Analyses of Domain Adaptation using Optimal Transport

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 Steve Evans
Assistant Professor Avi Feller

Spring 2020

Analyses of Domain Adaptation using Optimal Transport

Abstract

Analyses of Domain Adaptation using Optimal Transport

by

Yannik Pitcan

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peter Bartlett, Chair

This dissertation consists of two papers. The outline is as follows.

In chapter 1, we introduce domain adaptation, with particular emphasis on the generalization bounds for the unsupervised joint distribution domain adaptation problem. Previous work involved a theoretical analysis of the joint distribution optimal transport problem, but the generalization error required an exponentially large number of samples in order to be meaningful. This discussion is used to motivate the next two chapters, which revolve around different methods of regularizing optimal transport.

In chapter 2, we discuss entropic regularized optimal transport, otherwise known as the Sinkhorn divergence. This was introduced by Marco Cuturi as a means of regularizing the Wasserstein distance because it is more tractable computationally. In this chapter, we introduce some sample complexity bounds and also demonstrate a potential pathway to utilizing these to obtain a generalization bound for future work.

Next, in chapter 3, we study domain adaptation using optimal transport in Reproducing Kernel Hilbert Spaces. We introduce alternative means of regularization. Instead of using an entropic regularization, which is used in the Sinkhorn divergence, we regularize using dual potentials in an RKHS. In this chapter, we investigate some of the properties of this regularization methods and discuss the first main result, which is a sample complexity bound that outperforms that of unregularized optimal transport.

Chapter 4 diverges from domain adaptation and discusses some current work in prior elicitation, introducing a new framework in the form of a least squares minimization problem. There we provide non-asymptotic sample complexity bounds for M-estimators to demonstrate why our approach is theoretically viable for prior elicitation.

Finally, we discuss some directions for future work in chapter 5.

This is dedicated to my parents.

Contents

List of Figures

List of Tables

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 incredible advisor and suggested this topic and several of the key insights. Without him, I do not know where I would be today.

I would also like to thank my colleagues Soren Kuenzel and Alexander Tsigler. Both individuals gave me many suggestions and tips that helped immensely.

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

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 this 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 [32].

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 [32].

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.

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

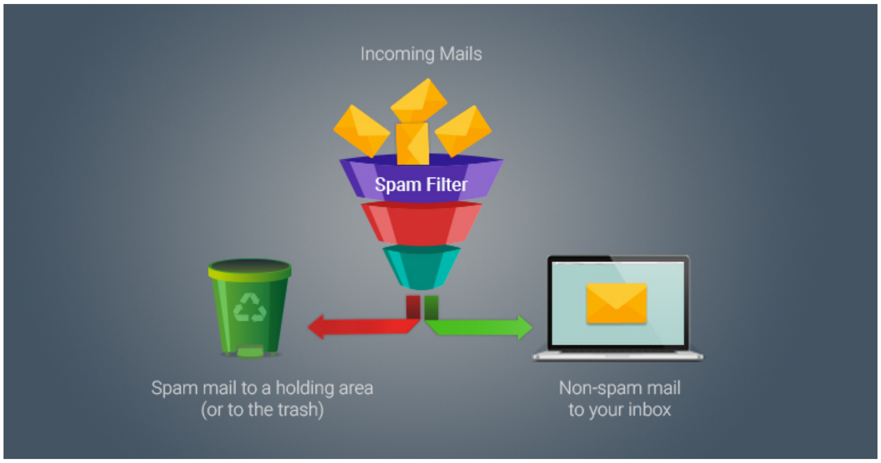

Figure 1.1: One application of transfer learning: spam filtering.

to measure how well it performs on the target domain. Formally, we describe this as follows

Theorem 1.2.1 (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. 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. $X_S = X_T$ but $t_S \neq t_T$.
- Transductive transfer learning. $X_S \neq X_T$ but $t_S = t_T$.
- Unsupervised transfer learning. $t_S \neq t_T$ and $X_S \neq X_T$.

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 $P(X, Y) = P(X)P(Y|X)$.

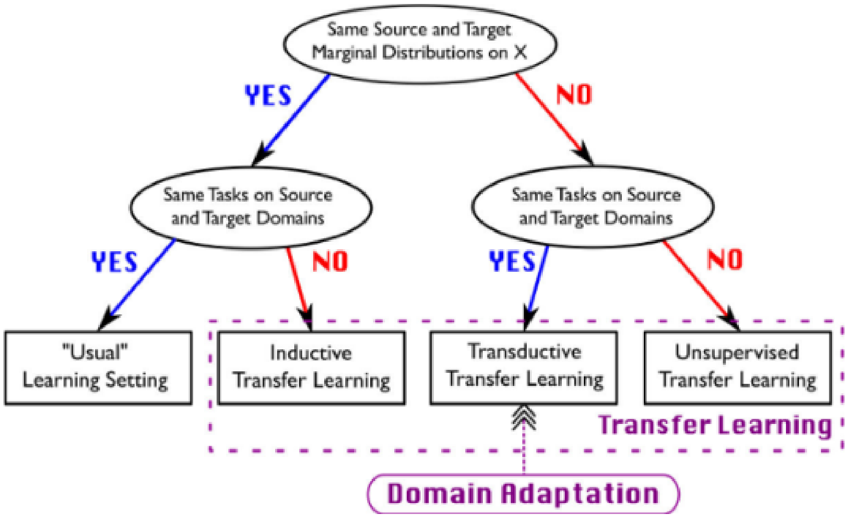
Same Source and Target Marginal Distributions on X
YES
NO
Same Tasks on Source and Target Domains
Same Tasks on Source and Target Domains
YES
NO
YES
NO
"Usual" Learning Setting
Inductive Transfer Learning
Transductive Transfer Learning
Unsupervised Transfer Learning
Transfer Learning
Domain Adaptation

Figure 1.2: Positioning of Domain Adaptation compared to other learning techniques (Redko).

- $Y \rightarrow X$ 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 [32]. 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 $X \rightarrow Y$ 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 [23] 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:

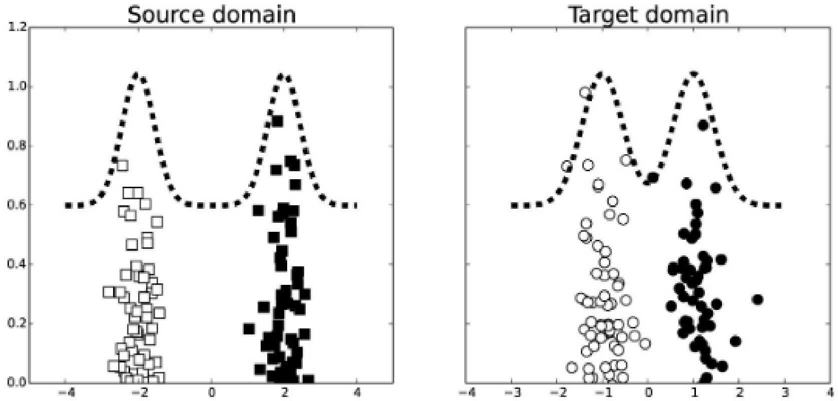

Figure 1.3: Covariate shift

$$
\begin{aligned}
R_T^l(h) &= \mathrm{E}_{(x,y)\sim T} l(h(x), y) \\
&= \mathrm{E}_{(x,y)\sim T} \frac{S(x,y)}{S(x,y)} l(h(x), y) \\
&= \mathrm{E}_{(x,y)\in X\times Y} T(x,y) \frac{S(x,y)}{S(x,y)} l(h(x), y) \\
&= \mathrm{E}_{(x,y)\sim S} \frac{T(x,y)}{S(x,y)} l(h(x), y) \\
&= \mathrm{E}_{(x,y)\sim S} \frac{P(X_T)}{P(X_S)} l(h(x), y)
\end{aligned}
$$

where the last equality uses the fact that $P(Y_T|X_T) = P(Y_S|X_S)$.

- 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

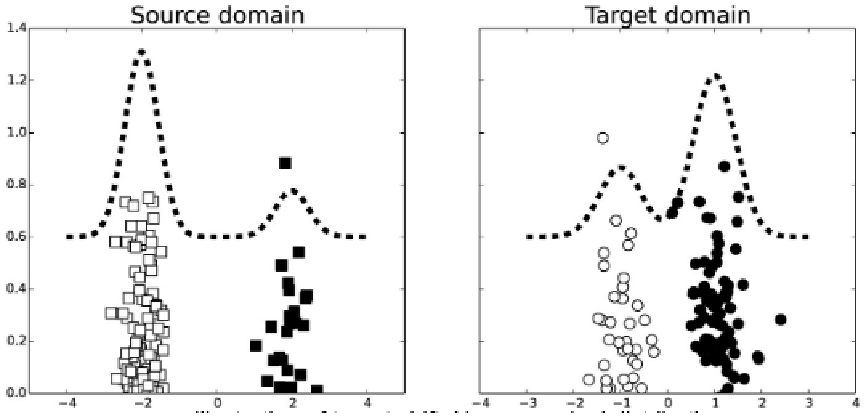

Figure 1.4: Target shift

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

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

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 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 define a dissimilarity measure 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, and the choice of metric can impact the behavior of the labeling function. [32]

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. This is done when a divergence measure is too difficult to compute empirically. Additionally, we investigates 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 by setting and proposing the L^1-distance [5].

First, we provide some definitions.

Definition 1. *Rademacher complexity*

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,

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

Definition 2. *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 3. *VC Dimension*

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

Definition 4. *$\mathcal{H}$-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.

Theorem 1.2.2. *Generalization bound with respect to $\mathcal{H}$-divergence [5]*

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

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

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 5. *Symmetric difference hypothesis divergence*

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

Theorem 1.2.3. *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 then obtained a bound for risk on the target domain that involved the empirical divergence.

Theorem 1.2.4. *Let $\lambda^* = \min_{h\in\mathcal{H}} R_S(h) + R_T(h)$ be the minimum joint risk. With probability at least $1-\delta$:*

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

One sees here that the bound relies on a notion of divergence between the two domains, as stated earlier, along with a divergence between the hypothesis and true labeling function.

Of note here 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 by Mohri and Mansour to use Renyi and $\mathcal{Y}$-discrepancy distances.

Definition 6. *Renyi divergence*

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

where α denotes its order. When $\alpha = 1$, the Renyi divergence is equivalent to the Kullback-Leibler divergence.

Definition 7. *$\mathcal{Y}$-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 $y \in Y$.

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

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

Here, the sand distribution and shape of the pit are represented by distributions μ and ν, respectively.

Moreover, one cannot change the size of a sand particle, so the sand cannot be concentrated at a single point in the pit. In other words, 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 can also recognize this as 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 a transport map. Normalizing μ and ν to be probability measures, the Monge problem finds the optimal transport map minimizing transport costs [27].

Definition 8. *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. Currently, 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

$$\begin{aligned} \pi(A \times Y) &= \mu(A), \quad A \subseteq X \\ \pi(X \times B) &= \nu(B), \quad B \subseteq Y \end{aligned}$$

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

Definition 9. *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 X and Y.

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 10. *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 Sinkhorn Divergence

Previously, we gave a glimpse at domain adaptation and briefly discussed optimal transport, without tying the two concepts together. Thus, we give a more in-depth introduction here to how optimal transport and domain adaptation are linked. Then, we introduce entropic-based optimal transport, otherwise known as Sinkhorn divergence [20]. Primarily, we focus on sample complexity results involving this type of divergence.

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 11. *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 12. *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 13. *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.

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

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 last line is a byproduct of the reproducing property.

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

Definition 14. *Discrete Wasserstein [32]*

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 energy needed to move a probability mass from* x_S^i *to* x_T^j*.*

Definition 15 (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), the expected loss is defined as*

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

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 $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}^+$.

We also use a Γ operator to represent expectation, i.e. $\Gamma f = E_\pi[f(X)]$ where $X \sim \pi$.

The following interpretation of expected loss using RKHS properties was used to derive earlier bounds.

Definition 16 (RKHS interpretation)**.** *Assume* $l \in \mathcal{H}_{k^q}$ *where* $\mathcal{H}_{k^q}$ *is an RKHS with kernel* $k^q : \Omega \times \Omega \to \mathbb{R}$ *induced by* $\phi : \Omega \to \mathcal{H}_{k^q}$ *and* $k^q(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_{k^q}}$*.*

With this definition, it is immediate that,

$$\epsilon_S(h, f_S) = \mathrm{E}_{x \sim \mu_S}[l(h(x), f_S(x))] = \mathrm{E}_{x \sim \mu_S}\left[\langle \phi(x), l \rangle_{\mathcal{H}_{k^q}}\right]$$

and

$$\epsilon_T(h, f_T) = \mathrm{E}_{y \sim \mu_T}[l(h(y), f_T(y))] = \mathrm{E}_{y \sim \mu_T}\left[\langle \phi(y), l \rangle_{\mathcal{H}_{k^q}}\right].$$

Prior Work

First, we introduce some past results pertaining to risk bounds with respect to the Wasserstein distance. In this section, we will show that the choice of the cost function is key in deriving theoretical bounds.

First Theoretical Bounds [32]

The key assumption here is the cost function. Another assumption is that the true labeling function f lies within a unit ball of an RKHS, i.e.

$$\mathcal{F} = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\},$$

where $\mathcal{H}_k$ is an RKHS with kernel k.

In this scenario, we have the following specifications:

- Let $\mu_S, \mu_T \in \mathcal{P}(X)$ be two probability measures on $\mathbb{R}^d$.

- $$c(x,x') = \|\phi(x) - \phi(x')\|_{\mathcal{H}_{k_l}}.$$
  - $\mathcal{H}_{k_l}$ is an RKHS.
  - $k_l : \Omega \times \Omega \to \mathbb{R}$ is a kernel function such that
    $$k_l(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_{k_l}}$$
- $$l(h(x), f(x)) = |h(x) - f(x)|^q,\ q > 0$$
  and thus the loss function is convex, bounded, symmetric, and satisfies triangle inequality
- If Ω is separable and $k_l(x,x') \in [0,K]$ for some $K \in \mathbb{R}$, then for all $x, x' \in \Omega$,
  $$\mathrm{E}_D[\sqrt{k_l(x,x')}] < \infty$$
  for $D = S$ or T.

Before we continue, let us briefly discuss the use of the aforementioned cost function. It can be seen that:

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

There also exists a one-to-one relationship between the choice of a positive-definite kernel k and the cost function c.

Secondly, $l_{h,f} : x \to l(h(x), f(x))$ belongs to an RKHS. $(h,f) \in \mathcal{F}^2$ and l is a nonlinear mapping of $\mathcal{H}_k$.

Lemma 2.1.1. *If the above assumptions hold, then, for all $h, f \in \mathcal{H}_{k_l}$,*

$$\epsilon_T(h,f) \le \epsilon_S(h,f) + W_1(\mu_S, \mu_T).$$

With the use of a concentration inequality on Wasserstein distances [9], empirical risk bounds are obtained.

Theorem 2.1.2. *Let μ be a probability measure on $\mathbb{R}^d$ such that*

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

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

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 $N \geq N_0 \max\{\epsilon^{-(d'+2),1}\}$,

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

Joint Distribution Domain Adaptation (JDOT)

JDOT is the approach taken in [14]. In this setting, one works with unsupervised domain adaptation between joint distributions. The inspiration behind this method is that the assumption that conditional distributions are preserved, i.e. $P_S(Y|T(X)) \approx P_T(Y|T(X))$, may not necessarily hold true.

- Now the cost function used is

$$\alpha d(x_s, x_t) + L(y_s, y_t).$$

- The unsupervised domain adaptation problem is studied here, so the target labels represented by y_t are not known. Thus, one cannot find an optimal coupling.
- The existence of an optimal coupling is not necessary since the goal here is to estimate a mapping on the target data.

2.2 Entropic Regularization

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 $O(n^3 \log(n))$.

One way of addressing the intractability of OT is by using an entropic regularization.

Definition 17 (Entropic Regularization of Wasserstein)**.**

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

where

$$H(\pi|\alpha \otimes \beta) = \int_{\mathcal{X}\times\mathcal{Y}} \log(\frac{d\pi(x,y)}{d\alpha(x)d\beta(y)})d\pi(x,y).$$

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 [20].

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

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

Since $W_\epsilon(\alpha, \alpha) \neq 0$, we can normalize this by defining the Sinkhorn divergence as in the definition below.

Definition 18 (Sinkhorn Divergence)**.**

$$\bar{W}_\epsilon(\alpha, \beta) = W_\epsilon(\alpha, \beta) - \frac{1}{2}(W_\epsilon(\alpha, \alpha) + W_\epsilon(\beta, \beta))$$

2.3 Preliminary Generalization Bounds

Our claim is that we can get guarantees when doing a dual minimization of the Wasserstein distance with the specified cost function $\alpha d(x_s, x_t) + l(y_s, f(x_t))$ from S to T_f.

$$(\hat{\Gamma}, \hat{f}) = \arg\min_{\Gamma, f} \Gamma\left[\alpha d(x_s, x_t) + l(y_s, f(x_t))\right]$$

over all couplings that preserve the marginals on S (x_s, y_s) and T_X (x_t). Let $\hat{\Gamma}$ be the optimal joint distribution over (x_s, y_s, x_t) above. Now define $\hat{\hat{\Gamma}}$ on (x_s, y_s, x_t, y_t) to be a coupling with the same marginal over (x_s, y_s, x_t) as $\hat{\Gamma}$ and the correct marginals on T (x_t, y_t). Let Γ^* be an arbitrary coupling on the same variables which preserves the marginal distributions on S $(x_s, y_s)$and T (i.e., x_t, y_t). For now, let's assume we're dealing with absolutely continuous distributions here so we don't have existence issues. Fix $f^* \in F$.

Error bounds with respect to Wasserstein distance

Theorem 2.3.1.

$$err_T(\hat{f}) \leq err_T(f^*) + W(S, T) + \hat{\hat{\Gamma}}\left(-\alpha d(x_s, x_t) + l(y_s, y_t)\right)$$

where $\hat{\hat{\Gamma}}$ is the expectation operator with respect to a coupling with matching marginals on T and the same marginal on (x_s, y_s, x_t) as $\hat{\Gamma}$.

Proof.

$$
\begin{aligned}
err_T(\hat{f}) &= \hat{\hat{\Gamma}} l(y_t, \hat{f}(x_t)) \\
&\leq \Gamma^* \left[\alpha d(x_s, x_t) + l(y_t, f^*(x_t)) + l(y_t, y_s)\right] - \hat{\hat{\Gamma}} \left[\alpha d(x_s, x_t) - l(y_s, y_t)\right] \\
&= err_T(f^*) + \alpha \left(\Gamma^* d(x_s, x_t) - \hat{\hat{\Gamma}} d(x_s, x_t)\right) + \Gamma^* l(y_s, y_t) + \hat{\hat{\Gamma}} l(y_s, y_t).
\end{aligned}
$$

There, by choosing Γ^* as the optimal coupling between S and T, that is,

$$W(S, T) = \Gamma^* \left(\alpha d(x_s, x_t) + l(y_s, y_t)\right),$$

then we get the upper bound

$$err_T(\hat{f}) \leq err_T(f^*) + W(S, T) + \hat{\hat{\Gamma}} \left(-\alpha d(x_s, x_t) + l(y_s, y_t)\right).$$

□

Sample Dependent Sinkhorn Bounds

The following two theorems can be used to bound the empirical Sinkhorn from the Wasserstein distance [20] and we list them for convenience. These can be combined with the above result to establish generalization bounds with respect to the Sinkhorn divergence, but we leave this as an open problem.

Theorem 2.3.2. *Let α and β be probability measures on $\mathcal{X}$ and $\mathcal{Y}$ subsets of $\mathbb{R}^d$ such that $|X| = |Y| \leq D$ and assume that c is L-Lipschitz w.r.t x and y. Then*

$$W_\epsilon(\alpha, \beta) - W(\alpha, \beta) \leq 2\epsilon d \log\left(\frac{\epsilon^2 LD}{\sqrt{d}\epsilon}\right)$$

where $\mathcal{X}$ and $\mathcal{Y}$ are subsets of $\mathbb{R}^d$ with diameters at most D and c is L-lipschitz w.r.t x and y.

and

Theorem 2.3.3. *Let $\hat{\alpha_n}$ and $\hat{\beta}_n$ be empirical measures for α and β with size n for each.*

$$|W_\epsilon(\hat{\alpha_n}, \hat{\beta}_n) - W_\epsilon(\alpha, \beta)| \leq 6B\frac{\lambda K}{\sqrt{n}} + C\sqrt{\frac{2\log\frac{1}{\delta}}{n}}$$

with probability at least $1 - \delta$.

Chapter 3

An Alternative Means of Regularization for Domain Adaptation Problems

3.1 Introduction

Previously, we explored the applications of entropic regularization for the Wasserstein distance (Sinkhorn) in domain adaptation. In particular, we derived a new generalization bound for target error with respect to the Sinkhorn divergence. However, for domain adaptation, entropic regularization may not be the ideal route to take.

When we prescribe a divergence measure for domain adaptation, the scenario we seek to penalize against is when the source S and target T distributions are not identical. However, with the entropic regularization, one is penalizing against S and T being independent. In this chapter, we propose an alternative regularization that may be more suitable for domain adaptation problems.

If S and T are identical in distribution, which is the ideal setting in machine learning, then there exists an identity mapping between the two. Thus, it makes sense to use a regularization that penalizes the deviation between the transport map and the identity map.

3.2 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 3.2.1. *Brenier's Theorem*

Let μ and ν be absolutely continuous probability measures on $\mathbb{R}^d$ with respect to the Lebesgue measure and ν has bounded support. There exists a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that its gradient, $\nabla\phi$, is the optimal transport map from μ to ν.

We call ϕ a Kantorovich potential.

Corollary 3.2.1.1. *Existence of Solution to Monge Problem*

Under the above assumptions, $\nabla\phi$ uniquely solves the Monge problem.

$$\int_X |x - \nabla\phi(x)|^2 \, d\mu(x) = \int_{T_\# \mu = \nu} |x - T(x)|^2 \, d\mu(x)$$

Key Optimization Problem

We wish to examine the following optimization problem, where H is a reproducing kernel hilbert space and $c(x, y)$ denotes an arbitrary cost function in x and y.

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

We will circle back to this after discussing some convex analysis preliminaries crucial to analyzing this going forward.

3.3 Convex Analysis Prerequisites

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

Definition 19. *Let $f : X \to \mathbb{R} \cup +\infty$ be a lower semicontinuous function. 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 20. *The subdifferential of a lower semi-continuous convex function ϕ at $x \in dom\phi$ is defined by*

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

Corollary 3.3.0.1. *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 3.3.1 (Fenchel-Young inequality)**.**

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

Corollary 3.3.1.1 (Fenchel-Young equality)**.**

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

iff

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

Theorem 3.3.2. *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,

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

Weak duality $p \geq d$ follows immediately. Strong duality $p = d$ requires $\partial v(0) \neq \emptyset$, i.e.

$$0 \in \text{int dom } v = \text{int}[\text{dom } g - \text{dom} f].$$

□

Strong duality criterion. Here we prove $0 \in \text{int dom } v = \text{int}[\text{dom } g - \text{dom} f]$ implies $\partial v(0) \neq \emptyset$. Let $-y^* \in \partial v(0)$ we have

$$f(x) + g(x+y) \geq v(y) \geq p - \langle y^*, y \rangle$$

Letting $u = x + y$, we have

$$p \leq f(x) - \langle y^*, x \rangle + g(u) + \langle y^*, u \rangle$$

and taking infimum with respect to x, u, we have

$$p \leq -f^*(y^*) - g^*(-y^*) \leq d \leq p.$$

This concludes the proof. □

3.4 Derivation of Primal Formulation

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

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

The arguments that follow are similar to those used to prove Kantorovich duality.

Let

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

.

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.

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

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

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

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

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

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

When finite, the last line (when π is not a coupling) can be rewritten as

$$\frac{\sup_{\|u\|=1}(\int u\,ds - \int u\,d\pi)^2}{2\lambda} + \frac{\sup_{\|v\|=1}(\int v\,dt - \int v\,d\pi)^2}{2\lambda}$$

if finite. Otherwise, $f^*(\pi) = +\infty$.

The dual problem is $\sup_{\pi\in E^*}\left[-f^*(\pi) - g^*(-\pi)\right]$, which, from the above, is

$$\sup_{\pi\in E^*}\left\{-\int_{X\times Y} c\,d\pi - \int_X \varphi\,d\mu - \int_Y \psi\,d\nu + \int_{X\times Y}(\varphi(x)+\psi(y))\,d\pi + \lambda(\|\varphi\|_H^2 + \|\psi\|_H^2)\right\}.$$

What does the dual of f look like?

If the previous claim holds,

$$f^*(\pi) = \frac{1}{4\lambda}\left(Q(s,\pi_1) + Q(t,\pi_2)\right)$$

where

$$Q(s,\pi_1) = \int k(x,y)\,ds(x)ds(y) + \int k(x,y)\,d\pi_1(x)d\pi_1(y) - 2\int k(x,y)\,ds(x)d\pi_1(y)$$

and similarly for $Q(t,\pi_2)$.

Then our dual problem is

$$\begin{aligned} \sup_{\pi\in E^*}(-f^*(\pi) - g^*(-\pi)) &= \sup_{\pi\in M_+}\left[-\frac{1}{4\lambda}(Q(s,\pi_1) + Q(t,\pi_2)) - \int\left(c(x,y) - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2\right)\,d\pi\right] \\ &= \inf_{\pi\in M_+}\left[\frac{1}{4\lambda}(Q(s,\pi_1) + Q(t,\pi_2)) + \int\left(c(x,y) - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2\right)\,d\pi\right] \end{aligned}$$

From here on, we assume our cost function is the square-loss, i.e. $c(x,y) = \frac{1}{2}\|x-y\|^2$. It immediately follows that $c(x,y) - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 = x^T y$

The question now is how to minimize

$$\frac{1}{4\lambda}(Q(s,\pi_1) + Q(t,\pi_2)) + \int x^T y \, d\pi(x,y).$$

Inner product space of signed measures

Define the inner product with respect to signed measures m, n on $\mathcal{X}$ as

$$\langle m, n \rangle = \int \tilde{k}(u,v) \, dm(u) \, dn(v).$$

and

$$\|s \otimes \pi_2 - \pi\|^2 = Q(s,\pi_1) = \|s - \pi_1\|^2.$$

If we're looking at $\mathcal{X}^2$, then we have

$$\langle m, n \rangle = \int \tilde{k}((u_1,u_2),(v_1,v_2)) \, dm(u_1,u_2) \, dn(v_1,v_2)$$

and

$$\|s \otimes \pi_2 - \pi\|^2 = \int \tilde{k}((u_1,u_2),(v_1,v_2)) \, d(s(u_1)\pi_2(u_2) - \pi(v_1,v_2))$$

Then our problem reduces to showing

$$\inf_m \{\langle m, s-\pi_1 \rangle + \lambda \langle m, m \rangle\} = -\frac{1}{4\lambda}\langle s-\pi_1, s-\pi_1 \rangle.$$

Proof.

$$\begin{aligned}\langle m, s-\pi_1 \rangle + \lambda\langle m, m\rangle &= \lambda(\langle m, m \rangle + \langle m, \frac{s-\pi_1}{\lambda}\rangle) \\ &= \lambda(\|m + \frac{s-\pi_1}{2\lambda}\|^2 - \|\frac{s-\pi_1}{2\lambda}\|^2)\end{aligned}$$

Choosing $m = \frac{\pi_1 - s}{2\lambda}$, we obtain the desired minimum. □

Discrete Setting

If we take π_1 and π_2 to be discrete measures, we can represent them by $\pi_1 = \Pi 1$ and $\pi_2^T = 1^T \Pi$, i.e. taking the row and column marginals of Π.

Our optimization problem then becomes

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

3.5 An Alternative Optimization Problem

$f(\Pi,\lambda) = \frac{(\Pi\vec{1}-s)^T K(\Pi\vec{1}-s) + (\Pi^T\vec{1}-t)^T K(\Pi^T\vec{1}-t)}{4\tau} - Tr[\Pi K] + \lambda(1-\vec{1}^T\Pi 1) - v_1 E_1^T \Pi E_1 - v_2 E_2^T \Pi E_1 - v_3 E_1^T \Pi E_2 - v_4 E_2^T \Pi E_2$

where $E_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $E_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ (assuming this is the 2D setting).

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

If $K = I$, then by Karsh Kuhn Tucker conditions, we have

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

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

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

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

And it follows that

$v_1 + 1 = v_3$

$v_4 + 1 = v_2$

and also

$\frac{1}{2\tau} - 1 = v_1 + v_2 + 2\lambda = v_3 + v_4 + 2\lambda.$

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.

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

Example Code

```
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(
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(
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]]
```

3.6 Error Bounds

We seek to determine bounds on the difference between our 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 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 [**Sonthalia2020**] but here, we generalize from the discrete setting.

Theorem 3.6.1.

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

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

3.7 Sample Complexity Bound

Introduction

Previously, we investigated the dual form of the dual norm regularized optimal transport and examined some of its fundamental properties. Now, we transition to establishing finite sample guarantees. 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 $K = \max_{X\times Y} c(x, y)$.

Proof Technique

The proof technique here is inspired by work on entropic optimal transport sample complexity [25]. 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^*, of*

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

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.

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

Assuming $c(x, y) \geq 0$, we get

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

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*

$$\text{(9) } |S(P_n,\ Q) - S(P,\ Q)| \leq 2 \sup_{u \in \mathcal{F}} |E_P u - E_{P_n} u|.$$

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 $f,\ f_n \in \mathcal{F}$.

Strong duality implies 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 $(f_n,\ g_n)$ for their respective dual problems, 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

$$|S(P,\ Q) - S(P_n,\ Q)| = |\mathcal{A}^{P,Q}(f,\ g) - \mathcal{A}^{P_n,Q}(f_n,\ g_n)|$$

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

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

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

and similarly,

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

□

Corollary 3.7.0.1. *Let P, Q, P_n, and Q_n be probability distributions. Then*

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

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.

From here onwards, define $K = \max(K(P, Q), K(P_n, Q))$. 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, . . . , X_n, we have the bound*

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

Proof. Denote

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

and $k(x, x) = 1$ if $k(x, y) = e^{-\sigma^2\|x-y\|_2^2}$. Thus $|f(x)| \leq \|f\|_{\mathcal{H}}$ for all $x \in X$. □

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

Theorem 3.7.1. *[35] 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*

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

Using this result, we then have, by the use of a chaining bound [21],

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

Here, $c_{p,\delta,d}$ and $c'_{p,\delta,d}$ denote constants with respect to n. One can notice here that the upper bound is $O(1/n)$.

Chapter 4

Domain adaptation using Monge mappings

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. The first result was also discovered concurrently by [17].

Linear Monge Mappings

Previous work [17] estimated the linear Monge mapping between Gaussian distributions and used this to provide a domain adaptation generalization bound.

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.

Notation

Let A and B be positive matrices.

Definition 21. *An $n \times n$ symmetric matrix M is a positive matrix if and only if $x^{'}Mx > 0$ for all $x \in \mathbb{R}^n$.*

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

Definition 22 (Fidelity)**.** *If A and B are positive matrices, then the fidelity $F(A, B) = tr(A^{1/2}BA^{1/2})^{1/2}$.*

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.

Lemma 4.0.1. *Let μ_1 and μ_2 be probability measures with zero means and positive-definite covariance matrices A, B respectively. Then the optimal linear transport map coincides with the optimal transport map.*

Minimizing over set of linear maps [8]

Let x and y be random vectors with values in $\mathbb{C}^n$, each having zero mean WLOG, and with covariance matrices A and B, respectively. This last statement means that

$$A = [E(\overline{x}_i x_j)],\ B = [E(\overline{y}_i y_j)].$$

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

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

$$\begin{bmatrix} [E(\overline{x}_i x_j)] & [E(\overline{x}_i y_j)] \\ [E(\overline{y}_i x_j)] & [E(\overline{y}_i y_j)] \end{bmatrix} = \begin{bmatrix} A & M \\ M^* & B \end{bmatrix}$$

We seek to minimize the following:

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

This is equivalent to the following optimization problem:
$\max\{|\text{tr}M| : C = \begin{bmatrix} A & M \\ M^* & B \end{bmatrix} \geq 0\}$.
The value of the maximum is $F(A,\ B)$. So

$$\begin{aligned} \min E\|x - y\|^2 &= \text{tr}(A + B) - 2\text{tr}(A^{1/2}BA^{1/2})^{1/2} \\ &= d^2(A,\ B). \end{aligned}$$

Let x be a vector with mean 0 and covariance matrix A. Then for any $T \in \mathrm{FM}(n)$ we have

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

Hence,

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

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 and $y = Tx$, with $T = A^{-1}\#B$.

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

$$\begin{aligned} E(\overline{y}_i y_j) \ &= \ E\sum_{k,l} t_{ik}t_{kl}\overline{x}_k x_l \\ &= \ \sum_{k,l} t_{ik}t_{kl}a_{kl} = (TAT)_{ij}. \end{aligned}$$

If T is the optimal transport map from A to B, then $TAT = B$. This shows that the covariance matrix of the vector y is B.

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.

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

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 $A = Cov(TX) = (\frac{1}{n-1})(TX)^T C(TX)$, $B = \Sigma_v$, and $C = (1 - \frac{1}{n}J) = C^T$ is a centering matrix.

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

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

Differentiating the Bures distance part I

For PSD matrices a drastic simplification is possible:

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

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 Higham

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

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

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

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 $dA \to dT$.

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

In the above derivation, the function

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

was 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

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

Chapter 5

Asymptotics for Prior Elicitation

This chapter is a self-contained paper submitted for publication at NeurIPS.

5.1 Statistical Elicitation

Elicitation is the process of forming a probability distribution from a person's knowledge and beliefs. We will focus on the case of elicitation to obtain a prior that will be used in a subsequent machine learning task. Although most of the results will be applicable to other motivations for elicitation, narrowing the language will simplify our discussion.

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 in collaboration in the structuring and decomposition step. Then, during the elicitation phase there is further iteration over three steps: 1. elicit summaries, 2. fit a distribution, and 3. assess adequacy. The elicitation process is our focus for the presented work, as the fitting and assessment steps are the primary role of the automated tool.

In higher dimensions, summaries are less intuitive and even cumbersome to communicate. Therefore, we will, in our automated facilitator-statistican discussion, shift from eliciting summaries to eliciting samples. Sample based elicitation has been applied in an experimental setting successfully for fitting beta distributions.

NEED TO ADD REFERENCES TO SOME SPECIFIC EXPERIMENTS

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.

In the Human Computer Interactive and Viszualization communities, elicitation has been studied in amazon turkers either to eval *REST OF SENTENCE? - NEED REFERENCES*

Toward the study of Bayesian modeling in a broad sense, HCI researchers have built tools for eliciting specific forms of priors. *REFERENCES*

Observing how statisticians set priors revealed that the choice of visualization can impact how experienced Bayesian statisticians choose to set a prior. In designing a more general prior elicitation tool, it will be important to understand what forms of information will facilitate good inferences and to balance these forms with what pscyhological insights exist regarding how experts choose matching interfaces. Further research has examined what information experts are able to express reliably, and how visualization impact the broad strategies of the expert. *REFERENCES* In this work, we consider the learnability of classes of priors from different forms of evidence toward making tool design choices.

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. While in general, learning from correlated samples can degrade performance relative to i.i.d samples, learning from coresets constructed with a diverse sampling strategy has been shown improve learning rates *REFERENCE*. 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.

A Least-Squares Based Approach to Elicitation

Proposed Objective Function

Assume that we elicit i.i.d. observations x_i with corresponding sample likelihoods z_i for $i = 1, \ldots, n$. Assuming we have a parametric model class, our proposed method of estimating θ involves minimizing an objective function, which we illustrate below.

Let $Q((\vec{x}, \vec{z}), \theta) = \sum_i (l(x_i, \theta) - z_i)^2$

Our proposed optimization problem is

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

where x_i, z_i is the ith sample and likelihood.

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

and let $\psi((x_i, z_i), \theta) = l(x_i, \theta) - z_i$.

$\hat{\theta}$ is a solution to

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

and

θ_0 solves

$$\mathrm{E}_{\theta_0} \left[(l(x_i, \theta) - z_i) \frac{\partial}{\partial\theta} l(x_i, \theta) \right] = 0$$

Asymptotic Analysis

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

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

If

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

is finite and nonzero and

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

,

then

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

where

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

Finite Sample

To do an elicitation, we will need to obtain a finite number of samples from an expert. While the large sample results give confidence in the general tractability of the problem, the finite sample results are important to understanding the realistic feasibility of implementing an automated elicitation tool.

For the finite sample result, the assumptions are the following [30]:

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 $\mathbb{R}$.

For brevity, we will use P and E throughout defined as $\mathrm{P} := \mathrm{P}_{\theta_0}$ and $\mathrm{E} := \mathrm{E}_{\theta_0}$.

For $x \in \mathcal{X}$ and $\theta \in \theta$, consider the function

$$\ell_{X,Z}(\theta) = -(\ell_x(\theta)-z)^2$$

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 $x \in \mathcal{X}_{>0}$ the likelihood $l_x(\theta)$ are 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. $\mathrm{E}\sup|\ell'''_{X,Z}(\theta)|^3 < \infty$.

$$\theta \in [\theta_0-\delta,\theta_0+\delta]$$

Suppose that the above conditions hold and that $\ell_{X,Z}(\theta)$ is concave in $\theta \in \Theta$, for each $x, z \in \mathcal{X} \times \mathcal{Z}$.

Then

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

for all real z, and

$$|\mathrm{P}(\sqrt{n\frac{I_2(\theta_0)^2}{I_1(\theta_0)}}(\hat{\theta}-\theta_0) \leq z) - \Phi(z)| \leq \frac{C_\omega}{z^3\sqrt{n}}$$

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

5.3 Experiments

To validate our learnability results, we performed experiments using synthetic data. Throughout, we simulate the batch data that would be obtained from an elicitation by sampling a "target" distribution and then computing the likelihood. Next, we use these data to learn the parameters of the proposed distribution class. We assess the quality of the elicitation with the $L2$ error of the learned parameters and the KL divergence for the learned distribution. Last, we generate a number of batches of data each of a fixed size and then attempt to learn from that.

Validating the estimator and the learning rate

First, we use exact samples without noise. We generate sample, likelihood pairs from the target distribution in fixed sizes and multiple batches. We used a batch size of 20. Then, we ran the optimization process to see how well can we learn the true distribution from these samples as we increase the number of samples in those batches.

We varied the number of samples from 2 to 40 and had 20 batches. We wanted to see for how many batches out of 20 are we able to learn the true distribution. We saw that the rate of success of elicitation was directly related to the sample size as shown below. Second, we added noise the likelihoods, assuming that our expert can give realistic samples, but may be variably good.

Proof of Theoretical Bound

Here, we provide a theoretical proof of the main asymptotic result. This follows the technique introduced in [30] for maximum likelihood estimators. Pinelis briefly states that his result could be extended to M-estimators, and in the following, we fully exposit the proof for the general class of M-estimators, which requires adjustments to the assumptions in [30].

This proof is organized as follows.

1. We describe the general problem setting and assumptions required.
2. We demonstrate tight bracketing of our M-estimator between two functions of the sum of independent random vectors.
3. We present uniform and nonuniform optimal-order bounds on the convergence rate in the multivariate delta method [31].
4. We apply the general bounds in the multivariate delta method such that we can make bracketing work.
5. We bound the remainder and show this is asymptotically negligible under certain conditions.

Setting and Assumptions

Let $X, X_1, X_2, \dots$ 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 $X, X_1, X_2, \dots$ are i.i.d. with respect to each of the measures 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 $\theta \in \Theta$, $P_\theta X^{-1}$ of X has a density p_θ with respect to a measure μ on $\mathcal{B}$.

Because the extended real line $[-\infty, \infty]$ is compact, for each $n \in \mathbb{N}$ and point $x = x_n = (x_1, \dots, x_n) \in \mathcal{X}^n$, the function $\Theta \ni \theta \mapsto \ell_{X,Z}(\theta) = \sum_{i=1}^n -(l_{x_i}(\theta) - z_i)^2$ has at least one generalized maximizer $\hat{\theta}_n(x)$ in the closure of Θ.

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 $\mathbb{R}$.

For convenience, we provide the assumptions for ℓ again 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. $\mathrm{E}\sup|\ell'''_{X,Z}(\theta)|^3 < \infty$.

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

$$0 = \ell'_{\mathrm{x}}(\hat{\theta}) = \ell'_{\mathrm{x}}(\theta_0) + (\hat{\theta} - \theta_0)\ell''_{\mathrm{x}}(\theta_0) + \frac{(\hat{\theta} - \theta_0)^2}{2}\ell'''_{\mathrm{x}}(\theta_0 + \xi(\hat{\theta} - \theta_0)) \tag{5.2}$$

$$= n(\overline{Z} - (\hat{\theta} - \theta_0)\overline{U} + \frac{(\hat{\theta} - \theta_0)^2}{2}\overline{R}) \tag{5.3}$$

for some $\xi \in (0, 1)$ as a function of the X_i's, where $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$, $\overline{U} = \frac{1}{n}\sum_{i=1}^n U_i$, $\overline{R} := \frac{1}{n}\sum_{i=1}^n R_i$, $\overline{R^*} := \sum_{i=1}^n R_i^*$,

$$Z_i = \ell'_{X_i}(\theta_0), \quad U_i = -\ell''_{X_i}(\theta_0) \tag{5.4}$$

$$R_i = \ell''_{X_i}(\theta_0 + \xi(\hat{\theta} - \theta_0)) \in [-R_i^*, R_i^*], \quad R_i^* = \sup_{\theta \in [\theta_0 - \delta, \theta_0 + \delta]} |\ell'''_{X_i}(\theta)|. \tag{5.5}$$

Looking at (5.2) and (5.3), one has a quadratic equation for $\hat{\theta}$.
On the event G one has

$$\hat{\theta} - \theta_0 = \frac{\overline{Z}}{\overline{U}} \text{ if } \overline{R} = 0 \,\&\, \overline{U} \neq 0,$$
$$\hat{\theta} - \theta_0 \in \{d_+,\ d_-\} \text{ if } \overline{R} \neq 0,$$

where

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

One defines a "bad event" by letting
$B := B_1 \cup B_2$, where
$B_1 := \{\overline{R} \neq 0, \hat{\theta} - \theta_0 = d_+\} \cup \{\overline{U} \leq 0\}$ and $B_2 := \{\overline{U}^2 \leq 2|\overline{Z}|\overline{R^*}\}$.
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),

$$\mathrm{P}(G \cap B_1) \leq \mathrm{P}(\overline{U} \leq 0 \text{ or } \frac{\overline{U}}{\overline{R^*}} \leq \delta) = \mathrm{P}(\frac{\overline{U}}{\overline{R^*}} \leq \delta) = \mathrm{P}(\sum_{i=1}^n (U_i - \delta R_i^*) \leq 0). \tag{5.6}$$

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

$$\mathrm{E}U_1 > 0,\ \mathrm{E}|Z_1|^3 < \infty,\ \mathrm{E}|U_1|^3 < \infty,\ \mathrm{E}(R_1^*)^3 < \infty.$$

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 (5.6) with Markov's inequality to have

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

where $\mathrm{C} := (\mathrm{E}|Y_1|^3 + \sqrt{8}/\pi(\mathrm{E}Y_1^2)^{3/2})/\delta_1^3$, 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.

Now, one notices B_2 implies at least one of the following events:

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

So,

$$\mathrm{P}(B_2) \leq \mathrm{P}(B_{21}) + \mathrm{P}(B_{22}) + \mathrm{P}(B_{23}). \tag{5.7}$$

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

$$\mathrm{P}(G \cap B) \leq \mathrm{P}(G \cap B_1) + \mathrm{P}(B_2) \leq \frac{\mathrm{C}}{n^{3/2}}, \tag{5.8}$$

where 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 $d_- = \dfrac{2\overline{Z}}{,U + \sqrt{\overline{U}^2 - 2\overline{ZR}}}$. Here, the condition $\overline{U} > 0$ is so the denominator of the latter ratio is nonzero. Thus, on the event $G\backslash B$ one has

$$\overline{U} > 0 \text{ and } \hat{\theta} - \theta_0 = \frac{2\overline{Z}}{\overline{U} + \sqrt{\overline{U}^2 - 2\overline{ZR}}} \in [T_-,\ T_+] \tag{5.9}$$

where

$$T_\pm \ := \frac{2\overline{Z}}{\overline{U} + \sqrt{\overline{U}^2 \mp 2|\overline{Z}|\overline{R^*}}}. \tag{5.10}$$

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

Theorem 5.3.1 (Smoothness Condition)**.**

$$|f(\mathrm{x}) - L(\mathrm{x})| \leq \frac{M_\epsilon}{2}\|\mathrm{x}\|^2 \textit{ for all } \mathrm{x} \in \mathbb{R}^d \textit{ with } \|\mathrm{x}\| \leq \epsilon. \tag{5.11}$$

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 $V, V_1, \ . \ . \ . \ , V_n$ 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

$$\tilde{\sigma} := \|L(V)\|_2, v_3 := \|V\|_3, \text{ and } \varsigma_3 := \frac{\|L(V)\|_3}{\tilde{\sigma}}. \tag{5.12}$$

Theorem 5.3.2. *Suppose that the smoothness condition holds and that $\tilde{\sigma} > 0$ and $v_3 < \infty$. Then for all $z \in \mathbb{R}$*

$$|\mathrm{P}(\frac{f(\overline{V})}{\tilde{\sigma}/\sqrt{n}} \leq z) - \Phi(z)| \leq \frac{\mathrm{C}}{\sqrt{n}}, \tag{5.13}$$

where 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*

$$z \in (0,\ \omega\sqrt{n}], \tag{5.14}$$

one has

$$|\mathrm{P}(\frac{f(\overline{V})}{\tilde{\sigma}/\sqrt{n}} \leq z) - \Phi(z)| \leq \frac{\mathrm{C}_\omega}{z^3\sqrt{n}} \tag{5.15}$$

where C_ω is a positive, finite, and only depends on f through the smoothness condition, the moments $\tilde{\sigma}$, ς_3, and v_3, and ω.

Applying bracketing

Now let $d = 3$ and then let

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

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

$$f_\pm(\mathrm{x}) = f_\pm(x_1,\ x_2,\ x_3) = \frac{2x_1}{x_2 + \mathrm{E}U_1 + \sqrt{(x_2 + \mathrm{E}U_1)^2 \mp 2|x_1||x_3 + \mathrm{E}R_1^*|}} \tag{5.16}$$

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

Clearly, $f_\pm(0) = 0$,

$$L_\pm(\mathrm{x}) := f'_\pm(0)(\mathrm{x}) = \frac{x_1}{\mathrm{E}U_1} = \frac{x_1}{I_2(\theta_0)} \tag{5.17}$$

for $\mathrm{x} = (x_1,\ x_2,\ x_3) \in \mathbb{R}^3$, and the smoothness condition (5.11) 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 (5.16) is bounded away from 0 for $\mathrm{x} = (x_1,\ x_2,\ x_3)$ in a neighborhood of 0.

Next, let

$$V_i := (Z_i,\ U_i - \mathrm{E}U_i,\ R_i^* - \mathrm{E}R_i^*) \tag{5.18}$$

for $i = 1, \ldots, n$, with Z_i, U_i, R_i^* as defined in (5.5) and (5.4) . Then, by (5.12), (5.17) , and condition 2 , for $f = f\pm$,

$$\tilde{\sigma} = \sqrt{\frac{\mathrm{E}Z_1^2}{I_2(\theta_0)^2}} = \frac{\sqrt{I_1(\theta_0)}}{I_2(\theta_0)} > 0 \tag{5.19}$$

and $v_3^3 = \mathrm{E}\|V\|^3 < \infty$ by the third and fourth conditions. This shows that all the required conditions for (5.3.2) are satisfied for $f = f \pm \cdot$.

Moreover, by (5.18), (5.16), and (5.10),

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

on the event $G \backslash B$. So, by the inclusion relation in (5.9) (which holds on the event $G \backslash B = (G^{\mathrm{c}} \cup B)^{\mathrm{c}}$, where c denotes the complement) and (5.19) , inequality (5.13) in 5.3.2 implies

$$\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \le z) \le \mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)f_-(\overline{V}) \le z) + \mathrm{P}(G^{\mathrm{c}} \cup B)$$

$$\le \Phi(z) + \frac{\mathrm{C}}{\sqrt{n}} + \mathrm{P}(G^{\mathrm{c}} \cup B)$$

and, quite similarly,

$$\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \le z) \ge \mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)f_+(\overline{V}) \le z) - \mathrm{P}(G^{\mathrm{c}} \cup B)$$

$$\ge \Phi(z) - \frac{\mathrm{C}}{\sqrt{n}} - \mathrm{P}(G^{\mathrm{c}} \cup B)\ ,$$

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 (5.8) that

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

for all real z. Quite similarly, but using (5.15) instead of (5.13) , one has

$$|\mathrm{P}(\sqrt{n/I_1(\theta_0)}I_2(\theta_0)(\hat{\theta} - \theta_0) \le z) - \Phi(z)| \le \frac{\mathrm{C}}{z^3\sqrt{n}} + \mathrm{P}(|\hat{\theta} - \theta_0| > \delta) \tag{5.21}$$

for z as in (5.14).

Given rather standard regularity 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 $\frac{\mathrm{c}}{\sqrt{n}}$, and even with the ("error" term $\frac{\mathrm{c}}{z^3\sqrt{n}}-$ under condition (5.14) . Some details on this can be found in the following section.

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. $$\mathrm{E}\frac{\exp(\ell_{X,Z}(\theta_0 \pm h))}{\exp(\ell_{X,Z}(\theta_0))} < 1.$$

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 , $\mathrm{E}\ell''_X(\theta_0) \neq 0$. 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 $\ell_{X,Z}(\theta)$. So, if $\hat{\theta} > \theta_0 + \delta$, then $\ell_{X,Z}(\theta_0 + h) \geq \ell_{X,Z}(\theta_0)$.

Therefore,

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

$$\leq \mathrm{E}\prod_{i=1}^{n} \sqrt{\frac{\exp(\ell_{X_i,Z_i}(\theta_0 + h))}{\exp(\ell_{X_i,Z_i}(\theta_0)}} = \lambda_+^n,$$

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,

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

for $\lambda := \max(\lambda_+,\ \lambda_-) \in [0, 1)$.

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.

6.2 Introduction

Optimal transport is a framework for defining meaningful metrics between probability measures [38, 29]. These metrics find a wide range of applications, such as generative modeling [19, 12], Bayesian inference [34], imitation learning [15], graph matching and averaging [40, 39]. Multi-marginal optimal transport [18] 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 [13], clustering [7], and computing barycenters of measures [4].

Following the establishment of key theoretical results, including by Gangbo and Świech [18], Agueh and Carlier [1], and Pass [28], research is shifting toward applications. This motivates a need for practical algorithms for the multi-marginal setting [24]. Standard approaches based on linear programming and entropic regularization scale exponentially with the number of measures, and/or the dimension of the space [6, 36]. 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* [11] 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 [11, 10, 22, 26, 16, 33] and Gromov–Wasserstein [37] settings between pairs of measures, but has not yet been applied to settings with more than two measures.

In this paper, 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 [18] is a class of optimization problems for comparing multiple measures $\mu_1, \dots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$, all supported on the metric space $(\mathbb{R}^d, ||\cdot||_2)$. The most common such problem is computing the multi-marginal Wasserstein distance, defined

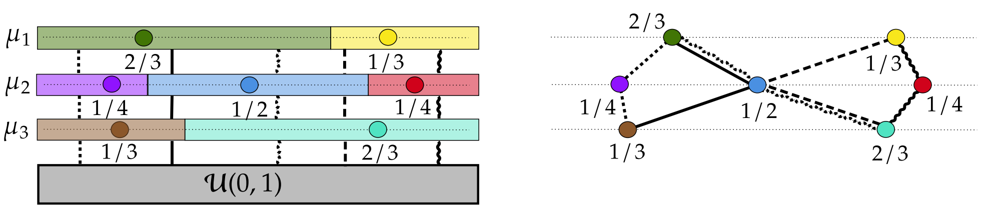

Figure 6.1: Illustration of the optimal coupling's structure on $\mathbb{R}$ between discrete measures μ_1, μ_2 and μ_3. Points are samples of each measures, with weights next to them. Left: histogram of measures (horizontal); joint samples are obtained by sampling a (black) line uniformly (drawn vertically), 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

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

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 Świech [18] and Agueh and Carlier [1], given by

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

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 2-Euclidean cost in the case $P = 2$ (up to constants), referred to as $\mathcal{W}$. Algorithms for estimating (6.3) from a set of samples scale exponentially with the number of measures P and/or the dimension d of the ground space [4, 2, 6].

$\mathcal{MW}$ 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 random axis, which gives [11, 10]

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

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 [11] 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 [**topstatprop**, 11], 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 [16, 15]. This immediately raises the question whether $\mathcal{SW}$ extends to the multi-marginal case so that it preserves its key appealing properties.

Sliced Multi-Marginal Wasserstein Distance

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

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

where $\mu_j^\theta = M^\theta_\#(\mu_j)$ 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 4. *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 4 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 (4) scales in $\mathcal{O}(KP^2 N \log N)$. Furthermore, as the sliced-Wasserstein metric is upper-bounded by the Wasserstein [11], an immediate consequence of Proposition 4 is that

$$\mathcal{SMW}^2(\mu_1, \dots, \mu_P) \overset{(4)}{=} \frac{1}{2} \sum_{i,j=1}^P \beta_i \beta_j \mathcal{SW}^2(\mu_i, \mu_j) \leq \frac{1}{2} \sum_{i,j=1}^P \beta_i \beta_j \mathcal{W}^2(\mu_i, \mu_j).$$

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 5. *$\mathcal{SMW}$ is a generalized metric.*

In particular, this means that $\mathcal{SMW}$ is (i) non-negative, (ii) zero if and only if all measures are identical, (iii) permutation-equivariant, and (iv) satisfies a generalized triangle inequality involving multiple measures. 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 6. *If $\mu_1, \ldots, \mu_P \in \mathcal{M}(\mathbb{R}^d)$, and assuming $\mathcal{W}^2$ has sample complexity $\rho(N)$ on $\mathbb{R}$, then,*

$$\mathrm{E}[\mathcal{SMW}^2(\mu_1, \ldots, \mu_P) - \mathcal{SMW}^2(\mu_1, \ldots, \mu_P)]^2 \leq \frac{1}{2}\rho(N),$$

where μ_p refers to empirical measures with N samples.

Proposition **??** 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.

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

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

where

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.

Chapter 7

Conclusions and Future Work

Here, we provide some suggestions for future directions to take the presented research works.

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] 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.

[6] Jean-David Benamou et al. “Iterative Bregman Projections for Regularized Transportation Problems”. In: *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[7] José Bento and Liang Mi. “Multi-Marginal Optimal Transport Defines a Generalized Metric”. In: *arXiv:2001.11114* (2020).

[8] 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.

[9] 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].

[10] Nicolas Bonneel et al. “Sliced and Radon Wasserstein Barycenters of Measures”. In: *Journal of Mathematical Imaging and Vision* 51.1 (2015), pp. 22–45.

[11] Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation.* 2013.

[12] Charlotte Bunne et al. “Learning Generative Models across Incomparable Spaces”. In: *ICML*. 2019.

[13] Jiezhang Cao et al. "Multi-marginal Wasserstein GAN". In: *NeurIPS*. 2019.

[14] Nicolas Courty et al. "Joint distribution optimal transportation for domain adaptation". In: *Advances in Neural Information Processing Systems* 2017-December.Nips (2017), pp. 3731–3740. ISSN: 10495258. arXiv: `1705.08848`.

[15] Robert Dadashi et al. "Primal Wasserstein Imitation Learning". In: *arXiv:2006.04678* (2020).

[16] Ishan Deshpande et al. "Max-Sliced Wasserstein Distance and Its Use for GANs". In: *CVPR*. 2019.

[17] 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`.

[18] Wilfrid Gangbo and Andrzej Świech. "Optimal maps for the multidimensional Monge-Kantorovich problem". In: *Communications on Pure and Applied Mathematics* 51.1 (1998), pp. 23–45.

[19] Aude Genevay, Gabriel Peyre, and Marco Cuturi. "Learning Generative Models with Sinkhorn Divergences". In: *AISTATS*. 2018.

[20] Aude Genevay et al. "Sample Complexity of Sinkhorn divergences". In: (Oct. 2018). arXiv: `1810.02733`. URL: `http://arxiv.org/abs/1810.02733`.

[21] Evarist Gine and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. 2016. ISBN: 9781107043169. DOI: `10.1017/cbo9781107337862`.

[22] Soheil Kolouri et al. "Generalized Sliced Wasserstein Distances". In: *NeurIPS*. 2019.

[23] 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).

[24] Tianyi Lin et al. "On the Complexity of Approximating Multimarginal Optimal Transport". In: *arXiv:1910.00152* (2019).

[25] 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`.

[26] Khai Nguyen et al. "Distributional Sliced-Wasserstein and Applications to Generative Modeling". In: *ICLR*. 2021.

[27] 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`.

[28] Brendan Pass. "Multi-Marginal Optimal Transport: Theory and Applications". In: *arXiv:1406.0026* (2014).

[29] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* (2019).

[30] 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`.

[31] 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`.

[32] 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`.

[33] Mark Rowland et al. "Orthogonal Estimation of Wasserstein Distances". In: *AISTATS*. 2019.

[34] 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.

[35] 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`.

[36] N. Tupitsa et al. "Multimarginal Optimal Transport by Accelerated Alternating Minimization". In: *CDC* (2020), pp. 6132–6137.

[37] Titouan Vayer et al. "Sliced Gromov-Wasserstein". In: *NeurIPS*. 2019.

[38] Cédric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

[39] Hongteng Xu, Dixin Luo, and Lawrence Carin. "Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching". In: *NeurIPS*. 2019.

[40] Hongteng Xu et al. "Gromov-Wasserstein Learning for Graph Matching and Node Embedding". In: *ICML*. 2019.