Learning and Inference with Wasserstein Metrics

by

Charles Frogner

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

Signature redacted

Author .....
Department of Brain and Cognitive Sciences
January 18, 2018

Signature redacted

Certified by.....
Tomaso Poggio
Eugene McDermott Professor of Brain and Cognitive Sciences
Thesis Supervisor

Signature redacted

Accepted by .....
Matthew Wilson
Chairman, Department Committee on Graduate Theses

Learning and Inference with Wasserstein Metrics

by

Charles Frogner

Submitted to the Department of Brain and Cognitive Sciences
on January 18, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis develops new approaches for three problems in machine learning, using tools from the study of optimal transport (or Wasserstein) distances between probability distributions. Optimal transport distances capture an intuitive notion of similarity between distributions, by incorporating the underlying geometry of the domain of the distributions. Despite their intuitive appeal, optimal transport distances are often difficult to apply in practice, as computing them requires solving a costly optimization problem. In each setting studied here, we describe a numerical method that overcomes this computational bottleneck and enables scaling to real data.

In the first part, we consider the problem of multi-output learning in the presence of a metric on the output domain. We develop a loss function that measures the Wasserstein distance between the prediction and ground truth, and describe an efficient learning algorithm based on entropic regularization of the optimal transport problem. We additionally propose a novel extension of the Wasserstein distance from probability measures to unnormalized measures, which is applicable in settings where the ground truth is not naturally expressed as a probability distribution. We show statistical learning bounds for both the Wasserstein loss and its unnormalized counterpart. The Wasserstein loss can encourage smoothness of the predictions with respect to a chosen metric on the output space. We demonstrate this property on a real-data image tagging problem, outperforming a baseline that doesn't use the metric.

In the second part, we consider the probabilistic inference problem for diffusion processes. Such processes model a variety of stochastic phenomena and appear often in continuous-time state space models. Exact inference for diffusion processes is generally intractable. In this work, we describe a novel approximate inference method, which is based on a characterization of the diffusion as following a gradient flow in a space of probability densities endowed with a Wasserstein metric. Existing methods for computing this Wasserstein gradient flow rely on discretizing the underlying domain of the diffusion, prohibiting their application to problems in more than several dimensions. In the current work, we propose a novel algorithm for computing a Wasserstein gradient flow that operates directly in a space of continuous functions,

free of any underlying mesh. We apply our approximate gradient flow to the problem of filtering a diffusion, showing superior performance where standard filters struggle.

Finally, we study the ecological inference problem, which is that of reasoning from aggregate measurements of a population to inferences about the individual behaviors of its members. This problem arises often when dealing with data from economics and political sciences, such as when attempting to infer the demographic breakdown of votes for each political party, given only the aggregate demographic and vote counts separately. Ecological inference is generally ill-posed, and requires prior information to distinguish a unique solution. We propose a novel, general framework for ecological inference that allows for a variety of priors and enables efficient computation of the most probable solution. Unlike previous methods, which rely on Monte Carlo estimates of the posterior, our inference procedure uses an efficient fixed point iteration that is linearly convergent. Given suitable prior information, our method can achieve more accurate inferences than existing methods. We additionally explore a sampling algorithm for estimating credible regions.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor of Brain and Cognitive Sciences

Acknowledgments

Thanks to my advisor, Tomaso Poggio, for welcoming me into his group and supporting me throughout these years. It was Tommy's class on statistical learning theory that got me hooked on machine learning to begin with and convinced me to give graduate school a shot. I'm grateful for his insights and his flexibility in accomodating my ever-evolving interests.

Thanks to Avi Pfeffer, Harry Lewis, and David Parkes for their support and advice when I was sorting out what to do with myself after college.

Thanks to everyone in the Poggio lab, past and present, for being colleagues and friends, in addition to being some of the sharpest and most industrious people I know. It's been a real pleasure and an education.

Thanks to my friends and classmates for all of the fun times and enlightening discussions. I couldn't have picked a nicer, more interesting, more talented group and you've made it a memorable few years.

And thanks to my parents and my sister for all their love and support. It means a lot to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis studies new approaches to some old problems in machine learning. Multi-output learning, diffusion processes, and ecological inference, in fact, have been studied since before the term "machine learning" was coined [92]. Yet they represent core concerns for machine learning today: modern supervised learning problems, such as image categorization, semantic segmentation, and speech recognition, frequently involve predicting entire sets of labels all together, while systems that represent uncertainty about the state of the world rely heavily on probabilistic modeling. There is ample demand for innovation.

For a fresh perspective on these problems, we delve into a rich and still-evolving set of tools coming from the study of optimal transport of probability measures. The topic has a long history, going back to Monge in 1781 [109]. Monge studied the problem of moving a pile of dirt to fill a hole, in such a way as to minimize the work done in moving it. If $\mathcal{X}$ is the domain on which we've piled the dirt, μ is the mass distribution of the dirt (a measure on $\mathcal{X}$), and $T : \mathcal{X} \to \mathcal{X}$ is our plan for moving it, the amount of work done is the total distance traveled, weighted by the mass moved,

$$W(T) = \int_{\mathcal{X}} \|x - T(x)\| d\mu(x).$$

Monge's problem was to find a map T such that the push-forward of μ via T exactly matches the shape of the hole, and such that the work $W(T)$ is minimized.

We can see the flexibility of optimal transport already from Monge's problem. For registering two images, for example, we can imagine transporting pixels from one image to another so as to minimize the distortion [111]. For transferring colors from one image to another, we might represent the distribution of colors in each image in a color space (such as RGB) and transport one color distribution to match the other [59]. And for domain adaptation in machine learning, we might transport the data distribution from one domain to match that in the other domain [119] [137]. There are numerous possiblities.

In this thesis, we rely on a relaxation of Monge's problem due to Kantorovich [85] [86], which we describe in Section 1.2. This relaxation is the basis for a distance function between probability measures, called the Wasserstein distance, defined by the minimal amount of work done in transporting one measure to match the other. The Wasserstein distance will be fundamental to the work in Chapters 2 and 3. It is in many respects the natural way to define a distance between two probability measures, and its application in machine learning is the subject of much recent work. We survey the machine learning applications of the Wasserstein distance in Section 1.3. In Section 1.4, we conclude the chapter with an overview of the contributions of this thesis.

1.1 Notation

Unless otherwise noted, $\mathcal{X}$ and $\mathcal{Y}$ are separable complete metric spaces. $\mathcal{M}_+(\mathcal{X})$ is the set of nonnegative Radon measures on $\mathcal{X}$ and $\mathcal{P}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$, $\mathcal{P}(\mathcal{X}) = \{\mu \in \mathcal{M}_+(\mathcal{X}) | \mu(\mathcal{X}) = 1\}$. Given a joint probability measure π on the product space $\mathcal{X} \times \mathcal{Y}$, its marginals are the measures $\mathrm{P}_1 \pi \in \mathcal{P}(\mathcal{X})$ and $\mathrm{P}_2 \pi \in \mathcal{P}(\mathcal{Y})$ defined by

$$(\mathrm{P}_1 \pi)(A) = \pi(A \times \mathcal{Y}), \quad (\mathrm{P}_2 \pi)(B) = \pi(\mathcal{X} \times B),$$

for $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$ measurable subsets of $\mathcal{X}$ and $\mathcal{Y}$.

The entropy of a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ is

$$H(\mu) = -\int_{\mathcal{X}} d\mu(x) \log d\mu(x)$$

with $d\mu$ the density with respect to the Lebesgue measure on $\mathcal{X}$. We let $H(\mu) = +\infty$ when such a density does not exist. The KL divergence between nonnegative measures $\mu \in \mathcal{M}_+(\mathcal{X})$ and $\nu \in \mathcal{M}_+(\mathcal{X})$ is

$$\mathrm{KL}(\mu \| \nu) = \int_{\mathcal{X}} d\mu(x)(\log \frac{d\mu}{d\nu}(x) - 1) + d\nu(x)$$

with $d\mu, d\nu$ the densities with respect to the Lebesgue measure on $\mathcal{X}$ and $\frac{d\mu}{d\nu}$ the relative density of μ with respect to ν. We define $\mathrm{KL}(\mu \| \nu) = +\infty$ whenever any of the derivatives $d\mu, d\nu$ or $\frac{d\mu}{d\nu}$ do not exist.

$\mathbb{R}_+$ is the set of nonnegative reals, while $\mathbb{R}_{++}$ are positive reals. Δ^d is the d-dimensional simplex, $\Delta^d = \{\mathbf{u} \in \mathbb{R}_+^d | \sum_{i=1}^d \mathbf{u}_i = 1\}$. For matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$, $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathcal{F}}$ is the Frobenius inner product.

1.2 Optimal transport and Wasserstein metric

Optimal transport

Suppose we have two probability measures μ and ν, defined on domains $\mathcal{X}$ and $\mathcal{Y}$, respectively. The theory of **optimal transport** [156] studies ways of relocating the mass of μ to match that of ν. It defines a **transport plan** π that is a measure on the product space $\mathcal{X} \times \mathcal{Y}$, whose values determine the amount of mass transferred between any pair of measurable subsets of the respective domains. Concretely, $\pi(A \times B)$ is the amount of mass transferred between subsets A and B, and π satisfies the condition that its **marginals** match μ and ν,

$$(\mathrm{P}_1 \pi)(A) = \mu(A) \quad (\mathrm{P}_2 \pi)(B) = \nu(B), \tag{1.1}$$

for all measurable $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$. Equivalently, for all pairs of test functions $\varphi \in L^1(d\mu)$ and $\psi \in L^1(d\nu)$,

$$\int_{\mathcal{X}\times\mathcal{Y}} (\varphi(x) + \psi(y)) d\pi(x,y) = \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y). \tag{1.2}$$

Note that π is necessarily a probability measure. Another way to look at it is that π specifies a joint probability distribution whose marginals are μ and ν.

For any pair of probability measures μ and ν, define the set of valid transport plans to be $\Pi(\mu,\nu)$,

$$\Pi(\mu,\nu) = \\ \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) | \, \mathrm{P}_1 \pi(A) = \mu(A) \;\&\; \mathrm{P}_2 \pi(B) = \nu(B), \; \forall \text{ measurable } A \subseteq \mathcal{X}, B \subseteq \mathcal{Y}\}. \tag{1.3}$$

$\Pi(\mu,\nu)$ is always nonempty, containing at least the product measure $\mu \otimes \nu$. In fact, it may contain an infinity of possible transport plans.

Optimal transport is concerned with identifying a transport plan that minimizes a total **cost** of transporting the mass. Concretely, we define a nonnegative, measurable function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{+\infty\}$ whose value $c(x,y)$ gives the cost of transporting a unit of mass from location $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. The total cost for a plan π, then, is

$$C(\pi) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y), \tag{1.4}$$

and optimizing this cost subject to the marginal constraints is known as the **Kantorovich optimal transport problem** [85] [86],

$$\pi_* = \operatorname*{arginf}_{\pi\in\Pi(\mu,\nu)} C(\pi). \tag{1.5}$$

Example 1.2.1 (Discrete measures)**.** We will frequently be concerned with optimal transport of discrete measures, in which the distributions being compared are supported at a finite number of discrete locations $\{\mathbf{x}^{(i)}\}_{i=1}^m \subseteq \mathcal{X}$ and $\{\mathbf{y}^{(j)}\}_{j=1}^n \subseteq \mathcal{Y}$. In

this case, the measures can written as weighted sums of delta functions,

$$\mu = \sum_{i=1}^{m} \mathbf{u}_i \delta_{\mathbf{x}^{(i)}} \quad \nu = \sum_{j=1}^{n} \mathbf{v}_j \delta_{\mathbf{y}^{(j)}}, \tag{1.6}$$

with $\mathbf{u} \in \Delta^m$ and $\mathbf{v} \in \Delta^n$ the vectors of weights. Any valid transport plan π then has a similar representation,

$$\pi = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{T}_{ij} \delta_{(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})}, \tag{1.7}$$

with $\mathbf{T} \in \Delta^{m \times n}$ the matrix of weights that define the transport plan. In fact, we can equally well represent the optimal transport problem in terms of $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{T}$, rather than μ, ν, and π. The marginal constraints (1.1) become

$$\mathbf{T}\mathbf{1} = \mathbf{u}, \quad \mathbf{T}^\intercal \mathbf{1} = \mathbf{v}, \tag{1.8}$$

with $\mathbf{1}$ the all-ones vector of appropriate dimension, and the set of valid plans is the intersection of these constraints with the nonnegative orthant,

$$\Pi(\mathbf{u}, \mathbf{v}) = \{\mathbf{T} \in \mathbb{R}_+^{m \times n} | \mathbf{T}\mathbf{1} = \mathbf{u}, \mathbf{T}^\intercal \mathbf{1} = \mathbf{v}\}. \tag{1.9}$$

$\Pi(\mathbf{u}, \mathbf{v})$ is called the **transport polytope** [1], and it is closed, bounded and convex. The marginal constraints (1.8) confine $\Pi(\mathbf{u}, \mathbf{v})$ to an affine subspace of dimension $(m-1)(n-1)$ [2]. The polytope is non-empty, containing at least the plan $\mathbf{u}\mathbf{v}^\intercal$.

Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be a cost function and let $\mathbf{C} \in \mathbb{R}_+^{m \times n}$ be the matrix whose entries are $\mathbf{C}_{ij} = c(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})$. The optimal transport problem (1.5) can then be written

$$\mathbf{T}_* = \underset{\mathbf{T} \in \Pi(\mathbf{u}, \mathbf{v})}{\operatorname{argmin}} \langle \mathbf{C}, \mathbf{T} \rangle_{\mathcal{F}}. \tag{1.10}$$

(1.10) is a linear program with $m + n - 1$ linear equality and mn linear inequality

[1] In statistics, $\Pi(\mathbf{u}, \mathbf{v})$ is sometimes known as the set of contingency tables or two-way tables with fixed margins [52].

[2] $(m-1)(n-1) = mn - (m+n-1)$. Note that one of the affine constraints is redundant, as conservation of mass determines one of the marginal values.

constraints. As a linear minimization over a convex compact set, it admits at least one solution which is an extreme point of the set [3]. This solution need not be unique.

Kantorovich duality

Kantorovich [85] defined a dual formulation of the optimal transport problem (1.5) that coincides with the primal at optimality. For dual variables $(\varphi, \psi) \in L^1(d\mu) \times L^1(d\nu)$, the dual objective is

$$D(\varphi, \psi) = \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y). \tag{1.11}$$

The duality is as follows.

Theorem 1 (Kantorovich duality, [156] Thm. 1.3). *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{+\infty\}$ be lower semi-continuous. Define C and Π as in (1.4) and (1.3), respectively, and let $\Psi(c)$ be the set of measurable pairs $(\varphi, \psi) \in L^1(d\mu) \times L^1(d\nu)$ satisfying*

$$\varphi(x) + \psi(y) \leq c(x, y), \tag{1.12}$$

for $d\mu$-almost all $x \in \mathcal{X}$ and $d\nu$-almost all $y \in \mathcal{Y}$. Then

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = \sup_{(\varphi,\psi) \in \Psi(c)} D(\varphi, \psi). \tag{1.13}$$

The infimum in (1.13) is attained.

Example 1.2.2 (Discrete duality). In the case of discrete measures $\mu = \sum_{i=1}^m \mathbf{u}_i \delta_{\mathbf{x}^{(i)}}$ and $\nu = \sum_{j=1}^n \mathbf{v}_j \delta_{\mathbf{y}^{(j)}}$, Kantorovich duality reduces to the standard linear programming duality [20]. In this case, we have dual vectors $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^n$ and a dual objective

$$D(\alpha, \beta) = \alpha^\intercal \mathbf{u} + \beta^\intercal \mathbf{v}, \tag{1.14}$$

[3] An extreme point of a convex set is one that cannot be written as a convex combination of any other pair of points from the set.

which is maximized over α and β lying in the polyhedron

$$\Psi(\mathbf{C}) = \{(\alpha, \beta) \in \mathbb{R}^m \times \mathbb{R}^n | \alpha_i + \beta_j \leq \mathbf{C}_{ij}, \forall i, j\}. \tag{1.15}$$

The duality is expressed

$$\min_{\mathbf{T} \in \Pi(\mathbf{u}, \mathbf{v})} \langle \mathbf{C}, \mathbf{T} \rangle_{\mathcal{F}} = \max_{(\alpha, \beta) \in \Psi(\mathbf{C})} D(\alpha, \beta), \tag{1.16}$$

with both optima attained.

Wasserstein metric

We are particularly interested in the case where μ and ν are probability measures on a single domain $\mathcal{X}$ and the cost c derives from a metric on the domain. In this case, the cost of the optimal transport plan can be used to define a metric on probability measures, called the Wasserstein metric.

Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a metric on $\mathcal{X}$, and define the cost $c(x, y) = d(x, y)^p$, for $p \in [0, +\infty)$. The associated optimal transport cost is

$$\mathcal{T}_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi(x, y). \tag{1.17}$$

We define the **Wasserstein distance** between μ and ν as follows.

Definition 1.2.1 (Wasserstein distance)**.**

1. For $p \in [1, +\infty)$, $\mathcal{W}_p(\mu, \nu) = \mathcal{T}_p(\mu, \nu)^{1/p}$.

2. For $p \in [0, 1)$, $\mathcal{W}_p(\mu, \nu) = \mathcal{T}_p(\mu, \nu)$.

The case $p = 1$ is sometimes called the **Earth Mover's Distance** [125]. In the discrete setting, we will abuse notation and write $\mathcal{W}_p(\mathbf{u}, \mathbf{v})$ for the Wasserstein distance between the discrete measures having $\mathbf{u}$ and $\mathbf{v}$ as their weight vectors.

The Wasserstein distance is in fact a metric on probability measures, subject to a moment condition.

Theorem 2 (Wasserstein is a metric, [156] Thm. 7.3). *Let $\mathcal{P}_p(\mathcal{X})$ be the set of probability measures on $\mathcal{X}$ having finite moments of order p,*

$$\mathcal{P}_p(\mathcal{X}) = \left\{ \mu \in \mathcal{P}(\mathcal{X}) | \int_{\mathcal{X}} d(x, x_0)^p d\mu(x) < +\infty, \forall x_0 \in \mathcal{X} \right\}. \tag{1.18}$$

Then for all $p \in [0, +\infty)$, $\mathcal{W}_p$ is a metric on $\mathcal{P}_p(\mathcal{X})$.

Example 1.2.3 (Total variation distance). In the case of $p = 0$, we have $d(x, y)^p = \mathbb{1}_{x \neq y}$. In other words, the cost is 0 if $x = y$ and 1 otherwise. Then the Wasserstein distance coincides with total variation,

$$\mathcal{W}_0(\mu, \nu) = \frac{1}{2} \|\mu - \nu\|_{\mathrm{TV}}, \tag{1.19}$$

for any $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

1.3 Wasserstein metrics in machine learning

Wasserstein metrics are potentially useful wherever one wants to compare probability measures defined on a metric space. Such settings occur frequently in machine learning. In natural language processing, computer vision, and bioinformatics, for example, data of interest are often represented as histograms or "bags" of features, such as SIFT features for images [105], bags of words or topic allocations for text [80] [25], and counts of n-grams for sequences [100]. In all of these cases we may have some notion of relatedness or similarity between the features, which can be encoded as a metric. We will give more examples in what follows.

In such settings, Wasserstein metrics differentiate themselves from other divergences – such as the Hellinger, χ_2, total variation, or Kullback-Leibler – in that they take into consideration the underlying metric of the domain. They do so in an intuitive way: the Wasserstein distance between two distributions for $p = 1$ (also known as the Earth Mover's Distance) is the minimal total distance traveled when moving the mass in one distribution to match the other, for example. Whereas pointwise divergences, such as those mentioned, will assign a large distance to measures with

distinct supports, Wasserstein distances will be large or small depending on the underlying metric distance between their supports. The Wasserstein distance between two point masses, for any $p \geq 1$, is simply their metric distance, for example.

Computational cost

A major obstacle to the application of Wasserstein distances is the fact that computing them requires solving a nontrivial optimization problem. The vast majority of existing applications of optimal transport distances involve discrete measures, for which the distance computation is a linear program with a number of constraints scaling as the product of the cardinalities of the supports of the two measures. This linear program can be solved via the network simplex method [62], interior point methods [113], or through more specialized algorithms such as network flow [2] [118]. For general underlying metrics, the complexity of interior point methods is $\mathcal{O}(n^3 \log n)$, where n is the larger of the two support cardinalities, while these and the network simplex are supercubic in practice [118] [125]. The execution time for computing a single distance rapidly becomes impractical, requiring minutes [4] for measures of cardinality larger than 1000.

Two approaches have been suggested, for practical applications with discrete measures. The first is to restrict to special metrics and special arrangements of the underlying support points that allow for faster algorithms. When the domain is $\mathcal{X} = \mathbb{R}$, for example, the optimal transport plan for the Euclidean cost is a monotone rearrangement whose computation is $\mathcal{O}(n \log n)$. For more general domains, Pele and Werman [118] use underlying metrics that are thresholded, saturating at a constant maximum value, and show that one can reduce the complexity of the network flow computation by an order of magnitude. Ling and Okada [101] assume the underlying metric is L^1 and the points lie on a Manhattan network, showing that the number of variables in the linear program can be reduced to $\mathcal{O}(n)$, and achieving time complexity scaling as n^2 in practice. Other methods for the L^1 metric approximate the Wasserstein distance by embedding the input measures in a different norm [79] [106], or by using

[4] Wall clock time from [45] run on a single core 2.66 Ghz Xeon.

sums of wavelet coefficients computed on the difference measure [139], for example.

The second approach is to regularize the optimal transport problem. In particular, entropic regularization is the focus of much current work, as it enables very efficient algorithms. Entropic regularization of the transport problem has a long history, going back at least to Schrödinger [135] (see also [99]), and it has seen a variety of applications outside machine learning [162] [146] [65]. It was introduced to the machine learning community by Cuturi [45].

The entropic regularizer penalizes the negative entropy of the transport plan, yielding a modified transport problem

$$\pi_\gamma = \operatorname*{arginf}_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) - \gamma H(\pi), \tag{1.20}$$

with $H(\pi) = -\int_{\mathcal{X}\times\mathcal{Y}} d\pi(x,y) \log d\pi(x,y)$ the entropy and $\gamma > 0$ the regularization parameter. While the original transport problem (1.5) may have many solutions, the regularized problem is strictly convex, ensuring uniqueness of the optimum.

Although the entropy regularizer acts as a barrier function enforcing positivity, it is perhaps not the most obvious choice in the sense of interior point methods [113]. The regularizer is effective instead due to its algebraic properties. In particular, the regularized transport problem (1.20) can be written as a projection, with respect to the Kullback-Leibler divergence, of a Gibbs measure onto the set of valid transport plans $\Pi(\mu,\nu)$. When the Lebesgue integral on $\mathcal{X}\times\mathcal{Y}$ is well-defined, the initial Gibbs measure ξ has a closed form,

$$\xi(A\times B) = \int_{A\times B} \exp\left(-\frac{1}{\gamma} c(x,y)\right), \tag{1.21}$$

for measurable $A \subseteq \mathcal{X}$, $B \subseteq \mathcal{Y}$. The regularized transport problem is expressed

$$\pi_\gamma = \operatorname*{arginf}_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi \| \xi). \tag{1.22}$$

In the case of discrete measures, we are solving a convex program,

$$\mathbf{T}_\gamma = \underset{\mathbf{T}\in\Pi(\mathbf{u},\mathbf{v})}{\operatorname{argmin}} \langle \mathbf{C}, \mathbf{T} \rangle_{\mathcal{F}} - \gamma H(\mathbf{T}) = \underset{\mathbf{T}\in\Pi(\mathbf{u},\mathbf{v})}{\operatorname{argmin}} \mathrm{KL}(\mathbf{T} \| \mathbf{K}), \tag{1.23}$$

with $H(\mathbf{T}) = -\langle \mathbf{T}, \log \mathbf{T} \rangle_{\mathcal{F}}$ and $\mathbf{K} = \exp\left(-\frac{1}{\gamma}\mathbf{C}\right)$ the kernel associated to the Gibbs measure (1.21).

As this is a KL projection onto the intersection of affine constraints, the regularized optimal transport problem is amenable to a very efficient method of optimization, called Bregman's method [29]. In Bregman's method, one alternately projects the Gibbs measure onto the marginal constraints for μ and for ν, until a fixed point is reached. In particular, we begin with $\mathbf{K}$ and alternately project the matrix onto the row and column sum constraints. Key to the efficiency is that each one of these marginal projections is simply a left- or right-multiplication by a diagonal matrix. The resulting algorithm is exactly Sinkhorn's algorithm for matrix scaling [142] [143] [144], which is known to converge linearly to the fixed point [63] [90]. Moreover, it is easily parallelized [45]. More recently, [67] have improved upon the efficiency of Sinkhorn's algorithm via stochastic optimization.

Wasserstein embeddings and nearest neighbors

Wasserstein distances are directly applicable in nearest neighbor algorithms, in which objects of interest (such as images and text) are embedded in a space of histograms and queries are made for the objects that are nearest in the sense of the Wasserstein distance in the embedding space. The Earth Mover's Distance (EMD) – i.e. the $1-$Wasserstein distance – has been widely applied in this context. EMD has been used for image and texture retrieval [125] [106] [163], comparing histograms of color, texture, shape and spatial position information, and for image keypoint matching [117] [41] [101], in which the compared histograms represent local image regions. [97] uses the EMD between image descriptors for both retrieval and nearest neighbor classification. [96] compute the EMD between normalized bag-of-words representations of text documents and use it for nearest neighbor classification.

Wasserstein embeddings are used in variety of other contexts, as well. For edge and corner detection, [129] use the 1-Wasserstein distance between histograms describing adjacent image regions to make judgments of local contrast and orientation. For computing transport distances between large numbers of images, [160] suggest embedding into a space in which the Euclidean distance approximates the $2-$Wasserstein distance. [13] [115] [154] apply this idea to biomedical image analysis, embedding images showing cell morphology before applying PCA and LDA in this space.

Central to using a Wasserstein embedding is the definition of the underlying ground metric relating histogram bins. [46] show how to learn the metric from a set of histograms and a similarity matrix for those histograms.

Wasserstein PCA

Defining an equivalent of PCA in Wasserstein space is challenging. In order to adapt the standard definition of PCA on Riemannian manifolds, one needs to define a bijection locally between the manifold of probability measures and the Wasserstein tangent space, and no such bijection is known. [24] and [26] approach the problem as an optimization over curves in $\mathcal{P}(\mathcal{X})$, focusing on restricted settings: $\mathcal{P}(\mathbb{R})$ in the first case and a parameterized family of densities in the second. [160] proposes embedding into a space in which the Euclidean distance approximates the Wasserstein distance, before performing PCA. [136] define a principal geodesic analysis (PGA), in which one finds parameterized geodesics that minimize the $2-$Wasserstein distance to the measures in the dataset, and solve it approximately. [36] take a similar approach but solve the PGA problem exactly, using proximal methods.

Wasserstein kernels

The Wasserstein distance has also been used to define kernels for a variety of applications. [165] [45] [49] [53] suggest to use a generalized Gaussian kernel,

$$\kappa(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{1}{\sigma}\mathcal{W}_p(\mathbf{u}, \mathbf{v})^p\right), \tag{1.24}$$

for $\sigma > 0$ the bandwidth and histograms $\mathbf{u} \in \Delta^m$ and $\mathbf{v} \in \Delta^n$ representing the objects being compared. [165] and [45] use this kernel for classifying images, while [49] does the same for classifying EEG signals, and [53] for brain connectomes. Note that the generalized Gaussian kernel with a Wasserstein distance is not positive definite in general, as this requires the metric space induced by the Wasserstein distance to have zero curvature [58]. To address this, [93] use a *sliced Wasserstein distance* [121] [28], defined by integrating Wasserstein distances between one-dimensional projections of the input distributions, and show that the generalized Gaussian kernel using the sliced Wasserstein distance (with $p = 2$) is positive definite, before showing applications to kernel k-means, kernel PCA, and kernel SVM classification. [44] alternatively suggests to use the permanent of the transport polytope $\Pi(\mathbf{u}, \mathbf{v})$ in place of the generalized Gaussian kernel, guaranteeing positive definiteness. [66] show that certain types of metrics underlying the Wasserstein distance can guarantee positive definiteness, as well.

Wasserstein barycenters

Wasserstein distances have also been used in objectives for variational problems in machine learning. The prototypical example is the Wasserstein barycenter problem [1], which is that of computing the "mean" of a set of probability measures, defined as a measure that minimizes the sum of Wasserstein distances to the original measures. [121] propose the *sliced Wasserstein distance*, defined by integrating Wasserstein distances between one-dimensional projections of the input distributions, and apply it to compute the barycenter of discrete measures on $\mathbb{R}^d$ (possibly limited to $d < 4$). [28] look at more general spaces and additionally define a Radon barycenter that combines barycenters computed on one-dimensional projections. [47] give a barycenter method for discrete measures on general metric spaces, using entropy-regularized Wasserstein distances, while [48] and [17] give alternative methods for the same formulation, achieving greater efficiency. [7] give a sparse linear programming formulation of the discrete barycenter problem. [149] describe a fully data-parallel barycenter method that operates on continuous measures. Wasserstein barycenters appear in

a number of applications, including shape and texture interpolation [146], fusion of posterior distributions in Bayesian inference [148], clustering distributions [164], and multi-target tracking [14]. [27] study the problem of projecting a discrete probability measure onto a weighted barycenter of some fixed measures.

Generative models

Wasserstein metrics are well-suited to density estimation and generative modeling. In many generative settings, the model defines a nonlinear mapping from a low-dimensional random vector into a high-dimensional observation space, implying a singular density that has positive probability only on a low-dimensional subset of the observation space. The Wasserstein metric is natural as a fitting criterion in such a setting, as it uses the geometry of the space. The resulting estimator is called the minimum Kantorovich estimator [11] [68] and has been implemented in restricted Boltzmann machines using entropic regularization [110], in a variety of parametric statistical models using an approximate Bayesian method [19], in neural networks using entropic regularization [69], and in neural networks via a rough approximation of the Kantorovich dual problem [9]. The last of these is also known as the Wasserstein GAN (Generative Adversarial Network), due to an interpretation of the Kantorovich dual problem (1.13) as optimizing an "adversarial" discriminative mapping. In particular, for the $1-$Wasserstein distance, the Kantorovich dual has a particular form which is

$$\mathcal{W}_1(\mu,\nu) = \sup_{\|\varphi\|_{\text{Lip}}\leq 1} \int_{\mathcal{X}} \varphi(x) d(\mu-\nu)(x). \tag{1.25}$$

Letting μ be the empirical distribution to which we are fitting the model and $\nu = \nu_\theta$ the model density, we see that we are optimizing a $1-$Lipschitz function φ to have positive values where $\mu(x) > \nu_\theta(x)$ and negative values where $\mu(x) < \nu_\theta(x)$ – in other words, it should be maximally discriminative between the two distributions. If we optimize the model ν_θ to minimize the $1-$Wasserstein distance, then, we have a minimax problem with the inner optimization over discriminative mappings φ. Note that the authors in [9] do not actually compute (1.25) but rather make a rough approx-

imation in which, rather than optimizing φ over 1-Lipschitz functions, they instead optimize φ over neural networks whose weights have a box constraint. Although their formulation is effective for fitting the model, there is no evidence this approximates the Wasserstein distance in any meaningful way.

Other variational problems

Other applications to variational problems have been described. [60], for example, propose a variation on Fisher discriminant analysis that maximizes a ratio of inter-class and intra-class Wasserstein distances. And several works have used a Wasserstein distance as the fitting criterion for dictionary learning [130] [48] [123] [134].

1.4 Contributions of this thesis

This thesis addresses three problems in machine learning. We develop new methods for each problem, using techniques from the study of Wasserstein distances. We characterize these new methods both experimentally and theoretically. In two cases we also propose novel numerical methods, for computing a new variant of the Wasserstein distance and for computing an object called a Wasserstein gradient flow.

In Chapter 2, we consider the problem of multi-output learning. We introduce a new method for incorporating into the learning problem a metric or similarity structure on the dimensions of the output, via a Wasserstein loss function. The Wasserstein loss measures the Wasserstein distance between predictions and ground truth, with the underlying ground metric chosen by the user. While the Wasserstein distance is constrained to comparing normalized probability vectors, in many learning problems it is more natural to express the outputs as unnormalized vectors. We propose a novel extension of the Wasserstein distance that operates on unnormalized measures, which can likewise function as a loss function for learning. We derive novel statistical learning bounds for this new, unnormalized Wasserstein loss function. We also characterize the Wasserstein loss empirically. First, we demonstrate that it has a "semantic smoothing" effect, in which the predictions are distributed over output

categories that are nearby to the ground truth, with respect to the underlying ground metric. We then demonstrate that, with a well-chosen ground metric, the Wasserstein loss can improve prediction performance on real data, applying it to an image tagging problem using the Yahoo Flickr Creative Commons dataset, and showing that it outperforms a baseline that doesn't use a metric on the outputs.

In Chapter 3, we consider the problem of probabilistic inference for diffusion processes. A diffusion process is a continuous-time, continuous-space Markov process whose time-dependent density evolves according to a PDE called the Fokker-Planck equation. Exact inference involves solving this PDE, which is intractable in general. Current methods for approximate inference either make parametric assumptions or are limited to low-dimensional domains. We propose a novel method for approximate inference that computes what is called a Wasserstein gradient flow, which is the limit of implicit Euler steps (in a space of probability densities) taken with respect to the Wasserstein metric. Computing this Euler step is an infinite-dimensional optimization over probability densities, and we propose a novel finite-dimensional approximation that nevertheless computes a continuous density. We derive a dual formulation of the Euler step that can be interpreted as maximizing an expectation of a functional on the dual variables, and this enables us to approximate the dual by Monte Carlo sampling. We show that this stochastic approximation necessarily has a finite-dimensional solution, and prove that it converges with increasing numbers of samples to the exact solution for the dual. We also show that the stochastic approximation can be stablized by Tikhonov regularization, and demonstrate empirically the quality of the resulting approximation for different numbers of Monte Carlo samples. We apply our inference method to the problem of filtering a hidden diffusion process, showing that the proposed method can produce more accurate posterior state distributions than existing methods, including classical filters such as the unscented and extended Kalman filters.

In Chapter 4, we consider the ecological inference problem. Mathematically, this is the problem of recovering a joint distribution of two random variables, given only their marginal distributions. It appears frequently in social sciences, in which one wants

to combine aggregate measurements of a population (the marginal distributions) into a more refined characterization (the joint distribution). A prominent example is in studying election data, where one wants to characterize the political preferences of different demographic groups, but only has access to aggregate census data and vote counts. As ecological inference is an ill-posed problem, one has to apply prior assumptions to select a solution. We propose a novel framework for ecological inference that incorporates a variety of different priors, and enables efficient inference of the most probable solution, via a linearly-convergent fixed point iteration. We examine methods for estimating the prior distribution from side information, showing empirically that the proposed inference method using the estimated prior can be significantly more accurate than existing methods, on both synthetic and real election data. We additionally propose a method for interval estimation, in which we determine the boundary of the highest probability density credible region, for a desired threshold probability level.

Chapter 2

Learning with a Wasserstein Loss

2.1 Introduction

We consider the problem of learning to predict a non-negative measure over a finite set. This problem includes many common machine learning scenarios. In multiclass classification, for example, one often predicts a vector of scores or probabilities for the classes. And in semantic segmentation [103], one can model the segmentation as being the support of a measure defined over the pixel locations. Many problems in which the output of the learning machine is both non-negative and multi-dimensional might be cast as predicting a measure.

We specifically focus on problems in which the output space has a natural metric or similarity structure, which is known (or estimated) *a priori*. In practice, many learning problems have such structure. In the ImageNet Large Scale Visual Recognition Challenge [ILSVRC] [128], for example, the output dimensions correspond to 1000 object categories that have inherent semantic relationships, some of which are captured in the WordNet hierarchy that accompanies the categories. Similarly, in the keyword spotting task from the IARPA Babel speech recognition project, the outputs correspond to keywords that likewise have semantic relationships. In what follows, we will call the similarity structure on the label space the *ground metric* or *semantic similarity*.

Using the ground metric, we can measure prediction performance in a way that

is sensitive to relationships between the different output dimensions. For example, confusing dogs with cats might be more severe an error than confusing breeds of dogs. A loss function that incorporates this metric might encourage the learning algorithm to favor predictions that are, if not completely accurate, at least semantically similar to the ground truth.

In this paper, we develop a loss function for multi-label learning that measures the *Wasserstein distance* between a prediction and the target label, with respect to a chosen metric on the output space. The Wasserstein distance is defined as the cost of the optimal transport plan for moving the mass in the predicted measure to match that in the target, and has been applied to a wide range of problems, including barycenter estimation [47], label propagation [147], and clustering [40]. To our knowledge, this paper represents the first use of the Wasserstein distance as a loss for supervised learning.

We briefly describe a case in which the Wasserstein loss improves learning performance. The setting is a multi-class classification problem in which label noise arises from confusion of semantically near-equivalent categories. Figure 2-1 shows such a case from the ILSVRC, in which the categories *Siberian husky* and *Eskimo dog* are nearly indistinguishable. We synthesize a toy version of this problem by identifying categories with points in the Euclidean plane and randomly switching the training labels to nearby classes. The Wasserstein loss yields predictions that are closer to the ground truth, robustly across all noise levels, as shown in Figure 2-2. The standard multiclass logistic loss is the baseline for comparison. Section 2.10.1 describes the experiment in more detail.

Siberian husky Eskimo dog

Figure 2-1: Semantically near-equivalent classes in ILSVRC

The main contributions of this work are as follows. We formulate the problem of learning with prior knowledge of the ground metric, and propose the Wasserstein loss as an alternative to traditional information divergence-based loss functions. Specifi-

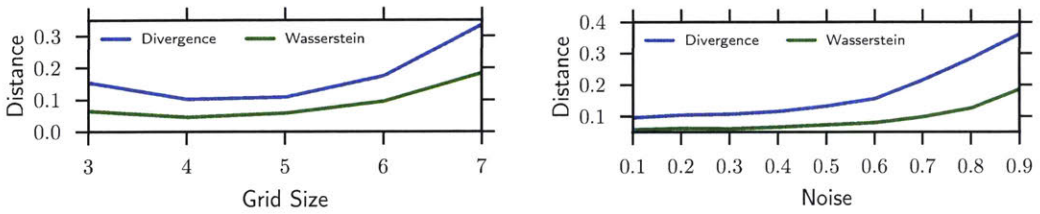


Figure 2-2: The Wasserstein loss encourages predictions that are similar to ground truth, robustly to incorrect labeling of similar classes (see Section 2.10.1). Shown is Euclidean distance between prediction and ground truth vs. (left) number of classes, averaged over different noise levels and (right) noise level, averaged over number of classes. Baseline is the multiclass logistic loss.

cally, we focus on empirical risk minimization (ERM) with the Wasserstein loss, and describe an efficient learning algorithm based on entropic regularization of the optimal transport problem. We also describe a novel extension to unnormalized measures that is similarly efficient to compute. We then justify ERM with the Wasserstein loss by showing a statistical learning bound. Finally, we evaluate the proposed loss on both synthetic examples and a real-world image annotation problem, demonstrating benefits for incorporating an output metric into the loss.

2.2 Related work

Decomposable loss functions like KL Divergence and ℓ_p distances are very popular for probabilistic [103] or vector-valued [6] predictions, as each component can be evaluated independently, often leading to simple and efficient algorithms. The idea of exploiting smoothness in the label space according to a prior metric has been explored in many different forms, including regularization [127] and post-processing with graphical models [38]. Optimal transport provides a natural distance for probability distributions over metric spaces. In [47, 48], the optimal transport is used to formulate the Wasserstein barycenter as a probability distribution with minimum total Wasserstein distance to a set of given points on the probability simplex. [147] propagates histogram values on a graph by minimizing a Dirichlet energy induced by optimal transport. The Wasserstein distance is also used to formulate a metric for

comparing clusters in [40], and is applied to image retrieval [126], contour matching [74], and many other problems [139, 55]. However, to our knowledge, this is the first time it is used as a loss function in a discriminative learning framework. The closest work to this paper is a theoretical study [12] of an estimator that minimizes the optimal transport cost between the empirical distribution and the estimated distribution in the setting of statistical parameter estimation.

2.3 Basics

2.3.1 Problem setup and notation

We consider the problem of learning a map from $\mathcal{X} \subseteq \mathbb{R}^D$ into the space $\mathcal{Y} = \mathbb{R}_+^K$ of nonnegative measures over a finite set $\mathcal{K}$ of size $|\mathcal{K}| = K$. Assume $\mathcal{K}$ possesses a metric $d_{\mathcal{K}}(\cdot, \cdot)$, which is called the *ground metric*. $d_{\mathcal{K}}$ measures semantic similarity between dimensions of the output, which correspond to the elements of $\mathcal{K}$. We perform learning over a hypothesis space $\mathcal{H}$ of predictors $h_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by $\theta \in \Theta$. These might be linear logistic regression models, for example.

In the standard statistical learning setting, we get an i.i.d. sequence of training examples $\mathcal{S} = ((x_1, y_1), \ldots, (x_N, y_N))$, sampled from an unknown joint distribution $\mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$. Given a measure of performance (a.k.a. *risk*) $\mathcal{E}(\cdot, \cdot)$, the goal is to find the predictor $h_\theta \in \mathcal{H}$ that minimizes the expected risk $\mathbf{E}[\mathcal{E}(h_\theta(x), y)]$. Typically $\mathcal{E}(\cdot, \cdot)$ is difficult to optimize directly and the joint distribution $\mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$ is unknown, so learning is performed via *empirical risk minimization*. Specifically, we solve

$$\min_{h_\theta \in \mathcal{H}} \left\{ \hat{\mathbf{E}}_{\mathcal{S}}[\ell(h_\theta(x), y) = \frac{1}{N} \sum_{i=1}^{N} \ell(h_\theta(x_i), y_i) \right\} \tag{2.1}$$

with loss function $\ell(\cdot, \cdot)$ acting as a surrogate of $\mathcal{E}(\cdot, \cdot)$, and the empirical mean replacing the expectation.

2.3.2 Exact Wasserstein loss

Information divergence-based loss functions are widely used in learning with probability-valued outputs. Along with other popular measures like Hellinger distance and χ^2 distance, these divergences treat the output dimensions independently, ignoring any metric structure on $\mathcal{K}$.

When the ground truth y and the output of h are both *probability* measures, lying in the simplex $\Delta^{\mathcal{K}}$, we can straightforwardly define a Wasserstein loss, which incorporates a ground metric in the output space and measures the Wasserstein distance between the predictions and ground truth.

Definition 2.3.1 (Exact Wasserstein Loss)**.** For any $h_\theta \in \mathcal{H}$, $h_\theta : \mathcal{X} \to \Delta^{\mathcal{K}}$, let $h_\theta(\kappa|x) = h_\theta(x)_\kappa$ be the predicted value at element $\kappa \in \mathcal{K}$, given input $x \in \mathcal{X}$. Let $y(\kappa)$ be the ground truth value for κ given by the corresponding label y. Then we define the *exact Wasserstein loss* as

$$W_p^p(h(\cdot|x), y(\cdot)) = \inf_{T \in \Pi(h(x),y)} \langle T, M \rangle \tag{2.2}$$

where $M \in \mathbb{R}_+^{K \times K}$ is the distance matrix $M_{\kappa,\kappa'} = d_{\mathcal{K}}^p(\kappa, \kappa')$, and the set of valid transport plans is

$$\Pi(h(x), y) = \{T \in \mathbb{R}_+^{K \times K} : T\mathbf{1} = h(x),\ T^\top \mathbf{1} = y\} \tag{2.3}$$

where $\mathbf{1}$ is the all-one vector.

2.4 Efficient learning

To do learning, we optimize the empirical risk minimization functional (2.1) by gradient descent (Algorithm 1). Doing so requires evaluating a descent direction for the loss, with respect to the predictions $h(x)$. Unfortunately, computing a subgradient of the exact Wasserstein loss (2.2), is quite costly, as follows.

Algorithm 1 Learning by stochastic gradient descent

Given training set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{R}^p \times \Delta^{\mathcal{K}}$, initial parameters $\theta^{(0)}$, number of epochs T, minibatch size N_b, step sizes $\{\eta_t\}_{t=1}^{T}$.
for $t = 1, \ldots, T$ **do**
  Sample uniformly a minibatch $\mathcal{S}^t = \{(x_j, y_j)\}_{j=1}^{N_b} \subset \mathcal{S}$.
  Compute subgradients of the loss $\frac{\partial \mathcal{W}_p^p(h(x_j), y_j)}{\partial h(x_j)}$, $j = 1, \ldots, N_b$.
  Compute subgradient of empirical risk with respect to parameters $\frac{\partial \hat{\mathbf{E}}_{\mathcal{S}^{\sqcup}}}{\partial \theta} = \frac{1}{N_b} \sum_{j=1}^{N_b} \left(\frac{\partial \mathcal{W}_p^p(h(x_j), y_j)}{\partial h(x_j)} \right) \cdot \left(\frac{\partial h(x_j)}{\partial \theta} \right)$.
  Update parameters $\theta^{(t)} = \theta^{(t-1)} - \eta_t \frac{\partial \hat{\mathbf{E}}_{\mathcal{S}^t}}{\partial \theta}$.
end for

2.4.1 Subgradients of the exact Wasserstein loss

The exact Wasserstein loss (2.2) is a linear program and a subgradient of its solution can be computed using Lagrange duality. The dual LP of (2.2) is

$$ {}^{d}W_p^p(h(x), y) = \sup_{\alpha, \beta \in C_M} \alpha^\top h(x) + \beta^\top y, \quad C_M = \{(\alpha, \beta) \in \mathbb{R}^{K \times K} : \alpha_\kappa + \beta_{\kappa'} \leq M_{\kappa,\kappa'}\}. \tag{2.4} $$

As (2.2) is a linear program, at an optimum the values of the dual and the primal are equal (see, e.g. [21]), hence the dual optimal α is a subgradient of the loss with respect to its first argument.

Computing α is costly, as it entails solving a linear program with $O(K^2)$ contraints, with K being the dimension of the output space. This cost can be prohibitive when optimizing by gradient descent.

2.4.2 Entropic regularization

Cuturi [45] proposes a smoothed transport objective that enables efficient approximation of both the transport matrix in (2.2) and the subgradient of the loss. [45] introduces an entropic regularization term that results in a strictly convex problem:

$$ {}^{\lambda}W_p^p(h(\cdot|x), y(\cdot)) = \inf_{T \in \Pi(h(x), y)} \langle T, M \rangle - \frac{1}{\lambda} H(T), \quad H(T) = -\sum_{\kappa,\kappa'} T_{\kappa,\kappa'} \log T_{\kappa,\kappa'}. \tag{2.5} $$

Algorithm 2 Gradient of the smoothed Wasserstein loss

Given $h(x), y \in \Delta^{\mathcal{K}}$, $\lambda > 0$, $\mathbf{K} = \exp(-\lambda M - 1)$.
$u \leftarrow \mathbf{1}$
while u has not converged **do**
  $u \leftarrow h(x) \oslash (\mathbf{K}(y \oslash \mathbf{K}^\top u))$
end while
$\partial W_p^p / \partial h(x) \leftarrow \frac{\log u}{\lambda} - \frac{\log u^\top \mathbf{1}}{\lambda \mathcal{K}} \mathbf{1}$

Importantly, the transport matrix that solves (2.5) is a *diagonal scaling* of a matrix $\mathbf{K} = e^{-\lambda M - 1}$:

$$T^* = \mathsf{diag}(u)\mathbf{K}\mathsf{diag}(v) \tag{2.6}$$

for $u = e^{\lambda \alpha}$ and $v = e^{\lambda \beta}$, where α and β are the Lagrange dual variables for (2.5).

Identifying such a matrix subject to equality constraints on the row and column sums is exactly a *matrix balancing* problem, which is well-studied in numerical linear algebra and for which efficient iterative algorithms exist [91]. [45] and [47] use the well-known Sinkhorn-Knopp algorithm.

2.4.3 Learning with the smoothed loss

When the output vectors $h(x)$ and y lie in the simplex, (2.5) can be used directly in place of (2.2), as (2.5) can approximate the exact Wasserstein distance closely for large enough λ [45]. In this case, the gradient α of the objective can be obtained from the optimal scaling vector u as $\alpha = \frac{\log u}{\lambda} - \frac{\log u^\top \mathbf{1}}{\lambda K} \mathbf{1}$. [1] A Sinkhorn iteration for the gradient is given in Algorithm 2.

2.5 Relaxed transport

For many learning problems, a normalized output assumption is unnatural. In image segmentation, for example, the target shape is not naturally represented as a histogram. And even when the prediction and the ground truth are constrained to the

[1] Note that α is only defined up to a constant shift: any upscaling of the vector u can be paired with a corresponding downscaling of the vector v (and vice versa) without altering the matrix T^*. The choice $\alpha = \frac{\log u}{\lambda} - \frac{\log u^\top \mathbf{1}}{\lambda K} \mathbf{1}$ ensures that α is tangent to the simplex.

simplex, the observed label can be subject to noise that violates the constraint.

There is more than one way to generalize optimal transport to unnormalized measures, and this is a subject of active study [39]. We will develop here a novel objective that deals effectively with the difference in total mass between $h(x)$ and y while still being efficient to optimize.

We propose a novel relaxation that extends the Wasserstein loss to unnormalized measures. By replacing the equality constraints on the transport marginals in (2.2) with soft penalties with respect to KL divergence, we get an approximate transport problem.

Definition 2.5.1 (Relaxed Wasserstein Loss)**.** For any $h_\theta \in \mathcal{H}$, $h_\theta : \mathcal{X} \to \mathbb{R}_+^{\mathcal{K}}$, let $h_\theta(\kappa|x) = h_\theta(x)_\kappa$ be the predicted value at element $\kappa \in \mathcal{K}$, given input $x \in \mathcal{X}$. Let $y(\kappa)$ be the ground truth value for κ given by the corresponding label y. Then we define the *relaxed Wasserstein loss* as

$$^{\gamma_a,\gamma_b}\mathcal{W}_{\mathrm{KL}}(h(\cdot|x), y(\cdot)) = \inf_{T \in \mathbb{R}_+^{\mathcal{K}\times\mathcal{K}}} \langle T, M\rangle + \gamma_a \,\mathrm{KL}(T\mathbf{1}\|h(\cdot|x)) + \gamma_b \,\mathrm{KL}(T^\intercal \mathbf{1}\|y) \quad (2.7)$$

where $M \in \mathbb{R}_+^{K\times K}$ is the distance matrix $M_{\kappa,\kappa'} = d_{\mathcal{K}}^p(\kappa,\kappa')$, and $\mathrm{KL}(w\|z) = w^\top \log(w \oslash z) - \mathbf{1}^\top w + \mathbf{1}^\top z$ is the *generalized KL divergence* between $w, z \in \mathbb{R}_+^K$. Here $\oslash$ represents element-wise division and $\mathbf{1}$ is the all-one vector.

As with the exact Wasserstein loss, computing subgradients is computationally costly. We again approach this by adding an entropic regularizer, obtaining an unconstrained objective.

$$^{\lambda,\gamma_a,\gamma_b}\mathcal{W}_{\mathrm{KL}}(h(\cdot|x), y(\cdot)) = \min_{T \in \mathbb{R}_+^{K\times K}} \langle T, M\rangle - \frac{1}{\lambda} H(T) + \gamma_a \,\mathrm{KL}(T\mathbf{1}\|h(x)) + \gamma_b \,\mathrm{KL}(T^\top \mathbf{1}\|y). \quad (2.8)$$

As with the previous formulation, the optimal transport matrix with respect to (2.8) is a diagonal scaling of the matrix $\mathbf{K}$.

Proposition 1. *The transport matrix T^* optimizing* (2.8) *satisfies $T^* = \mathsf{diag}(u)\mathbf{K}\mathsf{diag}(v)$, where $u = (h(x) \oslash T^*\mathbf{1})^{\gamma_a\lambda}$, $v = \left(y \oslash (T^*)^\top \mathbf{1}\right)^{\gamma_b\lambda}$, and $\mathbf{K} = e^{-\lambda M - 1}$.*

Proof. The first order condition for T^* optimizing (2.8) is

$$\begin{aligned}
&M_{ij} + \frac{1}{\lambda}\left(\log T^*_{ij} + 1\right) + \gamma_a \left(\log T^* \mathbf{1} \oslash h(x)\right)_i + \gamma_b \left(\log (T^*)^\top \mathbf{1} \oslash y\right)_j = 0. \\
\Rightarrow &\log T^*_{ij} + \gamma_a \lambda \log \left(T^* \mathbf{1} \oslash h(x_i)\right)_i + \gamma_b \lambda \log \left((T^*)^\top \mathbf{1} \oslash y_j\right)_j = -\lambda M_{ij} - 1 \\
\Rightarrow &T^*_{ij} \left(T^* \mathbf{1} \oslash h(x)\right)_i^{\gamma_a \lambda} \left((T^*)^\top \mathbf{1} \oslash y\right)_j^{\gamma_b \lambda} = \exp\left(-\lambda M_{ij} - 1\right) \\
\Rightarrow &T^*_{ij} = \left(h(x) \oslash T^* \mathbf{1}\right)_i^{\gamma_a \lambda} \left(y \oslash (T^*)^\top \mathbf{1}\right)_j^{\gamma_b \lambda} \exp\left(-\lambda M_{ij} - 1\right)
\end{aligned}$$

Hence T^* (if it exists) is a diagonal scaling of $\mathbf{K} = \exp\left(-\lambda M - 1\right)$. □

And the optimal transport matrix is a fixed point for a Sinkhorn-like iteration. [2]

Proposition 2. *$T^* = \mathsf{diag}(u)\mathbf{K}\mathsf{diag}(v)$ optimizing (2.8) satisfies: i) $u = h(x)^{\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}} \odot (\mathbf{K}v)^{-\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}}$, and ii) $v = y^{\frac{\gamma_b \lambda}{\gamma_b \lambda + 1}} \odot \left(\mathbf{K}^\top u\right)^{-\frac{\gamma_b \lambda}{\gamma_b \lambda + 1}}$, where $\odot$ represents element-wise multiplication.*

Proof. Let $u = \left(h(x) \oslash T^* \mathbf{1}\right)^{\gamma_a \lambda}$ and $v = \left(y \oslash (T^*)^\top \mathbf{1}\right)^{\gamma_b \lambda}$, so $T^* = \mathsf{diag}(u)\mathbf{K}\mathsf{diag}(v)$. We have

$$\begin{aligned}
T^* \mathbf{1} &= \mathsf{diag}(u)\mathbf{K}\mathsf{diag}(v) \\
\Rightarrow (T^* \mathbf{1})^{\gamma_a \lambda + 1} &= h(x)^{\gamma_a \lambda} \odot \mathbf{K}v
\end{aligned}$$

where we substituted the expression for u. Re-writing $T^* \mathbf{1}$,

$$\begin{aligned}
(\mathsf{diag}(u)\mathbf{K}v)^{\gamma_a \lambda + 1} &= \mathsf{diag}(h(x)^{\gamma_a \lambda})\mathbf{K}v \\
\Rightarrow u^{\gamma_a \lambda + 1} &= h(x)^{\gamma_a \lambda} \odot (\mathbf{K}v)^{-\gamma_a \lambda} \\
\Rightarrow u &= h(x)^{\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}} \odot (\mathbf{K}v)^{-\frac{\gamma_a \lambda}{\gamma_a \lambda + 1}}.
\end{aligned}$$

[2] Note that, although the iteration suggested by Proposition 2 is observed empirically to converge (see Figure 2-3c, for example), we have not proven a guarantee that it will do so.

Algorithm 3 Gradient of the relaxed Wasserstein loss

Given $h(x), y \in \mathbb{R}_+^K$, $\lambda, \gamma_a, \gamma_b > 0$, $\mathbf{K} = \exp(-\lambda M - 1)$.
$u \leftarrow \mathbf{1}$
while u has not converged **do**
$$u \leftarrow h(x)^{\frac{\gamma_a\lambda}{\gamma_a\lambda+1}} \oslash \left(\mathbf{K}\left(y \oslash \mathbf{K}^\top u\right)^{\frac{\gamma_b\lambda}{\gamma_b\lambda+1}}\right)^{\frac{\gamma_a\lambda}{\gamma_a\lambda+1}}$$
end while
$v \leftarrow y^{\frac{\gamma_b\lambda}{\gamma_b\lambda+1}} \oslash \left(\mathbf{K}^\top u\right)^{\frac{\gamma_b\lambda}{\gamma_b\lambda+1}}$
$\partial \mathcal{W}_{\mathrm{KL}} / \partial h(x) \leftarrow \gamma_a \left(\mathbf{1} - (\mathsf{diag}(u)\mathbf{K}\mathsf{diag}(v)) \oslash h(x)\right)$

A symmetric argument shows that $v = y^{\frac{\gamma_b\lambda}{\gamma_b\lambda+1}} \odot (\mathbf{K}^\top u)^{-\frac{\gamma_b\lambda}{\gamma_b\lambda+1}}$. □

Unlike the previous formulation, (2.8) is unconstrained with respect to $h(x)$. The gradient is given by $\nabla_{h(x)} \mathcal{W}_{\mathrm{KL}}(h(\cdot|x), y(\cdot)) = \gamma_a \left(\mathbf{1} - T^*\mathbf{1} \oslash h(x)\right)$. The iteration is given in Algorithm 3.

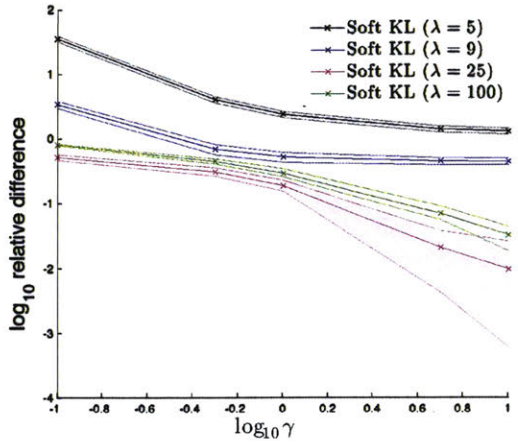
When restricted to normalized measures, the relaxed problem (2.8) approximates smoothed transport (2.5). Figure 2-3a shows, for normalized $h(x)$ and y, the relative distance between the values of (2.8) and (2.5) [3]. For λ large enough, (2.8) converges to (2.5) as γ_a and γ_b increase.

(2.8) also retains two properties of smoothed transport (2.5). Figure 2-3b shows that, for normalized outputs, the relaxed loss converges to the unregularized Wasserstein distance as λ, γ_a and γ_b increase [4]. And Figure 2-3c shows that convergence of the iterations in (2) is nearly independent of the dimension K of the output space.

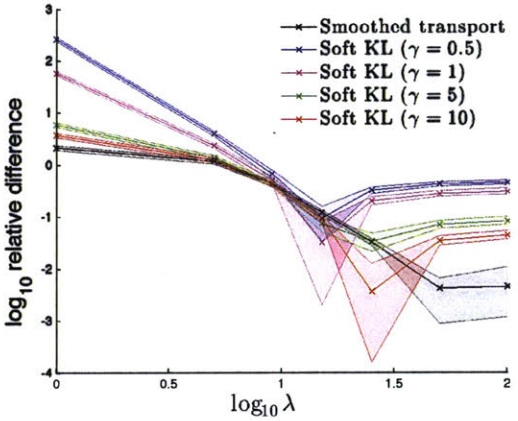
2.6 Statistical properties of the loss

In this section, we establish statistical learning bounds for the exact Wasserstein loss function (2.2) and its relaxed counterpart (2.7) [5].

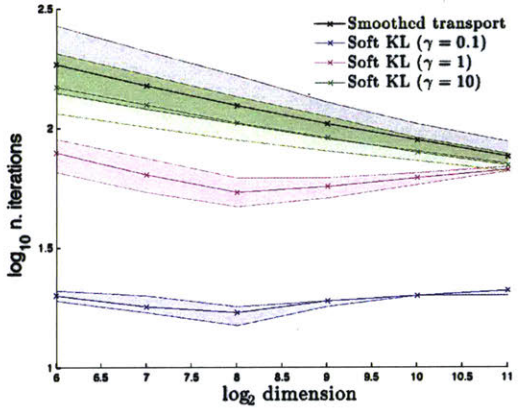
[3] In figures 2-3a-c, $h(x)$, y and M are generated as described in [45] section 5. In 2-3a-b, $h(x)$ and y have dimension 256. In 2-3c, convergence is defined as in [45]. Shaded regions are 95% intervals.
[4] The unregularized Wasserstein distance was computed using `FastEMD` [118].
[5] Note that this section extends the results in [64] to encompass the relaxed transport loss (2.7).



(a) Convergence to smoothed transport.



(b) Approximation of exact Wasserstein.



(c) Convergence of alternating projections ($\lambda = 50$).

Figure 2-3: The relaxed transport problem (2.8) for unnormalized measures.

2.6.1 Main results

Let $S = ((x_1, y_1), \ldots, (x_N, y_N))$ be i.i.d. samples and $h_{\hat{\theta}}$ be the empirical risk minimizer

$$h_{\hat{\theta}} = \operatorname*{argmin}_{h_\theta \in \mathcal{H}} \left\{ \hat{\mathbf{E}}_S \left[W_p^p(h_\theta(\cdot|x), y) \right] = \frac{1}{N} \sum_{i=1}^{N} W_p^p(h_x\theta(\cdot|x_i), y_i) \right\}.$$

Further assume $\mathcal{H} = \mathfrak{s} \circ \mathcal{H}^o$ is the composition of a softmax $\mathfrak{s}$ and a base hypothesis space $\mathcal{H}^o$ of functions mapping into $\mathbb{R}^K$. The softmax layer outputs a prediction that lies in the simplex $\Delta^{\mathcal{K}}$.

Theorem 3 (Consistency of ERM with exact Wasserstein loss). *For $p = 1$, and any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\mathbf{E}\left[W_1^1(h_{\hat{\theta}}(\cdot|x), y) \right] \leq \inf_{h_\theta \in \mathcal{H}} \mathbf{E}\left[W_1^1(h_\theta(\cdot|x), y) \right] + 32KC_M\mathfrak{R}_N(\mathcal{H}^o) + 2C_M\sqrt{\frac{\log(1/\delta)}{2N}} \tag{2.9}$$

with the constant $C_M = \max_{\kappa,\kappa'} M_{\kappa,\kappa'}$. $\mathfrak{R}_N(\mathcal{H}^o)$ is the Rademacher complexity *[10] measuring the complexity of the hypothesis space $\mathcal{H}^o$.*

The Rademacher complexity $\mathfrak{R}_N(\mathcal{H}^o)$ for commonly used models like neural networks and kernel machines [10] decays with the training set size. This theorem guarantees that the expected Wasserstein loss of the empirical risk minimizer approaches the best achievable loss for $\mathcal{H}$.

As an important special case, minimizing the empirical risk with Wasserstein loss is also good for multiclass classification. Let $y = \mathfrak{e}_\kappa$ be the "one-hot" encoded label vector for the groundtruth class.

Proposition 3. *In the multiclass classification setting, for $p = 1$ and any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\mathbf{E}_{x,\kappa}\left[d_{\mathcal{K}}(\kappa_{\hat{\theta}}(x), \kappa) \right] \leq \inf_{h_\theta \in \mathcal{H}} K\,\mathbf{E}[W_1^1(h_\theta(x), y)] + 32K^2C_M\mathfrak{R}_N(\mathcal{H}^o) + 2C_MK\sqrt{\frac{\log(1/\delta)}{2N}} \tag{2.10}$$

where the predictor is $\kappa_{\hat{\theta}}(x) = \operatorname{argmax}_\kappa h_{\hat{\theta}}(\kappa|x)$, with $h_{\hat{\theta}}$ being the empirical risk minimizer.

Note that instead of the classification error $\mathbf{E}_{x,\kappa}[\mathbb{1}_{\kappa_{\hat{\theta}}(x)\neq\kappa}]$, we actually get a bound on the expected semantic distance between the prediction and the groundtruth.

For the relaxed Wasserstein loss (2.7), we get an equivalent bound. Suppose the hypothesis space $\mathcal{H}$ is the product of component spaces $\mathcal{H}$, i.e. $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_K$.

Theorem 4 (Consistency of ERM with relaxed Wasserstein loss). *For $p = 1$, $\gamma_a, \gamma_b > 0$ and any $\delta > 0$, with probability at least $1 - \delta$, it holds that*

$$\mathbf{E}\left[{}^{\gamma_a,\gamma_b}\mathcal{W}_{\mathrm{KL}}(h_{\hat{\theta}}(\cdot|x), y)\right] \leq$$
$$\inf_{h_\theta \in \mathcal{H}} \mathbf{E}\left[{}^{\gamma_a,\gamma_b}\mathcal{W}_{\mathrm{KL}}(h_\theta(\cdot|x), y)\right] + 4\sqrt{2K}(C_M + \gamma)\sum_{k=1}^{K}\mathfrak{R}_N(\mathcal{H}_k) + 2C_M\sqrt{\frac{\log(1/\delta)}{2N}} \tag{2.11}$$

with the constant $C_M = \max_{\kappa,\kappa'} M_{\kappa,\kappa'}$, $\gamma = \max\{\gamma_a, \gamma_b\}$, $h_{\hat{\theta}}$ the empirical risk minimizer, and $\mathfrak{R}_N(\mathcal{H})$ the Rademacher complexity *[10] measuring the complexity of the hypothesis space $\mathcal{H}$.*

2.7 Empirical

2.7.1 Impact of the ground metric

In this section, we show that the Wasserstein loss encourages smoothness with respect to an artificial metric on the MNIST handwritten digit dataset. This is a multi-class classification problem with output dimensions corresponding to the 10 digits, and we apply a ground metric $d_p(\kappa, \kappa') = |\kappa - \kappa'|^p$, where $\kappa, \kappa' \in \{0, \ldots, 9\}$ and $p \in [0, \infty)$. This metric encourages the recognized digit to be *numerically* close to the true one. We train a model independently for each value of p and plot the average predicted probabilities of the different digits on the test set in Figure 2-4.

Note that as $p \to 0$, the metric approaches the $0-1$ metric $d_0(\kappa, \kappa') = \delta_{\kappa\neq\kappa'}$, which treats all incorrect digits as being equally unfavorable. In this case, as can be seen in the figure, the predicted probability of the true digit goes to 1 while the probability for all other digits goes to 0. As p increases, the predictions become more

evenly distributed over the neighboring digits, converging to a uniform distribution as $p \to \infty$ [6].

2.7.2 Tagging Flickr images

We apply the Wasserstein loss to a real world multi-label learning problem, using the recently released Yahoo/Flickr Creative Commons 100M dataset [153]. [7] Our goal is *tag prediction*: we select 1000 descriptive tags along with two random sets of 10,000 images each, associated with these tags, for training and testing. We derive a distance metric between tags by using `word2vec` [108] to embed the tags as unit vectors, then taking their Euclidean distances. To extract image features we use `MatConvNet` [155]. Note that the set of tags is highly redundant and often many semantically equivalent or similar tags can apply to an image. The images are also partially tagged, as different users may prefer different tags. We therefore measure the prediction performance by the *top-K cost*, defined as $C_K = 1/K \sum_{k=1}^{K} \min_j d_{\mathcal{K}}(\hat{\kappa}_k, \kappa_j)$, where $\{\kappa_j\}$ is the set of groundtruth tags, and $\{\hat{\kappa}_k\}$ are the tags with highest predicted probability. The standard AUC measure is also reported.

We find that a linear combination of the Wasserstein loss W_p^p and the standard multiclass logistic loss KL yields the best prediction results. Specifically, we train a linear model by minimizing $W_p^p + \alpha\mathsf{KL}$ on the training set, where α controls the relative weight of KL. Note that KL taken alone is our baseline in these experiments. Figure 2-5a shows the top-K cost on the test set for the combined loss and the baseline KL loss. We additionally create a second dataset by removing redundant labels from the original dataset: this simulates the potentially more difficult case in which a single user tags each image, by selecting one tag to apply from amongst each cluster of applicable, semantically similar tags. Figure 3b shows that performance for both algorithms decreases on the harder dataset, while the combined Wasserstein loss continues to outperform the baseline.

[6] To avoid numerical issues, we scale down the ground metric such that all of the distance values are in the interval $[0, 1)$.

[7] The dataset used here is available at `http://cbcl.mit.edu/wasserstein`.

In Figure 2-6, we show the effect on performance of varying the weight α on the KL loss. We observe that the optimum of the top-K cost is achieved when the Wasserstein loss is weighted more heavily than at the optimum of the AUC. This is consistent with a semantic smoothing effect of Wasserstein, which during training will favor mispredictions that are semantically similar to the ground truth, sometimes at the cost of lower AUC [8]. We finally show selected images from the test set in Figures 2-7 and 2-8. These illustrate cases in which both baseline and Wasserstein loss result in predictions that are semantically relevant in varying degrees, despite overlapping very little with the ground truth.

2.8 Conclusion

In this chapter we have described a loss function for learning to predict a non-negative measure over a finite set, based on the Wasserstein distance. Although optimizing with respect to the exact Wasserstein loss is computationally costly, an approximation based on entropic regularization is efficiently computed. We described a learning algorithm based on this regularization and we proposed a novel extension of the regularized loss to unnormalized measures that preserves its efficiency. We also described a statistical learning bound for the loss. The Wasserstein loss can encourage smoothness of the predictions with respect to a chosen metric on the output space, and we demonstrated this property on a real-data tag prediction problem, showing improved performance over a baseline that doesn't incorporate the metric.

An interesting direction for future work may be to explore the connection between the Wasserstein loss and Markov random fields, as the latter are often used to encourage smoothness of predictions, via inference at prediction time.

[8] The Wasserstein loss can achieve a similar trade-off by choosing the metric parameter p, as discussed in Section 2.7.1.

However, the relationship between p and the smoothing behavior is complex and it can be simpler to implement the trade-off by combining with the KL loss.

2.9 Proofs of statistical properties

We establish the proofs of Theorems 3 and 4 in this section.

2.9.1 Preliminaries

For simpler notation, for a sequence $S = ((x_1, y_1), \ldots, (x_N, y_N))$ of i.i.d. training samples, we denote the empirical risk $\hat{R}_S$ and risk R as

$$\hat{R}_S(h_\theta) = \hat{\mathbf{E}}_S \left[W_p^p(h_\theta(\cdot|x), y(\cdot)) \right], \quad R(h_\theta) = \mathbf{E} \left[W_p^p(h_\theta(\cdot|x), y(\cdot)) \right] \tag{2.12}$$

Lemma 1. *Let $h_{\hat{\theta}}, h_{\theta^*} \in \mathcal{H}$ be the minimizer of the empirical risk $\hat{R}_S$ and expected risk R, respectively. Then*

$$R(h_{\hat{\theta}}) \leq R(h_{\theta^*}) + 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)|$$

Proof. By the optimality of $h_{\hat{\theta}}$ for $\hat{R}_S$,

$$\begin{aligned} R(h_{\hat{\theta}}) - R(h_{\theta^*}) &= R(h_{\hat{\theta}}) - \hat{R}_S(h_{\hat{\theta}}) + \hat{R}_S(h_{\hat{\theta}}) - R(h_{\theta^*}) \\ &\leq R(h_{\hat{\theta}}) - \hat{R}_S(h_{\hat{\theta}}) + \hat{R}_S(h_{\theta^*}) - R(h_{\theta^*}) \\ &\leq 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| \end{aligned}$$

□

Therefore, to bound the risk for $h_{\hat{\theta}}$, we need to establish uniform concentration bounds for the Wasserstein loss. Towards that goal, we define a space of loss functions induced by the hypothesis space $\mathcal{H}$ as

$$\mathcal{L} = \left\{ \ell_\theta : (x, y) \mapsto W_p^p(h_\theta(\cdot|x), y(\cdot)) : h_\theta \in \mathcal{H} \right\} \tag{2.13}$$

The uniform concentration will depends on the "complexity" of $\mathcal{L}$, which is measured by the empirical *Rademacher complexity* defined below.

Definition 2.9.1 (Rademacher Complexity [10]). Let $\mathcal{G}$ be a family of mapping from $\mathcal{Z}$ to $\mathbb{R}$, and $S = (z_1, \ldots, z_N)$ a fixed sample from $\mathcal{Z}$. The *empirical Rademacher complexity* of $\mathcal{G}$ with respect to S is defined as

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbf{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{n} \sigma_i g(z_i) \right] \tag{2.14}$$

where $\sigma = (\sigma_1, \ldots, \sigma_N)$, with σ_i's independent uniform random variables taking values in $\{+1, -1\}$. σ_i's are called the Rademacher random variables. The *Rademacher complexity* is defined by taking expectation with respect to the samples S,

$$\mathfrak{R}_N(\mathcal{G}) = \mathbf{E}_S \left[\hat{\mathfrak{R}}_S(\mathcal{G}) \right] \tag{2.15}$$

Theorem 5. *For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $\ell_\theta \in \mathcal{L}$,*

$$\mathbf{E}[\ell_\theta] - \hat{\mathbf{E}}_S[\ell_\theta] \leq 2\mathfrak{R}_N(\mathcal{L}) + \sqrt{\frac{C_M^2 \log(1/\delta)}{2N}} \tag{2.16}$$

with the constant $C_M = \max_{\kappa,\kappa'} M_{\kappa,\kappa'}$.

By the definition of $\mathcal{L}$, $\mathbf{E}[\ell_\theta] = R(h_\theta)$ and $\hat{\mathbf{E}}_S[\ell_\theta] = \hat{R}_S[h_\theta]$. Therefore, this theorem provides a uniform control for the deviation of the empirical risk from the risk.

Theorem 6 (McDiarmid's Inequality). *Let $S = \{X_1, \ldots, X_N\} \subset \mathcal{X}$ be N i.i.d. random variables. Assume there exists $C > 0$ such that $f : \mathcal{X}^N \to \mathbb{R}$ satisfies the following stability condition*

$$|f(x_1, \ldots, x_i, \ldots, x_N) - f(x_1, \ldots, x_i', \ldots, x_N)| \leq C \tag{2.17}$$

for all $i = 1, \ldots, N$ and any $x_1, \ldots, x_N, x_i' \in \mathcal{X}$. Then for any $\varepsilon > 0$, denoting $f(X_1, \ldots, X_N)$ by $f(S)$, it holds that

$$\Pr\left(f(S) - \mathbf{E}[f(S)] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{NC^2}\right) \tag{2.18}$$

Lemma 2. *Let the constant* $C_M = \max_{\kappa,\kappa'} M_{\kappa,\kappa'}$, *then* $0 \le W_p^p(\cdot,\cdot) \le C_M$.

Proof. For any $h(\cdot|x)$ and $y(\cdot)$, let $T^* \in \Pi(h(x), y)$ be the optimal transport plan that solves (??), then

$$W_p^p(h(x), y) = \langle T^*, M \rangle \le C_M \sum_{\kappa,\kappa'} T_{\kappa,\kappa'} = C_M$$

□

Proof of Theorem 5. For any $\ell_\theta \in \mathcal{L}$, note the empirical expectation is the empirical risk of the corresponding h_θ:

$$\hat{E}_S[\ell_\theta] = \frac{1}{N}\sum_{i=1}^{N} \ell_\theta(x_i, y_i) = \frac{1}{N}\sum_{i=1}^{N} W_p^p(h_\theta(\cdot|x_i), y_i(\cdot)) = \hat{R}_S(h_\theta)$$

Similarly, $\mathbf{E}[\ell_\theta] = R(h_\theta)$. Let

$$\Phi(S) = \sup_{\ell\in\mathcal{L}} \mathbf{E}[\ell] - \hat{\mathbf{E}}_S[\ell] \tag{2.19}$$

Let S' be S with the i-th sample replaced by (x_i', y_i'), by Lemma 2, it holds that

$$\Phi(S) - \Phi(S') \le \sup_{\ell\in\mathcal{L}} \hat{\mathbf{E}}_{S'}[\ell] - \hat{\mathbf{E}}_S[\ell] = \sup_{h_\theta\in\mathcal{H}} \frac{W_p^p(h_\theta(x_i'), y_i') - W_p^p(h_\theta(x_i), y_i)}{N} \le \frac{C_M}{N}$$

Similarly, we can show $\Phi(S') - \Phi(S) \le C_M/N$, thus $|\Phi(S') - \Phi(S)| \le C_M/N$. By Theorem 6, for any $\delta > 0$, with probability at least $1-\delta$, it holds that

$$\Phi(S) \le \mathbf{E}[\Phi(S)] + \sqrt{\frac{C_M^2 \log(1/\delta)}{2N}} \tag{2.20}$$

To bound $\mathbf{E}[\Phi(S)]$, by Jensen's inequality,

$$\mathbf{E}_S[\Phi(S)] = \mathbf{E}_S\left[\sup_{\ell\in\mathcal{L}} \mathbf{E}[\ell] - \hat{\mathbf{E}}_S[\ell]\right] = \mathbf{E}_S\left[\sup_{\ell\in\mathcal{L}} \mathbf{E}_{S'}\left[\hat{\mathbf{E}}_{S'}[\ell] - \hat{\mathbf{E}}_S[\ell]\right]\right] \le \mathbf{E}_{S,S'}\left[\sup_{\ell\in\mathcal{L}} \hat{E}_{S'}[\ell] - \hat{E}_S[\ell]\right]$$

Here S' is another sequence of i.i.d. samples, usually called *ghost samples*, that is only used for analysis. Now we introduce the Rademacher variables σ_i, since the role

of S and S' are completely symmetric, it follows

$$\begin{aligned}
\mathbf{E}_S[\Phi(S)] &\leq \mathbf{E}_{S,S',\sigma}\left[\sup_{\ell\in\mathcal{L}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i(\ell(x_i', y_i') - \ell(x_i, y_i))\right] \\
&\leq \mathbf{E}_{S',\sigma}\left[\sup_{\ell\in\mathcal{L}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i \ell(x_i', y_i')\right] + \mathbf{E}_{S,\sigma}\left[\sup_{\ell\in\mathcal{L}} \frac{1}{N}\sum_{i=1}^{N} -\sigma_i \ell(x_i, y_i)\right] \\
&= \mathbf{E}_S\left[\hat{\mathfrak{R}}_S(\mathcal{L})\right] + \mathbf{E}_{S'}\left[\hat{\mathfrak{R}}_{S'}(\mathcal{L})\right] \\
&= 2\mathfrak{R}_N(\mathcal{L})
\end{aligned}$$

The conclusion follows by combining (2.19) and (2.20). □

To finish the proof of Theorem 3, we combine Lemma 1 and Theorem **??**, and relate $\mathfrak{R}_N(\mathcal{L})$ to $\mathfrak{R}_N(\mathcal{H})$ via the following generalized Talagrand's lemma [98].

Lemma 3. *Let $\mathcal{F}$ be a class of real functions, and $\mathcal{H} \subset \mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_K$ be a K-valued function class. If $\mathfrak{m} : \mathbb{R}^K \to \mathbb{R}$ is a $L_\mathfrak{m}$-Lipschitz function and $\mathfrak{m}(0) = 0$, then $\hat{\mathfrak{R}}_S(\mathfrak{m} \circ \mathcal{H}) \leq 2L_\mathfrak{m} \sum_{k=1}^{K} \hat{\mathfrak{R}}_S(\mathcal{F}_k)$.*

2.9.2 Exact Wasserstein loss

We can use the above results to obtain a bound on the Rademacher complexity of the composition of the loss with the hypothesis, in terms of the complexity of the base hypothesis class. To do so, we need to show a Lipschitz property for the loss. In the case of the exact Wasserstein loss (2.2), this is as follows.

Theorem 7 (Theorem 6.15 of [157]). *Let μ and ν be two probability measures on a Polish space $(\mathcal{K}, d_\mathcal{K})$. Let $p \in [1, \infty)$ and $\kappa_0 \in \mathcal{K}$. Then*

$$W_p(\mu, \nu) \leq 2^{1/p'} \left(\int_\mathcal{K} d_\mathcal{K}(\kappa_0, \kappa) d|\mu - \nu|(\kappa)\right)^{1/p}, \quad \frac{1}{p} + \frac{1}{p'} = 1 \tag{2.21}$$

Corollary 1. *The Wasserstein loss is Lipschitz continuous in the sense that for any*

$h_\theta \in \mathcal{H}$, *and any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$W_p^p(h_\theta(\cdot|x), y) \leq 2^{p-1} C_M \sum_{\kappa \in \mathcal{K}} |h_\theta(\kappa|x) - y(\kappa)| \tag{2.22}$$

In particular, when $p = 1$, *we have*

$$W_1^1(h_\theta(\cdot|x), y) \leq C_M \sum_{\kappa \in \mathcal{K}} |h_\theta(\kappa|x) - y(\kappa)| \tag{2.23}$$

We cannot apply Lemma 3 directly to the Wasserstein loss class, because the Wasserstein loss is only defined on probability distributions, so 0 is not a valid input. To get around this problem, we assume the hypothesis space $\mathcal{H}$ used in learning is of the form

$$\mathcal{H} = \{\mathfrak{s} \circ h^o : h^o \in \mathcal{H}^o\} \tag{2.24}$$

where $\mathcal{H}^o$ is a function class that maps into $\mathbb{R}^K$, and $\mathfrak{s}$ is the softmax function defined as $\mathfrak{s}(o) = (\mathfrak{s}_1(o), \ldots, \mathfrak{s}_K(o))$, with

$$\mathfrak{s}_k(o) = \frac{e^{o_k}}{\sum_j e^{o_j}}, \quad k = 1, \ldots, K \tag{2.25}$$

The softmax layer produce a valid probability distribution from arbitrary input, and this is consistent with commonly used models such as Logistic Regression and Neural Networks. By working with the log of the groundtruth labels, we can also add a softmax layer to the labels.

Lemma 4 (Proposition 2 of [70]). *The Wasserstein distances* $W_p(\cdot, \cdot)$ *are metrics on the space of probability distributions of* $\mathcal{K}$, *for all* $1 \leq p \leq \infty$.

Proposition 4. *The map* $\iota : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}$ *defined by* $\iota(y, y') = W_1^1(\mathfrak{s}(y), \mathfrak{s}(y'))$ *satisfies*

$$|\iota(y, y') - \iota(\bar{y}, \bar{y}')| \leq 4 C_M \|(y, y') - (\bar{y}, \bar{y}')\|_2 \tag{2.26}$$

for any $(y, y'), (\bar{y}, \bar{y}') \in \mathbb{R}^K \times \mathbb{R}^K$. *And* $\iota(0, 0) = 0$.

Proof. For any $(y, y'), (\bar{y}, \bar{y}') \in \mathbb{R}^K \times \mathbb{R}^K$, by Lemma 4, we can use triangle inequality on the Wasserstein loss,

$$|\iota(y, y') - \iota(\bar{y}, \bar{y}')| = |\iota(y, y') - \iota(\bar{y}, y') + \iota(\bar{y}, y') - \iota(\bar{y}, \bar{y}')| \leq \iota(y, \bar{y}) + \iota(y', \bar{y}')$$

Following Corollary 1, it continues as

$$|\iota(y, y') - \iota(\bar{y}, \bar{y}')| \leq C_M \left(\|\mathfrak{s}(y) - \mathfrak{s}(\bar{y})\|_1 + \|\mathfrak{s}(y') - \mathfrak{s}(\bar{y}')\|_1 \right) \tag{2.27}$$

Note for each $k = 1, \ldots, K$, the gradient $\nabla_y \mathfrak{s}_k$ satisfies

$$\|\nabla_y \mathfrak{s}_k\|_2 = \left\| \left(\frac{\partial \mathfrak{s}_k}{\partial y_j} \right)_{j=1}^K \right\|_2 = \left\| (\delta_{kj} \mathfrak{s}_k - \mathfrak{s}_k \mathfrak{s}_j)_{j=1}^K \right\|_2 = \sqrt{\mathfrak{s}_k^2 \sum_{j=1}^K \mathfrak{s}_j^2 + \mathfrak{s}_k^2 (1 - 2\mathfrak{s}_k)} \tag{2.28}$$

By mean value theorem, $\exists \alpha \in [0, 1]$, such that for $y_\theta = \alpha y + (1 - \alpha)\bar{y}$, it holds that

$$\|\mathfrak{s}(y) - \mathfrak{s}(\bar{y})\|_1 = \sum_{k=1}^K \left| \langle \nabla_y \mathfrak{s}_k |_{y = y_{\alpha_k}}, y - \bar{y} \rangle \right| \leq \sum_{k=1}^K \|\nabla_y \mathfrak{s}_k |_{y = y_{\alpha_k}}\|_2 \|y - \bar{y}\|_2 \leq 2\|y - \bar{y}\|_2$$

because by (2.28), and the fact that $\sqrt{\sum_j \mathfrak{s}_j^2} \leq \sum_j \mathfrak{s}_j = 1$ and $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, it holds

$$\begin{aligned} \sum_{k=1}^K \|\nabla_y \mathfrak{s}_k\|_2 &= \sum_{k : \mathfrak{s}_k \leq 1/2} \|\nabla_y \mathfrak{s}_k\|_2 + \sum_{k : \mathfrak{s}_k > 1/2} \|\nabla_y \mathfrak{s}_k\|_2 \\ &\leq \sum_{k : \mathfrak{s}_k \leq 1/2} \left(\mathfrak{s}_k + \mathfrak{s}_k \sqrt{1 - 2\mathfrak{s}_k} \right) + \sum_{k : \mathfrak{s}_k > 1/2} \mathfrak{s}_k \leq \sum_{k=1}^K 2\mathfrak{s}_k = 2 \end{aligned}$$

Similarly, we have $\|\mathfrak{s}(y') - \mathfrak{s}(\bar{y}')\|_1 \leq 2\|y' - \bar{y}'\|_2$, so from (2.27), we know

$$|\iota(y, y') - \iota(\bar{y}, \bar{y}')| \leq 2C_M (\|y - \bar{y}\|_2 + \|y' - \bar{y}'\|_2) \leq 2\sqrt{2} C_M \left(\|y - \bar{y}\|_2^2 + \|y' - \bar{y}'\|_2^2 \right)^{1/2}$$

then (2.26) follows immediately. The second conclusion follows trivially as $\mathfrak{s}$ maps the zero vector to a uniform distribution. □

Proof of Theorem 3. Consider the loss function space preceded with a softmax layer

$$\mathcal{L} = \{\iota_\theta : (x, y) \mapsto W_1^1(\mathfrak{s}(h_\theta^o(x)), \mathfrak{s}(y)) : h_\theta^o \in \mathcal{H}^o\}$$

We apply Lemma 3 to the $4C_M$-Lipschitz continuous function ι in Proposition 4 and the function space

$$\underbrace{\mathcal{H}^o \times \ldots \times \mathcal{H}^o}_{K \text{ copies}} \times \underbrace{\mathcal{I} \times \ldots \times \mathcal{I}}_{K \text{ copies}}$$

with $\mathcal{I}$ a singleton function space with only the identity map. It holds

$$\hat{\mathfrak{R}}_S(\mathcal{L}) \leq 8C_M \left(K\hat{\mathfrak{R}}_S(\mathcal{H}^o) + K\hat{\mathfrak{R}}_S(\mathcal{I})\right) = 8KC_M\hat{\mathfrak{R}}_S(\mathcal{H}^o) \tag{2.29}$$

because for the identity map, and a sample $S = (y_1, \ldots, y_N)$, we can calculate

$$\hat{\mathfrak{R}}_S(\mathcal{I}) = \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{I}} \frac{1}{N} \sum_{i=1}^N \sigma_i f(y_i)\right] = \mathbf{E}_\sigma \left[\frac{1}{N} \sum_{i=1}^N \sigma_i y_i\right] = 0$$

The conclusion of the theorem follows by combining (2.29) with Theorem 5 and Lemma 1. □

2.9.3 Relaxed Wasserstein loss

For the relaxed Wasserstein loss (2.7), we will again relate $\mathfrak{R}_N(\mathcal{L})$ to $\mathfrak{R}_N(\mathcal{H})$ via a variant of the Talagrand's lemma (Lemma 3). We first show a Lipschitz property for the relaxed loss.

We start by proving an analog to Theorem 8 for the relaxed Wasserstein distance.

Theorem 8. *Let μ and ν be two nonnegative measures on a Polish space $(\mathcal{X}, d)$. Let $p \in [1, \infty)$ and $x_0 \in \mathcal{X}$. Let $\gamma = \max\{\gamma_a, \gamma_b\}$. Then*

$${}^{\gamma_a,\gamma_b}\mathcal{W}_{\mathrm{KL}}^p(\mu, \nu) \leq 2^{p-1} \int_{\mathcal{X}} \left(d(x_0, x)^p + \frac{\gamma}{2^{p-1}}\right) d|\mu - \nu|(x). \tag{2.30}$$

Proof. For fixed marginals μ and ν, we define a transport plan π that attempts to distribute any unshared mass from the smaller of the two marginals uniformly across the larger of the two. Define the total quantity of unshared mass by $a = (\mu - \nu)_+(\mathcal{X})$, $b = (\mu - \nu)_-(\mathcal{X})$, and let $\beta = \max\{a, b\}$. Then the transport plan is defined

$$\pi(A \times B) = (\mu(A \cap B) - (\mu - \nu)_+(A \cap B)) + \frac{1}{\beta}(\mu - \nu)_+(A)(\mu - \nu)_-(B), \quad (2.31)$$

for any measurable sets A, B. Let $\mathrm{P}_1\,\pi$ and $\mathrm{P}_2\,\pi$ denote the marginalization operations, $\mathrm{P}_1\,\pi(A) = \pi(A \times \mathcal{X})$, $\mathrm{P}_2\,\pi(B) = \pi(\mathcal{X} \times B)$. Then we have that

$${}^{\gamma_a,\gamma_b}\mathcal{W}^p_{\mathrm{KL}}(\mu, \nu) \leq \int_{\mathcal{X}} d(x, y)^p d\pi(x, y) + \gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi, \mu) + \gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi, \nu).$$

We start by bounding the first term. Let $x_0 \in \mathcal{X}$.

$$\begin{aligned}
\int_{\mathcal{X}} d(x, y)^p d\pi(x, y) &= \frac{1}{\beta} \int_{\mathcal{X}} d(x, y)^p d(\mu - \nu)_+(x) d(\mu - \nu)_-(y) \\
&\leq \frac{2^{p-1}}{\beta} \int_{\mathcal{X}} (d(x, x_0)^p + d(x_0, y)^p) d(\mu - \nu)_+(x) d(\mu - \nu)_-(y) \\
&= 2^{p-1} \left[\frac{b}{\beta} \int_{\mathcal{X}} d(x, x_0)^p d(\mu - \nu)_+(x) + \frac{a}{\beta} \int_{\mathcal{X}} d(x_0, y)^p d(\mu - \nu)_-(y) \right] \\
&\leq 2^{p-1} \int_{\mathcal{X}} d(x, x_0)^p d\left[(\mu - \nu)_+ + (\mu - \nu)_- \right](x) \\
&= 2^{p-1} \int_{\mathcal{X}} d(x, x_0)^p d|\mu - \nu|(x).
\end{aligned}$$

The second step follows from the triangle inequality for d and the identity $(u + v)^p \leq 2^{p-1}(u^p + v^p)$.

What remains is to bound the KL terms. We have the following expressions for the two marginals of π.

$$\begin{aligned}
\mathrm{P}_1\,\pi(A) &= \mu(A) - (\mu - \nu)_+(A) + \frac{b}{\beta}(\mu - \nu)_+(A) = \mu(A) - \frac{\beta - b}{\beta}(\mu - \nu)_+(A) \\
\mathrm{P}_2\,\pi(B) &= \mu(B) - (\mu - \nu)_+(B) + \frac{a}{\beta}(\mu - \nu)_-(B).
\end{aligned}$$

There are three cases.

1. $a = b$. Then $\beta = a = b$, so $\mathrm{P}_1 \pi = \mu$ and $\mathrm{P}_2 \pi = \nu$. Both KL divergence terms are 0.

2. $a > b$. Then $\mathrm{P}_2 \pi = \nu$, while $\mathrm{P}_1 \pi = \mu - \frac{\beta - b}{\beta}(\mu - \nu)_+$. From the definition of the generalized KL divergence, we have

$$\mathrm{KL}(\mathrm{P}_1 \pi, \mu) = \int_{\mathcal{X}} \left[\frac{d \mathrm{P}_1 \pi}{d\mu} \log \frac{d \mathrm{P}_1 \pi}{d\mu} + 1 - \frac{d \mathrm{P}_1 \pi}{d\mu} \right] (x) d\mu(x).$$

We can bound the derivative $\frac{d \mathrm{P}_1 \pi}{d\mu}$,

$$\begin{aligned} \frac{d \mathrm{P}_1 \pi}{d\mu}(x) &= d \left(\frac{\mu - \frac{\beta - b}{\beta}(\mu - \nu)_+}{\mu} \right) (x) \\ &= 1 - \frac{\beta - b}{\beta} d \left(\frac{(\mu - \nu)_+}{\mu} \right) (x) \\ &\leq 1, \end{aligned}$$

as $b < \beta$ and $d(\mu - \nu)_+(x) \geq 0$, for all $x \in \mathcal{X}$ such that $d\mu(x) > 0$.

As a result, $\log \frac{d \mathrm{P}_1 \pi}{d\mu}(x) \leq 0$ and so

$$\frac{d \mathrm{P}_1 \pi}{d\mu}(x) \log \frac{d \mathrm{P}_1 \pi}{d\mu}(x) + 1 - \frac{d \mathrm{P}_1 \pi}{d\mu}(x) \leq 1 - \frac{d \mathrm{P}_1 \pi}{d\mu}(x).$$

Integrating both sides against μ,

$$\begin{aligned} \mathrm{KL}(\mathrm{P}_1 \pi, \mu) &\leq \int_{\mathcal{X}} \left(1 - \frac{d \mathrm{P}_1 \pi}{d\mu} \right) d\mu(x) \\ &= \int_{\mathcal{X}} \left| 1 - \frac{d \mathrm{P}_1 \pi}{d\mu}(x) \right| d\mu(x) \\ &= \int_{\mathcal{X}} d|\mu - \mathrm{P}_1 \pi|(x) \end{aligned}$$

with the second step following from $1 - \frac{d \mathrm{P}_1 \pi}{d\mu}(x) \geq 0$. The quantity in the last

step is

$$\int_{\mathcal{X}} d|\mu - \mathrm{P}_1\,\pi|(x) = \int_{\mathcal{X}} d\left|\frac{\beta - b}{\beta}(\mu - \nu)_+\right|(x)$$
$$\leq \int_{\mathcal{X}} d|\mu - \nu|(x),$$

as $\frac{\beta-b}{\beta} \leq 1$ and $\|(\mu-\nu)_+\|_1 \leq \|\mu - \nu\|_1$.

3. $a < b$. In this case, we have $\mathrm{P}_1\,\pi = \mu$, while $\mathrm{P}_2\,\pi = \mu - (\mu-\nu)_+ + \frac{a}{\beta}(\mu-\nu)_-$. As $\nu = \mu - (\mu-\nu)_+ + (\mu-\nu)_-$, we can rewrite this $\mathrm{P}_2\,\pi = \nu - \frac{\beta-a}{\beta}(\mu-\nu)_-$. The rest follows by analogy to the case that $a > b$, obtaining

$$\mathrm{KL}(\mathrm{P}_2\,\pi, \nu) \leq \int_{\mathcal{X}} d|\nu - \mathrm{P}_2\,\pi|(x) \leq \int_{\mathcal{X}} d|\mu - \nu|(x).$$

Combining the two bounds, we get the theorem. □

Corollary 2. *The relaxed Wasserstein loss* (2.7) *is bounded by the total variation distance: for any* $h_\theta \in \mathcal{H}$ *and any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$^{\gamma_a,\gamma_b}\mathcal{W}^p_{\mathrm{KL}}(h_\theta(\cdot|x), y) \leq (2^{p-1}C_M + \gamma)\sum_{\kappa\in\mathcal{K}} |h_\theta(\kappa|x) - y(\kappa)|, \tag{2.32}$$

with $\gamma = \max\gamma_a, \gamma_b$.

In what follows, we will use a triangle inequality for the relaxed Wasserstein distance.

Theorem 9. *Let* μ, μ', ν *be nonnegative measures on a Polish space* $(\mathcal{X}, d)$. *Let* $p \in [1, +\infty)$. *Then*

$$^{\gamma_a,\gamma_b}\mathcal{W}^p_{\mathrm{KL}}(\mu, \mu') \leq 2^{p-1}\left(^{\gamma'_a,\gamma'_b}\mathcal{W}^p_{\mathrm{KL}}(\mu, \nu) + ^{\gamma'_a,\gamma'_b}\mathcal{W}^p_{\mathrm{KL}}(\nu, \mu')\right), \tag{2.33}$$

with $\gamma'_a = \gamma_a/2^{p-1}$, $\gamma'_b = \gamma_b/2^{p-1}$.

Proof. Let π_{12}, π_{23} be nonnegative measures on $\mathcal{X} \times \mathcal{X}$, realizing the infimum in the definition of the relaxed distance, for the pairs (μ, ν) and (ν, μ'), respectively. We

have

$$
\begin{aligned}
{}^{\gamma_a,\gamma_b}\mathcal{W}^p_{\mathrm{KL}}(\mu,\nu) &= \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_{12}(x,y) + \gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi_{12},\mu) + \gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi_{12},\nu),\\
{}^{\gamma_a,\gamma_b}\mathcal{W}^p_{\mathrm{KL}}(\nu,\mu') &= \int_{\mathcal{X}\times\mathcal{X}} d(y,z)^p d\pi_{23}(y,z) + \gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi_{23},\nu) + \gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi_{23},\mu').
\end{aligned}
$$

Let π be a nonnegative measure on $\mathcal{X}\times\mathcal{X}\times\mathcal{X}$ having bivariate marginals π_{12} and π_{23}. Existence of π is discussed in [140] Thm. 5. Let π_{13} be the remaining bivariate marginal of π. Then we have

$$
{}^{\gamma_a,\gamma_b}\mathcal{W}^p_{\mathrm{KL}}(\mu,\mu') \le \int_{\mathcal{X}\times\mathcal{X}} d(x,z)^p d\pi_{13}(x,z) + \gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi_{13},\mu) + \gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi_{13},\mu').
$$

The first term is bounded as

$$
\begin{aligned}
\int_{\mathcal{X}\times\mathcal{X}} d(x,z)^p d\pi_{13}(x,z) &= \int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} d(x,z)^p d\pi(x,y,z)\\
&\le 2^{p-1}\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} (d(x,y)^p + d(y,z)^p) d\pi(x,y,z)\\
&= 2^{p-1}\left(\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi_{12}(x,y) + \int_{\mathcal{X}\times\mathcal{X}} d(y,z)^p d\pi_{23}(y,z)\right).
\end{aligned}
$$

For the remaining terms, we note that $\mathrm{P}_1\,\pi_{13} = \mathrm{P}_1\,\pi_{12}$ and $\mathrm{P}_2\,\pi_{13} = \mathrm{P}_2\,\pi_{23}$, and so

$$
\begin{aligned}
\gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi_{13},\mu) = \gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi_{12},\mu) &\le \gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi_{12},\mu) + \gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi_{12},\nu),\\
\gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi_{13},\mu') = \gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi_{23},\mu') &\le \gamma_a \,\mathrm{KL}(\mathrm{P}_1\,\pi_{23},\nu) + \gamma_b \,\mathrm{KL}(\mathrm{P}_2\,\pi_{23},\mu').
\end{aligned}
$$

Adding the first of these to the first term of the bound above gives $2^{p-1}\,{}^{\gamma'_a,\gamma'_b}\mathcal{W}^p_{\mathrm{KL}}(\mu,\nu)$, with $\gamma'_a = \gamma_a/2^{p-1}$ and $\gamma'_b = \gamma_b/2^{p-1}$. Similarly, the second of these combined with the second term of the bound above yield $2^{p-1}\,{}^{\gamma'_a,\gamma'_b}\mathcal{W}^p_{\mathrm{KL}}(\nu,\mu')$. □

The relaxed Wasserstein loss satisfies a Lipschitz property with respect to the Euclidean metric.

Proposition 5 (Lipschitz condition for the relaxed loss). *Let $h_\theta, h_{\theta'} \in \mathcal{H}$ and $(x,y) \in$*

$\mathcal{X} \times \mathcal{Y}$. *Let* $p = 1$. *Then*

$$\left|{}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), y) - {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y)\right| \leq (C_M + \gamma)\sqrt{K}\|h_\theta(\cdot|x) - h_{\theta'}(\cdot|x)\|_2 \tag{2.34}$$

with $\gamma = \max\{\gamma_a, \gamma_b\}$

Proof. Suppose ${}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), y) \geq {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y)$. Then

$$\begin{aligned}
\left|{}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), y) - {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y)\right| &= {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), y) - {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y) \\
&\leq {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), h_{\theta'}(\cdot|x)) + {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y) - {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y) \\
&= {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), h_{\theta'}(\cdot|x)),
\end{aligned}$$

with the second step from the triangle inequality (Theorem 9). If instead ${}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y) \geq {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), y)$, we similarly get

$$\left|{}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), y) - {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y)\right| \leq {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), h_\theta(\cdot|x)).$$

In either case, then, Corollary 2 gives

$$\left|{}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(\cdot|x), y) - {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_{\theta'}(\cdot|x), y)\right| \leq (C_M + \gamma)\|h_\theta(\cdot|x) - h_{\theta'}(\cdot|x)\|_1,$$

with $\gamma = \max\{\gamma_a, \gamma_b\}$. The ℓ^1 term on the right hand side is bounded by $\sqrt{K}\|h_\theta(\cdot|x) - h_{\theta'}(\cdot|x)\|_2$. □

We use a version of the Talagrand contraction inequality (Lemma 3), due to Maurer [107].

Lemma 5 (Vector contraction inequality [107]). *Let* $\mathcal{F}$ *be a class of real functions, and* $\mathcal{H} \subset \mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_K$ *be a* K*-valued function class. If* $\mathfrak{m}_i : \mathbb{R}^K \to \mathbb{R}$ *for* $i = 1, \ldots, N$ *are* L*-Lipschitz, then*

$$\hat{\mathfrak{R}}_S(\mathfrak{m} \circ \mathcal{H}) \leq \sqrt{2}L \sum_{k=1}^{K} \hat{\mathfrak{R}}_S(\mathcal{F}_k). \tag{2.35}$$

We're then ready to show the main theorem for the relaxed loss (Theorem 4).

Proof of Theorem 4. Consider the class generated by composing the relaxed 1−Wasserstein loss function with hypotheses from $\mathcal{H}$,

$$ {}^{\gamma_a,\gamma_b}\mathcal{L} = \left\{ \ell_\theta : (x,y) \mapsto {}^{\gamma_a,\gamma_b}\mathcal{W}^1_{\mathrm{KL}}(h_\theta(x), y) | h_\theta \in \mathcal{H} \right\}_{i=1}^N . \tag{2.36} $$

For any fixed label $y \in \mathbb{R}^K_+$, by Proposition 5, the loss function ${}^{\gamma_a,\gamma_b}\mathcal{W}(\cdot, y)$ is $(C_M + \gamma)\sqrt{K}$-Lipschitz. By Lemma 5, then, the Rademacher complexity of the composed class ${}^{\gamma_a,\gamma_b}\mathcal{L}$ is bounded in terms of the complexities of the component hypothesis classes $\mathcal{H}_k$:

$$ \hat{\mathfrak{R}}_S({}^{\gamma_a,\gamma_b}\mathcal{L}) \leq \sqrt{2}(C_M + \gamma)\sqrt{K} \sum_{k=1}^K \hat{\mathfrak{R}}_S(\mathcal{H}_k). \tag{2.37} $$

From Theorem 5, then we have for any $\ell_\theta \in {}^{\gamma_a,\gamma_b}\mathcal{L}$,

$$ \mathbf{E}[\ell_\theta] - \hat{\mathbf{E}}_S[\ell_\theta] \leq 2\sqrt{2K}(C_M + \gamma) \sum_{k=1}^K \mathfrak{R}_N(\mathcal{H}_k) + \sqrt{\frac{C_M^2 \log(1/\delta)}{2N}}. \tag{2.38} $$

The bound in Theorem 4 follows by applying Lemma 1. □

2.10 Experimental details

2.10.1 Label noise

We simulate the phenomenon of label noise arising from confusion of semantically similar classes, as follows. Consider a multiclass classification problem, in which the classes correspond to the vertices on a $D \times D$ lattice on the 2D plane. The Euclidean distance in $\mathbb{R}^2$ is used to measure the semantic similarity between labels. The examples within each class are sampled from an isotropic Gaussian distribution centered at the corresponding vertex. Given a noise level $t \in [0, 1]$, we choose with probability t to flip the label for each training sample to one of the neighboring

categories [9], chosen uniformly at random. Figure 2-9 shows the training set for a 3×3 lattice with noise levels $t = 0.1$ and $t = 0.5$, respectively.

Figure 2-2 is generated as follows. We repeat 10 times for noise levels $t = 0.1, 0.2, \ldots, 0.9$ and $D = 3, 4, \ldots, 7$. For each combination, we train a multiclass linear logistic regression classifier by SGD, using either the standard KL-divergence loss[10] or the proposed Wasserstein loss[11]. Performance is measured by the mean Euclidean distance between the predicted class and the true class, on the test set. Figure 2-2 compares the performance of the two loss functions.

2.10.2 Tagging Flickr images

From the tags in the Yahoo Flickr Creative Commons dataset, we filtered out those not occurring in the WordNet[12] database, as well those whose dominant lexical category was "noun.location" or "noun.time." We also filtered out by hand nouns referring to geographical location or nationality, proper nouns, numbers, photography-specific vocabulary, and several words not generally descriptive of visual content (such as "annual" and "demo"). From the remainder, the 1000 most frequently occurring tags were used.

We list some of the 1000 selected tags here. The 50 most frequently occurring tags: *travel, square, wedding, art, flower, music, nature, party, beach, family, people, food, tree, summer, water, concert, winter, sky, snow, street, portrait, architecture, car, live, trip, friend, cat, sign, garden, mountain, bird, sport, light, museum, animal, rock, show, spring, dog, film, blue, green, road, girl, event, red, fun, building, new, cloud.* ...and the 50 least frequent tags: *arboretum, chick, sightseeing, vineyard, animalia, burlesque, key, flat, whale, swiss, giraffe, floor, peak, contemporary, scooter, society, actor, tomb, fabric, gala, coral, sleeping, lizard, performer, album, body, crew, bathroom, bed, cricket, piano, base, poetry, master, renovation, step, ghost, freight,*

[9] Connected vertices on the lattice are considered neighbors, and the Euclidean distance between neighbors is set to 1. The lattice is 4-connected.

[10] This corresponds to maximum likelihood estimation of the logistic regression model.

[11] In this special case, this corresponds to weighted maximum likelihood estimation, c.f. Section ??.

[12] `http://wordnet.princeton.edu`

champion, cartoon, jumping, crochet, gaming, shooting, animation, carving, rocket, infant, drift, hope.

The complete features and labels can also be downloaded from the project website[13]. We train a multiclass linear logistic regression model using a linear combination of the Wasserstein loss and the KL divergence-based loss,

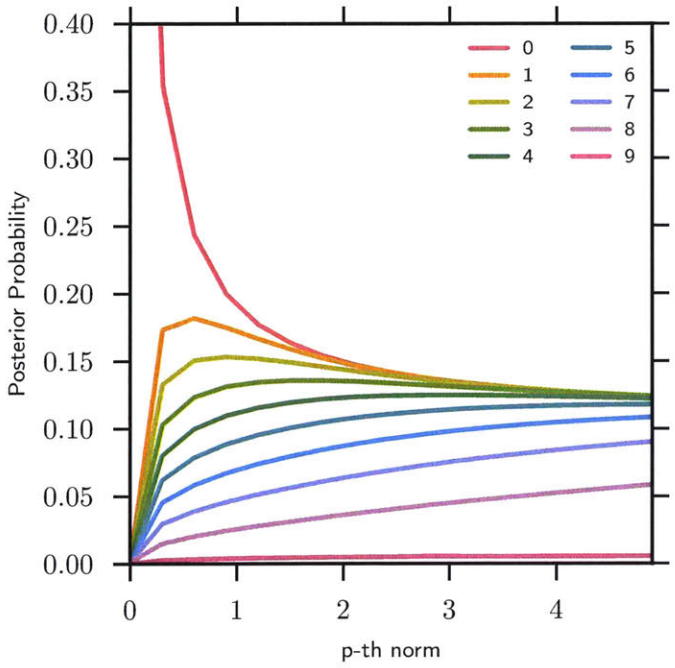
$$\mathcal{L}(h(x), y) = \mathcal{W}_p^p(h(x), y) + \alpha \, \mathrm{KL}(h(x) \| y),$$

with $\alpha > 0$ a fixed weight. The Wasserstein loss between the prediction and the normalized groundtruth is computed as described in Algorithm 2, using 10 iterations of the Sinkhorn-Knopp algorithm. Based on inspection of the ground metric matrix, we use p-norm with $p = 13$, and set $\lambda = 50$. This ensures that the matrix $\mathbf{K}$ is reasonably sparse, enforcing semantic smoothness only in each local neighborhood. We train using stochastic gradient descent (Algorithm 1) with a mini-batch size of 100, adding a momentum term with weight 0.7 to each iteration, and running for 100,000 iterations. As a baseline, we use the KL loss function alone, with identical training and test data.

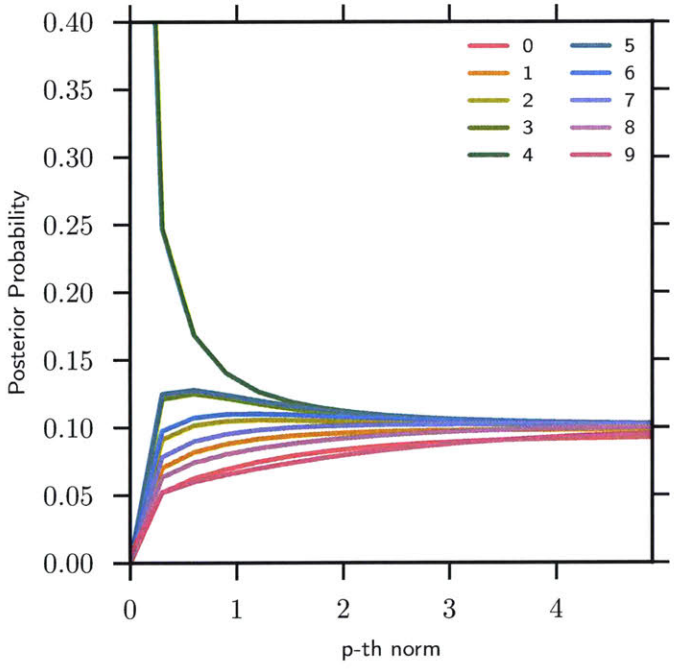
To create the dataset with reduced redundancy, for each image in the training set we compute the pairwise semantic distance for the ground truth tags, and cluster them into "equivalent" tag-sets with a threshold of semantic distance 1.3. Within each tag-set, one tag is selected randomly to represent the tag set, and the rest are discarded.

Figures 2-7 and 2-8 show test images and predictions randomly picked from the test set.

[13] `http://cbcl.mit.edu/wasserstein/`

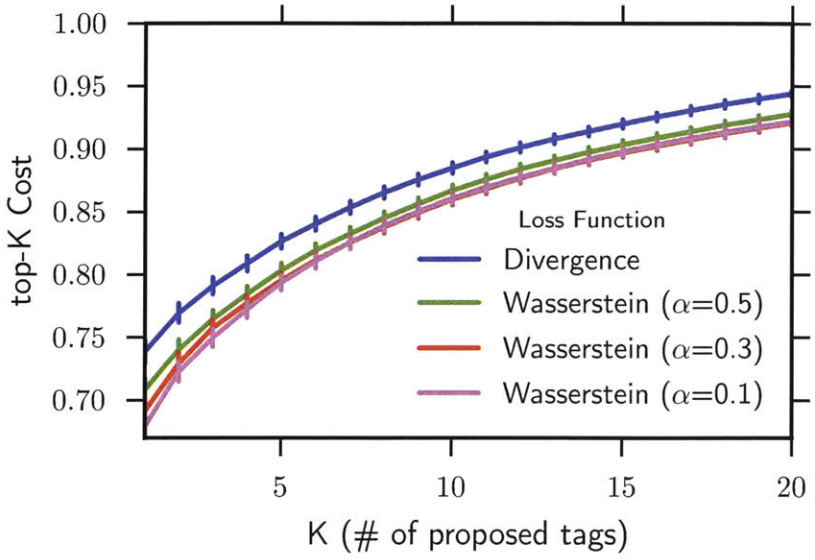


(a) Posterior predictions for images of digit 0.

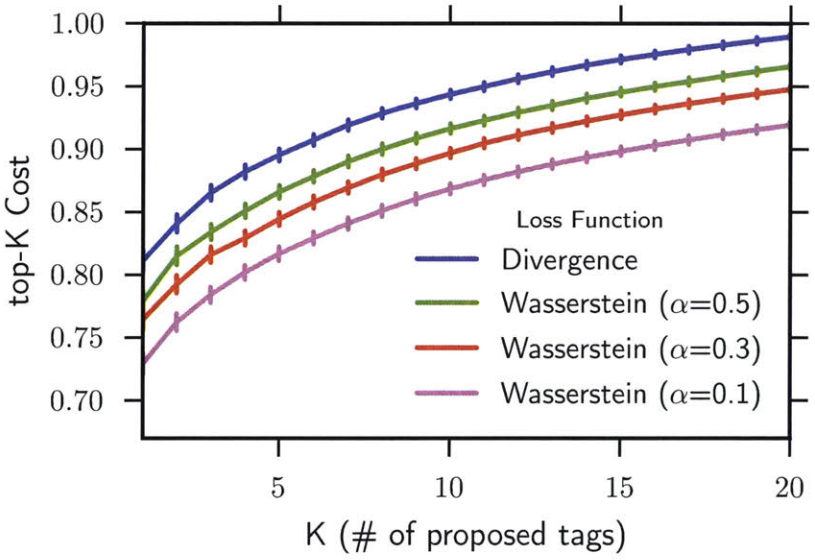


(b) Posterior predictions for images of digit 4.

Figure 2-4: MNIST example. Each curve shows the predicted probability for one digit, for models trained with different p values for the ground metric.

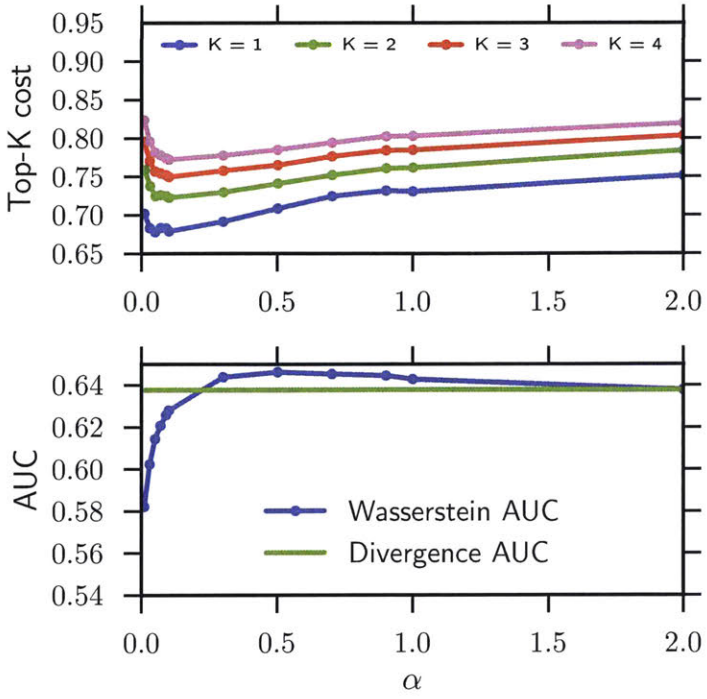


(a) Original Flickr tags dataset.

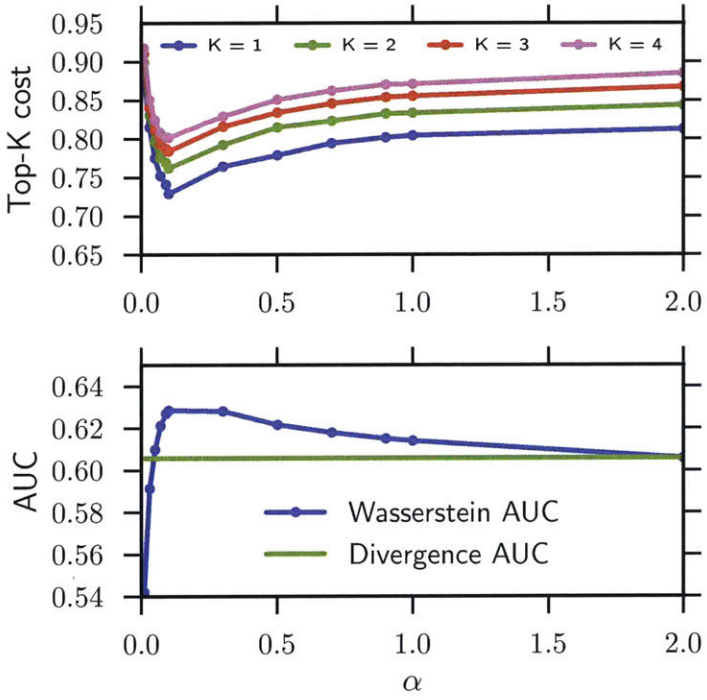


(b) Reduced-redundancy Flickr tags dataset.

Figure 2-5: Top-K cost comparison of the proposed loss (Wasserstein) and the baseline (Divergence).



(a) Original Flickr tags dataset.



(b) Reduced-redundancy Flickr tags dataset.

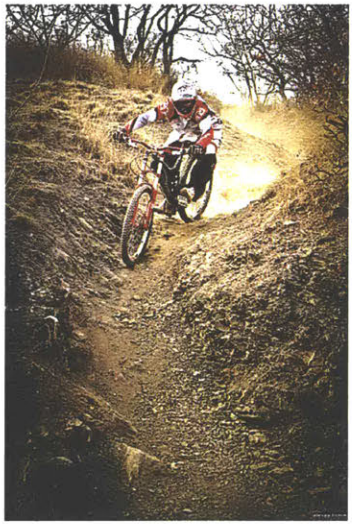
Figure 2-6: Trade-off between semantic smoothness and maximum likelihood.

(a) **Flickr user tags**: street, parade, dragon; **our proposals**: people, protest, parade; **baseline proposals**: music, car, band.

(b) **Flickr user tags**: water, boat, reflection, sunshine; **our proposals**: water, river, lake, summer; **baseline proposals**: river, water, club, nature.

(c) **Flickr user tags**: zoo, run, mark; **our proposals**: running, summer, fun; **baseline proposals**: running, country, lake.

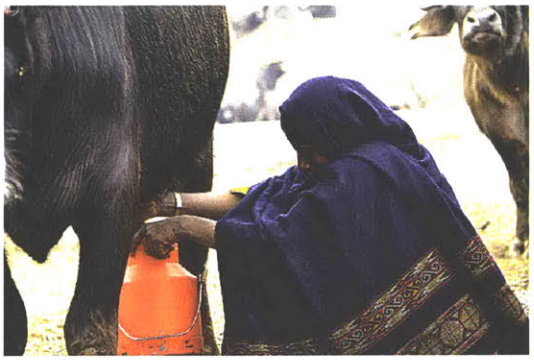
(d) **Flickr user tags**: travel, architecture, tourism; **our proposals**: sky, roof, building; **baseline proposals**: art, sky, beach.



(e) **Flickr user tags**: spring, race, training; **our proposals**: road, bike, trail; **baseline proposals**: dog, surf, bike.

Figure 2-7: Examples of images in the Flickr dataset. We show the groundtruth tags and as well as tags proposed by our algorithm and the baseline.

(a) **Flickr user tags**: family, trip, house; **our proposals**: family, girl, green; **baseline proposals**: woman, tree, family.

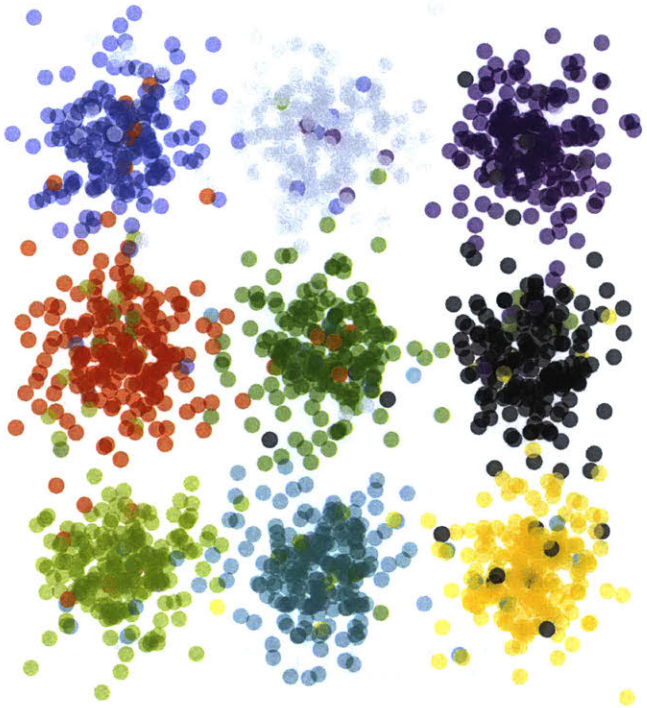


(b) **Flickr user tags**: education, weather, cow, agriculture; **our proposals**: girl, people, animal, play; **baseline proposals**: concert, statue, pretty, girl.

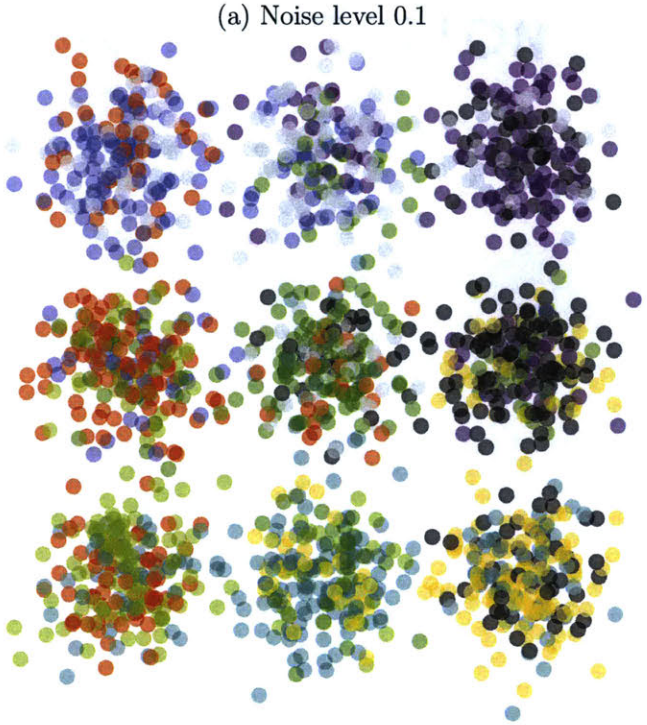
(c) **Flickr user tags**: garden, table, gardening; **our proposals**: garden, spring, plant; **baseline proposals**: garden, decoration, plant.

(d) **Flickr user tags**: nature, bird, rescue; **our proposals**: bird, nature, wildlife; **baseline proposals**: ature, bird, baby.

Figure 2-8: More examples of images in the Flickr dataset. We show the groundtruth tags and as well as tags proposed by our algorithm and baseline.



(a) Noise level 0.1



(b) Noise level 0.5

Figure 2-9: Illustration of training samples on a 3x3 lattice with different noise levels.

Chapter 3

Approximate inference with Wasserstein gradient flows

3.1 Introduction

Diffusion processes are ubiquitous in science and engineering. They arise when modeling dynamical systems driven by random fluctuations, such as interest rates and asset prices in finance, reaction dynamics in chemistry, population dynamics in ecology, and in numerous other settings. In signal processing and machine learning, diffusion processes provide the dynamics underlying classic filtering and smoothing methods such as the Kalman filter.

Inference for general diffusions is an outstanding challenge. Exact, closed-form solutions for the time-dependent probability density of the diffusion are typically unavailable, and numerous approximations have been proposed, including parametric approximations, particle or sequential Monte Carlo methods [43] [57], MCMC methods [122] [71] and variational approximations [8]. Each poses a different tradeoff between fidelity of the approximation and computational burden.

In this paper, we investigate an approximate inference method for nonlinear diffusions. It is based on a characterization, due to Jordan, Kinderlehrer and Otto [81], of the diffusion process as following a *gradient flow with respect to a Wasserstein metric* on probability densities. Concretely, they define a time discretization of the diffusion

process in which the approximate probability density ρ_k at the kth timestep solves a variational problem,

$$\rho_k = \underset{\rho \in \mathcal{P}(\mathcal{X})}{\operatorname{argmin}} \mathcal{W}(\rho, \rho_{k-1}) + \tau f(\rho) \tag{3.1}$$

with $\mathcal{W} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ being the Wasserstein distance, $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ a free energy functional defining the diffusion process, and $\tau > 0$ the size of the timestep [1]. This discrete process is shown to converge, as $\tau \to 0$, to the exact diffusion process.

For reasonable values of the timestep τ, the time-discretized Wasserstein gradient flow in (3.1) gives a close approximation to the density of the diffusion. In Figure 3-1, we apply an approximation of the Wasserstein gradient flow to a simple diffusion, initialized with a bimodal density. We see that it follows the exact density closely.

Existing methods for computing the Wasserstein gradient flow as given in (3.1) rely on discretization of the domain of the diffusion, which prohibits their application to diffusions in spaces with more than a few dimensions. Central to the current work is a novel method for computing the gradient flow that is *discretization-free*, operating directly on continuous densities. This method extends recent work on computing optimal transport between continuous densities [67].

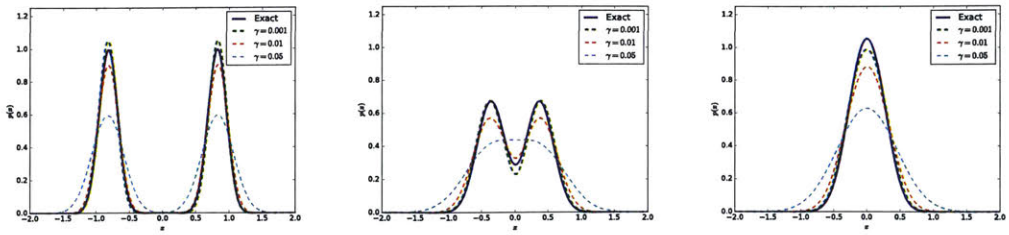
The rest of this paper is organized as follows. In Section 2 we review diffusion processes and introduce the Wasserstein gradient flow. In Section 3 we derive a smoothed dual formulation of the Wasserstein gradient flow, and in Section 4 we use this dual formulation to derive a novel inference algorithm for continuous domains. In Section 5 we investigate theoretical properties. In Section 6 we validate the proposed algorithm on a nonlinear filtering problem, before concluding.

3.2 Background and related work

3.2.1 Diffusions, free energy, and the Fokker-Planck equation

We consider a continuous-time stochastic process X_t taking values in a smooth manifold $\mathcal{X}$, for $t \in [t_i, t_f]$, and having single-time marginal densities $\rho_t : \mathcal{X} \to \mathbb{R}$ with re-

[1] $\mathcal{P}(\mathcal{X})$ is the space of probability densities defined on domain $\mathcal{X}$, having finite second moments.



(a) $t = 0.1$. (b) $t = 0.5$ (c) $t = 1$.

Figure 3-1: Regularized Wasserstein gradient flow (Section 3.3) approximates closely an Ornstein-Uhlenbeck diffusion, initialized with a bimodal density. Both the regularization (γ) and the discrete timestep (τ) are sources of error. Shaded region is the true density. ($\tau = 0.1$)

spect to a reference measure on $\mathcal{X}$. We are specifically interested in diffusion processes whose single-time marginal densities obey a diffusive partial differential equation,

$$\frac{\partial \rho_t}{\partial t} = \operatorname{div}\left[\rho_t \nabla f'(\rho_t)\right], \tag{3.2}$$

with $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ a functional on densities and f' its gradient for the $L^2(\mathcal{X})$ metric.

f is the *free energy* and defines the diffusion entirely. An important example, which will be our primary focus, is the *advection-diffusion process*, which is typically characterized as obeying an Itô stochastic differential equation,

$$dX_t = -\nabla w(X_t)dt + \beta^{-1/2} d\mathbf{W}_t \tag{3.3}$$

with ∇w being the gradient of a potential function $w : \mathcal{X} \to \mathbb{R}$, determining the advection or drift of the system, and $\beta^{-1/2} > 0$ the magnitude of the diffusion, which is driven by a Wiener process having stochastic increments $d\mathbf{W}_t$ (see [89] for a formal introduction) [2]. The advection-diffusion has marginal densities obeying a *Fokker-Planck* equation,

$$\frac{\partial \rho_t}{\partial t} = \beta^{-1} \Delta \rho_t + \operatorname{div}[\rho_t \nabla w], \tag{3.4}$$

which is a diffusive PDE with free energy functional $f(\rho) = \langle w, \rho \rangle_{L^2(\mathcal{X})} + \beta^{-1} \langle \rho, \log \rho \rangle_{L^2(\mathcal{X})}$,

[2] We assume sufficient conditions for existence of a strong solution to (3.3) are fulfilled [114] Thm. 5.2.1.

for scalar potential $w \in L^2(\mathcal{X})$. The advection-diffusion is *linear* whenever ∇w is linear in its argument.

We note that the current work applies to those diffusions that can be rendered into the form (3.2) via a change a variables. In particular, in the case of advection-diffusion, these are the *reducible* diffusions and include nearly all diffusions in one dimension [3].

3.2.2 Approximate inference for diffusions

Determining the predictive distribution for a diffusion is generally intractable. Given an initial density at time t_i, the goal is to determine the single-time marginal density ρ_t at some time $t > t_i$. Exact inference entails solving the foward PDE (3.2), for which closed-form solutions are seldom available.

Domain discretization

In certain cases, an Eulerian discretization of the domain, i.e. a fixed mesh, is available. Here one can apply standard numerical integration methods such as Chang and Cooper's [37] or entropic averaging [116] for integrating the Fokker-Planck PDE. A number of Eulerian methods have been proposed for Wasserstein gradient flows, as well, including finite element [32] and finite volume methods [34]. Entropic regularization of the problem yields an efficient iterative method [120]. Lagrangian discretizations, which follow moving particles or meshes, have also been explored [35] [161] [31] [18].

Particle simulation

One approach to inference approximates the predictive density by a weighted sum of delta functions, $\rho_t(\mathbf{x}) = \sum_{i=1}^N \mathbf{w}_i \delta_{\mathbf{x}_t^{(i)}=\mathbf{x}}$, at locations $\mathbf{x}_t^{(i)} \in \mathcal{X}$. Each delta function represents a "particle," and can be obtained by sampling an initial location $\mathbf{x}_{t_i}$ according to ρ_{t_i}, then forward simulating a trajectory from that location, according to the diffusion. Standard simulation methods such as Euler-Maruyama discretize the

time interval $[t_i, t]$ and update the particle's location recursively [89]. For a fixed time discretization, such methods are biased in the sense that, with increasing number of particles, they converge only to an approximation of the true predictive density. To address this, one can use a rejection sampling method [23] [22] to sample exactly (with no bias) from the distribution over trajectories. Density estimation can be used to extrapolate the predictive density beyond the particle locations [54] [77].

Parametric approximations

One can also approximate the predictive density by a member of a parametric class of distributions. This parametric density might be chosen by matching moments or another criterion. The extended Kalman filter [84] [95], for example, chooses a Gaussian density whose mean and covariance evolve according to a first order Taylor approximation of the dynamics. Sigma point methods such as the unscented Kalman filter [83] [82] [131] select a deterministic set of points $\mathbf{x}_t^{(i)} \in \mathcal{X}$ that evolve according to the exact dynamics of the process, such that the mean and covariance of the true predictive density is well-approximated by finite sums involving only these points. The mean and covariance so computed then define a Gaussian approximation. Gauss-Hermite [141], Gaussian quadrature and cubature methods [133] [132] correspond to different mechanisms for choosing the sigma points $\mathbf{x}_t^{(i)}$.

Beyond Gaussian approximations, mixtures of Gaussians have been used as well to approximate the predictive density [5] [151] [152]. Variational methods attempt to minimize a divergence between the chosen approximate density and the true predictive density. These can include Gaussian approximations [8] [4] as well as more general exponential families and mixtures [158] [150]. And for a broad class of diffusions, closed-form expansions in a function basis are available [3].

3.2.3 Wasserstein gradient flow

Let $f : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ be a free energy functional on densities. If we endow the space of probability densities $\mathcal{P}(\mathcal{X})$ with a Wasserstein metric then we can define a *gradient*

flow of f with respect to this Wasserstein metric, as the limit of implicit Euler steps. If $\nu \in \mathcal{P}$ is a density and $\tau > 0$ is the stepsize, we evolve ν by solving

$$\nu' = \text{prox}_{\tau f}^{\mathcal{W}} \nu = \underset{\mu \in \mathcal{P}(\mathcal{X})}{\text{argmin}}\, \mathcal{W}(\mu, \nu) + \tau f(\mu). \tag{3.5}$$

Here $\text{prox}_{\tau f}^{\mathcal{W}}$ is the proximal operator for f with respect to the Wasserstein metric.

Jordan, Kinderlehrer and Otto [81] show that in the limit $\tau \to 0$ [3], the sequence of densities obtained from (3.5) converges to the solution of the Fokker-Planck PDE (3.4), for a particular free energy functional,

$$f(\mu) = \langle w, \mu \rangle_{L^2(\mathcal{X})} + \beta^{-1} \langle \mu, \log \mu - 1 \rangle_{L^2(\mathcal{X})},$$

with $w \in L^2(\mathcal{X})$ the potential energy and $\beta > 0$ the inverse dispersion coefficient. Since then, many diffusive PDEs of the form (3.2) have been derived as Wasserstein gradient flows, including the heat equation on Riemannian manifolds [56].

3.3 Smoothed dual formulation for Wasserstein gradient flow

3.3.1 Entropy-regularized Wasserstein gradient flow

We start by introducing an entropy-regularized proximal operator for the gradient step, which uses a regularized Wasserstein distance. This is the Sinkhorn distance, introduced by Cuturi in [45], and its use for computing gradient flows has been studied by Peyré [120]. For $\mu, \nu \in \mathcal{P}(\mathcal{X})$, the regularized Wasserstein distance is

$$\mathcal{W}_\gamma(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}) - \gamma H(\pi). \tag{3.6}$$

[3] It is assumed that the cost c underlying the Wasserstein metric is the squared distance on the domain $\mathcal{X}$. [81] also assumes a growth condition on the free energy f, namely $\|\nabla w(\mathbf{x})\| \leq C(w(\mathbf{x}) + 1), \forall \mathbf{x} \in \mathcal{X}$.

with $c : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ the transport cost, $\Pi(\mu, \nu)$ the set of transport plans having marginals μ and ν, and $H(\pi) = -\langle \pi, \log \pi - 1 \rangle_{L^2(\mathcal{X}\times\mathcal{X})}$ the entropy functional. Given a free energy functional f, we define the primal objective $P_\nu^{\gamma,\tau} : \mathcal{P}(\mathcal{X}) \to [0, +\infty)$,

$$P_\nu^{\gamma,\tau}(\mu) \triangleq \mathcal{W}_\gamma(\mu, \nu) + \tau f(\mu), \tag{3.7}$$

for $\gamma \geq 0$, and $\tau > 0$. The primal formulation for the regularized Wasserstein gradient flow is

$$\operatorname{prox}_{\tau f}^{\mathcal{W}_\gamma} \nu = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} P_\nu^{\gamma,\tau}(\mu). \tag{3.8}$$

For $\gamma > 0$, the map $\mu \mapsto \mathcal{W}_\gamma(\mu, \nu)$ is strictly convex and coercive such that, assuming a convex functional f in (3.7), the proximal operator is uniquely defined.

Carlier et al. [33] show that this regularized Wasserstein gradient flow converges, in an appropriate limit of $\gamma, \tau \to 0$, to the solution for (3.2). Here the entropic regularizer functions, in part, as a barrier function for the positive octant, and ensures strict convexity of the minimization in (3.8). It also enables an unconstrained dual formulation, as shown in the following section.

Figure 3-2: Free energy expressions for advection-diffusion

$f(\mu) = \langle w, \mu \rangle_{L^2(\mathcal{X})} + \beta^{-1} \langle \mu, \log \mu - 1 \rangle_{L^2(\mathcal{X})}$
$f^*(z) = \beta^{-1} \int_{\mathcal{X}} \exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right)$
$\left(\nabla f^*(z)\right)(\mathbf{x}) = \exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right)$
$\left(\nabla^2 f^*(z)\right)(\mathbf{x}, \mathbf{y}) =$
$\begin{cases} \beta \exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right) & \mathbf{x} = \mathbf{y} \\ 0 & \text{otherwise} \end{cases}$

Note that we give all formulas in terms of a general free energy f. Table 3-2 gives concrete expressions for the free energy and its conjugate, in the case of an advection-diffusion system.

3.3.2 Smoothed dual formulation

We are interested in a dual formulation for the proximal operator (3.8). First, for dual variables $g, h \in L^2(\mathcal{X})$, we define functions $a(\mathbf{x}) = \exp\left(\frac{1}{\gamma} g(\mathbf{x})\right)$ and $b(\mathbf{y}) =$

$\exp\left(\frac{1}{\gamma}h(\mathbf{y})\right)$. Then the dual objective $D_\nu^{\gamma,\tau} : L^2(\mathcal{X}) \times L^2(\mathcal{X}) \to \mathbb{R}$ is

$$D_\nu^{\gamma,\tau}(g,h) \triangleq -\tau f^*(-\frac{1}{\tau}g) + \langle h, \nu\rangle_{L^2(\mathcal{X})} - \gamma\langle \mathcal{K}, a \otimes b\rangle_{L^2(\mathcal{X}\times\mathcal{X})}, \tag{3.9}$$

with f^* the convex conjugate [4] and $\mathcal{K}(\mathbf{x},\mathbf{y}) \triangleq \exp\left(-\frac{1}{\gamma}c(\mathbf{x},\mathbf{y})\right)$ the Gibbs kernel for the cost function c. We have the following.

Proposition 6 (Duality). *Let $\nu \in \mathcal{P}(\mathcal{X})$ and $f : \mathcal{P}(\mathcal{X}) \to [0,+\infty)$ a convex, lower semicontinuous and proper functional. Define $P_\nu^{\gamma,\tau}$ as in* (3.7) *and $D_\nu^{\gamma,\tau}$ as in* (3.9). *Assume $\gamma > 0$. Then*

$$\min_{\mu\in\mathcal{P}(\mathcal{X})} P_\nu^{\gamma,\tau}(\mu) = \max_{g\in L^2(\mathcal{X}),h\in L^2(\mathcal{X})} D_\nu^{\gamma,\tau}(g,h). \tag{3.10}$$

Suppose f is strictly convex and let g_*, h_* maximize $D_\nu^{\gamma,\tau}$. Then

$$\mu_* = \nabla f^*(-\frac{1}{\tau}g_*) \tag{3.11}$$

minimizes $P_\nu^{\gamma,\tau}$.

Proof. For $\mathcal{W}_\gamma(\cdot,\nu)$ and f both convex, lower semicontinuous and proper, Fenchel duality has that

$$\min_{\mu\in L^2(\mathcal{X})} \mathcal{W}_\gamma(\mu,\nu) + \tau f(\mu) = \max_{g\in L^2(\mathcal{X})} -\mathcal{W}_\gamma(\cdot,\nu)^*(g) - \tau f^*(-\frac{1}{\tau}g), \tag{3.12}$$

with $\mathcal{W}_\gamma(\cdot,\nu)^*$ and f^* the convex conjugates,

$$\mathcal{W}_\gamma(\cdot,\nu)^*(g) = \max_{\mu\in L^2(\mathcal{X})} \langle \mu, g\rangle_{L^2(\mathcal{X})} - \mathcal{W}_\gamma(\mu,\nu), \tag{3.13}$$

$$(\tau f)^*(-g) = \tau f^*(-\frac{1}{\tau}g) = \max_{\mu\in L^2(\mathcal{X})} -\langle \mu, g\rangle_{L^2(\mathcal{X})} - \tau f(\mu). \tag{3.14}$$

[4] $f^*(z) = \inf_\mu \langle \mu, z\rangle_{L^2(\mathcal{X})} - f(\mu)$.

The Lagrangian dual formulation [48] for $\mathcal{W}_\gamma(\mu,\nu)$ has

$$\mathcal{W}_\gamma(\mu,\nu) = \max_{\alpha,\beta\in L^2(\mathcal{X})} \langle\alpha,\mu\rangle_{L^2(\mathcal{X})} + \langle\beta,\nu\rangle_{L^2(\mathcal{X})} - \gamma\langle\Lambda^{\alpha,\beta},\mathcal{K}\rangle_{L^2(\mathcal{X}\times\mathcal{X})},$$

with $\alpha,\beta \in L^2(\mathcal{X})$ the Lagrangian dual variables for the marginal constraints, $\mathcal{K}(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{1}{\gamma}c(\mathbf{x},\mathbf{y})\right)$ the Gibbs kernel, and $\Lambda^{\alpha,\beta}$ defined by $\Lambda^{\alpha,\beta}(\mathbf{x},\mathbf{y}) = \exp\left(\frac{1}{\gamma}(\alpha(\mathbf{x})+\beta(\mathbf{y}))\right)$.

We can rewrite the conjugate $\mathcal{W}_\gamma(\cdot,\nu)^*$.

$$\begin{aligned}
\mathcal{W}_\gamma(\cdot,\nu)^*(g) &= \max_{\mu\in\mathcal{P}(\mathcal{X})} \langle g,\mu\rangle_{L^2(\mathcal{X})} - \max_{\alpha,\beta\in L^2(\mathcal{X})} \langle\alpha,\mu\rangle_{L^2(\mathcal{X})} + \langle\beta,\nu\rangle_{L^2(\mathcal{X})} - \gamma\langle\Lambda^{\alpha,\beta},\mathcal{K}\rangle_{L^2(\mathcal{X}\times\mathcal{X})} \\
&= \max_{\mu\in\mathcal{P}(\mathcal{X})} -\max_{\alpha,\beta\in L^2(\mathcal{X})} \langle\alpha',\mu\rangle_{L^2(\mathcal{X})} + \langle\beta,\nu\rangle_{L^2(\mathcal{X})} - \gamma\langle\Lambda^{\alpha'+g,\beta},\mathcal{K}\rangle_{L^2(\mathcal{X}\times\mathcal{X})} \\
&= \max_{\mu\in\mathcal{P}(\mathcal{X})} -\mathcal{W}_\gamma^{c'}(\mu,\nu)
\end{aligned}$$

where $\alpha'(\mathbf{x}) = \alpha(\mathbf{x}) - g(\mathbf{x})$ and we write $\mathcal{W}_\gamma^{c'}(\mu,\nu)$ for the regularized Wasserstein distance with respect to the cost $c'(\mathbf{x},\mathbf{y}) = c(\mathbf{x},\mathbf{y}) - g(\mathbf{x})$. The last expression reduces to an optimal transport problem constrained only in one marginal,

$$\mathcal{W}_\gamma(\cdot,\nu)^*(g) = -\min_{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X}),\int_\mathcal{X}\pi(\cdot,\mathbf{y})=\nu} \langle c',\pi\rangle_{L^2(\mathcal{X}\times\mathcal{X})} - \gamma H(\pi), \tag{3.15}$$

whose Lagrangian dual formulation is

$$\mathcal{W}_\gamma(\cdot,\nu)^*(g) = -\max_{h\in L^2(\mathcal{X})} \min_{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X})} \langle c',\pi\rangle_{L^2(\mathcal{X}\times\mathcal{X})} - \gamma H(\pi) + \langle h,\nu - \int_\mathcal{X} \pi(\mathbf{x},\cdot)d\mathbf{x}\rangle_{L^2(\mathcal{X})}, \tag{3.16}$$

with $h \in L^2(\mathcal{X})$ the Lagrangian dual variable for the marginal constraint $\pi(\mathcal{X}\times B) = \nu(B)$, for all measurable $B \subseteq \mathcal{X}$. From the first order conditions for (3.16), the optimal transport plan π_* satisfies

$$\pi_*^{g,h}(\mathbf{x},\mathbf{y}) = \exp\left(\frac{1}{\gamma}(g(\mathbf{x}) + h(\mathbf{y}) - c(\mathbf{x},\mathbf{y}))\right),$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Plugging $\pi_*^{g,h}$ back into (3.16) yields

$$\mathcal{W}_\gamma(\cdot, \nu)^*(g) = -\max_{h \in L^2(\mathcal{X})} \langle h, \nu \rangle_{L^2(\mathcal{X})} - \gamma \left\langle \Lambda^{g,h}, \mathcal{K} \right\rangle_{L^2(\mathcal{X} \times \mathcal{X})},$$

with $\Lambda^{g,h}$ defined as above. Plugging this into (3.12) yields the result (3.10).

Suppose $g_*, h_* \in L^2(\mathcal{X})$ optimize the dual objective $D_\nu^{\gamma,\tau}$. Then μ_* optimal for $P_\nu^{\gamma,\tau}$ satisfies

$$\mu_* \in \partial(\tau f)^*(-g_*).$$

When f is strictly convex, this is $\mu_* = \nabla(\tau f)^*(-g_*) = \nabla f^*(-\frac{1}{\tau} g_*)$. □

Importantly, we have replaced the linearly-constrained optimization in the primal (3.8) with an unconstrained problem (3.10). The term involving the Gibbs kernel acts as a soft constraint, encouraging $g(\mathbf{x}) + h(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y})$.

3.4 Discretization-free inference

Our target is the predictive distribution of a diffusion: given an initial density ρ_t, we want to evolve it forward by a time increment Δt, to obtain the solution for the diffusion (3.2) at time $t + \Delta t$. We propose to approximate this by one step of the Wasserstein gradient flow (3.5), with stepsize $\tau = \Delta t$.

We will address one significant problem: there is no published method for solving (3.8) on a continuous domain $\mathcal{X}$, which involves optimizing over continuous densities. We will address this by leveraging the smoothed dual formulation.

3.4.1 Representation

To represent continuous functions, we assume they lie in a compact subset $\mathcal{G}$ of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ defined on $\mathcal{X}$, having kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$. Let $g \in \mathcal{H}$. From the reproducing property of $\mathcal{H}$, we have that pointwise evaluation is a linear functional such that $g(\mathbf{x}) = \langle g, \kappa(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$, for all $\mathbf{x} \in \mathcal{X}$.

Algorithm 4 Stochastic approximation to Wasserstein gradient flow

Given: initial density ρ_t, constant $\gamma > 0$, timestep $\tau > 0$, regularizer $\lambda \geq 0$.
Choose sampling densities μ_0, ν_0 on $\mathcal{X}$.
Sample independently N pairs $(\mathbf{x}_i, \mathbf{y}_i) \sim \mu_0 \otimes \nu_0$.
Solve $g_*, h_* = \operatorname{argmax}_{g,h \in \mathcal{G}} D^{\gamma,\tau}_{\rho_t,N}(g,h) - \frac{\lambda}{2}\left(\|g\|^2_{\mathcal{H}} + \|h\|^2_{\mathcal{H}}\right)$.
The evolved density is $\rho_{t+\tau} = \nabla f^*\left(-\frac{1}{\tau} g_*\right)$.

3.4.2 Expectation maximization

Key to our approach is reformulation of the proximal operator (3.8) for the flow as maximizing an expectation over $\mathcal{X} \times \mathcal{X}$, which will allow for a Monte Carlo approximation. Specifically, we choose *reference densities* $\mu_0, \nu_0 \in \mathcal{P}(\mathcal{X})$, supported everywhere in $\mathcal{X}$, and express the dual objective (3.9) as

$$D^{\gamma,\tau}_{\nu}(g,h) = \mathbf{E}_{X,Y}\, d^{\gamma,\tau}_{\nu}(X,Y,g,h) \tag{3.17}$$

for random variables X, Y distributed as μ_0 and ν_0, respectively, where the integrand $d^{\gamma,\tau}_{\nu}$ is

$$d^{\gamma,\tau}_{\nu}(\mathbf{x},\mathbf{y},g,h) = -\tau \frac{\bar{f}^*(-\frac{1}{\tau} g(\mathbf{x}))}{\mu_0(\mathbf{x})} + h(\mathbf{y}) \frac{\nu(\mathbf{y})}{\nu_0(\mathbf{y})} - \gamma \frac{a(\mathbf{x}) b(\mathbf{y})}{\mu_0(\mathbf{x}) \nu_0(\mathbf{y})} \mathcal{K}(\mathbf{x},\mathbf{y}). \tag{3.18}$$

Here, the term $\bar{f}^*$ arises when we express the conjugate functional f^* in $D^{\gamma,\tau}_{\nu}$ in integral form,

$$f^*(z) = \int_{\mathcal{X}} \bar{f}^*(z(\mathbf{x}))\, d\mathbf{x}.$$

In the case of an advection-diffusion, for example, this is

$$\bar{f}^*(z(\mathbf{x})) = \beta^{-1} \exp\left(\beta(z(\mathbf{x}) - w(\mathbf{x}))\right)$$

for $w : \mathcal{X} \to [0, +\infty)$ the advection potential.

3.4.3 Stochastic approximation

Computing the Wasserstein gradient step for continuous density ν is equivalent to maximizing the expectation in (3.17). Unfortunately, this problem need not have a finite-dimensional solution, making its exact computation intractable. To obtain a tractable problem, we propose a Monte Carlo approximation to the expectation in (3.17). Specifically, we can sample pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{X}$ according to joint density $\mu_0 \otimes \nu_0$, and compute the empirical mean,

$$D^{\gamma,\tau}_{\nu,N}(g,h) = \frac{1}{N}\sum_{i=1}^{N} d^{\gamma,\tau}_{\nu}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, g, h), \tag{3.19}$$

in place of the expectation.

As we will show, maximizing (3.19) over $\mathcal{G} \times \mathcal{G}$ is in fact a finite-dimensional problem, as the solution is constrained to a finite-dimensional subspace of $\mathcal{H} \times \mathcal{H}$. First, however, we note that the problem as given is *unstable* for small to moderate N: the maximizer can vary significantly depending on the particular underlying sample. Figures 3-3a, 3-3d, and 3-3g show this variability for an example problem, in which we fix an initial distribution (a mixture of two Gaussians) and compute the Wasserstein gradient flow for multiple samples from $\mu_0 \otimes \nu_0$, using $N = 200$, 400, and 1000 for the stochastic approximation [5]. At $N = 200$ and 400, in particular, the solution varies significantly.

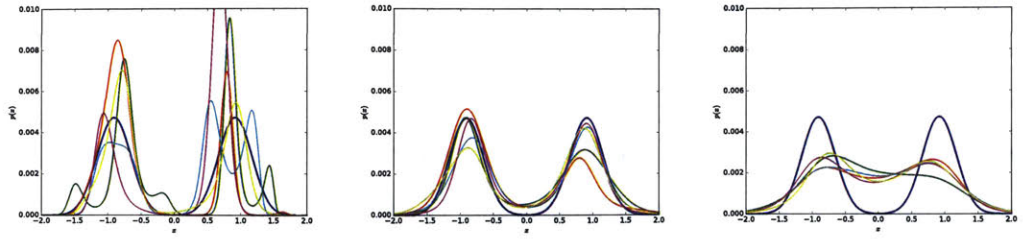
To stabilize the problem, we introduce Tikhonov regularization, penalizing the RKHS norm of the solution. The regularized problem is

$$(g_*, h_*) = \operatorname*{argmax}_{g,h\in\mathcal{G}} D^{\gamma,\tau}_{\nu,N}(g,h) - \frac{\lambda}{2}\left(\|g\|^2_{\mathcal{H}} + \|h\|^2_{\mathcal{H}}\right), \tag{3.20}$$

with parameter $\lambda > 0$. The resulting objective is λ-strongly concave.

Figure 3-3 shows the reduction in variability obtained by regularizing, for the same problem as above. The middle column shows roughly optimal choices for λ, minimizing the total ℓ^2 error with respect to the exact solution, with qualitatively

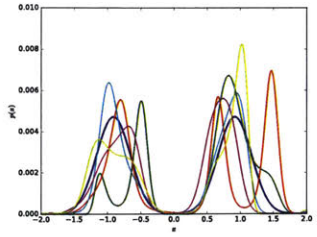
[5] In Figures 3-3 and 3-4, we sample points from μ_0 and ν_0 being standard normal distributions. We use a Gaussian kernel. The exact Wasserstein flow is computed by a Dykstra's method [120] on a discrete grid. See Section 3.8 for more details.



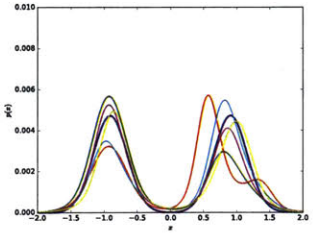
(a) Unregularized solution ($N = 200$).

(b) Regularized solution ($N = 200$, $\lambda = 0.1$).

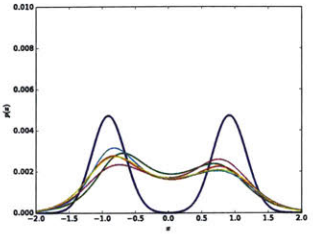
(c) Regularized solution ($N = 200$, $\lambda = 1$).



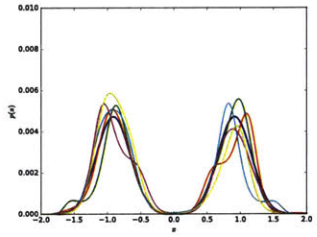
(d) Unregularized solution ($N = 400$).



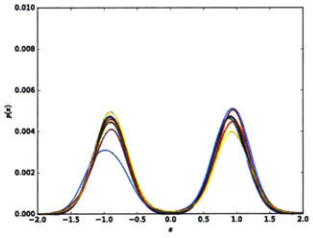
(e) Regularized solution ($N = 400$, $\lambda = 0.0316$).



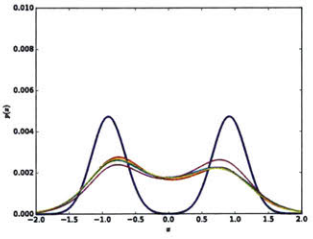
(f) Regularized solution ($N = 400$, $\lambda = 1$).



(g) Unregularized solution ($N = 1000$).



(h) Regularized solution ($N = 1000$, $\lambda = 0.01$).



(i) Regularized solution ($N = 1000$, $\lambda = 1$).

Figure 3-3: Regularization stabilizes stochastic approximation. Shaded region indicates the exact gradient flow solution.

less variability. Regularization incurs a bias, of course; Figure 3-4 shows the bias-variance tradeoff for this example problem. The bias here is the ℓ^2 error of the mean output distribution over many independent underlying samples, while the "variance" is the standard deviation of the output distributions around this mean distribution. The figure demonstrates that, for small to moderate N, we can obtain a reduction in variance while introducing minimal bias. At larger N ($N = 1000$), quality of the unregularized solution is such that regularization yields no improvement in total error.

Importantly, the optimizer for (3.20) has a finite-dimensional representation.

Proposition 7 (Representation). *Let $\nu \in \mathcal{P}(\mathcal{X})$ and $\gamma, \tau, N > 0$. Let $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N \subset \mathcal{X} \times \mathcal{X}$. Then there exist $g_*, h_* \in \mathcal{G}$ maximizing (3.20) such that*

$$(g_*, h_*) = \sum_{i=1}^{N} \left(\alpha_g^{(i)} \kappa(\mathbf{x}^{(i)}, \cdot), \alpha_h^{(i)} \kappa(\mathbf{y}^{(i)}, \cdot) \right),$$

for some sequences of scalar coefficients $\{\alpha_g^{(i)}\}_{i=1}^N$ and $\{\alpha_h^{(i)}\}_{i=1}^N$.

Proof. Let $\mathcal{H}_N \subset$ be the linear span of the functions $\kappa(\mathbf{x}^{(i)}, \cdot)$, and $\mathcal{H}_N^\perp$ its orthogonal complement. For any $g \in \mathcal{H}$, we can decompose it as $g = g^\parallel + g^\perp$, with $g^\parallel \in \mathcal{H}_N$ and $g^\perp \in \mathcal{H}_N^\perp$. Moreover, $D_{\nu,N}^{\gamma,\tau}(g, h) = D_{\nu,N}^{\gamma,\tau}(g^\parallel, h)$, as $D_{\nu,N}^{\gamma,\tau}$ depends on its first argument only via the evaluation functional at each point,

$$g(\mathbf{x}^{(i)}) = \langle \kappa(\mathbf{x}^{(i)}, \cdot), g \rangle_{\mathcal{H}} = \langle \kappa(\mathbf{x}^{(i)}, \cdot), g^\parallel \rangle_{\mathcal{H}}.$$

Hence if $D_{\nu,N}^{\gamma,\tau}$ is maximized by g_*, it is also maximized by $g_*^\parallel \in \mathcal{H}_N$. The same argument holds for h_*.

The regularization terms decompose as $\|g\|_{\mathcal{H}}^2 = \|g^\parallel\|_{\mathcal{H}}^2 + \|g^\perp\|_{\mathcal{H}}^2$ and $\|h\|_{\mathcal{H}}^2 = \|h^\parallel\|_{\mathcal{H}}^2 + \|h^\perp\|_{\mathcal{H}}^2$. Hence for any given value of the empirical term $D_{\nu,N}^{\gamma,\tau}$, the g and h attaining that value that also have maximum total objective value must lie in the subspace $\mathcal{H}_N$. This is true of g_*, h_* maximizing the total objective. □

The regularized stochastic approximation problem can be solved by a standard iterative method such as conjugate gradient. Algorithm 4 outlines the resulting inference method. And Algorithm 5 shows the computation of the objective and its gradient.

Algorithm 5 Objective and gradient computation for stochastic approximation

Given: initial density ρ_t, constant $\gamma > 0$, timestep $\tau > 0$, regularizer $\lambda \geq 0$, sampling densities μ_0, ν_0, sample points $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})_{i=1}^N$, parameters $\alpha_g, \alpha_h \in \mathbb{R}^N$.

$g(\mathbf{x}^{(i)}) \leftarrow \sum_{j=1}^N \kappa(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \alpha_g^{(j)}$, $\forall i$.

$h(\mathbf{y}^{(i)}) \leftarrow \sum_{j=1}^N \kappa(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) \alpha_h^{(j)}$, $\forall i$.

$R^{\gamma,\tau,\lambda}_{\rho_t,N}(\alpha_g, \alpha_h) \leftarrow \sum_{i=1}^N \left(-\tau \frac{\bar{f}^*(-\frac{1}{\tau} g(\mathbf{x}^{(i)}))}{\mu_0(\mathbf{x}^{(i)})} + h(\mathbf{y}^{(i)}) \frac{\rho_t(\mathbf{y}^{(i)})}{\nu_0(\mathbf{y}^{(i)})} - \gamma \frac{a(\mathbf{x}^{(i)}) b(\mathbf{y}^{(i)})}{\mu_0(\mathbf{x}^{(i)}) \nu_0(\mathbf{y}^{(i)})} \mathcal{K}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \right) -$
$\frac{\lambda}{2} \sum_{i,j=1}^N \left(\alpha_g^{(i)} \alpha_g^{(j)} \kappa(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + \alpha_h^{(i)} \alpha_h^{(j)} \kappa(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) \right)$.

$\frac{\partial}{\partial \alpha_g^{(j)}} R^{\gamma,\tau,\lambda}_{\rho_t}(\alpha_g, \alpha_h) \leftarrow \sum_{i=1}^N \left(\frac{\nabla f^*(-\frac{1}{\tau} g(\mathbf{x}^{(i)}))}{\mu_0(\mathbf{x}^{(i)})} - \frac{a(\mathbf{x}^{(i)}) b(\mathbf{y}^{(i)})}{\mu_0(\mathbf{x}^{(i)}) \nu_0(\mathbf{y}^{(i)})} \mathcal{K}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \right) \kappa(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) -$
$\lambda \sum_{i=1}^N \kappa(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \alpha_g^{(i)}$

$\frac{\partial}{\partial \alpha_h^{(j)}} R^{\gamma,\tau,\lambda}_{\rho_t}(\alpha_g, \alpha_h) \leftarrow \sum_{i=1}^N \left(\frac{\rho_t(\mathbf{y}^{(i)})}{\nu_0(\mathbf{y}^{(i)})} - \frac{a(\mathbf{x}^{(i)}) b(\mathbf{y}^{(i)})}{\mu_0(\mathbf{x}^{(i)}) \nu_0(\mathbf{y}^{(i)})} \mathcal{K}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \right) \kappa(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) -$
$\lambda \sum_{i=1}^N \kappa(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) \alpha_h^{(i)}$

3.5 Properties

3.5.1 Consistency

Both the unregularized (3.19) and the regularized stochastic problem (3.20) yield consistent approximations to the entropic-regularized Wasserstein gradient step (3.8), in the sense that, as we increase the number of samples (and correspondingly decrease the regularization parameter), the solution converges to that of the original expectation maximization (3.17).

To show this, we make the following assumptions.

A1 $\mathcal{X} \times \mathcal{X}$ is compact.

A2 μ_0 and ν_0 are bounded away from zero: $\min_{\mathbf{x} \in \mathcal{X}} \mu_0(\mathbf{x}) = U_0^{\min} > 0$, $\min_{\mathbf{y} \in \mathcal{X}} \nu_0(\mathbf{y}) = V_0^{\min} > 0$.

A3 $\mathcal{G}$ is compact and convex, with $\|g\|_{\mathcal{H}} \leq H$ for all $g \in \mathcal{G}$.

A4 $\mathcal{H}$ has reproducing kernel κ that is bounded: $\max_{\mathbf{x} \in \mathcal{X}} \sqrt{\kappa(\mathbf{x}, \mathbf{x})} = K < \infty$.

A5 $\bar{f}^*$ is convex and L_{f^*}-Lipschitz.

The assumptions guarantee that the stochastic dual objective (3.19) is L-Lipschitz.

Proposition 8 (Lipschitz property for $d_\nu^{\gamma,\tau}$). *Let $d_\nu^{\gamma,\tau}$ be defined as in (??) and suppose Assumptions* **A1-A5** *hold. Let $U^{\max} = \max_{\mathbf{x}\in\mathcal{X}, g\in\mathcal{H}} \frac{\nabla f^*(-\frac{1}{\tau}g(\mathbf{x}))}{\mu_0(\mathbf{x})}$ and $V^{\max} = \max_{\mathbf{y}\in\mathcal{X}} \frac{\nu(\mathbf{y})}{\nu_0(\mathbf{y})}$. Then for all $g, g', h, h' \in \mathcal{H}$, $d_\nu^{\gamma,\tau}$ satisfies*

$$|d_\nu^{\gamma,\tau}(\mathbf{x},\mathbf{y},g,h) - d_\nu^{\gamma,\tau}(\mathbf{x},\mathbf{y},g',h')| \leq L\|(g(\mathbf{x}),h(\mathbf{y})) - (g'(\mathbf{x}),h'(\mathbf{y}))\|_1$$

with constant L defined by $L = \max\{U^{\max}, V^{\max}\} + \frac{\exp\left(\frac{2}{\gamma}KH\right)}{U_0^{\min}V_0^{\min}}$.

Proof. Note that $U^{\max}$ and $V^{\max}$ are finite by assumptions **A2** and **A5**.

By **A3-A4**, we have that $K = \min_{\mathbf{x}\in\mathcal{X}} \sqrt{\kappa(\mathbf{x},\mathbf{x})} < \infty$, and $\mathcal{G}\times\mathcal{G}$ is bounded, such that $\|g\|_\mathcal{H}, \|h\|_\mathcal{H} \leq H$. Therefore $|g(\mathbf{x})|, |h(\mathbf{y})| \leq KH$, because by the reproducing property

$$\begin{aligned}
|g(\mathbf{x})| &= |\langle \kappa(\mathbf{x},\cdot), g\rangle_\mathcal{H}| \\
&\leq \|\kappa(\mathbf{x},\cdot)\|_\mathcal{H}\|g\|_\mathcal{H} \\
&\leq K\|g\|_\mathcal{H}, \\
&\leq KH,
\end{aligned}$$

with the second step from Cauchy-Schwarz. The analogous result holds for $|h(\mathbf{y})|$.

$d_\nu^{\gamma,\tau}$ has derivatives

$$\frac{\partial d_\nu^{\gamma,\tau}}{\partial g(\mathbf{x})} = \frac{\nabla f^*(-\frac{1}{\tau}g(\mathbf{x}))}{\mu_0(\mathbf{x})} - \frac{a(\mathbf{x})b(\mathbf{y})}{\mu_0(\mathbf{x})\nu_0(\mathbf{y})}\mathcal{K}(\mathbf{x},\mathbf{y})$$

in $g(\mathbf{x})$ and

$$\frac{\partial d_\nu^{\gamma,\tau}}{\partial h(\mathbf{y})} = \frac{\nu(\mathbf{y})}{\nu_0(\mathbf{y})} - \frac{a(\mathbf{x})b(\mathbf{y})}{\mu_0(\mathbf{x})\nu_0(\mathbf{y})}\mathcal{K}(\mathbf{x},\mathbf{y})$$

in $h(\mathbf{y})$. From Assumptions **A2-A4** these are bounded,

$$\begin{aligned}
\left|\frac{\partial d_\nu^{\gamma,\tau}}{\partial g(\mathbf{x})}\right| &\leq U^{\max} + \frac{\exp\left(\frac{2}{\gamma}KH\right)}{U_0^{\min}V_0^{\min}} \triangleq L_g \\
\left|\frac{\partial d_\nu^{\gamma,\tau}}{\partial h(\mathbf{y})}\right| &\leq V^{\max} + \frac{\exp\left(\frac{2}{\gamma}KH\right)}{U_0^{\min}V_0^{\min}} \triangleq L_h.
\end{aligned}$$

Letting $L = \max\{L_g, L_h\}$, this implies

$$|d_\nu^{\gamma,\tau}(\mathbf{x},\mathbf{y},g,h) - d_\nu^{\gamma,\tau}(\mathbf{x},\mathbf{y},g',h')| \leq L\|(g(\mathbf{x}),h(\mathbf{y})) - (g'(\mathbf{x}),h'(\mathbf{y}))\|_1,$$

for all $(g,h),(g',h') \in \mathcal{G}\times\mathcal{G}$ and $(\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{X}$. □

Note that assumption **A5** is satisfied by an advection-diffusion, so long as we assume w is bounded below, as

$$\max_{g\in\mathcal{G},\mathbf{x}\in\mathcal{X}} \left|\nabla f^*(-\frac{1}{\tau}g(\mathbf{x}))\right| = \max_{g\in\mathcal{G},\mathbf{x}\in\mathcal{X}} \exp(-\frac{\beta}{\tau}g(\mathbf{x}) - w(\mathbf{x})) \leq \exp(\frac{\beta}{\tau}KH - \beta W)$$

with $W = \min_{\mathbf{x}\in\mathcal{X}} w(\mathbf{x})$.

Under the assumptions, then, we get uniform convergence of the stochastic dual objective (3.19) to its expectation (3.17), and this suffices to guarantee consistency.

Proposition 9 (Consistency of unregularized stochastic approximation). *Let $D_\nu^{\gamma,\tau}$ and $D_{\nu,N}^{\gamma,\tau}$ be defined as in* (3.17) *and* (3.19)*, respectively, with $\gamma,\tau,N>0$, and suppose Assumptions* **A1**-**A5** *hold. Let (g_N,h_N) optimize $D_{\nu,N}$ and (g_∞,h_∞) optimize $D_\nu^{\gamma,\tau}$. Then for any $\delta>0$, with probability at least $1-\delta$ over the sample of size N,*

$$D_\nu^{\gamma,\tau}(g_\infty,h_\infty) - D_\nu^{\gamma,\tau}(g_N,h_N) \leq \mathcal{O}\left(\sqrt{\frac{(HKL)^2\log(1/\delta)}{N}}\right). \tag{3.21}$$

Proof. Note that $d_\nu^{\gamma,\tau}$ is jointly convex in $g(\mathbf{x})$ and $h(\mathbf{y})$, and these are in linear in g and h, respectively. They can be written $g(\mathbf{x}) = \langle g, \kappa(\mathbf{x},\cdot)\rangle_\mathcal{H}$ with $\|\kappa(\mathbf{x},\cdot)\|_\mathcal{H} \leq K$ and $\|g\|_\mathcal{H} \leq H$, and similarly for $h(\mathbf{y})$, with the same bounds.

From [138] Thm. 1, then, we have uniform convergence of the empirical functional to its expectation, such that with probability $1-\delta$

$$\sup_{g,h\in\mathcal{H}} \left|D_\nu^{\gamma,\tau}(g,h) - D_{\nu,N}^{\gamma,\tau}(g,h)\right| \leq \mathcal{O}\left(\sqrt{\frac{(HKL)^2\log(1/\delta)}{N}}\right),$$

for any $g, h \in \mathcal{G}$. This implies

$$D_\nu^{\gamma,\tau}(g_\infty, h_\infty) - D_{\nu,N}^{\gamma,\tau}(g_\infty, h_\infty) + D_{\nu,N}^{\gamma,\tau}(g, h) - D_\nu^{\gamma,\tau}(g, h) \leq \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right)$$

$$\Rightarrow D_\nu^{\gamma,\tau}(g_\infty, h_\infty) - D_\nu^{\gamma,\tau}(g, h) \leq \left(D_{\nu,N}^{\gamma,\tau}(g_\infty, h_\infty) - D_{\nu,N}^{\gamma,\tau}(g, h)\right) + \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right)$$

$$\leq \left(D_{\nu,N}^{\gamma,\tau}(g_N, h_N) - D_{\nu,N}^{\gamma,\tau}(g, h)\right) + \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right)$$

for any $g, h \in \mathcal{G}$. In particular, it's true for $g = g_N$ and $h = h_N$, which yields the statement. □

λ-strong convexity of the regularized stochastic dual problem (3.20) actually guarantees a faster $\mathcal{O}(\frac{1}{N})$ convergence rate, for fixed λ. This allows us to get $\mathcal{O}(\frac{1}{\sqrt{N}})$ convergence to the unregularized solution, with decreasing λ.

Proposition 10 (Consistency of regularized stochastic approximation). *Let (g_λ, h_λ) optimize the λ-regularized objective (3.20) for a sample of size N, with $\lambda = \Theta\left(\sqrt{\frac{(KL)^2 \log(1/\delta)}{H^2 N}}\right)$. Let (g_∞, h_∞) optimize $D_\nu^{\gamma,\tau}$. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the sample,*

$$D_\nu^{\gamma,\tau}(g_\infty, h_\infty) - D_\nu^{\gamma,\tau}(g_\lambda, h_\lambda) \leq \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right). \tag{3.22}$$

Proof. Define the regularized expected objective

$$R_\nu^{\gamma,\tau,\lambda}(g, h) = D_\nu^{\gamma,\tau}(g, h) - \frac{\lambda}{2}(\|g\|_{\mathcal{H}}^2 + \|h\|_{\mathcal{H}}^2),$$

and $R_{\nu,N}^{\gamma,\tau,\lambda}$ its empirical counterpart. The regularizer is λ-strongly convex, jointly in g and h.

Let $(g_{\lambda,\infty}, h_{\lambda,\infty})$ optimize $R_\nu^{\gamma,\tau,\lambda}$. From [138] Theorem 2, then, we have for a fixed

λ

$$R_\nu^{\gamma,\tau,\lambda}(g_{\lambda,\infty},h_{\lambda,\infty})-R_\nu^{\gamma,\tau,\lambda}(g,h)\leq 2\left(R_{\nu,N}^{\gamma,\tau,\lambda}(g_\lambda,h_\lambda)-R_{\nu,N}^{\gamma,\tau,\lambda}(g,h)\right)+\mathcal{O}\left(\frac{(KL)^2\log(1/\delta)}{\lambda N}\right),$$

for any $g,h\in\mathcal{G}$. Letting $g=g_\lambda$ and $h=h_\lambda$, and rearranging, we get

$$\begin{aligned}R_\nu^{\gamma,\tau,\lambda}(g_\lambda,h_\lambda)+\mathcal{O}\left(\frac{(KL)^2\log(1/\delta)}{\lambda N}\right)&\geq R_\nu^{\gamma,\tau,\lambda}(g_{\lambda,\infty},h_{\lambda,\infty})\\&\geq R_\nu^{\gamma,\tau,\lambda}(g_\infty,h_\infty)\\&=D_\nu^{\gamma,\tau}(g_\infty,h_\infty)-\frac{\lambda}{2}(\|g_\infty\|_{\mathcal{H}}^2+\|h_\infty\|_{\mathcal{H}}^2).\end{aligned}$$

Hence

$$D_\nu^{\gamma,\tau}(g_\lambda,h_\lambda)+\mathcal{O}\left(\frac{(KL)^2\log(1/\delta)}{\lambda N}\right)\geq D_\nu^{\gamma,\tau}(g_\infty,h_\infty)-\frac{\lambda}{2}(\|g_\infty\|_{\mathcal{H}}^2+\|h_\infty\|_{\mathcal{H}}^2).$$

Rearranging,

$$D_\nu^{\gamma,\tau}(g_\infty,h_\infty)-D_\nu^{\gamma,\tau}(g_\lambda,h_\lambda)\leq\frac{\lambda}{2}(\|g_\infty\|_{\mathcal{H}}^2+\|h_\infty\|_{\mathcal{H}}^2)+\mathcal{O}\left(\frac{(KL)^2\log(1/\delta)}{\lambda N}\right).$$

Bounding $\|g\|_{\mathcal{H}}^2+\|h\|_{\mathcal{H}}^2\leq 2H^2$ and plugging in $\lambda=\Theta\left(\sqrt{\frac{(KL)^2\log(1/\delta)}{H^2N}}\right)$ yields the bound. □

3.5.2 Computational complexity

Complexity of first order descent methods for the stochastic dual problem is dominated by evaluation of the functions g and h at each iteration, for each sample $(\mathbf{x}_i,\mathbf{y}_i)_{i=1}^N$. Each pointwise evaluation of g at a point $\mathbf{x}$ (and analogously for h at $\mathbf{y}$) requires evaluating the sum $\sum_{i=1}^N\kappa(\mathbf{x},\mathbf{x}_i)\alpha_i$, with α_i being the coefficients parameterizing the function. Hence straightforward serial evaluation of g and h at each

iteration is $\mathcal{O}(N^2)$. These sums, however, are trivially parallelizable. Moreover, for certain kernels (notably Gaussian kernels), the serial complexity can be reduced to $\mathcal{O}(N)$, by applying a fast multipole method such as the fast Gauss transform [75].

3.6 Application: nonlinear filtering

We demonstrate the Wasserstein gradient flow approximation in a continuous-time, nonlinear filtering task.

3.6.1 Continuous-discrete filtering

We focus on diffusions that are observed partially or indirectly. A partially-observed diffusion defines a measurement process that samples the diffusion at discrete times. This is a discrete-time stochastic process Y_k taking values in the measurement domain $\mathcal{Y}$, at times t_k, which is related to the underlying diffusion X_t by

$$Y_k = h(X_{t_k}) + \mathbf{v}_k$$

with $h : \mathcal{X} \to \mathcal{Y}$ the measurement function and $\mathbf{v}_k \sim \mathcal{N}(0, \sigma_Y^2)$ noise.

Given a sequence of measurements $Y_k = \mathbf{y}_k$, for $k = 0, \ldots, K$, the *continuous-discrete filtering* problem is that of determining the underlying state X_t for present and future times $t \geq t_K$. Letting $\mathbf{y}_{0:K}$ be the set of measurements up to time t_K, the goal is to evaluate the distribution over states $\Pr(X_t = \mathbf{x}_t | \mathbf{y}_{0:K})$. For future times $t > t_K$, this is the *marginal prior* or *predictive* distribution over states, defined by the dynamics of the diffusion process, satisfying the forward PDE (3.2) with initial density $\Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K})$. At the measurement time $t = t_K$, this is the *marginal posterior*, conditional upon the measurements, and is defined by a recursive update equation

$$\Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K}) = \frac{\Pr(Y_K = \mathbf{y}_K | X_{t_K} = \mathbf{x}_{t_K}) \Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K-1})}{\Pr(Y_K = \mathbf{y}_K)}.$$

The term $\Pr(X_{t_K} = \mathbf{x}_{t_K} | \mathbf{y}_{0:K-1})$ is the predictive distribution given the measurements up to time t_{K-1}. We assume an initial distribution $\Pr(X_{t_0} = \mathbf{x}_{t_0})$ is given.

3.6.2 Double-well system

Here we are tracking a particle that diffuses in a double-well potential, given by $w(\mathbf{x}) = -2\mathbf{x}^2 + 2\mathbf{x}^4$. The particle's dynamics are described by an Itô stochastic differential equation,

$$d\mathbf{x}_t = (4\mathbf{x} - 8\mathbf{x}^3)dt + d\mathbf{W}_t, \tag{3.23}$$

having unit dispersion coefficient $\beta = 1$. This diffusion exhibits discrete switching, as the particle occasionally switches between the two potential wells (Figure 3-5b), with frequency controlled by the dispersion coefficient. The particle's location is observed noisily at discrete timepoints, via an observation function h,

$$\mathbf{y}_k = h(\mathbf{x}_{t_k}) + \mathbf{v}_k \tag{3.24}$$

with $\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \sigma)$ additive Gaussian observation noise having standard deviation σ. We look at two observation regimes:

1. direct observation: $h(\mathbf{x}_{t_k}) = \mathbf{x}_{t_k}$ and $\sigma = 1$, and

2. quadratic observation: $h(\mathbf{x}_{t_k}) = \mathbf{x}_{t_k}^2$ and $\sigma = 0.1$.

Figures 3-6a and 3-6b show typical posterior distributions obtained by solving the filtering equations on a discrete grid [6]. Under direct observation, the posterior is unimodal but often skewed and non-Gaussian. The quadratic observation loses information about the sign of the location $\mathbf{x}_{t_k}$, leading to bimodal posterior densities.

3.6.3 Results

We apply the Wasserstein gradient flow to approximate the predictive density of the diffusion, which at measurement times is multiplied pointwise with the likelihood

[6] We use Chang and Cooper's method [37] on a regularly-spaced grid of 1000 points in $[-3, 3]$.

$\Pr(\mathbf{y}_k|\mathbf{x}_{t_k})$ to obtain an unnormalized posterior density.

We use five methods as baselines for comparison. The first computes the exact predictive density by numerically integrating the Fokker-Planck equation (3.4) on a fine grid – this allows us to compare computed posteriors to the exact, true posterior. The second and third are the Extended and Unscented Kalman filters, which maintain Gaussian approximations to the posterior. The fourth method is a Gaussian sum filter [5], which approximates the posterior by a mixture of Gaussians. And the fifth baseline is a bootstrap particle filter, which samples particles according to the transition density $\Pr(\mathbf{x}_{t_k}|\mathbf{x}_{t_{k-1}})$, by numerical forward simulation of the SDE (3.23) [7].

We simulate 20 observations at a time interval of $\Delta t = 0.5$, and compute the posterior density by each of the methods. Figures 3-7 and 3-8 show examples of the posterior evolution, with the exact solution shaded in blue and the various approximate methods overlaid. Qualitatively, the Wasserstein gradient flow approximation closely captures the non-Gaussian and bimodal shapes of the posteriors, while the other methods struggle. The extended Kalman filter, in particular, is very sensitive to initialization and tends to follow only a single mode of the posterior, and the Gaussian approximation in both the extended and unscented Kalman filters fails to capture both the skew of the posterior in the directly observed case and the bimodality of the quadratic observation case. The Gaussian sum filter, although in theory having the capacity to represent both unimodal and bimodal posteriors quite accurately, suffers from underlying linearization of the dynamics (leading it to over- or under-shoot the true location of the mode[s]) and introduces spurious asymmetry of the modes in the quadratic observation case. Not shown is the bootstrap particle filter, which can approximate an arbitrary posterior, if given enough samples.

Figure 3-9 shows quantitatively the fidelity of the estimated posterior to that computed by exact numerical integration, repeating the filtering experiment 100 times. We use the $1-$Wasserstein distance between the two densities [8], as it allows us to

[7] For foward simulation, we use an Euler's method with timestep 10^{-3}.

[8] The 1-Wasserstein distance is used with the Euclidean distance as the underlying metric. We compute it using `fastemd`.

treat the different posterior representations equivalently. Our Wasserstein gradient flow approximation, the Kalman filters and the Gaussian sum filter all give continuous representations of the posterior density, which can be evaluated at the same grid points as the exact method, while the particle filter represents the posterior by a collection of weighted delta functions, not straightforwardly extrapolated to the grid points.

The Wasserstein gradient flow approximation consistently outperforms the baselines, both qualitatively and quantitatively, achieving smaller Wasserstein distance to the true posterior.

3.7 Conclusion

We present a novel approximate inference method for diffusion processes, based on the Wasserstein gradient flow formulation of the diffusion. In this formulation, the time-dependent density of the diffusion is derived as the limit of implicit Euler steps that follow the gradients of a particular free energy functional. Existing methods for computing Wasserstein gradient flows rely on discretization of the domain of the diffusion, prohibiting their application to domains in more than several dimensions. We propose instead a discretization-free inference method that computes the Wasserstein gradient flow directly in a space of continuous functions. We characterize approximation properties of the proposed method and evaluate it on a nonlinear filtering task, finding superior performance to standard methods for filtering diffusions.

3.8 Experimental details

3.8.1 Instability of stochastic approximation

Figure 3-3 shows example solutions for the regularized stochastic approximation. Here the solutions are overlaid on the ground truth density obtained by computing a discrete (unregularized) Wasserstein gradient flow on a fine grid of 1000 points on $[-3, 3]$, using a Dykstra's method [120]. The stochastic approximation is computed indepen-

dently 5 times for each figure. We use time step $\tau = 0.1$ and Wasserstein regularization $\gamma = 0.01$, and vary the regularizer λ. The potential is the quadratic $w(\mathbf{x}) = \mathbf{x}^2$, and the initial density ν is a mixture of two Gaussians, with centers at $-1, +1$ and both having standard deviation 0.1. For sampling, we use standard normal distributions. We compute the stochastic approximation using a BFGS algorithm, stopping when the norm of the gradient is less than 10^{-2}.

3.8.2 Bias-variance tradeoff

Figure 3-4 shows the bias-variance tradeoff for $N = 200, 400, 1000$. Here the bias is computed as the RMSE of the mean distribution obtained from 100 independent runs of the stochastic approximation algorithm. From these 100 runs, we compute 100 bootstrap samples of size 100, and for each run we evaluate its resulting probability density μ at a set of 1000 grid points on $[-3, 3]$ and normalize the result to sum to 1. The bias is the RMSE of mean distribution taken within each bootstrap sample, with respect to the distribution obtained by computing a discrete (unregularized) Wasserstein gradient flow on the same grid, using a Dykstra's algorithm [120]. Similarly, the variance is computed as the RMSE of the estimated probability density with respect to the mean distribution, within each bootstrap sample, and the total error is the RMSE of the estimated density with respect to the ground truth unregularized Wasserstein flow. The figures show the bootstrapped median and 95% intervals for all three values.

Here the underlying potential function is the quadratic $w(\mathbf{x}) = \mathbf{x}^2$, and we use dispersion parameter $\beta = 1$, time step $\tau = 0.5$, and Wasserstein regularization $\gamma = 0.01$. The initial density is a mixture of two Gaussians, centered at $-1, +1$ and having standard deviation 0.1. We compute the stochastic approximation using a BFGS algorithm, stopping when the norm of the gradient is less than 10^{-2}.

3.8.3 Filtering

Problem setup and data generation

Latent state trajectories are generating from the SDE model

$$d\mathbf{x}_t = (4\mathbf{x} - 8\mathbf{x}^3)dt + d\mathbf{W}_t$$

which is an advection-diffusion with potential $w(\mathbf{x}) = -2\mathbf{x}^2 + 2\mathbf{x}^4$ and inverse dispersion coefficient $\beta = 1$.

The latent system is observed at a time interval of $\Delta t = 0.5$, with additive Gaussian noise having standard devation σ, in one of two regimes:

1. direct observation: $h(\mathbf{x}_{t_k}) = \mathbf{x}_{t_k}$ and $\sigma = 1$, and
2. quadratic observation: $h(\mathbf{x}_{t_k}) = \mathbf{x}_{t_k}^2$ and $\sigma = 0.1$.

State trajectories are generated by simulating the SDE using an Euler-Maruyama method with timestep 10^{-3}.

Baselines

Discretized numerical integration. We construct a regularly-spaced grid of 1000 points on the interval $[-3, 3]$, and use Chang and Cooper's method [37] to integrate the Fokker-Planck equation for the dynamics. We use a timestep of 10^{-3} for the integration.

When filtering, we obtain the posterior state distribution by first propagating forward the posterior at the previous observation time, via integrating the Fokker-Planck equation, then multiplying the resulting distribution pointwise by the observation likelihood and normalizing to sum to one.

Extended Kalman filter. The extended Kalman filter is implemented as described in [30]. We use Scipy's `odeint` to integrate the ODE for the mean and covariance. The EKF is initialized with a Gaussian of whose mean is drawn from a normal distribution having mean 0 and standard deviation 0.1, and whose variance is 10^{-4}.

Unscented Kalman filter. The unscented Kalman filter is implemented as described in [131]. We use Scipy's `odeint` to integrate the ODE for the mean and covariance. The UKF is initialized with a Gaussian of mean 0 and variance 10^{-4}. We use parameters $\alpha = \frac{1}{2}$, $\beta = 2$, $\kappa = 1$. (β here refers to the parameter in [131], rather than the inverse dispersion coefficient in the main text.)

Gaussian sum filter. We implement a Gaussian sum filter as described in [5]. The filter is initialized with a mixture of eight Gaussians, having means drawn independently from a normal distribution with mean 0 and standard deviation 0.1, and each having variance 10^{-4}.

Bootstrap particle filter. The bootstrap particle filter is implemented as described in [73]. For propagating particles forward in time, we simulate the system dynamics using an Euler-Maruyama method with timestep 10^{-3}. We resample trajectories after each observation.

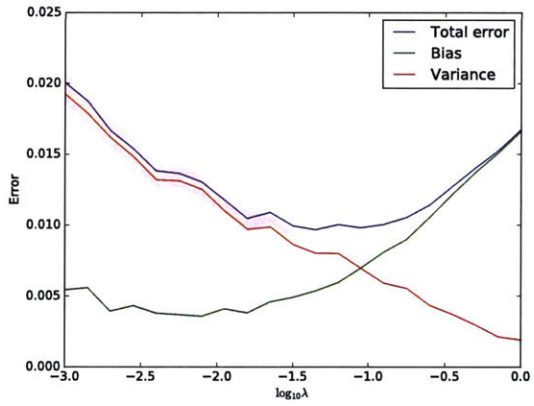
Example posterior evolution

Figures 3-7 and 3-8 show an example of the evolution of the posterior distribution for consecutive timesteps. We simulate system trajectories and observations as described above and use the regularized stochastic approximation algorithm for the Wasserstein gradient flow to propagate the posterior at one observation time to the next. The resulting distribution is multiplied pointwise by the likelihood to obtain an unnormalized posterior. We compute the normalizer for the posterior by Monte Carlo integration with 10000 samples. The sampling distribution for the stochastic approximation is taken from a UKF initialized with mean and variance computed from the previous timestep's posterior. We use $\gamma = 0.01$ and $\lambda = 1e - 2$ in both observation regimes. We compute the stochastic approximation using a BFGS algorithm, stopping when the norm of the gradient is less than 10^{-2}.

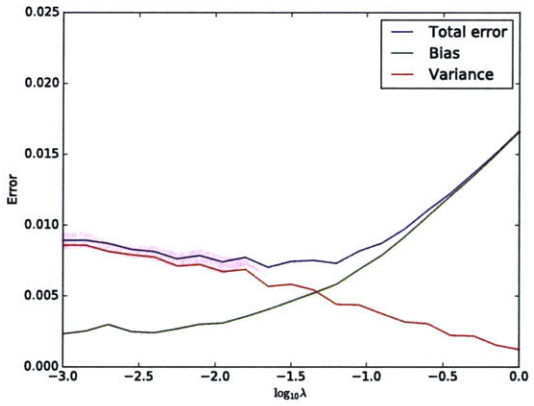
We additionally overlay posterior distributions for the baseline algorithms. The distribution obtained from discretized numerical integration is shaded in blue. For visualization, all distributions are sampled on a grid and normalized to sum to one.

Quantitative comparison of methods

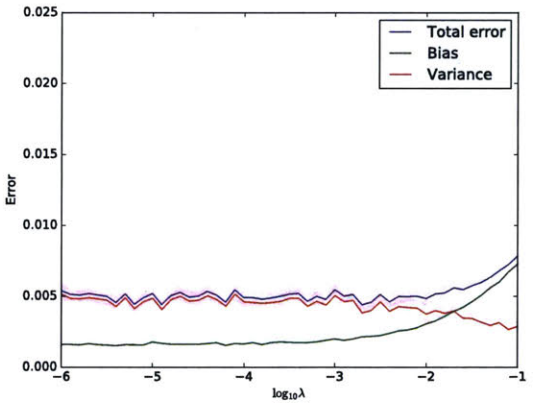
We simulate 100 independent latent state trajectories and their observations. For each we obtain posterior distributions for the proposed Wasserstein gradient flow approximation and the baseline methods, as described above. For all but the bootstrap particle filter, we sample the resulting distributions on the same grid as was used for discretized numerical integration and normalize to sum to one. We compute the $1-$Wasserstein distance, with Euclidean distance as the cost, between the exact distribution from discretized numerical integration and the approximate distribution from the given method. For the bootstrap filter, we compute the same distance, using the original posterior (being a normalized sum of delta functions).



(a) Bias-variance tradeoff ($N = 200$).

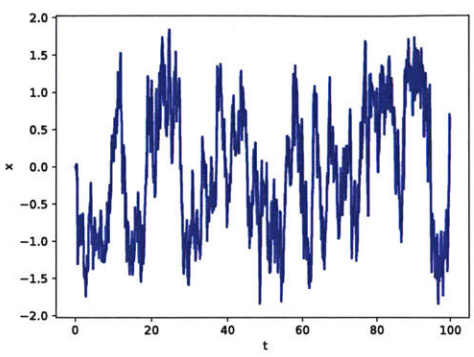


(b) Bias-variance tradeoff ($N = 400$).

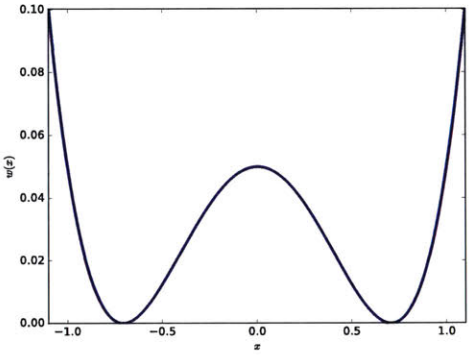


(c) Bias-variance tradeoff ($N = 1000$).

Figure 3-4: Regularization parameter λ induces a bias-variance tradeoff. Note that the x-axis scale is shifted for $N = 1000$. For large enough N, regularization has no impact on total accuracy, up to a threshold value of λ (roughly $\lambda = 10^{-3}$ when $N = 1000$).

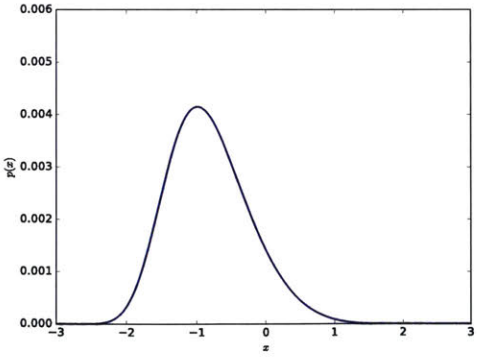


(a) Example trajectory from the double well system.

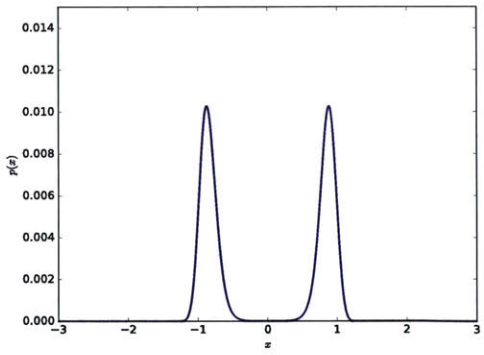


(b) Double well potential.

Figure 3-5: Diffusion in a double well potential.

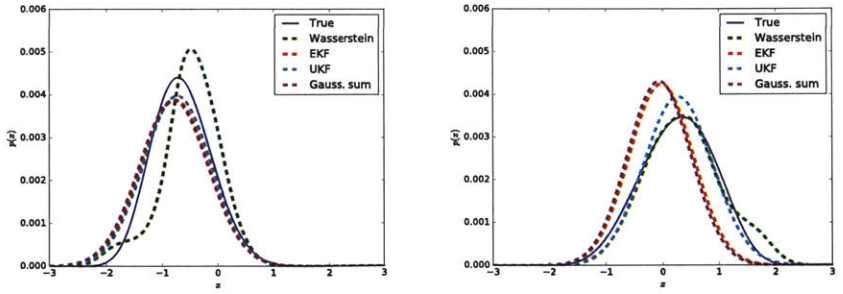


(a) Directly observed.



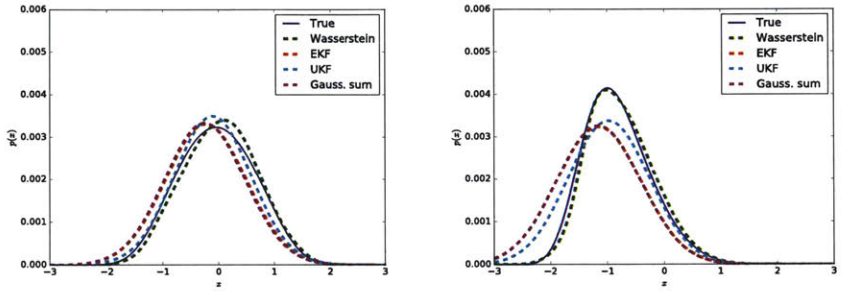
(b) Quadratic observation.

Figure 3-6: Filtering a double well diffusion, example posteriors.



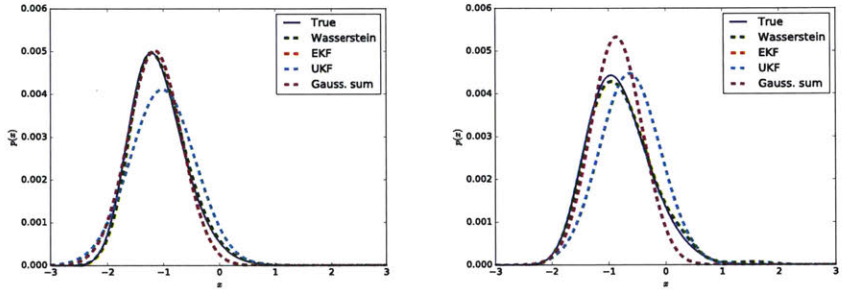
(a) $t = 0.5$.

(b) $t = 1$.



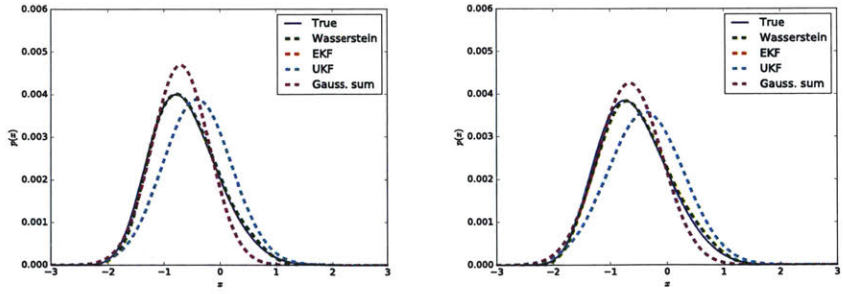
(c) $t = 1.5$.

(d) $t = 2$.



(e) $t = 2.5$.

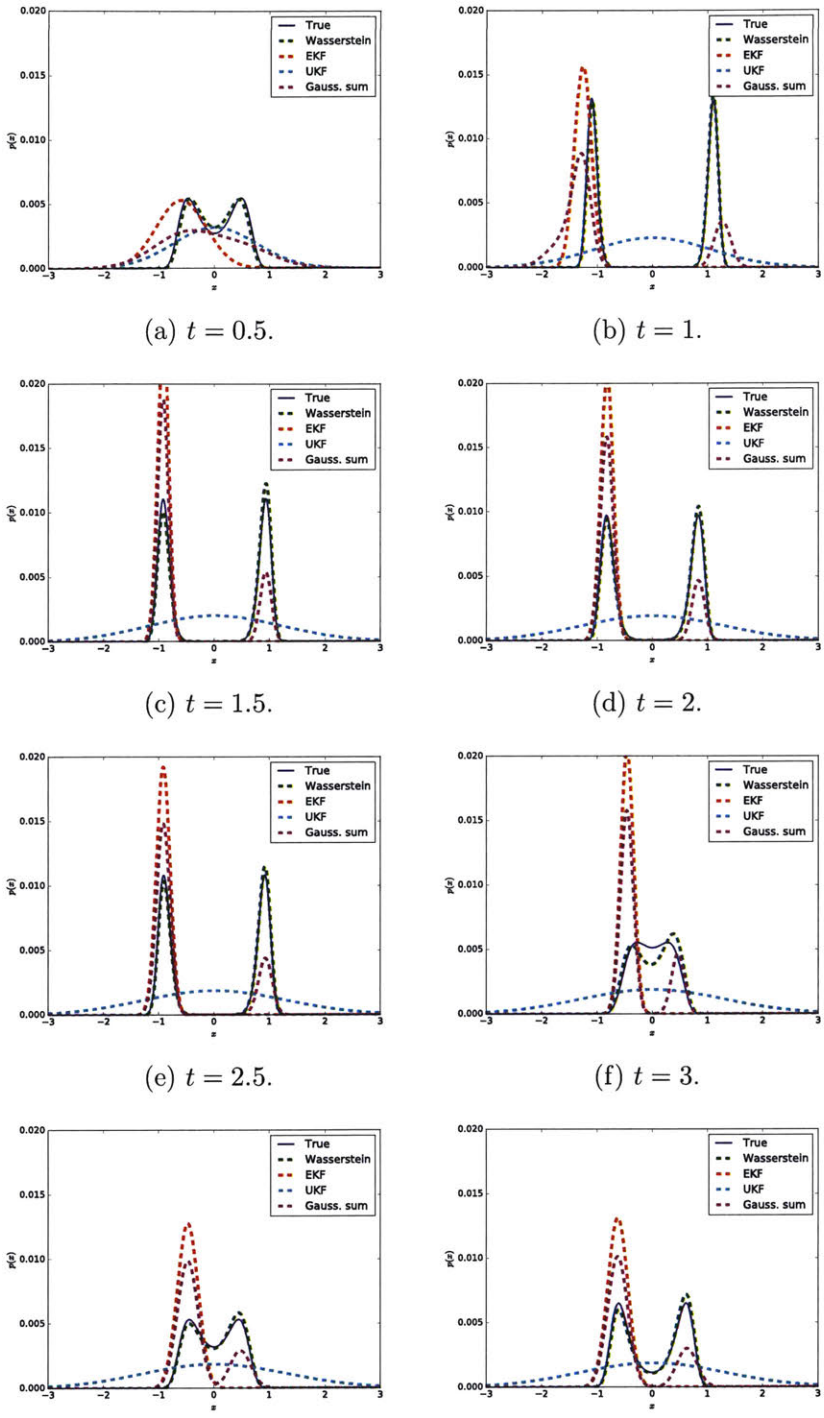
(f) $t = 3$.



(g) $t = 3.5$.

(h) $t = 4$.

Figure 3-7: Double well potential with direct observations. Evolution of the posterior density, with estimates from the various methods overlaid. Shaded region is the exact solution.



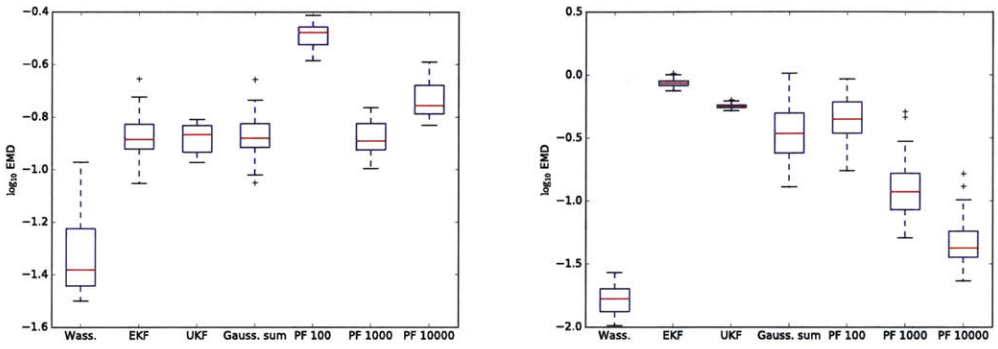
(a) $t = 0.5$. (b) $t = 1$.

(c) $t = 1.5$. (d) $t = 2$.

(e) $t = 2.5$. (f) $t = 3$.

(g) $t = 3.5$. (h) $t = 4$.

Figure 3-8: Double well potential with quadratic observations. Evolution of the posterior density, with estimates from the various methods overlaid. Shaded region is the exact solution.



(a) Direct observations.

(b) Quadratic observations.

Figure 3-9: Wasserstein distance to the true posterior.

Chapter 4

Ecological inference

In certain federal redistricting cases, the prosecution needs to establish that disenfranchisement has occurred: specifically, a protected group's vote has been diluted, such that their preferences are no longer represented. This is an empirical problem, as the lawyers need to demonstrate that there is a significant difference between the voting preferences of the protected group and those of new majority. Such data are typically inaccessible, however: voting occurs by secret ballot, such that only total vote counts are available, and not their demographic breakdown. The solution is to make what is called an ecological inference: to reason backwards from the aggregate data (total vote counts and census data) and infer the voting behaviors of individual subgroups of the population.

Ecological inference appears frequently in epidemiology, economics and social sciences, where one is limited to aggregate surveys of a population and wishes to combine these surveys for a more refined characterization. It is a well-known source of error in interpreting statistical data, as well. If European countries having larger minority populations tend also to have more votes for liberal candidates, we might (wrongly) be tempted to conclude that minority voters prefer liberal candidates. Something that is true for the group needs not be true for the individual. Such reasoning is called the ecological fallacy.

Mathematically, the inference problem is as follows. Given two partitions of a population into m and n groups, we want to infer a table $\pi \in \mathbb{R}^{m \times n}$ in which the

element π_{ij} gives the proportion of the population that belongs simultaneously to the ith and jth groups – in the voting rights case, these are the demographic groups and the political parties. We only have access, however, to the marginals of the table, $\pi\mathbf{1}$ and $\pi^\intercal\mathbf{1}$, which are the aggregate proportions of the groups. The problem is fundamentally ill-posed: for any pair of observed marginals there are many possible tables that might fit these marginals. Methods for solving the problem must somehow distinguish better and worse amongst the possible solutions – they require additional prior assumptions.

A number of methods have been proposed to solve the ecological inference problem, and each has its drawbacks. Simplest is the "neighborhood" model, which assumes (possibly wrongly) that the two aggregate measurements are entirely independent (i.e. the table is the product of its marginals). Goodman's regression method [72] does not require independence and is efficient to compute, but it is also not constrained to produce tables with entries inside the valid range of proportions ($[0, 1]$), making the results sometimes difficult to interpret. Hierarchical Bayesian models proposed by King [87], Rosen [124], Wakefield [159] and others represent more complex prior assumptions, but rely on inefficient Markov Chain Monte Carlo inference methods and can be sensitive to hyperparameter selection [159].

In this work, we propose a novel framework for ecological inference, which encompasses a variety of priors and allows for efficient computation of the most probable solution. Unlike previous methods, which rely on Monte Carlo estimates of the posterior distribution over tables, our inference procedure uses an efficient fixed point iteration that is linearly convergent and requires memory scaling as the size of the table. The method also generalizes naturally to tables having more than two observed marginals. Importantly, with the right prior, the inferred tables can be more accurate than those from existing methods, on both synthetic and real data. We additionally give a method for interval estimation.

This chapter is organized as follows. In Section 4.1, we define the ecological inference problem and discuss existing methods. In Section 4.2, we define the model underlying our framework. In Section 4.3 we derive an efficient algorithm for MAP

estimation of the table. In Section 4.4 we discuss estimation of the parameters of the prior distribution. In Section 4.5 we derive a method for interval estimation. In 4.6 we evaluate our methods on synthetic and real data, before concluding.

4.1 Background

4.1.1 The ecological inference problem

Let X and Y be random variables defined on discrete domains $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^m$ and $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^n$, respectively. Each represents a distinct measurement of an underlying population, in which we sample uniformly at random an individual from the population and assign that individual to one of a set of fixed categories. These might be demographic groups for X and political party affiliations for Y, for example. The value of the random variable X or Y is the category to which the individual belongs, and its distribution is determined by the proportion of individuals in the population belonging to each category.

Let $\mathbf{u} \in \Delta^m$ and $\mathbf{v} \in \Delta^n$ represent the distributions of X and Y, respectively, with elements $\mathbf{u}_i = \Pr(X = \mathbf{x}_i)$ and $\mathbf{v}_j = \Pr(Y = \mathbf{y}_j)$. Each vector gives the proportions of individuals falling into the various categories.

The two measurements X and Y need not be independent. In fact, for a pair of random measurements there can be infinitely many *joint distributions* that are consistent with both measurements simultaneously. We represent their joint distribution by a matrix $\pi \in \Delta^{m \times n}$, with $\Delta^{m \times n}$ the simplex of $m \times n$ nonnegative real matrices whose elements sum to 1, and $\pi_{ij} \triangleq \Pr(X = \mathbf{x}_i, Y = \mathbf{y}_j)$.

A joint distribution π is consistent with the measurements X and Y if its *marginals* match the distributions $\mathbf{u}$ and $\mathbf{v}$, meaning $\mathbf{u}_i = \sum_{j=1}^n \pi_{ij},\ \forall i$, and $\mathbf{v}_j = \sum_{i=1}^m \pi_{ij},\ \forall j$. We can concisely write this as $\pi\mathbf{1} = \mathbf{u}$ and $\pi^\intercal\mathbf{1} = \mathbf{v}$, with $\mathbf{1}$ the all-ones vector of the correct dimension. The set of all joint distributions consistent with the given marginals we denote $\Pi(\mathbf{u}, \mathbf{v})$, such that

$$\Pi(\mathbf{u}, \mathbf{v}) = \{\pi \in \Delta^{m \times n} | \pi\mathbf{1} = \mathbf{u},\ \pi^\intercal\mathbf{1} = \mathbf{v}\}.$$

The **ecological inference problem** consists of recovering the joint distribution π, given only the measurement distributions $\mathbf{u} \in \Delta^m$ and $\mathbf{v} \in \Delta^n$. In other words, given two aggregate categorizations of the population, we would like to recover their intersection. The problem is clearly ill-posed, as we have an entire set $\Pi(\mathbf{u}, \mathbf{v})$ of valid joint distributions for the given marginals. The challenge will be entirely how to define and compute a "best" joint distribution.

4.1.2 Related work

A number of methods have been proposed for ecological inference.

Neighborhood model

The most basic model is possibly the "neighborhood" model, in which the joint table π is assumed to be the product of its marginals,

$$\pi = \mathbf{u}\mathbf{v}^\intercal.$$

This is equivalent to an assumption that the marginal variables are independent.

Goodman's regression

Goodman's regression [72] was among the first proposed methods for ecological inference and, along with its generalization to arbitrary-sized tables [94] [102] [88] [76], has been widely used in practice [87]. The model makes three assumptions. First is that the individual tables $\pi^{(i)}$ decompose as

$$\pi^{(i)} = \mathsf{diag}(\mathbf{u}^{(i)})\rho, \quad \forall i$$

for some row-wise proportionality matrix $\rho \in \mathbb{R}^{m \times n}$ that is shared between the instances (i). The second is that ρ is row-stochastic, such that the row marginal constraint holds,

$$\pi^{(i)}\mathbf{1} = \mathsf{diag}(\mathbf{u}^{(i)})\rho = \mathbf{u}^{(i)}, \quad \forall i.$$

And the last assumption gives a regression model for the column marginal,

$$\mathbf{v}^{(i)} = \rho^{\intercal}\mathbf{u}^{(i)} + \varepsilon^{(i)}, \quad \forall i,$$

with $\varepsilon^{(i)}$ a normally distributed error. Note that the model is linear in the coefficients ρ_{jk}, so we can efficiently solve via linearly-constrained ordinary least squares. Note also that the assumption of constant ρ is restrictive and often unrealistic.

King's method

King [87] proposed a method aimed at addressing deficiencies in Goodman's and related models. King's method assumes that the tables $\pi^{(i)}$ decompose as

$$\pi^{(i)} = \mathsf{diag}(\mathbf{u}^{(i)})\rho^{(i)}, \quad \forall i \tag{4.1}$$

with $\rho^{(i)}$ rowwise proportionality matrices, constrained to be row-stochastic, $\rho^{(i)}\mathbf{1} = \mathbf{1}$, $\forall i$, such that the rowwise constraints $\pi^{(i)}\mathbf{1} = \mathbf{u}^{(i)}, \forall i$ are satisfied. Note that $\rho^{(i)}$ is not shared across instances. Furthermore, the column-wise constraints are enforced exactly via

$$\rho^{(i)\intercal}\mathbf{u}^{(i)} = \mathbf{v}^{(i)}, \quad \forall i, \tag{4.2}$$

unlike in Goodman's model. Finally, King imposes a prior distribution on $\rho^{(i)}$, assuming it is sampled from a truncated normal distribution, restricted to the hypercube $[0,1]^{m\times n}$:

$$\rho^{(i)} \sim \mathcal{T}_{[0,1]^{m\times n}}\mathcal{N}(\mu, \Sigma).$$

This truncation prevents uninterpretable values for entries of $\rho^{(i)}$, which in Goodman's case can fall outside $[0,1]$. By assuming independence of the observations $\mathbf{u}^{(i)}, \mathbf{v}^{(i)}$ conditional upon the parameters μ, Σ, the row- and column-wise constraints (4.1) and (4.2) imply posterior predictive distributions $\Pr(\rho^{(i)}|\mathbf{u}^{(i)}, \mathbf{v}^{(i)})$ via Bayesian averaging of the prior $\Pr(\rho^{(i)}|\mu, \Sigma)$ with a prior on the parameters $\Pr(\mu, \Sigma)$. This posterior is computed by Monte Carlo simulation.

Hierarchical Bayesian models

Models such as Rosen [124] and Wakefield [159] carry King's idea further, by defining a hierarchical distribution on $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$. Rosen [124] gives a multinomial-Dirichlet model, which assumes the marginal data are expressed as counts. The model has the column marginal $\mathbf{v}^{(i)}$ multinomial distributed, with the rowwise proportions $\rho^{(i)}$ governed by a Dirichlet distribution and the Dirichlet parameters i.i.d. Gamma,

$$\begin{aligned}
\mathbf{v}^{(i)} &\sim \text{Multinomial}(N^{(i)}, \rho^{(i)\intercal}\mathbf{u}^{(i)}), \quad \forall i \\
\rho^{(i)}_{j,\cdot} &\sim \text{Dirichlet}(\alpha^{(i)}_{j,\cdot}), \quad \forall i, j \\
\alpha^{(i)}_{j,k} &\sim \text{Gamma}(\lambda_1, \lambda_2), \quad \forall i, j, k.
\end{aligned}$$

The model is fit using a Metropolis-within-Gibbs algorithm.

Optimal transport

The theory of optimal transport [156] relates probability distributions by defining a transport plan that reassigns the mass in one distribution to match the other. The optimal transport plan minimizes the total cost of moving the mass, with respect to a given cost function. Specifically, for distributions given by vectors $\mathbf{u}$ and $\mathbf{v}$, we solve for a soft assignment matrix $\pi_* \in \Pi(\mathbf{u}, \mathbf{v})$, such that

$$\pi_* = \underset{\pi \in \Pi(\mathbf{u},\mathbf{v})}{\text{argmin}} \langle \pi, \mathbf{C} \rangle_{\mathcal{F}}$$

with $\mathbf{C} \in \mathbb{R}_+^{m \times n}$ the cost matrix having $\mathbf{C}_{ij}$ the cost for transporting a unit of mass from the ith to the jth location, and $\Pi(\mathbf{u}, \mathbf{v})$ the polytope of nonnegative matrices having $\mathbf{u}$ and $\mathbf{v}$ as row and column marginals.

There is significant similarity between optimal transport and ecological inference, in that both involve inferring a matrix with given marginals. A regularized form of optimal transport has, in fact, recently been applied to ecological inference, with quite positive results [112]. The authors there do not compare performance to existing methods. Separately, the algorithm we present in Section 4.3 owes a great deal to

methods derived for the optimal transport setting, in [17] [50].

4.2 Probabilistic model

4.2.1 Well-behaved priors

Our model relies on a prior distribution over tables, which we allow to come from a general class of distributions, being those that are *separable* and *log-concave of Legendre type*, with support in $\Delta^{m\times n}$. With two additional technical assumptions, these distributions are sufficiently well-behaved to enable the efficient optimization in Section 4.3.

Let $\mathcal{P}(\mathbf{C})$ be a family of distributions over $\text{int}(\Delta^{m\times n})$, parameterized by $\mathbf{C} \in \mathbb{R}^{m\times n}$, and let $\Pr(\pi|\mathbf{C})$ denote the density with respect to the Lebesgue measure. Define $Q(\pi) = -\log \Pr(\pi|\mathbf{C})$ the negative log density, for tables $\pi \in \text{dom}\, Q \subseteq \mathbb{R}^{m\times n}$. Formally, we assume the following.

$$\begin{aligned} &(A1)\ \ Q \text{ is separable and Legendre type.} \\ &(A2)\ \ \text{int}(\Delta^{m\times n}) \subseteq \text{dom}\, Q. \\ &(A3)\ \ \text{dom}\, Q^* \text{ is open.} \\ &(A4)\ \ \mathbf{0} \in \text{dom}\, Q^*. \end{aligned} \tag{4.3}$$

Here Q^* is the convex conjugate [1]. For certain priors (such as the Dirichlet prior, Example 4.3.1) we will drop the assumption A4.

We say $\Pr(\pi|\mathbf{C})$ is *separable* if it decomposes as

$$\Pr(\pi|\mathbf{C}) = \prod_{ij} f(\pi_{ij}|\mathbf{C}_{ij}),$$

with $f : \mathbb{R} \times \mathbb{R} \to [0, +\infty]$ a one-dimensional density.

[1] The convex conjugate $Q^* : \text{dom}\, Q^* \to \mathbb{R}$ is defined

$$Q^*(\mathbf{u}) = \sup_{\mathbf{x}\in\text{int}(\text{dom}\, Q)} \langle \mathbf{u}, \mathbf{x}\rangle - Q(\mathbf{x}).$$

$\Pr(\pi|\mathbf{C})$ is *log-concave of Legendre type* if its negative log is convex of Legendre type. Namely, the negative log is closed, proper, essentially smooth and strictly convex on the interior of its domain [2]. Separability and Legendreness of $\Pr(\pi|\mathbf{C})$ are satisfied a number of common distributions, including the (truncated) normal, exponential, gamma, lognormal, beta, and Dirichlet distributions.

One aspect of Legendre-type functions we will exploit is duality between the domain and range of the gradient ∇Q. Specifically, for a Legendre-type Q, the gradient of Q and that of the convex conjugate Q^* define a bijection between $\operatorname{int}(\operatorname{dom} Q)$ and $\operatorname{int}(\operatorname{dom} Q^*)$, with $\nabla Q^* = (\nabla Q)^{-1}$. We formulate, for example, the MAP estimation procedure in Section 4.3 as an optimization over the dual space $\operatorname{dom} Q^*$, and recover the primal solution via the map ∇Q^*.

4.2.2 Model

We assume a common prior distribution $\mathcal{P}(\mathbf{C})$ for N tables $\pi^{(i)}$, which satisfies the regularity properties (4.3). The tables represent different but related instances of the problem, corresponding for example to different geographic regions in the voter preference example. We assume the distribution's parameters $\mathbf{C} \in \mathbb{R}^{m \times n}$ are shared across instances. The model specifies

$$\begin{aligned} &\pi^{(i)} \sim \mathcal{P}(\mathbf{C}), \\ &\pi^{(i)} \perp\!\!\!\perp \pi^{(j)} \mid \mathbf{C}, \forall i \neq j. \end{aligned} \tag{4.4}$$

Conditioned on the observed vectors $\mathbf{u}^{(i)} \in \Delta^m$ and $\mathbf{v}^{(i)} \in \Delta^n$, we will draw inferences about the posterior density $\Pr(\pi|\mathbf{u}^{(i)}, \mathbf{v}^{(i)}, \mathbf{C})$, which is the truncation of the prior to the polytope $\Pi(\mathbf{u}, \mathbf{v})$. This ensures that marginals of the table $\pi^{(i)}$ are consistent with the observations.

[2] Bauschke and Borwein [15] Def. 2.8 and surrounding gives a formal treatment of Legendre type functions.

4.3 Maximum a priori estimation

Each observed marginal vector defines an affine constraint on the table $\pi^{(i)}$: for marginals $\mathbf{u}^{(i)}$ and $\mathbf{v}^{(i)}$, we have $\pi^{(i)}\mathbf{1} = \mathbf{u}^{(i)}$ and $\pi^{(i)\intercal}\mathbf{1} = \mathbf{v}^{(i)}$. Together with the constraints that the entries of the table be nonnegative, these affine constraints define a convex polytope $\Pi(\mathbf{u}, \mathbf{v})$ of tables satisfying the constraints. Maximum a priori estimation finds a table $\pi_*^{(i)}$ that maximizes the posterior $\Pr(\pi^{(i)}|\mathbf{u}^{(i)}, \mathbf{v}^{(i)}, \mathbf{C})$ over this polytope $\Pi(\mathbf{u}, \mathbf{v})$. On its face, MAP estimation is nontrivial, solving for each table $\pi^{(i)}$ a nonlinear objective with $mn + m + n$ linear constraints:

$$\begin{aligned}\pi^{(i)} &= \operatorname*{argmax}_{\pi \in \mathbb{R}_+^{m \times n}} \Pr(\pi|\mathbf{u}^{(i)}, \mathbf{v}^{(i)}, \mathbf{C}) \\ &= \operatorname*{argmax}_{\pi \in \Pi(\mathbf{u}^{(i)}, \mathbf{v}^{(i)})} \Pr(\pi|\mathbf{C}).\end{aligned}$$

4.3.1 MAP estimation is a Bregman projection

Key to the tractability of MAP estimation for priors satisfying the assumptions (4.3) is the fact that it can be formulated as minimization of a Bregman divergence over the polytope of marginal constraints. This is stated in Proposition 11.

Proposition 11 (MAP is divergence minimization)**.** *Let* $\Pr(\pi|\mathbf{C})$ *be a prior density over tables* $\pi \in \mathbb{R}^{m \times n}$*, satisfying the regularity properties* (4.3)*. Define* $Q(\pi) \triangleq -\log \Pr(\pi|\mathbf{C})$ *the negative log density. Then the posterior density* $\Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C})$ *has a unique maximum* π_* *that satisfies*

$$\pi_* = \operatorname*{argmin}_{\pi \in \Pi(\mathbf{u}, \mathbf{v})} \mathcal{D}_Q(\pi, \nabla Q^*(\mathbf{0})) \tag{4.5}$$

with Q^* *the convex conjugate and* $\mathcal{D}_Q$ *the Bregman divergence with respect to* Q*.*

Proof. We first note that Q attains its global minimum at $\pi_* = \nabla Q^*(\mathbf{0}) \in \operatorname{int}(\operatorname{dom} Q)$, as the assumptions A1, A3 and A4 from (4.3) imply that $\nabla Q(\pi_*) = \mathbf{0}$ via the bijective relation $(\nabla Q)^{-1} = \nabla Q^*$, and Q is strictly convex on $\operatorname{int}(\operatorname{dom} Q)$, so the critical point is a global minimum.

The posterior density $\Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C})$ is the truncation of the prior $\Pr(\pi|\mathbf{C})$ to the polytope $\Pi(\mathbf{u}, \mathbf{v})$. Dessein [50] in Section 3.2, Lemma 2, shows that the restriction of Q to $\Pi(\mathbf{u}, \mathbf{v})$ attains its global minimum uniquely at the Bregman projection of the unrestricted global minimum $\pi_* = \nabla Q^*(\mathbf{0})$ onto $\Pi(\mathbf{u}, \mathbf{v})$. The Lemma holds so long as $\nabla Q(\pi_*) = \mathbf{0}$ and $\Pi(\mathbf{u}, \mathbf{v}) \cap \operatorname{int}(\operatorname{dom} Q) \neq \emptyset$; the first is satisfied as noted above, and the second is satisfied by assumption A2 of (4.3). □

In other words, the MAP estimate exists and is unique, and is a Bregman projection onto the marginal constraints $\Pi(\mathbf{u}, \mathbf{v})$. Section 4.3.2 gives a method of alternating Bregman projections, called Dykstra's method, for computing the MAP estimate.

Example 4.3.1 (Dirichlet prior)**.** The Dirichlet distribution is a natural prior to use in the setting of contingency table estimation, as it is supported on the simplex $\Delta^{m \times n}$ of nonnegative tables whose total mass is equal to one, which are those representing valid frequency distributions. It appears in hierarchical models for the ecological inference problem [124] [159].

The Dirichlet distribution, in fact, is not quite regular: it fails assumption A4 from Section 4.2.1, and so Proposition 11 does not apply. The problem is that the negative log density (which has domain $\mathbb{R}_{++}^{m \times n}$) does not attain its global optimum – there is no finite $\pi \in \mathbb{R}_{++}^{m \times n}$ such that $\nabla Q(\pi) = \mathbf{0}$. We therefore have no starting point for the Bregman projection in Proposition 11.

The Dirichlet distribution is still tractable, however, in the following sense. For any small $\epsilon > 0$, we can find a table $\pi_\epsilon \in \operatorname{int}(\operatorname{dom} Q)$ such that $\|\nabla Q(\pi_\epsilon)\|_2 < \epsilon$. This, it turns out, is sufficient to guarantee that the Bregman projection of π_ϵ onto $\Pi(\mathbf{u}, \mathbf{v})$ is $\varepsilon-$close to the MAP solution, for an appropriate ε. This is stated formally in Proposition 12.

Proposition 12 ($\epsilon-$MAP estimation)**.** *Let* $\Pr(\pi|\mathbf{C})$ *be a prior density over tables* $\pi \in \mathbb{R}^{m \times n}$*, satisfying the regularity properties A1, A2 and A3 from* (4.3)*. Define* $Q(\pi) \triangleq -\log \Pr(\pi|\mathbf{C})$ *the negative log density, and suppose there exists* $\pi_\epsilon \in \operatorname{int}(\operatorname{dom} Q)$ *such*

that $\|\nabla Q(\pi_\epsilon)\|_2 < \epsilon$. *Let* π'_ϵ *be its Bregman projection,*

$$\pi'_\epsilon = \operatorname*{argmin}_{\pi \in \Pi(\mathbf{u},\mathbf{v})} \mathcal{D}_Q(\pi, \pi_\epsilon). \tag{4.6}$$

Then the posterior density $\Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C})$ *has a unique maximum* π_* *that satisfies*

$$Q(\pi'_\epsilon) - Q(\pi_*) < \sqrt{2}\epsilon. \tag{4.7}$$

Proof. Let π'_ϵ be the Bregman projection of π_ϵ onto $\Pi(\mathbf{u}, \mathbf{v})$ with respect to Q. $\Pi(\mathbf{u}, \mathbf{v})$ is a closed, convex set, and assumption A2 of (4.3) implies that $\Pi(\mathbf{u}, \mathbf{v}) \cap \operatorname{int}(\operatorname{dom} Q) \neq \emptyset$, so the Bregman projection is well-defined. The Bregman projection is characterized by the relation

$$\langle \pi - \pi'_\epsilon, \nabla Q(\pi_\epsilon) - \nabla Q(\pi'_\epsilon) \rangle \leq 0, \tag{4.8}$$

for all $\pi \in \Pi(\mathbf{u}, \mathbf{v}) \cap \operatorname{int}(\operatorname{dom} Q)$. From the definition of the Bregman divergence, we have that $\mathcal{D}_Q(\pi, \pi'_\epsilon) > 0$ for all $\pi \in \operatorname{int}(\operatorname{dom} Q)$, so

$$\begin{aligned} Q(\pi) - Q(\pi'_\epsilon) &> \langle \pi - \pi'_\epsilon, \nabla Q(\pi'_\epsilon) \rangle \\ &\geq \langle \pi - \pi'_\epsilon, \nabla Q(\pi_\epsilon) \rangle, \end{aligned}$$

with the second inequality deriving from (4.8). Q is strictly convex on $\operatorname{int}(\operatorname{dom} Q)$ and $\Pi(\mathbf{u}, \mathbf{v})$ is closed and convex, so Q has a unique minimum on $\Pi(\mathbf{u}, \mathbf{v}) \cap \operatorname{int}(\operatorname{dom} Q)$. Let π_* be this minimum. Inverting the previous inequality, we have

$$\begin{aligned} Q(\pi'_\epsilon) - Q(\pi_*) &< \langle \pi'_\epsilon - \pi_*, \nabla Q(\pi_\epsilon) \rangle \\ &\leq \|\pi'_\epsilon - \pi_*\|_2 \|\nabla Q(\pi_\epsilon)\|_2, \end{aligned}$$

by Cauchy-Schwarz. By assumption $\|\nabla Q(\pi_\epsilon)\|_2 < \epsilon$, while π'_ϵ and π_* both lie in the simplex $\Delta^{m \times n}$, meaning that $\|\pi'_\epsilon - \pi_*\|_2 \leq \sqrt{2}$. Combining these yields the bound (4.7). □

4.3.2 Dykstra's method of alternating projections

Casting MAP estimation as a Bregman projection (Propositions 11 and 12) suggests that we can apply efficient general methods for Bregman projections to compute the solution. In particular, we will use the Dykstra-Bregman ("Dykstra's") method [29] of alternating projections, which has been applied in the matrix balancing [144] and optimal transport [45] [17] settings to obtain fast-converging iterative solvers.

Dykstra's method [29] obtains the Bregman projections (4.5) and (4.6) onto $\Pi(\mathbf{u}, \mathbf{v})$ by decomposing the polytope into the intersection of three convex sets defined by the constraints,

$$\begin{aligned} \mathcal{C}_+ &= \mathbb{R}_+^{m \times n}, \\ \mathcal{C}_{\mathbf{u}} &= \{\pi \in \mathbb{R}^{m \times n} : \pi \mathbf{1} = \mathbf{u}\}, \\ \mathcal{C}_{\mathbf{v}} &= \{\pi \in \mathbb{R}^{m \times n} : \pi^\intercal \mathbf{1} = \mathbf{v}\}, \end{aligned} \tag{4.9}$$

such that $\Pi(\mathbf{u}, \mathbf{v}) = \mathcal{C}_+ \cap \mathcal{C}_{\mathbf{u}} \cap \mathcal{C}_{\mathbf{v}}$. The method alternates Bregman projections onto the constraints $\mathcal{C}_+$, $\mathcal{C}_{\mathbf{u}}$ and $\mathcal{C}_{\mathbf{v}}$ taken individually. In this case, a theorem of Bauschke and Lewis [16] guarantees the alternating projections converge linearly to the Bregman projection onto $\Pi(\mathbf{u}, \mathbf{v})$. Algorithm 6 gives the generic form of Dykstra's method for this problem, with $\mathrm{P}_{\mathcal{C}}^{Q}$ indicating the Bregman projection onto $\mathcal{C}$. Note that the initial table (whose projection we are computing) depends on the particular prior density used: it will be either $\nabla Q^*(\mathbf{0})$ from (4.5) or π_ϵ from (4.6).

The MAP inference problem boils down to computing Bregman projections onto the constraints individually. Bregman projections are rarely computable in closed form, but for affine constraints we will show that they have a form suitable for numerical optimization. Write the Lagrangians for the projections of a table π' onto $\mathcal{C}_{\mathbf{u}}$ and $\mathcal{C}_{\mathbf{v}}$,

$$\mathcal{L}_{\mathbf{u}}(\pi, \alpha) = Q(\pi) - \langle \nabla Q(\pi'), \pi \rangle + \alpha^\intercal \left(\pi \mathbf{1} - \mathbf{u}\right), \tag{4.10}$$

$$\mathcal{L}_{\mathbf{v}}(\pi, \beta) = Q(\pi) - \langle \nabla Q(\pi'), \pi \rangle + \beta^\intercal \left(\pi^\intercal \mathbf{1} - \mathbf{v}\right). \tag{4.11}$$

For Q which is convex of Legendre type, the gradient map $\nabla Q : \operatorname{int}(\operatorname{dom} Q) \to \operatorname{int}(\operatorname{dom} Q^*)$ is a bijection, with the gradient of the Fenchel conjugate ∇Q^* being the inverse of ∇Q. Applied to the first order conditions for (4.10) and (4.11), we get

$$\pi_{\mathbf{u}} = \mathrm{P}^{Q}_{\mathcal{C}_{\mathbf{u}}} \pi' = \nabla Q^* \left(\nabla Q(\pi') - \alpha \mathbf{1}^\intercal \right), \tag{4.12}$$

$$\pi_{\mathbf{v}} = \mathrm{P}^{Q}_{\mathcal{C}_{\mathbf{v}}} \pi' = \nabla Q^* \left(\nabla Q(\pi') - \mathbf{1} \beta^\intercal \right), \tag{4.13}$$

for $\pi_{\mathbf{u}}$ the projection of π' onto $\mathcal{C}_{\mathbf{u}}$, and $\pi_{\mathbf{v}}$ that onto $\mathcal{C}_{\mathbf{v}}$. Computing $\pi_{\mathbf{u}}$ reduces to finding α such that the original constraint holds,

$$\nabla Q^* \left(\nabla Q(\pi') - \alpha \mathbf{1}^\intercal \right) \mathbf{1} = \mathbf{u}, \tag{4.14}$$

and analogously for $\pi_{\mathbf{v}}$ and β. Dhillon and Tropp [51] suggest a method for finding α. As Q is Legendre, Q^* is strictly convex, and α satisfying (4.14) is the unique optimum for a strictly convex problem,

$$\alpha_* = \operatorname*{argmin}_{\alpha \in \mathbb{R}^m} J_{\mathbf{u}}(\alpha) = Q^* \left(\nabla Q(\pi') - \alpha \mathbf{1}^\intercal \right) + \mathbf{u}^\intercal \alpha, \tag{4.15}$$

which can be addressed by standard techniques from numerical optimization. The gradient of (4.15) exists, and the first order condition is exactly (4.14). When the Hessian of Q^* is available, for Q that is separable we have

$$\nabla^2 J_{\mathbf{u}}(\alpha) = \operatorname{diag}\left(\nabla^2 Q^* \left(\nabla Q(\pi') - \alpha \mathbf{1}^\intercal \right) \mathbf{1} \right). \tag{4.16}$$

The analogous equations hold for optimization with respect to β.

The projection onto nonnegativity constraints $\mathcal{C}_+$, for Q separable, has a simple form. Projecting π' results in

$$\pi_+ = \mathrm{P}_{\mathcal{C}_+} \max\{0, \pi'\}. \tag{4.17}$$

Algorithm 7 gives a realization of Dykstra's method as an iteration on the dual

Algorithm 6 Dykstra's method for MAP estimation

Input: $\mathbf{u} \in \Delta^m$, $\mathbf{v} \in \Delta^n$, $\pi_0 \in \operatorname{dom} Q$
$\pi \leftarrow \mathrm{P}_{\mathcal{C}_+}(\pi_0)$
repeat
  $\pi \leftarrow \mathrm{P}_{\mathcal{C}_+}\left(\mathrm{P}^Q_{\mathcal{C}_\mathbf{u}}\,\pi\right)$
  $\pi \leftarrow \mathrm{P}_{\mathcal{C}_+}\left(\mathrm{P}^Q_{\mathcal{C}_\mathbf{v}}\,\pi\right)$
until π converges

Algorithm 7 Dykstra's method for MAP estimation, dual parameterization

Input: $\mathbf{u} \in \Delta^m$, $\mathbf{v} \in \Delta^n$, $\Theta_0 \in \operatorname{dom} Q^*$
$\Theta \leftarrow \max\{\nabla Q(\mathbf{0}), \Theta_0\}$
repeat
  $\alpha_* \leftarrow \operatorname{argmin}_{\alpha \in \mathbb{R}^m} Q^*(\Theta - \alpha\mathbf{1}^\intercal) + \mathbf{u}^\intercal\alpha$
  $\Theta \leftarrow \max\{\nabla Q(\mathbf{0}), \Theta - \alpha_*\mathbf{1}^\intercal\}$
  $\beta_* \leftarrow \operatorname{argmin}_{\beta \in \mathbb{R}^n} Q^*(\Theta - \mathbf{1}\beta^\intercal) + \mathbf{v}^\intercal\beta$
  $\Theta \leftarrow \max\{\nabla Q(\mathbf{0}), \Theta - \mathbf{1}\beta_*^\intercal\}$
until Θ converges
$\pi_* \leftarrow \nabla Q^*(\Theta)$

variable $\Theta = \nabla Q(\pi)$, alternating projections onto the affine constraints $\mathcal{C}_\mathbf{u}$, $\mathcal{C}_\mathbf{v}$ with the nonnegativity constraint $\mathcal{C}_+$. Note that the projections (4.12) and (4.13) can be viewed as linearly updating the dual representation of the original table π'. We therefore need only represent Θ to compute the Dykstra iterations.

For solving (4.15), Algorithm 8 gives a Newton-Raphson method to compute the projection onto $\mathcal{C}_\mathbf{u}$. An analogous method works for $\mathcal{C}_\mathbf{v}$. Note that for some priors $\int(\operatorname{dom} Q^*)$ is a bounded subset of $\mathbb{R}^{m\times n}$, in which case backtracking can be used to ensure the bounds are respected.

Algorithm 8 Newton-Raphson method for the projection $\mathrm{P}_{\mathcal{C}_\mathbf{u}}$ onto a marginal constraint

Input: $\mathbf{u} \in \Delta^m$, $\Theta \in \operatorname{int}(\operatorname{dom} Q^*)^{m\times n}$, $\kappa > 0$
$\alpha \leftarrow \mathbf{0}$
repeat
  $\alpha \leftarrow \alpha - \kappa\left(\mathbf{u} - \nabla Q^*(\Theta - \alpha\mathbf{1}^\intercal)\,\mathbf{1}\right) \oslash \nabla^2 Q^*(\Theta - \alpha\mathbf{1}^\intercal)\,\mathbf{1}$
until α converges
$\Theta_* \leftarrow \Theta - \alpha\mathbf{1}^\intercal$

4.3.3 Complexity

The proposed approach to MAP estimation leverages fast convergence of Dykstra's method to achieve computational efficiency. Under the conditions in Section 4.2.1, alternating Bregman projections converge linearly in norm to the unique MAP solution [16]. Note that, for separable priors, the iterations have complexity $\mathcal{O}(mn)$.

Figure 4-1a shows the number of iterations to converge for the proposed method, using a Dirichlet prior. Note that the number of iterations to converge is roughly independent of the dimension of the table. We measure convergence by the norm of the difference with the converged table π_* after a large number of iterations, thresholding at $\|\pi - \pi_*\|_2 < 10^{-8}$. We observe convergence typically within 20 iterations [3].

Figure 4-1b shows the wall-clock time [4] for the proposed and existing methods, using synthetic data [5]. The proposed method is substantially faster than the existing Bayesian methods – the Dirichlet-Multinomial and King's methods – both of which rely on MCMC sampling of the posterior [6]. Goodman's regression (which essentially computes a linear regression) is the fastest by an order of magnitude.

4.3.4 Tertiary or higher-order relationships

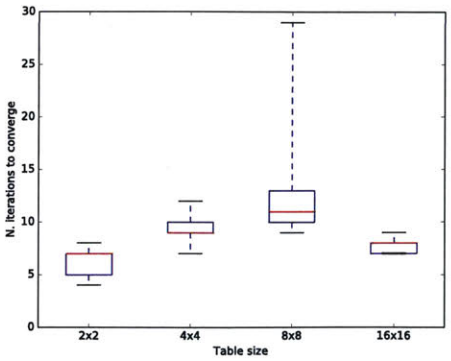
The inference problem we consider naturally generalizes to multidimensional tables relating more than two marginals. In the example of voting prefences, we might want to relate several types of aggregate demographic information (such as gender

[3] We conduct 1000 trials, in each trial sampling a table uniformly from $\Delta^{m\times n}$ and using its marginals as inputs for the proposed MAP inference method, using a Dirichlet prior with parameters $\mathbf{C}$ chosen uniformly from $[0, 10]^{m\times n}$. Figure 4-1a shows percentiles $2.5, 25, 50, 75, 97.5$ for the number of iterations.

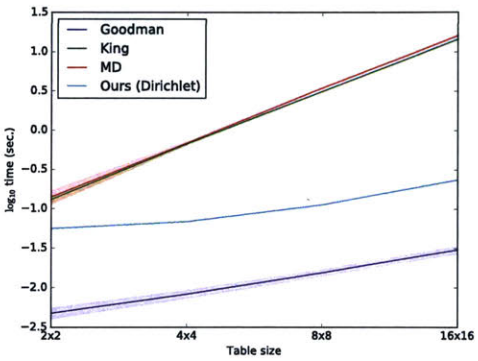
[4] Wall-clock time measured on a MacBook Pro with a single 2.9 GHz Intel Core i7 processor. The proposed method was implemented in MATLAB, while the Goodman's, King's and Dirichlet-Multinomial methods were taken from the `ei` and `eiPack` packages for R.

[5] We conduct 100 trials in which we sample 100 tables from a Dirichlet prior having parameters $\mathbf{C}$ chosen uniformly at random from $[0, 10]^{m\times n}$. Figure 4-1b shows the mean time ($\pm$ one standard deviation) to infer these 100 tables from their marginals. For our method we assume a Dirichlet prior with uninformative parameters ($\mathbf{C} \propto \mathbf{1}\mathbf{1}^\intercal$).

[6] For MCMC, the computation time depends on the total number of samples. In both cases, we use a burnin period of 1000 samples and no thinning of the remaining 1000 samples, as increasing these parameters had no impact on the performance of the posterior mean estimator on the given data.



(a) Convergence of Dykstra's method for MAP estimation.



(b) Wall-clock time.

and race) to vote counts, necessitating that we infer a multidimensional table each of whose one-dimensional marginals matches a particular type of aggregate data.

The MAP estimation method outlined in this section extends straightforwardly to the case of more than two marginals. Noting that each step of the Dykstra's method (Algorithm 6) is a projection onto one of the marginal constraints followed by projection onto the nonnegative orthant, we extend naturally by doing so while cycling through all of the marginal constraints in order. Let $\{\mathbf{u}_k\}_{k=1}^K$ be a be a set of K marginals given as input data. Each marginal associates to an affine constraint $\mathcal{C}_{\mathbf{u}_k} = \{\pi \in \mathbb{R}^{m_1 \times \cdots \times m_K} : \sum_{j_\ell, \ell \neq k} \pi_{j_1,\ldots,j_k,\ldots,j_K} = (\mathbf{u}_k)_{j_k} \forall j_k\}$. With these K constraints, the alternating projection in Algorithm 6 becomes:

repeat

$k \leftarrow (k+1) \mod K+1$

$\pi \leftarrow \mathrm{P}_{\mathcal{C}_+}\left(\mathrm{P}^{Q}_{\mathcal{C}_{\mathbf{u}_k}} \pi\right)$

until π converges

This straightforward extension preserves the efficiency of the two-marginal case, converging linearly to the MAP table [16].

4.4 Estimating the prior

The prior distribution in (4.4) depends on a set of parameters $\mathbf{C}$, which encode prior preferences for certain tables π over others. In the case of an ecological inference about voting preferences, $\mathbf{C}$ might encode the intrinsic tendency for certain groups of people to vote for certain political parties, which then informs our inference for specific observed census and vote counts. This prior knowledge can derive from a number of sources, depending on the availability of side information extrinsic to the inference problem.

4.4.1 Estimation with fully-observed tables

The richest side information we can use would be fully-observed tables sampled from a shared prior distribution (as in the model (4.4)), or a proxy thereof. In the case of voting preferences, for example, the individual-level data we care about is the proportion of voters from each group that vote for each political party. A good proxy might be the corresponding proportions of voter registrations for each party.

Regardless the source, given a set of one or more observed tables $\xi^{(j)} \in \Delta^{m \times n}$, we can estimate the parameters $\mathbf{C}$ by maximum likelihood. For a prior $\Pr(\xi|\mathbf{C})$, this is $\mathbf{C}_*$ satisfying

$$\mathbf{C}_* = \operatorname*{argmin}_{\mathbf{C} \in \operatorname{dom} f} \sum_{j=1}^{N} -\log \Pr(\xi^{(j)}|\mathbf{C}).$$

A negative log density that is convex of Legendre type is smooth with a unique minimum in $\operatorname{int}(\operatorname{dom} f)$, and we can optimize by standard methods.

4.4.2 Estimation with polling data

We can also imagine sampling individuals from the population we're studying and recording their individual-level data. In the voting preference example, this would

mean polling individual voters and recording both their demographic information and their vote.

For this type of polling data, we can extend our existing model (4.4) to include a multinomial likelihood, which describes the distribution of counts for the different individual-level categories, conditional on the particular table π of proportions. More formally, for a single population this is

$$\begin{aligned} \pi &\sim \mathcal{P}(\mathbf{C}), \\ \mathbf{z} &\sim \text{Multinomial}(\pi, N), \end{aligned} \tag{4.18}$$

with N the number of sampled individuals and $\mathbf{z} \in \mathbb{Z}_+^{m \times n}$ the matrix of counts, drawn according to the proportions in table π. In order to estimate the prior parameters $\mathbf{C}$, we optimize the negative log of the marginal likelihood,

$$-\log \Pr(\mathbf{z}|\mathbf{C}) = -\log \int_{\Delta^{m \times n}} \Pr(\mathbf{z}|\pi) \Pr(\pi|\mathbf{C}) d\pi. \tag{4.19}$$

For most families of priors $\mathcal{P}$, the marginal likelihood is not available in closed form. We can either apply a method of numerical integration or attempt to optimize an upper bound on (4.19).

Example 4.4.1 (Dirichlet prior)**.** In the case of the Dirichlet prior, the marginal likelihood has a closed form. This is

$$\Pr(\mathbf{z}|\mathbf{C}) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{i=1}^{m} \prod_{j=1}^{n} \frac{\Gamma(\mathbf{z}_{ij} + \mathbf{C}_{ij})}{\Gamma(\mathbf{C}_{ij})}, \tag{4.20}$$

with $\alpha = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{C}_{ij}$. We can optimize the negative log directly, by standard methods.

4.5 Interval estimation

4.5.1 Credible region

For interval estimation, we compute the highest posterior density (HPD) credible region. Given observed marginals $\mathbf{u} \in \Delta^m$ and $\mathbf{v} \in \Delta^n$, the HPD α-credible region for the underlying table is a subset of the polytope $\Pi(\mathbf{u}, \mathbf{v})$ of valid tables, having total posterior probability mass α, and containing the tables π with the largest posterior density $\Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C})$.

An HPD α-credible region is bounded by an isocontour of the posterior density and can therefore be represented by the corresponding density value. As this is a single parameter, the problem of computing the credibility region reduces to a search over possible bounding density values, to find the value whose bounded region contains the desired fraction α of the probability mass. We need only compute this fraction for each density value.

We propose a Monte Carlo approximation for the fraction of mass in the bounded region. First, note that the posterior density can be expressed in terms of the restriction of the prior $\Pr(\pi|\mathbf{C})$ to the polytope $\Pi(\mathbf{u}, \mathbf{v})$, suitably normalized,

$$\Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C}) = Z(\mathbf{u}, \mathbf{v})^{-1} \Pr(\pi|\mathbf{C}) I_{\Pi(\mathbf{u},\mathbf{v})}(\pi), \tag{4.21}$$

with $Z(\mathbf{u}, \mathbf{v})$ the normalizer,

$$Z(\mathbf{u}, \mathbf{v}) = \int_{\Pi(\mathbf{u},\mathbf{v})} \Pr(\pi|\mathbf{C}). \tag{4.22}$$

Define the credible region $\mathcal{R}_\gamma = \{\pi \in \Pi(\mathbf{u}, \mathbf{v}) : \Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C}) > \gamma\}$, for threshold density γ. Substituting the expressions (??) and (??), the total posterior mass in $\mathcal{R}_\gamma$ can be expressed

$$\int_{\mathcal{R}_\gamma} \Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C}) = \frac{\int_{\mathcal{R}_\gamma} \Pr(\pi|\mathbf{C})}{\int_{\Pi(\mathbf{u},\mathbf{v})} \Pr(\pi|\mathbf{C})}. \tag{4.23}$$

Suppose we have a uniform sample $\{\pi^{(k)}\}_{k=1}^N$ from $\Pi(\mathbf{u}, \mathbf{v})$. Then we can approximate

Algorithm 9 Credible region estimation

Input: $\mathbf{u} \in \Delta^m$, $\mathbf{v} \in \Delta^n$, $\mathbf{C} \in \mathbb{R}_+^{m \times n}$, $N \in \mathbb{Z}_+$, $\beta > 0$, $\alpha \in [0, 1]$
Sample: $\pi^{(k)} \sim \text{Unif}(\Pi(\mathbf{u}, \mathbf{v}))$, $1 \leq k \leq N$
$\tau_{\min} \leftarrow 0$, $\tau_{\max} \leftarrow \max_k \Pr(\pi^{(k)}|\mathbf{C})$
repeat
  $\tau \leftarrow \frac{\tau_{\min} + \tau_{\max}}{2}$
  $I_\tau \leftarrow \{k : \Pr(\pi^{(k)}|\mathbf{C}) > \tau\}$
  $\alpha_{\text{est}} \leftarrow \frac{\sum_{\ell \in I_\tau} \Pr(\pi^{(\ell)}|\mathbf{C})}{\sum_{k=1}^N \Pr(\pi^{(k)}|\mathbf{C})}$
  $(\tau_{\min}, \tau_{\max}) \leftarrow \begin{cases} (\tau_{\min}, \tau) & \alpha_{\text{est}} \leq \alpha \\ (\tau, \tau_{\max}) & \alpha_{\text{est}} > \alpha \end{cases}$
until $\tau_{\max} - \tau_{\min} < \beta$
Output: $\{\pi^{(k)} : \Pr(\pi^{(k)}|\mathbf{C}) > \tau\}$

the integrals in (4.23) by

$$\int_{\mathcal{R}_\gamma} \Pr(\pi|\mathbf{C}) \approx \frac{\text{vol}(\Pi(\mathbf{u}, \mathbf{v}))}{N} \sum_{\ell=1}^N \Pr(\pi^{(\ell)}|\mathbf{C}) I_{\mathcal{R}_\gamma}(\pi^{(\ell)}), \tag{4.24}$$

$$\int_{\Pi(\mathbf{u},\mathbf{v})} \Pr(\pi|\mathbf{C}) \approx \frac{\text{vol}(\Pi(\mathbf{u}, \mathbf{v}))}{N} \sum_{k=1}^N \Pr(\pi^{(k)}|\mathbf{C}), \tag{4.25}$$

with $\text{vol}(\Pi(\mathbf{u}, \mathbf{v}))$ the volume of the polytope. The standard estimator for (4.23) is then the ratio

$$\int_{\mathcal{R}_\gamma} \Pr(\pi|\mathbf{u}, \mathbf{v}, \mathbf{C}) \approx \frac{\sum_{\ell=1}^N \Pr(\pi^{(\ell)}|\mathbf{C}) I_{\mathcal{R}_\gamma}(\pi^{(\ell)})}{\sum_{k=1}^N \Pr(\pi^{(k)}|\mathbf{C})}. \tag{4.26}$$

For large n the bias should be neglible, going as $\mathcal{O}(\frac{1}{n})$. Algorithm 9 gives the resulting algorithm for computing the HPD α-credible region, which uses the estimated total mass (4.26) and does a binary search for a threshold τ. Note that the threshold used in the algorithm is equivalent to a posterior density threshold of $\gamma = Z(\mathbf{u}, \mathbf{v})^{-1}\tau$. Note also that we rely on a uniform sample from the polytope $\Pi(\mathbf{u}, \mathbf{v})$, which is computed as described in the following section.

4.5.2 Generating uniform samples from $\Pi(\mathbf{u}, \mathbf{v})$

Our algorithm for computing the credible region relies on uniform sampling of the polytope $\Pi(\mathbf{u}, \mathbf{v})$. Uniform sampling of convex polytopes is hard in general, with the

best known algorithms scaling as $\mathcal{O}(n^3)$ per sample [104]. Of the existing methods, the "hit and run" algorithm first proposed by Smith [145] is thought to be fastest in practice [42]. Here we propose a hit and run sampler whose distribution converges to the uniform distribution on $\Pi(\mathbf{u}, \mathbf{v})$.

The standard hit and run algorithm is a Markov Chain Monte Carlo algorithm that samples a sequence of points: for each sampled point the next point is chosen by sampling a uniform random direction and distance to move from the previous point, while staying within the polytope being sampled.

We note that, given a table $\pi_* \in \Pi(\mathbf{u}, \mathbf{v})$ (taken as a vector in $\mathbb{R}^{mn}$ and a uniform random direction $\mathbf{z} \in \mathcal{S}^{mn-1}$, the table $\pi_* + \lambda \mathbf{z}$ for $\lambda \in \mathbb{R}_+$ is almost surely not in $\Pi(\mathbf{u}, \mathbf{v})$. This is because $\Pi(\mathbf{u}, \mathbf{v})$ is a lower-dimensional affine subset of $\mathbb{R}^{mn}$. Our hit and run sampler therefore has to restrict the directions sampled.

We decompose the polytope $\Pi(\mathbf{u}, \mathbf{v})$ into the intersection of an affine set defined by the marginal constraints, $\mathcal{C}_{\mathbf{u},\mathbf{v}} = \{\pi \in \mathbb{R}^{m \times n} : \pi \mathbf{1} = \mathbf{u}, \pi^\intercal \mathbf{1} = \mathbf{v}\}$, with the nonnegative orthant $\mathbb{R}_+^{m \times n}$. If we write the marginal constraints in matrix form, as $\mathbf{A} \operatorname{vec}(\pi) = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$, with each row of $\mathbf{A}$ encoding one of the row or column constraints, we have that the valid directions for updating a point $\pi_* \in \Pi(\mathbf{u}, \mathbf{v})$, while staying in the polytope, are exactly those lying in the nullspace of $\mathbf{A}$: the directions $\mathbf{z} \in \ker \mathbf{A}$ are exactly those that guarantee $\pi_* + \lambda \mathbf{z} \in \mathcal{C}_{\mathbf{u},\mathbf{v}}$, for any $\lambda \in \mathbb{R}_+$. Our sampler therefore needs only to choose λ such that $\pi_* + \lambda \mathbf{z}$ lies in the nonnegative orthant. The bounds for such a λ are easy to compute: these are $\lambda \in [0, \min_{\mathbf{z}_k < 0} \frac{\operatorname{vec}(\pi)_k}{-\mathbf{z}_k}]$. Algorithm 10 gives the resulting hit and run sampler, which converges to a uniform sample of $\Pi(\mathbf{u}, \mathbf{v})$.

Figure 4-2 shows the autocorrelation [7] of the sequence produced by our hit and run sampler, given randomly generated input marginals. Note first that the sample sequence mixes, with the autocorrelation going to zero. For 2×2 tables the sampler mixes rapidly, with negligible autocorrelation beyond a lag of 10 samples. The mixing time increases substantially for larger tables.

[7] We measure autocorrelation of a sequence $\mathbf{x}(t)$ by the mean cosine of the angle between the centered sequence $\bar{\mathbf{x}}(t) = \mathbf{x}(t) - \frac{1}{K} \sum_{k=1}^{K} \mathbf{x}(t_k)$ and its τ-shifted version $\bar{\mathbf{x}}(t - \tau)$. Shaded regions are $\pm$ one standard error over 100 replicates.

Algorithm 10 Hit and run sampler for Unif $(\Pi(\mathbf{u}, \mathbf{v}))$

Input: $\pi^{(0)} \in \Pi(\mathbf{u}, \mathbf{v})$, $K \in \mathbb{Z}_+$
$\mathbf{A} \leftarrow \begin{bmatrix} I_m \otimes \mathbf{1}_n^\intercal \\ \mathbf{1}_m^\intercal \otimes I_n \end{bmatrix}$, $k \leftarrow 0$
$\{\mathbf{e}_j\}_{j=1}^d \leftarrow$ orthonormal basis for $\ker \mathbf{A}$
repeat
  $\alpha \sim \text{Unif}(\mathcal{S}^{d-1})$
  $\mathbf{z} \leftarrow \sum_{j=1}^d \alpha_j \mathbf{e}_j$
  $\lambda_{\max} \leftarrow \min_{\mathbf{z}_\ell < 0} \frac{\text{vec}(\pi^{(k)})_\ell}{-\mathbf{z}_\ell}$
  $\lambda \sim \text{Unif}([0, \lambda_{\max}])$
  $\text{vec}(\pi^{(k+1)}) \leftarrow \text{vec}(\pi^{(k)}) + \lambda \mathbf{z}$
  $k \leftarrow k + 1$
until $k > K$
Output: $\{\pi^{(k)}\}_{k=1}^K$

4.6 Empirical

4.6.1 Estimating the prior: synthetic data

We investigate three different settings of the ecological inference problem, using synthetic data. In all settings we first estimate the prior parameters (using one of the methods from Section 4.4) before using these parameters to perform inference. We assume a Dirichlet prior in all cases, and sample the data from this prior. In the first setting, we assume that there is a collection of problem instances that share a single prior (this is the model in Section 4.2), and that we get to observe fully the tables in a subset of these instances. The task is to estimate the prior and use it to infer the tables in the remaining instances.

In the second setting, we have a single problem instance, meaning a single table that we wish to infer, and we are able to poll N individuals from the underlying population and determine their group memberships. The task is then to use this polling data to estimate the prior, which is used to infer the table. (This is the model in Section 4.4.2.)

In the third setting, we again have a collection of problem instances that share a single prior, but now we assume we will poll N individuals from the total combined

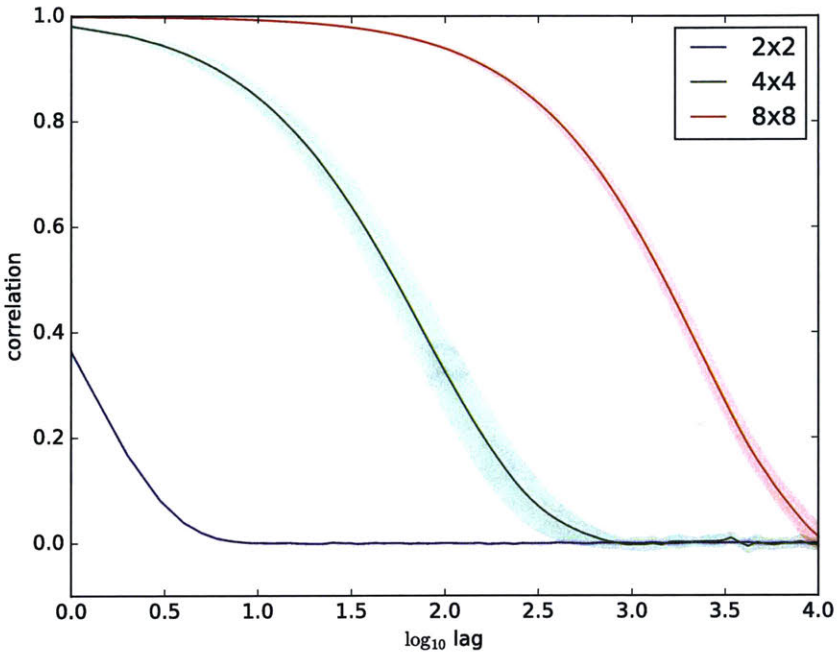


Figure 4-2: Autocorrelation of hit and run sampler for $\Pi(\mathbf{u}, \mathbf{v})$.

populations underlying these instances [8]. This polling data is used to estimate a single prior, which we then use to infer all of the tables. This setting simulates, for example, conducting a statewide poll to estimate a common prior for the voting behavior in many counties within the state.

Figure 4-3a shows the results for the first setting. We compare MAP inference with the maximum likelihood prior (Section 4.4.1) against two baselines. The first is the neighborhood model (Section 4.1.2), which is straightforward, requires no parameters to be estimated, and reflects an inaccurate prior. The second baseline is MAP inference with the true prior, which represents the best possible performance. Most importantly, we see that even a single observed table yields a significantly more accurate inference than with the neighborhood model, while performance converges to that of the true prior with increasing number of observed tables.

Figure 4-3b shows the results for the second setting. We compare MAP inference

[8] We assume uniform random ratios of population sizes between the instances.

with the maximum marginal likelihood prior (Section 4.4.2) against three baselines. The first two are the neighborhood model and the true prior, as before. The third baseline estimates the table by simply taking the proportions directly from the polling data. Given that it incorporates extra constraints (from the observed marginals), we expect MAP inference with a sufficiently good prior to outperform the direct method (the last baseline). This is indeed the case: the prior we estimate by maximum marginal likelihood is good enough, yielding a more accurate inferred table for all numbers of polled individuals.

Figure 4-3c shows the results for the third setting. As in the second setting, we compare MAP inference with the maximum marginal likelihood prior (Section 4.4.2) against three baselines. The first two are the neighborhood model and the true prior, as before, while the third baseline estimates the table by simply taking the proportions directly from the polling data. We make two observations. First, the directly estimated proportions, which approximate the average table over the combined populations underlying the instances, converge to a worse error than is achieved by MAP inference using the true prior. In other words, we can do better via MAP inference than guessing the average table. And second, the prior estimated from polling data, despite deriving from the combined population, achieves an error equivalent to that using the true prior.

In all three settings, we find that it is beneficial to use the methods of Section 4.4 to estimate the prior, before performing inference. In a real world setting, of course, applicability of these methods depends on the availability of side information (such as polling data).

4.6.2 Florida election

We examine a dataset of real voter registration data from the 2012 US presidential election, in Florida [78]. The data contains both demographic information and party registrations for roughly 10 million individual voters from 68 Florida counties. This provides us with real ground truth data for validating the methodology proposed here, and has been used for the same purpose elsewhere [61] [78] [112].

Given the total demographic proportions and the total party registrations within each county, the task is to infer the joint distribution of demographics and party registrations. We apply the proposed MAP inference method (Section 4.3), using a Dirichlet prior with two different sets of parameters. The first specifies an "uninformative" prior [9], which is nearly uniform over the simplex of tables. The second "feature" prior is adapted from [112]: we define a feature vector for each demographic group and for each political party, by computing the average age, gender, and vote in 2008 within the group, then use these feature vectors to define a similarity score between pairs of groups. Specifically, if $\mathbf{x}^{(i)}$ is the feature vector for the ith demographic group and $\mathbf{y}^{(j)}$ is the feature vector for the jth political party, we define the similarity $\mathbf{C}_{ij}$ by

$$\mathbf{C}_{ij} = \exp\left(-2\|\mathbf{x}^{(i)} - \mathbf{y}^{(j)}\|_2\right). \tag{4.27}$$

$\mathbf{C}$ is then used as the Dirichlet prior parameter for MAP inference.

As baselines, we compare against four major existing methods: the neighborhood model, Goodman's regression [72], King's method [87], and the Dirichlet-multinomial model of Rosen [124]. We additionally include two of the optimal transport-based models described in Muzellec et al. [112]. For the optimal transport methods, we use the suggested ground metric values from [112], given by

$$\mathbf{M}_{ij} = \sqrt{2 - 2\exp(-5\|\mathbf{x}^{(i)} - \mathbf{y}^{(j)}\|_2)}, \tag{4.28}$$

with $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(j)}$ the same feature vectors as above.

Table 4.1 shows the results, in terms of accuracy with respect to the ground truth tables [10]. Interestingly, the best performing inference method is also the simplest, being the neighborhood model – this indicates that independence of the two measurements (demographic and political party) is not an unreasonable assumption in this dataset. The next best performing is the MAP inference method proposed here, with the prior parameters making almost no difference in the performance. The remaining

[9] The parameters are $\mathbf{C}_{ij} = 1$ for all i, j.

[10] For methods that aren't guaranteed to return true probability tables, we report the generalized KL divergence.

baseline methods perform somewhat worse in terms of accuracy, with Goodman's regression the worst in terms of absolute error and optimal transport the worst in terms of KL divergence.

We additionally examine the impact of polling data on the inference problem. We assume that we are able to conduct a statewide poll, in which we sample N members of the population and assess both their demographic group and their party registration. We use the method of Section 4.4.2 to estimate the maximum marginal likelihood prior (assuming a Dirichlet prior), and use the estimated statewide prior to perform MAP inference of the countywide tables.

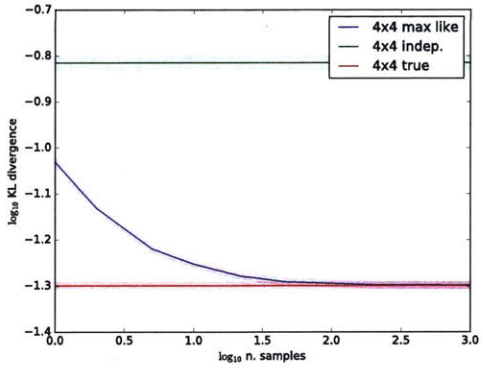
Figure 4-4 shows the result. In Figure 4-4a we see the mean absolute error of the inferred tables with respect to ground truth, and in Figure 4-4b the mean KL divergence. For comparison, we show performance for both the best performing baseline (the neighborhood model) and the Dirichlet model with a "uniform" prior. We make two observations. First, the model of a common prior amongst the many instances (Section 4.2) is a useful one on this real dataset, as MAP inference in this model, using the performing estimated prior, significantly outperforms the best baseline method in terms of accuracy. And second, it only requires 100 poll respondents for the estimated prior to outperform the best baseline in terms of absolute error (1000 respondents for KL).

Table 4.1: Accuracy of inferred tables for existing and proposed methods, Florida election data ($N = 68$).

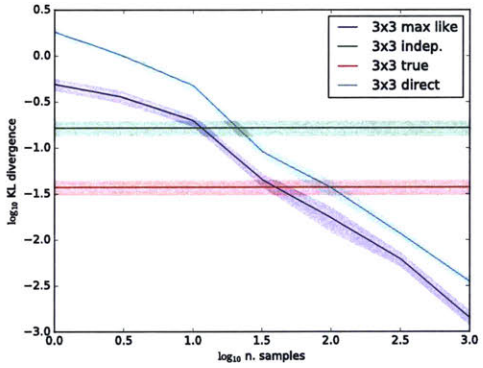
Method	Absolute error	KL divergence
Goodman's regression	0.0301 ± 0.0101	0.208 ± 0.090
Multinomial-Dirichlet	0.0205 ± 0.0594	0.178 ± 0.101
King's method	0.0149 ± 0.0076	0.107 ± 0.095
Optimal transport	0.0211 ± 0.0061	4.83 ± 1.93
Entropy-regularized OT	0.0234 ± 0.0023	0.349 ± 0.089
Neighborhood	$\mathbf{0.0076 \pm 0.0037}$	$\mathbf{0.0492 \pm 0.0287}$
Ours (uniform prior)	0.0100 ± 0.0047	0.0688 ± 0.0399
Ours (feature prior)	0.0101 ± 0.0044	0.0663 ± 0.0362

4.7 Conclusion

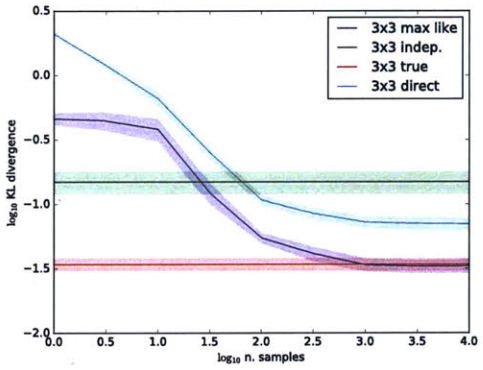
We described a novel method for ecological inference – the determination of individual behaviors from aggregate measurements. Our method flexibly incorporates a variety of prior distributions and admits an efficient fixed point iteration for computing the most probable solution. We demonstrate that, with a suitable prior, our method is more accurate than existing methods for ecological inference. We additionally propose a method for interval estimation. Ecological inference is a common sticking point when analyzing data in social sciences and elsewhere, and we hope that our proposed method can enable more transparent specification of priors and more accurate inferences.



(a) Performance of MAP estimate using max. likelihood parameters from fully-observed tables.

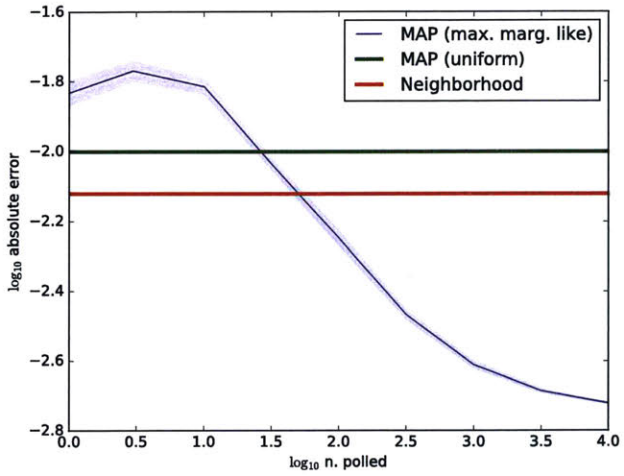


(b) Performance of MAP estimate using max. marginal likelihood parameters from polling data for a single table.

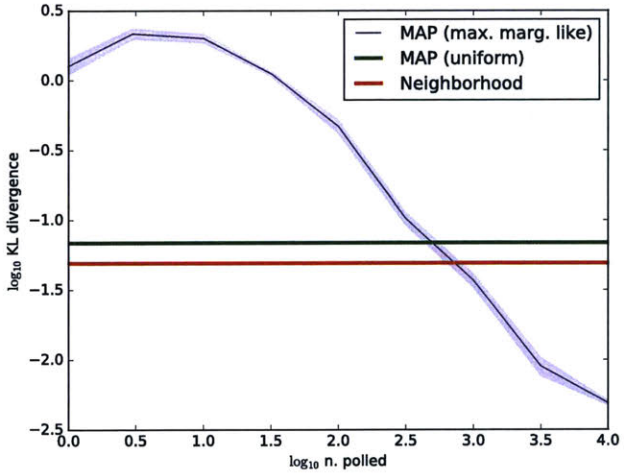


(c) Performance of MAP estimate using max. marginal likelihood parameters from polling data combining many tables.

Figure 4-3: Performance of prior estimation methods.



(a) Performance (absolute error) of MAP estimate using max. marginal likelihood parameters from statewide polling data.



(b) Performance (KL divergence) of MAP estimate using max. marginal likelihood parameters from statewide polling data.

Figure 4-4: Performance (Florida election data) of prior estimation from polling data.

Bibliography

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Network flows: theory, algorithms, and applications. 1993.

[3] Yacine Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906–937, April 2008.

[4] Juha Ala-Luhtala, Simo Särkkä, and Robert Piché. Gaussian filtering and variational approximations for Bayesian smoothing in continuous-discrete stochastic dynamic systems. *Signal Processing*, 2015.

[5] Daniel Alspach and Harold Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, 17(4):439–448, 1972.

[6] Mauricio A. Alvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2011.

[7] Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete wasserstein barycenters: optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, 2016.

[8] Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor. Variational Inference for Diffusion Processes. *NIPS*, 2007.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[10] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, March 2003.

[11] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006.

[12] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum kantorovich distance estimators. *Stat. Probab. Lett.*, 76(12):1298–1302, 1 July 2006.

[13] Saurav Basu, Soheil Kolouri, and Gustavo K Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014.

[14] Marcus Baum, K Peter, D Uwe, et al. On wasserstein barycenters and mmospa estimation. *IEEE Signal Processing Letters*, 22(10):1511–1515, 2015.

[15] Heinz H Bauschke, Jonathan M Borwein, et al. Legendre functions and the method of random bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

[16] Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[17] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[18] Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the monge–ampère operator. *Numerische Mathematik*, 134(3):611–636, 2016.

[19] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Inference in generative models using the wasserstein distance. *arXiv preprint arXiv:1701.05146*, 2017.

[20] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[21] Dimitris Bertsimas, John N. Tsitsiklis, and John Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, Boston, third printing edition, 1997.

[22] Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O Roberts. A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, 10(1):85–104, 2008.

[23] Alexandros Beskos, Gareth O Roberts, et al. Exact simulation of diffusions. *The Annals of Applied Probability*, 15(4):2422–2444, 2005.

[24] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo López, et al. Geodesic pca in the wasserstein space by convex pca. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.

[25] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[26] Emmanuel Boissard, Thibaut Le Gouic, Jean-Michel Loubes, et al. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[27] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

[28] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[29] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[30] Robert Grover Brown and Patrick Y. C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering.* John Wiley and Sons, 1997.

[31] Chris J Budd, MJP Cullen, and EJ Walsh. Monge–ampére based moving mesh methods for numerical weather prediction, with applications to the eady problem. *Journal of Computational Physics*, 236:247–270, 2013.

[32] Martin Burger, José Antonio Carrillo de la Plata, and Marie-Therese Wolfram. A mixed finite element method for nonlinear diffusion equations. 2009.

[33] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of Entropic Schemes for Optimal Transport and Gradient Flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, April 2017.

[34] José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258, 2015.

[35] José A Carrillo and J Salvador Moll. Numerical simulation of diffusive and aggregation phenomena in nonlinear continuity equations by evolving diffeomorphisms. *SIAM Journal on Scientific Computing*, 31(6):4305–4329, 2009.

[36] Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, and Nicolas Papadakis. Log-pca versus geodesic pca of histograms in the wasserstein space. *arXiv preprint arXiv:1708.08143*, 2017.

[37] JS Chang and G Cooper. A practical difference scheme for fokker-planck equations. *Journal of Computational Physics*, 6(1):1–16, 1970.

[38] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[39] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced Optimal Transport: Geometry and Kantorovich Formulation. *arXiv.org*, August 2015.

[40] Michael H Coen, M Hidayath Ansari, and Nathanael Fillmore. Comparing Clusterings in Space. *ICML*, pages 231–238, 2010.

[41] Robert T Collins and Weina Ge. Csdd features: Center-surround distribution distance for feature extraction and matching. In *European Conference on Computer Vision*, pages 140–153. Springer, 2008.

[42] Ben Cousins and Santosh Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016.

[43] Dan Crisan and Terry Lyons. A particle approximation of the solution of the kushner–stratonovitch equation. *Probability Theory and Related Fields*, 115(4):549–578, 1999.

[44] Marco Cuturi. Permanents, transportation polytopes and positive definite kernels on histograms. In *International Joint Conference on Artificial Intelligence, IJCAI*, volume 1, 2007.

[45] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[46] Marco Cuturi and David Avis. Ground metric learning. *Journal of Machine Learning Research*, 15(1):533–564, 2014.

[47] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[48] Marco Cuturi and Gabriel Peyré. A Smoothed Dual Approach for Variational Wasserstein Problems. *SIAM J. Imaging Sci.*, 9(1):320–343, 2016.

[49] Mohammad Reza Daliri. Kernel earth mover's distance for eeg classification. *Clinical EEG and neuroscience*, 44(3):182–187, 2013.

[50] Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *arXiv preprint arXiv:1610.06447*, 2016.

[51] Inderjit S Dhillon and Joel A Tropp. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.

[52] Persi Diaconis and Bradley Efron. Testing for independence in a two-way table: new interpretations of the chi-square statistic. *The Annals of Statistics*, pages 845–874, 1985.

[53] Yulia Dodonova, Mikhail Belyaev, Anna Tkachev, Dmitry Petrov, and Leonid Zhukov. Kernel classification of connectomes based on earth mover's distance between graph spectra. *arXiv preprint arXiv:1611.08812*, 2016.

[54] Garland B Durham and A Ronald Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338, 2002.

[55] Herbert Edelsbrunner and Dmitriy Morozov. Persistent homology: Theory and practice. In *Proceedings of the European Congress of Mathematics*, 2012.

[56] Matthias Erbar. The heat equation on manifolds as a gradient flow in the wasserstein space. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 46, pages 1–23. Institut Henri Poincaré, 2010.

[57] Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O Roberts. Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):755–777, 2008.

[58] Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3032–3042, 2015.

[59] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[60] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *arXiv preprint arXiv:1608.08063*, 2016.

[61] Seth R Flaxman, Yu-Xiang Wang, and Alexander J Smola. Who supported obama in 2012?: Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298. ACM, 2015.

[62] Lester Randolph Ford Jr and Delbert Ray Fulkerson. *Flows in networks*. Princeton university press, 2015.

[63] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[64] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[65] Alfred Galichon and Bernard Salanié. Matching with trade-offs: Revealed preferences over competing characteristics. *Preprint SSRN-1487307*, 2010.

[66] Andrew Gardner, Christian A Duncan, Jinko Kanno, and Rastko R Selmic. On the definiteness of earth mover's distance and its relation to set intersection. *IEEE Transactions on Cybernetics*, 2017.

[67] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[68] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Gan and vae from an optimal transport point of view. *arXiv preprint arXiv:1706.01807*, 2017.

[69] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Sinkhorn-autodiff: Tractable wasserstein learning of generative models. *arXiv preprint arXiv:1706.00292*, 2017.

[70] Clark R. Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):231–240, 1984.

[71] Andrew Golightly and Darren J Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693, 2008.

[72] Leo Goodman. Ecological regressions and the behavior of individuals. *American Sociological Review*, 18:663–665, 1953.

[73] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

[74] Kristen Grauman and Trevor Darrell. Fast contour matching using approximate earth mover's distance. In *CVPR*, 2004.

[75] Leslie Greengard and John Strain. The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94, 1991.

[76] Bernard Grofman, Michael Migalski, and Nicholas Noviello. The "totality of circumstances test" in section 2 of the 1982 extension of the voting rights act: A social science perspective. *Law & Policy*, 7(2):199–223, 1985.

[77] A Stan Hurn, Kenneth A Lindsay, and Vance L Martin. On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations. *Journal of Time Series Analysis*, 24(1):45–63, 2003.

[78] Kosuke Imai and Kabir Khanna. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272, 2016.

[79] Piotr Indyk and Nitin Thaper. Fast Image Retrieval via Embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision.* ICCV, 2003.

[80] Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms.* Kluwer Academic Publishers, 2002.

[81] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, January 1998.

[82] Simon Julier, Jeffrey Uhlmann, and Hugh F Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on automatic control*, 45(3):477–482, 2000.

[83] Simon J Julier, Jeffrey K Uhlmann, and Hugh F Durrant-Whyte. A new approach for filtering nonlinear systems. In *American Control Conference, Proceedings of the 1995*, volume 3, pages 1628–1632. IEEE, 1995.

[84] Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1):95–108, 1961.

[85] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk SSSR*, volume 37, pages 199–201, 1942.

[86] Leonid Vitalievich Kantorovich. On a problem of monge. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 3, pages 225–226, 1948.

[87] Gary King. *A solution to the ecological inference problem.* Princeton, NJ: Princeton University Press, 1997.

[88] Paul Kleppner. *Chicago divided: The making of a black mayor.* northern illinois University Press, 1985.

[89] Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations.* Springer Science & Business Media, April 2013.

[90] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[91] Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):drs019–1047, October 2012.

[92] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30(2-3):271–274, 1998.

[93] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

[94] J Morgan Kousser. Ecological regression and the analysis of past politics. *The Journal of Interdisciplinary History*, 4(2):237–262, 1973.

[95] Harold Kushner. Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control*, 12(5):546–556, 1967.

[96] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[97] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

[98] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics. Springer Berlin Heidelberg, 2011.

[99] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[100] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific, 2001.

[101] Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 29(5):840–853, 2007.

[102] James W Loewen. Social science in the courtroom: Statistical techniques and research methods for winning class-action suits. 1982.

[103] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[104] László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

[105] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[106] Qin Lv, Moses Charikar, and Kai Li. Image similarity search with compact data structures. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 208–217. ACM, 2004.

[107] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

[108] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[109] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[110] Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 3718–3726, 2016.

[111] Oleg Museyko, Michael Stiglmayr, Kathrin Klamroth, and Günter Leugering. On the application of the monge–kantorovich problem to image registration. *SIAM Journal on Imaging Sciences*, 2(4):1068–1097, 2009.

[112] Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *AAAI*, pages 2387–2393, 2017.

[113] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[114] Bernt Oksendal. *Stochastic Differential Equations*. An Introduction with Applications. Springer Science & Business Media, April 2013.

[115] John A Ozolek, Akif Burak Tosun, Wei Wang, Cheng Chen, Soheil Kolouri, Saurav Basu, Hu Huang, and Gustavo K Rohde. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Medical image analysis*, 18(5):772–780, 2014.

[116] Lorenzo Pareschi and Mattia Zanella. Structure preserving schemes for nonlinear fokker-planck equations and applications. *arXiv preprint arXiv:1702.00088*, 2017.

[117] Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. *Computer Vision–ECCV 2008*, pages 495–508, 2008.

[118] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *Computer vision, 2009 IEEE 12th international conference on*, pages 460–467. IEEE, 2009.

[119] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems*, pages 4197–4205, 2016.

[120] Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[121] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[122] Gareth O Roberts and Osnat Stramer. On inference for partially observed nonlinear diffusion models using the metropolis–hastings algorithm. *Biometrika*, 88(3):603–621, 2001.

[123] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638, 2016.

[124] Ori Rosen, Wenxin Jiang, Gary King, and Martin A Tanner. Bayesian and frequentist inference for ecological inference: The r× c case. *Statistica Neerlandica*, 55(2):134–156, 2001.

[125] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[126] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000.

[127] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[128] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[129] Mark A. Ruzon and Carlo Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1281–1295, 2001.

[130] Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1873–1880. IEEE, 2009.

[131] Simo Sarkka. On unscented kalman filtering for state estimation of continuous-time nonlinear systems. *IEEE Transactions on automatic control*, 52(9):1631–1641, 2007.

[132] Simo Särkkä and Juha Sarmavuori. Gaussian filtering and smoothing for continuous-discrete dynamic systems. *Signal Processing*, 93(2):500–510, 2013.

[133] Simo Särkkä and Arno Solin. On continuous-discrete cubature kalman filtering. *IFAC Proceedings Volumes*, 45(16):1221–1226, 2012.

[134] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Maurice Ngolè Mboula, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning. *arXiv preprint arXiv:1708.01955*, 2017.

[135] Erwin Schrödinger. *Über die Umkehrung der naturgesetze.* Verlag Akademie der wissenschaften in kommission bei Walter de Gruyter u. Company, 1931.

[136] Vivien Seguy and Marco Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pages 3312–3320, 2015.

[137] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

[138] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[139] Sameer Shirdhonkar and David W Jacobs. Approximate earth mover's distance in linear time. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[140] Rae Shrott. Universally measurable spaces: an invariance theorem and diverse characterizations. *Fundamenta Mathematicae*, 121(2):169–176, 1984.

[141] Hermann Singer. Generalized gauss–hermite filtering. *AStA Advances in Statistical Analysis*, 92(2):179–195, 2008.

[142] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

[143] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[144] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[145] Robert L Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

[146] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66–66, July 2015.

[147] Justin Solomon, Raif M Rustamov, Leonidas J Guibas, and Adrian Butscher. Wasserstein Propagation for Semi-Supervised Learning. In *ICML*, pages 306–314, 2014.

[148] Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015.

[149] Matthew Staib, Sebastian Claici, Justin Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. *arXiv preprint arXiv:1705.07443*, 2017.

[150] Tobias Sutter, Arnab Ganguly, and Heinz Koeppl. A variational approach to path estimation and parameter inference of hidden diffusion processes. *Journal of Machine Learning Research*, 17(190):1–37, 2016.

[151] Gabriel Terejanu, Puneet Singla, Tarunraj Singh, and Peter D Scott. A novel gaussian sum filter method for accurate solution to the nonlinear filtering problem. In *Information Fusion, 2008 11th International Conference on*, pages 1–8. IEEE, 2008.

[152] Gabriel Terejanu, Puneet Singla, Tarunraj Singh, and Peter D Scott. Adaptive gaussian sum filter for nonlinear bayesian estimation. *IEEE Transactions on Automatic Control*, 56(9):2151–2156, 2011.

[153] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[154] Akif Burak Tosun, Oleksandr Yergiyev, Soheil Kolouri, Jan F Silverman, and Gustavo K Rohde. Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. *Cytometry Part A*, 87(4):326–333, 2015.

[155] A. Vedaldi and K. Lenc. MatConvNet – Convolutional Neural Networks for MATLAB. *CoRR*, abs/1412.4564, 2014.

[156] Cédric Villani. *Topics in Optimal Transportation.* American Mathematical Soc., 2003.

[157] Cédric Villani. *Optimal Transport: Old and New.* Springer Berlin Heidelberg, 2008.

[158] Michail D Vrettas, Manfred Opper, and Dan Cornford. Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. *Physical Review E*, 91(1):012148, 2015.

[159] Jon Wakefield. Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(3):385–445, July 2004.

[160] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

[161] Michael Westdickenberg and Jon Wilkening. Variational particle schemes for the porous medium equation and for the system of isentropic euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(1):133–166, 2010.

[162] Alan Geoffrey Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

[163] Dong Xu, Tat-Jen Cham, Shuicheng Yan, and Shih-Fu Chang. Near duplicate image identification with patially aligned pyramid matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[164] Jianbo Ye, Panruo Wu, James Z Wang, and Jia Li. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.

[165] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238, 2007.