\Pi E_2$

where $E_1=\begin{pmatrix}1\\0\end{pmatrix}$ and $E_2=\begin{pmatrix}0\\1\end{pmatrix}$.

$\frac{df}{d\Pi}=2(\Pi\vec{1}-s)\vec{1}^T+2(\Pi^T\vec{1}-t)\vec{1}^T-I-\tau\vec{1}\vec{1}^T=2[(\Pi+\Pi^T)\vec{1}-(s+t)]\vec{1}^T-I-\tau 11^T-v_1\begin{pmatrix}1&0\\0&0\end{pmatrix}-v_2\begin{pmatrix}0&0\\1&0\end{pmatrix}-v_3\begin{pmatrix}0&1\\0&0\end{pmatrix}-v_4\begin{pmatrix}0&0\\0&1\end{pmatrix}$

By Karush Kuhn Tucker conditions, we have

$\frac{\pi_{11}-\pi_{22}-(s_1+t_1)+1}{2\lambda}-(\tau+v_1+1)=0$

$\frac{\pi_{22}-\pi_{11}-(s_2+t_2)+1}{2\lambda}-(\tau+v_2)=0$

$\frac{\pi_{11}-\pi_{22}-(s_1+t_1)+1}{2\lambda}-(\tau+v_3)=0$

$\frac{\pi_{22}-\pi_{11}-(s_2+t_2)+1}{2\lambda}-(\tau+v_4+1)=0$

And it follows that
$v_1 + 1 = v_3$
$v_4 + 1 = v_2$
and also
$\frac{1}{2\lambda} - 1 = v_1 + v_2 + 2\tau = v_3 + v_4 + 2\tau$.
If $\pi_{11}, \pi_{22} = \frac{s_1+t_1}{2}, \frac{s_2+t_2}{2}$, then $\pi_{11} + \pi_{22} = 1$ and thus $\pi_{12} = \pi_{21} = 0$. Letting $v_1 = v_4 = 0$ and $v_2 = v_3 = 1$, this satisfies slackness constraints.

So the joint probability matrix has all its mass concentrated on the diagonal and furthermore, the diagonal elements are averages of the source and target probabilities. This was also checked empirically – code is provided in the appendix. We also empirically note this result seems to hold when $\|S\| = \|T\| \geq 3$ as well. Extending this proof to S and T of arbitrary size and analyzing the significance of these solutions are left for future work.

3.8 Error Bounds

In this section, we have our main results where we examine the tradeoffs between OT and another dual regularized OT. The optimization problem considered here is slightly different from what we introduced earlier, but this is easier to use in order to demonstrate some statistical properties. First, we will compare theoretical regularized OT versus unregularized OT, and then we examine the sample complexity of regularized OT in the subsequent section.

Consider the RKHS dual regularized optimal transport problem

$$S(P,\ Q) = \sup_{f\in H, g\in H} \int f(x)\mathrm{d}P(x) + \int g(y)\mathrm{d}Q(y) - \lambda(\|f\|_H + \|g\|_H), \quad f \oplus g \leq c$$

and the unregularized OT,

$$OT(P,\ Q) = \sup_{f\in H, g\in H} \int f(x)\mathrm{d}P(x) + \int g(y)\mathrm{d}Q(y), \quad f \oplus g \leq c.$$

We will next see that the difference between these two quantities is dependent on the norms of the dual potentials for the unregularized and regularized optimal transport problem. The proof here is similar to that of [45] but here, we generalize from the discrete setting.

Theorem 3.8.1. *Let f_0, g_0 be dual solutions to $OT(P,Q)$ and f^*, g^* be corresponding solutions to $S(P,Q)$. Then*

$$\lambda(\|f^*\|_H + \|g^*\|_H) \leq OT(P,Q) - S(P,Q) \leq \lambda(\|f_0\|_H + \|g_0\|_H)$$

Proof. Since f^*, g^* satisfy the constraint $f^* \oplus g^* \leq c$, we have

$$\int f^*(x)\mathrm{d}P(x) + \int g^*(y)\mathrm{d}Q(y) \leq \int f_0(x)\mathrm{d}P(x) + \int g_0(y)\mathrm{d}Q(y) = OT(P,Q)$$

Subtracting $\lambda(\|f^*\| + \|g^*\|)$ from both sides gives us

$$\begin{aligned} S(P,Q) &= \int f^*(x)\mathrm{d}P(x) + \int g^*(y)\mathrm{d}Q(y) - \lambda(\|f^*\|_H + \|g^*\|_H) \\ &\leq \int f_0(x)\mathrm{d}P(x) + \int g_0(y)\mathrm{d}Q(y) - \lambda(\|f^*\|_H + \|g^*\|_H) \\ &= OT(P,Q) - \lambda(\|f^*\| + \|g^*\|) \end{aligned}$$

Thus we have $OT(P,Q) - S(P,Q) \geq \lambda(\|f^*\| + \|g^*\|)$.
Next, note that

$$-\int f_0(x)\mathrm{d}P(x) - \int g_0(y)\mathrm{d}Q(y) + \lambda(\|f_0\|_H + \|g_0\|_H) \geq -S(P,Q).$$

Rearranging this gives us $OT(P,Q) - S(P,Q) \leq \lambda(\|f_0\|_H + \|g_0\|_H)$.
Combining these two inequalities, we get our desired bound. □

3.9 Sample Complexity Bound

Introduction

Previously, we investigated the dual form of the dual norm regularized optimal transport and examined some of its fundamental properties. In particular, we show that our convergence rate is faster than that of unregularized OT under certain conditions.

Our quantity can be expressed as follows, where we regularize by the norms of the dual potentials f, g in an RKHS H,

$$S(P,\ Q) = \sup_{f\in L_1(P)\cap H, g\in L_1(Q)\cap H} \int f(x)\mathrm{d}P(x) + \int g(y)\mathrm{d}Q(y) - \lambda(\|f\|_H + \|g\|_H).$$

Throughout this section, we assume P and Q are measures on compact subsets $\mathcal{X}$ and $\mathcal{Y}$ in $\mathbb{R}^d$. And assume the cost function c is continuous in $\mathcal{X}\times\mathcal{Y}$. Let $K = \max_{X\times Y} c(x,y)$.

Also assume the RKHS H exhibits a Gaussian radial basis function kernel.

Proof Technique

The proof technique here is inspired by work on entropic optimal transport sample complexity [30]. First, we use a simple argument to remove the regularization terms. Then we examine the empirical process over some set that the optimal potentials belong to. Key to our results are the establishment of uniform bounds of the potential functions. And since these potentials are in an RKHS, we can exploit the structure of the RKHS to establish empirical process bounds. After controlling the potentials, we use a chaining bound.

Uniform bounds on the optimal potentials

Proposition 1. *The optimal f, g, denoted f^*, g^*, that maximize*

$$\int f(x)\mathrm{d}P(x) + \int g(y)\mathrm{d}Q(y) - \lambda(\|f\|_H + \|g\|_H)$$

subject to the constraint that $f \oplus g \leq c$ lie in a RKHS ball of radius $W(P,Q)/\lambda$, where

$$W(P,Q) = \min_{\pi \in \Pi(P,Q)} \int c\, d\pi.$$

Proof. Let f_0, g_0 be arbitrary potentials that satisfy the constraint $f_0 \oplus g_0 \leq c$.

$$\begin{aligned}
\lambda(\|f^*\| + \|g^*\|) &\leq \lambda(\|f^*\| + \|g^*\|) + \sup_{f,g s.t. f\oplus g \leq c} \left\{ \int f\, dP + \int g\, dQ \right\} - \left(\int f^*\, dP + \int g^*\, dQ \right) \\
&\leq \lambda(\|f^*\| + \|g^*\|) + W(P,Q) - \left(\int f^*\, dP + \int g^*\, dQ \right) \\
&\leq \left\{ \lambda(\|f_0\| + \|g_0\|) + W(P,Q) - \left(\int f_0\, dP + \int g_0\, dQ \right) \right\}
\end{aligned}$$

Assuming $c(x,y) \geq 0$, we can choose $f_0 = 0$ and $g_0 = 0$ and thus we get

$$\lambda(\|f^*\| + \|g^*\|) \leq W(P,Q)$$

which implies that f^* and g^* lie in an RKHS ball of radius $W(P,Q)/\lambda$. □

We denote by $\mathcal{F}$ the set of functions in the RKHS ball with radius $\frac{K}{\lambda}$. The following proposition shows that it suffices to control an empirical process indexed by this set.

Proposition 2. *Let P, Q, and P_n be probability distributions. Then*

$$|S(P_n,\ Q) - S(P,\ Q)| \leq 2 \sup_{u \in \mathcal{F}} |E_P u - E_{P_n} u|.$$

Proof. We define the operator $\mathcal{A}^{\alpha,\beta}(u,\ v)$ for the pair of probability measures $(\alpha,\ \beta)$ and functions $(u,\ v) \in L_1(\alpha) \otimes L_1(\beta)$ as:

$\mathcal{A}^{\alpha,\beta}(u,\ v) = \int u(x)\mathrm{d}\alpha(x) + \int v(y)\mathrm{d}\beta(y) - \lambda(\|u\|_H + \|v\|_H)$.

Denote by $(f_n,\ g_n)$ a pair of optimal potentials for $(P_n,\ Q)$ and $(f,\ g)$ for $(P,\ Q)$, respectively. We can choose smooth optimal potentials $(f,\ g)$ and $(f_n,\ g_n)$ to lie in the RKHS balls with radii $W(P,Q)/\lambda$ and $W(P_n,Q)/\lambda$ respectively for all $x, y \in \mathbb{R}^d$. And $W(P,Q) \leq K$ by construction and similarly for $W(P_n,Q)$. Thus $f,\ f_n \in \mathcal{F}$.

Note that $S(P,\ Q) = \mathcal{A}^{P,Q}(f,\ g)$ and $S(P_n,\ Q) = \mathcal{A}^{P_n,Q}(f_n,\ g_n)$.

Moreover, by the optimality of $(f,\ g)$ and $(f_n,\ g_n)$, we obtain

$$\mathcal{A}^{P,Q}(f_n,\ g_n) - \mathcal{A}^{P_n,Q}(f_n,\ g_n) \leq \mathcal{A}^{P,Q}(f,\ g) - \mathcal{A}^{P_n,Q}(f_n,\ g_n) \leq \mathcal{A}^{P,Q}(f,\ g) - \mathcal{A}^{P_n,Q}(f,\ g).$$

From the above bound, we see that

$$\begin{aligned}|S(P,\ Q) - S(P_n,\ Q)| &= |\mathcal{A}^{P,Q}(f,\ g) - \mathcal{A}^{P_n,Q}(f_n,\ g_n)| \\ &\leq |\mathcal{A}^{P,Q}(f,\ g) - \mathcal{A}^{P_n,Q}(f,\ g)| + |\mathcal{A}^{P,Q}(f_n,\ g_n) - \mathcal{A}^{P_n,Q}(f_n,\ g_n)|.\end{aligned}$$

All that is left is bounding the differences $|\mathcal{A}^{P,Q}(f,\ g) - \mathcal{A}^{P_n,Q}(f,\ g)|$ and $|\mathcal{A}^{P,Q}(f_n,\ g_n) - \mathcal{A}^{P_n,Q}(f_n,\ g_n)|$.

$$\begin{aligned}\left|\mathcal{A}^{P,Q}(f,\ g) - \mathcal{A}^{P_n,Q}(f,\ g)\right| &= \left|\int f(x)(\mathrm{d}P(x) - \mathrm{d}P_n(x))\right| \\ &\leq \sup_{u\in\mathcal{F}} |\int u(x)(\mathrm{d}P(x) - \mathrm{d}P_n(x))|.\end{aligned}$$

and similarly,

$$\begin{aligned}\left|\mathcal{A}^{P,Q}(f_n,\ g_n) - \mathcal{A}^{P_n,Q}(f_n,\ g_n)\right| &= \left|\int f_n(x)(\mathrm{d}P(x) - \mathrm{d}P_n(x))\right| \\ &\leq \sup_{u\in\mathcal{F}} |\int u(x)(\mathrm{d}P(x) - \mathrm{d}P_n(x))|.\end{aligned}$$

□

Corollary 3.9.0.1. *Let P, Q, P_n, and Q_n be probability distributions. Then*

$$|S(P_n, Q_n) - S(P, Q)| \lesssim 2\sup_{u\in\mathcal{F}} |\int u(x)(\mathrm{d}P(x) - \mathrm{d}P_n(x))| + 2\sup_{u\in\mathcal{F}} |\int u(x)(\mathrm{d}Q(x) - \mathrm{d}Q_n(x))|$$

PROOF. By the triangle inequality,

$$|S(P_n,\ Q_n) - S(P,\ Q)| \leq |S(P,\ Q) - S(P_n,\ Q)| + |S(P_n,\ Q) - S(P_n,\ Q_n)|.$$

almost surely. □

Bounding the empirical process

Denote by $N(\varepsilon,\ \mathcal{F}^s,\ L_2(P_n))$ the covering number with respect to the (random) metric $L_2(P_n)$ defined by

$$\|f\|_{L_2(P_n)} = \left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)^2\right)^{1/2}$$

The empirical process bounds established in this chapter rely on our reproducing kernel Hilbert space (RKHS) having a special structure. Particularly, the reason we assumed that the RKHS exhibits a Gaussian radial basis function kernel will soon be clear.

Our first preliminary result is the following proposition:

Proposition 3. *Let $\mathcal{H}_\sigma$ be a Gaussian radial basis function RKHS with the kernel defined as $k(x,y) = e^{-\sigma^2\|x-y\|_2^2}$. If P_n is an empirical distribution, then, given the sample $X_1, \ldots, X_n$, we have the bound*

$$\max_{f\in\mathcal{F}} \|f\|_{L_2(P_n)}^2 \leq \|f\|_{\mathcal{H}_\sigma}^2 \leq \frac{K^2}{\lambda^2} \ .$$

Proof. Denote

$$f(x) = \langle f(\cdot), K(x,\cdot)\rangle \leq \|f\|_{\mathcal{H}_\sigma} \max_{x\in X} \sqrt{k(x,x)}\, \forall f \in \mathcal{H}_\sigma, \forall x \in \mathcal{X}$$

and notice that $k(x,x) = 1$ if $k(x,y) = e^{-\sigma^2\|x-y\|_2^2}$. Thus $|f(x)| \leq \|f\|_{\mathcal{H}_\sigma}$ for all $x \in X$. This proposition actually holds for any RKHS exhibiting a kernel where $k(x,x) = 1$ for all x. □

Assuming the RKHS is Gaussian like in the previous proposition, then we have the covering number bound.

Theorem 3.9.1. *[47] Let $\sigma \geq 1$, $X \subset \mathbb{R}^d$ be a compact subset with nonempty interior, and $H_\sigma(X)$ be the RKHS of the Gaussian RBF kernel k_σ on X. Then for all $0 < p \leq 2$ and all $\delta > 0$, there exists a constant $c_{p,\delta,d} > 0$ independent of σ such that for all $\epsilon > 0$ we have*

$$\sup_{T\in(X\times Y)^n} \log N(\varepsilon,\ \mathcal{F},\ L_2(P_n)) \leq c_{p,\delta,d}\sigma^{(1-p/2)(1+\delta)d}\epsilon^{-p}.$$

Define

$$\|P - P_n\|_{\mathcal{F}} := \sup_{u\in\mathcal{F}} \left(\int u(x)\, (\mathrm{d}P(x) - \mathrm{d}P_n(x)) \right).$$

Using 3.9.1, we then have, by the use of a chaining bound [24],

$$\begin{aligned}
E\|P - P_n\|_{\mathcal{F}}^2 &\lesssim \frac{1}{n} E\Big(\int_0^{\sqrt{\max_{f\in\mathcal{F}}\|f\|_{L_2(P_n)}^2}} \sqrt{\log 2N(\epsilon,\mathcal{F},L_2(P_n))}\mathrm{d}\epsilon\Big)^2 \\
&\leq \frac{1}{n} E\Big(\int_0^{K/\lambda} \sqrt{1 + c_{p,\delta,d}\sigma^{(1-p/2)(1+\delta)d}\epsilon^{-p}}\mathrm{d}\epsilon\Big)^2 \\
&\leq \frac{c'_{p,\delta,d}\sigma^{(1-p/2)(1+\delta)d}}{n} E\Big(\int_0^{K/\lambda} \epsilon^{-p/2}\mathrm{d}\epsilon\Big)^2 \\
&= \frac{c'_{p,\delta,d}\sigma^{(1-p/2)(1+\delta)d}}{n(1-p/2)^2} \left(\frac{K}{\lambda}\right)^{2-p}
\end{aligned}$$

Here, $c_{p,\delta,d}$ and $c'_{p,\delta,d}$ denote constants with respect to n.

And by an application of Cauchy-Schwarz,

$$E\|P - P_n\|_{\mathcal{F}} \leq \left[E\|P - P_n\|_{\mathcal{F}}^2\right]^{1/2}$$

and we obtain the main result of this chapter, which is

Theorem 3.9.2. *Let $\sigma \geq 1$, $X \subset \mathbb{R}^d$ be a compact subset with nonempty interior, and let $H = H_\sigma(X)$ be the RKHS of the Gaussian RBF kernel k_σ on X. And let P_n and Q_n be the empirical distributions for P and Q. Then for all $0 < p \leq 2$ and all $\delta > 0$, there exists a constant $c_{p,\delta,d} > 0$ independent of σ such that for all $\epsilon > 0$ we have*

$$E\|P - P_n\|_{\mathcal{F}} \leq \frac{\sqrt{c_{p,\delta,d}\sigma^{(1-p/2)(1+\delta)d}}}{\sqrt{n}(1-p/2)} \left(\frac{K}{\lambda}\right)^{1-p/2}$$

and by an application of 3.9.0.1, we have that

$$|S(P_n, Q_n) - S(P, Q)| \leq \frac{\sqrt{c_{p,\delta,d}\sigma^{(1-p/2)(1+\delta)d}}}{\sqrt{n}(1-p/2)} \left(\frac{K}{\lambda}\right)^{1-p/2}.$$

Conclusion

We showed that our dual regularized optimal transport exhibits an $O(n^{-1/2})$ sample complexity, which is much stronger than the $O(n^{-1/d})$ rate of unregularized OT. One upside to our bounds is that there is no requirement that our probability measures must be subgaussian unlike those in [30]. And, by the results from 3.8.1, we can control the difference between regularized and unregularized OT via the size of the potential functions.

Of note is that the Gaussian RBF kernel assumption is not needed to obtain sample complexity bounds – it is sufficient to use an RKHS with a kernel where $k(x, x) = 1$ and there exist covering number bounds for non-Gaussian RKHSes as shown in [54]. However, the bounds will not be as sharp since those bounds were with respect to L_∞ instead of $L_2(P_n)$ and a future direction for this work is to obtain sharper bounds that are less dependent on the structure of the RKHS. Another possible direction is seeing how to relax the compactness assumptions.

Chapter 4

Domain adaptation using Monge mappings

4.1 Introduction

Previously, we discussed two methods of regularized optimal transport for domain adaptation:

- Entropic regularization
- Regularization with respect to dual potentials

However, the emphasis earlier was not on finding the explicit mapping between datasets. One may instead seek to directly estimate the Monge map between source and target datasets. In this chapter, we show how to establish the optimal linear transport map under arbitrary distributions with finite second moments, demonstrate that this is also the optimal transport map, and propose another regularization scheme that incorporates the explicit transport map.

We show why estimating the linear Monge map may be of importance if we know the second-order moments. And, as stated previously, the optimal linear transport map is the optimal transport map in this setting. This result was also discovered concurrently but independently in [20].

4.2 Background

Let A and B be positive matrices. Most of these definitions and propositions are taken from [9].

Definition 25. *An $n \times n$ symmetric matrix M is a positive matrix if and only if $x^{'}Mx > 0$ for all $x \in \mathbb{R}^n$ where $x \neq 0$.*

And we also introduce the concept of fidelity measure between matrices.

Definition 26 (Fidelity)**.** *If A and B are positive matrices, then the fidelity $F(A,B) = tr(A^{1/2}BA^{1/2})^{1/2}$.*

Definition 27 (Bures distance)**.** $d(A,B) = [tr(A) + tr(B) - 2tr(A^{1/2}BA^{1/2})^{1/2}]^{1/2}$.

Definition 28 (Operator norm)**.** *Define $\|\cdot\|$ as the operator norm such that*

$$\|A\| = \sup_{\|x\|=1} \|Ax\| = \sup_{\|x\|\leq 1} \|Ax\|.$$

The next two propositions and lemma build up on each other and can be seen in [9] along with much of the other background details here.

Proposition 4. *The operator A is contractive, i.e. $\|A\| \leq 1$ if and only if the operator $\begin{bmatrix} I & A \\ A^* & I \end{bmatrix}$ is positive.*

Proof. We do a singular value decomposition on A ($A = USV$). Then

$$\begin{aligned} \begin{bmatrix} I & A \\ A^* & I \end{bmatrix} &= \begin{bmatrix} I & USV \\ V^*SU^* & I \end{bmatrix} \\ &= \begin{bmatrix} U & O \\ O & V^* \end{bmatrix} \begin{bmatrix} I & S \\ S & I \end{bmatrix} \begin{bmatrix} U^* & O \\ O & V \end{bmatrix}. \end{aligned}$$

This is unitarily equivalent to $\begin{bmatrix} I & S \\ S & I \end{bmatrix}$, which is also (unitarily) equivalent to the direct sum

$$\begin{bmatrix} 1 & s_1 \\ s_1 & 1 \end{bmatrix} \oplus \begin{bmatrix} 1 & s_2 \\ s_2 & 1 \end{bmatrix} \oplus \cdots \oplus \begin{bmatrix} 1 & s_n \\ s_n & 1 \end{bmatrix}$$

where $s_1, \ldots, s_n$ are the singular values of A. These 2×2 matrices are all positive if and only if $s_1 \leq 1$ (i.e., $\|A\| \leq 1$). □

Proposition 5. *Let A, B be positive. Then the matrix $\begin{bmatrix} A & X \\ X^* & B \end{bmatrix}$ is positive $\Leftrightarrow$ if $X = A^{1/2}KB^{1/2}$ for some contraction K.*

Proof.

$$\begin{aligned} \begin{bmatrix} A & X \\ X^* & B \end{bmatrix} &\sim \begin{bmatrix} A^{-1/2} & O \\ O & B^{-1/2} \end{bmatrix} \begin{bmatrix} A & X \\ X^* & B \end{bmatrix} \begin{bmatrix} A^{-1/2} & O \\ O & B^{-1/2} \end{bmatrix} \\ &= \begin{bmatrix} I & A^{-1/2}XB^{-1/2} \\ B^{-1/2}X^*A^{-1/2} & I \end{bmatrix} \end{aligned}$$

Let $K = A^{-1/2}XB^{-1/2}$. This matrix is positive if and only if K is a contraction. □

Lemma 4.2.1.

$$F(A,B) = \max_{X>0} \left\{ |\operatorname{tr} X| : A \geq XB^{-1}X^* \right\}$$

Proof. Note that

$$A \geq MB^{-1}M^* \Leftrightarrow \begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \geq 0 \Leftrightarrow M = A^{1/2}KB^{1/2}$$

for some contraction K.

By the Schwarz inequality we have

$$\begin{aligned} |\operatorname{tr} M| = & |\operatorname{tr}\left(A^{1/2}KB^{1/2}\right)| \leq \left\|A^{1/2}K\right\|_2 \left\|B^{1/2}\right\|_2 \\ & \leq \left\|A^{1/2}\right\|_2 \left\|B^{1/2}\right\|_2 = \sqrt{\operatorname{tr} A \operatorname{tr} B} \end{aligned}$$

If

$$\begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \geq 0$$

then for every $X > 0$ we have

$$\begin{aligned} 0 \leq & \begin{bmatrix} X^{1/2} & O \\ O & X^{-1/2} \end{bmatrix} \begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \begin{bmatrix} X^{1/2} & O \\ O & X^{-1/2} \end{bmatrix} \\ = & \begin{bmatrix} X^{1/2}AX^{1/2} & X^{1/2}MX^{-1/2} \\ X^{-1/2}M^*X^{1/2} & X^{-1/2}BX^{-1/2} \end{bmatrix} \end{aligned}$$

Hence

$$\left|\operatorname{tr} X^{1/2}MX^{-1/2}\right| \leq \sqrt{\operatorname{tr}\left(X^{1/2}AX^{1/2}\right) \operatorname{tr}\left(X^{-1/2}BX^{-1/2}\right)}$$

In other words,

$$|\operatorname{tr} M| \leq \sqrt{\operatorname{tr}(AX) \operatorname{tr}\left(BX^{-1}\right)}$$

This is true for all M satisfying the condition $A \geq MB^{-1}M^*$ and for all $X > 0$.

So

$$\begin{aligned} \max\left\{ |\operatorname{tr} M| : A \geq MB^{-1}M^* \right\} & \leq \min_{X>0} \sqrt{\operatorname{tr}(AX) \operatorname{tr}\left(BX^{-1}\right)} \\ & = F(A,B) \end{aligned}$$

Let $M = (AB)^{1/2} = A\left(A^{-1}\#B\right)$. Then

$$\begin{aligned} MB^{-1}M^* &= (AB)^{1/2}B^{-1}(BA)^{1/2} = B^{-1}B(AB)^{1/2}B^{-1}(BA)^{1/2} \\ &= B^{-1}(BA)^{1/2}(BA)^{1/2} = B^{-1}BA = A \end{aligned}$$

The maximum on the left hand side of 4.2.1 is attained when $M = (AB)^{1/2}$ and it is equal to $\operatorname{tr}\left(A^{1/2}BA^{1/2}\right)^{1/2} = F(A,B)$. This completes the proof. □

Also let $A\#B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}$ be the matrix geometric mean. With this established, we can introduce our main lemma.

4.3 Optimal Linear Transport Mapping

Definition 29 (Optimal linear transport map)**.** *Let X and Y be sets with measures μ and ν respectively. Then the optimal linear transport map $T : X \to Y$ is the one minimizes*

$$E \left\| x - Tx \right\|^2$$

over linear maps T.

Theorem 4.3.1 (Optimal Linear Transport is Optimal Transport)**.** *Let μ and ν be probability measures with zero means and positive-definite covariance matrices A, B respectively such that $\nu = \tilde{T}_{\#}\mu$ for a linear positive-definite $\tilde{T}$. And define our cost function as $c(x, y) := \left\| x - y \right\|^2$. Then the optimal linear transport map coincides with the optimal transport map.*

Proof. The first half of this proof demonstrates the existence of an optimal linear transport map by construction and is adapted from [9] from complex to real matrices. The second half of the proof shows that this optimal linear transport map is the optimal transport map by an application of 3.4.1.

Let $x \sim \mu$ and $y \sim \nu$ be random vectors with values in $\mathbb{R}^n$, each having zero mean WLOG, and with covariance matrices A and B, respectively.

$$A_{ij} = [E(x_i x_j)],\ B_{ij} = [E(y_i y_j)].$$

We want to find x and y for which $E\|x - y\|^2$ is minimal.

The covariance matrix of the vector $(x,\ y)$ is

$$\begin{bmatrix} [E(x_i x_j)] & [E(x_i y_j)] \\ [E(y_i x_j)] & [E(y_i y_j)] \end{bmatrix} = \begin{bmatrix} A & M \\ M^* & B \end{bmatrix}$$

We aim to minimize the following:

$$\begin{aligned} E\|x - y\|^2 &= E(\sum_{i=1}^{n}(|x_i|^2 + |y_i|^2 - 2x_i y_i)) \\ &= \sum_{i=1}^{n} E(|x_i|^2 + |y_i|^2 - 2x_i y_i) \\ &= \mathrm{tr}(A + B) - 2trM. \end{aligned}$$

This is equivalent to the following optimization problem:

$\max\{|\mathrm{tr}M| : C = \begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \geq 0\}$.

The value of the maximum is $F(A,\ B)$ by 4.2.1. So

$$\min E\|x - y\|^2 \ = \ \mathrm{tr}(A + B) - 2\mathrm{tr}(A^{1/2} B A^{1/2})^{1/2} = \ d^2(A,\ B).$$

Let x be a vector with mean 0 and covariance matrix A. Then for any $n \times n$ real-valued matrix, we have

$$\begin{aligned} E(\langle x,\ Tx\rangle) &= E(\sum_{i,j} t_{ij}x_i x_j) = \sum_{i,j} t_{ij}E(x_i x_j) \\ &= \sum_{i,j} t_{ij}a_{ij} = \mathrm{tr}TA. \end{aligned}$$

Hence,

$$\begin{aligned} E\|x - Tx\|^2 &= E(\|x\|^2 + \|Tx\|^2 - 2\langle x,\ Tx\rangle) \\ &= \mathrm{tr}A + \mathrm{tr}T^*TA - 2\mathrm{tr}TA \\ &= \mathrm{tr}A + \mathrm{tr}TAT^* - 2\mathrm{tr}A^{1/2}TA^{1/2} \end{aligned}$$

If we choose $T = A^{-1}\#B$, then we see that tr $A^{1/2}TA^{1/2} =$ tr $(A^{1/2}BA^{1/2})^{1/2}$, and that $\mathrm{tr}TAT^* = \mathrm{tr}B$.

Thus, for this choice of T, we have

$$E\|x - Tx\|^2 = tr(A + B) - 2tr(A^{1/2}BA^{1/2})^{1/2} = \ d^2(A,\ B)\ .$$

Thus the problem

$$\min E\|x - y\|^2$$

where x, y are vectors with mean zero and covariance matrices A and B, respectively, has as its solution the pairs $(x,\ y)$, where x is any vector centered at zero with covariance A and $y = Tx$, with $T = A^{-1}\#B$.

Let x be a vector with covariance matrix A, and let $y = Tx$. Then

$$\begin{aligned} E(y_i y_j) \ &= \ E\sum_{k,l} t_{ik}t_{kl}x_k x_l \\ &= \ \sum_{k,l} t_{ik}t_{kl}a_{kl} = (TAT)_{ij}. \end{aligned}$$

If $T = A^{-1}\#B$, then $TAT^* = B$. This shows that the covariance matrix of the vector y is B.

Now we have shown that T is the optimal linear transport map, but this is not enough to show T is the optimal transport map because we have only shown that $T = A^{-1}\#B$ preserves means and second moments. This does not necessarily mean $T = \tilde{T}$, i.e. that it maps the distribution μ into ν.

The insight we had was that we can combine the above result with Brenier's theorem. This is because $T = A_{\#}^{-1}B$ is the gradient of a convex function, $\frac{1}{2}x^TTx$. And recall that theorem 3.4.1 by Brenier states that the gradient of a convex function is the optimal transport map. Thus T is the optimal transport map, i.e. $T = \tilde{T}$. □

One could extend this result to nonzero means if we allow for affine maps ($\nu = \tilde{T}_{\#}\mu + b$) where b is a constant vector. The proof is mostly the same except we first center μ and ν.

4.4 Proposed Regularized Optimization

One may want to look at a regularization of the combination of the squared Bures distance and the expected distance moved under the transport map T by the squared Hilbert-Schmidt norm of T. The motivation behind this is we want to match the second moments between source and target distributions (squared Bures distance) while controlling the 'size' of the transport map with the Hilbert-Schmidt norm.

Definition 30 (Hilbert-Schmidt Norm)**.** *For a matrix A,*

$$\|\mathrm{A}\|_{HS} := \sqrt{\sum_{i,j} a_{ij}^2}.$$

The reason for this section is just to propose a conjectured regularization and that one can feasibly solve for the mapping.

Definition 31 (Regularization With Transport Map)**.**

$$\min_T Q(T) \text{ where } Q(T) = \lambda E \|TX - X\|^2 + d^2(Cov(TX), \Sigma_v) + \mu \|T\|_{HS}^2$$

Differentiating $Q(T)$, we get

$$2\lambda(TX - X)X^T + (\frac{2}{n-1})CTX(I - (BA)^{-1/2}B)X^T + 2\mu T$$

where $A = Cov(TX) = (\frac{1}{n-1})(TX)^T C(TX)$, $B = \Sigma_v$, the covariance matrix on the target distribution, and $C = (1 - \frac{1}{n}J) = C^T$ is a centering matrix. See 7.2 for more details on how this is derived.

Solving for $\hat{T}$ such that $Q'(\hat{T}) = 0$, we get

$$\lambda XX^T = \lambda \hat{T} XX^T + \mu \hat{T} + \frac{1}{n-1} C\hat{T}[X(I - (BA)^{-1/2}B)]X^T.$$

This is a Sylvester equation that has a closed form solution.

Chapter 5

Asymptotics for Prior Elicitation

5.1 Statistical Elicitation

In this section, the main focus of the study will be on the case of elicitation to obtain a prior that will be used in a subsequent machine learning task. Elicitation is the process of forming a probability distribution from a person's knowledge and beliefs. However, most of the results will apply to other motivations for elicitation.

Classically, elicitation is a human-centered process with multiple roles: The *modeler* will ultimately do the modeling, with the elicited prior. The *facilitator* has a strategy and asks questions to gather information to use for inference. The *expert* has the knowledge that the facilitator will use. A *statistician* will train the expert on probability and provide feedback.

An individual may fill multiple roles; for example, a single individual commonly fills both the statistician and facilitator roles. The expert may also be the modeler who will ultimately use the elicited prior.

Elicitation is a multi-stage process, typified by the following steps. The modeler and the statistician will determine the target value during the structuring and decomposition steps.

Then, during the elicitation phase, there is further iteration over three steps, which we list below.

1. Elicit summaries.
2. Fit a distribution.
3. Assess adequacy.

It is important to highlight that these three steps are the backbone in fitting and assessment of the automated tool. In higher dimensions, summaries are less intuitive and even

cumbersome to communicate. In automated facilitator-statistician discussion, we will shift from eliciting summaries to eliciting samples. Sample based elicitation has been applied for fitting beta distributions.

Literature on elicitation focuses on making inferences from the type of information provided by elicitation and the related psychological literature. The psychology of elicitation relates to how people characterize uncertainty (not consistently) to what information is actually needed in order to make inferences about uncertainty that are themselves useful for further inferences.

5.2 Sample Based Elicitation

Prior work using sample-based elicitation used a method of moments technique with weighted samples. We propose a similar elicitation procedure, but consider a distinct inference technique. Moving to a least squares based approach with a stated objective function for the prior enables cases where the moments do not exist.

In order to build a general automated elicitation tool, we need to consider how the tool will learn from the expert. In the end, this learning will be an online process which learns from each sample sequentially and then presents the updated model to the user for feedback.

In elicitation, the examples will be provided by a human expert along with given instructions. For our work, we assume this will result in samples that are more diverse than a sample directly from the distribution for two reasons. First, an expert is unlikely to give an example that is very close to a previous sample, which leads to the generation of a representative sample that explains the range of their belief. Second, the instructions can prompt the user to provide examples that are both likely and unlikely. Therefore, by examining the i.i.d case, we are obtaining a worst-case estimate of the learnability of the problem.

In this section, we present our main analytical results. First, we will introduce our least squares based objective function. Next, we will consider the large sample behavior of the proposed estimator by evaluating the consistency of the estimator, and we will show the conditions under which we achieve asymptotic normality. Third, we present a finite sample result.

5.3 A Least-Squares Based Approach to Elicitation

Proposed Objective Function

Assume that we elicit i.i.d. observations x_i with corresponding sample log-likelihoods z_i for $i = 1, \ldots, n$. Let $\ell(x, \theta)$ be the log-likelihood function evaluated at parameter θ and obser-

vation x. Assuming we have a parametric model class, our proposed method of estimating θ involves minimizing an objective function, which we illustrate below. Let

$$Q((\vec{x},\vec{z}),\theta) = \sum_i \left(\ell(x_i,\theta) - z_i\right)^2.$$

Our proposed optimization problem is

$$\hat{\theta} = \arg\min_\theta \sum_i \left(\ell(x_i,\theta) - z_i\right)^2 = \arg\min_\theta Q((\vec{x},\vec{z}),\theta),$$

where x_i, z_i is the ith sample and likelihood.

$$\frac{dQ}{d\theta} = 2\sum_i \left((\ell(x_i,\theta) - z_i)\frac{\partial}{\partial\theta} l(x_i,\theta) \right)$$

and let $\psi((x_i,z_i),\theta) = \ell(x_i,\theta) - z_i$.
$\hat{\theta}$ is a solution to

$$\sum_i \left((\ell(x_i,\theta) - z_i)\frac{\partial}{\partial\theta} l(x_i,\theta) \right) = 0$$

and θ_0 solves

$$\mathrm{E}_\theta \left[(\ell(x_i,\theta) - z_i)\frac{\partial}{\partial\theta} l(x_i,\theta) \right] = 0$$

Asymptotic Analysis

The following details are immediate applications of results in [49], which are also provided in the appendix in section 7.3.
Let Ω be the parameter space with an open set ω such that θ_0, the true parameter value, is an interior point.

Consistency

Then if $\left((\ell(x_i,\theta) - z_i)\frac{\partial}{\partial\theta}\ell(x_i,\theta)\right)$ is monotone in θ, continuous in a neighborhood of θ_0, and θ_0 is an isolated root, $\hat{\theta} \xrightarrow{\mathcal{P}} \theta_0$.

Asymptotic Normality

Note that

$$\frac{\partial}{\partial\theta}\psi((x,z),\theta) = \left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right]^2 + (\ell(x,\theta) - z)\frac{\partial^2}{\partial\theta^2}\ell(x,\theta).$$

If

$$\mathrm{E}_{\theta_0}\left[\left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right]^2 + (\ell(x,\theta) - z)\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\right]$$

is finite and nonzero and

$$\mathrm{E}_{\theta_0}\left[\left\{\left[\frac{\partial}{\partial\theta}\ell(x,\theta)\right]^2+(\ell(x,\theta)-z)\frac{\partial^2}{\partial\theta^2}\ell(x,\theta)\right\}^2\right]<\infty$$

then

$$\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{\mathcal{L}}\ \mathcal{N}(0,\sigma^2_{\hat{\theta}})$$

where

$$\sigma^2_{\hat{\theta}}=\frac{\mathrm{E}_{\theta_0}[\psi^2((X,Z),\theta_0)]}{(\mathrm{E}_{\theta_0}[\frac{\partial}{\partial\theta}\psi((X,Z),\theta)|_{\theta=\theta_0}])^2}.$$

5.4 Main Result

Setting and Assumptions

Let $X, X_1, X_2, \ldots$ be random variables mapping from $(\Omega,\mathcal{A})$ to $(\mathcal{X},\mathcal{B})$ and let $(P_\theta)_{\theta\in\Theta}$ be a parametric family of probability measures such that $X, X_1, X_2, \ldots$ are i.i.d. drawn from some P_θ with $\theta\in\Theta$. Importantly, $\Theta\subseteq\mathbb{R}$, i.e. the parameter space Θ is a subset of the real line.
Let E_θ be the expectation with respect to P_θ. For each $\theta\in\Theta$, X has a density p_θ with respect to a measure μ on $\mathcal{B}$. And assume $\ell_{\mathbf{x},\mathbf{z}}$, the log-likelihood function, is continuous in θ.

Since the extended real line $[-\infty,\infty]$ is compact, for each $n\in\mathbb{N}$ and points $\mathbf{x}=\mathbf{x}_n=(x_1,\ldots,x_n)\in\mathcal{X}^n$ and $\mathbf{z}=\mathbf{z}_n=(z_1,\ldots,z_n)\in\mathcal{Z}^n$, the function $\Theta\ni\theta\mapsto\ell_{\mathbf{x},\mathbf{z}}(\theta)=\sum_{i=1}^n-(\ell_{x_i}(\theta)-z_i)^2$ has at least one generalized maximizer $\hat{\theta}_n(\mathbf{x},\mathbf{z})$ in the closure of Θ in the sense that $\sup_{\theta\in\Theta}\ell_{\mathbf{x},\mathbf{z}}(\theta)=\limsup_{\theta\to\hat{\theta}_n(\mathbf{x},\mathbf{z})}\ell_{\mathbf{x},\mathbf{z}}(\theta)$.

Picking, for each $\mathbf{x}=(x_1,\ldots,x_n)\in\mathcal{X}^n$ and $\mathbf{z}=\mathbf{z}_n=(z_1,\ldots,z_n)\in\mathcal{Z}^n$, any one of such generalized maximizers $\hat{\theta}_n(\mathbf{x},\mathbf{z})$, one obtains a map $\Omega\ni\omega\mapsto\hat{\theta}_n(\mathbf{X}(\omega),\mathbf{Z}(\omega))$, where $\mathbf{X}:=\mathbf{X}_n:=(X_1,\ldots,X_n)$ and $\mathbf{Z}:=\mathbf{Z}_n:=(Z_1,\ldots,Z_n)$. Any such map will be denoted here by $\hat{\theta}_n(\mathbf{X},\mathbf{Z})$ (or by $\hat{\theta}_n$ or $\hat{\theta}$).
Let $\theta_0\in\Theta$ be the expert's target value of the parameter θ, such that

$$[\theta_0-\delta,\ \theta_0+\delta]\subseteq\Theta^\circ$$

for some real $\delta>0$, where Θ° denotes the interior of the subset Θ of $\mathbb{R}$.
And

$$\ell_{\mathbf{X},\mathbf{Z}}(\theta):=-\sum_{i=1}^n(\ell(X_i,\theta)-Z_i)^2.$$

for $\theta \in \theta$, $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

Sufficient assumptions for ℓ are below.

1. The set $\mathcal{X}_{>0} := \{x \in \mathcal{X} : p_\theta(x) > 0\}$ is the same for all $\theta \in [\theta_0 - \delta,\ \theta_0 + \delta]$, and for each $x \in \mathcal{X}_{>0}$ $\ell_x(\theta)$ is thrice differentiable in θ at each point $\theta \in [\theta_0 - \delta,\ \theta_0 + \delta]$.
2. $\mathrm{E}\ell'_{X,Z}(\theta_0)^2 = I_1(\theta_0)$ and $-\mathrm{E}\ell''_{X,Z}(\theta_0) = I_2(\theta_0) \in (0, \infty)$.
3. $\mathrm{E}|\ell'_{X,Z}(\theta_0)|^3 + \mathrm{E}|\ell''_{X,Z}(\theta_0)|^3 < \infty$.
4. $\mathrm{E}\sup|\ell'''_{X,Z}(\theta)|^3 < \infty$.
5. $\ell_{x,z}(\theta)$ is concave in $\theta \in \Theta$, for each $x, z \in \mathcal{X} \times \mathcal{Z}$.
6. $$\mathrm{E}\frac{\exp(\ell_{X,Z}(\theta_0 \pm h))}{\exp(\ell_{X,Z}(\theta_0))} < 1$$

Theorem 5.4.1 (Finite-Sample Bound)**.** *Suppose that the above conditions hold. Then*

$$\left|\mathrm{P}\left(\sqrt{n\frac{I_2(\theta_0)^2}{I_1(\theta_0)}}(\hat{\theta} - \theta_0) \le z\right) - \Phi(z)\right| \le \frac{\mathfrak{C}}{\sqrt{n}}$$

for all real z, and

$$\left|\mathrm{P}(\sqrt{n\frac{I_2(\theta_0)^2}{I_1(\theta_0)}}(\hat{\theta} - \theta_0) \le z) - \Phi(z)\right| \le \frac{\mathfrak{C}_\omega}{z^3\sqrt{n}}$$

for $z \in (0, \omega\sqrt{n}]$ for any $\omega \in (0, \infty)$. $\mathfrak{C}_\omega$ is a finite expression that depends on ω and neither $\mathfrak{C}$ and $\mathfrak{C}_\omega$ depend on n or z.

5.5 Proof of Theoretical Bound

Here, we provide a theoretical proof of Theorem 5.4.1. This follows closely the technique introduced in [39] for maximum likelihood estimators. Here, we extend the result in [39] to M-estimators, and in particular, we fully exposit the proof for our proposed estimator. The proof requires substantial new adjustments to the assumptions, which were not discussed in [39].

This proof is organized as follows.

1. We demonstrate tight bracketing of our M-estimator between two functions of the sum of independent random vectors. The key here is to show $\hat{\theta}-\theta_0$ is bounded on a specific subset.

2. We present uniform and nonuniform optimal-order bounds on the convergence rate in the multivariate delta method [40].

3. We apply the general bounds in the multivariate delta method such that we can make bracketing work.
$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta}-\theta_0)\leq z)-\Phi(z)|$$
is bounded above by the sum of a "error" term and a remainder term. We show here the error term is $O(1/\sqrt{n})$.

4. We bound the remainder term and show this is asymptotically negligible under certain conditions. In particular, we show a sufficient condition for this to be exponentially decreasing with respect to n.

Tight Bracketing

Without loss of generality (w.l.o.g.), $\mathcal{X}_{>0}=\mathcal{X}$. Then on the event
$$G:=\{\hat{\theta}\in[\theta_0-\delta,\ \theta_0+\delta]\} \tag{5.1}$$
(G for "good event"), one must have
$$0=\ell'_{\mathbf{X},\mathbf{Z}}(\hat{\theta})=\ell'_{\mathbf{X},\mathbf{Z}}(\theta_0)+(\hat{\theta}-\theta_0)\ell''_{\mathbf{X},\mathbf{Z}}(\theta_0)+\frac{(\hat{\theta}-\theta_0)^2}{2}\ell'''_{\mathbf{X},\mathbf{Z}}(\theta_0+\xi(\hat{\theta}-\theta_0)) \tag{5.2}$$
$$=n(\overline{K}-(\hat{\theta}-\theta_0)\overline{U}+\frac{(\hat{\theta}-\theta_0)^2}{2}\overline{R}) \tag{5.3}$$
for some $\xi\in(0,1)$ as a function of the X_i's and Z_i's, where $\overline{K}=\frac{1}{n}\sum_{i=1}^n K_i$, $\overline{U}=\frac{1}{n}\sum_{i=1}^n U_i$, $\overline{R}:=\frac{1}{n}\sum_{i=1}^n R_i$, $\overline{R^*}:=\frac{1}{n}\sum_{i=1}^n R_i^*$,
$$K_i=\ell'_{X_i,Z_i}(\theta_0),\quad U_i=-\ell''_{X_i,Z_i}(\theta_0) \tag{5.4}$$
$$R_i=\ell'''_{X_i,Z_i}(\theta_0+\xi(\hat{\theta}-\theta_0))\in[-R_i^*,R_i^*],\quad R_i^*=\sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\ell'''_{X_i,Z_i}(\theta)|. \tag{5.5}$$

On the event G one has
$$\hat{\theta}-\theta_0=\frac{\overline{K}}{\overline{U}}\text{ if }\overline{R}=0\quad\&\quad\overline{U}\neq 0,$$
$$\hat{\theta}-\theta_0\in\{d_+,\ d_-\}\text{ if }\overline{R}\neq 0,$$

where

$$d_{\pm} := \frac{\overline{U} \pm \sqrt{\overline{U}^2 - 2\overline{KR}}}{\overline{R}}.$$

One defines a "bad event" by letting
$B := B_1 \cup B_2$, where

$$B_1 := \{\overline{R} \neq 0, \hat{\theta} - \theta_0 = d_+\} \cup \{\overline{U} \leq 0\} \text{ and } B_2 := \{\overline{U}^2 \leq 2|\overline{K}|\overline{R^*}\}.$$

In the rest of this section, we prove the following two lemmas:

Lemma 5.5.1.

$$\mathrm{P}(G \cap B) \leq \mathrm{P}(G \cap B_1) + \mathrm{P}(B_2) \leq \frac{\mathfrak{C}}{n^{3/2}}, \tag{5.6}$$

where $\mathfrak{C}$ depends on ℓ, the measure μ, and the choice of θ_0– but not on n.

and

Lemma 5.5.2. *On $G \backslash B$ one has*

$$\overline{U} > 0 \text{ and } \hat{\theta} - \theta_0 = \frac{2\overline{K}}{\overline{U} + \sqrt{\overline{U}^2 - 2\overline{KR}}} \in [T_-,\ T_+] \tag{5.7}$$

where

$$T_{\pm} := \frac{2\overline{K}}{\overline{U} + \sqrt{\overline{U}^2 \mp 2|\overline{K}|\overline{R^*}}}. \tag{5.8}$$

Proof. On the event $B_1 \cap \{\overline{U} > 0\}$, one sees $|\hat{\theta} - \theta_0| = |d_+| \geq \overline{U}/|\overline{R}| \geq \overline{U}/\overline{R^*}$
By (5.1),

$$\mathrm{P}(G \cap B_1) \leq \mathrm{P}(\overline{U} \leq 0 \text{ or } \frac{\overline{U}}{\overline{R^*}} \leq \delta) = \mathrm{P}(\frac{\overline{U}}{\overline{R^*}} \leq \delta) = \mathrm{P}(\sum_{i=1}^{n}(U_i - \delta R_i^*) \leq 0). \tag{5.9}$$

And by the assumptions for ℓ and the definitions for K_i, U_i, R_i, and R_i^*,

$$\mathrm{E}U_1 > 0,\ \mathrm{E}|K_1|^3 < \infty,\ \mathrm{E}|U_1|^3 < \infty,\ \mathrm{E}(R_1^*)^3 < \infty.$$

Therefore, $\mathrm{E}R_1^* < \infty$. Choose $\delta > 0$ to be small enough such that

$$\delta_1 := \mathrm{E}(U_i - \delta R_i^*) > 0.$$

Then, letting $Y_i := (U_i - \delta R_i^*) - \mathrm{E}(U_i - \delta R_i^*)$, we use (5.9) with Markov's inequality and 7.4.1 to have

$$\begin{aligned}\mathrm{P}(G \cap B_1) \leq \mathrm{P}(\sum_{i=1}^{n} Y_i \leq -n\delta_1) &\leq \frac{1}{(n\delta_1)^3}\mathrm{E}|\sum_{i=1}^{n} Y_i|^3 \\ &\leq \frac{n\mathrm{E}|Y_1|^3 + \sqrt{8/\pi}(n\mathrm{E}Y_1^2)^{3/2}}{(n\delta_1)^3} \leq \frac{\mathfrak{C}}{n^{3/2}}\end{aligned}$$

where $\mathfrak{C} := (\mathrm{E}|Y_1|^3 + \sqrt{8}/\pi(\mathrm{E}Y_1^2)^{3/2})/\delta_1^3$, which depends on $\delta_1 > 0, \mathrm{E}Y_1^2 < \infty$, and $\mathrm{E}|Y_1|^3 < \infty$. However, this does not depend on n.

B_2 implies at least one of the following events:

$$\begin{aligned}B_{21} &= \{\overline{U} \leq \frac{1}{2}\mathrm{E}U_1\} \\ B_{22} &= \{\overline{R^*} \geq 1 + \mathrm{E}R_1^*\}, \text{ or} \\ B_{23} &= \{|\overline{K}| \geq \frac{1}{8}(\mathrm{E}U_1)^2/(1 + \mathrm{E}R_1^*)\}.\end{aligned}$$

Therefore,

$$\mathrm{P}(B_2) \leq \mathrm{P}(B_{21}) + \mathrm{P}(B_{22}) + \mathrm{P}(B_{23}). \tag{5.10}$$

The bounding of each of the probabilities $\mathrm{P}(B_{21})$, $\mathrm{P}(B_{22})$, $\mathrm{P}(B_{23})$ is quite similar to the bounding of $\mathrm{P}(G \cap B_1)$ – because

$$\begin{aligned}\mathrm{P}(B_{21}) &= \mathrm{P}(\sum_{i=1}^{n} Y_{i,21} \leq -n\delta_{21}), \\ \mathrm{P}(B_{22}) &= \mathrm{P}(\sum_{i=1}^{n} Y_{i,22} \geq n\delta_{22}) \text{ ,and} \\ \mathrm{P}(B_{23}) &= \mathrm{P}(\sum_{i=1}^{n} |Y_{i,23}| \geq n\delta_{23}).\end{aligned}$$

It follows that

$$\mathrm{P}(G \cap B) \leq \mathrm{P}(G \cap B_1) + \mathrm{P}(B_2) \leq \frac{\mathfrak{C}}{n^{3/2}}, \tag{5.11}$$

where $\mathfrak{C}$ depends on ℓ, the measure μ, and the choice of θ_0- but not on n.

On the other hand, if $\overline{R} \neq 0$ and $\overline{U} > 0$, then $d_- = \dfrac{2\bar{K}}{\bar{U} + \sqrt{\bar{U}^2 - 2\bar{K}\bar{R}}}$. Here, the condition $\overline{U} > 0$ ensures that the denominator of the latter ratio is nonzero. Thus, on the event $G \backslash B$ one has

$$\overline{U} > 0 \text{ and } \hat{\theta} - \theta_0 = \frac{2\overline{K}}{\overline{U} + \sqrt{\overline{U}^2 - 2\overline{K}\overline{R}}} \in [T_-,\ T_+] \tag{5.12}$$

where

$$T_\pm \ := \frac{2\overline{K}}{\overline{U} + \sqrt{\overline{U}^2 \mp 2|\overline{K}|\overline{R^*}}}. \tag{5.13}$$

Thus we have our bracketing of $\hat{\theta} - \theta_0$ between T_- and T_+. □

General uniform and nonuniform bounds on the rate of convergence to normality for smooth nonlinear functions of sums of independent random vectors

Denote the standard normal distribution function (d.f.) by Φ. For any $\mathbb{R}^d$-valued random vector ζ,

$$\|\zeta\|_p := (\mathrm{E}\|\zeta\|^p)^{1/p} \text{ for any real } p \geq 1,$$

where $\| \ . \ \|$ denotes the Euclidean norm on $\mathbb{R}^d$.

Take any Borel-measurable functional $f : \mathbb{R}^d \to \mathbb{R}$ satisfying the following smoothness condition: there exist $\epsilon \in (0,\ \infty), M_\epsilon \in (0,\ \infty)$, and a linear functional $L : \mathbb{R}^d \to \mathbb{R}$ such that

Definition 32 (Smoothness Condition)**.**

$$|f(\mathrm{x}) - L(\mathrm{x})| \leq \frac{M_\epsilon}{2}\|\mathrm{x}\|^2 \textit{ for all } \mathrm{x} \in \mathbb{R}^d \textit{ with } \|\mathrm{x}\| \leq \epsilon. \tag{5.14}$$

Thus, $f(0) = 0$ and L necessarily coincides with the first Fréchet derivative, $f'(0)$, of the function f at 0. Moreover, for the smoothness condition to hold, it is enough that

$$M_\epsilon \geq M_\epsilon^* := \sup\{\frac{1}{\|\mathrm{x}\|^2}|\frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\mathrm{x} + t\mathrm{x})|_{t=0}| \ : \ \mathrm{x} \in \mathbb{R}^d,\ 0 < \|\mathrm{x}\| \leq \epsilon\}.$$

Notice that f does not need to be twice differentiable at 0. One example is if $d = 1$ and $f(x) = \dfrac{x}{1 + |x|}$ for $x \in \mathbb{R}$.

Let $V, V_1, \ldots, V_n$ be i.i.d. random vectors in $\mathbb{R}^d$, with $\mathrm{E}V = 0$ and

$$\overline{V} := \frac{1}{n}\sum_{i=1}^{n} V_i.$$

And let

$$\tilde{\sigma} := \|L(V)\|_2, v_3 := \|V\|_3, \text{ and } \varsigma_3 := \frac{\|L(V)\|_3}{\tilde{\sigma}}. \tag{5.15}$$

Theorem 5.5.3. *Suppose that the smoothness condition holds and that $\tilde{\sigma} > 0$ and $v_3 < \infty$. Then for all $z \in \mathbb{R}$*

$$|\mathrm{P}(\frac{f(\overline{V})}{\tilde{\sigma}/\sqrt{n}} \leq z) - \Phi(z)| \leq \frac{\mathfrak{C}}{\sqrt{n}}, \tag{5.16}$$

where $\mathfrak{C}$ is a finite positive expression that depends only on the function f and the moments $\tilde{\sigma}$, ς_3, and v_3. Moreover, for any $\omega \in (0, \infty)$ and for all

$$z \in (0, \omega\sqrt{n}], \tag{5.17}$$

one has

$$|\mathrm{P}(\frac{f(\overline{V})}{\tilde{\sigma}/\sqrt{n}} \leq z) - \Phi(z)| \leq \frac{\mathfrak{C}_\omega}{z^3\sqrt{n}} \tag{5.18}$$

where $\mathfrak{C}_\omega$ is a positive, finite, and only depends on f through the smoothness condition, the moments $\tilde{\sigma}$, ς_3, and v_3, and ω. This is in [40].

Applying bracketing

In this section, we apply the previous bounds from Theorem 5.5.3 for a particular function that we construct that satisfies the smoothness condition (5.14). Specifically, we prove the following lemma:

Lemma 5.5.4. *For all real z,*

$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \leq z) - \Phi(z)| \leq \frac{\mathfrak{C}}{\sqrt{n}} + \mathrm{P}(|\hat{\theta} - \theta_0| > \delta). \tag{5.19}$$

And

$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \leq z) - \Phi(z)| \leq \frac{\mathfrak{C}}{z^3\sqrt{n}} + \mathrm{P}(|\hat{\theta} - \theta_0| > \delta) \tag{5.20}$$

for $z \in (0, \omega\sqrt{n}]$.

Proof. Let $d = 3$ and

$$\mathcal{D} := \{\mathrm{x} = (x_1,\ x_2,\ x_3) \in \mathbb{R}^3\ :\ x_2 + \mathrm{E}U_1 > 0,\ (x_2 + \mathrm{E}U_1)^2 > 2|x_1||x_3 + \mathrm{E}R_1^*|\}.$$

By (5.5) and assumptions 2 and 4 for ℓ, $\mathrm{E}U_1 = I_2(\theta_0) \in (0,\ \infty)$ and $\mathrm{E}R_1^* \in [0,\ \infty)$. So, for some real $\epsilon > 0$, the set $\mathcal{D}$ contains the ϵ-neighborhood of the origin 0 of $\mathbb{R}^3$.

Define functions $f_\pm : \mathbb{R}^3 \to \mathbb{R}$ by the formula

$$f_\pm(\mathrm{x}) = f_\pm(x_1,\ x_2,\ x_3) = \frac{2x_1}{x_2 + \mathrm{E}U_1 + \sqrt{(x_2 + \mathrm{E}U_1)^2 \mp 2|x_1||x_3 + \mathrm{E}R_1^*|}} \tag{5.21}$$

for $\mathrm{x} = (x_1,\ x_2,\ x_3) \in \mathcal{D}$, and let $f_\pm(\mathrm{x}) := 0$ if $\mathrm{x} \in \mathbb{R}^3 \backslash \mathcal{D}$.
Clearly, $f_\pm(0) = 0$,

$$L_\pm(\mathrm{x}) := f'_\pm(0)(\mathrm{x}) = \frac{x_1}{\mathrm{E}U_1} = \frac{x_1}{I_2(\theta_0)} \tag{5.22}$$

for $\mathrm{x} = (x_1,\ x_2,\ x_3) \in \mathbb{R}^3$, and the smoothness condition (5.14) holds for some ϵ and M_ϵ in $(0,\ \infty)$ –because, as was noted above, $\mathrm{E}U_1 = I_2(\theta_0) \in (0,\ \infty)$ and $\mathrm{E}R_1^* \in [0,\ \infty)$, and hence the denominator of the ratio in (5.21) is bounded away from 0 for $\mathrm{x} = (x_1,\ x_2,\ x_3)$ in a neighborhood of 0.

Next, let

$$V_i := (K_i,\ U_i - \mathrm{E}U_i,\ R_i^* - \mathrm{E}R_i^*) \tag{5.23}$$

for $i = 1, \ldots, n$, with K_i, U_i, R_i^* as defined in (5.5) and (5.4) . Then, by (5.15), (5.22) and condition 2, for $f = f_\pm$,

$$\tilde{\sigma} = \sqrt{\frac{\mathrm{E}K_1^2}{I_2(\theta_0)^2}} = \frac{\sqrt{I_1(\theta_0)}}{I_2(\theta_0)} > 0 \tag{5.24}$$

and $v_3^3 = \mathrm{E}\|V\|^3 < \infty$ by the third and fourth conditions. This shows that all the required conditions for (5.5.3) are satisfied for $f = f_\pm\cdot$.
Moreover, by (5.23), (5.21), and (5.5.2),

$$T_\pm = f_\pm(\overline{V})$$

on the event $G\backslash B$. So, by the inclusion relation in (5.12) (which holds on the event $G\backslash B = (G^{\mathrm{c}} \cup B)^{\mathrm{c}}$) and (5.24), inequality (5.16) in Theorem 5.5.3 implies

$$\begin{aligned}\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \le z) &\le \mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)f_-(\overline{V}) \le z) + \mathrm{P}(G^{\mathrm{c}} \cup B) \\ &\le \Phi(z) + \frac{\mathfrak{C}}{\sqrt{n}} + \mathrm{P}(G^{\mathrm{c}} \cup B)\end{aligned}$$

and

$$\begin{aligned}\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta}-\theta_0)\le z) &\ge \mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)f_+(\overline{V})\le z) - \mathrm{P}(G^{\mathrm{c}}\cup B)\\ &\ge \Phi(z) - \frac{\mathfrak{C}}{\sqrt{n}} - \mathrm{P}(G^{\mathrm{c}}\cup B)\ ,\end{aligned}$$

for all real z. Note that $\mathrm{P}(G^{\mathrm{c}}\cup B) = \mathrm{P}(G^{\mathrm{c}}) + \mathrm{P}(G\cap B)$. It follows now by (5.1) and (5.5.1) that

$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta}-\theta_0)\le z) - \Phi(z)| \le \frac{\mathfrak{C}}{\sqrt{n}} + \mathrm{P}(|\hat{\theta}-\theta_0|>\delta) \tag{5.25}$$

for all real z. Quite similarly, but using (5.18) instead of (5.16) , one has

$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta}-\theta_0)\le z) - \Phi(z)| \le \frac{\mathfrak{C}}{z^3\sqrt{n}} + \mathrm{P}(|\hat{\theta}-\theta_0|>\delta) \tag{5.26}$$

for $z \in (0, \omega\sqrt{n}]$ as in (5.17). □

Under certain conditions, the remainder term $\mathrm{P}(|\theta-\theta_0|>\delta)$ typically decreases exponentially fast in n and thus is negligible as compared with the "error" term $\frac{\mathfrak{C}}{\sqrt{n}}$, and even with the error term $\frac{\mathfrak{C}}{z^3\sqrt{n}}-$ under condition (5.17).
In the next section, we will prove that the remainder term is negligible under special assumptions.

Bounding the remainder

Before we proceed, we use the following assumptions:

1. $\ell_{x,z}(\theta)$ is concave in $\theta\in\theta$, for each $x\in\mathcal{X}$ and $z\in\mathcal{Z}$
2. $$\mathrm{E}\left[\frac{\exp(\ell_{X,Z}(\theta_0\pm h))}{\exp(\ell_{X,Z}(\theta_0))}\right] < 1.$$

Suppose that the $\ell_{x,z}(\theta)$ is concave in $\theta\in\theta$, for each $x\in\mathcal{X}$ and $z\in\mathcal{Z}$. By assumption 2, $\mathrm{E}\ell''_{X,Z}(\theta_0)\neq 0$. Hence, $\mathrm{P}(\ell_{X,Z}(\theta_0+h)\neq\ell_{X,Z}(\theta_0))>0$ for some $h\in(0,\ \delta)$. The concavity of $\ell_{x,z}(\theta)$ in θ implies that of $\ell_{\mathbf{X},\mathbf{Z}}(\theta)$. So, if $\hat{\theta}>\theta_0+\delta$, then $\ell_{\mathbf{X},\mathbf{Z}}(\theta_0+h)\ge\ell_{\mathbf{X},\mathbf{Z}}(\theta_0)$. This

is because $\hat{\theta}$ maximizes $\ell_{\mathbf{X},\mathbf{Z}}$ and by concavity, ℓ is increasing when $\theta < \hat{\theta}$.
Therefore,

$$\mathrm{P}(\hat{\theta} > \theta_0 + \delta) \leq \mathrm{P}(\ell_{X,Z}(\theta_0 + h) \geq \ell_{X,Z}(\theta_0)) = \mathrm{P}\left(\prod_{i=1}^{n} \sqrt{\frac{\exp(\ell_{X_i,Z_i}(\theta_0 + h))}{\exp(\ell_{X_i,Z_i}(\theta_0)}} \geq 1\right)$$
$$\leq \mathrm{E}\prod_{i=1}^{n} \sqrt{\frac{\exp(\ell_{X_i,Z_i}(\theta_0 + h))}{\exp(\ell_{X_i,Z_i}(\theta_0)}} = \lambda_+^n,$$

where

$$\lambda_+ := \mathrm{E}\sqrt{\frac{\exp(\ell_{X,Z}(\theta_0 + h))}{\exp(\ell_{X,Z}(\theta_0)}} < \sqrt{\mathrm{E}\frac{\exp(\ell_{X,Z}(\theta_0 + h))}{\exp(\ell_{X,Z}(\theta_0))}} < 1;$$

the inequality here is an instance of a strict version of the Cauchy-Schwarz inequality, which holds because $\mathrm{P}(\ell_{X,Z}(\theta_0 + h) \neq \ell_{X,Z}(\theta_0)) > 0$. Similarly, $\mathrm{P}(\hat{\theta} < \theta_0 - \delta) \leq \lambda_-^n$ for some $\lambda_- \in [0, 1)$, and so,

$$\mathrm{P}(|\hat{\theta} - \theta_0| > \delta) \leq 2\lambda^n \quad (6.1)$$

for $\lambda := \max(\lambda_+,\ \lambda_-) \in [0, 1)$.
Now recall Lemma 5.5.4, that said the following:
For all real z,

$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \leq z) - \Phi(z)| \leq \frac{\mathfrak{C}}{\sqrt{n}} + \mathrm{P}(|\hat{\theta} - \theta_0| > \delta). \quad (5.27)$$

And

$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \leq z) - \Phi(z)| \leq \frac{\mathfrak{C}}{z^3\sqrt{n}} + \mathrm{P}(|\hat{\theta} - \theta_0| > \delta) \quad (5.28)$$

for $z \in (0, \omega\sqrt{n}]$.

But since we showed the remainder is exponentially decreasing in n, the remainder is dominated by $\frac{\mathfrak{C}}{\sqrt{n}}$ or $\frac{\mathfrak{C}}{z^3\sqrt{n}}$. Combining the remainder bounds with the bracketing earlier, we have finished our proof.

5.6 Conclusion and Future Considerations

In this chapter, we proposed an elicitation approach involving minimizing a least-squares objective functions and demonstrated that this has promising asymptotic and finite-sample properties under certain assumptions. One can immediately note that the proof can be easily generalized to any M-estimator by replacing $\ell_{X,Z}(\theta)$ with another objective function.

Future work may entail relaxing the log-concavity assumptions on $\ell_{x,z}$. One route to consider may be seeing if this could be relaxed if we were to consider a compact parameter

space Θ. Another condition to investigate may be if one can obtain similar finite sample bounds for a multivariate parameter. In this chapter, $\theta \in \Theta \subset \mathbb{R}$ was necessary for our tight bracketing results to hold.

Chapter 6

Sliced Mixed-Marginal Wasserstein

6.1 Abstract

Multi-marginal optimal transport enables one to compare multiple probability measures, which increasingly finds application in multi-task learning problems. One practical limitation of multi-marginal transport is computational scalability in the number of measures, samples and dimensionality. In this work, we propose a multi-marginal optimal transport paradigm based on random one-dimensional projections, whose (generalized) distance we term the *sliced multi-marginal Wasserstein distance*. To construct this distance, we introduce a characterization of the one-dimensional multi-marginal Kantorovich problem and use it to highlight a number of properties of the sliced multi-marginal Wasserstein distance. In particular, we show that (i) the sliced multi-marginal Wasserstein distance is a (generalized) metric that induces the same topology as the standard Wasserstein distance, (ii) it admits a dimension-free sample complexity, (iii) it is tightly connected with the problem of barycentric averaging under the sliced-Wasserstein metric. We conclude by illustrating the sliced multi-marginal Wasserstein on multi-task density estimation and multi-dynamics reinforcement learning problems.

This chapter consists of joint work done with Alex Terenin, Samuel Cohen, Sesh Kumar, and Marc Deisenroth at University College of London.

6.2 Introduction

Optimal transport is a framework for defining meaningful metrics between probability measures [51, 36]. These metrics find a wide range of applications, such as generative modeling [22, 14], Bayesian inference [46], imitation learning [18], graph matching and averaging [53, 52]. Multi-marginal optimal transport [21] studies ways of comparing more than two probability measures in a geometrically meaningful way. Multi-marginal distances defined using this paradigm are often useful in settings where sharing geometric structure is useful, such as

multi-task learning. In particular, they have been applied for training multi-modal generative adversarial networks [15], clustering [8], and computing barycenters of measures [4].

Following the establishment of key theoretical results, including by Gangbo and Święch [21], Agueh and Carlier [1], and Pass [35], research is shifting toward applications. This motivates a need for practical algorithms for the multi-marginal setting [28]. Standard approaches based on linear programming and entropic regularization scale exponentially with the number of measures, and/or the dimension of the space [7, 48]. A number of recent works have therefore studied settings, where multi-marginal transport problems can be efficiently solved via low-rank structures on the underlying cost function [4], but exponential cost in the dimension remains [2, 3].

In parallel, a number of works on *sliced transport* [12] developed techniques for scalable transport, which (i) derive a closed form for a problem in a single dimension, and (ii) extend it into higher dimensions via random linear projections (slicing) and thereby inherit the complexity of the one-dimensional problem. This strategy has been shown effective in the classical Wasserstein [12, 11, 26, 33, 19, 42] and Gromov–Wasserstein [50] settings between pairs of measures, but has not yet been applied to settings with more than two measures.

In this chapter, we address this gap and propose *sliced multi-marginal transport*, providing a scalable analog of the multi-marginal Wasserstein distance. To do so, we derive a closed-form expression for multi-marginal Wasserstein transport in one dimension, which lifts to a higher-dimensional analog via slicing. This one-dimensional closed-form expression can be computed with a complexity of $\mathcal{O}(PN \log N)$, where P is the number of measures and N is the number of samples per measure. Sliced multi-marginal Wasserstein ($\mathcal{SMW}$) can be estimated by Monte Carlo in $\mathcal{O}(KPN \log N)$, where K is the number of Monte Carlo samples.

Furthermore, we study $\mathcal{SMW}$'s theoretical properties. We prove that (i) it is a generalized metric, whose associated topology is the topology of weak convergence, (ii) its sample complexity is dimension free, just like the sliced Wasserstein case involving two measures, and (iii) sliced multi-marginal transport is closely connected with the problem of barycentric averaging under the sliced Wasserstein metric. We also showcase applications, where we focus on multi-task learning on probability spaces, where sharing knowledge across tasks can be beneficial and sliced multi-marginal Wasserstein can be used as a regularizer between task-specific models. We demonstrate this on a multi-task density estimation problem, where individual estimation tasks are corrupted and shared structure is needed to solve the problem, as well as a reinforcement learning problem, where certain agents receive no reward and must instead learn from other agents to solve their given task.

6.3 Background

Multi-marginal optimal transport [21] is a class of optimization problems for comparing multiple measures $\mu_1, \ldots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$, all supported on the metric space $(\mathbb{R}^d, ||\cdot||_2)$. The most common such problem is computing the multi-marginal Wasserstein distance, defined

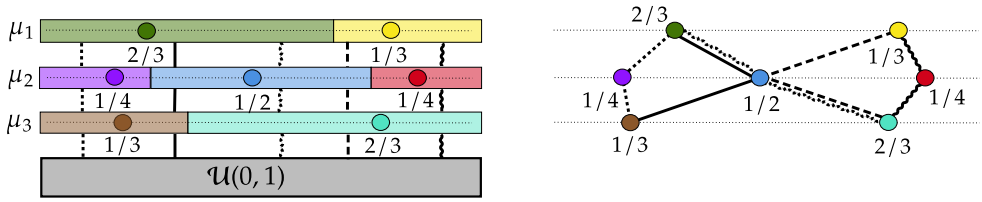


Figure 6.1: Illustration of the optimal coupling's structure on $\mathbb{R}$ between discrete measures μ_1, μ_2 and μ_3. Points are samples of each measures, with weights next to them. Left: histogram of measures (horizontal); joint samples are obtained by sampling a (black) line uniformly (drawn vertically), and picking points that are associated with the bin intersected by that line. Right: Corresponding triples of points that are aligned according to the coupling are linked by a pair of lines.

as

$$\mathcal{MW}^2(\mu_1, \ldots, \mu_P) = \min_{\pi \in \Pi(\mu_1, \ldots, \mu_P)} \int_{(\mathbb{R}^d)^P} c(x_1, \ldots, x_P)\, d\pi(x_1, \ldots, x_P), \tag{6.1}$$

where $c : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \to \mathbb{R}$ is a cost function and $\Pi(\mu_1, \ldots, \mu_P)$ is the set of probability measures in $\mathcal{M}((\mathbb{R}^d)^P)$ with marginals $\mu_1, \ldots, \mu_P$. We focus on the barycentric cost of Gangbo and Święch [21] and Agueh and Carlier [1], given by

$$c(x_1, \ldots, x_P) = \sum_{p=1}^{P} \beta_p \left\| x_p - \sum_{j=1}^{P} \beta_j x_j \right\|^2, \quad \beta_1, \ldots, \beta_P \geq 0, \quad \sum_{p=1}^{P} \beta_p = 1.$$

Above, the $\beta_1, \ldots, \beta_P$ are fixed.

This cost was originally motivated from an economics-inspired perspective, but is also often preferable because it leads to connections with barycentric averaging [1], giving it a simple interpretation. It also recovers the Wasserstein distance with squared Euclidean cost in the case $P = 2$ (up to constants), referred to as $\mathcal{W}$. Algorithms for estimating (6.1) from a set of samples scale exponentially with the number of measures P and/or the dimension d of the ground space [4, 2, 7].

$\mathcal{MW}$ is useful in multi-task settings for regularizing measures $\mu_1, \ldots, \mu_P$ by adding $\mathcal{MW}(\mu_1, \ldots, \mu_P)$ to a multi-task loss. It can also be used in a setting, where we aim for a model output μ to be close to a given set of measures $\nu_1, \ldots, \nu_P$, which can be done by introducing a loss of the form $\mathcal{MW}(\mu, \nu_1, \ldots, \nu_P)$ and minimizing it with respect to μ.

Sliced transport. With the usual Euclidean-type cost structures, the Wasserstein distance between pairs of one-dimensional discrete measures can be computed efficiently using *sorting* with $\mathcal{O}(N \log N)$ complexity. More generally, we can consider the average distance

between measures projected onto $\mathbb{R}$ along a random axis, which gives [12, 11]

$$\mathcal{SW}^2(\mu,\nu) = \int_{S_{d-1}} \mathcal{W}^2\big(M^\theta_\#(\mu), M^\theta_\#(\nu)\big)\, d\Theta(\theta),$$

where $M^\theta(x) = x^T\theta$, $(.)_\#$ denotes the push-forward of measures, and Θ is the uniform distribution on the unit sphere S_{d-1}. We sample from $M^\theta_\#(\mu)$ by sampling from μ and projecting onto θ.

A fundamental result by Bonnotte [12] is that $\mathcal{SW}$ is a metric that metrizes the topology of weak convergence—the *exact same* topology as $\mathcal{W}$. $\mathcal{SW}$ can be estimated via Monte Carlo and preserves the computational complexity of estimating $\mathcal{W}$ on $\mathbb{R}$, which is $\mathcal{O}(N \log N)$. Owing to the Monte Carlo nature, the sample complexity of $\mathcal{SW}$ is dimension free [12, 32], in contrast with the exponential dependency of the Wasserstein distance on dimension. The combination of good computational and statistical properties makes $\mathcal{SW}$ an attractive choice for minimization problems on measure spaces, including generative modeling and imitation learning [19, 18]. This immediately raises the question whether $\mathcal{SW}$ extends to the multi-marginal case so that it preserves its key appealing properties.

Definition 33. *For a measure $\mu \in \mathcal{M}(\mathbb{R})$, define its CDF $C_\mu : \mathbb{R} > [0,1]$ as*

$$C_\mu(x) = \int_{-\infty}^{x} d\mu(y)\ \forall x.$$

Also, define its pseudo-inverse $C^{-1}_\mu : [0,1] \to \mathbb{R} \cup \{-\infty\}$ as

$$C^{-1}_\mu(r) = \min_x \{x \in \mathbb{R} \cup \{-\infty\} : C_\mu(x) \geq r\}.$$

This function is a generalization of the quantile function.

Proposition 6. *If $\mu_1, \dots, \mu_P \in M(\mathbb{R})$ and $\mathcal{U}(0,1)$ is the uniform measure, then*

$$\mathcal{MW}^2(\mu_1,\dots,\mu_P) = \int_0^1 \sum_{p=1}^{P} \beta_p \left| C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) \right|^2 dx, \tag{6.2}$$

and the optimal coupling solving (6.1) *is of the form*

$$\pi^\star = (C^{-1}_{\mu_1}, \dots, C^{-1}_{\mu_P})_\#\ \mathcal{U}(0,1).$$

Sliced Multi-Marginal Wasserstein Distance

To define the sliced multi-marginal Wasserstein distance, we average the expressions given in (6.2) along one-dimensional random projections, which gives

$$\mathcal{SMW}^2(\mu_1,\dots,\mu_P) = \int_{S_{d-1}} \int_0^1 \sum_{p=1}^{P} \beta_p \left| C^{-1}_{\mu^\theta_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu^\theta_j}(x) \right|^2 dx\, d\Theta(\theta), \tag{6.3}$$

where $\mu^\theta_j = M^\theta_\#(\mu_j)$ for $j = 1, \dots, P$. $\mathcal{SMW}$ in (6.3) can be estimated via Monte Carlo in $O(KPN \log N)$, where K is the number of Monte Carlo samples (projections).

Topological properties We now study $\mathcal{SMW}$'s topological properties. We first show that $\mathcal{SMW}$ is the weighted mean of sliced Wasserstein distances between pairs of measures.

Proposition 7. *Let $\mu_1, \dots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$. We have that*

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P) = \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \mathcal{SW}^2(\mu_i, \mu_j).$$

Proposition 7 is useful in deriving statistical and topological properties of $\mathcal{SMW}$. It is however more efficient to estimate it via our closed-form formula for multi-marginal transport – see (6.3). This leads to a computational complexity of $O(KPN \log N)$, whereas naively implementing (7) scales in $\mathcal{O}(KP^2N \log N)$. Furthermore, as the sliced-Wasserstein metric is upper-bounded by the Wasserstein [12], an immediate consequence of Proposition 7 is that

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P) \overset{(7)}{=} \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \mathcal{SW}^2(\mu_i, \mu_j) \leq \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \mathcal{W}^2(\mu_i, \mu_j).$$

This shows that $\mathcal{SMW}$ gives rise to the topology of weak convergence—one of the key properties that made $\mathcal{SW}$ an attractive choice in the first place. We now study metric properties of $\mathcal{SMW}$.

Proposition 8. *$\mathcal{SMW}$ is a generalized metric. In particular, this means that $\mathcal{SMW}$ is*

- *non-negative,*
- *zero if and only if all measures are identical,*
- *permutation-equivariant, and*
- *satisfies a generalized triangle inequality involving multiple measures.*

These are all proven at the end of this chapter in Definition 34. Hence, $\mathcal{SMW}$ is well-behaved topologically-wise as it is a generalized metric inducing weak convergence. We continue by studying $\mathcal{SMW}$'s statistical properties.

Statistical Properties In the following proposition, we assess the impact of the number of samples and random projections used to estimate $\mathcal{SMW}$.

Proposition 9. *If $\mu_1, \dots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$, and assuming $\mathcal{W}^2$ has sample complexity $\rho(N)$ on $\mathbb{R}$, i.e.*

$$\mathrm{E}[\mathcal{W}^2(\mu_1, \dots, \mu_P) - \mathcal{W}^2(\hat{\mu}_1, \dots, \hat{\mu}_P)]^2 \leq \rho(N),$$

then

$$\mathrm{E}[\mathcal{SMW}^2(\mu_1,\dots,\mu_P) - \mathcal{SMW}^2(\hat{\mu}_1,\dots,\hat{\mu}_P)]^2 \leq \frac{1}{2}\rho(N),$$

where $\hat{\mu}_p$ refers to empirical measures with N samples.

Proposition 9 shows that the sample complexity of $\mathcal{SMW}$ is dimension-free—this stands in contrast to the sample complexity of the multi-marginal Wasserstein, which is exponential in the dimension. In practice, we use Monte Carlo sampling to compute $\mathcal{SMW}$, which introduces additional error. To understand this error, we examine $\mathcal{SMW}$'s projection complexity.

Proposition 10. *Let $\mu_1,\dots,\mu_P \in \mathcal{M}(\mathbb{R}^d)$, and define $\overline{\mathcal{SMW}}$ the approximation obtained by uniformly picking L projections on S_{d-1}, then*

$$\mathrm{E}\left[\overline{\mathcal{SMW}}^2(\mu_1,\dots,\mu_P) - \mathcal{SMW}^2(\mu_1,\dots,\mu_P)\right]^2 \leq L^{-1/2}\mathrm{Var}_\theta[\mathcal{MW}^2(\mu_1^\theta,\dots,\mu_P^\theta)],$$

where θ follows the uniform distribution on S_{d-1} and $\mu_p^\theta = M_\#^\theta(\mu_p)$.

This shows that the quality of Monte Carlo estimates of $\mathcal{SMW}$ is controlled by number of projections and the variance of evaluations of the base multi-marginal Wasserstein in 1D.

Connection to Barycenters We now study connections of $\mathcal{SMW}$ to the problem of barycentric averaging, which extends the notion of a *mean* to more general settings. Let $\mathcal{D} : \mathcal{M}(\mathbb{R}^d) \times \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}$ be a divergence on the space of probability measures. Recall that the *barycenter* of P measures $\mu_1,\dots,\mu_P$ is defined as

$$\mu^\star := \arg\min_{\mu\in\mathcal{M}(\mathbb{R}^d)} \mathcal{F}(\mu), \quad \mathcal{F}_\mathcal{D}^{\mu_1,\dots,\mu_P}(\mu) := \sum_{p=1}^P \mathcal{D}(\mu_p,\mu).$$

Barycentric averaging is well-studied from theoretical and computational view-points, notably under the squared Wasserstein [17], sliced Wasserstein [11] and Gromov–Wasserstein [37] metrics.

Proposition 11. *Let $\mu_1,\dots,\mu_P \in \mathcal{M}(\mathbb{R}^d)$, $\beta_p \geq 0$, $\sum_{p=1}^P \beta_p = 1$. Furthermore, let $\hat{\beta}_p$ be augmented multi-marginal weights, so that for $m \in [0,1]$ it holds that $\hat{\beta}_p = m\beta_p$ for $p = 1,\dots,P$, $\sum_{p=1}^{P+1}\hat{\beta}_p = 1$, and $D = \mathcal{SW}^2$. Then*

$$\arg\min_{\mu\in\mathcal{M}(\mathbb{R}^d)} \mathcal{SMW}^2_{\hat{\beta}}(\mu_1,\dots,\mu_P,\mu) = \arg\min_{\mu\in\mathcal{M}(\mathbb{R}^d)} \mathcal{F}_{\mathcal{D},\beta}^{\mu_1,\dots,\mu_P}(\mu),$$

where β is the weight vector of $\mathcal{F}_\mathcal{D}^{\mu_1,\dots,\mu_P}$ and $\hat{\beta}$ is the weight vector of $\mathcal{SMW}$.

Proposition 11 reveals a connection between sliced multi-marginal transport and barycenters under the sliced-Wasserstein: the measure that is closest to $\mu_1,\dots,\mu_P$ in $\mathcal{SMW}$ is actually the barycenter of such measures under $\mathcal{SW}$. We continue by studying differentiability of $\mathcal{SMW}$ as a loss function.

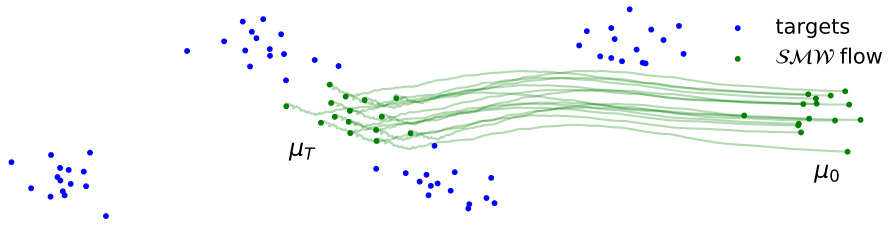


Figure 6.2: Gradient flow $\partial\mu_t = -\nabla\mathcal{SMW}^2(\mu_t, \nu_1, \ldots, \nu_P)$ starting from a randomly initialized Gaussian μ_0. It is solved iteratively following Bonneel et al. [11].

Differentiability Sliced Wasserstein variants are desirable candidate losses for learning on probability spaces thanks to their smoothness properties. We show $\mathcal{SMW}$ inherits these properties.

Proposition 12. *Let $\mu_1, \ldots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$ be discrete measures with N atoms, which we gather into matrices $\{\mathbf{X}^{(p)}\}_{p=1}^P$ where $\mathbf{X}^{(p)} \in \mathbb{R}^{N\times d}$, and similarly define $\mu_\mathbf{X}$ with N atoms X. Assume $\mathbf{X}$ has distinct points. Then $\mathcal{SMW}^2$ is differentiable with gradient*

$$\nabla_X \mathcal{SMW}^2(\mu_1, \ldots, \mu_P, \mu_\mathbf{X}) = \beta_{P+1} \sum_{p=1}^{P} \beta_p \int_{S_{d-1}} \mathbf{X}_\theta - \left(\mathbf{X}_\theta^{(p)} \circ \sigma_{\mathbf{X}_\theta} \circ \sigma^{-1}_{\mathbf{X}_\theta^{(p)}}\right) d\Theta(\theta),$$

where $\sigma_\mathbf{X}$ is the permutation that sorts atoms of $\mathbf{X}$ and $\boldsymbol{X}_\theta \in \mathbb{R}^N$, such that $(\boldsymbol{X}_\theta)_i = \langle \boldsymbol{x}_i, \theta\rangle$.

Proposition 12 shows that $\mathcal{SMW}^2$ is differentiable almost everywhere, and is hence well-suited for multi-task learning, as it allows to compare multiple task-representative probability measures. We illustrate this in Figure 6.2. Here, we consider the problem $\min_\mu \mathcal{SMW}^2(\mu, \nu_1, \ldots, \nu_4)$, amounting to estimating the sliced barycenter of $\nu_1, \ldots, \nu_4$ (see Proposition 11), and solve it iteratively via the gradient flow $\partial\mu_t = -\nabla\mathcal{SMW}^2(\mu_t, \nu_1, \ldots, \nu_4)$, following Bonneel et al. [11] in the pairwise case.

6.4 Multi-Task Learning with Sliced Multi-marginal Optimal Transport

In the previous section, we proposed a multi-marginal metric between probability measures, which avoids exponential computational and statistical complexities and is thus practical for applications where a large number of samples N, number of measures P, or dimension d is of interest. $\mathcal{SMW}$ allows us to evaluate the closeness of probability measures $\mu_1, \ldots, \mu_P$, which makes it a good candidate regularizer in multi-task learning settings over probability

spaces, by encouraging shared global structure across tasks through closeness in sliced multi-marginal geometry. We now sketch potential areas of applications of $\mathcal{SMW}$ in the context of multi-task learning on spaces of probability measures, and illustrate examples in density estimation and multi-dynamics reinforcement learning.

Density Estimation with Shared Structure

Consider P target measures $\mu_1, \dots, \mu_P$, which we aim to approximate by parametric models $\nu_1, \dots, \nu_P$, such as for instance generative adversarial networks. In applications, it is often the case that these measures are affected by issues related to *distributional shift* [5], which prevents us from obtaining accurate empirical samples of $\mu_1, \dots, \mu_P$. One way to counteract these issues is to introduce a shared structure between the measures, which can be enforced through $\mathcal{SMW}$ regularization.

For example, consider empirical estimates $\hat{\mu}_1, \dots, \hat{\mu}_P$ of $\mu_1, \dots, \mu_P$, which are corrupted because no data is available in certain regions of each measure's support. Here, reconstruction of $\mu_1, \dots, \mu_P$ is only possible through the use of shared structure on the generative models $\nu_1, \dots, \nu_P$, which we can enforce by using $\mathcal{SMW}(\nu_1, \dots, \nu_P)$ as a regularizer. This results in the optimization problem

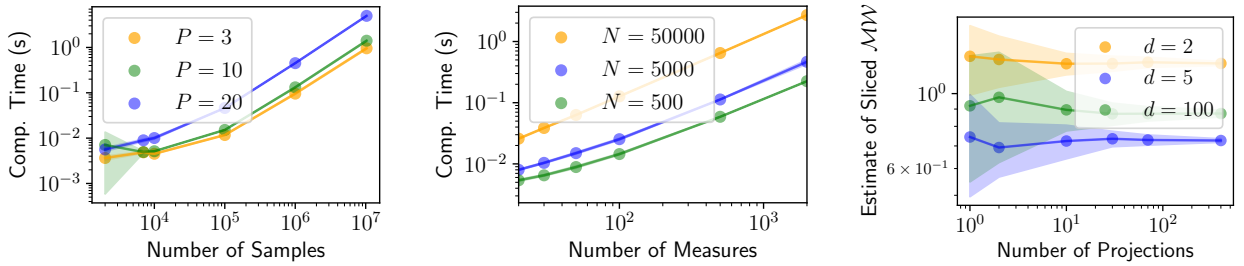
$$\arg\min_{\nu_1,\dots,\nu_P} \sum_{p=1}^{P} \underbrace{\mathcal{SW}^2(\mu_p, \nu_p)}_{\text{local loss}} + \gamma \underbrace{\mathcal{SMW}^2(\nu_1, \dots, \nu_P)}_{\text{global loss (shared)}},$$

where $\mathcal{SW}^2(\mu_p, \nu_p)$ ensures that the respective generative models $(\nu_p)_{p=1}^P$ approximates targets $(\mu_p)_{p=1}^P$, and $\mathcal{SMW}^2(\nu_1, \dots, \nu_P)$ ensures shared structure is present in the loss.

Multi-Dynamics Reinforcement Learning with Shared Structure

We now consider the problem of reinforcement learning in settings where the dynamics change. In order to speed up learning, we use $\mathcal{SMW}$ to share structure across different environments in this multi-dynamics reinforcement learning problem. Sharing knowledge is not only useful to bias (and thereby speed up) learning, but it is also useful in settings, where agents are ill informed, e.g., due to sparse reward signals. With a shared structure, these agents can learn from other agents. Here, the challenge is in effectively utilizing information from other agents in spite of differences in their respective environments. In the following, we focus on this setting.

Consider P identical-task agents in finite-horizon Markov decision processes $(\mathcal{S}, \mathcal{A}, \mathcal{T}_p, r_p^{\text{env}})$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space, both shared by all agents, $T_p(\mathbf{x}_t^{(p)}, \mathbf{a}_t^{(p)}) = \mathbf{x}_{t+1}^{(p)}$ is the transition model of agent p, which varies across agents, and r_p^{env} is the environment's reward function. Since different agents' tasks are identical, sharing structure can be beneficial. We consider the case, where some agents receive rewards $r_p^{\text{env}} = 0$. These agents are *uninformed* and can only learn via a shared structure that allows to transfer knowledge



(a) Computational time (log-log scale, mean ± standard deviation over 5 runs) for computing the sliced multi-marginal distance in seconds against the number of samples for various P.

(b) Computational time (log-log scale, mean ± standard deviation over 5 runs) for computing the sliced multi-marginal distance in seconds against the number of measures, $d = 10$ for various N.

(c) Mean $\mathcal{SMW}$ (± standard deviations) sliced multi-marginal distance against the number of projections for $P = 5$ measures with $N = 250$ samples.

Figure 6.3: Properties of the sliced multi-marginal distance. (a) computational time as a function of the number of samples; (b) computational time as a function of the number of measures; (c) accuracy as a function of the number of projections

from other agents. Structure sharing is done by augmenting the agent-specific reward function with a global multi-task reward term. In particular, define the augmented reward R_p as

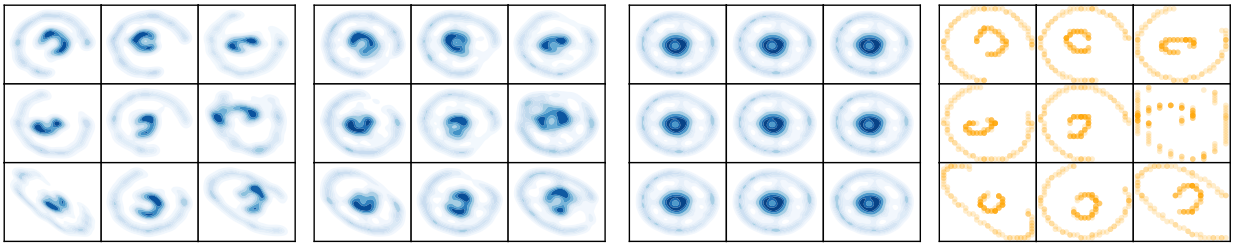
$$R_p(\mathbf{x}_t^{(p)}, \mathbf{X}) = \underbrace{r_p^{\text{env}}(\mathbf{x}_t^{(p)})}_{\substack{\text{agent specific} \\ \text{(local)}}} + \gamma \underbrace{r^{\text{mul}}(\mathbf{x}_t^{(p)}, \mathbf{X})}_{\substack{\text{multi-task reward} \\ \text{(shared/global)}}},$$

where $\mathbf{X} = \{\mathbf{x}_t^{(p)}\}_{p,t=1}^{P,T}$ is the collection of all states of every agent at all time steps, $r_p^{\text{env}}(\mathbf{x}_t^{(p)})$ is the single-task reward of the p^{th} environment and $r^{\text{mul}}(\mathbf{x}_t^{(p)}, \mathbf{X})$ is a (multi-task) reward signal. The former provides task-specific information about the task to be solved by agent p, while the latter allows for agents to share structure through the history of their state trajectories. If $r_p^{\text{env}} = 0$ for a given agent, then this agent can only learn through the shared structure arising from the shared reward r^{mul}. Finally, γ is a regularizer that controls the influence of shared structure on the overall learning.

We now describe the shared reward r^{mul}. Denote $\mu_p = \frac{1}{T}\sum_{t=1}^{T} \delta_{\mathbf{x}_t^{(p)}}$, which allows us to interpret the rollout of agent p as a discrete probability measure supported on the state space. Then,

$$r^{\text{mul}}(\mathbf{x}_t^{(p)}, \mathbf{X}) = -\frac{\beta_p}{K} \sum_{k=1}^{K} \left| \langle \mathbf{x}_t^{(p)} - \sum_{j=1}^{P} \beta_j \mathbf{x}_{\eta_{p,j,k}(t)}^{(j)}, \theta_k \rangle \right|^2,$$

where $\eta_{p,j,k}$ returns the index of the atom in μ_j that is aligned with state $\mathbf{x}_t^{(p)}$ after projecting on (Monte Carlo-sampled) (θ_k) and sorting all projected states. Intuitively, the reward signal attributed to the state $\mathbf{x}_t^{(p)}$ of agent p at time t is computed by projecting all measures onto



(a) $\gamma = 0$ (b) $\gamma = 0.3$ (c) $\gamma = 25$ (d) Corrupted Targets

Figure 6.4: Multi-task density estimation experiment applied on corrupted nested ellipses (plotted in orange), using $\mathcal{SW}^2$ as pairwise loss and $\mathcal{SMW}^2$ as regularizer. Learned models are plotted in blue. We use regularization coefficients $\gamma = 0$ in (a), $\gamma = 0.3$ in (b), $\gamma = 25$ in (c).

K vectors, gathering all states that are aligned with $\mathbf{x}_t^{(p)}$ for each projection θ_k, and summing squared distances between them.

Remark. *The barycentric cost structure with non-uniform weights β is particularly attractive in this setting, as it allows to give more weight to the communication arising from agents that perform well in their own environment. For instance, we can use Boltzmann weights*

$$\beta_p \propto \exp\Big(\alpha \sum_{t=1}^{T} r_p^{env}(\mathbf{x}_t^{(p)})\Big),$$

where α is a temperature. It gives more weight in the reward to agents performing best.

We train all agents simultaneously by maximizing

$$\mathbb{E}_{\pi_1,\dots,\pi_P}\Big[\sum_{p=1}^{P}\sum_{t=1}^{T} R_p(\mathbf{x}_t^{(p)})\Big] = \mathbb{E}_{\pi_1,\dots,\pi_P}\Big[\sum_{p=1}^{P}\sum_{t=1}^{T} \underbrace{r_p^{\text{env}}(\mathbf{x}_t^{(p)}) + \gamma \mathcal{SMW}^2(\mu_1,\dots,\mu_P)}_{=R_p(\mathbf{x}_t^{(p)},\mathbf{X})}\Big]$$

with respect to the parameters of policies π_p, $p = 1,\dots,P$. Note that the extra term in the augmented reward regularizes the objective via the sliced multi-marginal Wasserstein distance. $\mathcal{SMW}$ thus enforces closeness of agents' trajectories which allows to share structure across agents.

6.5 Experiments

We now illustrate the behavior of sliced multi-marginal transport in simple multi-task learning setups.

Scalability

Number of Samples (N). We study the impact of the number of samples on the computational time to compute the sliced multi-marginal distance in (6.3). In particular, we compute $\mathcal{SMW}$ between $P = 3, 10, 20$ measures in $\mathbb{R}^{10}$, $\mu_p \sim \mathcal{N}(m_p, \eta^2 I)$, where $p = 1, \ldots, P$ for a fixed number of projections $K = 10$. Figure 6.3(a) shows the $\mathcal{O}(N \log N)$ scaling of $\mathcal{SMW}$. This enables computation of multi-marginal distances with over 10^7 samples and a large number of measures.

Number of Measures (P). We now examine scaling with respect to the number of measures P. Figure 6.3(b) shows the time required to compute $\mathcal{SMW}$ against $N = 500, 5000, 50000$ measures. We observe the expected linear scaling of $\mathcal{SMW}$.

Number of Projections (K). Finally, we consider the impact of the number of projections on the estimation of $\mathcal{SMW}$ for dimensions $d = 2, 5, 20$. We set $N = 250$, and $P = 5$. Monte Carlo estimation is used to estimate $\mathcal{SMW}$. Figure 6.3(c) shows the expected variance shrinkage as the number of projection grows, while the estimated mean converges to $\mathcal{SMW}$ with rate $\mathcal{O}(\frac{1}{\sqrt{K}})$ and constant factors depending on dimension.

Multi-Task Density Estimation

We consider the multi-task density estimation setting of Section 6.4. Each target measures consist of a nested ellipse with corrupted samples. In particular, parts of each individual ellipse have been removed from each measure's support. Using the multi-task learning setup allows for sharing knowledge of the structure of the target tasks across problems—namely, that all target measures have the overall shape of nested ellipses. Figures 6.4(a)–6.4(c) show the models obtained by multi-task training with regularization coefficients $\gamma = 0,\ 0.3,\ 25$. When $\gamma = 0$, measures are learned individually without any structure sharing. $\nu_1, \ldots, \nu_P$ hence collapse to the corrupted measures $\mu_1, \ldots, \mu_P$. When structure is introduced ($\gamma > 0$) knowledge of the inherent nested ellipse structure is shared across tasks, which leads to solutions that have such structure (holes are filled), but that still preserve the task-specific orientations and ellipse width/height as long as the structure coefficient η is not too large. The latter causes the learned measures to be too close to each other. These effects can be seen in Figure 6.4(c). When this happens, all learned measures collapse to the barycenter.

Multi-Dynamics Reinforcement Learning

We consider a multi-task RL application in the setting of Section 6.4. In particular, we consider $P = 5$ pendulum swing-up tasks with different dynamics (gravities $g \in \{8, 9, 10, 11, 12\}$ m/s^2). States consist of angle and angular velocities, and actions of are torques. Environment rewards are dense as implemented in OpenAI Gym [13], and following Dadashi et al. [18], we transform the shared reward r^{mul} via $f(y) = e^{-5y}$. Two out of five agents do not receive any reward. All other agents share the same reward function. We consider agents trained with and without $\mathcal{SMW}$-based regularization, and consider the uniform and non-uniform

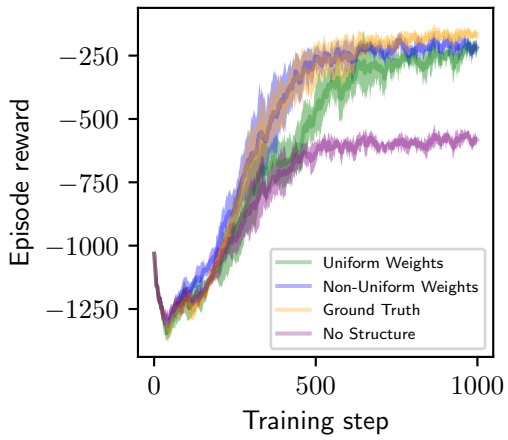


Figure 6.5: Multi-task ($P = 5$) RL experiment. Environments have different dynamics (different gravities), and 2/5 agents have no environmental reward. Without shared structure, these agents do not solve their respective tasks (orange). By contrast, with shared structure, all agents learn accurate policies (green, blue), on par with agents trained without corrupted rewards (blue). Training curves (mean ± standard deviation averaged over 5 runs) are shown.

barycentric weights β; see Section 6.4 for more details. To facilitate learning, we lower-bound the weights of agents without reward. Policies are learned using Q-learning with function approximation on state observations.

Figure 6.5 shows the results. Training without regularization ($\gamma = 0$, blue curve) does not allow the two agents without environment rewards ($r_p^{\text{env}} = 0$) to solve their respective tasks. By contrast, with regularization, all agents (even those with no environment reward) solve their respective tasks (green, blue) as well as if all agents were receiving environmental rewards (orange). Agents with non-uniform regularization significantly outperform agents with uniform weights, showing that giving more weight in the regularizer to stronger agents is helpful.

Overall, this demonstrates that knowledge transfer via the shared reward structure can be effective. In particular, the regularization-based rewards encourage the state trajectories of all agents to be close under the sliced multi-marginal geometry. Hence, agents without environment rewards learn to *follow* agents trained with environment rewards. This is possible because of similarity of environments and of agent goals, so that agent rollouts share geometric structure.

6.6 Conclusion

In this work, we proposed a scalable multi-marginal optimal transport distance. Our main idea is to derive a closed-form formula for multi-marginal optimal transport in 1D in the

general case and to extend it into a higher-dimensional metric via slicing. We show it is well-behaved topologically, and in particular that it is a generalized metric. We also show it is well-behaved statistically with dimension-free sample complexity (modulo a caveat arising from projection complexity). We derive a range of other results illustrating the simple and intuitive geometric structure of sliced multi-marginal transport. Finally, we propose areas of applications of sliced multi-marginal transport in the context of multi-task learning on probability spaces, and concrete instantiations in density estimation, and reinforcement learning. We hope these contributions enable practitioners in reinforcement learning, generative modeling and other areas to share structure across tasks in a geometrically-motivated way. Our work relies on the assumption that tasks live on the same space, and share structure. Future work would extend our approach to allow for multi-task learning on incomparable spaces, enabling structure sharing in more general set-ups, for instance via Gromov–Wasserstein-like techniques.

6.7 Proofs

Closed-form Formulas for Multimarginal Optimal Transport

Recall Definition 33, which states that for a measure $\mu \in \mathcal{M}(\mathbb{R})$, the CDF $C_\mu : \mathbb{R} \to [0,1]$ is

$$C_\mu(x) = \int_{-\infty}^{x} d\mu(y) \ \forall x.$$

And its pseudo-inverse $C_\mu^{-1} : [0,1] \to \mathbb{R} \cup \{-\infty\}$ is

$$C_\mu^{-1}(r) = \min_x \{x \in \mathbb{R} \cup \{-\infty\} : C_\mu(x) \geq r\}.$$

1D Multi-Marginal

Proposition 13. *If $\mu_1, \ldots, \mu_P \in \mathcal{M}(\mathbb{R})$ and $\mathcal{U}(0,1)$ is the uniform measure, then*

$$\mathcal{MW}^2(\mu_1, \ldots, \mu_P) = \int_0^1 \sum_{p=1}^{P} \beta_p \left| C_{\mu_p}^{-1}(x) - \sum_{j=1}^{P} \beta_j C_{\mu_j}^{-1}(x) \right|^2 dx,$$

and the optimal coupling solving (6.1) *is of the form*

$$\pi^\star = (C_{\mu_1}^{-1}, \ldots, C_{\mu_P}^{-1})_{\#} \, \mathcal{U}(0,1).$$

Proof. Our aim is to provide a closed form formula for

$$\mathcal{MW}^2(\mu_1, \ldots, \mu_P) = \min_{\pi \in \Pi(\mu_1, \ldots, \mu_P)} \int_{(\mathbb{R}^d)^P} \sum_{p=1}^{P} \beta_p |x_p - \sum_j \beta_j x_j|^2 \, d\pi(x_1, \ldots, x_P),$$

where $\Pi(\mu_1, \ldots, \mu_P)$ is the set of probability measures in $\mathcal{M}((\mathbb{R}^d)^P)$ with marginals $\mu_1, \ldots, \mu_P$.

First, notice

$$\int_{(\mathbb{R}^d)^P} \sum_{p=1}^{P} \beta_p \|x_p - \sum_j \beta_j x_j\|^2 d\pi(x_1, \ldots, x_P) = \sum_{p=1}^{P} \beta_p \int_{\mathbb{R}^d} |x_p|^2 d\mu_p - 2 \sum_{p,j=1}^{P} \beta_p \beta_j \int_{(\mathbb{R}^d)^2} x_p x_j \, d\pi_{pj}(x_p, x_j),$$

where π_{pj} corresponds to marginalizing π onto all components but p, j. This can be formalized by defining the map $L_{pj}(x_1, \ldots, x_P) = (x_p, x_j)$ and $\pi_{pj} = L_{pj\#}\pi$.

Now define $\pi^\star = (C_{\mu_1}^{-1}, \ldots, C_{\mu_P}^{-1})_{\#}\mathcal{U}(0,1)$.

Claim: $\pi^\star$ is optimal

First observe $L_{pj\#}\pi^\star = (C_{\mu_p}^{-1}, C_{\mu_j}^{-1})_\# \mathcal{U}(0,1)$ by marginalization. Note this is the optimal coupling between pairs μ_p, μ_j, see [36] (this can easily be obtained by observing that plugging in $(C_{\mu_p}^{-1}, C_{\mu_j}^{-1})_\# \mathcal{U}(0,1)$ into the Wasserstein objective achieves the minimum – it is also a valid coupling, thus it has to be the optimal coupling.)

Now, note that

$$\arg \max_{\gamma \in \Pi(\mu_p, \mu_j)} \int_{(\mathbb{R}^d)^2} x_p x_j \, d\gamma = \arg \min_{\gamma \in \Pi(\mu_p, \mu_j)} \int_{(\mathbb{R}^d)^2} |x_p - x_j|^2 \, d\gamma,$$

and also that for any multimarginal coupling $\pi \in \Pi(\mu_1, \dots, \mu_P)$, π_{pj} is a pairwise coupling in $\Pi(\mu_p, \mu_j)$ by the transfer lemma.

We can hence deduce that $\forall \pi \in \Pi(\mu_1, \dots, \mu_P)$

$$\int_{(\mathbb{R}^d)^2} x_p x_j d\pi_{pj} \leq \int_{(\mathbb{R}^d)^2} x_p x_j d\pi_{pj}^\star \quad \forall p, j = 1, \dots, P,$$

because both π_{pj} and $\pi_{pj}^\star$ are couplings of μ_p, μ_j and $\pi_{pj}^\star$ is optimal.

Therefore, it holds that

$$\begin{aligned}
\int_{(\mathbb{R}^d)^P} \sum_{p=1}^P \beta_p \|x_p - \sum_j \beta_j x_j\|^2 \, d\pi^\star(x_1, \dots, x_P) &= \sum_{p=1}^P \beta_p \int_{\mathbb{R}^d} |x_p|^2 d\mu_p - 2 \sum_{p,j=1}^P \beta_p \beta_j \int_{(\mathbb{R}^d)^2} x_p x_j \, d\pi_{pj}^\star(x_p, x_j) \\
&\leq \sum_{p=1}^P \beta_p \int_{\mathbb{R}^d} |x_p|^2 d\mu_p - 2 \sum_{p,j=1}^P \beta_p \beta_j \int_{(\mathbb{R}^d)^2} x_p x_j \, d\pi_{pj}(x_p, x_j) \\
&= \int_{(\mathbb{R}^d)^P} \sum_{p=1}^P \beta_p \|x_p - \sum_j \beta_j x_j\|^2 \, d\pi(x_1, \dots, x_P),
\end{aligned}$$

which proves the claim that $\pi^\star$ is the optimal multi-marginal coupling. We now compute the distance by plugging in the optimal coupling:

$$\begin{aligned}
\mathcal{MW}^2(\mu_1, \dots, \mu_P) &= \int_{(\mathbb{R}^d)^P} \sum_{p=1}^P \beta_p |x_p - \sum_j \beta_j x_j|^2 \, d\pi^\star(x_1, \dots, x_P) \\
&= \int_{(\mathbb{R}^d)^P} \sum_{p=1}^P \beta_p |x_p - \sum_j \beta_j x_j|^2 \, d(C_{\mu_1}^{-1}, \dots, C_{\mu_p}^{-1})_\# \mathcal{U}(0,1) \\
&= \int_0^1 \sum_{p=1}^P \beta_p |C_{\mu_p}^{-1}(x) - \sum_j \beta_j C_{\mu_j}^{-1}(x)|^2 \, dx.
\end{aligned}$$

□

Generalized Metric Properties

Definition 34. *Assume $\mu_p \in \mathcal{M}(\mathbb{R}^d)$, where $p = 1, \ldots, P$, and let $D : \mathcal{M}(\mathbb{R}^d) \times \ldots \times \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}$ be a multi-marginal Wasserstein metric with barycentric weights β. Then, D is a generalized metric if the following properties hold:*

1. $D(\mu_1, \ldots, \mu_P) \geq 0$
2. $D(\mu_1, \ldots, \mu_P) = 0 \Leftrightarrow \mu_1 = \ldots = \mu_P$
3. $D(\mu_1, \ldots, \mu_P) = D_\sigma(\mu_{\sigma(1)}, \ldots, \mu_{\sigma(P)})$, $\forall \sigma \in \mathbb{S}_P$ *where D_σ denotes that the barycentric weights β are permuted by σ and $\mathbb{S}_P$ is the group of permutations of order P.*
4. $\forall \mu \in \mathcal{M}(\mathbb{R}^d) : D(\mu_1, \ldots, \mu_P) \leq \sum_{p=1}^{P} D(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots \mu_P)$

Proposition 14. *$\mathcal{MW}$ is a generalized metric on the restriction $\mathcal{M}(\mathbb{R})$.*

Proof. Property (1), i.e., positivity is clear because

$$\mathcal{MW}^2(\mu_1, \ldots, \mu_P) = \int_0^1 \sum_{p=1}^{P} \beta_p \Big| C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) \Big|^2 dx \geq 0$$

Next, we prove property (2).
We begin by proving the forward implication ($\Rightarrow$).

$$\mathcal{MW}(\mu_1, \ldots, \mu_P) = 0 \tag{6.4}$$

$$\Rightarrow \Big(\int_0^1 \sum_{p=1}^{P} \beta_p \Big| C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) \Big|^2 dx \Big)^{\frac{1}{2}} = 0 \tag{6.5}$$

$$\Rightarrow \int_0^1 \sum_{p=1}^{P} \beta_p \Big| C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) \Big|^2 dx = 0 \tag{6.6}$$

$$\Rightarrow C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) = 0 \quad \forall p = 1, \ldots, P, \ \forall x \in [0, 1] \tag{6.7}$$

Now assume for contradiction that $\exists m, n, x : \quad C^{-1}_{\mu_m}(x) \neq C^{-1}_{\mu_n}(x)$, then:

$$C^{-1}_{\mu_m}(x) = \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x), \qquad C^{-1}_{\mu_n}(x) = \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) \tag{6.8}$$

$$\Leftrightarrow C^{-1}_{\mu_m}(x) - C^{-1}_{\mu_n}(x) = \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) = 0 \tag{6.9}$$

which is a contradiction, therefore $C^{-1}_{\mu_m}(x) = C^{-1}_{\mu_n}(x) \quad \forall m, n, x$, thus $\mu_1 = \ldots = \mu_P$

We continue by proving the backward implication ($\Leftarrow$).

If $\mu_1 = \ldots = \mu_P$, then $C^{-1}_{\mu_p}(x) = C^{-1}_{\mu_{p'}}(x) \quad \forall x,\ \forall p, p' = 1, \ldots, P$.

Therefore, $C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) = 0 \quad \forall p = 1, \ldots, P,\ \forall x \in [0, 1]$. Thus,

$$\mathcal{MW}(\mu_1, \ldots, \mu_P) = \Big(\int_0^1 \sum_{p=1}^{P} \beta_p |C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x)|^2 \, dx \Big)^{\frac{1}{2}} = 0. \tag{6.10}$$

We continue with permutation invariance (3),

$$\begin{aligned}
\mathcal{MW}(\mu_1, \ldots, \mu_P) &= \Big(\int_0^1 \sum_{p=1}^{P} \beta_p \Big| C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) \Big|^2 \, dx \Big)^{\frac{1}{2}} && (6.11)\\
&= \Big(\int_0^1 \sum_{p=1}^{P} \beta_p \Big| C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_{\sigma(j)} C^{-1}_{\mu_{\sigma(j)}}(x) \Big|^2 \, dx \Big)^{\frac{1}{2}} && (6.12)\\
&= \Big(\int_0^1 \sum_{p=1}^{P} \beta_{\sigma(p)} \Big| C^{-1}_{\mu_{\sigma(p)}}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_{\sigma(j)}}(x) \Big|^2 \, dx \Big)^{\frac{1}{2}} && (6.13)\\
&= \mathcal{MW}_\sigma(\mu_{\sigma(1)}, \ldots, \mu_{\sigma(P)}) && (6.14)
\end{aligned}$$

Equalities holds because sums are invariant under any permutation σ.

We finally prove the generalized triangle inequality (4). Note the slight abuse of notation that $p + 1$ component does not exist when $p = P$.

We begin by proving the case $P \geq 3$. Firstly, we rewrite the multi-marginal functional in the following way:

$$\begin{aligned}
\mathcal{MW}^2(\mu_1, \ldots, \mu_P) &= \sum_{p=1}^{P} \beta_p \int_0^1 \Big| C^{-1}_{\mu_p}(x) - \sum_{j=1}^{P} \beta_j C^{-1}_{\mu_j}(x) \Big|^2 \, dx && (6.15)\\
&= \frac{1}{2} \sum_{p,p'=1}^{P} \beta_p \beta_{p'} \int_0^1 \Big| C^{-1}_{\mu_p}(x) - C^{-1}_{\mu_{p'}}(x) \Big|^2 \, dx && (6.16)\\
&= \frac{1}{2} \sum_{p,p'=1}^{P} \beta_p \beta_{p'} \int_0^1 f^2_{p,p'}(x) \, dx && (6.17)
\end{aligned}$$

where $f_{p,p'}(x) = \Big| C^{-1}_{\mu_p}(x) - C^{-1}_{\mu_{p'}}(x) \Big|$. The results holds because

$$\sum_{m,n=1}^{P} \beta_m \beta_n |C^{-1}_{\mu_m}(x) - C^{-1}_{\mu_n}(x)|^2 = \sum_{m=1}^{P} \beta_m \Big| C^{-1}_{\mu_m}(x) - \sum_{n=1}^{P} \beta_n C^{-1}_{\mu_j}(x) \Big|^2, \tag{6.18}$$

which holds because

$$\sum_{m=1}^{P} \beta_m \Big| x_m - \sum_{n=1}^{P} \beta_n x_n \Big|^2 \tag{6.19}$$

$$= \sum_{m=1}^{P} \beta_m \Big[|x_m|^2 + |\sum_{n=1}^{P} \beta_n x_n|^2 - 2 \sum_{n=1}^{P} \beta_n x_m x_n \Big] \tag{6.20}$$

$$= \sum_{m=1}^{P} \beta_m |x_m|^2 + \sum_{m,n=1} \beta_m \beta_n x_m x_n - 2 \sum_{m,n=1}^{P} \beta_m \beta_n x_m x_n \tag{6.21}$$

$$= \sum_{m=1}^{P} \beta_m |x_m|^2 - \sum_{m,n=1} \beta_m \beta_n x_m x_n \tag{6.22}$$

$$= \sum_{m,n=1}^{P} \beta_m \beta_n |x_m|^2 - \sum_{m,n=1} \beta_m \beta_n x_m x_n \tag{6.23}$$

$$= \sum_{m,n=1}^{P} \beta_m \beta_n (\frac{1}{2}|x_m|^2 + \frac{1}{2}|x_n|^2 - x_m x_n) \tag{6.24}$$

$$= \frac{1}{2} \sum_{m,n=1}^{P} \beta_m \beta_n |x_m - x_n|^2. \tag{6.25}$$

Therefore, we have

$$\sum_{p=1}^{P} \mathcal{MW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P) = \frac{1}{2} \sum_{p=1}^{P} \sum_{m,n \neq p}^{P} \beta_m \beta_n \int_0^1 f_{n,m}^2(x) dx + C, \tag{6.26}$$

where $C > 0$.

We now show that $\int_0^1 \sum_{p=1}^{P} \sum_{m,n\neq p}^{P} \beta_m \beta_n f_{n,m}^2(x)\, dx \geq \sum_{p,p'=1}^{P} \beta_p \beta_{p'} \int_0^1 f_{p,p'}^2(x)\, dx$. This can be observed by noting that all $\int_0^1 f_{p,p'}^2(x)\, dx$ terms on the RHS appear on the LHS. Indeed, for any m', n', $\int_0^1 f_{m',n'}^2(x)\, dx$ appears in the $p \neq m', n'$ summation, which always holds for some p as $P \geq 3$.

Therefore, we have shown that

$$\mathcal{MW}^2(\mu_1, \ldots, \mu_P) \leq \sum_{p=1}^{P} \mathcal{MW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P) \tag{6.27}$$

Also,

$$\mathcal{MW}^2(\mu_1, \ldots, \mu_P) \leq \sum_{p=1}^{P} \mathcal{MW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P) \tag{6.28}$$

$$\Rightarrow \mathcal{MW}(\mu_1, \ldots, \mu_P) \leq \sqrt{\sum_{p=1}^{P} \mathcal{MW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P)} \tag{6.29}$$

$$\leq \sum_{p=1}^{P} \sqrt{\mathcal{MW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P)} \tag{6.30}$$

$$= \sum_{p=1}^{P} \mathcal{MW}(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P) \tag{6.31}$$

which proves the result. The case $P = 2$ has been proved via different approaches (e.g. [36]). □

Proposition 8. *$\mathcal{SMW}$ is a generalized metric on the restriction $\mathcal{M}(\mathbb{R}^d)$.*

Proof. Property (1) holds by definition due to positivity of $\mathcal{MW}$ on $\mathbb{R}$ and the definition of the sliced multi-marginal distance.

Property (2) is more delicate. We begin with the forward direction ($\Rightarrow$).

We extend the proof of Nadjahi et al. [31] to the multi-marginal case. Define Θ as the uniform distribution on S_{d-1}. Define 'for (Θ-almost-every) θ' as $\forall\Theta$-a-e-θ. Firstly, the following holds:

$$\mathcal{SMW}(\mu_1, \ldots, \mu_P) = 0 \tag{6.32}$$

$$\Rightarrow \Big(\frac{1}{\text{Vol}(S_{d-1})} \int_{S_{d-1}} \mathcal{MW}^2(M_{\theta\#}\mu_1, \ldots, M_{\theta\#}\mu_P) d\Theta(\theta)\Big)^{\frac{1}{2}} = 0 \tag{6.33}$$

$$\Rightarrow \mathcal{MW}(M_{\theta\#}\mu_1, \ldots, M_{\theta\#}\mu_P) \ \forall\Theta\text{-a-e-}\theta \tag{6.34}$$

$$\Rightarrow M_{\theta\#}\mu_1 = \ldots = M_{\theta\#}\mu_P \ \forall\Theta\text{-a-e-}\theta \tag{6.35}$$

Next, we define the Fourier transform of any measure μ on $\mathcal{M}(\mathbb{R}^s), s \geq 1$ at any $w \in \mathbb{R}^s$:

$$\mathcal{F}[\mu](w) = \int_{\mathbb{R}^s} e^{-i\langle w,x\rangle}\, d\mu(x). \tag{6.36}$$

Therefore, using properties of push-forwards, the following holds:

$$\mathcal{F}[M_{\theta\#}\mu](t) = \int_{\mathbb{R}} e^{-itu}\, dM_{\theta\#}\mu(u) = \int_{\mathbb{R}^s} e^{-it\langle\theta,x\rangle}\, d\mu(x) = \mathcal{F}[\mu](t\theta). \tag{6.37}$$

As $\forall\Theta$-a-e-θ, $M_{\theta\#}\mu_1 = \ldots = M_{\theta\#}\mu_P$, then $\mathcal{F}[M_{\theta\#}\mu_1] = \ldots = \mathcal{F}[M_{\theta\#}\mu_P]$, which implies that $\mathcal{F}[\mu_1] = \ldots = \mathcal{F}[\mu_P]$. By injectivity of the Fourier transform, we conclude that $\mu_1 = \ldots = \mu_P$.

We continue with the backward direction ($\Leftarrow$).
We assume $\mu_1 = \ldots = \mu_P$, which implies the following:

$$\mu_1 = \ldots = \mu_P \tag{6.38}$$
$$\Rightarrow M_{\theta\#}\mu_1 = \ldots = M_{\theta\#}\mu_P \quad \forall\Theta\text{-a-e-}\theta \tag{6.39}$$
$$\Rightarrow \mathcal{MW}^2(M_{\theta\#}\mu_1, \ldots, M_{\theta\#}\mu_P) = 0 \quad \forall\Theta\text{-a-e-}\theta \tag{6.40}$$
$$\Rightarrow \mathcal{SMW}(\mu_1, \ldots, \mu_P) = \Big(\frac{1}{\text{Vol}(S_{d-1})}\int_{S_{d-1}} \mathcal{MW}^2(M_{\theta\#}\mu_1, \ldots, M_{\theta\#}\mu_P)\, d\Theta(\theta)\Big)^{\frac{1}{2}} = 0. \tag{6.41}$$

We now prove Property (3)

$$\mathcal{SMW}(\mu_1, \ldots, \mu_P) = \Big(\frac{1}{\text{Vol}(S_{d-1})}\int_{S_{d-1}} \mathcal{MW}^2(M_{\theta\#}\mu_1, \ldots, M_{\theta\#}\mu_P)\, d\Theta(\theta)\Big)^{\frac{1}{2}} \tag{6.42}$$
$$= \Big(\frac{1}{\text{Vol}(S_{d-1})}\int_{S_{d-1}} \mathcal{MW}_\sigma^2(M_{\theta\#}\mu_{\sigma(1)}, \ldots, M_{\theta\#}\mu_{\sigma(P)}) d\Theta(\theta)\Big)^{\frac{1}{2}} \tag{6.43}$$
$$= \mathcal{SMW}_\sigma(\mu_{\sigma(1)}, \ldots, \mu_{\sigma(P)}) \tag{6.44}$$

We finally end by proving Property (4), the generalized triangle inequality.
Earlier, we showed that

$$\mathcal{MW}^2(\mu_1, \ldots, \mu_P) \leq \sum_{p=1}^{P} \mathcal{MW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P). \tag{6.45}$$

This implies that

$$\mathcal{SMW}^2(\mu_1, \ldots, \mu_P) \tag{6.46}$$
$$= \frac{1}{\text{Vol}(S_{d-1})}\int_{S_{d-1}} \mathcal{MW}^2(M_{\theta\#}\mu_1, \ldots, M_{\theta\#}\mu_P)\, d\Theta(\theta) \tag{6.47}$$
$$\leq \sum_{p=1}^{P} \frac{1}{\text{Vol}(S_{d-1})}\int_{S_{d-1}} \mathcal{MW}^2(M_{\theta\#}\mu_1, \ldots, M_{\theta\#}\mu_{p-1}, M_{\theta\#}\mu, M_{\theta\#}\mu_{p+1}, \ldots, M_{\theta\#}\mu_P)\, d\Theta(\theta) \tag{6.48}$$
$$= \sum_{p=1}^{P} \mathcal{SMW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P). \tag{6.49}$$

Therefore, we conclude that

$$\mathcal{SMW}^2(\mu_1, \ldots, \mu_P) \leq \sum_{p=1}^{P} \mathcal{SMW}^2(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P) \tag{6.50}$$
$$\Rightarrow \mathcal{SMW}(\mu_1, \ldots, \mu_P) \leq \sum_{p=1}^{P} \mathcal{SMW}(\mu_1, \ldots, \mu_{p-1}, \mu, \mu_{p+1}, \ldots, \mu_P) \tag{6.51}$$

directly in the same way as in the proof of Proposition the generalized triangle inequality for $\mathcal{MW}$. □

Mathematical Properties

Proposition 7.

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P) = \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \mathcal{SW}^2(\mu_i, \mu_j)$$

Proof.

$$\begin{aligned}
\mathcal{SMW}^2(\mu_1, \dots, \mu_P) &= \frac{1}{\text{Vol}(S_{d-1})} \int_{S_{d-1}} \int_{\mathbb{R}^d} \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j |x_i - x_j|^2 \, d\pi^{\star\theta}(x_1, \dots, x_P) \, d\Theta(\theta) \\
&= \frac{1}{2\text{Vol}(S_{d-1})} \sum_{i,j=1}^{P} \beta_i \beta_j \int_{S_{d-1}} \int_{\mathbb{R}\times\mathbb{R}} |x_i - x_j|^2 \, d\pi_{ij}^{\star\theta}(x_i, x_j) \, d\Theta(\theta) \\
&= \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \frac{1}{\text{Vol}(S_{d-1})} \int_{S_{d-1}} \mathcal{W}^2(M_{\theta\#}\mu_i, M_{\theta\#}\mu_j) \, d\Theta(\theta),
\end{aligned}$$

where $\pi^{\star\theta}$ is the optimal coupling between $M_{\theta\#}\mu_1, \dots, M_{\theta\#}\mu_P$ and $M_\theta(x) = \langle x, \theta \rangle$. Similarly to proofs of closed-form formulas for multi-marginal Kantorovich transport, we know that $\pi_{ij}^{\star\theta}$ is the optimal coupling between $M_{\theta\#}\mu_i, M_{\theta\#}\mu_j$. As a result, it holds that

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P) = \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \mathcal{SW}^2(\mu_i, \mu_j).$$

□

Corollary 6.7.0.1.

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P) \leq \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \mathcal{W}^2(\mu_i, \mu_j)$$

Proof. By Proposition 7, it holds that

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P) = \frac{1}{2} \sum_{i,j=1}^{P} \beta_i \beta_j \mathcal{SW}^2(\mu_i, \mu_j).$$

Also, Bonnotte [12] shows that

$$\mathcal{SW}^2(\mu, \nu) \leq \mathcal{W}^2(\mu, \nu) \quad \forall \mu, \nu.$$

The result follows directly. □

Sample/Projection Complexity

We now study $\mathrm{E}[\mathcal{SMW}^2(\mu_1,\dots,\mu_P) - \mathcal{SMW}^2(\hat{\mu}_1,\dots,\hat{\mu}_P)]^2$ where $\hat{\mu}_p$'s refers to empirical measures with n samples. Then the following result holds:

Proposition 15. *If $\mu_1,\dots,\mu_P \in \mathcal{M}(\mathbb{R}^d)$, and assuming $\mathcal{W}^2$ has sample complexity $\rho(N)$ on $\mathbb{R}$, then*

$$E[\mathcal{SMW}^2(\mu_1,\dots,\mu_P) - \mathcal{SMW}^2(\hat{\mu}_1,\dots,\hat{\mu}_P)]^2 \leq \frac{1}{2}\rho(N).$$

This result shows the sample complexity is dimension free.

Proof. We conclude from Proposition 7

$$\mathcal{SMW}^2(\mu_1,\dots,\mu_P) - \mathcal{SMW}^2(\hat{\mu}_1,\dots,\hat{\mu}_P) = \frac{1}{2}\sum_{i,j=1}^{P} \beta_i\beta_j \Big(\mathcal{SW}^2(\mu_i,\mu_j) - \mathcal{SW}^2(\hat{\mu}_i,\hat{\mu}_j)\Big).$$

If $\mathcal{W}^2$ on $\mathbb{R}$ has sample complexity $\rho(N)$, then $\mathcal{SW}^2$ on $\mathbb{R}^d$ also has sample complexity $\rho(N)$, i.e., its sample complexity is dimension free. The proof relies on an application of Jensen's inequality and is a special case of Nadjahi et al. [32].

$$\begin{aligned}
E\Big|\mathcal{SW}^2(\mu,\nu) - \mathcal{SW}^2(\hat{\mu}_n,\hat{\nu}_n)\Big| &= E\Big|\int_{S_{d-1}} \{\mathcal{W}^2(\theta^*_\#\mu,\theta^*_\#\nu) - \mathcal{W}^2(\theta^*_\#\hat{\mu}_n,\theta^*_\#\hat{\nu}_n)\}\, d\Theta(\theta)\Big| \\
&\leq E\left\{\int_{S_{d-1}} \Big|\mathcal{W}^2(\theta^*_\#\mu,\theta^*_\#\nu) - \mathcal{W}^2(\theta^*_\#\hat{\mu}_n,\theta^*_\#\hat{\nu}_n)\Big|\, d\Theta(\theta)\right\} \\
&\leq \int_{S_{d-1}} E\Big|\mathcal{W}^2(\theta^*_\#\mu,\theta^*_\#\nu) - \mathcal{W}^2(\theta^*_\#\hat{\mu}_n,\theta^*_\#\hat{\nu}_n)\Big|\, d\Theta(\theta) \\
&\leq \int_{S_{d-1}} \rho(N)\, d\Theta(\theta) = \rho(N)
\end{aligned}$$

Hence,

$$\begin{aligned}
E\Big|\mathcal{SMW}^2(\mu_1,\dots,\mu_P) - \mathcal{SMW}^2(\hat{\mu}_1,\dots,\hat{\mu}_P)\Big| &= E\Big|\frac{1}{2}\sum_{i,j=1}^{P} \beta_i\beta_j\Big(\mathcal{SW}^2(\mu_i,\mu_j) - \mathcal{SW}^2(\hat{\mu}_i,\hat{\mu}_j)\Big)\Big| \\
&\leq \frac{1}{2}\sum_{i,j=1}^{P} \beta_i\beta_j E\Big|\mathcal{SW}^2(\mu_i,\mu_j) - \mathcal{SW}^2(\hat{\mu}_i,\hat{\mu}_j)\Big| \\
&\leq \frac{1}{2}\sum_{i,j=1}^{P} \beta_i\beta_j\rho(N) = \frac{1}{2}\rho(N).
\end{aligned}$$

□

Here we also derive projection complexity results.

Proposition 10. *Let $\mu_1, \dots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$, and define $\overline{\mathcal{SMW}}$ the approximation obtained by uniformly picking L projections on S_{d-1}, then*

$$\mathrm{E}\left[\overline{\mathcal{SMW}}^2(\mu_1, \dots, \mu_P) - \mathcal{SMW}^2(\mu_1, \dots, \mu_P)\right]^2 \leq L^{-1/2}\mathrm{Var}_\theta\Big[\mathcal{MW}^2\big(\mu_1^\theta, \dots, \mu_P^\theta\big)\Big],$$

where θ follows the uniform distribution on S_{d-1} and $\mu_p^\theta = M_{\#}^\theta(\mu_p)$.

Proof. We bound the error arising from the Monte Carlo approximation of $\mathcal{SMW}$, similarly to Nadjahi et al. [32] in the pairwise case. In particular, define

$$\delta = \int_{S_{d-1}} \mathcal{MW}^2(M_{\theta\#}\mu_1, \dots, M_{\theta\#}\mu_P)\, d\Theta(\theta).$$

Then we have that

$$\begin{aligned}
&E_{\theta\sim\sigma}|\overline{\mathcal{SMW}}^2(\mu_1, \dots, \mu_P) - \mathcal{SMW}^2(\mu_1, \dots, \mu_P)| \\
&\qquad \leq \left\{E_{\theta\sim\sigma}|\overline{\mathcal{SMW}}^2(\mu_1, \dots, \mu_P) - \mathcal{SMW}^2(\mu_1, \dots, \mu_P)|^2\right\}^{\frac{1}{2}} \\
&\qquad \leq L^{-1/2}\int_{S_{d-1}} \Big\{\mathcal{MW}^2(M_{\theta\#}\mu_1, \dots, M_{\theta\#}\mu_P) - \delta\Big\}^2 d\Theta(\theta) \\
&\qquad = L^{-1/2}\mathrm{Var}_\theta\Big[\mathcal{MW}^2\big(\mu_1^\theta, \dots, \mu_P^\theta\big)\Big],
\end{aligned}$$

which holds due to the same Monte-Carlo concentration inequality as in the proof of Theorem 6 of Nadjahi et al. [32]. □

Equivalence to Sliced Barycenters and Weak Convergence

Proposition 11. *Let $\mu_1, \dots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$, $\sum_{p=1}^P \beta_p = 1$. Furthermore, let $\hat{\beta}_p$ be augmented multi-marginal weights, so that for $m \in [0, 1]$ it holds that $\hat{\beta}_p = m\beta_p$ for $p = 1, \dots, P$, $\sum_{p=1}^{P+1} \hat{\beta}_p = 1$, and $\mathcal{D} = \mathcal{SW}^2$. Then*

$$\operatorname*{arg\,min}_{\mu\in\mathcal{M}(\mathbb{R}^d)} \mathcal{SMW}^2(\mu_1, \dots, \mu_P, \mu) = \operatorname*{arg\,min}_{\mu\in\mathcal{M}(\mathbb{R}^d)} \mathcal{F}(\mu),$$

where β is the weight vector of $\mathcal{F}$ and $\hat{\beta}$ is the weight vector of $\mathcal{SMW}$.

Proof.

$$\begin{aligned}
\operatorname*{arg\,min}_{\mu\in\mathcal{M}(\mathbb{R}^d)} \mathcal{SMW}^2(\mu_1, \dots, \mu_P, \mu) &= \operatorname*{arg\,min}_{\mu\in\mathcal{M}(\mathbb{R}^d)} \sum_{p=1}^P \hat{\beta}_p\hat{\beta}_{P+1}\mathcal{SW}^2(\mu, \mu_p) \\
&= \operatorname*{arg\,min}_{\mu\in\mathcal{M}(\mathbb{R}^d)} \sum_{p=1}^P \beta_p\mathcal{SW}^2(\mu_p, \mu) \\
&= \operatorname*{arg\,min}_{\mu\in\mathcal{M}(\mathbb{R}^d)} \mathcal{F}(\mu).
\end{aligned}$$

□

Differentiability

Proposition 12. *Let $\mu_1, \dots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$ be discrete measures with N atoms, which we gather into matrices $\{\boldsymbol{X}^{(p)}\}_{p=1}^P$, and similarly define $\mu_{\boldsymbol{X}}$ with atoms $\boldsymbol{X}$. Assume $\boldsymbol{X}$ has distinct points. Then $\mathcal{SMW}^2$ is differentiable with gradient*

$$\nabla_{\boldsymbol{X}} \mathcal{SMW}^2(\mu_1, \dots, \mu_P, \mu_{\boldsymbol{X}}) = \beta_{P+1} \sum_{p=1}^P \beta_p \int_{S_{d-1}} \boldsymbol{X}_\theta - \left(\boldsymbol{X}_\theta^{(p)} \circ \sigma_{\boldsymbol{X}_\theta} \circ \sigma^{-1}_{\boldsymbol{X}_\theta^{(p)}}\right) d\Theta(\theta),$$

where σ_X is the permutation that sorts atoms of $\boldsymbol{X}$.

Proof. Define $\sigma_{\mathbf{Y}}$ be the permutation of $\{1, \dots, N\}$ that sorts atoms of $\mathbf{Y}$. Also, define $\mathbf{X}_\theta \in \mathbb{R}^N$, such that $(\mathbf{X}_\theta)_i = \langle \mathbf{x}_i, \theta \rangle$.

Then

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P, \mu_{\mathbf{X}}) = \sum_{p=1}^P \beta_{P+1}\beta_p \mathcal{SW}^2(\mu_{\mathbf{X}}, \mu_p) + C(\mu_1, \dots, \mu_P).$$

Hence,

$$\begin{aligned}
\nabla_{\mathbf{X}} \mathcal{SMW}^2(\mu_1, \dots, \mu_P, \mu_{\mathbf{X}}) &= \nabla_{\mathbf{X}} \sum_{p=1}^P \beta_{P+1}\beta_p \mathcal{SW}^2(\mu_{\mathbf{X}}, \mu_p) \\
&= \sum_{p=1}^P \beta_{P+1}\beta_p \int_{S_{d-1}} \mathbf{X}_\theta - \mathbf{X}_\theta^{(p)} \circ \left(\sigma_{\mathbf{X}_\theta} \circ \sigma^{-1}_{\mathbf{X}_\theta^{(p)}}\right) d\theta.
\end{aligned}$$

The last equality is due to Bonneel et al. [11]. □

6.8 Additional Experimental Details

We now provide further experimental details. All experiments ran on CPU, besides the benchmarking experiments, which ran on a single P100 GPU.

Ellipses - Multi-Task Density Estimation

We set the batch size to 150, and parametrize each measure ν_p as a discrete measure with 150 atoms which we optimize over via stochastic gradient descent. We set the number of projections to 20.

Multi-Task Reinforcement Learning

The horizon is set to $T = 200$. The learning rate is set to 2.5×10^{-4}, and the batch size to optimize the Q-function to 32. The Q-network is a 2-layer MLP with tanh activation. We use $f(x) = e^{-5x}$ to rescale the reward function following Dadashi et al. [18], we set the number of projections to $K = 50$ and $\gamma = 1$. Also, we set $\alpha = \frac{1}{30}$. Our implementation extends the repository `https://github.com/xtma/simple-pytorch-rl` to the multi-task setting, and leverages OpenAI gym environments [13].

Gradient Flow experiment

We follow the setup of Bonneel et al. [11]. In particular, we discretize the flow to numerically estimate it via gradient descent $X^{(l+1)} = X^{(l+1)} - \nabla \mathcal{SMW}^2(\mu_1, \ldots, \mu_P, \mu_{X^{(l)}})$, and plot the location of particles for $l = 0, \ldots, T$ where T is the number of steps (200), which approximates the gradient flow. We estimate $\mathcal{SMW}$ with 30 projections. Each measure (including the initial measure μ_0) consists of samples from isotropic Gaussians, and the initial measure.

Chapter 7

Appendix

In the following, we provide some sample code that demonstrates 3.7 empirically.

7.1 Example Code for 3.7

```
import cvxpy as cp
import numpy as np

n = 2
#b = 10

PP = cp.Variable((n,n),"PP")
KK = [[4,1],[1,4]]
#s = np.array([[.5, .5]]).T
#t = np.array([[.2,.8]]).T
s = np.array([[.3, .7]]).T
t = np.array([[.2,.8]]).T
e = np.ones((n,1))
x = PP.T@e - s
y = PP@e - t
for b in range(1,21):
obj = (1/4/b) * (cp.quad_form(x,KK) +
cp.quad_form(y,KK)) - cp.trace(KK@PP)
prob = cp.Problem(cp.Minimize(obj),[PP>=0,cp.sum(PP)==1])
obj=prob.solve()
print("status:",prob.status)
print("obj:",obj)
print(PP.value)
```

```
n = 3
PP = cp.Variable((n,n),"PP")
KK = [[1,0,0],[0,1,0],[0,0,1]]
s = np.array([[.1, .4, .5]]).T
t = np.array([[.4, .2, .4]]).T
e = np.ones((n,1))
x = PP.T@e - s
y = PP@e - t
for b in range(1,21):
obj = (1/4/b) * (cp.quad_form(x,KK) +
cp.quad_form(y,KK)) - cp.trace(KK@PP)
prob = cp.Problem(cp.Minimize(obj),[PP>=0,cp.sum(PP)==1])
obj=prob.solve()
print("status:",prob.status)
print("obj:",obj)
print(PP.value)
```

Output after running on Ubuntu machine.

```
yannik@yannik-ubuntu:~/OTDA$ python optimization_implementation.py
status: optimal
obj: -3.9925
[[ 2.50000000e-01  1.22249411e-23]
[-1.23247236e-22  7.50000000e-01]]
status: optimal
obj: -3.99625
[[ 2.50000000e-01 -1.74316142e-22]
[ 6.32939582e-23  7.50000000e-01]]
status: optimal
obj: -3.9975
[[ 2.50000000e-01 -1.16215745e-22]
[-2.16851043e-22  7.50000000e-01]]
status: optimal
obj: -3.998125
[[2.50000000e-01 5.50834936e-23]
[5.59387059e-23 7.50000000e-01]]
status: optimal
obj: -3.9985
[[2.50000000e-01 5.92447828e-23]
[1.62799830e-22 7.50000000e-01]]
status: optimal
obj: -3.99875
```

```
[[ 2.50000000e-01 -1.05815269e-22]
[-1.16229217e-22  7.50000000e-01]]
status: optimal
obj: -3.9989285714285714
[[ 2.50000000e-01 -1.91865798e-24]
[ 1.12940857e-22  7.50000000e-01]]
status: optimal
obj: -3.9990624999999995
[[2.50000000e-01 2.21829294e-22]
[1.11237621e-22 7.50000000e-01]]
status: optimal
obj: -3.9991666666666665
[[2.50000000e-01 1.68413892e-22]
[5.36304987e-23 7.50000000e-01]]
status: optimal
obj: -3.99925
[[ 2.50000000e-01 -1.11021317e-22]
[-1.11023280e-22  7.50000000e-01]]
status: optimal
obj: -3.999318181818182
[[ 2.50000000e-01 -1.66641110e-22]
[-5.54035982e-23  7.50000000e-01]]
status: optimal
obj: -3.999375
[[ 2.50000000e-01 -1.11238505e-22]
[ 2.16099333e-25  7.50000000e-01]]
status: optimal
obj: -3.999423076923077
[[ 2.50000000e-01 -1.11130073e-22]
[ 1.07889446e-25  7.50000000e-01]]
status: optimal
obj: -3.9994642857142857
[[ 2.50000000e-01 -5.66878028e-23]
[ 5.66878108e-23  7.50000000e-01]]
status: optimal
obj: -3.9995
[[ 2.50000000e-01  1.12085206e-22]
[-1.06311748e-24  7.50000000e-01]]
status: optimal
obj: -3.9995312499999995
[[2.50000000e-01 5.51862768e-23]
[5.58360336e-23 7.50000000e-01]]
```

```
status: optimal
obj: -3.9995588235294117
[[2.50000000e-01 5.46823694e-23]
[5.63399411e-23 7.50000000e-01]]
status: optimal
obj: -3.9995833333333333
[[ 2.50000000e-01 -1.11130414e-22]
[-1.10914183e-22  7.50000000e-01]]
status: optimal
obj: -3.999605263157895
[[ 2.50000000e-01 -5.52883039e-23]
[-1.66756293e-22  7.50000000e-01]]
status: optimal
obj: -3.999625
[[ 2.50000000e-01 -1.10718444e-22]
[-3.03850898e-25  7.50000000e-01]]
status: optimal
obj: -0.9825
[[ 2.50000000e-01  1.10709851e-22 -1.11209797e-22]
[ 2.22356962e-22  3.00000000e-01 -1.10897391e-22]
[-1.10834732e-22 -1.11147135e-22  4.50000000e-01]]
status: optimal
obj: -0.99125
[[ 2.50000000e-01 -1.18086022e-22  5.54360229e-23]
[-1.03958191e-22  3.00000000e-01 -4.85221266e-23]
[ 5.55859871e-23 -6.24999931e-23  4.50000000e-01]]
status: optimal
obj: -0.9941666666666666
[[ 2.50000000e-01  1.67542279e-24 -9.02920747e-25]
[-1.67529689e-24  3.00000000e-01 -2.57830146e-24]
[ 2.22947625e-22  2.57835866e-24  4.50000000e-01]]
status: optimal
obj: -0.995625
[[ 2.50000000e-01  1.07518204e-22 -6.07356262e-23]
[ 3.36570089e-22  3.00000000e-01 -5.72319709e-23]
[ 1.71758402e-22 -5.37898275e-23  4.50000000e-01]]
status: optimal
obj: -0.9965
[[ 2.50000000e-01  3.62041633e-24 -2.23532165e-22]
[-2.25665827e-22  3.00000000e-01 -1.16130663e-22]
[-1.09534272e-22 -2.16935737e-22  4.50000000e-01]]
status: optimal
```

```
obj: -0.9970833333333333
[[ 2.50000000e-01 -3.04602898e-25  1.08844672e-22]
[ 2.22349078e-22  3.00000000e-01  2.20171516e-22]
[ 2.24222382e-22  1.12895546e-22  4.50000000e-01]]
status: optimal
obj: -0.9975
[[ 2.50000000e-01  1.67389178e-22  5.46058619e-23]
[ 5.46549226e-23  3.00000000e-01 -1.76082074e-24]
[ 2.78460827e-22  1.76149241e-24  4.50000000e-01]]
status: optimal
obj: -0.9978125
[[ 2.50000000e-01  1.12313616e-22 -1.11893932e-22]
[ 1.09729020e-22  3.00000000e-01 -2.16476158e-24]
[-1.10148658e-22  2.16510806e-24  4.50000000e-01]]
status: optimal
obj: -0.9980555555555556
[[ 2.50000000e-01  1.53521320e-24  5.55069310e-23]
[-1.53498217e-24  3.00000000e-01  5.39719504e-23]
[ 5.55152244e-23  5.70504391e-23  4.50000000e-01]]
status: optimal
obj: -0.99825
[[ 2.50000000e-01 -5.78584577e-23  2.20132019e-22]
[-5.31633945e-23  3.00000000e-01  1.66970182e-22]
[ 1.12932966e-22  5.50760656e-23  4.50000000e-01]]
status: optimal
obj: -0.9984090909090909
[[ 2.50000000e-01 -1.10159194e-22  5.78533683e-23]
[-1.11884677e-22  3.00000000e-01 -5.40302524e-23]
[ 5.31672297e-23 -5.69909081e-23  4.50000000e-01]]
status: optimal
obj: -0.9985416666666667
[[ 2.50000000e-01  1.58478915e-24 -2.90879011e-25]
[-1.58779679e-24  3.00000000e-01 -1.12899214e-22]
[ 1.11314400e-22 -1.09143652e-22  4.50000000e-01]]
status: optimal
obj: -0.9986538461538461
[[ 2.50000000e-01  1.29692731e-24 -1.09306083e-22]
[-1.12320800e-22  3.00000000e-01 -1.10603467e-22]
[-3.34782807e-22 -1.11440161e-22  4.50000000e-01]]
status: optimal
obj: -0.99875
[[ 2.50000000e-01  3.05960762e-25 -1.10047970e-22]
```

```
[-3.07807840e-25   3.00000000e-01   6.67870216e-25]
[-9.73800980e-25  -1.11688797e-22   4.50000000e-01]]
status: optimal
obj: -0.9988333333333334
[[ 2.50000000e-01 -2.20472007e-22 -1.10265620e-22]
[-4.45663084e-22   3.00000000e-01 -2.22859007e-22]
[ 1.10264081e-22 -1.10205137e-22   4.50000000e-01]]
status: optimal
obj: -0.99890625
[[2.50000000e-01 1.67831187e-22 2.23235717e-22]
[2.76255783e-22 3.00000000e-01 5.54049384e-23]
[1.09830812e-22 5.56200424e-23 4.50000000e-01]]
status: optimal
obj: -0.9989705882352942
[[ 2.50000000e-01   2.33062237e-24   1.12105844e-22]
[-2.33398078e-24   3.00000000e-01  -1.24802428e-24]
[-1.08251257e-24   1.25052514e-24   4.50000000e-01]]
status: optimal
obj: -0.9990277777777777
[[ 2.50000000e-01 -5.37805096e-23   1.66737466e-22]
[ 5.37758046e-23   3.00000000e-01  -1.52711421e-24]
[ 5.53077093e-23   1.12553722e-22   4.50000000e-01]]
status: optimal
obj: -0.9990789473684211
[[ 2.50000000e-01 -5.39169018e-23   5.52124567e-23]
[-5.71093712e-23   3.00000000e-01  -1.12913039e-22]
[ 5.58078346e-23  -2.20147494e-22   4.50000000e-01]]
status: optimal
obj: -0.999125
[[ 2.50000000e-01   5.63739603e-23   1.67291982e-22]
[ 5.46420535e-23   3.00000000e-01  -1.05069787e-25]
[-1.67291107e-22   1.11132958e-22   4.50000000e-01]]
```

7.2 Differentiating Bures Distance

Differentiating the Bures distance part I

In the derivation, the function

$$\mathrm{Sym}(M) = \tfrac{1}{2}(M + M^T)$$

is utilized, as well as the trace/Frobenius product

$$P : M = \mathrm{Tr}(P^T M) = \mathrm{Tr}(M^T P) = M : P$$

These have the following interaction

$$P : \mathrm{Sym}(M) = \mathrm{Sym}(P) : M$$

For PSD matrices a drastic simplification is possible:

$$\mathrm{Tr}((A^{1/2} B A^{1/2})^{1/2}) = \mathrm{Tr}((BA)^{1/2})$$

In addition, there is a general result for the differential of the trace of any matrix function

$$d\,\mathrm{Tr}\big(f(X)\big) = f'(X^T) : dX$$

where f' is the ordinary derivative of the scalar function f; both f and f' are evaluated using their respective matrix arguments.

Combining these yields a straightforward solution for the problematic term

$$\begin{aligned}
\phi &= \mathrm{Tr}\Big((BA)^{1/2}\Big) \\
d\phi &= \tfrac{1}{2}\big((BA)^T\big)^{-1/2} : d(BA) \\
&= \tfrac{1}{2}(AB)^{-1/2} : B\,dA \\
&= \tfrac{1}{2}B(AB)^{-1/2} : dA \\
\frac{\partial \phi}{\partial A} &= \tfrac{1}{2}B(AB)^{-1/2} \;=\; \tfrac{1}{2}(BA)^{-1/2}B
\end{aligned}$$

Where the final equality is a theorem due to [25]

$$B \cdot f(AB) = f(BA) \cdot B$$

Therefore the gradient of the Bures Distance is

$$\begin{aligned}
\beta(A,B) &= \mathrm{Tr}\Big(A + B - 2(BA)^{1/2}\Big) \\
d\beta &= \Big(I - B(AB)^{-1/2}\Big) : dA \\
\frac{\partial \beta}{\partial A} &= I - B(AB)^{-1/2} \;=\; I - (BA)^{-1/2}B \\
&= I - A^{-1}(AB)^{1/2} \;=\; I - (BA)^{1/2}A^{-1}
\end{aligned}$$

Differentiating the Bures distance part II

Let J be the all-ones matrix and

$$\begin{aligned} C &= (I - \tfrac{1}{n}J) = C^T \qquad \big(\text{ Centering Matrix}\big) \\ B &= \Sigma_v \\ A &= \mathrm{Cov}(TX) \\ &= \big(\tfrac{1}{n-1}\big)\,(TX)^T C\,(TX) \end{aligned}$$

From earlier, the Bures distance function and its differential can be simplified to

$$\begin{aligned} \beta(A,B) &= \mathrm{Tr}\Big(A + B - 2(BA)^{1/2}\Big) \\ d\beta &= \Big(I - (BA)^{-1/2}B\Big) : dA \end{aligned}$$

Now change the differentiation variable from $dA \to dT$.

$$\begin{aligned} d\beta &= \Big(I - (BA)^{-1/2}B\Big) : \big(\tfrac{2}{n-1}\big)\ \mathrm{Sym}(X^T T^T C\, dT\, X) \\ &= \big(\tfrac{2}{n-1}\big)\Big(I - (BA)^{-1/2}B\Big) : (X^T T^T C\, dT\, X) \\ &= \big(\tfrac{2}{n-1}\big)\, CTX\Big(I - (BA)^{-1/2}B\Big)X^T : dT \\ \frac{\partial \beta}{\partial T} &= \big(\tfrac{2}{n-1}\big)\, CTX\Big(I - (BA)^{-1/2}B\Big)X^T \end{aligned}$$

7.3 M-Estimator Asymptotics

These theorems and proofs are adapted from [49] for the reader's convenience.

Theorem 7.3.1 (Consistency of M-Estimators)**.** *Let Θ be a subset of the real line and let Ψ_n be random functions and Ψ a fixed function of θ such that $\Psi_n(\theta) \to \Psi(\theta)$ in probability for every θ. Assume that each map $\theta \mapsto \Psi_n(\theta)$ is continuous and has exactly one zero $\hat{\theta}_n$, or is nondecreasing with $\Psi_n\left(\hat{\theta}_n\right) = o_P(1)$. Let θ_0 be a point such that $\Psi(\theta_0 - \varepsilon) < 0 < \Psi(\theta_0 + \varepsilon)$ for every $\varepsilon > 0$. Then $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$.*

Proof. If the map $\theta \mapsto \Psi_n(\theta)$ is continuous and has a unique zero at $\hat{\theta}_n$, then

$$\mathrm{P}\left(\Psi_n(\theta_0 - \varepsilon) < 0, \Psi_n(\theta_0 + \varepsilon) > 0\right) \le \mathrm{P}\Big(\theta_0 - \varepsilon < \hat{\theta}_n < \theta_0 + \varepsilon\Big)$$

The left side converges to one, because $\Psi_n(\theta_0 \pm \varepsilon) \to \Psi(\theta_0 \pm \varepsilon)$ in probability. Thus the right side converges to one as well, and $\hat{\theta}_n$ is consistent.

If the map $\theta \mapsto \Psi_n(\theta)$ is nondecreasing and $\hat{\theta}_n$ is a zero, then the same argument is valid. More generally, if $\theta \mapsto \Psi_n(\theta)$ is nondecreasing, then $\Psi_n(\theta_0 - \varepsilon) < -\eta$ and $\hat{\theta}_n \leq \theta_0 - \varepsilon$ imply $\Psi_n\left(\hat{\theta}_n\right) < -\eta$, which has probability tending to zero for every $\eta > 0$ if $\hat{\theta}_n$ is a near zero. This and a similar argument applied to the right tail shows that, for every $\varepsilon, \eta > 0$,

$$\mathrm{P}\left(\Psi_n(\theta_0 - \varepsilon) < -\eta, \Psi_n(\theta_0 + \varepsilon) > \eta\right) \leq \mathrm{P}\left(\theta_0 - \varepsilon < \hat{\theta}_n < \theta_0 + \varepsilon\right) + o(1)$$

For 2η equal to the smallest of the numbers $-\Psi(\theta_0 - \varepsilon)$ and $\Psi(\theta_0 + \varepsilon)$ the left side still converges to one. □

Suppose a sequence of estimators $\hat{\theta}_n$ is consistent for a parameter θ that ranges over an open subset of a Euclidean space. The next question of interest concerns the order at which the discrepancy $\hat{\theta}_n - \theta$ converges to zero. We now derive the asymptotic normality of M-estimators.

We can use a characterization of M-estimators by solving estimating equations.

Theorem 7.3.2 (Asymptotic Normality of M-Estimators)**.** *Let $X_1, \ldots, X_n$ be a sample from some distribution P, and let a random and a "true" criterion function be of the form:*

$$\Psi_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = \mathbb{E}_n \psi_\theta, \quad \Psi(\theta) = E\psi_\theta$$

Assume that the estimator $\hat{\theta}_n$ is a zero of Ψ_n and converges in probability to a zero θ_0 of Ψ. Also assume $\Psi_n\left(\tilde{\theta}_n\right)$ is $O_P(1)$.

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{D}} N\left(0, \frac{E\psi_{\theta_0}^2}{\left(E\psi'_{\theta_0}\right)^2}\right).$$

Proof. Because $\hat{\theta}_n \to \theta_0$, it makes sense to expand $\Psi_n\left(\hat{\theta}_n\right)$ in a Taylor series around θ_0. Assume for simplicity that θ is one-dimensional. Then

$$0 = \Psi_n\left(\hat{\theta}_n\right) = \Psi_n(\theta_0) + \left(\hat{\theta}_n - \theta_0\right)\Psi'_n(\theta_0) + \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)^2 \Psi''_n\left(\tilde{\theta}_n\right)$$

where $\tilde{\theta}_n$ is a point between $\hat{\theta}_n$ and θ_0. This can be rewritten as

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\Psi'_n(\theta_0) + \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)\Psi''_n\left(\tilde{\theta}_n\right)}.$$

If $E\psi_{\theta_0}^2$ is finite, then the numerator $-\sqrt{n}\Psi_n(\theta_0) = -n^{-1/2}\sum \psi_{\theta_0}(X_i)$ is asymptotically normal by the central limit theorem. The asymptotic mean and variance are $E\psi_{\theta_0} = \Psi(\theta_0) = 0$ and $E\psi_{\theta_0}^2$, respectively. Next consider the denominator. The first term $\Psi'_n(\theta_0)$ is an average

and can be analyzed by the law of large numbers: $\Psi'_n(\theta_0) \xrightarrow{\mathrm{P}} E\psi'_{\theta_0}$, provided the expectation exists. The second term in the denominator is a product of $\hat{\theta}_n - \theta = o_P(1)$ and $\psi''_n\left(\tilde{\theta}_n\right)$ and converges in probability to zero. Together with Slutsky's lemma, these observations yield

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{D}} N\left(0, \frac{E\psi^2_{\theta_0}}{\left(E\psi'_{\theta_0}\right)^2}\right).$$

□

Lemma 7.3.3 (Slutsky)**.** *Let X_n, X and Y_n be random vectors or variables. If $X_n \xrightarrow{\mathrm{D}} X$ and $Y_n \xrightarrow{\mathrm{P}} c$ for a constant c, then*

- $X_n + Y_n \xrightarrow{\mathrm{D}} X + c$;
- $Y_n X_n \xrightarrow{\mathrm{D}} cX$;
- $Y_n^{-1} X_n \xrightarrow{\mathrm{D}} c^{-1}X$ *provided* $c \neq 0$.

7.4 Rosenthal-type Inequality

This is also in [38]. Let $\mathcal{X}$ denote the class of all finite sequences $\mathbf{X} = (X_1, \ldots, X_n)$ of independent zero-mean random variables (r.v.'s). For any $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}$, let $S_{\mathbf{X}} := X_1 + \cdots + X_n$. Take any real number $p > 2$ and any positive real numbers A and B. Consider

$$\mathcal{X}_{p;A,B} := \left\{\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X} : \sum_1^n \mathrm{E}X_i^2 = B, \sum_1^n \mathrm{E}\left|X_i\right|^p = A\right\}$$

,

$$\mathcal{X}_{p;\leq A,\leq B} := \left\{\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X} : \sum_1^n \mathrm{E}X_i^2 \leq B, \sum_1^n \mathrm{E}\left|X_i\right|^p \leq A\right\}$$

$$\mathcal{X}_{p;X;A,B} := \{\mathbf{X} \in \mathcal{X}_{p;A,B} : \mathbf{X} \text{ is independent of } X\}$$

$$\mathcal{X}_{p;X;\leq A,\leq B} := \{\mathbf{X} \in \mathcal{X}_{p;\leq A,\leq B} : \mathbf{X} \text{ is independent of } X\}.$$

Theorem 7.4.1. *Suppose that $p \in (2, 3]$ and $\mathrm{E}|X|^p < \infty$. Then*

$$\begin{aligned}\sup_{\mathbf{X} \in \mathcal{X}_{p;X:\leq A,\leq B}} \mathrm{E}\left|X + S_{\mathbf{X}}\right|^p &= \sup_{\mathbf{X} \in \mathcal{X}_{p;X;A,B}} \mathrm{E}\left|X + S_{\mathbf{X}}\right|^p \\ &= A + \mathrm{E}\left|X + B^{1/2}Z\right|^p.\end{aligned}$$

Here, $Z \sim N(0, 1)$.

Bibliography

[1] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein Space." In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[2] Jason Altschuler and Enric Boix-Adsera. "Wasserstein Barycenters are NP-hard to Compute". In: *arXiv:2101.01100* (2021).

[3] Jason M. Altschuler and Enric Boix-Adserà. "Hardness results for Multimarginal Optimal Transport problems". In: *arXiv:2012.05398* (2020).

[4] Jason M. Altschuler and Enric Boix-Adserà. "Polynomial-time Algorithms for Multimarginal Optimal Transport Problems with Structure". In: *arXiv:2008.03006* (2020).

[5] Dario Amodei et al. "Concrete Problems in AI Safety". In: *arXiv:1606.06565* (2016).

[6] Shai Ben-David et al. "Analysis of representations for domain adaptation". In: *Advances in Neural Information Processing Systems* (2007), pp. 137–144. ISSN: 10495258. DOI: 10.7551/mitpress/7503.003.0022.

[7] Jean-David Benamou et al. "Iterative Bregman Projections for Regularized Transportation Problems". In: *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[8] José Bento and Liang Mi. "Multi-Marginal Optimal Transport Defines a Generalized Metric". In: *arXiv:2001.11114* (2020).

[9] Rajendra Bhatia, T. Jain, and Yongdo Lim. "On the Bures–Wasserstein distance between positive definite matrices". In: *Expositiones Mathematicae* 37.2 (2019), pp. 165–191. ISSN: 07230869. DOI: 10.1016/j.exmath.2018.01.002. arXiv: 1712.01504.

[10] François Bolley, Arnaud Guillin, and Cédric Villani. "Quantitative concentration inequalities for empirical measures on non-compact spaces". In: *Probability Theory and Related Fields* 137.3-4 (2007), pp. 541–593. ISSN: 01788051. DOI: 10.1007/s00440-006-0004-7. arXiv: 0503123 [math].

[11] Nicolas Bonneel et al. "Sliced and Radon Wasserstein Barycenters of Measures". In: *Journal of Mathematical Imaging and Vision* 51.1 (2015), pp. 22–45.

[12] Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. 2013.

[13] Greg Brockman et al. "OpenAI Gym". In: *arXiv:1606.01540* (2016). MIT License.

[14] Charlotte Bunne et al. "Learning Generative Models across Incomparable Spaces". In: *ICML*. 2019.

[15] Jiezhang Cao et al. "Multi-marginal Wasserstein GAN". In: *NeurIPS*. 2019.

[16] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *NeurIPS*. 2013.

[17] Marco Cuturi and Arnaud Doucet. "Fast Computation of Wasserstein Barycenters". In: *ICML*. 2014.

[18] Robert Dadashi et al. "Primal Wasserstein Imitation Learning". In: *arXiv:2006.04678* (2020).

[19] Ishan Deshpande et al. "Max-Sliced Wasserstein Distance and Its Use for GANs". In: *CVPR*. 2019.

[20] Rémi Flamary, Karim Lounici, and André Ferrari. "Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation". In: *arXiv* (2019). arXiv: 1905.10155.

[21] Wilfrid Gangbo and Andrzej Święch. "Optimal maps for the multidimensional Monge-Kantorovich problem". In: *Communications on Pure and Applied Mathematics* 51.1 (1998), pp. 23–45.

[22] Aude Genevay, Gabriel Peyre, and Marco Cuturi. "Learning Generative Models with Sinkhorn Divergences". In: *AISTATS*. 2018.

[23] Aude Genevay et al. "Sample Complexity of Sinkhorn divergences". In: (Oct. 2018). arXiv: 1810.02733. URL: http://arxiv.org/abs/1810.02733.

[24] Evarist Gine and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. 2016. ISBN: 9781107043169. DOI: 10.1017/cbo9781107337862.

[25] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2008, pp. xx+425. ISBN: 978-0-898716-46-7.

[26] Soheil Kolouri et al. "Generalized Sliced Wasserstein Distances". In: *NeurIPS*. 2019.

[27] R. Fergus L. Fei-Fei and P. Perona. "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories". In: *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshop on Generative-Model Based Vision* (2004).

[28] Tianyi Lin et al. "On the Complexity of Approximating Multimarginal Optimal Transport". In: *arXiv:1910.00152* (2019).

[29] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. "Multiple source adaptation and the Rényi divergence". In: *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009* (2009), pp. 367–374. arXiv: 1205.2628.

[30] Gonzalo Mena and Jonathan Weed. "Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem". In: *arXiv* (2019), pp. 1–23. ISSN: 23318422. arXiv: 1905.11882.

[31] Kimia Nadjahi et al. "Statistical And Topological Properties of Sliced Probability Divergences". In: *arXiv:2003.05783* (2020).

[32] Kimia Nadjahi et al. "Statistical and Topological Properties of Sliced Probability Divergences". In: *NeurIPS*. 2020.

[33] Khai Nguyen et al. "Distributional Sliced-Wasserstein and Applications to Generative Modeling". In: *ICLR*. 2021.

[34] Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. 2020. ISBN: 978-3-030-38437-1. DOI: 10.1007/978-3-030-38438-8. URL: http://link.springer.com/10.1007/978-3-030-38438-8.

[35] Brendan Pass. "Multi-Marginal Optimal Transport: Theory and Applications". In: *arXiv:1406.0026* (2014).

[36] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* (2019).

[37] Gabriel Peyré, Marco Cuturi, and Justin Solomon. "Gromov-Wasserstein Averaging of Kernel and Distance Matrices". In: *ICML*. 2016.

[38] Iosif Pinelis. "Exact Rosenthal-type bounds". In: *The Annals of Probability* 43.5 (Sept. 2015). ISSN: 0091-1798. DOI: 10.1214/14-aop942. URL: http://dx.doi.org/10.1214/14-AOP942.

[39] Iosif Pinelis. "Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators". In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1160–1179. ISSN: 19357524. DOI: 10.1214/17-EJS1264.

[40] Iosif Pinelis and Raymond Molzon. "Optimal-order bounds on the rate of convergence to normality in the multivariate delta method". In: *Electronic Journal of Statistics* 10.1 (2016), pp. 1001–1063. ISSN: 19357524. DOI: 10.1214/16-EJS1133. arXiv: 0906.0177.

[41] Ievgen Redko, Amaury Habrard, and Marc Sebban. *Theoretical Analysis of Domain Adaptation with Optimal Transport*. Tech. rep. 2017. URL: https://hal.archives-ouvertes.fr/hal-01613564.

[42] Mark Rowland et al. "Orthogonal Estimation of Wasserstein Distances". In: *AISTATS*. 2019.

[43] Vivien Seguy et al. "Large-Scale Optimal Transport and Mapping Estimation". In: (2018). arXiv: 1711.02283 [stat.ML].

[44] Dino Sejdinovic et al. "Equivalence of distance-based and RKHS-based statistics in hypothesis testing". In: *Annals of Statistics* 41.5 (2013), pp. 2263–2291. ISSN: 00905364. DOI: 10.1214/13-AOS1140. arXiv: 1207.6076.

[45] Rishi Sonthalia and Anna C. Gilbert. "Dual Regularized Optimal Transport". In: (2020). arXiv: 2012.03126. URL: http://arxiv.org/abs/2012.03126.

[46] Sanvesh Srivastava, Cheng Li, and David B. Dunson. "Scalable Bayes via Barycenter in Wasserstein Space". In: *Journal of Machine Learning Research* 19.1 (Jan. 2018), pp. 312–346.

[47] Ingo Steinwart and Clint Scovel. "Fast rates for support vector machines using Gaussian kernels". In: *Annals of Statistics* 35.2 (2007), pp. 575–607. ISSN: 00905364. DOI: 10.1214/009053606000001226. arXiv: arXiv:0708.1838v1.

[48] N. Tupitsa et al. "Multimarginal Optimal Transport by Accelerated Alternating Minimization". In: *CDC* (2020), pp. 6132–6137.

[49] A. W. van der Vaart. *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, June 2000. ISBN: 0521784506. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20%5C&path=ASIN/0521784506.

[50] Titouan Vayer et al. "Sliced Gromov-Wasserstein". In: *NeurIPS*. 2019.

[51] Cédric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

[52] Hongteng Xu, Dixin Luo, and Lawrence Carin. "Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching". In: *NeurIPS*. 2019.

[53] Hongteng Xu et al. "Gromov-Wasserstein Learning for Graph Matching and Node Embedding". In: *ICML*. 2019.

[54] Ding Xuan Zhou. "Capacity of reproducing kernel spaces in learning theory". In: *IEEE Transactions on Information Theory* 49.7 (2003), pp. 1743–1752. ISSN: 00189448. DOI: 10.1109/TIT.2003.813564.