Structure as Simplification:
Transportation Tools for Understanding Data

by

Sebastian Claici

B.Sc., University of Southampton (2014)
S.M., Massachusetts Institute of Technology (2016)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author . . . . . .
Department of Electrical Engineering and Computer Science
May 15, 2020

Certified by . . . . . .
Justin Solomon
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Structure as Simplification:

Transportation Tools for Understanding Data

by

Sebastian Claici

Submitted to the Department of Electrical Engineering and Computer Science
on May 15, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The typical machine learning algorithms looks for a pattern in data, and makes an assumption that the signal to noise ratio of the pattern is high. This approach depends strongly on the quality of the datasets these algorithms operate on, and many complex algorithms fail in spectacular fashion on simple tasks by overfitting noise or outlier examples.

These algorithms have training procedures that scale poorly in the size of the dataset, and their outputs are difficult to intepret. This thesis proposes solutions to both problems by leveraging the theory of optimal transport and proposing efficient algorithms to solve problems in: (1) quantization, with extensions to the Wasserstein barycenter problem, and a link to the classical coreset problem; (2) natural language processing where the hierarchical structure of text allows us to compare documents efficiently; (3) Bayesian inference where we can impose a hierarchy on the label switching problem to resolve ambiguities.

Thesis Supervisor: Justin Solomon
Title: Associate Professor of Electrical Engineering and Computer Science

Arbor invers am rămas, rupt din sferă
cu sfera aceasta aidoma, geamănă...
Şi totul îmi pare ştiut, dar nimica
din ce ştiu cu ce este nu se aseamănă.

Nichita Stănescu

Thought clambers up,
snail like, upon the wet rocks
hidden from sun and sight–

William Carlos Williams

Acknowledgments

It is difficult to acknowledge the whole cast of people whose support made this thesis possible. Some projects start off of a comment made in passing in a hallway years ago; some projects falter despite constant effort and best intentions.

I make here a best effort to remember everyone, and I apologize ahead of time if your name that should rightfully be here is nowhere to be found.

I thank my advisor, Justin Solomon, for his neverending patience with my mathematical beffudlements, and his complete honesty throughout the four years I spent working towards this thesis. His advice, research, life or otherwise, has been invaluable. I would be a worse person today had I never met him. I hope the gamble of taking on a robotics student in his first year as a professor has paid off.

I have been fortunate to work with collaborators who have kept me on the straight right path despite my obstinate meanderings. These are Edward Chien, Matthew Staib, Mikhail Bessmeltsev, Scott Schaeffer, Stefanie Jegelka, Hugo Lavenant, Mikhail Yurochkin, Farzaneh Mirzazadeh, Pierre Monteillier, Aude Genevay, and Soumya Ghosh.

Finally, but in a sense foremost, I thank my sister Sabina, my mother Camelia, and my father, Dorin for keeping me happy at the worst of times, and giving me a dose of Eastern European pessimism at the best of times. I thank my partner Catherine for the constant support: she endured several Boston winters for my sake, and was always willing to listen to my rambling complaints. I thank my friends for providing a constant source of entertainment: an ocean and a five hour time difference is no barrier to true friendship.

Contents

List of Figures

List of Tables

Notation

These are the symbols used throughout the thesis.

Symbol	Definition
$\mathbb{R}^d$	d-dimensional Euclidean space
$\mathscr{M}(X)$	space of measure on X
$\mathscr{P}(X)$	space of probability distributions on X
$C(X)$	space of continuous functions on X
$C_b(X)$	space of continuous and bounded functions on X
δ_x	Dirac mass concentrated at point x
$T_{\#}\mu$	image measure of μ through the map T
$\mu \llcorner V$	restriction of μ to the set V
$\|\cdot\|_p$	L_p norm in Euclidean space
$\|\cdot\|$	metric on whatever underlying space we are in

Table 1: List of symbols and notation.

1

Introduction

Machine learning algorithms find patterns in data. Given how central such algorithms have become to everyday tasks, an immense amount of effort is spent every year improving either their effectiveness on the tasks at hand, or their performance at scale. Many of the issues that plague learning algorithms stem from the poor data they operate on. Often, the reason for the underperformance is itself algorithmic in nature: many learning algorithms scale poorly in the size of the dataset, and so approximations must be made.

How can we improve data? That is one of two central questions asked in this thesis, and the angle we use to attack this question defines precisely what it means for an improvement to be both *close* to the initial dataset, and yet also *better* in some way.

The second question we ask is parallel to the first. Once we have our learning algorithm running on better data, how can we make sense of the patterns it is learning?

We tackle both questions through the lens of optimal transport, and we treat our data, whether images or text or anything else as distributions over a geometric space.

There are many ways of comparing probability distributions. The workhorse of machine learning has been the Kullback-Leibler divergence (or relative entropy) due to its computational ease, and its probabilistic interpretation. Kullback-Leibler is part of a family of f-divergences which all share the same disadvantage: The comparison is in terms of overlap between measures. This can lead to unintuitive results where two measures are equidistant to a third, despite appearances indicating otherwise. The optimal transport distances which takes a more geometric approach to the problem of measuring distances be-

tween measures, and avoids many of the issues with f-divergences at the cost of a heavier computational burden.

Unlike divergences which compare the amount of mass in each measure at each point in space, the optimal transport distance asks a physics-motivated question: How much effort is required to move one measure onto another? This point is illustrated in Figure 1-1 where we show the interpolant between two measures under a vertical (Kullback-Leibler) and horizontal (transport) measure of distance.

The focus of this thesis is broadly on computational aspects of the transport problem: How do we compute a transport plan efficiently if we allow ourselves some leeway in the exactness of the result? We will explore this problem across two main themes: changing the nature of the result, and changing the nature of the algorithm.

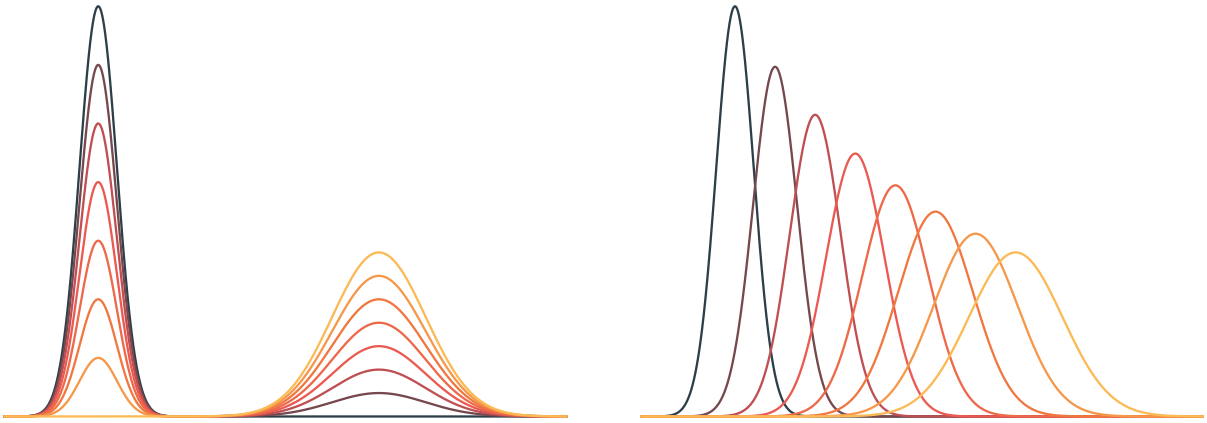
The first theme leads naturally to the notion of quantization of measures. We show some surprising connections to the definition of a *coreset* of a dataset, give a stochastic gradient algorithm to compute the quanta of a measure, and show how to extend this algorithm to compute the barycenter (or mean) of a set of distributions.

The second theme exploits or imposes a hierarchical structure on regular transport problems. As an example of a problem with existing hierarchy, we show how we can speed up and improve a popular document distance metric by exploiting the inherent topical structure of natural language documents. As an example of a problem where hierarchy can be imposed, we show that an old problem in Bayesian inference known as *label switching* can be ameliorated by lifting it into the space of distributions on distributions and using transport techniques to compute point estimates.

1.1 Overall approach

The overabundance of data can be both a boon and a bane. For modern learning algorithms, more data is almost synonymous with better performance, but data is inherently noisy, and the more of it we have, the more this noise is treated as part of the signal.

We want algorithms that can sanitise their input, uncover meaningful patterns, and which are fast and with easily understood outputs, but achieving everything at once for all the problems we are interested in is fanciful hope. Can we make progress if we break the task into smaller problems?



(a) (b)

Figure 1-1: Interpolation between two Gaussian distributions for $0 \leq t \leq 1$. (a) Interpolation under the l_2 metric: $\mu_t = (1-t)\mu_0 + t\mu_1$. (b) Interpolation under the W_2 metric: $\mu_t = ((1-t)\mathrm{Id} + tT)_{\#}\mu_0$.

There are three issues at hand:

Discretisation. For a problem to be understood by a computer, it must be presented as discrete data. Deciding on a discretisation method is a crucial first step in any tool that seeks to understand and model data. There is no universally best method; each cost function we seek to minimise, and each algorithm we wish to employ will prefer data in a specific form and with specific properties.

Modelling. Once data is properly discretised, we must model the problem we are about to solve in a way that is conducive to both performance and interpretation. It is not enough to know *what* the data is saying; we must also be able to understand *why* the data is saying so. Interpretability can be explicitly factored into a model by imposing structure on the data.

Optimisation. A model that has been adequately discretised is amenable to algorithms we can implement on a computer, but these algorithms must be practical for the model and discretisation to be useful. Practicality in many of the examples we present involves solving a non-convex optimisation problem efficiently, and we must always bear in mind the computational aspects of the problems we tackle.

The thread that links these challenges is that, in the ideal, if not in practice, they operate on data distributions. The development of tools to manipulate distributions is then of utmost importance for

the goal of understanding data.

1.2 The optimal transport problem

The tool we will use throughout the thesis is optimal transport. In this section, we give a basic introduction to the problem, and show how it connects to the goals outlined earlier: quantization and hierarchies. A longer discussion of optimal transport is given in Chapter 2.

The optimal transport setup is simple: We are given two distributions μ and ν, with μ supported on a space X, and ν supported on a space Y. We are also given a cost function $c(x, y)$ that measures the cost of moving one unit of mass from point $x \in X$ to point $y \in Y$. The problem is then to find a map $T : X \to Y$ that minimises

$$\int_X c(x, T(x)) \, \mathrm{d}\mu(x)$$

and such that $\nu(A) = \mu(T^{-1}(A))$ for all $A \subset Y$. Such a map T is known as a Monge map due to its introduction in the work of Gaspard Monge (Monge, 1781).

This problem does not always have a solution–one example is $\mu = \delta_x$ and $\nu = 1/2\delta_{y_1} + 1/2\delta_{y_2}$. The convex relaxation of Kantorovich (Kantorovich, 1942) removes the constraintthat T has to be a map between X and Y, and instead considers all distributions π over $X \times Y$. We now look for a solution of

$$\begin{aligned} &\inf_{\pi} \int_{X \times Y} c(x, y) \, \mathrm{d}\pi(x, y) \\ &\text{subject to } \begin{cases} \pi(A \times Y) = \mu(A), \ \forall A \subseteq X \\ \pi(X \times B) = \nu(B), \ \forall B \subseteq Y \\ \pi \geq 0. \end{cases} \end{aligned} \tag{1.1}$$

This formulation is a linear program in infinite dimensions, but existence of a solution depends on properties of μ, ν, and $c(\cdot, \cdot)$, and is given in detail in Santambrogio (2015); Villani (2008).

Problem (1.1) comes with an associated dual problem given by

$$\begin{aligned} &\sup_{f,g} \int_X f(x)\,\mathrm{d}\mu(x) + \int_Y g(y)\,\mathrm{d}\nu(y) \\ &\text{subject to } f(x) + g(y) \leq c(x,y). \end{aligned} \tag{1.2}$$

It is tricky to show that this is indeed the dual, or that the dual and primal solutions agree, but an intuitive understanding of how (1.2) relates to (1.1) can be arrived at by analogy to a real world problem. Suppose we have some material located at μ that we want to move to ν, and we are willing to pay at most $c(x,y)$ for someone to move one unit of mass from x to y. If we want to minimise our cost, we arrive at the primal; from the perspective of the shipment company, however, we can set prices $f(x)$ for picking up one unit at x, and $g(y)$ for dropping off one unit at y, and we wish to maximise our profit under the constraint that we cannot exceed the budget $c(x,y)$.

The dual problem can be turned into an unconstrained supremum problem if we introduce the c-transform of f, defined as $f^c(y) = \inf_x c(x,y) - f(x)$. The dual problem is then

$$\sup_f \int_X f(x)\,\mathrm{d}\mu(x) + \int_Y f^c(y)\,\mathrm{d}\nu(y).$$

The optimal transport cost defines a distance whenever $X = Y$ and $c(\cdot,\cdot)$ is a power of the metric $d(\cdot,\cdot)$ on X. The p-Wasserstein distance is given by

$$W_p(\mu,\nu) = \inf_\pi \left(\int_{X\times X} d(x,y)^p \,\mathrm{d}\pi(x,y) \right)^p \tag{1.3}$$

where π is subject to the same mass preserving constraints as in (1.1).

1.2.1 Flavours of transport

While the continuous problem (where μ and ν are absolutely continuous with respect to some base measure) is, in some sense, the most well behaved, computing a transport plan π is usually an intractable problem. The two transport problems we know how to solve are:

1. The semi-discrete problem where μ is a continuous distribution, and $\nu = \sum_{i=1}^n \alpha_i \delta_{y_i}$ is finitely

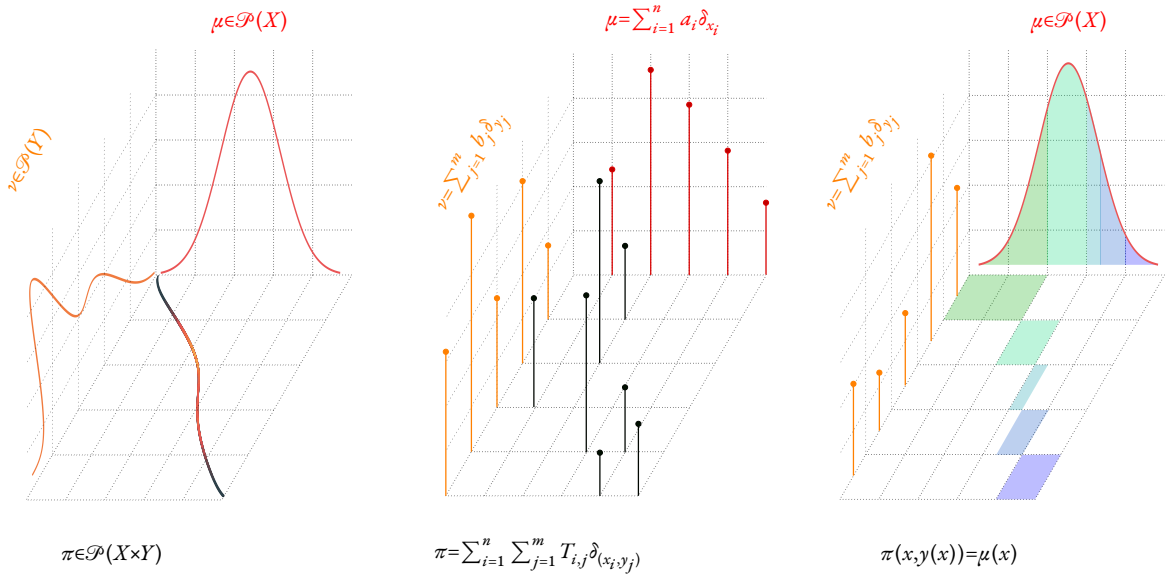


Figure 1-2: Important specialisations of the transport problem. (a) The continuous optimal transport problem requires both distributions to be absolutely continuous with respect to the volume measure. (b) The discrete optimal transport problem can be cast as a finite dimensional linear program and solved using standard matching algorithms. (c) The semi-discrete optimal transport problem asks for the distance between a continuous distribution and a discrete one. This problem can be solved efficiently using tools from computational geometry.

supported. Solving this problem requires specialised tools from computational geometry (Aurenhammer, 1987), and we can only hope for approximate solutions outside of a few special cases. An example of a semi-discrete problem is shown in Figure 1-2(a).

2. The discrete problem where both μ and ν are finitely supported. In this case the transport problem reduces to a minimum cost matching problem that can be solved using classical algorithms (Kuhn, 1955). An example of the discrete problem is shown in Figure 1-2(b).

The discrete transport problem is a finite dimensional linear program, and is amenable to polynomial time algorithms such as the Hungarian method (Kuhn, 1955). We discuss the semi-discrete problem in what follows.

1.2.2 Semi-discrete optimal transport

The semi-discrete transport problem asks for a transport plan from a measure μ that is absolutely continuous with respect to the Lebesgue measure, to a discrete measure $\nu = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$. The optimisation

variable is a finite dimensional vector $w \in \mathbb{R}^n$:

$$\sup_{w \in \mathbb{R}^n} \left\{ F(w) = \sum_{i=1}^{n} w_i \alpha_i + \int_X \min_{i=1,\dots,n} \left(c(x_i, y) - w_i \right) \, \mathrm{d}\mu(y) \right\}. \tag{1.4}$$

This is a finite dimensional optimisation problem that can be solved by gradient methods. The gradient with respect to w_i is given by

$$\frac{\partial F}{\partial w_i} = \alpha_i - \int_{V_i^w} \mathrm{d}\mu(y),$$

where $V_i^w = \{ y : c(x_i, y) - w_i \leq c(x_j, y) - w_j, \forall j \neq i \}$ is a power cell Aurenhammer (1987). When the integral of μ on V_i^w can be computed exactly, both first and second order methods are known to converge (Kitagawa et al., 2018).

One question we can ask given (1.4) is what happens when we can optimise over the α_i and x_i. This leads to a *quantization* problem, and forms part of our proposal detailed more in the following section.

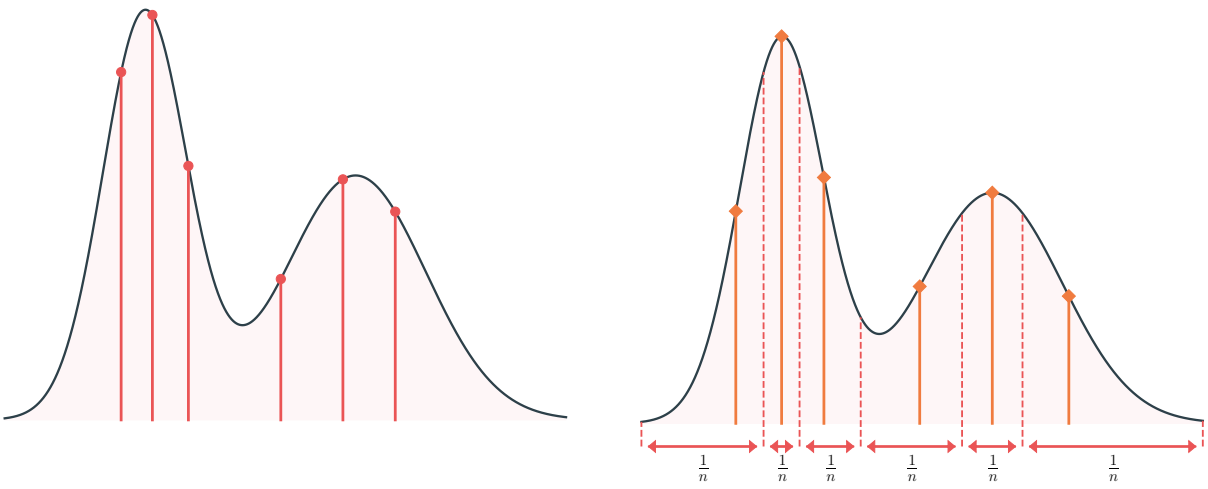
1.3 Quantization of measures

Let us return to the discrete transport problem:

$$\begin{aligned} &\min_{\pi} \sum_{i,j} \pi_{ij} \, c(x_i, y_j) \\ &\text{subject to } \begin{cases} \sum_j \pi_{ij} = a_i, \\ \sum_i \pi_{ij} = b_j, \pi \geq 0. \end{cases} \end{aligned} \tag{1.5}$$

We are minimising over all possible feasible plans π in this equation, but that may not be the only unknown. If we have control over the support points x_i and weights a_i, and wish to minimise for the distance between the measures $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$, we are looking for a *quantization* of ν. The quantization of a measure refers to an *optimal* discrete approximation with fixed number of points to a given measure ν. For the many notions of *optimality* here, see the excellent monograph (Pollard, 1982).

The optimal quantization of a measure is often impossible to compute exactly. As an example, quantizing a discrete measure under the 2-Wasserstein distance using k points leads to the k-means problem



(a) (b)

Figure 1-3: Quantiles and minimisers of problem (1.6)

and Lloyd's algorithm.

Our data is not always finite however. The quantization problem becomes significantly more important when we want to summarise a continuous *distribution* by a finite sample. Throughout the thesis we will focus on this problem, and show how accurate quantizations can be computed efficiently using semi-discrete optimal transport.

We begin with a simple illustration of this problem in one dimension.

1.3.1 Quantiles and quantization

If ν is a one dimensional distribution, and we must approximate ν by $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ where the only variable we have control over are the points x_i, what should we want the x_i to represent? One answer is that each x_i ought to represent exactly $\frac{1}{n}$ of the mass of ν. This leads us naturally to consider the quantiles of ν as one potential answer.

But a similar notion of quantization arises naturally from the transportation problem if we treat ν as continuous, set $c(x_i, y) = \|x_i - y\|^2$, and $a_i = \frac{1}{n}$. If we minimise over the x_i, problem (1.5) becomes

$$\min_{\{x_i\}} W_2^2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \nu\right). \tag{1.6}$$

Let us look at what happens when ν is one-dimensional. The dual problem to (1.6) is

$$\min_{\{x_i\}} \max_{w \in \mathbb{R}^n} \sum_{i=1}^{n} \frac{1}{n} w_i + \int_{-\infty}^{\infty} \min_i \left\{ \|y - x_i\|^2 - w_i \right\} \mathrm{d}\nu(y). \tag{1.7}$$

Of course, we can pull the min out of the integral, and integrate $\|y - x_i\|^2 - w_i$ over regions where x_i, w_i are the minimising pair. If we differentiate (1.7) with respect to the dual variables w_i, we recover a simple optimality condition

$$\frac{1}{n} = \int_{a_i}^{a_{i+1}} \mathrm{d}\nu$$

where (a_i, a_{i+1}) is the range on which $\|y - x_i\|^2 - w_i \leq \|y - x_j\|^2 - w_j$. If we order the x_i such that $x_1 \leq x_2 \leq \cdots \leq x_n$, then the optimality condition for w_1 reads

$$\frac{1}{n} = \int_{-\infty}^{a_1} \mathrm{d}\nu.$$

This is satisfied if a_1 is the first $(n+1)$-quantile of ν. Proceeding from a_1, we obtain that each a_i is one of the $(n+1)$-quantiles of ν.

While the one dimensional picture is clear, higher dimensions pose challenges. Quantiles are not easy to define in higher dimensions, and the definitions that are available to us rely on solving intractable problems (Carlier et al., 2016). However, we can still look for a solution of the underlying problem of approximating a measure by a discrete distribution.

1.3.2 Wasserstein measure coresets

This notion of quantization extends naturally to higher dimensions, and reveals a connection to a subfield of machine learning that studies sparsification of datasets. A *coreset*[1] is a small subset of a dataset that approximates the performance of an algorithm on that dataset. We can be more precise, and say that if $\mathcal{F}$ is some hypothesis set of functions that our algorithm can learn, and X is a dataset, we want the following

[1] The name is a portmanteau of the words *core* and *dataset* and should be read as the *core of a dataset.*

condition to hold for our coreset C:

$$\left| \sum_{x \in C} \mu_C(x) f(x) - \sum_{x \in X} \frac{1}{n} f(x) \right| \leq \varepsilon \sum_{x \in X} \frac{1}{n} f(x)$$

for any $f \in \mathcal{F}$. Here $\mu_C(x)$ is the weight of point x in the coreset. If the size of C is small relative to X, and we can achieve a small error ε, then our coreset accurately represents the dataset, and is much easier to train on.

Coreset constructions are typically discrete and refer to a particular algorithm. For example, there are algorithms that construct coresets for k-means, and algorithms that construct coresets for support vector machine classification, but no algorithm will do both well.

Our proposal is twofold.

First, we extend coreset language to the continuous setting, where the dataset X is replaced by a measure μ over the underlying space. This leads to a notion of a *measure* coreset defined as a measure ν for which

$$\left| \mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{x \sim \nu}[f(x)] \right| \leq \mathbb{E}_{x \sim \mu}[f(x)]$$

for all $f \in \mathcal{F}$.

The connection to optimal transport comes from the dual problems of W_1 and W_2. The W_1 distance is given as a supremum over 1-Lipschitz functions:

$$\sup_{f \in \mathrm{Lip}_1} \int_X f(x) \, \mathrm{d}\mu(x) - \int_X f(x) \, \mathrm{d}\nu(x),$$

while W_2 admits an inequality with respect to functions in the Sobolev space H^1 of functions with bounded quadratic growth. What this means for our coreset problem is that if $\mathcal{F} \subseteq \mathrm{Lip}_1$ or $\mathcal{F} \subseteq H^1$, then it is enough to construct a measure ν that is close to μ with respect to the W_1 or W_2 distances.

To make this problem concrete, let's put a prior on ν, and say that $\nu = {}^1\!/_n \sum_{i=1}^n \delta_{x_i}$. Our goal is then

to find the optimal location of the points x_i by minimising

$$W_1\left(\mu, \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}\right) \quad \text{or} \quad W_2\left(\mu, \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}\right).$$

We call these *measure coresets*, and we can show they typically outperform classical coresets on discrete problems, despite being generic in nature and construction.

1.3.3 Stochastic Wasserstein barycenters

Measure coresets are quantizations of a single distribution. A natural question to follow up is to ask for the quantization of multiple distributions. What single finitely supported distribution best approximates all the input distributions $\mu_1, \ldots, \mu_m$, all supported on the same space X.

If we pose this as a minimization problem over all distributions, we arrive at the Wasserstein barycenter definition:

$$\nu^* = \arg\min_{\nu} \sum_{j=1}^{m} \lambda_j W_2^2(\nu, \mu_j) \tag{1.8}$$

where the λ_j are non-negative weights that sum to 1. This notion of barycenter captures the geometric intuition of Figure 1-1, as the interpolant distributions can be seen as barycenters between μ_0 and μ_1 for different values of λ.

Computing ν^* is typically computationally intractable. Approximate solutions can be arrived at by making assumptions. The simplest assumption to make is that the μ_j and the output barycenter ν are supported on a finite grid, and the goal is then to figure out the amount of mass at each grid point that would minimise (1.8). The barycenter problem on a grid can be written as a linear program as in (Carlier et al., 2015), and faster algorithms can be obtained by entropically regularising the transport problems (Benamou et al., 2015a; Solomon et al., 2015), or by distributing computation across multiple machines (Staib et al., 2017).

This approach leads to an approximation of the true solution as the barycenter of distributions supported on a grid of n points need not be supported on the same grid. To avoid this issue, we propose an approach to compute free support barycenters by using samples from the input distributions.

We parameterise our barycenter as $\nu = 1/n \sum_{i=1}^{n} \delta_{x_i}$, and optimise for the positions of the x_i by solving

$$\arg\min_{x_1,\dots,x_n} \sum_{j=1}^{m} \lambda_j W_2^2 \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}, \mu_j \right). \tag{1.9}$$

If we write down this problem in terms of its dual, the gradient step with respect to the points x_i can be written as a sum of simple expectations under the μ_j. This makes the algorithm amenable to stochastic gradient methods, and thus allows us to solve (1.9) for distributions μ_j that are absolutely continuous with respect to some base measure, as long as we have sample access to the μ_j.

1.4 Hierarchical transport

The methods we have seen so far approximated distributions by point sets or parametric distributions under the Wasserstein distance, but the algorithms for computing these approximations require a solver for the original optimal transport problem.

Computing the transport distance is challenging and scales poorly with increasing size of the input distributions. One approach that we propose in this section to mitigate this problem is to exploit inherent structure in instances of the transport problem to reduce the complexity of the problem.

The first example of this that we will look at is in text comparison

1.4.1 Hierarchical optimal transport for document retrieval

If we were to ask somebody to compare two novels, a typical response would centre around thematic similarities and differences. We would likely hear the words: "This novel is *about* ...". But how can we teach a machine what a novel or news article is about, and how can we compare these themes to figure out which novels are similar to which?

The goal of this project is to extend results in document comparison proposed by Kusner et al. (2015a) to handle large documents and yield better, more interpretable results. The approach of Kusner et al. (2015a) treats documents as distributions over words of a common vocabulary by normalising word counts in each document. The distance between two documents is given by the optimal transport cost between

these distributions. Of course, to compute a transport cost requires a ground metric between the support points of the distributions–in this case, the words. The approach relies on a word embedding space (Mikolov et al., 2013; Pennington et al., 2014) for this purpose. The resulting distance is known as the word mover's distance (WMD) between documents, and has better interpretability than black-box models as the word-to-word matching allows one to interpret why two documents are similar.

WMD is costly to compute. Because each document-to-document distance is computed as the optimal transport cost between two distributions supported on a potentially very large set (vocabulary sizes for a novel range in the tens of thousands of unique words), the computational cost of a single comparison is frequently infeasible. The WMD between documents $d_1 = \sum_{v \in V} \alpha_v \delta_v$ and $d_2 = \sum_{v \in V} \beta_v \delta_v$ is

$$\mathrm{WMD}(d_1, d_2) = W_1\left(\sum_{v \in V} \alpha_v \delta_v, \sum_{v \in V} \beta_v \delta_v\right). \tag{1.10}$$

In our work, we think of each document as a distribution over topics instead of words. The total number of topics in a document collection is typically small (on the order of 30), which mitigates the computational cost of the transport distance. This leads to the same problem as before: What is a ground metric on the space on topics? Our approach is to view this as a hierarchical transport problem, namely:

$$\mathrm{HOTT}(d_1, d_2) = W_1\left(\sum_{t \in T} \mu_t \delta_t, \sum_{t \in T} \nu_t \delta_t\right) \tag{1.11}$$

where the inner sum is over topics, and the ground metric between topics t_1 and t_2 is the WMD between these two topics treated as distributions over words.

As topics can be maintained on the fly using methods such as latent Dirichlet allocation (Blei et al., 2003), this approach scales to large document collections, and is robust to document size.

1.4.2 Alleviating label switching in Bayesian inference

Hierarchies are natural in interpreting text documents. For an example of a problem where identifying the hierarchy is not obvious, but where doing so leads to an improved algorithm, consider the *label switching* problem in Bayesian inference. *Label switching* occurs in Bayesian inference when the posterior is invariant to some group acting on the parameters of the distribution. For example, an MCMC sampler

from the posterior of a mixture of three Gaussians does not care whether its samples from the posterior are ordered as first, second, third Gaussian or third, second, and first Gaussian.

We can see this issue most clearly if we look at the likelihood of the model. Let Θ be the set of all parameters of a model. In this case, our model is a mixture of Gaussians with identity covariance; the parameters are only the mean of each Gaussian and the weight. The likelihood is given by:

$$p(x|\Theta) = \sum_{k=1}^{K} \pi_k f(x; \mu_k). \tag{1.12}$$

For any permutation $\sigma(\cdot)$ of labels $k = 1, \dots, K$, let $\sigma(\Theta) = \{\theta_{\sigma(1)}, \dots, \theta_{\sigma(K)}\}$. The likelihood under σ is unchanged:

$$p(x|\sigma(\Theta)) = \sum_{k=1}^{K} \pi_{\sigma(k)} f(x; \mu_{\sigma(k)}) = p(x|\Theta).$$

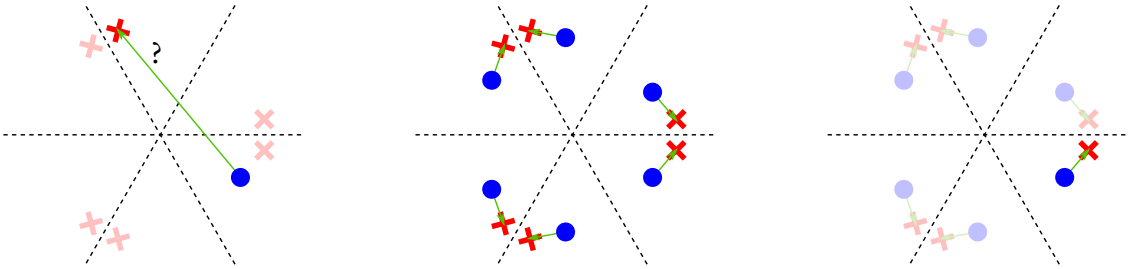
This leads to trouble down the road if we want to obtain expectations under the posterior distribution, as this permutation invariance of the likelihood, but not of the parameters can lead to meaningless averages.

The goal is then simple: To compute an expected value for the posterior in such a way as to avoid the label switching issue. Most approaches to this problem rely on finding a privileged sample to align everything to. Our proposal is different. Instead of aligning everything to a single sample, lift every sample to the space of distributions and encode all possible permutations in the distribution associated to each sample.

This procedure turns samples θ into distributions $1/|S_K| \sum_{\sigma \in S_K} \delta_{\sigma(\theta)}$, and allows us to use notions of Wasserstein barycenters to obtain a principled average. While this leads to a simple and naive algorithm, it is nearly impossible to compute a free support Wasserstein barycenter for distributions with such large support size (the symmetric group on K labels has $K!$ elements).

We can prove two facts that significantly help here: (1) The barycenter of distributions invariant under some group action is also invariant under the same group action; and (2) The barycenter is supported on the same number of support points as the input distribution.

We can now apply stochastic gradient methods to the problem of estimating an expectation, and work with a single point in the quotient space for an efficient algorithm that is highly adaptable to any group action and works for distributions supported on non-Euclidean domains. This idea is illustrated



(a) Ambiguity (b) Orbit empirical distribution (c) Quotient update

Figure 1-4: (a) Suppose we wish to update our estimate of the average (blue) given a new sample (red) from; due to label switching, other points (light shade) have equal likelihood to our sample, causing ambiguity. (b) Our proposal suggests an unambiguous update by constructing $|G|$-point orbits as empirical distributions and doing gradient descent with respect to the Wasserstein metric. (c) This algorithm is equivalent to moving one point, with a careful choice of update functions.

in Figure 1-4.

1.5 Overview

Only reading the section titles above, one would think we will present a scattering of applications of optimal transport with only a loose thread connecting them. However, if we think of these problems in terms of the algorithms used to solve them, a coherent picture emerges. The optimal transport problem is old by now, and research has centred on solving simple variants of it; this research has culminated in very efficient solvers for the discrete problem (Cuturi, 2013; Altschuler et al., 2017), and exhaustive theory and algorithms for the semi-discrete problem (Kitagawa et al., 2018; Mérigot, 2011).

To arrive at further improvement, we must change the nature of the problem. What happens in the semi-discrete problem if we allow for the support of the discrete distribution to change? How do we exploit the structure of the distributions in the discrete problem? These questions lead naturally to the problem of quantization and the idea of using hierarchical structure.

The goal of this thesis is to uncover where these new algorithms can be found, and how to develop them into useful tools. To this end, we provide details on the optimal transport problem in Chapter 2, including derivations of many of the identities mentioned in this section. In Part I, we describe the measure quantization problem and show how it links to classical coreset construction algorithms in machine learning; we then extend this definition to the problem of computing means in the space of probability

measures.

Quantization is a good first step towards simplifying complicated distributions, but at the end of the day, comparing two distributions still requires solving a complicated optimal transport problem that scales poorly in the size of the input. How can we make these algorithms faster? The common approach in the transport literature is to approximate the true solution by something that is smoother and easier to optimise for. This line of entropic regularised transport has seen great success, but it only gets you so far. Instead of developing new algorithms to solve the transport problem, we can instead rely on the data to guide us towards an efficient approximation of the solution. This leads to the idea of exploiting the hierarchical structure of the data to reduce the size of the problem we seek to solve. We explore this approach in Part II and show how it can be applied to problems in natural language processing and Bayesian inference.

We finish with a practical approach to computing the transport cost when computing geodesics is difficult. In Part III we extend the dynamical formulation of Benamou & Brenier (2000) to discrete surfaces such as triangle meshes.

There are two ways to read through this thesis. The first focuses on the algorithmic aspect, and sees the progression of approaches as more and more specialised tools for solving optimal transport problems, from the generic approximation of a measure in Part I to the specialised tools of Part II, and the extension of the dynamical formulation to triangulated surfaces in Part III.

The second reading revolves around how the constraints of the problem shape the approach. Part I imposes no constraints on the problem input, but requires a finite measure as an output. Part II relies heavily on an existing hierarchical structure in the data that we exploit for faster and more effective algorithms.

2

Optimal Transport

Wherein we present the optimal transport problem, the main workhorse of this thesis. Care is taken towards developing mathematical intuition on these results, and less on formal correctness.

We will present optimal transport as it was developed historically, starting from Monge's ideas developed in the late 18th century (Monge, 1781), passing to Kantorovich's convex relaxation of Monge's formulation and detailing Kantorovich duality (Kantorovich, 1942), and ending with a look at the space of distributions under the Wasserstein distance, including notions of Fréchet means (Agueh & Carlier, 2011).

The theory of optimal transport relies heavily on notions of duality. For details, we recommend the first chapters of either Santambrogio (2015) or Villani (2008). For our purposes, we will denote the space of continuous bounded functions on the space X as $C_b(X)$, and rely on the fact that it is the dual space to the space $\mathscr{M}(X)$ of measures on X.

2.1 Monge transport maps

The initial motivation for the optimal transport problem was resource allocation (Monge, 1781). Monge proposed a notion of distance that measured the work required to move earth from one site (the *deblais*) to another (the *remblais*). Everything else equal, the cost for moving a single particle of earth ought to be proportional to the mass of the particle multiplied by the distance it has to travel; the total cost is given by the sum of this amount over all particles. To write this in measure theoretic language, let's assume we

have some measure μ that represents the earth to be moved, and another measure ν that represents the location where we want to move it. Monge's formulation thinks of every point x in the support of μ as a particle with mass $\mu(x)$, and looks for a destination for that particle $x \to T(x)$. The goal is to minimise the total work given a notion of cost $c(\cdot,\cdot)$ between x and $T(x)$. This leads to the Monge problem:

Definition 2.1 (Monge transport). *Given measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and a cost function $c(x,y)$ for every $x \in X$, $y \in Y$, the Monge map is given by:*

$$\inf_T \int_X c(x, T(x))\, \mathrm{d}\mu(x), \tag{2.1}$$

with the constraint that the pushforward $T_{\#}\mu = \nu$.

This definition is easy enough to understand, but we run into a problem quickly: Such a map T may not exist at all. A simple example is $\mu = \delta_x$ and $\nu = 1/2\delta_{y_1} + 1/2\delta_{y_2}$. The Monge formulation does not allow for mass to be split, and thus no T would satisfy $T_{\#}\mu = \nu$. This is exactly the assumption that we relax to arrive at a convex relaxation where problems of existence and uniqueness are easier to tackle.

Let's look at a simple example where a Monge map exists and is easy to understand in terms of permutations. Let $\mu = \sum_{i=1}^n 1/n\, \delta_{x_i}$, and $\nu = \sum_{j=1}^n 1/n\, \delta_{y_j}$; in this case, Monge maps exist as the problem reduces to matching problem of assigning to each x_i exactly one y_j. In other words, we are looking for a permutation σ which minimises $\sum_{i=1}^n c(x_i, y_{\sigma(i)})$.

2.2 Kantorovich transport plans

We can look at (2.1) from a different perspective. Instead of looking for a transport map $T(x)$, we can ask how much mass is moved from point x to point y; if there is a way to move all of the mass at x to a single point y for all points x in the support of μ, then T exists. But, if we relax this constraint, we arrive at the Kantorovich formulation of optimal transport:

Definition 2.2 (Kantorovich transport). *Given measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and a cost function $c(x,y)$*

for every $x \in X$, $y \in Y$, the Kantorovich optimal transport problem is given by:

$$\begin{aligned} &\inf_{\pi} \quad \int_{X\times Y} c(x,y)\,\mathrm{d}\pi(x,y) \\ &\textit{subject to} \quad \begin{cases} \pi(A\times Y) = \mu(A) \\ \pi(X\times B) = \nu(B) \end{cases} \quad \forall A \subseteq X, B \subseteq Y. \end{aligned} \tag{2.2}$$

In other words, we look for a distribution π on the product space $X \times Y$ whose X and Y marginals agree with μ and ν, and which minimises the total transport cost.

The discrete problem is helpful for developing intuition. Let $\mu = \sum_{i=1}^{m} \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{n} \beta_j \delta_{y_j}$ be our two input distributions. The space of distributions on (x_i, y_j) is the space of doubly stochastic matrices in $\mathbb{R}^{m\times n}$, and we can rewrite (2.2) as a linear program with a matrix unknown and $m + n$ constraints:

$$\begin{aligned} &\min_{\pi\in\mathbb{R}^{m\times n}} \quad \sum_{i,j} \pi_{i,j} c(x_i, y_j) \\ &\text{subject to} \quad \begin{cases} \sum_{j=1}^{n} \pi_{i,j} = \alpha_i, \forall i \\ \sum_{i=1}^{m} \pi_{i,j} = \beta_j, \forall j \\ \pi_{i,j} \geq 0, \forall i, j \end{cases} . \end{aligned} \tag{2.3}$$

The underlying question here is how to best match supply $(\alpha_i \delta_{x_i})$ with demand $(\beta_j \delta_{y_j})$ under capacity constraints. This ties directly with network flow and stable marriage problems.

When $X = Y$, and the cost function is the metric $d(\cdot,\cdot)$ on X, the value of (2.2) is known as the 1-Wasserstein distance between μ and ν, and the value of (2.3) is known as the Earth mover's distance (or EMD for short).

2.2.1 Dual formulation

Problem (2.2) is a linear program with linear constraints. We can derive a dual form; first, the constraints admit a weak form as

$$\begin{cases} \int_X f(x)\,\mathrm{d}\mu(x) - \int_{X\times Y} f(x)\,\mathrm{d}\pi(x,y) = 0 \\ \int_Y g(y)\,\mathrm{d}\nu(y) - \int_{X\times Y} g(y)\,\mathrm{d}\pi(x,y) = 0. \end{cases} \qquad \forall f \in C_b(X), g \in C_b(Y)$$

If the constraints are not satisfied, then there is some choice of f and g such that

$$\int_X f(x)\,\mathrm{d}\mu(x) + \int_Y g(y)\,\mathrm{d}\nu(y) - \int_{X\times Y} (f(x)+g(y))\,\mathrm{d}\pi(x,y) = +\infty.$$

We can now turn (2.2) into an unconstrained problem:

$$\inf_\pi \int_{X\times Y} c(x,y)\,\mathrm{d}\pi(x,y) + \sup_{\substack{f\in C_b(X)\\ g\in C_b(Y)}} \int_X f(x)\,\mathrm{d}\mu(x) + \int_Y g(y)\,\mathrm{d}\nu(y) - \int_{X\times Y} (f(x)+g(y))\,\mathrm{d}\pi(x,y). \quad (2.4)$$

We would like to exchange the inf and sup to eliminate π. Proving that this is possible is not obvious, and we refer the reader to either Santambrogio (2015) or Villani (2008) for a full proof[1]. For now, let us assume that this is possible and proceed with our argument. The term that depends on π is

$$\inf_\pi \int_{X\times Y} c(x,y) - (f(x)+g(y))\,\mathrm{d}\pi(x,y) = \begin{cases} 0, & \text{if } f(x)+g(y) \le c(x,y) \\ -\infty, & \text{otherwise} \end{cases}. \quad (2.5)$$

Assuming the supremum in (2.4) exists, we can eliminate π using (2.5) and replace it with the constraint $f(x)+g(y) \le c(x,y)$. This leads to the Kantorovich dual formulation of optimal transport:

$$\begin{gathered} \sup_{\substack{f\in C_b(X)\\ g\in C_b(Y)}} \int_X f(x)\,\mathrm{d}\mu(x) + \int_Y g(y)\,\mathrm{d}\nu(y) \\ \text{subject to} f(x)+g(y) \le c(x,y), \forall x \in X, y \in Y. \end{gathered} \quad (2.6)$$

[1] The easiest, but least illuminating way to prove this is to use the duality theorem of Fenchel-Rockafeller.

When μ and ν are finite dimensional, this reduces to a linear program with $m + n$ variables, and mn constraints; compare this with the linear program for the primal problem which had mn variables and $m + n$ constraints.

We can make one more observation about (2.6) before moving on. Let's say we fix f and want to find the optimal g. The only constraint to worry about is $f(x) + g(y) \leq c(x, y)$, and since we want to maximise $\int g \, \mathrm{d}\nu$, we might as well set $g(y) = \inf_x c(x, y) - f(x)$ for all y as this maximises g pointwise. The value $\inf_x c(x, y) - f(x)$ is known as the c-transform of f and is typically written as f^c. If we make this substitution we arrive at an unconstrained form of the dual problem with a single function to optimise over:

$$\sup_{f \in C_b(X)} \int_X f(x) \, \mathrm{d}\mu(x) + \int_Y \inf_{x \in X} \left(c(x, y) - f(x) \right) \, \mathrm{d}\nu(y). \tag{2.7}$$

Equations (2.2) and (2.7) allow us to talk about the existence and uniqueness of transport plans in problems with only weak assumptions on the input distributions and underlying spaces X and Y. What we still do not know how to do in general is solve for this transport plan even if we know it exists and is unique. Gaussian distributions represent one of the few exceptions where not only are the distance and plan known, but computing them is a rather painless affair. We take a short detour to derive these identities.

2.2.2 Wasserstein distance between Gaussian distributions

The Gaussian distribution centred at mean μ with covariance Σ has density

$$p(x; m, \Sigma) = \frac{1}{\sqrt{2\pi} \det|\Sigma^{-1}|} \exp\left\{ -(x - m)^\intercal \Sigma^{-1} (x - m) \right\}. \tag{2.8}$$

The Wasserstein distance between Gaussians $\mathcal{N}(m_1, \Sigma_1)$ and $\mathcal{N}(m_2, \Sigma_2)$ is given by

$$W_2^2 \left(\mathcal{N}(m_1, \Sigma_1), \mathcal{N}(m_2, \Sigma_2) \right) = \|m_1 - m_2\|_2^2 + \mathrm{Tr}\left[\Sigma_1 + \Sigma_2 - 2 \left(\Sigma_2^{1/2} \Sigma_1 \Sigma_2^{1/2} \right) \right]. \tag{2.9}$$

Let's try to derive this equation for Gaussians centred at the origin. We can rewrite the 2-Wasserstein distance as

$$W_2^2(\mu,\nu) = \inf_{\pi} \mathbb{E}_{(X,Y)\sim\pi}\left[|X-Y|^2\right] \tag{2.10}$$

where π is a coupling on $X \times Y$ with marginals μ and ν. We expand the square inside the expectation for

$$\mathbb{E}_{X\sim\mu}\left[X^\intercal X\right] + \mathbb{E}_{Y\sim\nu}\left[Y^\intercal Y\right] - 2\mathbb{E}_{(X,Y)\sim\pi}\left[X^\intercal Y\right] = \mathrm{Tr}\left(\Sigma_1\right) + \mathrm{Tr}\left(\Sigma_2\right) - 2\mathrm{Tr}\left(C\right).$$

We define C from the covariance matrix of the joint law $(X,Y) \sim \pi$:

$$A = \begin{pmatrix} \Sigma_1 & C \\ C^\intercal & \Sigma_2 \end{pmatrix}.$$

To minimise $-2\mathrm{Tr}\,(C)$ subject to positive semi-definite constraints on A is thus equivalent to minimising (2.10) over laws π. Proving that this minimiser is achieved by the expression in (2.9) is not trivial, but a detailed proof can be found in Givens et al. (1984).

2.2.3 One dimensional transport

Even if we do not have Gaussian distributions, as long as our distributions are absolutely continuous with respect to the volume measure and one-dimensional, the transport plan between them can be computed exactly and easily. We will again provide an intuitive derivation of the result without worrying about formal correctness.

Let's assume that μ and ν have densities f and g, and let's look at a point $x \in \mathbb{R}$ in the support of μ. Notice that for any $y \le x$, it must hold that $T(y) \le T(x)$ as otherwise we can improve the total cost by swapping $T(x)$ and $T(y)$. Thus, for any x, the total mass up to x must have been mapped to somewhere to the left of $T(x)$. That is,

$$\int_{-\infty}^{x} f(x)\,\mathrm{d}x = \int_{-\infty}^{T(x)} g(x)\,\mathrm{d}x.$$

If we write $\mathrm{CDF}_f(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$ and $\mathrm{CDF}_f^{-1}(t) = \inf_{x\in\mathbb{R}}\left\{\mathrm{CDF}_f(x) \ge t\right\}$, then we can write the

condition above as

$$\mathrm{CDF}_f(x) = \mathrm{CDF}_g(T(x))$$

and hence

$$T(x) = \mathrm{CDF}_g^{-1}\left(\mathrm{CDF}_f(x)\right). \tag{2.11}$$

2.3 Semi-discrete transport

We turn now to a very important case of equation (2.7). If μ is a finitely supported distribution $\mu = \sum_{i=1}^m \alpha_i \delta_{x_i}$, then the first integral of (2.7) reduces to a sum, and the optimisation variable f is now a vector in $\mathbb{R}^m$. Let's call this vector v to distinguish it from the continuous case, and rewrite (2.7):

$$\max_{v \in \mathbb{R}^m} \left\{ F[v] = \sum_{i=1}^m \alpha_i v_i + \int_Y \min_i \left(c(x_i, y) - v_i\right) \mathrm{d}\nu(y) \right\}. \tag{2.12}$$

The min term in the integral above partitions space into convex regions that are generalisations of Voronoi regions. We will write $V_i^v = \{y : c(x_i, y) - v_i \leq c(x_j, y) - v_j, \forall j \neq i\}$ for this *power* region for point x_i with weight v_i.

We are in better shape to optimise (2.12) as v is just a vector in $\mathbb{R}^m$. Let's look at the gradient of the functional F with respect to v:

$$\frac{\partial F}{\partial v_i} = \alpha_i - \int_{V_i^v} \mathrm{d}\nu(y). \tag{2.13}$$

At optimality this tells us that the amount of mass of ν in the region V_i^v is equal to the mass at point x_i; the transport plan from ν to μ simply moves all mass within V_i^v to the point x_i.

We can thus solve (2.12) by gradient descent on the vector v. The semi-discrete problem, and approaches towards its solution have been popular in applied mathematics (Lévy, 2015; Lévy & Schwindt, 2018), fluid simulation (de Goes et al., 2015b; Mérigot & Mirebeau, 2016; Gallouët & Mérigot, 2017), and image processing (de Goes et al., 2011).

We restrict our attention to the metric $c(x, y) = |x - y|$. The main stumbling block in using gradient descent to solve (2.12) is computing the integral of ν on the region V_i^v. Computing the regions is a known problem in computational geometry (Aurenhammer, 1987) with existing implementations for 2 and 3

dimensions (The CGAL Project, 2020). The power diagram also gives us access to the boundaries of cells; let us say that points x_i and x_j are neighbours in the power diagram, and write $\partial V^v_{i,j}$ for the boundary between V^v_i and V^v_j. Let us further say that ν has a density $p(\cdot)$ so that we can write $\mathrm{d}\nu(y) = p(y)\,\mathrm{d}y$. This is enough to obtain second order derivatives of F:

$$\frac{\partial^2 F}{\partial v_i \partial v_j} = -\frac{1}{|x_i - x_j|} \int_{\partial V^v_{i,j}} p(x)\,\mathrm{d}x. \tag{2.14}$$

Because the semidiscrete transport problem is concave in v, Newton and quasi-Newton methods are known to converge (Kitagawa et al., 2018).

That said, while the regions are convex, few measures μ admit an exact form for $\int_{V^v_i} \mathrm{d}\nu$. One solution that we will explore later on is to sample from ν and compute stochastic gradients.

2.4 Wasserstein barycenters

Imagine we want to find a point $\bar{x}$ closest to a set of $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$. If we did not know that the solution is given by the Euclidean average of the points $\bar{x} = {}^1\!/_n \sum_{i=1}^n x_i$, we could pose this as an optimisation problem that asks for a solution to

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^n \|x - x_i\|^2. \tag{2.15}$$

Setting the derivative w.r.t. x to 0 gives us the Euclidean average. We can ask the same question of a set of distributions $\{\mu_1, \dots, \mu_n\}$. We no longer have a simple notion of average, but we can define one by analogy to what we did above.

Definition 2.3 (Wasserstein barycenter). *The Wasserstein barycenter of the distributions $\mu_1, \dots, \mu_n$ with weights $\lambda_1, \dots, \lambda_n$ summing to 1 is given by a solution of*

$$\inf_{\nu} \sum_{j=1}^n \lambda_j W_2^2(\mu_j, \nu). \tag{2.16}$$

The infimum in (2.16) exists and is unique if all of the μ_j have finite second moments, and at least one of them is absolutely continuous with respect to the volume measure on the underlying space X (Agueh

& Carlier, 2011).

Despite these appealing theoretical properties, computing a Wasserstein barycenter can only be done for a few specialised distributions. We are thus left with only a few strategies to make this computation tractable. We can discretise space and assume every distribution is finitely supported. The barycenter problem reduces to a finite dimensional linear program in this case with efficient algorithms (Carlier et al., 2015; Solomon et al., 2015; Benamou et al., 2015b; Cuturi & Doucet, 2014). We can approximate the Wasserstein distance by projecting the input distributions onto random lines which reduces the problem to a matching problem in one dimension (Rabin et al., 2011). Or we can assume that transport *maps* between the input distribution and a given distribution can be computed easily, and arrive at the barycenter through a fixed point scheme (Álvarez-Esteban et al., 2016).

In §5 we present a different approach that only assumes sample access to the input distributions and approximates the barycenter by a finite point set.

Part I

Quantization

3

Introduction to Quantization

Wherein we find the best choice of samples from a measure. Applications to learning problems show how it is sometimes better to simplify your data than to complicate your algorithm. An extension of this idea to summarising data from multiple sources follows.

Now that we have developed a common language, it is time for us to delve into the details of how we can use the tool of optimal transport to tackle fundamental problems in machine learning. We begin with two questions: What is the best approximation to a distribution? How do we evaluate the quality of samples? Answering these questions is the goal of this chapter.

We begin with the problem of evaluating the expectation of a function f under a distribution μ. To write this out:

$$\mathbb{E}_{x\sim\mu}\left[f(x)\right] = \int_X f(x)\,\mathrm{d}\mu(x).$$

It is almost always impossible to evaluate the integral on the right hand side, and we have to resort to approximations. For example, if we can sample from μ easily, then we can approximate the integral by a finite sum $\sum_{i=1}^{n} \frac{1}{n} f(x_i)$. How close is this approximation to the integral? That depends on f, X, and the number of points n. We shall see shortly that a reasonable error estimate is given by the Wasserstein distance between μ and the sample distribution:

$$\left| \int_X f(x)\,\mathrm{d}\mu(x) - \sum_{i=1}^{n} \frac{1}{n} f(x_i) \right| \leq L \cdot W_1\left(\mu, \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}\right). \tag{3.1}$$

Here L is the Lipschitz constant of the function f. If f is not Lipschitz, then other inequalities may hold.

The link between (3.1) and generalisation error, and between generalisation error and data sparsification are the subject of §4.

This approach to quantization works well if all of our data comes from the same source, but this is not always the case. Some examples of this problem in machine learning are in federated learning (Yurochkin et al., 2019a), and distributed posterior inference (Srivastava et al., 2015b). In the presence of multiple data sources, we could simply apply the approach of Chapter 4 to each data source, but this is costly and is more susceptible to dataset specific noise. Instead, we can pose the problem as an averaging problem. The same way we asked before what is the best finite approximation to a given distribution, we can ask what is the best finite approximation to a set of distributions and seek to minimise a functional that looks something like

$$\sum_{j=1}^{J} \left| \int_X f(x) \, \mathrm{d}\mu_j(x) - \sum_{i=1}^{n} \frac{1}{n} f(x_i) \right| . \tag{3.2}$$

over the points x_i. We are lead to the notion of a Wasserstein barycenter (Agueh & Carlier, 2011) of a set of distributions, and in §5 we give an algorithm to compute such an approximation.

This chapter is based on Claici et al. (2018) and Claici et al. (2020).

4

Wasserstein Measure Coresets

What is the best summary of a dataset, and what can we say about the performance of algorithms on the summary? The answer to these questions is the goal of a coreset. In this chapter, we extend the coreset idea to distributional data, and give simple yet effective algorithms to construct coresets for a large class of problems.

4.1 Introduction

How do we deal with too much data? Despite the common wisdom that more data is better, algorithms whose complexity scales with the size of the dataset are still routinely used in many areas of machine learning. While large datasets capture high frequency differences between data points, many algorithms only need a handful of *representative* samples that summarise the dataset.

Formalising a notion of *representative* requires care, however, since a representative sample for a clustering algorithm may differ from that for a classification algorithm. The notion of a data *coreset* was introduced to specify precisely a notion of data summarisation that is task dependent. Originally proposed for computational geometry, coresets have found their way into the learning literature for tasks ranging from clustering (Bachem et al., 2018b), classification (Tsang et al., 2005), neural network compression (Baykal et al., 2018), and Bayesian inference (Huggins et al., 2016; Campbell & Broderick, 2019).

Coreset construction is typically posed as a discrete optimisation problem: Given a fixed dataset and learning algorithm, how can we construct a smaller dataset on which that algorithm achieves similar per-

formance? This approach, however, ignores a key theme in machine learning. A dataset is an empirical sample from an underlying data distribution, and learning problems typically seek to minimise an expected loss against the distribution, not the dataset. The effectiveness of a coreset should thus be measured against the *distribution*, and not the *sample*. In other words, the coreset should be designed to guarantee good generalisation.

To address this oversight, we introduce *measure coresets*, which approximate the dataset by either a parametric continuous measure or a finitely supported one with a smaller number of points. Our formulation extends coreset language to smooth data distributions and recovers the original formulation when the distribution is supported on finitely many points. We specifically focus on *Wasserstein measure coresets*, which hinge on a natural connection between coreset language and optimal transport theory.

Contributions. We generalise the definition of a coreset to take into account the underlying data distribution, producing a *measure coreset*, with strong generalisation guarantees for a variety of learning problems. Our formulation reveals an elegant connection to optimal transport, allowing us to leverage relevant theoretical results to obtain generalisation error bounds for our coresets as well as stability under Lipschitz transformations. From a computational perspective, we provide stochastic algorithms for extracting measure coresets, yielding methods that are well-adapted to cases involving incoming streams of data. This allows us to construct coresets in an online manner, without having to store the whole dataset in memory. Besides, contrarily to existing methods which are specific to a given learning problem, our formulation is robust enough so that a given coreset can be used for different tasks.

4.1.1 Related work

We join the probabilistic language of optimal transport with the discrete setting of data compression via coresets.

Coresets. Initially introduced in computational geometry (Agarwal et al., 2005), coresets have found their way to machine learning research via importance sampling (Langberg & Schulman, 2010). Coreset applications are varied, and generic frameworks exist for their construction (Feldman & Langberg, 2011). Among the relevant recent applications are k-means and k-median clustering (Har-Peled & Mazumdar, 2004; Arthur & Vassilvitskii, 2007; Feldman et al., 2013; Bachem et al., 2018b), Bayesian inference (Camp-

bell & Broderick, 2018; Huggins et al., 2016), support vector machine training (Tsang et al., 2005), and neural network compression (Baykal et al., 2018).

While coresets are discrete, a sensitivity-based approach to importance sampling coresets was introduced in a continuous setting for approximating expectations under absolutely continuous measures with respect to the Lebesgue measure (Langberg & Schulman, 2010). For more information, see (Bachem et al., 2018b; Munteanu & Schwiegelshohn, 2018).

Another line of work closer to ours uses the theory of Reproducing Kernel Hilbert Spaces (RKHS) to design coresets, in particular kernel herding (Chen et al., 2010; Lacoste-Julien et al., 2015) and Stein points (Chen et al., 2018). These methods also take into account the underlying distribution of the data, but both require knowledge of that distribution (e.g., the density up to a normalising constant) while our approach simply assumes sample access.

Optimal transport (OT). The connection between optimal transport and quantization can be traced back to Pollard (1982), who studied asymptotic properties of k-means in the language of OT. More recently, Cuturi & Doucet (2014) proposed a more efficient version of transport-based quantization using entropy-regularised transport. Entropy-regularised transport (Cuturi, 2013) is a computationally efficient formulation of OT, which led to a wide range of machine learning applications; see recent surveys (Solomon, 2018; Peyré & Cuturi, 2018) for details. Recent results characterise its statistical behaviour (Genevay et al., 2019) and its ability to handle noisy datasets (Rigollet & Weed, 2018), which we can leverage to design robust coresets.

Our coreset construction algorithms are inspired by semi-discrete methods that compute transport from a continuous measure to a discrete one using power diagrams (Aurenhammer, 1987). Efficient algorithms that use computational geometry tools to perform gradient iterations to solve the Kantorovich dual problem have been introduced for 2D (Mérigot, 2011) and 3D (Lévy, 2015). Closer to our method are the algorithms by De Goes et al. (2012) and Claici et al. (2018), which solve a non-convex problem for the support of a discrete uniform measure that minimises transport cost to an input image (De Goes et al., 2012) or the barycenter of the input distributions (Claici et al., 2018). Stochastic approaches for semi-discrete transport, both standard and regularised, were tackled by Genevay et al. (2016).

4.2 Coresets: from discrete to continuous

4.2.1 Discrete coresets

A coreset is a *small summary* of a data set. *Small* usually refers to a the number of points in the coreset, which one hopes is much smaller than the data set size, but one can also think of this in terms of the number of bits required to store the coreset. The *summary* is often a weighted subset of the data, but can also refer to points that are not in the initial dataset but rather represent the original points well.

To make these notions more precise, we must define a coreset in terms of both the dataset and the cost function that the coreset is meant to perform well against. We can understand the definition as a learning problem, where our goal is to approximate the performance of a learning algorithm on a dataset X by its performance on the coreset C.

Let $\mathcal{F}$ be the hypothesis set for a learning problem. Every function $f \in \mathcal{F}$ maps from X to $\mathbb{R}$. Let μ_X be a weighting function on the points in X (this is typically uniform), and define the cost of f on (X, μ_X) as

$$\operatorname{cost}(X, \mu_X, f) = \sum_{x \in X} \mu_X(x) f(x). \tag{4.1}$$

A coreset is then defined by a set C and a weight function μ_C in such a way that $\operatorname{cost}(C, \mu_C, f)$ is close to $\operatorname{cost}(X, \mu_X, f)$. This leads to the following classical definition of a coreset (Bachem et al., 2017):

Definition 4.1 (Strong/weak ε-coreset). *The pair (C, μ_C) is a* strong *ε-coreset for the function family $\mathcal{F}$ if $C \subseteq X$ and*

$$\left|\operatorname{cost}(X, \mu_X, f) - \operatorname{cost}(C, \mu_C, f)\right| \leq \varepsilon \cdot \operatorname{cost}(X, \mu_X, f)$$

for all $f \in \mathcal{F}$. If we require that the inequality only holds at $f^* = \arg\min_{f \in \mathcal{F}} \operatorname{cost}(X, \mu_X, f)$, then we call (C, μ_C) a weak *ε-coreset.*

A coreset always exists for a dataset (X, μ_X) and family $\mathcal{F}$ as the original dataset (X, μ_X) satisfies Definition 4.1.

What distinguishes coresets from other notions of data sparsification is their dependence on the learn-

ing problem. For instance, there exist coresets for clustering (Bachem et al., 2018a,b), Bayesian inference (Campbell & Broderick, 2019), and classification (Baykal et al., 2017).

Example (k-means). The cost of a particular choice Q of k centres is given by $\sum_{x\in X} \min_{q\in Q} \|x-q\|^2$. To translate this into the language of Definition 4.1, we take $f_Q(x) = \min_{q\in Q} \|x-q\|^2$ and $\mu_X(x) = 1$ for all $x \in X$. The function family $\mathcal{F}$ is thus parameterised by the set of all possible choices of the centre set Q, and we wish to construct a coreset that performs well against all such choices (in the case of a strong coreset) or against the optimal k-means assignment (in the case of a weak coreset).

4.2.2 Measure coresets

So far we have used discrete language to describe coresets, but this belies the intent of coresets for learning problems. Typical learning problems are posed as minimisations in a hypothesis class of an *expectation* over a data distribution μ. The standard coreset definition is incompatible with this setting as it relies on the existence of a finite data set. To circumvent this issue, we define a *measure coreset* as a measure ν that produces similar results under $\mathcal{F}$ as μ:

Definition 4.2 (Measure Coreset). *We call ν a* strong *ε-measure coreset for μ if for all $f \in \mathcal{F}$*

$$\left|\mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(X)]\right| \leq \varepsilon. \tag{4.2}$$

In analogy to the discrete case, a *weak* ε-measure coreset is one for which the inequality holds at $f^* = \arg\min_{f\in\mathcal{F}} \mathbb{E}_\mu[f(X)]$. As in the case of discrete coresets, such a ν always exists, as $\nu = \mu$ satisfies the inequality.

Beyond the change to measure theoretic language, our definition differs from the typical coreset one in two ways. (1) The coreset ν can be an absolutely continuous measure, which means the size of the coreset can no longer be measured simply in the number of points. (2) We use absolute error instead of relative error; this connects our notion of coreset with generalisation error in learning problems in that we can see the coreset as *observed* data and the full measure as *out of sample* data. Absolute instead of relative error is uncommon in coreset language, but not unheard of; see (Reddi et al., 2015; Bachem et al., 2018a) for examples.

Under which constraints on ν, μ and $\mathcal{F}$ can we construct a measure coreset? We will show a connection to optimal transport and a resulting construction algorithm that aims at minimising a Wasserstein distance between the coreset ν and the target measure μ. Using optimal transport duality, we can qualify which learning problems admit measure coresets and the guarantees we can hope to achieve.

4.3 Sufficient conditions for coreset approximation

The link between our measure coreset formulation and the theory of optimal transport uses the notion of integral probability metrics (Müller, 1997):

Definition 4.3 (Integral Probability Metric). *Consider a class of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$. The integral probability metric $d_{\mathcal{F}}$ between two measures μ and ν is defined by*

$$d_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mu}[f(X)] - \mathbb{E}_{\nu}[f(X)] \right|. \tag{4.3}$$

Under mild assumptions on the set of functions $\mathcal{F}$, $d_{\mathcal{F}}$ defines a distance on the space of probability measures. We mention the following examples:

- 1-Wasserstein Distance: $\mathcal{F} = \{f \mid \|\nabla f\| \leq 1\}$ the space of 1-Lipschitz functions.
- Dual-Sobolev distance: $\mathcal{F} = \{f \mid \|f\|_{H^1(\mu)} \leq 1\}$ where H^1 is the Sobolev space $\{f \in L^2 \mid \partial_{x_i} f \in L^2\}$.
- Maximum Mean Discrepancy (MMD) (Gretton et al., 2007): $\mathcal{F} = \{f \mid \|f\|_{\mathcal{H}} \leq 1\}$ where $\mathcal{H}$ is a universal Reproducing Kernel Hilbert Space (RKHS).

The examples above allow us to derive a coreset condition for each of these function classes based on the Wasserstein distance or the MMD, explored in detail below.

Wasserstein distances. The p-Wasserstein distance between distributions μ and ν is given by the solution of a minimisation problem:

$$W_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p \, \mathrm{d}\pi(x, y), \tag{4.4}$$

where $\Pi(\mu,\nu) = \{\pi \in P(\mathcal{X}\times\mathcal{X}) \mid \pi(\mathrm{d}x\times\mathcal{X}) = \mu(\mathrm{d}x), \pi(\mathcal{X}\times\mathrm{d}y) = \nu(\mathrm{d}y)\}$ is the set of couplings with marginals μ and ν.

When $p = 1$, $W_1(\mu,\nu)$ can be rewritten via duality as a maximisation problem over the set of 1-Lipschitz functions (Santambrogio, 2015, §3.1). In particular, for $\mathcal{F} = \mathrm{Lip}_1(\mathcal{X})$,

$$d_{\mathcal{F}}(\mu,\nu) = \sup_{f\in\mathrm{Lip}_1} \int_{\mathcal{X}} f \,\mathrm{d}(\mu-\nu) = W_1(\mu,\nu).$$

When $p = 2$, $W_2(\mu,\nu)$ upper-bounds the dual Sobolev norm of $(\mu-\nu)$ if μ and ν have densities w.r.t the Lebesgue measure that are bounded above by some constant M. In particular, for any C^1 function f, define a semi-norm by

$$\|f\|_{H^1(\mu)} = \left(\int_{\mathcal{X}} |\nabla f(x)|^2 \,\mathrm{d}\mu(x)\right)^{\frac{1}{2}}.$$

This norm allows us to define a dual Sobolev norm on measures as

$$\|\nu\|_{H^{-1}(\mu)} = \sup_{\|f\|_{H^1(\mu)}\le 1} \int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x).$$

Using (Peyre, 2018, Equation (17)), we obtain that for $\mathcal{F} = \{f \mid \|f\|_{H^1(\mu)} \le 1\}$:

$$d_{\mathcal{F}}(\mu,\nu) = \|\mu-\nu\|_{H^{-1}(\mu)} \le \sqrt{M} W_2(\mu,\nu),$$

where M is the uniform bound on the densities of μ and ν.

Maximum mean discrepancy. When $\mathcal{F}$ is the unit ball of a RKHS, equation (4.3) defines a distance function known as the *maximum mean discrepancy* (Gretton et al., 2007). If $\kappa(\cdot,\cdot)$ is the reproducing kernel of the RKHS, we can rewrite (4.3) as an expectation over kernel evaluations

$$\begin{aligned}\mathrm{MMD}(\mu,\nu) =& \mathbb{E}_{\mu\otimes\mu}[\kappa(X,X')] + \mathbb{E}_{\nu\otimes\nu}[\kappa(Y,Y')] \\ &- 2\mathbb{E}_{\mu\otimes\nu}[\kappa(X,Y)].\end{aligned} \tag{4.5}$$

While our focus is on coresets under the Wasserstein distance, we mention that coresets that minimise the MMD have been constructed for kernel density estimation (Phillips & Tai, 2018). Generic construction algorithms for sampling to minimise MMD to a known fixed measure—known as *kernel herding*—have been given by Chen et al. (2010) and Lacoste-Julien et al. (2015).

Coreset condition. Using the properties of IPMs above, we summarise conditions for ν to be an ε-coreset for μ based on conditions on $\mathcal{F}$.

Proposition 4.1. *The measure ν is an ε-coreset for μ with function family $\mathcal{F}$ if:*

(i) $W_1(\mu, \nu) \leq \varepsilon$ *for* $\mathcal{F} \subseteq \mathrm{Lip}_1$*;*

(ii) $W_2(\mu, \nu) \leq \varepsilon/\sqrt{M}$ *for* $\mathcal{F} \subseteq H^1(\mu)$*, when μ and ν have densities with respect to the Lebesgue measure that are bounded above by M; or*

(iii) $\mathrm{MMD}(\mu, \nu) \leq \varepsilon$ *for* $\mathcal{F} \subseteq \mathscr{H}$.

We can extend the first two conditions to Lip_K and $\|f\|_{H^1(\mu)} \leq K$ by scaling f by the Lipschitz or Sobolev constant by a multiplicative K factor. In the remainder of this paper, we will focus on coresets based on Wasserstein distances and will call them *measure coresets* for simplicity. When more precision is required, we will denote by W_1 (resp. W_2, MMD) measure coreset a coreset with function family Lip_1 (resp. $H^1(\mu)$, $\mathscr{H}$).

4.4 Practical Wasserstein coreset constructions

While §4.3 gives a metric for measuring how close a distribution ν is to satisfying the coreset condition for a distribution μ, the question of how to compute such a ν remains.

In our definition, ν was unconstrained, but for it to be a useful coreset for a measure, we should be able to describe it using fewer bits than needed to describe the full measure μ. From a practical point of view, we should also be able to compute expectations under the coreset ν and at least approximate expectations under μ.

We make a few simplifications. We assume that we can sample from μ efficiently and that μ is supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$. This is true of any finite dataset. The simplest notion of a measure

coreset is a uniform distribution over a finite point set $x_1, \dots, x_n$. This leads to the following optimisation problem, which will be our focus in this section:

$$\min_{(x_1,\dots,x_n)} W_p\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \mu\right). \tag{$\mathscr{P}$}$$

It is also possible to formulate the problem using a continuous parametric density as a coreset. Given a family of parametric densities $(p_\theta)_{\theta\in\Theta}$ (e.g., Gaussian), we want to find the parametric distribution p_θ that best approximates a measure μ. This can be written simply as

$$\min_{\theta\in\Theta} W_p\left(p_\theta, \mu\right). \tag{4.6}$$

We experimented with this option using Gaussian mixtures, but the minimisation is highly non-convex, and gradient descent algorithms do not converge except in restricted settings (e.g., mixtures with equal weights). We find the simpler problem ($\mathscr{P}$) sufficient for the applications we consider and leave computation of more general coresets to future research.

4.4.1 Properties of empirical coresets

We address the problem of estimating n the number of points in a coreset n given ε for μ an arbitrary measure continuous. Namely, we ask how many samples n we need such that $W_p\left(\mu, \nu\right) \leq \varepsilon$ when $\nu = \sum_{i=1}^{n}\delta_{x_i}$.

Statistical bounds. There exist several theorems for finite sample rates of W_p, which each focus on specific hypotheses to marginally improve rates. We give a general statement:

Theorem 4.1 (Metric convergence, Kloeckner 2012; Brancolini et al. 2009; Weed et al. 2019). *Suppose μ is a compactly supported measure in $\mathbb{R}^d$ and ν_n is a uniform measure supported on n points drawn from μ. Then $W_p(\nu_n, \mu) \sim \Theta(n^{-1/d})$. Moreover, if μ has Hausdorff dimension $s < d$, then $W_p(\nu_n, \mu) \sim \Theta(n^{-1/s})$.*

Thus, both W_1 and W_2 have finite sample rate $O(n^{-1/d})$. If we assume that μ is supported on a lower dimensional manifold of dimension s, we get the improved rate $O(n^{-1/s})$.

Corollary 1. *If $\nu = \sum_{i=1}^{n}\delta_{x_i}$ with $n = \Theta(\varepsilon^{-s})$ is a globally optimal solution for ($\mathscr{P}$), then ν is a ε-measure*

coreset.

While we cannot guarantee this bound in practice since global optimality is NP-hard (Claici et al., 2018), empirically we observe that it holds and in fact is an overestimate of coreset size. Note that the theoretically required coreset size is independent of additional variables in the underlying problem, e.g., the number of means in k-means.

This bound improves over the best known deterministic coreset size for k-means and k-median of $O(k\varepsilon^{-d}\log n)$ (Har-Peled & Mazumdar, 2004), but we must be careful as our coreset bounds are given in absolute error. For k-means and k-medians, we are typically in the regime where the full data set has large cost (4.1), but if that does not hold, the coresets are no longer comparable.

Better randomised construction algorithms exist for both k-means/k-median and SVM with sizes that do not have such a strong dependence on dimension. Empirically, our coresets are competitive, and often better than specialised construction algorithms, especially in the small data regime (see Figures 4-3, 4-2 and 4-4).

One useful property of W_p coresets is that given an $\varepsilon-$coreset for a reference measure μ, we immediately have a $L\varepsilon-$coreset for the pushforward measure $f\#\mu$, where L is the Lipschitz constant of f.

Proposition 4.2. *(Coreset of pushforward measure) Consider a L-Lipschitz function f. If $\{x_i\}_{i=1}^n$ is a ε-measure coreset under W_p for μ, then $\{f(x_i)\}_{i=1}^n$ is a $L\varepsilon$-measure coreset under W_p for $f\#\mu$.*

Proof. f being L-Lipschitz implies that $\|f(x) - f(y)\|^p \leq L^p\|x-y\|^p \quad \forall (x,y) \in \mathcal{X}$. Thus, for all $\pi \in \Pi(\frac{1}{n}\delta x_i, \mu)$,

$$\begin{aligned}\int_{\mathcal{X}} \sum_{i=1}^n \|f(x_i) - f(x)\|^p \,\mathrm{d}\pi(x_i, x) \\ \leq L^p \int_{\mathcal{X}} \sum_{i=1}^n \|x_i - x\|^p \,\mathrm{d}\pi(x_i, x).\end{aligned}$$

Minimising over π on the right hand side and using the definition of a pushforward measure on the left gives

$$W_p^p\left(\frac{1}{n}\sum_{i=1}^n \delta_{f(x_i)}, f\#\mu\right) \leq L^p W_p^p\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \mu\right).$$

Since x_i is a W_p ε-measure coreset for μ, we have $W_p\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \mu\right) \leq \varepsilon$, yielding the desired bound. □

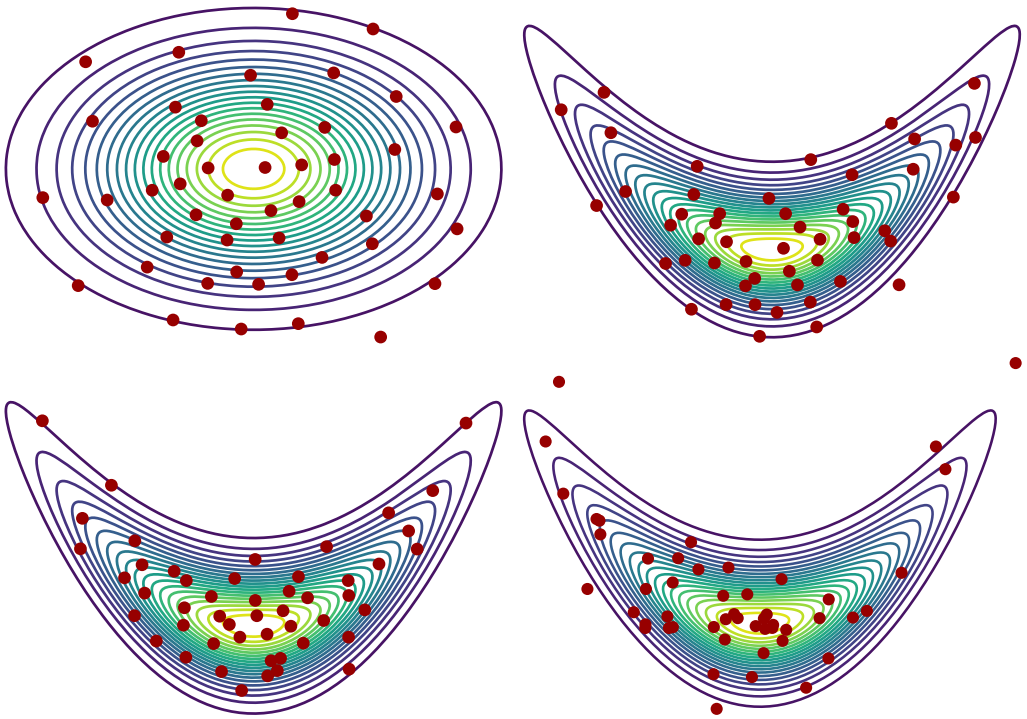


Figure 4-1: Coresets with 50 points for a Gaussian (top left) and the pushforward of a Gaussian through $f : (x, y) \mapsto (x, x^2+y)$. Top right is the image of the Gaussian coreset through f, bottom left is computed directly on the pushforward. A random sample is plotted bottom right.

Pushforward measures are ubiquitous in (deep) generative models, which have gained popularity for image generation through GANs Goodfellow et al. (2014) and VAEs Kingma & Welling (2014). Specifically, new data is generated by pushing uniform or Gaussian noise through a neural network f (Genevay et al., 2018). The above proposition suggests that if the pushforward function f has bounded variation, constructing a coreset for the source noise and pushing it through f is sufficient to find a 'good enough' coreset for the generative model without additional computations. This robustness property is illustrated by Figure 4-1, where the banana-shaped distribution is the pushforward of a normalised Gaussian $\mathcal{N}$ through $f : (x, y) \mapsto (x, x^2 + y)$. Even though the coreset obtained as the image of the coreset of the Gaussian through f performs slightly worse than the coreset computed directly on $f\#\mathcal{N}$, it represents the distribution in a more faithful way than a random sample.

We also have the following relationship between being a W_2 coreset and being a W_1 coreset:

Remark 1. *Let* $\{x_i\}_{i=1}^n$ *minimise* $W_2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \mu\right)$. *Using the inequality between* W_p *metrics,*

$$W_1\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \mu\right) \leq W_2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}\mu\right).$$

Thus, if we choose n *large enough such that* $W_2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \mu\right) \leq \varepsilon$, *then* $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ *is also a* W_1 ε*-measure coreset for* μ.

4.4.2 Entropy-regularised Wasserstein distances

The entropy-regularised Wasserstein distance is a popular approximation of the Wasserstein distance, as it is computable with faster algorithms (Cuturi, 2013). The entropically regularised p-Wasserstein distance is

$$W_{p,\eta}^p(\mu,\nu) = \operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p \mathrm{d}\pi(x,y) + \eta \mathrm{KL}(\pi\|\mu\otimes\nu). \tag{4.7}$$

As the KL term is nonnegative, $W_{p,\eta}^p$ upper-bounds W_p^p for all p, and thus any coreset under $W_{1,\eta}$ and $W_{2,\eta}$ is also a coreset under W_1 and W_2. Due to the entropic term, however, we have $W_{p,\eta}(\mu,\mu) = O(\eta)$ (Genevay et al., 2018), so even with a large number of samples n in the coreset, it is not always possible to get an ϵ-coreset for W_p using $W_{p,\eta}$. In practice, we observe that this regulariser yields mode collapse of the coreset, with the number of modes decreasing as η increases.

To alleviate this issue, Genevay et al. (2018) introduce Sinkhorn divergences, defined via

$$SD_{p,\eta}(\mu,\nu) = W_{p,\eta}(\mu,\nu) - \frac{1}{2}\left(W_{p,\eta}(\mu,\mu) + W_{p,\eta}(\nu,\nu)\right).$$

The additional terms ensure that $SD_{p,\eta}(\mu,\mu) = 0$. Interestingly, when η goes to infinity, Sinkhorn divergences converge to MMD defined in (4.5) with kernel $\kappa(x,y) = -\|x-y\|^p$ for $0 < p < 2$. While solving ($\mathcal{P}$) using $SD_{p,\eta}$ can be faster than with W_p, especially for larger coreset sizes, we do not have theoretical guarantees for the minimiser.

Algorithm 1 Compute an online W_1 coreset via SGD

Require: Measure μ, $n > 0$, mini batch size m, $\gamma > 0$
Ensure: Points $x_1, \dots, x_n$
1: Initialise $(x_1, \dots, x_n) \sim \mu$
2: for $k = 1, \dots$ do
3: Sample $(y_1, \dots, y_m) \sim \mu$
4: Update estimate of v^* using samples y_k.
5: Define generalised Voronoi regions $V_i(v^*)$.
6: Step: $x_i \leftarrow x_i - \frac{\gamma}{\sqrt{k}} \sum_{y_k \in V_i(v^*)} \frac{1}{|V_i(v^*)|} \frac{y_k - x_i}{\|y_k - x_i\|}$.

Algorithm 2 Compute an online W_2 coreset via SGD

Require: Measure μ, $n > 0$, mini batch size m, $\gamma > 0$
Ensure: Points $x_1, \dots, x_n$
1: Initialise $(x_1, \dots, x_n) \sim \mu$
2: for $k = 1, \dots$ do
3: Sample $(y_1, \dots, y_m) \sim \mu$
4: Update estimate of v^* using samples y_k.
5: Define generalised Voronoi regions $V_i(v^*)$.
6: Update: $x_i \leftarrow \sum_{y_k \in V_i(v^*)} \frac{1}{|V_i(v^*)|} y_k$

4.4.3 Algorithms

Recall that the goal of our measure coreset algorithms is to find a set of points $\{x_1, \dots, x_n\}$ that minimises some Wasserstein distance to a given distribution. Here, we detail how this goal is achieved by leveraging the dual of the Wasserstein problem. In particular, we give algorithms that compute coresets under the W_1 and W_2, via the updates specific to each setting.

Minimising W_1 and W_2. In the semi-discrete case, when $\nu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$, computing the Wasserstein distance can be cast as maximising an expectation:

$$W_p^p(\nu, \mu) = \max_{v \in \mathbb{R}^n} \mathbb{E}_\mu \Big[\min_i \left(\|X - x_i\|^p - v_i \right) + \frac{1}{n} \sum_{i=1}^{n} v_i \Big], \tag{4.8}$$

which can be optimised via stochastic gradient methods (Genevay et al., 2016; Claici et al., 2018). The gradients w.r.t. x_i can be written in terms of power diagrams:

$$\nabla_{x_i} W_1\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \mu\right) = \int_{V_i(v^*)} \frac{x - x_i}{\|x - x_i\|}\,\mathrm{d}\mu(x) \tag{4.9}$$

$$\nabla_{x_i} W_2^2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, \mu\right) = x_i - \int_{V_i(v^*)} x\,\mathrm{d}\mu(x) \tag{4.10}$$

where v^* is the solution of (4.8) and $V_i(v) = \{x : \|x - x_i\|^p - v_i \leq \|x - x_j\|^p - v_j, \forall j \neq i\}$ is the generalised Voronoi region of point x_i with $p = 1$ for W_1, and $p = 2$ for W_2.

Thus, a gradient step in the point positions x_i requires first solving (4.8) to get the optimal v, and then computing the gradients according to (4.9), (4.10). For W_2^2, the gradient step can be replaced by a fixed point iteration (Claici et al., 2018).

Minimising $W_{p,\eta}$ and $SD_{p,\eta}$. Due to the mode collapse inherent to large regularisation η mentioned in §4.4.2, Sinkhorn divergences empirically are better candidates to construct coresets. Following Genevay et al. (2018), we compute $\nabla_x SD_{p,\eta}$ using automatic differentiation of the objective. The resulting algorithm is identical to Algorithm 1, where $\nabla_x W_1$ gradient in line (6) is replaced by $\nabla_x SD_{p,\eta}$.

4.4.4 Convergence

We mention some observations on the convergence of our approach. The minimisation over the x variables is not convex due to inherent symmetries in the solution space, and $W_p(\cdot,\cdot)$ is not sufficiently smooth in the x variables to give precise convergence guarantees.

In Algorithms 1 and 2, we specify the number of points in the coreset. This parameter is unlike discrete coreset algorithms, which take ε as an input and return a coreset with enough points to satisfy the coreset inequality. Because our input is a measure that is absolutely continuous with respect to the Lebesgue measure, we do not have the luxury of this approach. An illustrative example is to consider $\varepsilon = 0$. In this case, a discrete coreset algorithm would simply return the original dataset. For a continuous μ, however, there is no finite distribution that has 0 error relative to μ.

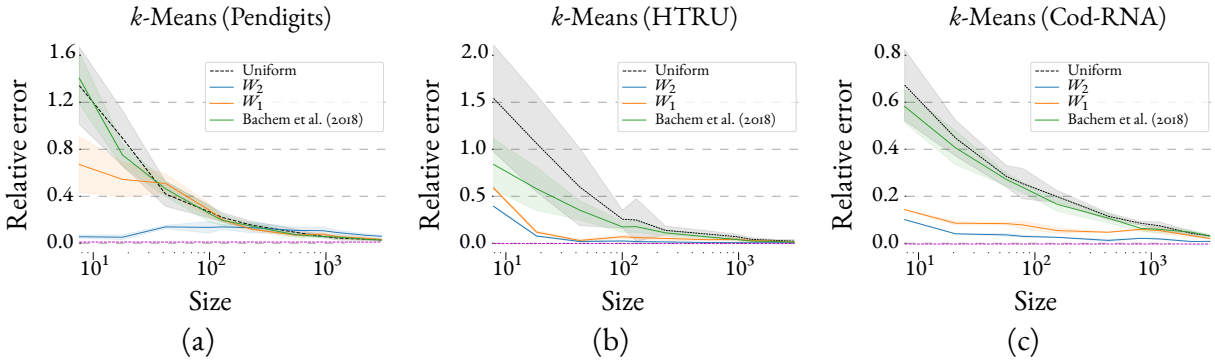


Figure 4-2: Coreset construction on for the k-means algorithm. We compute the k-means cost on the full data using means learned on the coreset. The y axis measures relative error to computing the cost using the means learned on the full data. Comparison is with Bachem et al. (2018a). We expect (and verify) that W_2 coresets perform better than W_1 coresets on this problem. (a) Pendigits dataset Keller et al. (2012); (b) HTRU dataset (Lyon et al., 2016); (c) Cod-RNA dataset (Uzilov et al., 2006)

4.4.5 Implementation details

Construction time depends strongly on the characteristics of the measure we are approximating. Most of the time is spent evaluating the expectations in (4.9), (4.10). Since we run the gradient ascent until $\|\nabla_w F\|_2 \leq \varepsilon$ and perform T fixed point iterations, the construction requires $O(T/\varepsilon)$ calls to an oracle that computes densities of the power cells $V_i(v)$.

The algorithms for W_1 and W_2 were implemented in C++ using the Eigen matrix library (Guennebaud et al., 2010) and run on an Intel i7-6700K processor with 4 cores and 32GB of system memory. Computing expectations under samples from μ can be trivially parallelised. The total coreset construction time ranges from a few seconds for small coresets on small datasets, to 5 minutes on large datasets where large coresets are required. The Sinkhorn divergence coresets were implemented in TensorFlow and run on the same architecture without GPU support. Since our code for W_p is in C++, we do not observe significant computational speedup when using Sinkhorn divergences in our experiments. As the resulting coresets are merely an approximation of W_p coresets, we do not display them in the experimental results.

All algorithms were run 20 times – we display the mean and standard deviations in our plots. Regarding the parameters in Algorithms 1 and 2, we use a step size $\gamma = 1$ and 100 iterations.

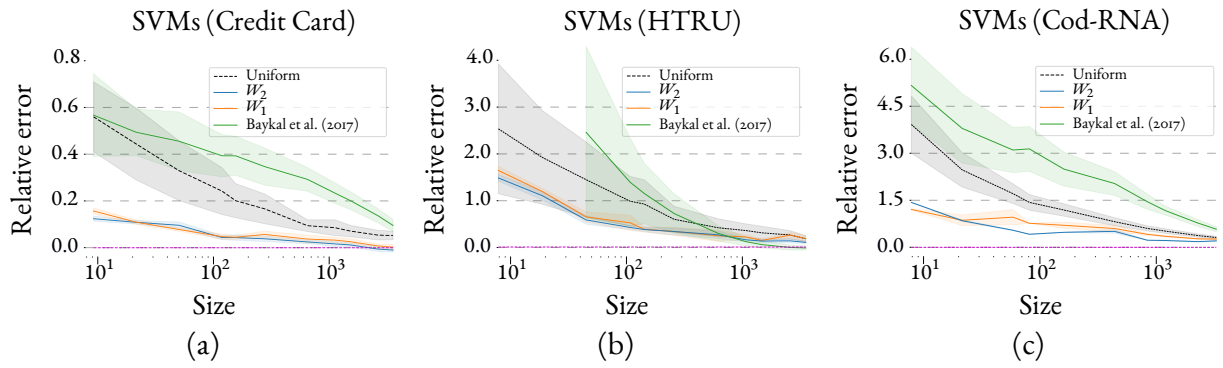


Figure 4-3: Coreset construction for SVM classification. We compute relative accuracy with respect to training a classifier on all the data. Comparison is with Baykal et al. (2017). Soft margin SVMs minimise a Lipschitz cost, and we expect both W_1 and W_2 coresets to perform well. (a) Credit card dataset (Yeh & Lien, 2009); (b) HTRU dataset (Lyon et al., 2016); (c) Cod-RNA dataset (Uzilov et al., 2006)

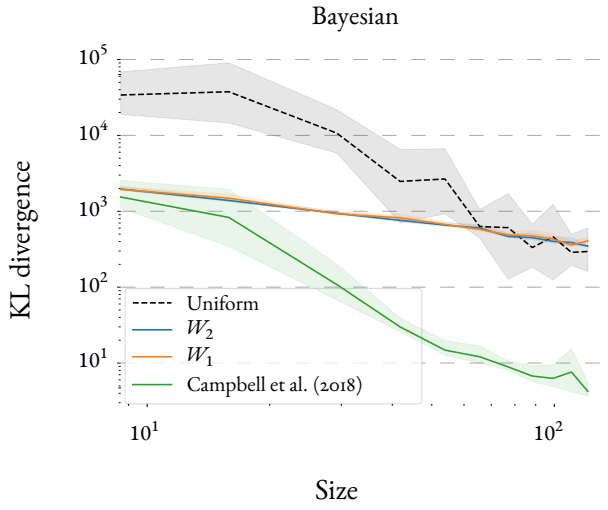


Figure 4-4: Coreset construction on a synthetic dataset (described in 4.5.3). The goal is to approximate the posterior distribution for a logistic regression model, and we report the KL divergence to the true posterior learned on the full data. Comparison is with Campbell & Broderick (2019). The log likelihood of the model is Lipschitz, and we expect similar performance from W_1 and W_2 coresets.

4.5 Comparison with classical coresets

We compare with classical coreset constructions on a few problems. Each of the three tasks we consider has a specialised coreset construction algorithm that does not extend to other problems. Our coresets, on the other hand, do not have this limitation, but broad applicability may come at the price of performance. Even so, our coresets perform better than uniform on the three tasks we have chosen (k-means clustering, SVM classification, posterior inference), and greatly *outperform* state-of-the-art algorithms for the first

two.

4.5.1 k-means clustering

The k-means objective for a fixed set of cluster centres Q is given by $J(Q) = \sum_{x \in X} \min_{q \in Q} \|x - q\|^2$.

When Q is a subset of a compact set, this cost has bounded Sobolev norm but is not Lipschitz. We expect W_1 coresets to perform worse than W_2 coresets on this problem. To measure performance, we compute coresets on the Pendigits dataset (Keller et al., 2012) and compute relative cost $1 - J(Q_c)/J(Q^*)$ of the centres learned on the coreset Q_c against the centres learned on the full data Q^*. We compare with the importance sampling method of Bachem et al. (2018a). The number of clusters we expect in the data is 10, one for each digit.

In this experiment, Bachem et al. (2018a) does not exhibit a clear advantage over uniform sampling. This suggests that their method is better suited to larger datasets. On the other hand, when using W_2 coresets, our method is on par with the minimal error for a coreset of 10 points. This is not surprising, as minimising ($\mathscr{P}$) with W_2 and $n = k$ support points is equivalent to minimising the k-means objective with balanced cluster assignments (Pollard, 1982; Cañas & Rosasco, 2012). This example demonstrates that our stochastic gradient descent approach is an efficient means of solving balanced k-means problems over large datasets, since we only access small-sized batches of the data at each iteration and never process the whole dataset at once.

4.5.2 Support vector machine classification

The soft margin SVM cost of a point x_i with label y_i is given by $y_i(w^\intercal x_i + b) - 1 + \xi_i$, where ξ_i is a slack variable associated to x_i. This cost is Lipschitz with a constant depending on the diameter of the set of allowable w's.

Because SVMs solve classification problems and our coresets approximate a dataset, our experimental setup here is slightly different than for k-means. Instead of constructing a coreset on the (x_i, y_i) pairs in the training data, we construct individual coresets for all data associated to a single label and merge them afterward. Hence, the coreset contains equal numbers of positive and negative samples. We hypothesise that this property and the tendency of coresets to remove large outliers explains why in Figure 4-3 our

coresets can yield better classifiers than training on the full data for large coreset size.

4.5.3 Bayesian inference

We construct a synthetic dataset for logistic regression by drawing $x_i \sim \mathcal{N}(0, I)$ and labelling the x_i by

$$\theta \sim \mathcal{N}(0, I) \qquad y_i \mid x_i, \theta \sim \text{Bern}\left(\frac{1}{1 + e^{-x_i^T \theta}}\right). \tag{4.11}$$

The goal is to construct a (weighted) coreset that approximates the log likelihood of the full data $\sum_i \log p(y_i|\theta)$. This cost is Lipschitz in this particular case. To agree with Campbell & Broderick (2019), instead of computing the relative log likelihood of our coreset against that of the full data, we use the coreset to infer the parameters of the posterior distribution and report KL divergence against the posterior learned on the entire dataset. Figure 4-4 shows results on a dataset of 20000 points drawn from a 5-dimensional Gaussian distribution. While we do not match the performance of Campbell & Broderick (2019), our coreset performs significantly better than a uniform sample.

4.5.4 Comparison with Kernel Herding

We have mentioned constructing coresets under the maximum mean discrepancy. Coresets under the MMD distance can be constructed using kernel herding, as shown in Chen et al. (2010); Lacoste-Julien et al. (2015). We give a qualitative comparison between W_2 coresets and samples obtained from herding on the mixture of Gaussian example from Chen et al. (2010) in Figure 4-5.

4.6 Discussion

Learning problems are frequently posed as finding the best hypothesis that minimises expected loss under a data distribution. However classic coreset theory ignores that the samples from the dataset are drawn from some distribution. We have introduced a notion of *measure coreset* whose goal is to minimise generalisation error of the coreset against the data distribution. Our definition is the natural one, and we can draw connections between this generalised notion of a coreset and optimal transport theory that leads to

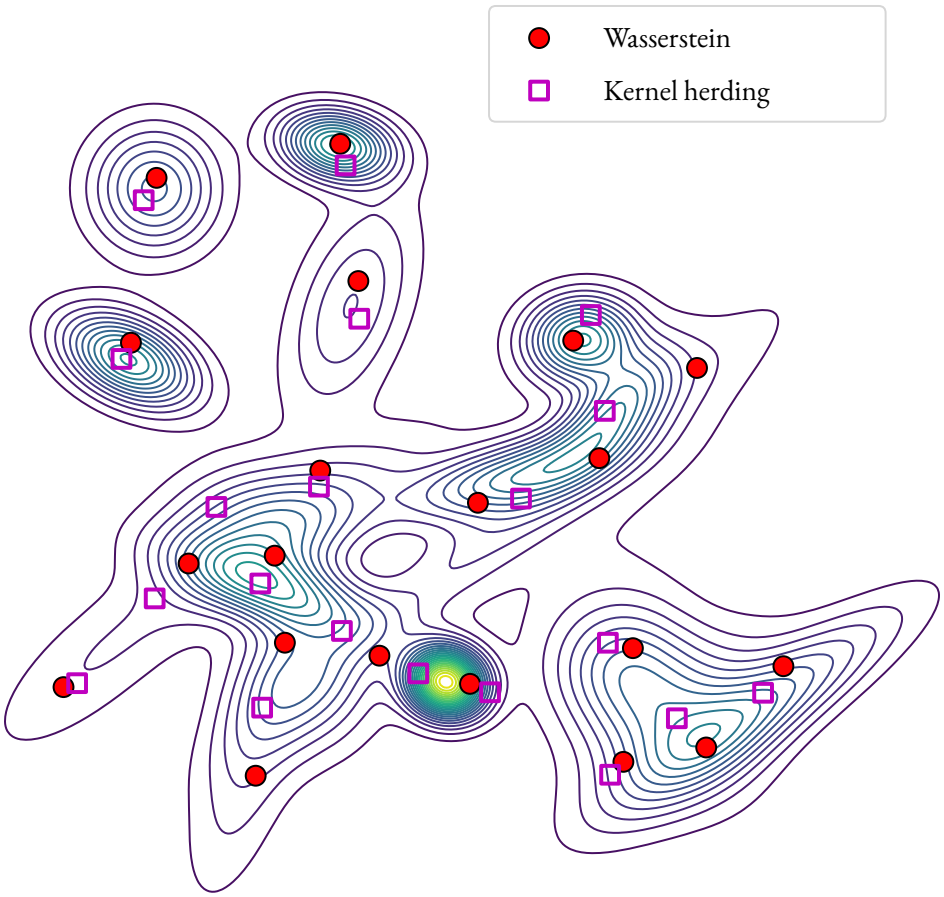


Figure 4-5: Comparison with kernel herding on a mixture of Gaussians. The first twenty points obtained from herding are plotted against a twenty point coreset under the W_2 distance.

online construction algorithms.

As our approach is exploratory, there are many avenues for future research. For one, our definitions rely on identities and inequalities that relate large function families to W_1 and W_2. If we cannot assume much about μ, then these relations cannot be refined. The theory in this chapter, however, does not sufficiently explain the effectiveness of our coreset constructions on the learning problems in §4.5.

Our algorithm's performance suggests several questions. There is a gap between the statistical knowledge we have about the sample complexity of W_1 and W_2 and the behaviour of Algorithms 1 and 2 in the few-samples regime. Additionally, a coreset condition similar to Proposition 4.1 for Sinkhorn divergences would allow us to leverage their improved sample complexity compared to Wasserstein distances, yielding tighter theoretical bounds for the number of points required to be an ϵ-measure coreset.

5

Stochastic Wasserstein Barycenters

The mean of a dataset often tells us something about the most likely value, and computing a mean value is central to much of learning theory. The notion of a mean in Euclidean space can be extended to means of distributions by posing it as a variational problem, but computing such means is challenging. In this chapter, we present an algorithm that recovers a quantization of the true barycenter.

5.1 Introduction

Several scenarios in machine learning require summarising a collection of probability distributions with shared structure but individual bias. For instance, multiple sensors might gather data from the same environment with different noise distributions; the samples they collect must be assembled into a single signal. As another example, a dataset might be split among multiple computers, each of which carries out MCMC Bayesian inference for a given model; the resulting "subset posterior" latent variable distributions must be reassembled into a single posterior for the entire dataset. In each case, the summarised whole can be better than the sum of its parts: noise in the input distributions cancels when averaging, while shared structure is reinforced.

The theory of *optimal transport* (OT) provides a promising and theoretically-justified approach to averaging distributions over a geometric domain. OT equips the space of measures with a distance metric known as the Wasserstein distance; the average, or *barycenter*, of a collection $\{\mu_j\}_{j=1}^N$ is then defined as a Fréchet mean minimising the sum of squared Wasserstein distances to the input distributions (Agueh &

Carlier, 2011). This mean is aware of the geometric structure of the underlying space. For example, the Wasserstein barycenter of two Dirac distributions δ_x and δ_y supported at points $x, y \in \mathbb{R}^n$ is a single Dirac delta at the centre point $\delta_{(x+y)/2}$ rather than the bimodal superposition $\frac{1}{2}(\delta_x + \delta_y)$ obtained by averaging algebraically.

If the input distributions are discrete, then the Wasserstein barycenter is computable in polynomial time by solving a large linear program (Anderes et al., 2016). Adding entropic regularisation yields elegant and efficient approximation algorithms (Genevay et al., 2016; Cuturi & Peyré, 2016; Cuturi & Doucet, 2014; Ye et al., 2017). These and other state-of-the-art methods typically suffer from any of a few drawbacks, mainly (1) poor behaviour as regularisation decreases, (2) required access to the distribution functions rather than sampling machinery, and/or (3) a fixed discretisation on which the input or output distribution is supported, chosen without knowledge of the barycenter's structure.

Given sample access to N distributions μ_j, we propose an algorithm that iteratively refines an approximation to the true Wasserstein barycenter. The support of our barycenter is adjusted in each iteration, adapting to the geometry of the desired output. Unlike most existing OT algorithms, we tackle the problem without regularisation, yielding a sharp result whose support is contained (to tolerance) within the support of the true barycenter even though we use stochastic optimisation rather than computational geometry.

Contributions. We give a straightforward parallelizable stochastic algorithm to approximate and sample from the Wasserstein barycenter of a collection of distributions, which does not rely on regularisation to make the problem tractable. We only employ samplers from the input distributions, and our technique is not restricted to input or output distributions supported on a fixed set of points. We verify convergence properties and showcase examples where our approach is inherently more suitable than competing approaches that require a fixed support.

5.2 Related work

OT has made significant inroads in computation and machine learning; see (Lévy & Schwindt, 2018; Peyré & Cuturi, 2018; Solomon, 2018) for surveys. Although most algorithms we highlight approximate

OT distances rather barycenters, they serve as potential starting points for barycenter computation.

Cuturi (2013) renewed interest in OT in machine learning through introduction of entropic regularisation. The resulting Sinkhorn algorithm is compact and efficient; it has been extended to barycenter problems through gradient descent (Cuturi & Doucet, 2014) or iterative projection (Benamou et al., 2015a). Improvements for structured instances enhance Sinkhorn's efficiency, e.g. via fast convolution (Solomon et al., 2015) or multiscale approximation (Schmitzer, 2016).

Our technique, however, is influenced more by *semidiscrete* methods, which compute OT distances to distributions supported on a finite set of points. Semidiscrete OT is equivalent to computing a power diagram (Aurenhammer, 1987; Aurenhammer et al., 1992), a generalisation of a Voronoi diagram whose cells receive the mass from each ∂. Algorithms by Mérigot (2011) in 2D and Lévy (2015) in 3D use computational geometry to extract gradients for the dual semidiscrete problem; Kitagawa et al. (2018) accelerate convergence via a second-order Newton method. Similar to our technique, De Goes et al. (2012) move the support of a discrete approximation to a distribution to reduce Wasserstein distance.

Recent stochastic techniques target learning applications. Genevay et al. (2016) propose a scalable stochastic algorithm based on the dual of the entropically-regularised problem; they are among the first to consider the setting of sample-based access to distributions but rely on entropic regularisation to smooth out the problem and approximate OT distances rather than barycenters. Staib et al. (2017) propose a stochastic barycenter algorithm from samples, but a finite, fixed set of support points must be provided a priori. Arjovsky et al. (2017) incorporate a coarse stochastic approximation of the 1-Wasserstein distance into a generative adversarial network (GAN); the 1-Wasserstein distance typically is not suitable for barycenter computation.

Further machine learning applications range from supervised learning to Bayesian inference. Schmitz et al. (2018) leverage OT theory for dictionary learning. Carrière et al. (2017) apply the Wasserstein distance to point cloud segmentation by developing a notion of distance on topological persistence diagrams. Courty et al. (2016) utilise the optimal transport plan for transfer learning on different domains. Srivastava et al. (2015a,b) use the Wasserstein barycenter to approximate the posterior distribution of a full dataset by the barycenter of the posteriors on smaller subsets; their method provably recovers the full posterior as the number of subsets increases.

5.3 Background and preliminaries

For measures $\mu_1, \ldots, \mu_N$, we can define the Wasserstein barycenter as the minimiser of the functional

$$F[\nu] = \frac{1}{N} \sum_{j=1}^{N} W_2^2(\nu, \mu_j). \tag{5.1}$$

When the input measures are discrete distributions, (5.1) is a linear program solvable in polynomial time.

If at least one of the measures μ_j is absolutely continuous with respect to the Lebesgue measure, then (5.1) admits a unique minimiser μ^* (Agueh & Carlier, 2011; Santambrogio, 2015). However, μ^* will also be absolutely continuous, implying that computational systems typically can only find an inexact finite approximation.

We study a discretization of this problem. Suppose $\Sigma \subset X$ consists of m points $\{x^i\}_{i=1}^m$, and define the functional

$$F[\Sigma] = \frac{1}{N} \sum_{j=1}^{N} W_2^2 \left(\frac{1}{m} \sum_{i=1}^{m} \delta_{x^i}, \mu_j \right). \tag{5.2}$$

We define the main problem.

Problem 1 (Semidiscrete approximation). *Find a minimiser of* $\Sigma \to F[\Sigma]$ *subject to the constraints* $\Sigma \subset X$, $|\Sigma| = m$.

Solving problem (1) for a single input measure is equivalent to finding the optimal m-point approximation to the input measure. We can use the solution as a set of supersamples from the input (Chen et al., 2010), or if the input distribution is a grayscale image, the solution yields a blue noise approximation to the image (De Goes et al., 2012).

5.4 Mathematical formulation

The OT problem (2.2) admits an equivalent dual problem

$$\sup_{f \in L^1(X)} \int_X f(x)\, \mathrm{d}\nu(x) + \int_X \overline{f}(y)\, \mathrm{d}\mu(y), \tag{5.3}$$

where f is the Kantorovich potential and $\overline{f}(x) := \inf_{y \in X}\{d(x,y)^2 - f(y)\}$ is the c-transform of f (Santambrogio, 2015; Villani, 2008).

Following Santambrogio (2015), if $\nu = \sum_{i=1}^m \frac{1}{m}\delta_{x^i}$ is a finite measure supported on $\Sigma = \{x^i\}_{i=1}^m$, then (5.3) becomes

$$\max_{v \in \mathbb{R}^m} \left\{ \sum_i \frac{1}{m} v^i + \int_X \overline{v}(y)\, \mathrm{d}\mu(y) \right\}, \tag{5.4}$$

where $v = (v^1, \dots, v^m)$. The key observation is that the function $f \in L^1(X)$ is replaced with a finite-dimensional $v \in \mathbb{R}^m$.

With this formula in mind, define

$$F_{\mathrm{OT}}[v, \Sigma; \mu] := \sum_i \frac{1}{m} v^i + \int_X \overline{v}(y)\, \mathrm{d}\mu(y). \tag{5.5}$$

Note that constant shifts in the v^i do not change the value of F_{OT}. F_{OT} has a simple derivative with respect to the v^i's:

$$\frac{\partial F_{\mathrm{OT}}}{\partial v^i} = \frac{1}{m} - \int_{V_v^i} \mathrm{d}\mu(y) \tag{5.6}$$

where V_v^i is the *power cell* of point x^i:

$$V_v^i = \{x \in X : d(x, x^i)^2 - v^i \le d(x, x^{i'})^2 - v^{i'}, \forall i'\}.$$

From here on we work with compact subsets of the Euclidean space $\mathbb{R}^D$ endowed with the Euclidean metric, $d(x,y) = \|x - y\|_2$. To differentiate with respect to the x^i's, notice that the first term in equa-

tion (5.5) does not depend on the positions of the points. We rewrite the second term as

$$\sum_{i=1}^{m} \int_{V_v^i} (d(y, x^i)^2 - v^i) \, \mathrm{d}\mu(y).$$

Using Reynolds' transport theorem to differentiate while accounting for boundary terms shows

$$\frac{\partial F_{\mathrm{OT}}}{\partial x^i} = x^i \int_{V_v^i} \mathrm{d}\mu(y) - \int_{V_v^i} y \, \mathrm{d}\mu(y). \tag{5.7}$$

Equation (5.6) confirms the intuition that each cell contains as much mass as its associated source point. We will leverage (5.7) to design a fixed-point iteration that moves each point to the centre of its power cell.

Each subproblem of (5.2) admits a different Kantorovich potential $v_j = (v_j^1, \dots, v_j^m)$, giving the following optimisation functional

$$F\left[\{v_j\}_{j=1}^N, \Sigma; \{\mu_j\}_{j=1}^N\right] = \frac{1}{N} \sum_{j=1}^{N} F_{\mathrm{OT}}[v_j, \Sigma; \mu_j] \tag{5.8}$$

Define

$$a_j^i = \int_{V_{v_j^i}} \mathrm{d}\mu(y) \qquad b_j^i = \frac{1}{a_j^i} \int_{V_{v_j^i}} y \, \mathrm{d}\mu(y).$$

With this notation in place, the partial derivatives are

$$\frac{\partial F}{\partial v_j^i} = \frac{1}{N}\left(\frac{1}{m} - a_j^i\right) \qquad \frac{\partial F}{\partial x^i} = \frac{1}{N} \sum_{j=1}^{N} a_j^i \left(x^i - b_j^i\right). \tag{5.9}$$

5.5 Optimisation

With our optimisation objective function in place, we now introduce our barycenter algorithm. To simplify nomenclature, from here on we refer to the dual potentials v_j as weights on the generalised Voronoi diagram. Our overall strategy is an alternating optimisation of F in (5.8):

- For fixed point positions, F is concave in the weights and is optimised using stochastic gradient ascent.

- For fixed weights, we apply a single fixed point iteration akin to Lloyd's algorithm (Lloyd, 1982).

5.5.1 Estimating Gradients

Each of a_j^i and b_j^i can be expressed as an expectation of a simple function with respect to the μ_j. We estimate these quantities by a simple Monte Carlo scheme.

In more detail, we can rewrite a_j^i and b_j^i as

$$a_j^i = \mathbb{E}_{y\sim\mu_j}\left[\mathbb{1}_{y\in V_{v_j}^i}\right] \qquad b_j^i = \mathbb{E}_{y\sim\mu_j}\left[y\cdot\mathbb{1}_{y\in V_{v_j}^i}\right].$$

Here, $\mathbb{1}$ indicates the indicator function of a set.

Since we have sample access to each μ_j, the expectations can be approximated by drawing K points independently $y_k \sim \mu_j$ and computing

$$\hat{a}_j^i = \frac{1}{K}\sum_{k=1}^{K}\mathbb{1}_{y_k\in V_{v_j}^i} \qquad \hat{b}_j^i = \frac{1}{K}\sum_{k=1}^{K} y_k\cdot\mathbb{1}_{y_k\in V_{v_j}^i}. \tag{5.10}$$

5.5.2 Concave Maximisation

The first step in our alternating optimisation maximises F over the weights v while the points x^i are fixed. We call this step of the algorithm an *ascent* step.

For a fixed set of points, the functional F is concave in the weights v_j, since it is the dual of the convex semidiscrete transport problem. To solve for the weights, we perform gradient ascent using the formula in (5.9) where a_j^i is approximated using $\hat{a}_j^i$. Note that the gradient for a set of weights v_j only requires computation of the density of a single measure μ_j, implying that the ascent steps can be decoupled across different measures.

Write $w^0 = v_j$ for the initial iterate. The simplest version of our algorithm updates

$$w^{k+1} = w^k + \alpha\frac{\partial F}{\partial v_j}[w^k].$$

The iterates converge when each point contains equal mass in its associated power cell.

F has a known Hessian as a function of the v_j that can be used in Newton's algorithm (Kitagawa et al., 2018). Computing the Hessian, however, is only possible with access to the density functions of the μ_j's as it requires computing a density of the measure on the boundary between two power cells. The boundary set is inherently lower dimensional than the problem space, and hence sample access to the μ_j is insufficient. Moreover, even had we access to the probability density functions, computing the Hessian would require the Delaunay triangulation of the point set, which is expensive in more than two dimensions.

In any event, choosing the step size α is important for convergence. Line search is difficult as we do not have access to true objective value at each iterate. Instead, we rely on Nesterov acceleration to improve performance (Nesterov, 1983). With acceleration, our iterates are

$$z^{k+1} = \beta z^k + \frac{\partial F}{\partial v_j}[w^k] \tag{5.11}$$

$$w^{k+1} = w^k + \alpha z^{k+1}. \tag{5.12}$$

In our experiments, we use $\alpha = 10^{-3}$ and $\beta = 0.99$. Convergence of the accelerated gradient method can be shown when $\alpha = 1/L$ where L is the Lipschitz constant of F; in §5.6, we give an estimate of this constant. Our convergence criterion for this step is $\|\nabla F\|_2^2 \leq \epsilon$.

5.5.3 Fixed Point Iteration

The second step of our optimisation is a fixed point iteration on the point positions. This step is similar to the point update in a k-means algorithm in that it snaps points to the centres of local cells, and we refer to it as a *snap* step.

To derive the iteration, we set the second gradient in (5.9) to zero:

$$\frac{\partial F}{\partial x^i} = 0 \Longrightarrow \frac{1}{N}\sum_{j=1}^{N} a_j^i (x^i - b_j^i) = 0$$

Algorithm 3 Adding a point to the current barycenter estimate Σ.

for $t = 1, 2, \dots, T$ do
  for $j = 1, 2, \dots, J$ do
    $z^0 \leftarrow 0$ ▷ Ascent on weights
    $w^0 \leftarrow v_j$
    while $\left\| \frac{\partial F}{\partial v_j} \right\| > \epsilon$ do
      Compute $\hat{a}_j^i$ according to equation (5.10)
      $z^{k+1} = \beta z^k + \frac{\partial F}{\partial v_j}[w^k]$
      $w^{k+1} = w^k + \alpha z^{k+1}$
    $v_j \leftarrow w^{\text{end}}$
  Compute $\hat{b}_j^i$ according to equation (5.10)
  for $x_i \in S$ do
    $x_i \leftarrow \frac{\sum_{j=1}^N \hat{a}_j^i \hat{b}_j^i}{\sum_{j=1}^N \hat{a}_j^i}$ ▷ Snap points

which leads to the point update

$$x^i = \frac{\sum_{j=1}^N a_j^i b_j^i}{\sum_{j=1}^N a_j^i}. \tag{5.13}$$

This suggests a fixed point iteration for the x^i's that can be decomposed into the following steps:

1. First find the barycenter of the power cells of each x^i with respect to each μ_j.

2. Then, average the points with weights given by the density of each measure in the cell.

If the concave maximisation has converged appropriately, and uniform areas a_j^i have been achieved, then the update step becomes a uniform average over the barycenters b_j^i with respect to each measure.

5.5.4 Global and Local Strategies

The *ascent* and *snap* steps can be used to refine a configuration of points Σ. Once the iterates converge, we have an m-point approximation to the barycenter that can be used as an initialisation for $m + 1$ point approximation in two ways. A new point x is sampled uniformly from X, and then we have a choice between (1) moving all points including the new one or (2) allowing only x to move.

These two approaches are codified in Algorithm 3 where the choice on the set S dictates which points move. The number of iterations of the outer loop is fixed beforehand. Typically, we see convergence in fewer than 20 steps, and empirically, we observe good performance even with $T = 1$. The two most natural choices for S are $S = \Sigma$ and $S = \{x\}$. If the barycenter is absolutely continuous with respect to the underlying Lebesgue measure, these two strategies converge at the same rate asymptotically (Brancolini et al., 2009). The latter, however, can generate spurious samples that are not in the support of the barycenter. Note that optimising the weights is regardless a global problem as moving or introducing just one point can change the volumes of the power cells of neighbouring points.

Both algorithms are highly parallelisable, since (1) the gradient estimates are expectations computed using Monte Carlo integration and (2) the gradient step in the weights decouples across distributions.

5.6 Analysis

We justify the use of uniform finitely-supported measures, and then prove that our algorithm converges to a minimum cost under mild assumptions.

We assume in this section that at least one of the distributions μ_j is absolutely continuous with respect to the Lebesgue measure, ensuring a unique Wasserstein barycenter.

5.6.1 Approximation Suitability

The simplest approach for absolutely continuous measures $\mu_j \in \mathcal{P}(X)$ is to sample p points from each of the J measures and solve for the true barycenter of the empirical distributions (Anderes et al., 2016). This approach likely approximates the barycenter as the number of samples increases, but requires solution of a linear program with $O(p^J)$ variables. As an alternative, Staib et al. (2017) propose a stochastic problem for approximating barycenters. They are able to prove a rate of convergence, but the support of their approximate barycenter is fixed to a finite set of points.

Our technique allows the support points to move during the optimisation procedure, empirically allowing a better approximation of the barycenter with fewer points. The following theoretical result shows that the use of uniform measures supported on a finite set of points can approximate the barycenter

arbitrarily well:

Theorem (Metric convergence, (Kloeckner, 2012; Brancolini et al., 2009)). *Suppose ν_m^* is a uniform measure supported on m points that minimises $\frac{1}{N}\sum_{j=1}^{N} W_2(\nu_m^*, \mu_j)$, and let $\bar{\mu}$ denote the true barycenter of the measures $\{\mu_j\}_{j=1}^N$. Then $W_2(\nu_m^*, \bar{\mu}) \leq Cm^{-1/D}$ where C depends on the underlying space X, the dimension D, and the metric $d(\cdot,\cdot)$.*

Note that this shows convergence in probability $\nu_m^* \rightharpoonup \bar{\mu}$ since the Wasserstein distance metrises weak convergence (Villani, 2008). Brancolini et al. (2009) also show asymptotic equivalence of the local and global algorithms.

While we cannot guarantee that our method converges to ν_m^*, these properties indicate that the *global* minimiser of our objective provides an effective approximant to the true barycenter as the number of support points $m \to \infty$.

5.6.2 Algorithmic Properties

The functional F is concave in the weights v_i^j with fixed point positions. We can investigate the convergence properties of the gradient ascent step of the algorithm. Specifically, what we are after is a Lipschitz constant for the gradient of F with respect to the weights. We will show that this does not hold generally.

Counterexample. *Assume X is a compact subset of $\mathbb{R}^D$. There are measures $\mu \in \mathcal{P}(X)$ for which the gradient of F is not Lipschitz continuous. A set of weights that satisfies $\frac{\partial F}{\partial v} = 0$ may not exist, and if it does, it may not be unique.*

Construction. We provide a counterexample for $D = 1$. Let $X = [-1, 1]$ with the standard metric and $\mu = \delta_0$. Let $\Sigma = \{-1, 1\}$ be the fixed positions, and take $v_1 = \{-\epsilon, 0\}$ and $v_2 = \{\epsilon, 0\}$ for small ϵ. Then $\|v_1 - v_2\|_1 = 2\epsilon$, but $\|\nabla F_v[v_1] - \nabla F_v[v_2]\|_1 = 2$.

Non-existence is shown in Figure 5-1. To see non-uniqueness, take $\mu = \frac{1}{2}\delta_{-\epsilon} + \frac{1}{2}\delta_{\epsilon}$ with Σ as before. Any set of weights in $(-\epsilon, \epsilon)^2$ minimises F_v. □

For mildly behaved measures μ the gradient of F with respect to v is Lipschitz continuous:

Lemma. *Assume X is a compact subset of $\mathbb{R}^D$, and μ is absolutely continuous with respect to the Lebesgue measure, with density function ρ. If the m points of Σ are distinct and $\rho \leq M$ almost everywhere for some*

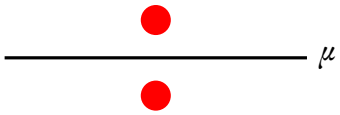


Figure 5-1: Non-existence of a set of weights. Let μ be the uniform measure on the line segment, and Σ be the two red points such that the line between them is orthogonal to the support of μ. There is no set of weights such that the mass of μ is split evenly between the two red points.

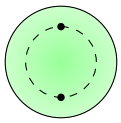


Figure 5-2: Non-unique minimiser on two points for the uniform measure defined on the unit disk. All antipodal points on the dashed circle at distance $2/\pi$ from the centre are valid minimisers.

constant M, then:

$$\|\nabla F_v[v_1] - \nabla F_v[v_2]\|_2 \leq \sqrt{m}\frac{MS}{2L}\|v_1 - v_2\|_2.$$

where S denotes the surface area of $\partial\mathrm{conv}(X)$ and L denotes the minimum pairwise distance between points in Σ.

Proof. Consider the ith component of the gradient difference:

$$\begin{aligned}\left|\frac{\partial F_v}{\partial v^i}[v_1] - \frac{\partial F_v}{\partial v^i}[v_2]\right| &= \left|\int_{V^i_{v_1}} \rho \,\mathrm{d}\lambda - \int_{V^i_{v_2}} \rho \,\mathrm{d}\lambda\right| \\ &\leq \frac{S\|v_1 - v_2\|_2}{2L} M.\end{aligned}$$

The second inequality follows as the area of a power cell is bounded by S and the faces of the cells change at a rate linear in $\|v_1 - v_2\|_2$. The rate is dependent on the distance between the points, so the constant L is required. The Lipschitz bound follows directly from considering all components of the gradient difference together. □

This lemma applies convergence for a step size that is the inverse of the Lipschitz constant. While the above requires absolute continuity of μ, we have found that our ascent steps and method often converge even when this is not satisfied (see Figures 5-4 and 5-6).

We may also show that our algorithm monotonically decreases $F[\Sigma]$ (defined in Equation (5.2)) after each pair of snap and then ascent steps for compact domain and absolutely continuous μ_j. For this purpose, recall that the transport cost for a map $T : X \to \Sigma$ sending measure μ_j to $\frac{1}{m}\sum_i \delta_{x^i}$ is:

$$\int_X d(x, T(x))^2 \, d\mu_j.$$

Fixing the power cells V_j^i after an ascent step, we may define $T_j(\Sigma)$ to be the transport cost for the map sending the power cells V_j^i to the point set Σ, and we may define $TC(\tilde{\Sigma}) = \frac{1}{N}\sum_j T_j$ to be the joint (average) transport cost. Letting $\tilde{\Sigma} = \{\tilde{x^i}\}$ denote the new positions after a snap step, we may now show:

Lemma. *For $X \subset \mathbb{R}^D$ compact, and μ_j absolutely continuous with respect to the Lebesgue measure for all j:*

$$F[\tilde{\Sigma}] \leq F[\Sigma].$$

Proof. By strong duality, we have the following equality for each j when the v have been optimised after an ascent step:

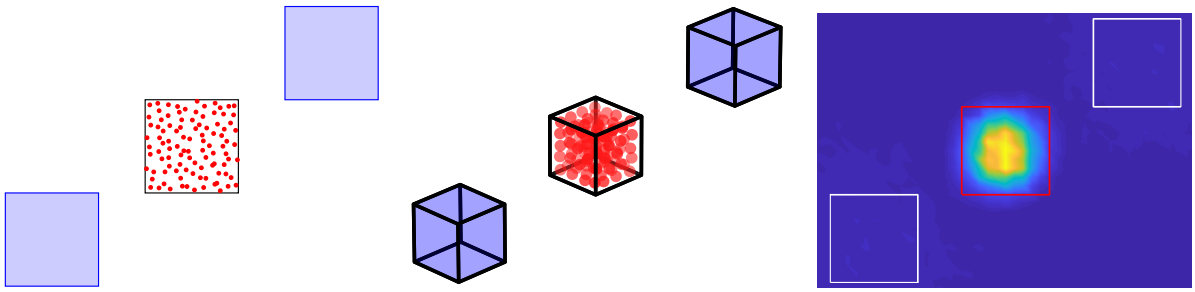
$$F_{OT}[v, \Sigma; \mu_j] = W_2^2\left(\frac{1}{m}\sum_{i=1}^{m} \delta_{x^i}, \mu_j\right).$$

This implies that $F[\Sigma] = TC(\Sigma)$ as W_2^2 is simply the optimal transport cost. We now argue that $TC(\tilde{\Sigma}) \leq TC(\Sigma)$. We may split up the integrals for transport cost over the power cells corresponding to each ith point. We differentiate $\sum_{j=1}^{N} \int_{V_j^i} \|x - p\|^2 \, d\mu_j$ with respect to p to find the point with lowest joint transport cost to the cells V_j^i. Setting this to o yields the following:

$$\sum_{j=1}^{N} a_j^i b_j^i - a_j^i p = 0$$

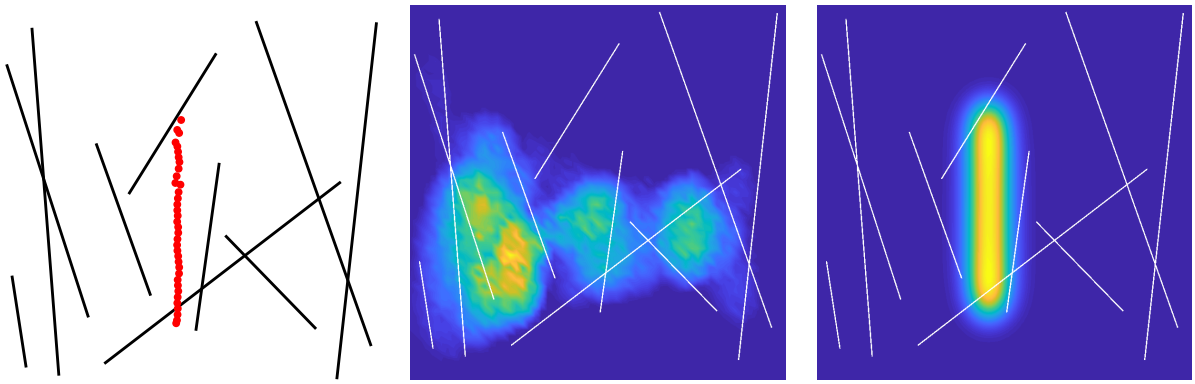
Note this is equivalent to the barycenter update step in Equation (5.13), and with convergence of the previous ascent step, we should have uniform a_j^i weights. This demonstrates that snapping to the uniform average of barycenters lowers TC, and we have that $F[\Sigma] = TC(\Sigma) \geq TC(\tilde{\Sigma}) \geq F[\tilde{\Sigma}]$. The last inequality follows as the next ascent step will find the optimal transport and decrease the transport cost. □

With joint transportation cost being non-negative, this implies that our objective function converges



(a) (b) (c)

Figure 5-3: Barycenter when $N = 2$ tested on two uniform distributions over unit squares. (a) Our output: the input distributions are shown in blue, while the output barycenter points are shown in red, with the limits of the true barycenter in black. (b) A similar example in three dimensions. (c) The output barycenter of Staib et al. (2017): note the output has non-zero measure outside the true barycenter.



(a) (b) (c)

Figure 5-4: Barycenter of sharp featured distributions. (a) 50 points from our algorithm yields a barycenter supported on a line. (b) The barycenter from Staib et al. (2017) using a grid of 20000 points. (c) Barycenter from Solomon et al. (2015) using a regulariser value of $\gamma = 0.1$; smaller regularisers were numerically unstable.

to a local minimum. This does not imply that our iterates converge, as there may not be a unique minimising point configuration (see Figure 5-2). Empirically, our iterates converge in all of our test cases. We note also that our formula bears some resemblance to the mean-shift algorithm and to Lloyd's algorithm, both of which which are also known to converge under some assumptions (Li et al., 2007; Bottou & Bengio, 1995).

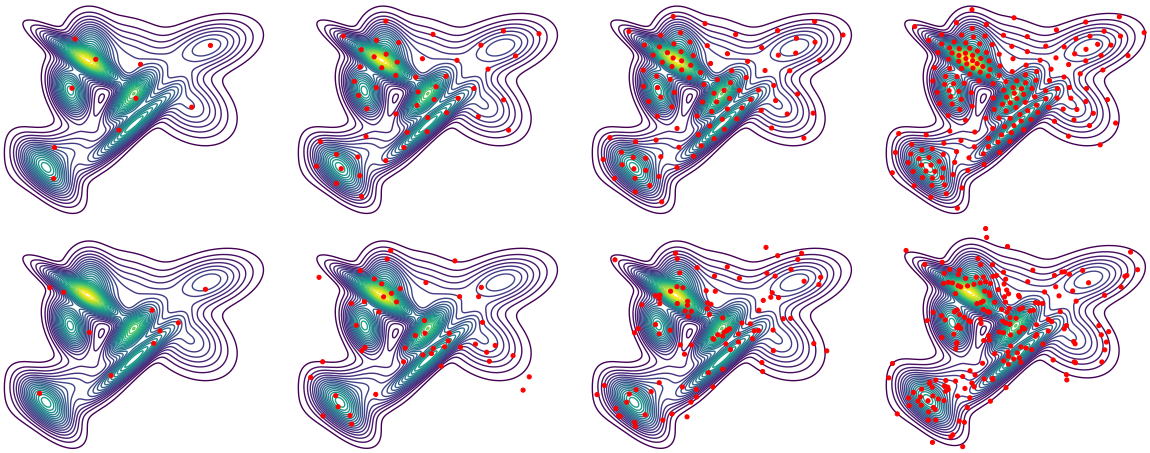


Figure 5-5: The n point approximation of a mixture of ten Gaussians. Top row: our method with 10, 50, 100, and 200 points. Bottom row: iid sampling with the same number of points.

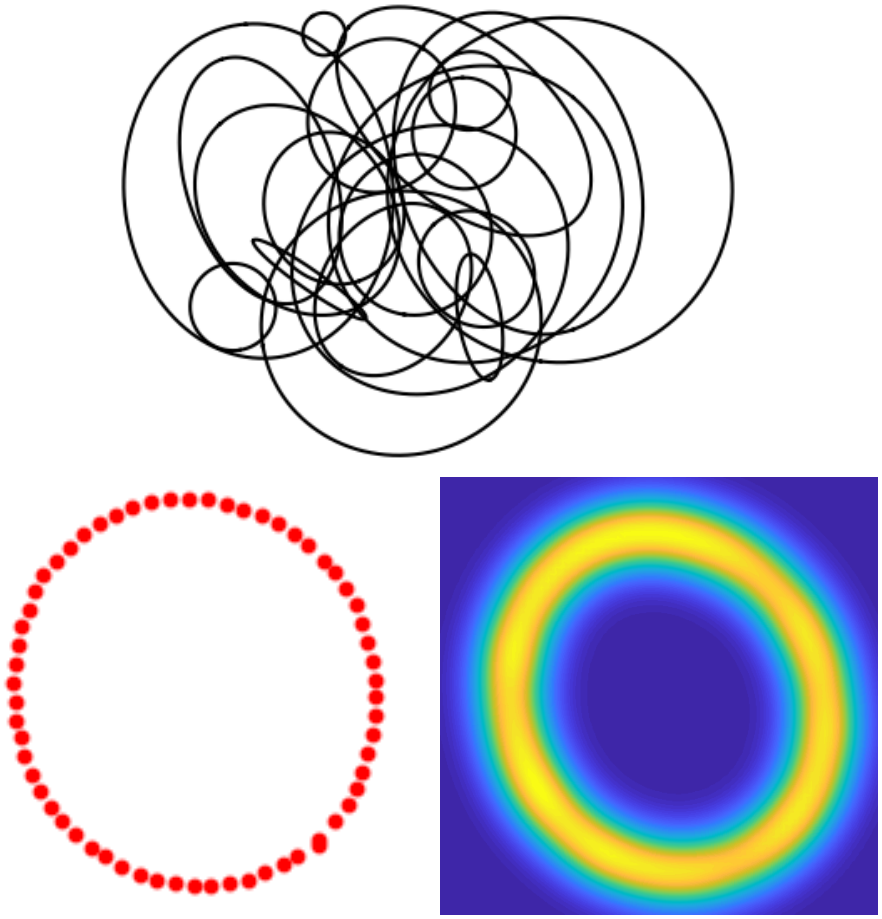
5.7 Experiments

We showcase the versatility of our method on several applications. We typically use between 16K and 256K samples per input distribution to approximate the power cell density and barycenter. The variance is due to different problem sizes and dimensionality of the input measures. We stop the gradient ascent step when $\|\nabla F\|_2^2 \leq 10^{-6}$. The snap step empirically converges in under 20 iterations, and several of our examples use only one step.

5.7.1 Distributions with Sharp Features

Our algorithm is well-suited to problems where the input distributions have very sharp features. We test against the algorithms in Staib et al. (2017) and Solomon et al. (2015) on two test cases: ten uniform distributions over lines in the 2D plane (Figure 5-4), and 20 uniform distributions over ellipses (Figure 5-6).

The results of Figures 5-4 and 5-6 show that our barycenter is more sharply supported than the results of competing methods. Our output agrees with that of Solomon et al. (2015), but our results more closely match expected behaviour. We strongly suspect that the true barycenter in Figure 5-4 is also a uniform measure on a line, while that in Figure 5-6 is a circle centred at the origin.



(a) (b)

Figure 5-6: Barycenter of randomly generated ellipses. Top: plot showing 20 ellipses with randomly drawn centre, semi-major and semi-minor axes, and skew. Bottom: (a) The output of our algorithm is a sharp distribution approximating a circle. (b) The output of Solomon et al. (2015) with a regulariser value of $\gamma = 0.1$.

5.7.2 The Case $N = 2$

In the case of two input measures μ_1 and μ_2, we expect the barycenter to be McCann's interpolant (Agueh & Carlier, 2011; McCann, 1997):

$$\mu_{1/2} := \left(\frac{1}{2}\mathrm{id} + \frac{1}{2}T\right)_{\#}\mu_0 = \left(\frac{1}{2}\mathrm{id} + \frac{1}{2}T^*\right)_{\#}\mu_1$$

where T is the optimal map, and T^* is the inverse map, while $\#$ denotes the pushforward of a measure.

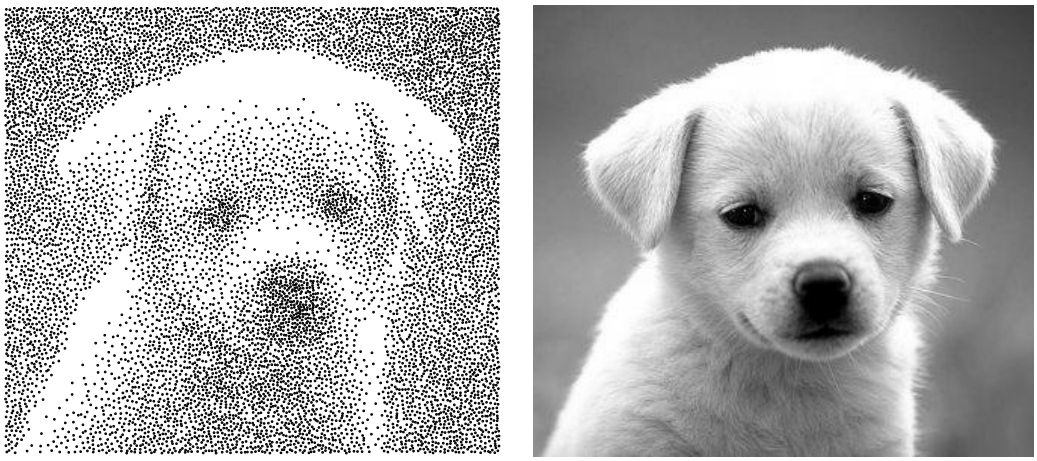


Figure 5-7: Blue noise sampling. Left: 10K samples from our algorithm. Right: Original image (approximately 90K pixels).

We test this on two uniform distributions on the unit square in Figure 5-3. The transport map in this case is transport of the entire distribution along a straight line. As expected from McCann's interpolant, we recover a uniform distribution on the unit square halfway between the two input distributions. We show our results alongside those of Staib et al. (2017). Notice that their output barycenter is not uniform, and that it has non-zero measure outside the true barycenter.

5.7.3 The Case $N = 1$

The case $N = 1$ bears interest as well. There are instances when sampling iid from a distribution yields samples that do not approximate the underlying distribution accurately. We showcase two applications in generating super samples from distributions, as well as approximating grayscale images through blue noise.

Super Samples Our method can be adapted to generate super samples from complex distributions (Chen et al., 2010). Figure 5-5 details our results on a mixture of ten Gaussians. Our method better approximates the shape of the underlying distribution due to negative autocorrelations: points move away from oversampled regions. The points drawn iid from the mixture tend to oversample around the larger modes and do not approximate density contours as well.

Blue Noise The term blue noise refers to an unstructured but even and isotropic distribution of points. It has been used in image dithering as it captures image intensity via local point density, without the need for varying point sizes as in halftoning.

De Goes et al. (2012) described the link between optimal transport and blue noise generation. We recover a stochastic version of their algorithm by taking μ a discrete distribution over the image pixels proportional to intensity. As our method is more general, we observe performance loss, but the output is of comparable quality (Figure 5-7).

5.8 Conclusion

We have proposed an algorithm for computing the Wasserstein barycenter of continuous measures using only samples from the input distributions. The algorithm decomposes into a concave maximisation and a fixed point iteration similar to the mean-shift and k-means algorithms. Our algorithm is easy to implement and parallelise, and it does not rely on a fixed-support grid. This allows us to recover much sharper approximations to the barycenter than previous methods. Our algorithm is general and versatile enough to be applied to other problems beyond barycenter computation.

There are several avenues for future work. Solving the concave maximisation problem is currently a bottleneck for our algorithm as we do not have access to the function value or the Hessian, but we believe multiscale methods can be adapted to our approach. The potential applications of this method extend beyond what was covered. One application we highlight is in developing coresets that minimise the distance to the empirical distribution on the input data.

6

Quantization: Discussion

We have seen how the simple idea of approximating a measure by a finite set of samples has deep implications with regards to learning theory, approximation, and summarisation.

At the core of these approaches lies the problem

$$\min_{x_1,\dots,x_n} W_p\left(\mu, \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}\right).$$

If the finite approximation is a good proxy for the measure μ, then many problems that had to deal with μ are now tractable, from Bayesian inference to classification methods. The strength of this approach is that once such an approximation has been computed, even simple algorithms can produce state-of-the-art results thus taking the burden away from the practitioner.

Where do we go from here? The algorithms in Chapter 4 and Chapter 5 rely on efficiently solving the semi-discrete transport problem, but this is often hard. How we get around this problem is answered by the following two chapters. In Part II we show how structure inherent in a learning problem can be used to speed up computation of the optimal transport cost, while in Part III we present an algorithm for computing the optimal transport cost in situations where the ground metric is costly to calculate.

Part II

Hierarchical Structure

7

Introduction to Hierarchical Structure

Wherein we solve the transportation faster by using thematic or hierarchical structure in the data. Applications to natural language processing and Bayesian inference. A discussion follows on how this approach generalises to much more.

Optimal transport is, by nature, a theory that operates at a fine-grained level. It deals with distributions by allocating resources from individual *points* to individual *points*. In the absence of additional information, this fine-grained approach is all we can work with, and much research has focused on how to make such algorithms faster.

However, distributional data often comes with additional structure which is ignored by these approaches. In this part we show examples of this structure that leads to significantly faster algorithms, better results, and more coarse-grained interpretability.

We can illustrate this with an example. Suppose someone asks us to compare two classics of American literature: Herman Melville's *Moby Dick*, and Nathaniel Hawthorne's *The Scarlet Letter*. If we treat this as an optimal transport problem, the atoms we can manipulate are individual words, and the mass at each atom is akin to word frequency. But no human being would compare these two novels by splitting hairs on each individual word.

What the algorithm is missing is thematically coherent regions of the text. When we say that both *Moby Dick* and *The Scarlet Letter* are novels of vengeance, we pack a lot of significance into that single word which a computer is oblivious towards.

We show how to use themes in optimal transport to both understand data better, and improve perfor-

mance on document-to-document distance computation in Chapter 8. At the heart of this approach is a Wasserstein distance on the space of distributions whose atoms are themselves distributions, hence the *hierarchical structure* of the title. This approach is most advantageous when the coarse-grained structure does not change significantly over time.

While the problem in Chapter 9 is completely different, we recognise the same patterns, and show how a hierarchical approach can alleviate a long standing problem in Bayesian inference.

This part is based on Yurochkin et al. (2019b) and Monteiller et al. (2019).

8

Hierarchical Optimal Topic Transport

Natural language documents have always been a challenging test bed for learning algorithms, and for good reason: semantics can change based on context, coarse grained structures such as narratives or themes are hard to codify, etc. In this chapter, we show how to compare documents using coarse level thematic structure by leveraging optimal transport and topic models. This leads to a simple, yet effective and fast algorithm that is easy to interpret.

8.1 Introduction

Topic models like latent Dirichlet allocation (LDA) (Blei et al., 2003) are major workhorses for summarising document collections. Typically, a topic model represents topics as distributions over the vocabulary (i.e., unique words in the corpus); documents are then modelled as distributions over topics. In this approach, words are vertices of a simplex whose dimension equals the vocabulary size and for which the distance between any pair of words is the same. More recently, word embeddings map words into high-dimensional space such that co-occurring words tend to be closer to each other than unrelated words (Mikolov et al., 2013; Pennington et al., 2014). Kusner et al. (2015a) combine the geometry of word embedding space with optimal transport to propose the *word mover's distance* (WMD), a powerful document distance metric limited mostly by computational complexity.

As an alternative to WMD, in this paper we combine hierarchical latent structures from topic models with geometry from word embeddings. We propose *hierarchical* optimal topic transport (HOTT)

document distances, which combine language information from word embeddings with corpus-specific, semantically-meaningful topic distributions from latent Dirichlet allocation (LDA) (Blei et al., 2003). This document distance is more efficient and more interpretable than WMD.

We give conditions under which HOTT gives a metric and show how it relates to WMD. We test against existing metrics on k-NN classification and show that it outperforms others on average. It performs especially well on corpora with longer documents and is robust to the number of topics and word embedding quality. Additionally, we consider two applications requiring pairwise distances. The first is visualization of the metric with t-SNE (van der Maaten & Hinton, 2008). The second is link prediction from a citation network, cast as pairwise classification using HOTT features.

Contributions. We introduce *hierarchical* optimal transport to measure dissimilarities between distributions with common structure. We apply our method to document classification, where topics from a topic modeller represent the shared structure. Our approach

- is computationally efficient, since HOTT distances involve transport with small numbers of sites;
- uses corpus-specific topic and document distributions, providing higher-level interpretability;
- has comparable performance to WMD and other baselines for k-NN classification; and
- is practical in applications where all pairwise document distances are needed.

8.2 Related work

Document representation and similarity assessment are key applications in learning. Many methods are based on the bag-of-words (BOW), which represents documents as vectors in $\mathbb{R}^{|V|}$, where $|V|$ is the vocabulary size; each coordinate equals the number of times a word appears. Other weightings include term frequency inverse document frequency (TF-IDF) (Luhn, 1957; Spärck Jones, 1972) and latent semantic indexing (LSI) (Deerwester et al., 1990). Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a hierarchical Bayesian model where documents are represented as admixtures of latent topics and admixture weights provide low-dimensional representations. These representations equipped with the l_2 metric comprise early examples of document dissimilarity scores.

Recent document distances employ more sophisticated methods. WMD incorporates word embeddings to account for word similarities (Kusner et al., 2015a) (see §8.3). Huang et al. (2016) extend WMD

to the supervised setting, modifying embeddings so that documents in the same class are close and documents from different classes are far. Due to computational complexity, these approaches are impractical for large corpora or documents with many unique words.

Wu & Li (2017) attempt to address the complexity of WMD via a topic mover's distance (TMD). While their k-NN classification results are comparable to WMD, they use significantly more topics, generated with a Poisson infinite relational model. This reduces semantic content and interpretability, with less significant computational speedup. They also do not leverage language information from word embeddings or otherwise. Xu et al. (2018) jointly learn topics and word embeddings, limiting the complexity to under a hundred words, which is not suited for natural language processing.

Wu et al. (2018) approximate WMD using a random feature kernel. In their method, the WMD from corpus documents to a selection of random short documents facilitates approximation of pairwise WMD. The resulting word mover's embedding (WME) has similar performance with significant speedups. Their method, however, requires parameter tuning in selecting the random document set and lacks topic-level interpretability. Additionally, they do not show full-metric applications. Lastly, Wan (2007), whose work predates (Kusner et al., 2015a), applies transport to blocks of text.

8.3 Background

Word mover's distance. Given an embedding of a vocabulary as $V \subset \mathbb{R}^n$, the Euclidean metric puts a geometry on the words in V. A corpus $D = \{d^1, d^2, \ldots d^{|D|}\}$ can be represented using distributions over V via a normalised BOW. In particular, $d^i \in \Delta^{l_i}$, where l_i is the number of unique words in a document d^i, and $d^i_j = c^i_j / |d^i|$, where c^i_j is the count of word v_j in d^i and $\left|d^i\right|$ is the number of words in d^i. The WMD between documents d^1 and d^2 is then $WMD(d^1, d^2) = W_1(d^1, d^2)$.

The complexity of computing WMD depends heavily on $l = \max(l_1, l_2)$; for longer documents, l may be a significant fraction of $|V|$. To evaluate the full metric on a corpus, the complexity is $O(|D|^2 l^3 \log l)$, since WMD must be computed pairwise. Kusner et al. (2015a) test WMD for k-NN classification. To circumvent complexity issues, they introduce a pruning procedure using a relaxed word mover's distance (RWMD) to lower-bound WMD. On the larger 20NEWS dataset, they additionally remove infrequent words by using only the top 500 words to generate a representation.

8.4 Hierarchical optimal transport

Assume a topic model produces corpus-specific topics $T = \{t_1, t_2, \dots, t_{|T|}\} \subset \Delta^{|V|}$, which are distributions over words, as well as document distributions $\bar{d}^i \in \Delta^{|T|}$ over topics. WMD defines a metric $WMD(t_i, t_j)$ between topics; we consider discrete transport over T as a metric space.

We define the hierarchical topic transport distance (HOTT) between documents d^1 and d^2 as

$$HOTT(d^1, d^2) = W_1\left(\sum_{k=1}^{|T|} \bar{d}_k^1 \delta_{t_k}, \sum_{k=1}^{|T|} \bar{d}_k^2 \delta_{t_k}\right),$$

where each Dirac delta δ_{t_k} is a probability distribution supported on the corresponding topic t_k and where the ground metric is WMD between topics as distributions over words. The resulting transport problem leverages topic correspondences provided by WMD in the base metric. This explains the *hierarchical* nature of our proposed distance.

Our construction uses transport *twice*: WMD provides topic distances, which are subsequently the costs in the HOTT problem. This hierarchical structure greatly reduces runtime, since $|T| \ll l$; the costs for HOTT can be precomputed once per corpus. The expense of evaluating pairwise distances is drastically lower, since pairwise distances between topics may be precomputed and stored. Even as document length and corpus size increase, the transport problem for HOTT remains the same size. Hence, full metric computations are feasible on larger datasets with longer documents.

When computing $WMD(t_i, t_j)$, we reduce computational time by truncating topics to a small amount of words carrying the majority of the topic mass and re-normalise. This procedure is motivated by interpretability considerations and estimation variance of the tail probabilities. On the interpretability side, LDA topics are often displayed using a few dozen top words, providing a human-understandable tag. Semantic coherence, a popular topic modelling evaluation metric, also is based on heavily-weighted words and was demonstrated to align with human evaluation of topic models (Newman et al., 2010). Moreover, any topic modelling inference procedure, e.g. Gibbs sampling (Griffiths & Steyvers, 2004), has estimation variance that may dominate tail probabilities, making them unreliable. Hence, we truncate to the top 20 words when computing WMD between topics. We empirically verify that truncation to any small

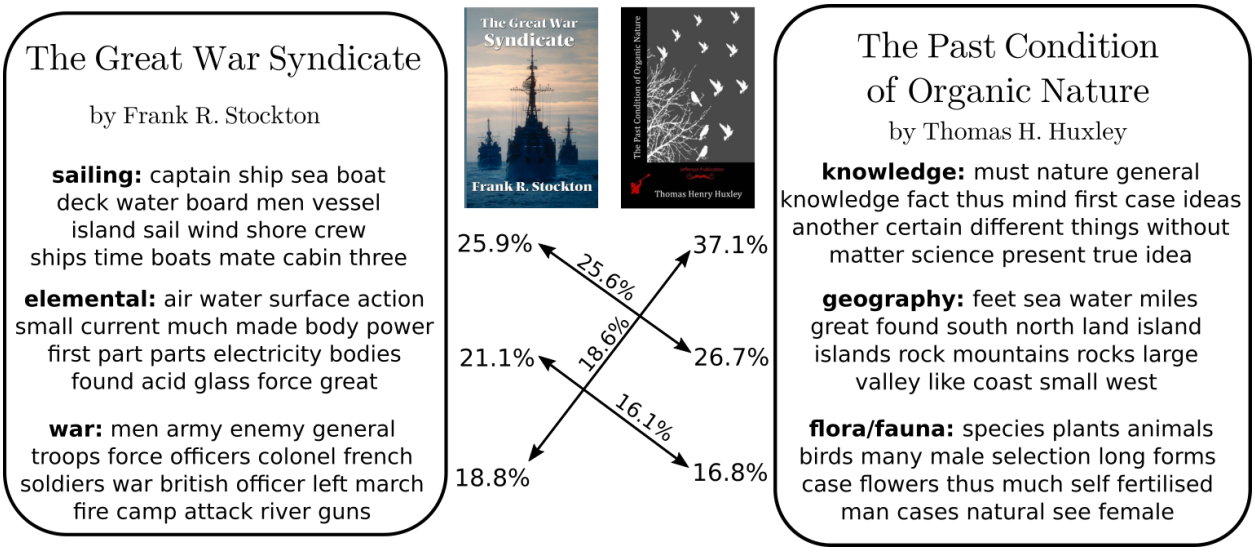


Figure 8-1: Topic transport interpretability. We show two books from GUTENBERG and their heaviest-weighted topics (bolded topic names are manually assigned). The first involves steamship warfare, while the second involves biology. Left and right column percentages indicate the weights of the topics in the corresponding texts. Percentages labelling the arrows indicate the transported mass between the corresponding topics, which match semantically-similar topics.

number of words performs equally well in §8.5.3.

In topic models, documents are assumed to be represented by a small subset of topics of size $\kappa_i \ll |T|$ (e.g., in Figure 8-1, *books* are majorly described by three topics), but in practice document topic proportions tend to be dense with little mass outside of the dominant topics. Williamson et al. (2010) propose an LDA extension enforcing sparsity of the topic proportions, at the cost of slower inference. When computing HOTT, we simply truncate LDA topic proportions at $1/(|T| + 1)$, the value below LDA's uniform topic proportion prior, and re-normalise. This reduces complexity of our approach without performance loss as we show empirically in §8.5.2 and §8.5.3.

Metric properties of HOTT. If each document can be uniquely represented as a linear combination of topics $d^i = \sum_{k=1}^{|T|} \bar{d}^i_k t_k$, and each topic is unique, then *HOTT* is a metric on document space. We present a brief proof in the supplementary material.

Topic-level interpretability. The additional level of abstraction promotes higher-level interpretability at the level of topics as opposed to dense word-level correspondences from WMD. We provide an example in Figure 8-1. This diagram illustrates two books from the GUTENBERG dataset and the semanti-

cally meaningful transport between their three most heavily-weighted topics. Remaining topics and less prominent transport terms account for the remainder of the transport plan not illustrated.

Relation to WMD. First we note that if $|T| = |V|$ and topics consist of single words covering the vocabulary, then HOTT becomes WMD. In well-behaved topic models, this is expected as $|T| \to |V|$. Allowing $|T|$ to vary produces different levels of granularity for our topics as well as a trade-off between computational speed and topic specificity. When $|T| \ll |V|$, we argue that WMD is upper bounded by HOTT and two terms that represent topic modelling loss. By the triangle inequality,

$$WMD(d^i, d^j) \leq W_1\left(d^i, \sum_{k=1}^{|T|} \bar{d}^i_k t_k\right) + W_1\left(\sum_{k=1}^{|T|} \bar{d}^i_k t_k, \sum_{k=1}^{|T|} \bar{d}^j_k t_k\right) + W_1\left(\sum_{k=1}^{|T|} \bar{d}^j_k t_k, d^j\right). \tag{8.1}$$

LDA inference minimises $\mathrm{KL}(d^i \| \sum_{k=1}^{|T|} \bar{d}^i_k t_k)$ over topic proportions $\bar{d}^i$ for a given document d^i; hence, we look to relate Kullback–Leibler divergence to W_1. In finite-diameter metric spaces, $W_1(\mu, \nu) \leq \mathrm{diam}(X)\sqrt{\frac{1}{2}\mathrm{KL}(\mu\|\nu)}$, which follows from inequalities relating Wasserstein distances to KL divergence (Otto & Villani, 2000). The middle term satisfies the following inequality:

$$W_1\left(\sum_{k=1}^{|T|} \bar{d}^i_k t_k, \sum_{k=1}^{|T|} \bar{d}^j_k t_k\right) \leq W_1\left(\sum_{k=1}^{|T|} \bar{d}^i_k \delta_{t_k}, \sum_{k=1}^{|T|} \bar{d}^j_k \delta_{t_k}\right), \tag{8.2}$$

where on the right we have $HOTT(d^1, d^2)$. The optimal topic transport on the right implies an equal-cost transport of the corresponding linear combinations of topic distributions on the left. The inequality follows since W_1 gives the *optimal* transport cost. Combining into a single inequality,

$$WMD(d^i, d^j) \leq HOTT(d^i, d^j) + \mathrm{diam}(X)\left[\sqrt{\frac{1}{2}\mathrm{KL}\left(d^j \middle\| \sum_{k=1}^{|T|} \bar{d}^j_k t_k\right)} + \sqrt{\frac{1}{2}\mathrm{KL}\left(d^i \middle\| \sum_{k=1}^{|T|} \bar{d}^i_k t_k\right)}\right].$$

WMD involves a large transport problem and Kusner et al. (2015a) propose relaxed WMD (RWMD), a relaxation via a lower bound (see also Atasu & Mittelholzer (2019) for a GPU-accelerated variant). We next show that RWMD is not always a good lower bound on WMD.

RWMD–Hausdorff bound. Consider the optimisation in (2.2) for calculating $WMD(d^1, d^2)$, and remove the marginal constraint on d^2. The resulting optimal Γ is no longer a transport plan, but rather

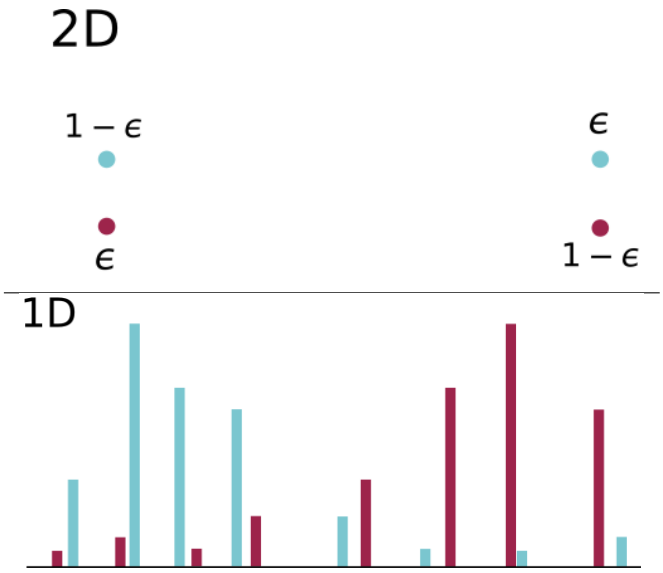


Figure 8-2: RWMD as a poor approximation to WMD

moves mass on words in d^1 to their nearest words in d^2, only considering the support of d^2 and not its density values. Removing the marginal constraint on d^1 produces symmetric behaviour; $RWMD(d^1, d^2)$ is defined to be the larger cost of these relaxed problems.

Suppose that word v_j is shared by d^1 and d^2. Then, the mass on v_j in d^1 and d^2 in each relaxed problems will not move and contributes zero cost. In the worst case, if d^1 and d^2 contain the same words, i.e., $\text{supp}(d^1) = \text{supp}(d^2)$, then $RWMD(d^1, d^2) = 0$. More generally, the closer the supports of two documents (over V), the looser RWMD might be as a lower bound.

Figure 8-2 illustrates two examples. In the 2D example, $1 - \epsilon$ and ϵ denote the masses in the teal and maroon documents. The 1D example uses histograms to represent masses in the two documents. In both, RWMD is nearly zero as masses do not have far to move, while the WMD will be larger thanks to the constraints.

To make this precise we provide the following tight upper bound: $RWMD(d^1, d^2) \leq d_H(\text{supp}(d^1), \text{supp}(d^2))$, the Hausdorff distance between the supports of d^1 and d^2. Let $X = \text{supp}(d^1)$ and $Y = \text{supp}(d^2)$; and let $RWMD_1(d^1, d^2)$ and $RWMD_2(d^1, d^2)$ denote the relaxed optimal values when the marginal constraints

on d^1 and d^2 are kept, respectively:

$$d_H(X,Y) = \max\left(\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right)$$
$$\geq \max\left(RWMD_1(d^1,d^2), RWMD_2(d^1,d^2)\right) = RWMD(d^1,d^2).$$

The inequality follows since the left argument of the max is the furthest mass must travel in the solution to $RWMD_1$, while the right is the furthest mass must travel in the solution to $RWMD_2$. It is tight if the documents have singleton support and whenever d^1 and d^2 are supported on parallel affine subspaces and are translates in a normal direction. A 2D example is in Figure 8-2.

The preceding discussion suggests that RWMD is not an appropriate way to speed up WMD for long documents with overlapping support, scenario where WMD computational complexity is especially prohibitive. The GUTENBERG dataset showcases this failure, in which documents frequently have common words. Our proposed HOTT does not suffer from this failure mode, while being significantly faster and as accurate as WMD. We verify this in the subsequent experimental studies. In the supplementary materials we present a brief empirical analysis relating HOTT and RWMD to WMD in terms of Mantel correlation and a Frobenius norm.

8.5 Experiments

We present timings for metric computation and consider applications where distance between documents plays a crucial role: k-NN classification, low-dimensional visualisation, and link prediction.

8.5.1 Computational timings

HOTT implementation. During training, we fit LDA with 70 topics using a Gibbs sampler (Griffiths & Steyvers, 2004). Topics are truncated to the 20 most heavily-weighted words and renormalised. The pairwise distances between topics $WMD(t_i, t_j)$ are precomputed with words embedded in $\mathbb{R}^{300}$ using *GloVe* (Pennington et al., 2014). To evaluate HOTT at testing time, a few iterations of the Gibbs sampler are run to obtain topic proportions $\bar{d}^i$ of a new document d^i. When computing HOTT between a pair

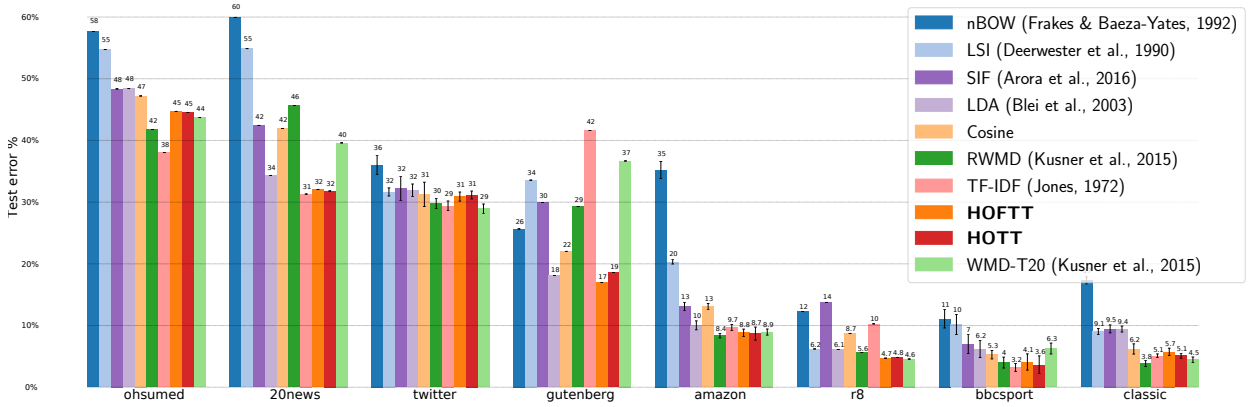


Figure 8-3: k-NN classification performance across datasets

of documents we truncate topic proportions at $1/(|T| + 1)$ and renormalise. Every instance of the OT linear program is solved using Gurobi (Gurobi Optimization, 2018).

We note that LDA inference may be carried out using any other approaches, e.g. stochastic/streaming variational inference (Hoffman et al., 2013; Broderick et al., 2013) or geometric algorithms (Yurochkin & Nguyen, 2016; Yurochkin et al., 2019c). We chose the MCMC variant (Griffiths & Steyvers, 2004) for its strong theoretical guarantees, simplicity and wide adoption in the topic modelling literature.

Topic computations. The preprocessing steps of our method—computing LDA topics and the topic to topic pairwise distance matrix—are dwarfed by the cost of computing the full document-to-document pairwise distance matrix. The complexity of base metric computation in our implementation is $O(|T|^2)$, since $\left|\text{supp}(t_i)\right| = 20$ for all topics, leading to a relatively small OT instance.

HOTT computations. All distance computations were implemented in Python 3.7 and run on an Intel i7-6700K at 4GHz with 32GB of RAM. Timings for pairwise distance computations are in Table 8.1 (right). HOTT outperforms RWMD and WMD in terms of speed as it solves a significantly smaller linear program. On the left side of Table 8.1 we summarise relevant dataset statistics: $|D|$ is the number of documents; $|V|$ is the vocabulary size; intersection over union (IOU) characterises average overlap in words between pairs of documents; $\text{AVG}(l)$ is the average number of unique words per document and $\text{AVG}(\kappa)$ is the average number of major topics (i.e., after truncation) per document.

Table 8.1: Dataset statistics and document pairs per second; higher is better. HOTT has higher throughput and excels on long documents with large portions of the vocabulary (as in GUTENBERG).

	Dataset statistics						Pairs per second				
Dataset	$\lvert D\rvert$	$\lvert V\rvert$	IOU	avg(l)	avg(κ)	Classes	RWMD	WMD	WMDT20	HOFTT	HOTT
BBCSPORT	737	3657	0.066	116.5	11.7	5	1494	526	1545	2016	2548
TWITTER	3108	1205	0.029	9.7	6.3	3	2664	2536	2194	1384	1552
OHSUMED	9152	8261	0.046	59.4	11.0	10	454	377	473	829	908
CLASSIC	7093	5813	0.017	38.5	8.7	4	816	689	720	980	1053
REUTERS8	7674	5495	0.06	35.7	8.7	8	834	685	672	918	989
AMAZON	8000	16753	0.019	44.3	9.0	4	289	259	253	927	966
20NEWS	13277	9251	0.011	69.3	10.5	20	338	260	384	652	699
GUTENBERG	3037	15000	0.25	4367	13.3	142	2	0.3	359	1503	1720

8.5.2 k-NN classification

We follow the setup of Kusner et al. (2015a) to evaluate performance of HOTT on k-NN classification.

Datasets. We consider 8 document classification datasets: BBC sports news articles (BBCSPORT) labelled by sport; tweets labelled by sentiments (TWITTER) (Sanders, 2011); Amazon reviews labelled by category (AMAZON); Reuters news articles labelled by topic (REUTERS) (we use the 8-class version and train-test split of Cachopo et al. (2007)); medical abstracts labelled by cardiovascular disease types (OHSUMED) (using 10 classes and train-test split as in Kusner et al. (2015a)); sentences from scientific articles labelled by publisher (CLASSIC); newsgroup posts labelled by category (20NEWS), with "by-date" train-test split and removing headers, footers and quotes;[1] and Project Gutenberg full-length books from 142 authors (GUTENBERG) using the author names as classes and 80/20 train-test split in the order of document appearance. For GUTENBERG, we reduced the vocabulary to the most common 15000 words. For 20NEWS, we removed words appearing in ≤ 5 documents.

Baselines. We focus on evaluating HOTT and a variation without topic proportion truncation (HOFTT: hierarchical optimal full topic transport) as alternatives to RWMD in a variety of metric-dependent tasks. As demonstrated by the authors, RWMD has nearly identical performance to WMD, while being more computationally feasible. Additionally, we analyse a naïve approach for speeding-up WMD where we truncate documents to their top 20 unique words (WMD-T20), making complexity comparable to HOTT (yet 20 $>$avg(κ) on all datasets). For k-NN classification, we also consider baselines that represent doc-

[1] https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

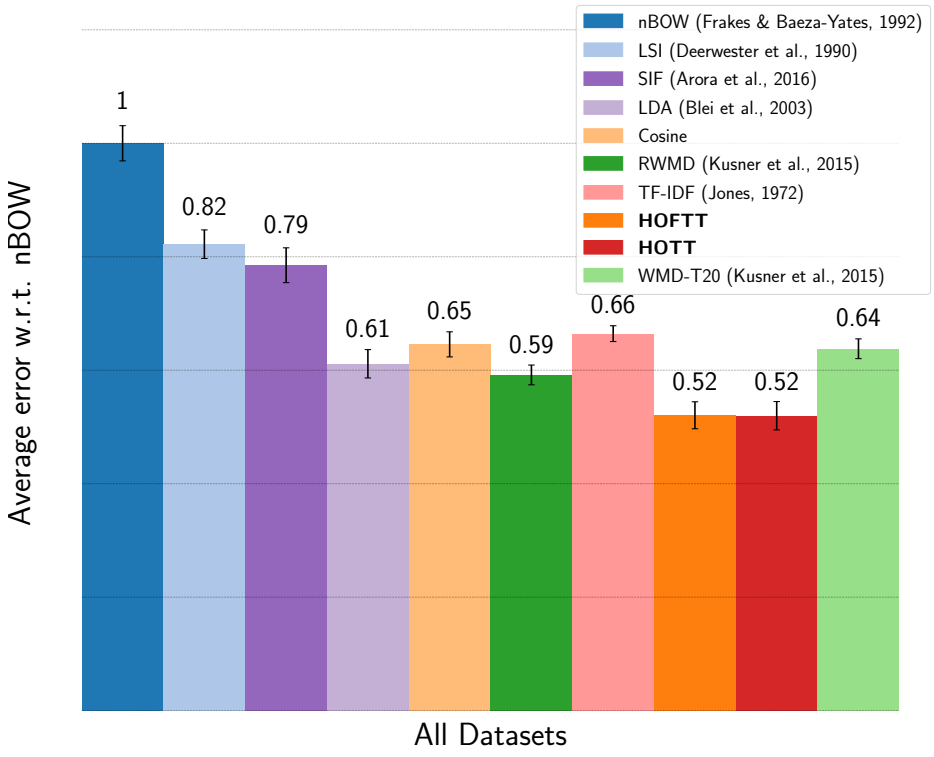


Figure 8-4: Aggregated k-NN classification performance normalised by nBOW

uments in vector form and use Euclidean distances: normalised bag-of-words (nBOW) (Frakes & Baeza-Yates, 1992); latent semantic indexing (LSI) (Deerwester et al., 1990); latent Dirichlet allocation (LDA) (Blei et al., 2003) trained with a Gibbs sampler (Griffiths & Steyvers, 2004); and term frequency inverse document frequency (TF-IDF) (Spärck Jones, 1972). We omit comparison to embedding via BOW weighted averaging as it was shown to be inferior to RWMD by Kusner et al. (2015a) (i.e., Word Centroid Distance) and instead consider smooth inverse frequency (SIF), a recent document embedding method by Arora et al. (2017). We also compare to bag-of-words, where neighbours are identified using cosine similarity (Cosine). We use same pre-trained *GloVe* embeddings for HOTT, RWMD, SIF and truncated WMD and set the same number of topics $|T| = 70$ for HOTT, LDA and LSI; we provide experiments testing parameter sensitivity.

Results. We evaluate each method on k-NN classification (Fig. 8-3). There is no uniformly best method, but HOTT performs best on average (Fig. 8-4) We highlight the performance on the Gutenberg dataset compared to RWMD. We anticipate poor performance of RWMD on Gutenberg, since

books contain more words, which can make RWMD degenerate (see §8.4 and Fig. 8-2). Also note strong performance of TF-IDF on OHSUMED and 20NEWS, which differs from results of Kusner et al. (2015a). We believe this is due to a different normalisation scheme. We used *TfidfTransformer* from scikit-learn (Pedregosa et al., 2011) with default settings. We conclude that HOTT is most powerful, both computationally (Table 8.1 right) and as a distance metric for k-NN classification (Figures 8-3 and 8-4), on larger corpora of longer documents, whereas on shorter documents both RWMD and HOTT perform similarly.

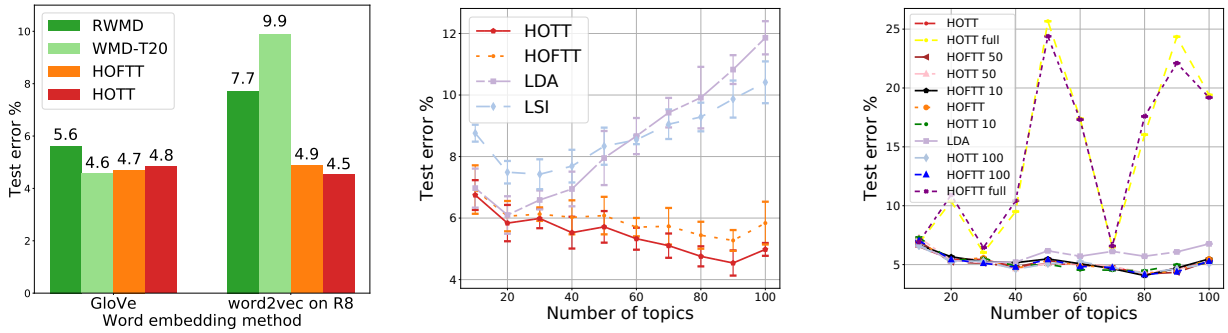
Another interesting observation is the effect of truncation: HOTT performs as well as HOFTT, meaning that truncating topic proportions of LDA does not prevent us from obtaining high-quality document distances in less computational time, whereas truncating unique words for WMD degrades its performance. This observation emphasises the challenge of speeding up WMD, i.e. WMD *cannot* be made computationally efficient using truncation without degrading its performance. WMD-T20 is slower than HOTT (Table 8.1) and performs noticeably worse (Figure 8-4). Truncating WMD further will make its performance even worse, while truncating less will quickly lead to impractical run-time.

In the supplement, we complement our results considering 2-Wasserstein distance, and stemming, a popular text pre-processing procedure for topic models to reduce vocabulary size. HOTT continues to produce best performance on average. We restate that in all main text experiments we used 1-Wasserstein (i.e. eq. (2.2)) and did not stem, following experimental setup of Kusner et al. (2015a).

8.5.3 Sensitivity analysis of HOTT

We analyse sensitivity of HOTT with respect to its components: word embeddings, number of LDA topics, and topic truncation level.

Sensitivity to word embeddings. We train *word2vec* (Mikolov et al., 2013) 200-dimensional embeddings on REUTERS and compare relevant methods with our default embedding (i.e., *GloVe*) and newly-trained *word2vec* embeddings. According to Mikolov et al. (2013), word embedding quality largely depends on data *quantity* rather than quality; hence we expect the performance to degrade. In Fig. 8-5(a), RWMD and WMD truncated performances drop as expected, but HOTT and HOFTT remain stable; this behaviour is likely due to the embedding-independent topic structure taken into consideration.



Embedding sensitivity on REUTERS

Topic number sensitivity on CLASSIC

Topic truncation sensitivity on REUTERS

Figure 8-5: Sensitivity of our approach with respect to hyperparameters.

Number of LDA topics. In our experiments, we set $|T| = 70$. When the $|T|$ increases, LDA resembles the nBOW representation; correspondingly, HOTT approaches the WMD. The difference, however, is that nBOW is a weaker baseline, while WMD is powerful document distance. Using the CLASSIC dataset, in Fig. 8-5(b) we demonstrate that LDA (and LSI) may degrade with too many topics, while HOTT and HOFTT are robust to topic overparameterization. In this example, better performance of HOTT over HOFTT is likely due relatively short documents of the CLASSIC dataset.

While we have shown that HOTT is not sensitive to the choice of the number of topics, it is also possible to eliminate this parameter by using LDA inference algorithms that learn number of topics (Yurochkin et al., 2017) or adopting Bayesian nonparametric topic modes and corresponding inference schemes (Teh et al., 2006; Wang et al., 2011; Bryant & Sudderth, 2012).

Topic truncation. Fig. 8-5(c) demonstrates k-NN classification performance on the REUTERS dataset with varying topic truncation: top 10, 20 (HOTT and HOFTT), 50, 100 words and no truncation (HOTT full and HOFTT full); LDA performance is given for reference. Varying the truncation level does not affect the results significantly, however no truncation results in unstable performance.

8.5.4 t-SNE metric visualisation

Visualising metrics as point clouds provides useful qualitative information for human users. Unlike k-NN classification, most methods for this task require long-range distances and a full metric. Here, we use

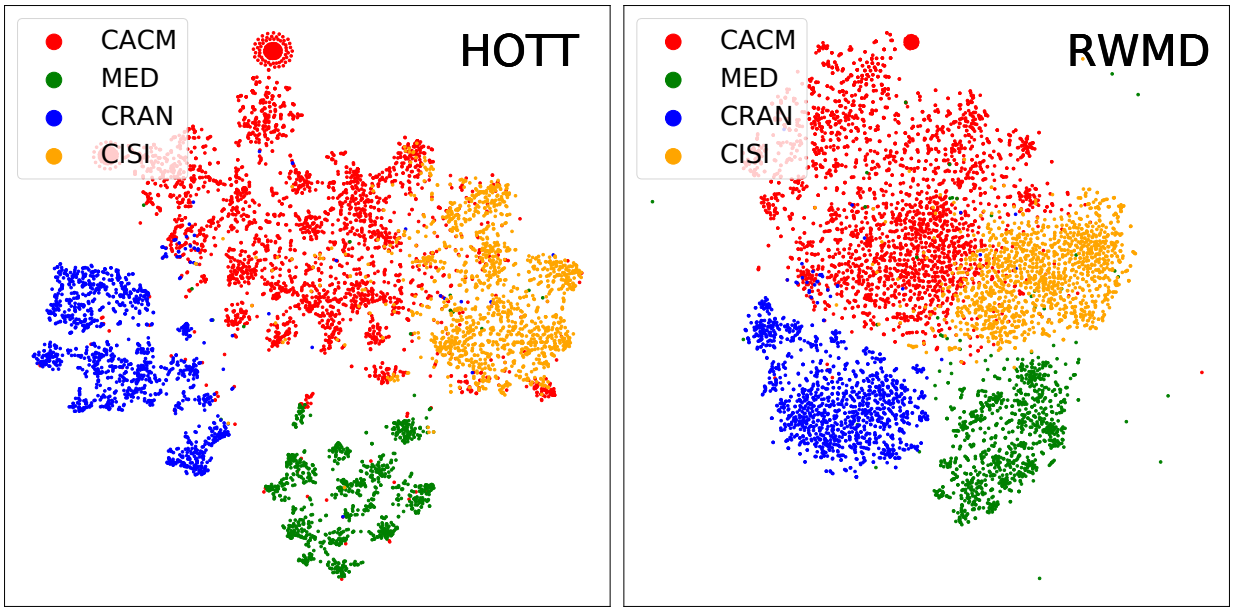


Figure 8-6: t-SNE on CLASSIC

t-SNE (van der Maaten & Hinton, 2008) to visualise HOTT and RWMD on the CLASSIC dataset in Fig. 8-6. HOTT appears to more accurately separate the labelled points (colour-coded). The supplementary material gives additional t-SNE results.

8.5.5 Supervised link prediction

We next evaluate HOTT in a different prediction task: supervised link prediction on graphs defined on text domains, here *citation networks*. The specific task we address is the Kaggle challenge of Link Prediction TU.[2] In this challenge, a citation network is given as an undirected graph, where nodes are research papers and (undirected) edges represent citations. From this graph, edges have been removed at random. The task is to reconstruct the full network. The dataset contains 27770 papers (nodes). The training and testing sets consist of 615512 and 32648 node pairs (edges) respectively. For each paper, the available data only includes publication year, title, authors, and abstract.

To study the effectiveness of a distance-based model with HOTT for link prediction, we train a linear SVM classifier over the feature set Φ, which includes the distance between the two abstracts ϕ_{dist} com-

[2] www.kaggle.com/c/link-prediction-tu

Table 8.2: Link prediction: using distance (rows) for node-pair representations (cols).

Distance	F1 Score				
	Φ_0	Φ_1	Φ_2	Φ_3	Φ_4
HOFTT	73.22	76.27	76.62	78.85	83.37
HOTT	73.19	76.03	76.24	78.64	83.25
RWMD	71.60	74.90	75.20	77.16	82.92
WMD-T20	67.22	63.38	65.20	70.38	81.84
None	—	61.13	64.27	67.72	81.68

puted via one of {HOFT, HOTT, RWMD, WMD-T20}. For completeness, we also examine excluding the distance totally. We incrementally grow the feature sets Φ as: $\Phi_0 = \{\phi_{dist}\}$, $\Phi_1 = \{\phi_{dist}\} \cup \{\phi_1\}$, $\Phi_n = \{\phi_{dist}\} \cup \{\phi_1, \ldots, \phi_n\}$ where ϕ_1 is the number of common words in the titles, ϕ_2 the number of common authors, and ϕ_3 and ϕ_4 the signed and absolute difference between the publication years.

Table 8.2 presents the results; evaluation is based on the F1-Score. Consistently, HOFTT and HOTT are more effective than RWMD and WMD-T20 in all tests, and not using any of the distances consistently degrades the performance.

8.6 Conclusion

We have proposed a hierarchical method for comparing natural language documents that leverages optimal transport, topic modelling, and word embeddings. Specifically, word embeddings provide global semantic language information, while LDA topic models provide corpus-specific topics and topic distributions. Empirically these combine to give superior performance on various metric-based tasks. We hypothesise that modelling documents by their representative topics is better for highlighting differences despite the loss in resolution. HOTT appears to capture differences in the same way a person asked to compare two documents would: by breaking down each document into easy to understand concepts, and then comparing the concepts.

There are many avenues for future work. From a theoretical perspective, our use of a nested Wasserstein metric suggests further analysis of this hierarchical transport space. Insight gained in this direction

may reveal the learning capacity of our method and inspire faster or more accurate algorithms. From a computational perspective, our approach currently combines word embeddings, topic models and OT, but these are all trained separately. End-to-end training that efficiently optimises these three components jointly would likely improve performance and facilitate analysis of our algorithm as a unified approach to document comparison.

Finally, from an empirical perspective, the performance improvements we observe stem directly from a reduction in the size of the transport problem. Investigation of larger corpora with longer documents, and applications requiring the full set of pairwise distances are now feasible. We also can consider applications to modelling of images or 3D data.

9

Alleviating Label Switching with Optimal Transport

What follows may seem to come from left field given what has come before, but at the heart of our proposal to alleviate a common issue in Bayesian inference is an algorithm that relies heavily on the hierarchical point of view we have explored in the previous chapter.

Label switching is a problem caused by the invariance of optimisation variables to a group action, such as permuting the labels of the variables. This leads to problems down the road when we want to compute statistics of these variables. Our approach lifts this problem into the space of measures on measures and develops tools to manipulate these higher order distributions that are both fast and effective.

9.1 Introduction

Mixture models are powerful tools for understanding multimodal data. In the Bayesian setting, to fit a mixture model to such data, we typically assume a prior number of components and optimise or sample from the posterior distribution over the component parameters. If prior components are exchangeable, this leads to an identifiability issue known as *label switching*. In particular, permuting the ordering of mixture components does not change the likelihood, since it produces the same model. The underlying problem is that a group acts on the parameters of the mixture model; posterior probabilities are invariant under the action of the group.

To formalise this intuition, suppose our input is a data set X and a parameter K denoting the number of mixture components. In the most common application, we want to fit a mixture of K Gaussians to

the data; our parameter set is $\Theta = \{\theta_1, \dots, \theta_K\}$ where $\theta_k = \{\mu_k, \Sigma_k, \pi_k\}$ gives the parameters of each component. The likelihood of $x \in X$ conditioned on Θ is $p(x|\Theta) = \sum_{k=1}^{K} \pi_k f(x; \mu_k, \Sigma_k)$, where $f(x; \mu_k, \Sigma_k)$ is the density function of $\mathcal{N}(\mu_k, \Sigma_k)$. Any permutation of the labels $k = 1, \dots, K$ yields the same likelihood. The prior is also permutation invariant. When we compute statistics of the posterior $p(\Theta|x)$, however, this permutation invariance leads to $K!$ symmetric regions in the posterior landscape. Sampling and inference algorithms behave poorly as the number of modes increases, and this problem is only exacerbated in this context since increasing the number of components in the mixture model leads to a super-exponential increase in the number of modes of the posterior. Previous methods such as the invariant losses of Celeux et al. (2000) and pivot alignments of Marin et al. (2005) do not identify modes in a principled manner.

To combat this issue, we leverage the theory of optimal transport. In particular, one way to avoid the multimodal nature of the posterior distribution is to replace each sample with its orbit under the action of the symmetry group seen as a distribution over $K!$ points. While this symmetrised distribution is invariant to group actions, we can not average several such distributions using standard Euclidean metrics. We use the notion of a Wasserstein barycenter to calculate a mean in this space, which we can project to a mean in the parameter space via the quotient map. We show conditions under which our optimisation can be performed efficiently on the quotient space, thus circumventing the need to store and manipulate orbit distributions with large support.

Contributions. We give a practical and simple algorithm to solve the *label switching* problem. To justify our algorithm, we demonstrate that a group-invariant Wasserstein barycenter exists when the distributions being averaged are group-invariant. We give conditions under which the Wasserstein barycenter can be written as the orbit of a single point, and we explain how failure modes of our algorithm correspond to ill-posed problems. We show that the problem can be cast as computing the expected value of the quotient distribution, and we give an SGD algorithm to solve it.

9.2 Related work

Mixture models. Gaussian mixture models are powerful for modelling a wide range of phenomena (McLachlan et al., 2019). These models assume that a sample is drawn from one of the latent states (or

components), but that the particular component assigned to any given sample is unknown. In a Bayesian setup, Markov Chain Monte Carlo can sample from the posterior distribution over the parameters of the mixture model. Hamiltonian Monte Carlo (HMC) has proven particularly successful for this task. Introduced for lattice quantum chromodynamics (Duane et al., 1987), HMC has become a popular option for statistical applications (Neal et al., 2011). Recent high-performance software offers practitioners easy access to HMC and other sampling algorithms (Carpenter et al., 2017).

Label switching. Label switching arises when we take a Bayesian approach to parameter estimation in mixture models (Diebolt & Robert, 1994). Jasra et al. (2005) and Papastamoulis (2016) overview the problem. Label switching can happen even when samplers do not explore all $K!$ possible modes, e.g., for Gibbs sampling. Documentation for modern sampling tools mentions that it arises in practice.[I] Label switching can also occur when using parallel HMC, since tools like `Stan` run multiple chains at once. While a single chain may only explore one mode, several chains are likely to yield different label permutations.

Jasra et al. (2005, §6) mention a few loss functions invariant to the different labellings. Most relevant is the loss proposed by Celeux et al. (2000, §5). Beyond our novel theoretical connections to optimal transport, in contrast to their method, our algorithm uses optimal rather than greedy matching to resolve elements of the symmetric group, applies to general groups and quotient manifolds, and uses stochastic gradient descent instead of simulated annealing. Somewhat ad-hoc but also related is the pivotal reordering algorithm (Marin et al., 2005), which uses a sample drawn from the distribution as a pivot point to break the symmetry; as we will see in our experiments, a poorly-chosen pivot seriously degrades the performance.

Optimal transport. Optimal transport (OT) has seen a surge of interest in learning, from applications in generative models (Arjovsky et al., 2017; Genevay et al., 2018), Bayesian inference (Srivastava et al., 2015a), and natural language (Kusner et al., 2015b; Alvarez-Melis & Jaakkola, 2018) to technical underpinnings for optimisation methods (Chizat & Bach, 2018). See Solomon (2018); Peyré & Cuturi (2018) for discussion of computational OT and Santambrogio (2015); Villani (2008) for theory.

The Wasserstein distance from optimal transport (§9.3.1) induces a metric on the space of probability distributions from the geometry of the underlying domain. This leads to a notion of a Wasserstein barycenter of several probability distributions (Agueh & Carlier, 2011). Scalable algorithms have been

[I] https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html

proposed for barycenter computation, including methods that exploit entropic regularisation (Cuturi & Doucet, 2014), use parallel computing (Staib et al., 2017), or apply stochastic optimisation (Claici et al., 2018)

9.3 Optimal transport under group actions

Before delving into technical details, we will illustrate our approach with a simple example. Assume we have some data to which we wish to fit a Gaussian mixture model with K components. We can now draw samples from the posterior distribution, and we would like to obtain a point estimate of the mean of the posterior. We draw two samples $\Theta^1 = (\theta_1^1, \dots, \theta_K^1)$ and $\Theta^2 = (\theta_1^2, \dots, \theta_K^2)$. We cannot average them due to the ambiguity of label switching; see Figure 9-1(a) and §B.1.3 of the supplementary for a simple example. However, we can explicitly encode this multimodality as a uniform distribution over all $K!$ states:

$$\frac{1}{K!}\sum_{\sigma\in S_K}\delta_{\sigma\cdot\Theta^1} \quad \text{and} \quad \frac{1}{K!}\sum_{\sigma\in S_K}\delta_{\sigma\cdot\Theta^2}$$

where S_K is the symmetry group on K points that acts by permuting the elements of Θ^1 and Θ^2. These distributions are now invariant to permutations, so we can ask if there exists an average in this space. In this section, we prove that this is possible through the machinery of optimal transport.

We provide theoretical results relevant to optimal transport between measures supported on the quotient space under actions of some group G. This theory is fairly general and requires only basic assumptions about the underlying space X and the action of G. For each theoretical result, we will use *italics* to highlight key assumptions, since they vary somewhat from proposition to proposition.

9.3.1 Preliminaries: Optimal transport

W_p induces a metric on the set $P_p(X)$ of measures with *finite* p-th moments (Villani, 2008). We will focus on $P_2(X)$, endowed with the metric W_2. This metric structure allows us to define meaningful statistics for sets of distributions. In particular, a Fréchet mean (or Wasserstein barycenter) of a set of distributions

$\nu_1, \ldots, \nu_n \in P_2(X)$ is defined as a minimiser

$$\mu^* = \operatorname*{arg\,min}_{\mu \in P_2(X)} \sum_{i=1}^{n} \frac{1}{n} W_2^2(\mu, \nu_i). \tag{9.1}$$

We follow Kim & Pass (2017) and generalise this notion slightly, by placing a measure itself on the space $P_2(X)$. We will use $P_2(P_2(X))$ to denote the space of probability measures on $P_2(X)$ that have finite second moments and let Ω be a member of this set. Then the following functional will be finite, which generalises (9.1) from finite sums to infinite sets of measures:

$$B(\mu) = \int_{P_2(X)} W_2^2(\mu, \nu) \, \mathrm{d}\Omega(\nu) = \mathbb{E}_{\nu \sim \Omega} \left[W_2^2(\mu, \nu) \right]. \tag{9.2}$$

In analogue to (9.1), a natural task is to search for a minimiser of the map $\mu \mapsto B(\mu)$. For existence of such a minimiser, we simply require that $\operatorname{supp}(\Omega)$ is tight.

Definition 9.1 (Tightness of measures). *A collection $\mathscr{C}$ of measures on X is called* tight *if for any $\varepsilon > 0$ there exists a compact set $K \subset X$ such that for all $\mu \in \mathscr{C}$, we have $\mu(K) > 1 - \varepsilon$.*

Here are three examples of tight collections: $P_2(X)$ if X is compact, the set of all Gaussian distributions with means supported on a compact space and of bounded variance, or any set of measures with a uniform bound on second moments (argued in §B.1.2 of the supplementary). This assumption is fairly mild and covers many application scenarios.

Prokhorov's theorem (deferred to the §B.1.1) implies the existence of a barycenter:

Theorem 9.1 (Existence of minimisers). *$B(\mu)$ has at least one minimiser in $P_2(X)$ if $\operatorname{supp}(\Omega)$ is tight.*

9.3.2 Optimal transport with group invariances

Let G be a *finite group* that acts by *isometries* on X. We define the set of measures invariant under group action $P_2(X)^G = \{\mu \in P_2(X) \mid g_\# \mu = \mu, \forall g \in G\}$, where the pushforward of μ by g is defined as $g_\# \mu(B) = \mu(g^{-1}(B))$ for B a measurable set. We are interested in the relation between the space $P_2(X)^G$ and the space of measures on the quotient space $P_2(X/G)$. If all of the measures in the support of Ω in (9.2) are invariant under group action, we can show that there exists a barycenter with the same property:

Lemma. *If $\Omega \in P_2(P_2(X)^G)$ is supported on the set of group-invariant measures on X and $supp(\Omega)$ is tight, then there exists a minimiser of $B(\mu)$ in $P_2(X)$ that is invariant under group action.*

Proof. Let $\mu \in P_2(X)$ denote the minimiser from Theorem 9.1. Define a new distribution $\mu_G = \frac{1}{|G|}\sum_{g\in G} g_\# \mu$. We verify that μ_G has the same cost as μ:

$$\mathbb{E}_{\nu\sim\Omega}\left[W_2^2\left(\frac{1}{|G|}\sum_{g\in G} g_\#\mu, \nu\right)\right] \leq \mathbb{E}_{\nu\sim\Omega}\left[\frac{1}{|G|}\sum_{g\in G} W_2^2(g_\#\mu, \nu)\right] \text{ by convexity of } \mu \mapsto W_2^2(\mu,\nu)$$

$$= \mathbb{E}_{\nu\sim\Omega}\left[\frac{1}{|G|}\sum_{g\in G} W_2^2(\mu, (g^{-1})_\#\nu)\right] \text{ since } g \text{ acts by isometry}$$

$$= \frac{1}{|G|}\sum_{g\in G} \mathbb{E}_{\nu\sim\Omega}[W_2^2(\mu,\nu)] = \mathbb{E}_{\nu\sim\Omega}[W_2^2(\mu,\nu)] \text{ by linearity of expectation and group invariance of } \nu.$$

But μ is a minimiser, so the inequality in line 1 must be an equality. □

Remark: If X is a compact Riemannian manifold and Ω gives positive weight to the set of absolutely continuous measures, then Theorem 3.1 of Kim & Pass (2017) provides uniqueness (and this may be extended to other non-compact cases with suitable decay conditions). However, in our setting, Ω is supported on samples, measures consisting of delta functions. In this case, a simple counterexample is presented in the supplementary (§B.1.4) which arises in the case where X consists of two points in $\mathbb{R}^2$ and S_2 acts to swap the points (S_K is the group of permutations of a finite set of K points). This is accompanied by a study of the case of K points in $\mathbb{R}^d$ (see supplementary §B.1.3), relevant to the mixture models where components are evenly weighted and identical with a single mean parameter. Via this study we see that counterexamples seem to require a high degree of symmetry, which is unlikely to happen in applied scenarios, and does not arise empirically in our experiments.

An analogous proof technique can be used to show the following lemma needed later:

Lemma. *If ν_1 and ν_2 are two measures invariant under group action, then there exists an optimal transport plan $\pi \in \Pi(\nu_1,\nu_2)$ that is invariant under the group action $g\cdot\pi(x,y) = \pi(g\cdot x, g\cdot y)$.*

The above suggests that we might instead search for barycenters in the quotient space. Consider:

Lemma (Lott & Villani 2009, Lemma 5.36). *Let $p : X \to X/G$ be the quotient map. The map $p_* : P_2(X) \to P_2(X/G)$ restricts to an isometric isomorphism between the set of $P_2(X)^G$ of G-invariant elements*

in $P_2(X)$ and $P_2(X/G)$.

We now introduce additional structure relevant to label switching. Assume that all measures $\nu \sim \Omega$ are the orbits of individual delta distributions, as they are samples of parameter values, i.e., $\nu = \frac{1}{|G|}\sum_{g\in G}\delta_{g\cdot x}$ for some $x \in X$. In the simple example of a mixture of two Gaussians from 1D data with means at $\mu_1, \mu_2 \in \mathbb{R}$, ν is of the following form $\nu = \frac{1}{2}\delta_{(\mu_1,\mu_2)} + \frac{1}{2}\delta_{(\mu_2,\mu_1)}$.

Under this assumption and by Lemmas 9.3.2 and 9.3.2, minimisation of $B(\mu)$ is equivalent to finding the Wasserstein barycenter of delta distributions on X/G. Letting $\Omega_* := p_{*\#}\Omega$, we aim to find:

$$\operatorname*{arg\,min}_{\mu\in P_2(X/G)} \mathbb{E}_{\delta_x\sim\Omega_*}\left[W_2^2(\mu,\delta_x)\right]. \tag{9.3}$$

From properties of Wasserstein barycenters (Carlier et al. 2015, Equation (2.9)), the support of μ lies in the set of solutions to

$$\min_{z\in X/G} \mathbb{E}_{\delta_x\sim\Omega_*}\left[d(x,z)^2\right] \tag{9.4}$$

where d is the metric on the quotient space X/G (see e.g. Santambrogio 2015, §5.5.5). As Ω has finite second moments, so does Ω_*, giving us existence of the expectation. The existence of minimisers of $z \to \mathbb{E}_{\delta_x\sim\Omega_*}\left[d(x,z)^2\right]$ is established in §B.2.1 of the supplementary, giving the following lemma:

Lemma. *The map $z \to \mathbb{E}_{\delta_x\sim\Omega_*}\left[d(x,z)^2\right]$ has a minimiser.*

Uniqueness of minimisers is not guaranteed (see §B.1.4 of supplementary), but we can rewrite (9.3) as:

$$\begin{aligned}\operatorname*{arg\,min}_{\mu\in P_2(X/G)} \mathbb{E}_{\delta_x\sim\Omega_*}\left[W_2^2(\mu,\delta_x)\right] &= \operatorname*{arg\,min}_{\mu\in P_2(X/G)} \int_{X/G}\int_{X/G} d(x,y)^2 \,\mathrm{d}\mu(y)\,\mathrm{d}\Omega_*(\delta_x) \\ &= \operatorname*{arg\,min}_{\mu\in P_2(X/G)} \int_{X/G}\int_{X/G} d(x,y)^2 \,\mathrm{d}\Omega_*(\delta_x)\,\mathrm{d}\mu(y).\end{aligned}$$

By Lemma 9.3.2, the term $y \to \int_{X/G} d(x,y)^2 \mathrm{d}\Omega_*(\delta_x)$ has a (potentially non-unique) minimiser. Call this function $b(y)$. We are left with

$$\operatorname*{arg\,min}_{\mu\in P_2(X/G)} \int_{X/G} b(y)\,\mathrm{d}\mu(y).$$

Any minimiser y^* of b leads to a minimising distribution $\mu = \delta_{y^*}$, and we can conclude

Theorem 9.2 (Single Orbit Barycenters). *There is a barycenter solution of* (9.2) *that can be written as* $\mu = \frac{1}{|G|} \sum_{g \in G} \delta_{g \cdot z^*}$.

Returning to our example of a Gaussian mixture model, we see that this theorem implies there is a barycenter (a mean in distribution space) that has the same form as the symmetrised sample distributions. Any point in the support of the barycenter is an estimate for the mean of the posterior distribution.

As an aside, we mention that our proofs do not require finite groups. In fact, we prove Theorem 9.2 for compact groups G endowed with a Haar measure in the supplement.

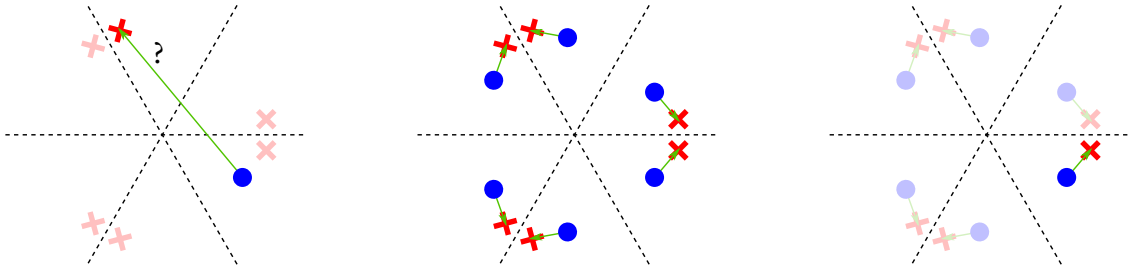
To summarise: Label switching leads to issues when computing posterior statistics because we work in the full space X, when we ought to work in the quotient space X/G. Theorem 9.2 relates means in X/G to barycenters of measures on X which gives us a principled method for computing statistics backed by a convex problem in the space of measures: take a quotient, find a mean in X/G, and then pull the result back to X. We will see below in concrete detail that we do not need to explicitly construct and average in X/G, but may leverage group invariance of the transport to perform these steps in X.

The crux of this theory is that the Wasserstein barycenter in the setting of Lemma 9.3.2 is a point estimate for the mean of the symmetrised posterior distribution. The results leading to Theorem 9.2 should be understood then as a reduction of the problem of finding an estimate of the mean to that of minimising a distance function on the quotient space; this latter minimisation problem can then be solved via Riemannian gradient descent.

9.4 Algorithms

Label switching usually occurs due to symmetries of certain Bayesian models. Posteriors with the label switching make it difficult to compute meaningful summary statistics, e.g. posterior expectations for the parameters of interest.

Any attempt to compute posterior statistics in this regime must account for the *orbits* of samples under the symmetry group. Continuing in the case of expectations, based on the previous section we can extract a meaningful notion of averaging by taking the image of each posterior sample under the



(a) Ambiguity (b) Orbit empirical distribution (c) Quotient update

Figure 9-1: (a) Suppose we wish to update our estimate of the average (blue) given a new sample (red) from Ω; due to label switching, other points (light shade) have equal likelihood to our sample, causing ambiguity. (b) Theorem 9.2 suggests an unambiguous update by constructing $|G|$-point orbits as empirical distributions and doing gradient descent with respect to the Wasserstein metric. (c) This algorithm is equivalent to moving one point, with a careful choice of update functions. This schematic arises for a mean-only model with three means in $\mathbb{R}$ (§B.1.3 of supplementary); $G = S_3$, with action is generated by reflection over the dashed lines.

symmetry group and computing a barycenter with respect to the Wasserstein metric. This resolves the ambiguity regarding which points in orbits should match, without symmetry-breaking heuristics like pivoting (Marin et al., 2005).

Algorithm 4 Riemannian Barycenter of Ω.

Require: Distribution Ω, exp and log maps on $\mathcal{M}$

Ensure: Estimate of the barycenter of Ω

1: Initialise the barycenter $p \sim \Omega$

2: for $t = 1, \ldots$ do

3: Draw $q \sim \Omega$

4: $-D_p c(p, q) := \log_p(q)$

5: $p \leftarrow \exp_p\left(-\frac{1}{t} D_p c(p, q)\right)$

In this section, we provide an algorithm for computing the W_2 barycenters above, extracting a symmetry-invariant notion of expectation for distributions with label switching. As input, we are given a sampler from a distribution Ω over a space $\mathcal{M}$ subject to label switching, as well as its (finite) symmetry group G. Our goal is to output a barycenter of the form $\frac{1}{|G|} \sum_{g \in G} \delta_{g \cdot x}$ for some $x \in \mathcal{M}$, using stochastic gradient descent on (9.2). Our approach can be interpreted two ways, echoing the derivation of Theorem 9.2:

- The most direct interpretation, shown in Figure 9-1(b), is that we push forward Ω to a distribu-

tion over empirical distributions of the form $\frac{1}{|G|}\sum_{g\in G}\delta_{g\cdot x}$, where $x \sim \Omega$, and then compute the barycenter as a $|G|$-point empirical distribution whose support points move according to stochastic gradient descent, similar to the method by Claici et al. (2018)

- Since $|G|$ can grow extremely quickly, we argue that this algorithm is *equivalent* to one that moves a single representative x, so long as the gradient with respect to x accounts for the objective function; this is illustrated in Figure 9-1(c).

Although our final algorithm has cosmetic similarity to pivoting and other algorithms that compute a single representative point, the details of our approach show an *equivalence* to a well-posed transport problem. Moreover, our stochastic gradient algorithm invokes a sampler from Ω in every iteration, rather than precomputing a finite sample, i.e. our algorithm deals with samples as they come in, rather than collecting multiple samples, and then trying to cluster or break the symmetry *a posteriori*.

Algorithm 5 Barycenter of Ω on quotient space

Require: Distribution Ω, exp and log maps on $\mathcal{M}$

Ensure: Barycenter $[(p_1, \dots, p_K)]$

1: Initialise the barycenter $(p_1, \dots, p_K) \sim \Omega$
2: for $t = 1, \dots$ do
3: Draw $(q_1, \dots, q_K) \sim \Omega$
4: Compute σ in (9.5)
5: for $i = 1, \dots, K$ do
6: $-D_{p_i} c(p_i, q_{\sigma(i)}) := \log_{p_i}(q_{\sigma(i)})$
7: $p_i \leftarrow \exp_{p_i}\left(-\frac{1}{t} D_{p_i} c(p_i, q_{\sigma(i)})\right)$

Gradient descent on the quotient space. For simplicity of exposition, we introduce a few additional assumptions on our problem; our algorithm can generalise to other cases, but these assumptions are the most relevant to the experiments and applications in §9.5. In particular, we assume we are trying to infer a mixture model with K components. The parameters of our model are tuples $(p_1, \dots, p_K)$, where $p_i \in \mathcal{M}$ for all i and some Riemannian manifold $\mathcal{M}$. We can think of the space of parameters as the product $\mathcal{M}^K$. Typically it is undesirable when two components match exactly in a mixture model, so we additionally

excise any tuple $(p_1, \dots, p_K)$ with any matching elements (together a set of measure zero). Representing parameters in a mixture model can be made through a point process, it is natural to work with the Kth ordered configuration space of $\mathscr{M}$ considered in physics and algebraic topology (Fadell & Husseini, 2012):

$$\mathrm{Conf}_K(\mathscr{M}) := \mathscr{M}^K \setminus \{(p_1, \dots, p_K) \mid p_i = p_j \text{ for some } i \neq j\} \subset \mathscr{M}^K.$$

Let $\Omega \in P(\mathrm{Conf}_K(M))$ be the Bayesian posterior distribution restricted to $\mathrm{Conf}_K(M)$ (assuming the posterior $P(\mathscr{M}^K)$ is absolutely continuous with respect to the volume measure, this restriction does essentially nothing). If $K = 1$, we can compute the expected value of Ω using a classical stochastic gradient descent (Algorithm 4). If $K > 1$, however, label switching may occur: There may be a group G acting on $\{1, 2, \dots, K\}$ that reindexes the elements of the product $\mathrm{Conf}_K(M)$ without affecting likelihood. This invalidates the expectation computed by Algorithm 4.

In this case, we need to work in the quotient $\mathrm{Conf}_K(M)/G$. Two key examples for G will be the symmetric group S_K of permutations and the cyclic group C_K of cyclic permutations. When $G = S_K$ we simply recover the Kth unordered configuration space, typically denoted $\mathrm{UConf}_K(M)$.

$\mathrm{UConf}_K(M)$ is a Riemannian manifold with structure inherited from the product metric on $\mathrm{Conf}_K(M)$ and has the property:

$$d_{\mathrm{UConf}_K(M)}([(p_1, \dots, p_K)], [(q_1, \dots, q_K)]) = \min_{\sigma \in S_K} d_{M^K}((p_1, \dots, p_K), (q_{\sigma(1)}, \dots, q_{\sigma(K)})). \tag{9.5}$$

The analogous fact holds for $\mathrm{Conf}_K(\mathscr{M})/G$ for other finite G via standard arguments (see e.g. Kobayashi (1995)). Thus, we may step in the gradient direction on the quotient by solving a suitable optimal transport matching problem.

Since G is finite, the map σ minimising (9.5) is computable algorithmically. When $G = C_K$, we simply enumerate all K cyclic permutations of $(q_1, \dots, q_K)$ and choose the one closest to p. When $G = S_K$, we can recover σ by solving a linear assignment problem with cost $\bar{c}_{ij} = d(p_i, q_j)^2$.

Algorithm 6 Barycenter for Gaussian Mixtures

Require: Distribution Ω

Ensure: Barycenter $p = (\mu_1^*, \Sigma_1^*) \dots, (\mu_K^*, \Sigma_K^*)$

1: Initialise the barycenter $p \sim \Omega$
2: for $t = 1, \dots$ do
3: Draw $((\mu_1, \Sigma_1) \dots, (\mu_K, \Sigma_K)) \sim \Omega$
4: Compute σ in (9.5)
5: for $i = 1, \dots, K$ do
6: $\mu_i^* = \mu_i^* - \eta(\mu_i^* - \mu_{\sigma(i)})$
7: $L_i^* = L_i^* - \eta(I - T^{\Sigma_i^* \Sigma_{\sigma*(i)}})L_i^*$

These properties suggest an adjustment of Algorithm 4 to account for G. Given a barycenter estimate $\mathrm{p} = (p_1, \dots, p_K)$ and a draw $\mathrm{q} = (q_1, \dots, q_K) \sim \Omega$: (1) align p and q by minimising the right-hand side of (9.5); (2) compute component-wise Riemannian gradients from p_i to $q_{\sigma(i)}$; and (3) step p toward q using the exponential map.

Algorithm 5 summarises our approach. It can be understood as stochastic gradient descent for z in (9.4), working in space $\mathrm{Conf}_K(M)$ rather than the quotient $\mathrm{Conf}_K(M)/G$. Theorem 9.2, however, gives an alternative interpretation. Construct a $|G|$-point empirical distribution $\mu = \frac{1}{|G|} \sum_{\sigma \in G} \delta_{\sigma \cdot \mathrm{p}}$ from the iterate p. After drawing $\mathrm{q} \sim \Omega$, we do the same to obtain $\nu \in P_2(\mathrm{Conf}_K(M))$. Then, our update can be understood as a stochastic Wasserstein gradient descent step of μ toward ν for problem (9.2). While this equivalent formulation would require $O(|G|)$ rather than $O(1)$ memory, it imparts the theoretical perspective in §9.3, in particular a connection to the (convex) problem of Wasserstein barycenter computation.

In the supplementary, we prove the following theorem:

Theorem 9.3 (Ordering Recovery). *If $\mathcal{M} = \mathbb{R}$, with the standard metric, then:*

$$\mathrm{UConf}_K(M) \cong \{(u_1, \dots, u_K) \in \mathrm{Conf}_K(\mathbb{R}) \mid u_1 < \dots < u_K\} \subset \mathbb{R}^K.$$

Additionally, the single-orbit barycenter of Theorem 9.2 is unique and our algorithm provably converges.

This setting occurs when one's mixture model consists of evenly weighted components with only a single mean parameter for each in $\mathbb{R}$. The result relates our method to the classical approach of ordering these means for correspondence and shows that it is well-justified. The convergence of our algorithm leverages the convexity of $\mathrm{UConf}_K(\mathcal{M})$. The supplementary contains additional discussion (§B.2.3) about such "mean-only" models in $\mathbb{R}^d$ for $d > 1$. They lack the niceness of the $d = 1$ case, due to positive curvature. This curvature is problematic for convergence arguments (as it leads to potential non-uniqueness of barycenters), but we empirically find that our algorithm converges to reasonable results.

Mixtures of Gaussians. One particularly useful example involves estimating the parameters of a Gaussian mixture over $\mathbb{R}^d$. For simplicity, assume that all the mixture weights are equal. The manifold $\mathcal{M}$ is the set of all (μ, Σ) pairs: $\mathcal{M} \cong \mathbb{R}^d \times \mathcal{P}^d$ with $\mathcal{P}^d$ the set of positive definite symmetric matrices. This space can be endowed with the W_2 metric:

$$d((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))^2 = W_2^2(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = \|\mu_1 - \mu_2\|_2^2 + \mathfrak{B}^2(\Sigma_1, \Sigma_2), \tag{9.6}$$

where $\mathfrak{B}^2$ is the squared Bures metric $\mathfrak{B}^2(\Sigma_1, \Sigma_2) = \mathrm{Tr}[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}]$.

As the mean components inherit the structure of Euclidean space, we only need to compute Riemannian gradients and exponential maps for the Bures metric. Muzellec & Cuturi (2018) leverage the Cholesky decomposition to parameterise $\Sigma_i = L_i L_i^\intercal$. The gradient of the Bures metric then becomes:

$$\nabla_{L_1} \frac{1}{2}\mathfrak{B}(\Sigma_1, \Sigma_2) = (I - T^{\Sigma_1\Sigma_2})L_1 \quad \text{with} \quad T^{\Sigma_1\Sigma_2} = \Sigma_1^{-\frac{1}{2}}(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}$$

The 2-Wasserstein exponential map for SPD matrices is $\exp_\Sigma(\xi) = (I + \mathscr{L}_\Sigma(\xi))\Sigma(I + \mathscr{L}_\Sigma(\xi))$ where $\mathscr{L}_\Sigma(\xi)$ is the solution of this Lyapunov equation : $\mathscr{L}_\Sigma(\xi)\Sigma + \Sigma\mathscr{L}_\Sigma(\xi) = \xi$.

9.5 Results

In §9.4, we gave a symmetry-invariant, simple, and efficient algorithm for computing a Wasserstein barycenter to summarise a distribution subject to label switching. To verify empirically that our algorithm can efficiently address label switching, we test on two natural examples: estimating the parameters of a Gaus-

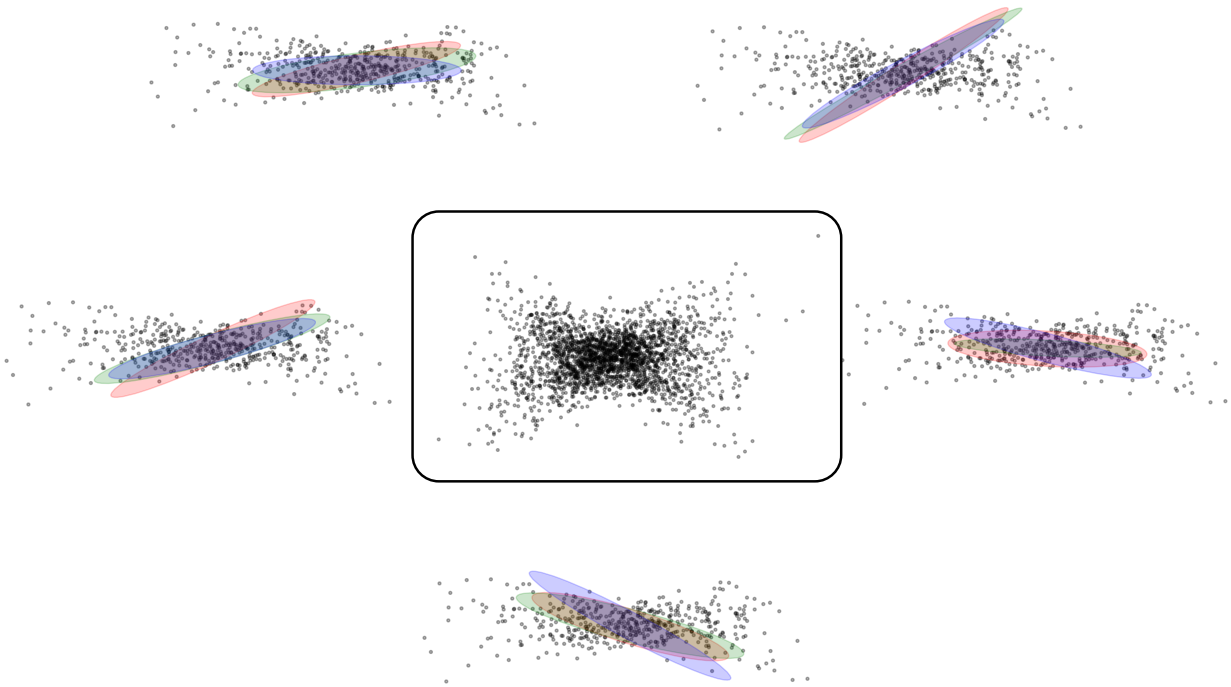


Figure 9-2: True covariances in blue, covariances from SGD in green and pivot in red. Original samples in the centre.

sian mixture model and a Bayesian instance of multi-reference alignment.

Estimating components of a Gaussian mixture. Our first scenario is estimating the parameters of a Gaussian mixture with $K > 1$ components. We use Hamiltonian Monte Carlo (HMC) to sample from the posterior distribution of a Gaussian mixture model. Naïve averaging does not yield a meaningful barycenter estimate, since the samples are not guaranteed to have the same label ordering. To resolve this ambiguity, we apply our method and two baselines: the pivotal reordering method (Marin et al., 2005) and Stephens' method (Stephens, 2000). The Stephens and Pivot methods relabel samples. Stephens minimises the Kullback–Leibler divergence between average classification distribution and classification distribution of each MCMC sample. Pivot aligns every sample to a prespecified sample (i.e. pivot) by solving a series of linear sum assignment problems. Pivot method requires pre-selecting a single sample for alignment — poor choice of the pivot sample leads to bad estimation quality, while making a "good" pivot choice may be highly non-trivial in practice. The default pivot choice is the MAP. Stephens method is more accurate, however it is expensive computationally and has large memory requirement.

To illustrate why pivoting fails, consider samples drawn from a mixture of five Gaussians with mean 0

and covariances $R_\theta M$ with $M = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0.1 \end{smallmatrix}\right)$ and R_θ a rotation of angle $\theta \in \{-\pi/12, -\pi/24, 0, \pi/12, \pi/24\}$ (Figure 9-2). The resulting pivot is uninformative for certain components. The underlying issue is that the pivot is chosen to maximise the posterior distribution. If this sample lies on the boundary of $\mathrm{Conf}_K(M)\,/S_K$, the pivot cannot be effectively used to realign samples. Quantitative results for this test case are in Table 9.1.

To get a better handle of the performance/accuracy trade-off for the three methods, we run an additional experiment. We draw samples from a mixture of five Gaussians over $\mathbb{R}^5$ with means $0.5e_i$, where $e_i \in \mathbb{R}^5$ is the i-th standard basis vector with $i \in \{1, \ldots, 5\}$, and covariances $0.4I_{5\times5}$. We implement HMC sampler using `Stan` (Carpenter et al., 2017), with four chains discarding 500 burn-in samples and keeping 500 per chain. Then we compare the three methods, increasing the number of samples to which they have access. We measure relative error as a function of wall clock time and number of samples (Figure 9-3). The resulting plots align with our intuition: pivoting obtains a suboptimal solution quickly, but if a more accurate solution is desired, it is better to run our SGD algorithm.

	Pivot	Stephens	SGD
Error (abs)	1.65	1.26	1.47
Time (s)	1.4	54	7.5

Table 9.1: Absolute error & timings

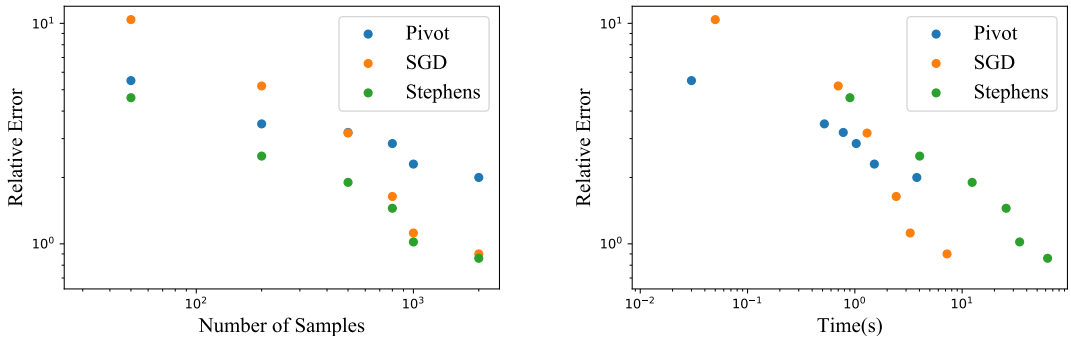


Figure 9-3: Relative error as a function of (a) number of samples and (b) time.

Multi-reference alignment. A different problem to which we can apply our methods is *multi-reference alignment* (Zwart et al., 2003; Bandeira et al., 2014). We wish to reconstruct a template signal $x \in \mathbb{R}^K$ given noisy and cyclically shifted samples $y \sim g \cdot x + \mathcal{N}(0, \sigma^2 I)$, where $g \in C_K$ acts by cyclic permutation. These observations correspond to a mixture model with K components $\mathcal{N}(g \cdot x, \sigma^2 I)$ for $g \in C_K$ (Bandeira et al., 2017). We simulated draws from this distribution using Markov Chain Monte Carlo

(MCMC), where each draw applies a random cyclic permutation and adds Gaussian noise (Figure 9-4a). The sampler we used was a Gibbs Sampler (Casella & George, 1992). To reconstruct the signal, we first compute a barycenter using Algorithm 5, giving a reference point to which we can align the noisy signals; we then average the aligned samples. Reconstructed signals for different σ's are in Figure 9-4b. To evaluate quantitatively, we compute the relative error of the reconstruction as a function of signal-to-noise ratio SNR $= \|x\|^2/K\sigma^2$ (Figure 9-4c).

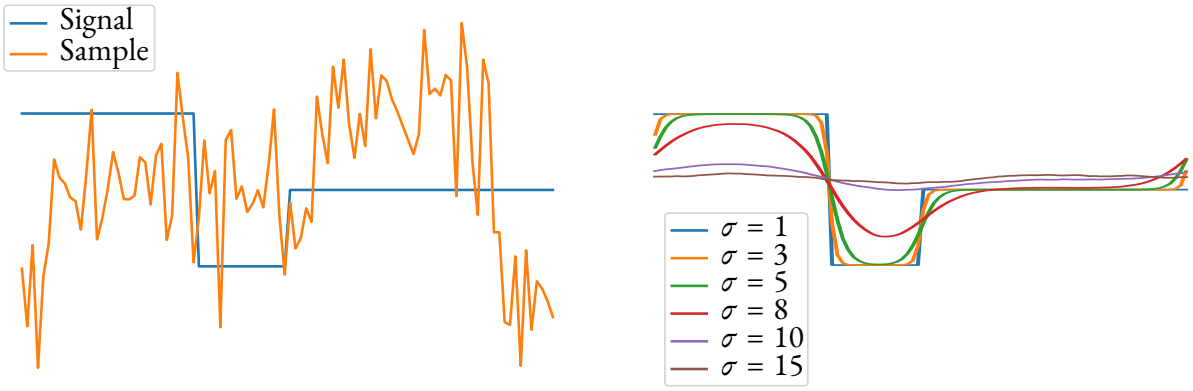


Figure 9-4: Reconstruction of a signal from shifted and noisy observations. (a) The true signal is plotted in blue against a random shifted and noisy draw from the MCMC chain. (b) Reconstructed signals at varying values of noise. (c) Relative error as a function of SNR.

9.6 Discussion and conclusion

The issue underlying label switching is the existence of a group acting on the space of parameters. This group-theoretic abstraction allows us to relate a widely-recognised problem in Bayesian inference to Wasserstein barycenters from optimal transport. Beyond theoretical interest, this connection suggests a well-posed and easily-solved optimisation method for alleviating label switching in practice.

The new structure we have revealed in the label switching problem opens several avenues for further inquiry. Most importantly, (9.4) yields a simple algorithm, but this algorithm is only well-understood when the Fréchet mean is unique. This leads to two questions: When can we prove uniqueness of the mean? More generally, are there efficient algorithms for computing barycenters in $P_2(X)^G$?

Finding faster algorithms for computing barycenters under the constraints of Lemma 9.3.2 provides an unexplored and highly-structured instance of the barycenter problem. Current approaches, such as those by Cuturi & Doucet (2014) and Claici et al. (2018) are too slow and not tailored to the demands of

our application, since each measure is supported on $K!$ points and the barycenter may not share support with the input measures. Moreover, after incorporating an HMC sampler or similar piece of machinery, our task likely requires taking the barycenter of an infinitely large set of distributions. The key to this problem is to exploit the symmetry of the support of the input measures and the barycenter.

10

Hierarchical Structure: Discussion

The examples we have seen of hierarchical structure may seem limited. Indeed, how could we generalise the approach in Chapter 9 to data that does not exhibit the same group invariance?

Recall, however, what we did in Chapter I where the goal was to approximate a distribution by a finite point set. One way to compute a distance between two distributions μ and ν is to compute finite approximations of both, and then compute the transport cost between the finitely supported measures.

Instead of simply computing the transport distance between the distributions we obtain, we could use a hierarchical approach for a better solution. Recall that each point in the quantization is assigned a region in the input distribution. If we can compute transport between such regions more easily than we can between the original distributions (or even approximate such a transport cost), then any problem that admits a quantization also admits a hierarchical approximation to the true transport cost.

Using the notation of Chapter 2, what we just described can be written as

$$W_p(\mu,\nu) \approx W_p\left(\sum_{i=1}^{n} \frac{1}{n}\delta_{\mu\llcorner V_i^v}, \sum_{j=1}^{n} \frac{1}{n}\delta_{\nu\llcorner V_j^w}\right).$$

The ground metric on the restrictions is itself transport, and, in certain cases, can be computed much faster than the full problem.

Part III

Optimal Transport on Discrete Surfaces

II

Introduction

Wherein we step into curved space and see how ideas from fluid dynamics can inform algorithms for solving the optimal transport problem. We show applications to solving measure valued Dirichlet problems, and computing gradient flows on the space of distributions.

Our journey has taken us from generic quantization tools through to specialised transport algorithms for data that exhibits certain structural patterns. What we have not discussed, and have often assumed was easy is computing the Wasserstein distance on a given metric space.

The reason we have avoided this discussion is that frequently the problem simply boils down to a linear program with n^2 variables for which there exist a slew of algorithms. But all of these algorithms assume that the cost matrix $C_{i,j} = |x_i - y_j|^p$ can be computed quickly.

What if this is not the case?

Unfortunately, the linear programming formulation of the optimal transport problem does not allow us to compute costs on the fly, but there are cases where this can be done.

A simple example is computing W_1 on a graph with the ground metric induced by the graph metric. We can take a page out of network flow algorithms, and think of transport as flow in and out of nodes in the graph. We are given measures μ and ν; let us measure the mass at node v. If we think of transport in physics terms, mass must flow into node v from μ and out of node v towards ν. The total of the mass flowing into v less the mass flowing out of v must equal the mass at v which is just $\mu_v - \nu_v$. If we write J for the flow along edges, V for the vertex set, E for the edge set of G, and c_e for the cost of edge e, we

recover Beckmann's problem:

$$\begin{aligned} \min_J \quad & \sum_{e \in E} J_e c_e \\ \text{subject to} \quad & \sum_{w \in N(v)} J_{(v,w)} = \mu_v - \nu_v. \end{aligned}$$

The more familiar version appears if we replace the finite graph with a surface M, and hence replace J with a vector field:

$$\begin{aligned} \min_J \quad & \int_M \|J\| \, \mathrm{d}x \\ \text{subject to} \quad & \nabla \cdot J(x) = \mu(x) - \nu(x) \\ & J(x) \cdot n(x) = 0 \end{aligned} \tag{11.1}$$

This is a simple formulation for computing W_1 when geodesic distances are not known beforehand, and it works for any manifold M, making this formulation particularly useful for problems in computer graphics(Solomon et al., 2014).

Can we do the same for W_2? Unfortunately, not readily. The issue at hand is that geodesic distances can be computed greedily by summing up costs along the shortest path, but the same is not true for squared geodesic costs. The right perspective (and formulation) was uncovered by Benamou & Brenier (2000), but their discretisation does not preserve Riemannian structure, and thus has limited practical uses. We fix this problem in what follows.

This chapter is based on Lavenant et al. (2018).

12

Dynamical Optimal Transport on Discrete Surfaces

12.1 Introduction

Probability distributions are key objects in geometry processing that can encode a variety of quantities, including uncertain feature locations on a surface, colour histograms, and physical measurements like the density of a fluid. A central problem related to distributions is that of *interpolation*: Given two probability distributions over a fixed domain, how can one transition smoothly from the first to the second?

Optimal transport gives one potential solution. This theory lifts the geometric structure of a surface to a Riemannian structure on the space of probability distributions over the surface, the latter being endowed with the so-called *Wasserstein* metric; the set of distributions equipped with this metric is sometimes called Wasserstein space. To interpolate between two probability distributions, one computes a *geodesic* in Wasserstein space between the two. This definition is sometimes referred to as McCann's displacement interpolation (McCann, 1997), applied to graphics e.g. in (Bonneel et al., 2011).

Even though optimal transport theory is now well-understood (Villani, 2003, 2008; Santambrogio, 2015), the interpolation problem remains challenging numerically. Related problems, like the computation of Wasserstein distances or barycenters in Wasserstein space, can be tackled by fast and scalable algorithms like entropic regularisation or semi-discrete methods, developed only a few years ago. Most of these methods, however, fail to reproduce the Riemannian structure of Wasserstein space and/or are prone to diffusion: The interpolation between two peaked probability distributions is more diffuse in the midpoint than optimal transport theory suggests. This drawback can inhibit application of transport

in computer graphics practice, in which blurry interpolants are often undesirable.

As an alternative, we define a Riemannian structure on the space of probability distributions over a discrete surface, designed to mimic that of the Wasserstein distance between distributions over a smooth manifold. Our construction is inspired by the Benamou–Brenier formula (Benamou & Brenier, 2000), previously discretised only on flat grids without structure preservation. This Riemannian structure automatically defines geodesics and distances between probability distributions. In particular, the geodesic problem can be recast as a convex problem and be tackled by iterative methods phrased using local operators familiar in geometry processing and finite elements (gradients, divergence and Laplacian on the surface). Our method does not require precomputation of pairwise distances between points on the surface.

Compared to other methods, our interpolation can be rephrased as a geodesic problem and numerically exhibits less diffusion when interpolating between peaked distributions. In cases where the sharpness captured by our method and predicted by optimal transport theory is undesirable visually, we provide a quadratic regulariser that *controllably* reduces congestion of the computed interpolant; unlike entropically-regularised transport, however, our optimisation problem does not degenerate or become harder to solve when the regularisation term vanishes. Although the computation of interpolants remains quite slow for meshes with more than a few thousand vertices and improving the scalability of numerical routines used to optimise our convex objective remains a challenging task for future work, we demonstrate application to tasks derived from transport, e.g. computation of harmonic mappings into Wasserstein space and integration of gradient flows.

In addition to our algorithmic contributions, we regard our work as a key theoretical step toward making optimal transport compatible with the language of discrete differential geometry (DDG). Our Riemannian metric induces a *true* geodesic distance—with a triangle inequality—on the space of distributions over a triangulated surface expressed using one value per vertex. Inspired by an analogous construction on graphs (Maas, 2011), we leverage a non-obvious observation that a strong contender for structure-preserving discrete transport on meshes actually involves a real-valued external time variable, rather than discretising transport as a linear program as in most previous work. The resulting geodesic problem naturally preserves convexity and other key properties from the theoretical case while suggesting

an effective computational technique.

12.2 Related work

12.2.1 Linear programming and regularisation

Landmark work by Kantorovich (Kantorovich, 1942) showed that optimal transport can be phrased as a linear programming problem. If both probability distributions have finite support, we end up with a finite-dimensional linear program solvable using standard convex programming techniques. A variety of solvers has been designed to tackle this linear program, which exploit the particular structure of the objective functional (Edmonds & Karp, 1972; Klein, 1967; Orlin, 1997). These methods, however, usually require as input the *pairwise* distance matrix, a dense matrix that scales quadratically in the size of the support and is difficult to evaluate if the points are on a curved space.

A landmark paper by Cuturi (Cuturi, 2013) reinvigorated interest in numerical transport by proposing adding an *entropic regulariser* to the problem, leading to the efficient *Sinkhorn* (or *matrix rebalancing*) algorithm. This algorithm, which involves iteratively rescaling the rows and columns of a kernel in the cost matrix, is highly parallelizable and well-suited to GPU architectures. When the cost matrix involves squared geodesic distances along a discrete surface, Solomon et al. (Solomon et al., 2015) showed that Sinkhorn iterations can be written in terms of heat diffusion operators, eliminating the need to store the cost matrix explicitly. While they are efficient, these entropically-regularised techniques suffer from diffusion, making them less relevant to problems in which measures are sharp or peaked. They also do not define true distances on the space of distributions over mesh vertices.

When the transport cost is equal to geodesic distance, i.e. the 1-Wasserstein distance, optimal transport is equivalent to the *Beckmann problem* (Santambrogio, 2015, Chapter 4), for which specific and efficient algorithms can be designed (Solomon et al., 2014; Li et al., 2018). These methods cannot be applied to the quadratic Wasserstein distance, which is needed to make transport-based interpolation nontrivial, namely to recover McCann's displacement interpolation (McCann, 1997). In particular, the optimal transport problem defining the 1-Wasserstein distance does not come with a time dependency allowing to define a smooth interpolation and suffers from non-uniqueness coming from the lack of strict convexity.

12.2.2 Semi-discrete optimal transport

When one of the distributions has a density w.r.t. Lebesgue while the other one is discrete, the transport problem can be reduced to a finite-dimensional convex problem whose number of unknowns scales with the cardinality of the support of the discrete distribution. Leveraging tools from computational geometry, this *semi-discrete* problem can be solved efficiently up to fairly large scale when the cost is Euclidean (Aurenhammer et al., 1998; Mérigot, 2011; De Goes et al., 2012; Lévy, 2015; Kitagawa et al., 2016).

Semi-discrete transport has been used to tackle problems for which the precise structure of the optimal transportation map is relevant, as in fluid dynamics (de Goes et al., 2015b; Mérigot & Mirebeau, 2016; Gallouët & Mérigot, 2017). It also has been used for approximating barycenters in the stochastic case (Claici et al., 2018) and as a measure of proximity for shape reconstruction (de Goes et al., 2011; Digne et al., 2014). Extensions of semi-discrete transport to curved spaces can be found in (de Goes et al., 2014; Mérigot et al., 2018). Although they can be fast and give explicit transport maps, these methods are not suited for the application we have in mind: They rely on the computation of transport maps between two probability distributions that are not of the same nature (one is discrete, the other has a density) and hence cannot be used to implement a distance or interpolation cleanly.

12.2.3 Fluid dynamic formulations

By switching from Lagrangian to Eulerian descriptions of transport, Benamou and Brenier (Benamou & Brenier, 2000) proved that optimal transport could be rephrased using fluid dynamics: Instead of computing a coupling, they show that transport with quadratic costs is equivalent to finding a *time-varying* sequence of distributions smoothly interpolating between the two measures. The problem that they obtain is convex and solved via the Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2011). Proof of the convergence of ADMM in the infinite-dimensional setting (i.e. when neither time nor the geometric domain is discretised) is provided in (Guittet, 2003; Hug et al., 2020). Papadakis et al. (Papadakis et al., 2014) reread the ADMM iterations as a proximal splitting scheme and show how one can build different algorithms to solve the convex problem. This fluid dynamic formulation also appears in mean field games (Benamou & Carlier, 2015).

In all of the above work, however, the authors work in a flat space and use finite difference discreti-

sations of the densities and velocity fields. Hence their work does not contain a clear indication about how to handle the problem on a discrete curved space, and theoretical properties of their models *after* discretization remain unverified.

The algorithm for approximating 1-Wasserstein distances presented by Solomon et al. (Solomon et al., 2014) achieves some of the objectives mentioned above. Their vector field formulation is in some sense dynamical, and their distance satisfies properties like the triangle inequality after discretisation. As mentioned above, however, their optimisation problem lacks strict convexity and is not suitable for interpolation.

12.2.4 Dynamical transport on graphs and meshes

Maas (Maas, 2011) defines a Wasserstein distance between probability distributions over the vertices of a graph. The (finite-dimensional) space of distributions in this case inherits a Riemannian metric with some structure preserved from the infinite-dimensional definition; for instance, the gradient flow of entropy corresponds to a notion of heat flow along the graph. A similar structure is proposed by Chow et al. (Chow et al., 2019), but they recover a different heat flow. Erbar et al. (Erbar et al., 2020) propose a numerical algorithm for approximating the discrete Wasserstein distance introduced by Maas, but the distributions they produce have a tendency to diffuse along the graph. This flaw is not related to their numerical method but rather comes from the very definition of their optimal transport distance. It is also not obvious what is the best way to adapt their construction to discrete surfaces rather than graphs.

12.2.5 Interpolation and geodesics

Optimal transport is not the only way to interpolate between probability distributions; for instance, Azencot et al. (Azencot et al., 2016) use a time-independent velocity field to advect functions and match them. Their method, however, cannot be understood as a geodesic curve in the space of distributions. In another direction, Heeren et al. (Heeren et al., 2012) have provided an efficient way to discretise in time geodesics in a high-dimensional space of thin shells. Their formulation is not well-suited for optimal transport where direct discretisation of the Benamou–Brenier formula is possible. Finally, methods like (Panozzo et al., 2013) provide a means of averaging points on discrete surfaces, although it is not clear

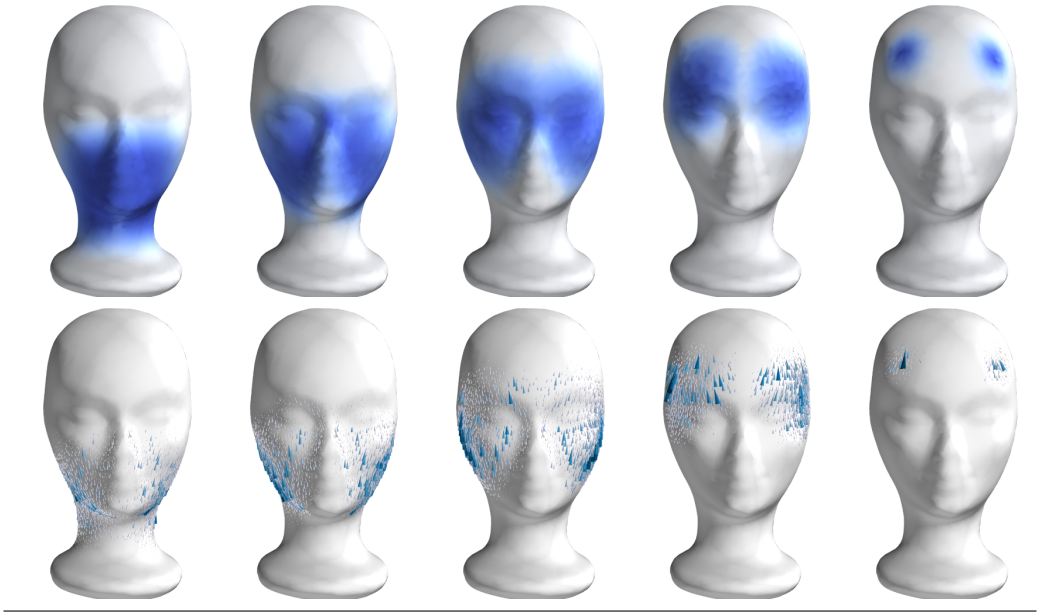


Figure 12-1: Top row: Interpolation of probability distributions. The left and right distributions are data and the middle ones are the output of our algorithm. Bottom row: Display of the momentum $\mathrm{m} = \mu \mathrm{v}$, where v is a time-dependent velocity-field advecting the left distribution on the right one. We have used the regularisation described in Subsection 12.5.4 with $\alpha = 0.1$.

how to extend them to the more general distribution case.

12.3 Optimal transport on a discrete surface

12.3.1 Optimal transport on manifolds

We begin by introducing briefly optimal transport theory on a smooth space. Let $\mathscr{M}$ be a connected and compact Riemannian manifold with metric $\langle\ ,\ \rangle$ and induced norm $\|\ \|$; define $d : \mathscr{M} \times \mathscr{M} \to \mathbb{R}_+$ to be geodesic distance.

Denote by $\mathscr{P}(\mathscr{M})$ the space of probability measures on $\mathscr{M}$. This space is endowed with the quadratic Wasserstein distance from optimal transport: If $\bar{\mu}^0, \bar{\mu}^1 \in \mathscr{P}(\mathscr{M})$, then the distance $W_2(\bar{\mu}^0, \bar{\mu}^1)$ between them is defined as

$$W_2^2(\bar{\mu}^0, \bar{\mu}^1) := \min_{\pi} \iint_{\mathscr{M}\times\mathscr{M}} \frac{1}{2} d(x,y)^2 \mathrm{d}\pi(x,y), \tag{12.1}$$

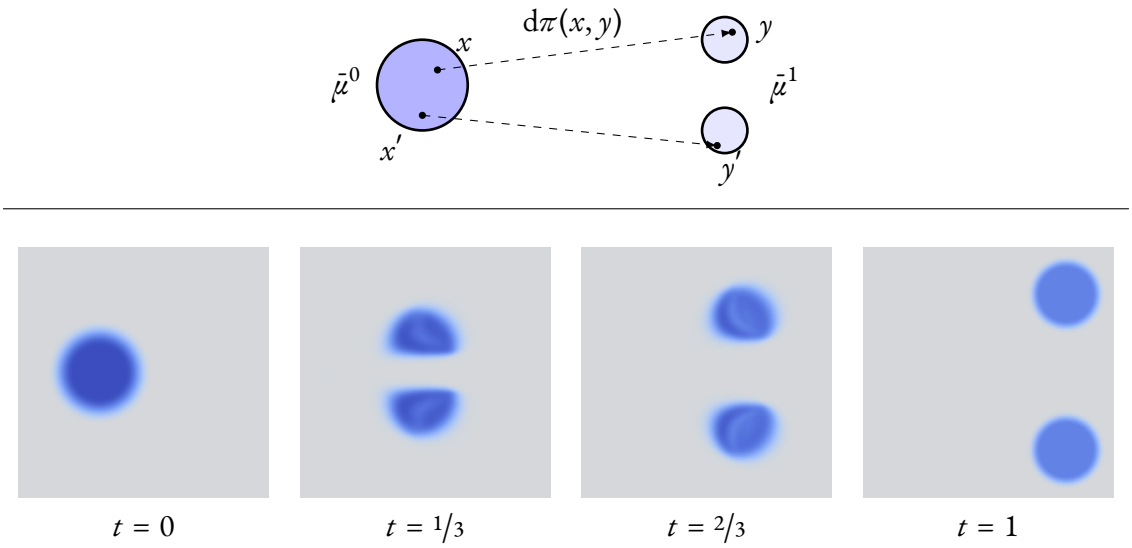


Figure 12-2: Top row: schematic view of the static formulation of optimal transport (12.1). The initial distribution $\bar{\mu}^0$ is on the left, and the final distribution $\bar{\mu}^1$ is on the right. The quantity $\mathrm{d}\pi(x,y)$ represents the amount of mass that is transported from x to y. The coupling π is chosen in such a way that the total cost is minimal. Bottom row: dynamical formulation between the same distributions (computed with the algorithm in Section 12.4). To go from the top to the bottom row, once one has the optimal π, a proportion $\mathrm{d}\pi(x,y)$ of particles follows the geodesic (in this case a straight line) between x and y with constant speed. The macroscopic result of all these motions is a time-varying sequence of distributions, displayed in blue.

where the minimum is taken over all probability measures π on the product space $\mathscr{M} \times \mathscr{M}$ whose first (resp. second) marginal is $\bar{\mu}^0$ (resp. $\bar{\mu}^1$).

The problem (12.1) can be interpreted as follows: $\mathrm{d}\pi(x, y)$ denotes the quantity of particles located at x that are sent to y, and the cost for such a displacement is $d(x, y)^2$. The constraint on the marginals enforces that π describes a way of moving the distribution of mass $\bar{\mu}^0$ onto $\bar{\mu}^1$. Thus, the variational problem (12.1) reads: Find the cheapest way π to send $\bar{\mu}^0$ onto $\bar{\mu}^1$, and the result (i.e. the minimal cost) is defined as the squared Wasserstein distance between $\bar{\mu}^0$ and $\bar{\mu}^1$. In some generic cases (Brenier, 1991; Gangbo & McCann, 1996), the optimal π is located on the graph of a map $T : \mathscr{M} \to \mathscr{M}$, which means that a particle $x \in \mathscr{M}$ is sent onto a unique location $y = T(x) \in \mathscr{M}$.

The space $(\mathscr{P}(\mathscr{M}), W_2)$ is a complete metric space (Santambrogio, 2015; Villani, 2003), and—at least formally—it has the structure of an (infinite-dimensional) Riemannian manifold. Revealing this manifold structure requires some manipulation and rephrasing of the original problem (12.1), detailed below.

As first noticed by Benamou and Brenier (Benamou & Brenier, 2000), the Wasserstein distance between $\bar{\mu}^0$ and $\bar{\mu}^1$ can be obtained by solving an alternative, physically-motivated optimisation problem:

$$W_2^2(\bar{\mu}^0, \bar{\mu}^1) = \begin{cases} \min_{\mu, \mathrm{v}} & \int_0^1 \int_{\mathscr{M}} \frac{1}{2} \|\mathrm{v}^t\|^2 \mathrm{d}\mu^t \mathrm{d}t \\ \text{s.t.} & \mu^0 = \bar{\mu}^0,\ \mu^1 = \bar{\mu}^1, \\ & \partial_t \mu + \nabla \cdot (\mu \mathrm{v}) = 0. \end{cases} \tag{12.2}$$

As we will have to deal with functions and vectors depending both on time and space, here and moving forward we adopt the following convention: Upper indices denote time, and lower indices denote space. Moreover, $t \in [0, 1]$ will denote an instant in time, and f will later denote a generic triangle (f for *face*) in a triangulation. In (12.2), the minimum is taken over all curves $\mu : [0, 1] \to \mathscr{P}(\mathscr{M})$ and all time-dependent velocity fields $\mathrm{v} : [0, 1] \times \mathscr{M} \to T\mathscr{M}$ such that the continuity equation $\partial_t \mu + \nabla \cdot (\mu \mathrm{v}) = 0$ is satisfied in the sense of distributions. The optimal curve μ is known as McCann's displacement interpolation (McCann, 1997).

The physical interpretation of this problem is as follows. Imagine probability distributions as distributions of mass, e.g. the density of a fluid. The curve μ represents an assembly of particles in motion, distributed as $\bar{\mu}^0$ at $t = 0$ and $\bar{\mu}^1$ at $t = 1$. At time t, a particle located at $x \in \mathscr{M}$ moves with velocity v_x^t.

The continuity equation $\partial_t \mu + \nabla \cdot (\mu v) = 0$ simply expresses the conservation of mass. For a given time t, the cost $\int_{\mathcal{M}} \frac{1}{2}\|v^t\|^2 d\mu^t$ is the total kinetic energy of all the particles. Hence, the cost minimised in (12.2), i.e. the integral w.r.t. time of the kinetic energy, is the *action* of the curve. As there is no congestion cost—that is, the particles do not interact with each other—(12.2) is the least-action principle for a pressureless gas.

Formulation (12.1) is static, since it directly determines the target for each particle at $t = 1$ given the arrangement at $t = 0$. On the other hand, (12.2) is dynamical, recovering a curve in $\mathcal{P}(\mathcal{M})$ interpolating smoothly between $\bar{\mu}^0$ and $\bar{\mu}^1$. To convert from the static to the dynamical formulation, one takes an optimal transport plan π from (12.1) and an assembly of particles distributed according to $\bar{\mu}^0$. If a particle located at $x \in \mathcal{M}$ at time $t = 0$ and is supposed, according to π, to be sent to $y \in \mathcal{M}$, then this particle follows a constant-speed geodesic along $\mathcal{M}$ from x to y. The optimal curve μ in (12.2) is exactly the resulting macroscopic motion of all the particles, illustrated in Figure 12-2.

Calling $m = \mu v$ the momentum and using the change of variables $(\mu, v) \leftrightarrow (\mu, m)$, problem (12.2) becomes convex, because the mapping $(\mu, v) \to 1/2\|v\|^2\mu$ is not jointly convex while $(\mu, m) \to 1/2\|m\|^2/\mu$ is. Its dual reads

$$W_2^2(\bar{\mu}^0, \bar{\mu}^1) = \begin{cases} \max_{\varphi} & \int_{\mathcal{M}} \varphi^1 d\bar{\mu}^1 - \int_{\mathcal{M}} \varphi^0 d\bar{\mu}^0 \\ \text{s.t.} & \partial_t \varphi + \frac{1}{2}\|\nabla \varphi\|^2 \leqslant 0 \text{ on } [0,1] \times \mathcal{M}, \end{cases} \tag{12.3}$$

where the maximisation is performed over real-valued functions $\varphi : [0,1] \times \mathcal{M} \to \mathbb{R}$ (Villani, 2008; Santambrogio, 2015). The relation $v = \nabla \varphi$ holds whenever v (resp. φ) is a minimiser (resp. maximiser) of the primal (resp. dual) problem. In particular, in (12.2), we can restrict ourselves to the set of v such that $v^t = \nabla \varphi^t$ for every $t \in [0,1]$.

Equation (12.2) defines a formal Riemannian structure on $\mathcal{P}(\mathcal{M})$ (Otto, 2001). Given $\mu \in \mathcal{P}(\mathcal{M})$ with a density bounded from below by a strictly positive constant, the tangent space $T_\mu \mathcal{P}(\mathcal{M})$ is identified as the set of functions $\delta\mu : \mathcal{M} \to \mathbb{R}$ with 0-mean: $\delta\mu$ is the partial derivative w.r.t. time of a curve whose value at time 0 is μ. If $\delta\mu \in T_\mu \mathcal{P}(\mathcal{M})$, we can compute $\varphi : \mathcal{M} \to \mathbb{R}$ the solution (unique up to translation by constants) of the elliptic equation

$$\nabla \cdot (\mu \nabla \varphi) = -\delta\mu. \tag{12.4}$$

Then, the norm of $\delta\mu$ is defined as

$$\|\delta\mu\|^2_{T_\mu\mathscr{P}(\mathscr{M})} := \frac{1}{2}\int_{\mathscr{M}} \|\nabla\varphi\|^2 \mathrm{d}\mu. \tag{12.5}$$

Endowed with this scalar product obtained from the polarisation identity $\langle x, y\rangle = \frac{1}{4}(\|x+y\|^2 - \|x-y\|^2)$, one can check, and the derivation appears in the supplemental material, that the Wasserstein distance W_2 can be interpreted as the geodesic distance induced by (12.4) and (12.5). This is precisely the content of the Benamou–Brenier formula (12.2).

One needs to assume $\mu \geqslant c > 0$ on $\mathscr{M}$ for the elliptic equation (12.4) to be well-posed. Nevertheless, one can still give a meaning to this Riemannian structure using tools from analysis in metric spaces (Ambrosio et al., 2008).

12.3.2 Discrete surfaces

The previous subsection contains only well-understood results. Let us now start our contribution: to mimic these definitions and properties when the manifold is replaced by a triangulated surface.

Instead of a smooth manifold $\mathscr{M}$, we consider the case where we only have access to a triangulated surface $S = (V, E, T)$, which consists of a set $V \subset \mathbb{R}^3$ of vertices, a set $E \subseteq V \times V$ of edges linking vertices, and a set $T \subseteq V \times V \times V$ of triangles containing exactly 3 vertices linked by 3 edges. For a given face $f \in T$, we denote by $V_f \subset V$ the set of vertices v such that $v \in f$; for a given vertex $v \in V$, we denote by $T_v \subset T$ the set of faces f such that $v \in f$. The area of a triangle $f \in T$ is denoted by $|f|$. Each vertex v is associated to a barycentric dual cell (see Figure 12-3) whose area, equal to $\frac{1}{3}\sum_{f\in T_v} |f|$, is denoted by $|v|$.

Following standard constructions from first-order finite elements (FEM), a scalar function on $\mathscr{M}$ will be seen as having one value per vertex, i.e. belonging to $\mathbb{R}^{|V|}$. A distribution $\mu \in \mathscr{M}$ will be also discretised by one value per vertex representing the density w.r.t. the volume measure. In other words, the volume of the dual cell centred at $v \in V$, measured with μ, is $|v|\mu_v$. We denote by $\mathscr{P}(S)$ the set of probability distributions on the discrete surface:

$$\mathscr{P}(S) := \left\{\mu \in \mathbb{R}^{|V|} \text{ s.t. } \mu_v \geqslant 0 \text{ for all } v \in V \text{ and } \sum_{v\in V} |v|\mu_v = 1\right\}. \tag{12.6}$$

For instance, the volume measure is represented by the vector in $\mathscr{P}(S)$ parallel to $(1, 1, \dots, 1)^\top$.

The set V of vertices can be interpreted as a discrete metric space, either by using directly the Euclidean distance on $\mathbb{R}^3$ or by some version of the discrete geodesic distance along S. Hence, a natural attempt to discretise the 2-Wasserstein distance would be to use (12.1) and replace d by the distance between vertices. As pointed out in (Maas, 2011; Gigli & Maas, 2013), however, this discretisation leads to a space without a smooth structure. For instance, there do not exist non-constant smooth (e.g., Lipschitz) curves valued in such a space; whereas in a space with a smooth structure (e.g. a Riemannian manifold), one expects the existence of non-constant Lipschitz curves, namely the (constant-speed) geodesics.

Let us briefly recall the argument. We take the simplest example of a space consisting of two points. If $X = \{x_0, x_1\}$ contains two points separated by a given distance ℓ, a probability distribution μ on X is characterised by a single number $\mu_{x_0} \in [0, 1]$, as $\mu_{x_1} = 1 - \mu_{x_0}$. If μ^t is a curve valued in $\mathscr{P}(X)$, one can compute $W_2(\mu^t, \mu^s) = \ell\sqrt{|\mu^t_{x_0} - \mu^s_{x_0}|}$. In particular, if μ is Lipschitz with Lipschitz constant L, our expression for W_2 implies $|\mu^t_{x_0} - \mu^s_{x_0}| \leqslant \frac{L^2}{\ell^2}|t - s|^2$. There is an exponent 2 on the r.h.s., but only 1 on the l.h.s.: it is precisely this discrepancy which is an issue. Indeed, dividing by $|t - s|$ on both side and letting $s \to t$, one sees that $t \mapsto \mu^t_{x_0}$ is differentiable everywhere with derivative 0, i.e. is constant.

For this reason, we prefer to discretise the Benamou–Brenier formulation (12.2), as it will automatically give a Riemannian structure on the space $\mathscr{P}(S)$. In this sense, the basic inspiration for our technique is the same as that of Maas (Maas, 2011), although on a triangulated surface we enjoy the added structure afforded by an embedded manifold approximation of the domain rather than an abstract graph.

As (12.2) involves velocity fields, a choice has to be made about their representation (de Goes et al., 2015a). To take full advantage of the triangulation, we want to use triangles and not only edges to define our objective functional. The latter choice leads to formulas similar to Maas (2011); Chow et al. (2019), which, as we say above, exhibit strongly diffuse geodesics. We prefer to represent vector fields on triangles. More precisely, a (piecewise-constant) velocity field v is represented as an element of $(\mathbb{R}^3)^{|T|}$, i.e. as one vector per triangle, with the constraint that v_f, which is a vector of $\mathbb{R}^3$, is parallel to the plane spanned by f, which means that our velocity fields lie in a subspace of dimension $2|T|$.

If $\varphi \in \mathbb{R}^{|V|}$ represents a real-valued function, we compute its gradient along the mesh using the first-order (piecewise-linear) finite element method (Brenner & Scott, 2007): For each triangle f, we compute

$\hat{\varphi}$, the unique affine function defined on f coinciding with φ on the vertices of f. Then, the gradient of φ in f is simply defined as the gradient of $\hat{\varphi}$ at any point of f; as the gradient is constant on each triangle, we need to store only one vector per triangle. Since this operator is linear, let us denote by $G \in \mathbb{R}^{3|T|\times|V|}$ its matrix representation. In particular, the Dirichlet energy of $\varphi \in \mathbb{R}^{|V|}$ is defined as

$$\mathrm{Dir}(\varphi) := \frac{1}{2}\sum_{f\in T} |f|\|(G\varphi)_f\|^2. \tag{12.7}$$

The sum is weighted by the areas of the triangles to discretise a surface integral. The first variation of this Dirichlet energy can be expressed in matrix form as $(G^\top M_T G)\varphi$, where $M_T \in \mathbb{R}^{3|T|\times 3|T|}$ is a diagonal weight matrix whose elements are the areas of the triangles. The matrix $G^\top M_T G$ is the so-called cotangent Laplace matrix of a triangulated surface (Pinkall & Polthier, 1993).

12.3.3 Dual problem on meshes

Let us introduce our discrete Benamou–Brenier formula by starting from its dual formulation (12.3). Since the objective functional is linear, its discrete counterpart is straightforward as both μ and φ are defined on vertices. On the other hand, in the constraint $\partial_t\varphi + \frac{1}{2}\|\nabla\varphi\|^2 \leqslant 0$, we would like to replace $\nabla\varphi$ by $G\varphi$ but then the two terms of the sum do not live on the same space.

The constraint $\partial_t\varphi + \frac{1}{2}\|\nabla\varphi\|^2 \leqslant 0$ is a priori not coercive. Suppose φ satisfies the constraint, and take another function ψ with the property that $\varphi + s\psi$ satisfies the constraint for arbitrarily large $s \geqslant 0$. Expanding the inequality $\partial_t(\varphi + s\psi) + \frac{1}{2}\|\nabla\varphi + s\nabla\psi\|^2 \leqslant 0$ and taking the limit $s \to +\infty$ shows that ψ satisfies this property if and only if $\|\nabla\psi\| = 0$ and $\partial_t\psi \leqslant 0$; these two conditions together imply that the objective functional in (12.3) is smaller when evaluated at $\varphi + s\psi$ rather than at φ. This is a property that we would like to keep at the discrete level. To do so, we enforce a discrete analogue of the constraint at each vertex of the mesh. To go from $\|G\varphi\|^2$, which is defined on triangles, to something defined on vertices, we *first* take the squared norm and *subsequently* average in space:[1]

Definition 12.1. *Let $\bar{\mu}_0, \bar{\mu}_1 \in \mathcal{P}(S)$. The discrete (quadratic) Wasserstein distance $W_d(\bar{\mu}_0, \bar{\mu}_1)$ is defined*

[1] If we do the opposite (averaging and then taking the square), there are spurious modes in the kernel of the quadratic part of the constraint, which leads to poor results when working with non-smooth densities.

as the solution of the following convex problem:

$$W_d^2(\bar{\mu}_0,\bar{\mu}_1) = \begin{cases} \sup_{\varphi} & \sum_{v\in V} |v|\varphi_v^1\bar{\mu}_v^1 - \sum_{v\in V}|v|\varphi_v^0\bar{\mu}_v^0 \\ s.t. & \partial_t\varphi_v^t + \dfrac{1}{2}\dfrac{\sum_{f\in T_v}|f|\|(G\varphi)_f^t\|^2}{3|v|} \leqslant 0 \\ & \text{for all } (t,v)\in[0,1]\times V, \end{cases} \tag{12.8}$$

where the unknown is a function $\varphi : [0,1]\times V \to \mathbb{R}$.

The denominator $3|v|$ is nothing else, by definition, than $\sum_{f\in T_v}|f|$. In particular, the value $\left(\sum_{f\in T_v}|f|\|(G\varphi)_f^t\|^2\right)(3|v|)^{-1}$ is the average, weighted by the areas of the triangles, of $\|(G\varphi)_f^t\|^2$ for $f\in T_v$. One can check that the same reasoning as above can be performed. Indeed, if $\varphi : [0,1]\times V\to\mathbb{R}$ satisfies the constraint in (12.8) and $\varphi + s\psi$ also satisfies it for arbitrarily large $s\geqslant 0$, it implies, taking $s\to+\infty$, that

$$\frac{1}{2}\frac{\sum_{f \text{ s.t. } v\in f}|f|\|(G\psi)_f^t\|^2}{3|v|} \leqslant 0. \tag{12.9}$$

This inequality must hold for all $(t,v)\in[0,1]\times V$. Thus, we conclude (and it is for this implication that it is important to average after taking squares) that $G\psi$ is identically 0. In other words, for all $t\in[0,1]$, the function ψ^t is constant over the discrete surface. Plugging this information back into the constraint in (12.8) and taking again $s\to+\infty$, we see that $\partial_t\psi\leqslant 0$. Hence, the value ψ^0 (which is constant over the surface) is larger than ψ^1. With this information ($G\psi = 0$ and $\partial_t\psi\leqslant 0$), the value of the objective functional must be smaller for $\varphi+s\psi$ than for φ as soon as $s\geqslant 0$.

12.3.4 Riemannian structure of the space of probabilities on a discrete surface

To recover an equation which looks like the primal formulation of the Benamou–Brenier formula (12.2), it is enough to write the dual of the discrete formulation (12.8). The latter formulation, as explained above, was important to justify the *choice* of the way we average quantities that do not live on the same grid.

We introduce additional notation to deal with the averaging of the density μ. If $\mu\in\mathcal{P}(S)$, we denote

by $\hat{\mu} \in \mathbb{R}^{|T|}$ the vector given by, for any $f \in T$,

$$\hat{\mu}_f = \frac{1}{3} \sum_{v \in V_f} \mu_v. \tag{12.10}$$

This is a natural way to average μ from vertices to triangles, appearing in the dual formulation given below:

Proposition 12.1. *The following identity holds:*

$$W_d^2(\bar{\mu}^0, \bar{\mu}^1) = \begin{cases} \min_{\mu, \mathrm{v}} & \int_0^1 \left(\sum_{f \in T} \frac{1}{2} \|\mathrm{v}_f^t\|^2 |f| \hat{\mu}_f^t \right) \mathrm{d}t \\ s.t. & \mu^0 = \bar{\mu}^0, \ \mu^1 = \bar{\mu}^1 \\ & \partial_t (M_V \mu_v^s) + (-G^\top M_T [\hat{\mu}^t \mathrm{v}^t])_v = 0 \\ & \textit{for all } (t, v) \in [0,1] \times V. \end{cases} \tag{12.11}$$

Recall that $M_T \in \mathbb{R}^{3|T| \times 3|T|}$ and $M_V \in \mathbb{R}^{|V| \times |V|}$ are diagonal matrices corresponding to multiplication by the area of the triangles and of the dual cells respectively. Then, $-G^\top M_T$ represents a discrete version of the (integrated) divergence operator, suggesting that the constraint can be interpreted as a discrete continuity equation. The derivation of this result, detailed in the supplemental material, relies on an inf-sup exchange, similar to the case of a smooth surface $\mathscr{M}$.

Proposition 12.1, very similar to (12.2), shows that W_d is the geodesic distance for a Riemannian structure on the space $\mathscr{P}(S)$, at least for non-vanishing densities. Let us detail the metric tensor for a density $\mu \in \mathscr{P}(S)$ with $\min_v \mu_v > 0$. As the set $\mathscr{P}(S)$ is a codimension-1 subset of the linear space $\mathbb{R}^{|V|}$, the tangent space at μ is naturally $\{x \in \mathbb{R}^{|V|} \text{ s.t. } \sum_{v \in V} |v| x_v = 0\}$. In analogy to (12.4), take $\delta\mu \in T_\mu \mathscr{P}(S)$. We call $\varphi \in \mathbb{R}^{|V|}$ a solution of

$$M_V \delta\mu = -(G^\top M_T M_{\hat{\mu}} G)\varphi, \tag{12.12}$$

where $M_{\hat{\mu}} \in \mathbb{R}^{3|T| \times 3|T|}$ is a diagonal matrix corresponding to multiplication on each triangle by $\hat{\mu}$. As $\hat{\mu} > 0$ everywhere on V, this equation is well-posed: The kernel of $G^\top M_T M_{\hat{\mu}} G$ is of dimension one (it consists only of the constant functions), and $M_V \delta\mu$ lies in the image of this operator. When the distribution μ is uniform, (12.12) boils down to a Poisson equation, as the operator $-(G^\top M_T M_{\hat{\mu}} G)$ is proportional to the

cotangent Laplacian.

One can then define the norm of $\partial\mu$ on the tangent space $T_\mu\mathcal{P}(S)$ as

$$\|\partial\mu\|^2_{T_\mu\mathcal{P}(S)} := \frac{1}{2}\sum_{f\in T}\|(G\varphi)_f\|^2|f|\hat{\mu}_f. \tag{12.13}$$

The function φ is unique up to an additive constant, which lies in the kernel of the matrix G, so this norm is well-defined.

To put everything in one formula, the scalar product $\langle\partial\mu,\partial\nu\rangle_{T_\mu\mathcal{P}(S)}$ between two elements of the tangent space at μ can be expressed as $(\partial\nu)^\top P_\mu(\partial\mu)$, where the matrix P_μ is expressed as

$$P_\mu = \frac{1}{2}M_V^\top G^{-\top}(M_{\hat{\mu}}M_T)^{-1}G^{-1}M_V. \tag{12.14}$$

One can check, and the derivation is provided in the supplemental material, that W_d is exactly the geodesic distance induced by this metric tensor.

Proposition 12.2. *The function $W_d : \mathcal{P}(S)\times\mathcal{P}(S)$ is a distance.*

Proof. It is a general fact that the geodesic distance on a manifold (defined by minimization over all possible trajectories) is a distance, see for instance (Jost, 2008, Section 1.4). □

A natural question is whether the space $(\mathcal{P}(S), W_d)$ converges to $(\mathcal{P}(\mathcal{M}), W_2)$ as S becomes a finer and finer discretisation of a manifold $\mathcal{M}$. For a discrete Wasserstein distance like the one of Maas (Maas, 2011), based on the graph structure of S—which corresponds to the case where velocity fields are discretised by their values on edges and a particular choice of scalar product—the answer is known to be positive in the case where $\mathcal{M}$ is the flat torus Gigli & Maas (2013); Trillos (2017) in the sense of Gromov–Hausdorff convergence of metric spaces, while a very recent work by Gladbach, Kopfer and Mass (Gladbach et al., 2018) has refined the analysis and exhibits necessary conditions for such a convergence to hold. The high technicality of the proofs of these results, however, indicates that the question for our particular definition is likely to be challenging and out of the scope of the present article.

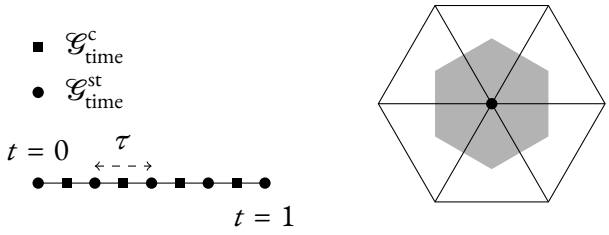


Figure 12-3: Left: temporal grids $\mathcal{G}^{\mathrm{c}}_{\mathrm{time}}$ and $\mathcal{G}^{\mathrm{st}}_{\mathrm{time}}$ for $N = 4$. Right: a vertex (•) surrounded by 6 adjacent triangles, the dual barycentric cell is in grey.

12.4 Time discretisation of the geodesic problem

12.4.1 Discrete geodesic

So far, we have defined a structure-preserving notion of optimal transport on a triangle mesh. While our model has many properties in common with the continuum version of transport, the resulting optimisation problem is infinite-dimensional since the unknown μ^t is indexed by a time variable $t \in [0, 1]$. Our next step is to derive a time discretisation that approximates this interpolant in practice. Put simply, we want to solve the geodesic problem, i.e., given $\bar{\mu}^0, \bar{\mu}^1 \in \mathcal{P}(S)$, we want to approximate the solution μ of (12.11). To this end, we discretise in time the dual problem (12.8).

Our infinite-dimensional problem can be classified as a second-order cone program (SOCP) (Boyd & Vandenberghe, 2004, Section 4.4.2); we choose a discretisation that preserves this structure. The main issue is that with a standard finite difference scheme, the derivative $\partial_t \varphi$ ends up on a grid staggered w.r.t. the one on which φ is defined. Hence, we average to define the constraint on a compatible grid. We apply the same idea as before: With the term involving $\|G\varphi\|^2$, we average *after* taking the square to avoid the introduction of any spurious null space.

Let N be the number of discretisation points in time. We consider two grids: the *staggered grid* $\mathcal{G}^{\mathrm{st}}_{\mathrm{time}} := \{k/N \,:\, k = 0, 1, \dots, N\}$ and the *centred grid* $\mathcal{G}^{\mathrm{c}}_{\mathrm{time}} := \{(k + 1/2)/N \,:\, k = 0, 1, \dots N - 1\}$, see Figure 12-3 The staggered grid has $N + 1$ elements whereas the centred one has only N. We call $\tau := 1/N$

the time step. The linear operator $D : \mathbb{R}^{\mathcal{G}^{\mathrm{st}}_{\mathrm{time}}} \to \mathbb{R}^{\mathcal{G}^{\mathrm{c}}_{\mathrm{time}}}$ defined by

$$(D\varphi)^t := \frac{\varphi^{t+\tau/2} - \varphi^{t-\tau/2}}{\tau}, \tag{12.15}$$

is a natural discretisation of the time derivative.

Next, we discretise $\varphi \in \mathbb{R}^{\mathcal{G}^{\mathrm{st}}_{\mathrm{time}} \times |V|}$ a function depending both on space and time. The constraint $\partial_t \varphi^t_v + \frac{1}{2} \frac{\sum_{f \in T_v} |f| \|(G\varphi)^t_f\|^2}{3|v|} \leqslant 0$ will be imposed on the centred grid $\mathcal{G}^{\mathrm{c}}_{\mathrm{time}}$. It is enough to replace $\partial_t \varphi$ by $D\varphi$. On the other hand, the term $\frac{1}{2} \frac{\sum_{f \in T_v} |f| \|(G\varphi)^t_f\|^2}{3|v|}$, which is defined on $\mathcal{G}^{\mathrm{st}}_{\mathrm{time}}$, will be also averaged in time. In other words, the fully discrete problem reads:

$$\boxed{\begin{array}{l} \text{Find } \varphi \in \mathbb{R}^{\mathcal{G}^{\mathrm{st}}_{\mathrm{time}} \times |V|} \text{ maximising} \\ \begin{cases} \displaystyle\sum_{v \in V} |v| \varphi^1_v \bar{\mu}^1_v - \sum_{v \in V} |v| \varphi^0_v \bar{\mu}^0_v \\ \text{s.t. } (D\varphi)^t_v + \dfrac{1}{2} \displaystyle\sum_{i \in \{-1,1\}} \frac{1}{2} \frac{\sum_{f \in T_v} |f| \|(G\varphi)^{t+i\tau/2}_f\|^2}{3|v|} \leqslant 0 \\ \text{for all } (t, v) \in \mathcal{G}^{\mathrm{c}}_{\mathrm{time}} \times V, \end{cases} \end{array}} \tag{12.16}$$

The constraint still stays quadratic, and hence the fully-discrete problem is still a SOCP.

12.4.2 Algorithm

To tackle (12.16) algorithmically, we follow Benamou and Brenier (Benamou & Brenier, 2000) by building an augmented Lagrangian and using the Alternating Direction Method of Multipliers (ADMM). The main issue is that the constraint is nonlocal—since it involves discrete derivatives—and nonlinear. We construct a splitting of the problem that decouples these two effects.

To this end, we introduce two additional variables A and B. We enforce the constraint $A = D\varphi$, and hence A is defined on the grid $\mathcal{G}^{\mathrm{c}}_{\mathrm{time}} \times V$. On the other hand, the variable B stores the values of $G\varphi$ but with some redundancy. Each $(G\varphi)^t_f$ appears in more than one inequality constraint in (12.16), and B is chosen so that each component of B appears in only one inequality constraint. In detail, B is defined on the grid $\mathcal{G}^{\mathrm{c}}_{\mathrm{time}} \times \{\pm 1\} \times T \times V$ with the constraint that $(f, v) \in T \times V$ is such that $v \in f$. We will impose the constraint that $\mathrm{B}^{t,i}_{f,v} = (G\varphi)^{t+i\tau/2}_f$ for all $(t, i, f, v) \in \mathcal{G}^{\mathrm{c}}_{\mathrm{time}} \times \{\pm 1\} \times T \times V$.

We introduce the notation $q = (A, \mathrm{B})$ and write $q = \Lambda\varphi$ if A, B satisfy the relations written above. Define

$$F(\varphi) = \sum_{v\in V} |v|\varphi_v^1\bar{\mu}_v^1 - \sum_{v\in V} |v|\varphi_v^0\bar{\mu}_v^0, \tag{12.17}$$

and C to be the function such that $C(A, \mathrm{B}) = C(q) = 0$ if

$$\begin{aligned}&\forall (t, v) \in \mathcal{G}_{\text{time}}^{\text{c}} \times V,\\ &\qquad A_v^t + \frac{1}{2}\sum_{i\in\{-1,1\}} \frac{1}{2}\frac{\sum_{f\in T_v} |f|\|\mathrm{B}_{v,f}^{t,i}\|^2}{3|v|} \leqslant 0\end{aligned} \tag{12.18}$$

and $-\infty$ otherwise. The discrete problem (12.16) can be written

$$\max_{q=\Lambda\varphi} F(\varphi) + C(q). \tag{12.19}$$

The idea is to introduce a Lagrange multiplier $\sigma = (\mu, \mathrm{m})$ associated to the constraint $q = \Lambda\varphi$ and to build the augmented Lagrangian

$$L(\varphi, q, \sigma) = F(\varphi) + C(q) + \langle \sigma, q - \Lambda\varphi\rangle - \frac{r}{2}\|q - \Lambda\varphi\|^2. \tag{12.20}$$

In this equation, $\langle \sigma, q - \Lambda\varphi\rangle = \langle \mu, A - D\varphi\rangle_V + \langle \mathrm{m}, \mathrm{B} - G\varphi\rangle_T$, where the scalar product $\langle\,,\,\rangle_V$ (resp. $\langle\,,\,\rangle_T$) is weighted by the areas of the vertices (resp. the triangles) and the time step τ.

The saddle points of the Lagrangian (12.20)—which do not depend on the parameter r—are precisely the solutions to the problem (12.16), and μ, the first component of σ associated to the constraint $A = D\varphi$, is an approximation of the time-continuous geodesic (12.11). On the other hand, the second component m is an approximation of the momentum μv.

To compute a saddle point, we use ADMM, which consists in iterations of the following form (Boyd et al., 2011):

1. Given q and σ, find φ that maximises L.
2. Given φ and σ, find q that maximises L.

3. Do a gradient descent step (with step r) to update σ.

The parameter $r > 0$ is arbitrary and tuned to speed up the convergence; see (Boyd et al., 2011) for discussion. In our case, details of the iterations are briefly presented below and summarised in Algorithm 7.

Algorithm 7 Geodesic Computation

```
function Geodesic(μ̄⁰, μ̄¹)
    Initialise φ, A, B, μ, m ← 0
    while PrimalResidual and DualResidual > ε do
        φ ← solution of (12.21)
        for s, v ∈ G^c_time × V do
            update A and B by solving (12.22)
        Update μ and m through (12.23)
    return μ
```

Maximisation w.r.t. φ The Lagrangian L is simply a quadratic function of φ, so its maximisation amounts to inverting the matrix $\mathcal{A}^\top \mathcal{A}$ which, in our case, behaves like a space-time Laplacian.

More precisely, writing $\varphi \in \mathbb{R}^{\mathcal{G}^{\mathrm{st}}_{\mathrm{time}} \times V}$ as a $(N+1) \times |V|$ matrix (with rows indexed by time and columns by space), the equation satisfied by a maximiser of L over φ reads

$$r\left[D^\top M_V D\varphi + 3(E^\top E)\varphi(G^\top M_T G)\right] = N(\bar{\mu}^1 I_{t=1} - \bar{\mu}^0 I_{t=0}) - D^\top M_V(\mu - r\mathcal{A}) - (\mathrm{m} - r\mathrm{B})M_T \tilde{G}^\top. \quad (12.21)$$

Again recall that the unknown here is φ; the remaining symbols are fixed matrices. In this equation, $E \in \mathbb{R}^{\mathcal{G}^{\mathrm{c}}_{\mathrm{time}} \times \mathcal{G}^{\mathrm{st}}_{\mathrm{time}}}$ stands for the averaging in time defined by $(E\varphi)^t = \frac{\varphi^{t-\tau/2} + \varphi^{t+\tau/2}}{2}$. The matrices $I_{t=0}$ and $I_{t=1} \in \mathbb{R}^{\mathcal{G}^{\mathrm{st}}_{\mathrm{time}} \times V}$ stand for the indicator of $t = 0$ (resp. $t = 1$), namely they contain zeros except on the first (resp. last) row which is full of ones. The factor 3 comes from the fact that each value of $(G\varphi)_f$ is duplicated 3 times in B, one for each vertex which belongs to f. The operator $\tilde{G}$ is almost the same as G but takes in account the fact that the values of $G\varphi$ are duplicated in B (hence in m): $\tilde{G}$ corresponds to the adjoint of the second component of the operator $\mathcal{A}$.

$(D^\top M_V D)$ is the discrete Laplacian in time, and $G^\top M_T G$ is the discrete Laplacian on S. In fact, (12.21) is a Poisson equation with a space-time Laplacian. Equation (12.21) admits more than one solution but they only differ by a constant whose value does not modify the value of L.

The linear operator to invert is the same at each iteration, and hence standard precomputation techniques can be used to speed up the application of its inverse.

Maximisation w.r.t. A, B The Lagragian L is also quadratic w.r.t. q, but there is a quadratic constraint on these two variables due to the presence of $C(q)$. Because of the redundancy in B, each component of A or B is subject to only one constraint. More precisely, we can check that one needs, for each $(t, v) \in \mathscr{G}^{\mathrm{c}}_{\mathrm{time}} \times V$, to minimise

$$|v| \left(A^t_v - (D\varphi)^t_v - \frac{1}{r}\mu^t_v \right)^2 + \frac{|f|}{2} \sum_{i\in\{\pm 1\}} \sum_{v\in V_f} \left\| \mathrm{B}^{t,i}_{f,v} - (G\varphi)^{t+i\tau/2}_f - \frac{1}{r}\mathrm{m}^{t,i}_{f,v} \right\|^2 \tag{12.22}$$

under the constraint (12.18). This minimization amounts to a Euclidean projection on the set of A, B satisfying (12.18), which can be carried out by solving a cubic equation in one variable, independently on each point of $\mathscr{G}^{\mathrm{c}}_{\mathrm{time}} \times V$. These equations are solved using Newton's method.

Dual update This gradient descent corresponds to the following operations:

$$\begin{aligned} \mu^t_v &\leftarrow \mu^t_v - r\left(A^t_v - (D\varphi)^t_v\right) \\ \mathrm{m}^{t,i}_{f,v} &\leftarrow \mathrm{m}^{t,i}_{f,v} - r\left(\mathrm{B}^{t,i}_{f,v} - (G\varphi)^{t+i\tau/2}_f\right), \end{aligned} \tag{12.23}$$

for any $(t, v) \in \mathscr{G}^{\mathrm{c}}_{\mathrm{time}} \times V$ and any $(t, i, f, v) \in \mathscr{G}^{\mathrm{c}}_{\mathrm{time}} \times \{\pm 1\} \times T \times V$.

12.5 Experiments

Recall that our main practical contribution is to be able interpolate between probability distributions using an optimal transport model that preserves structure from the non-discretised case. We will illustrate the robustness of our method: It can handle peaked distributions, and it lifts the intrinsic geometry of the discrete surface while being insensitive to the choice of mesh topology.

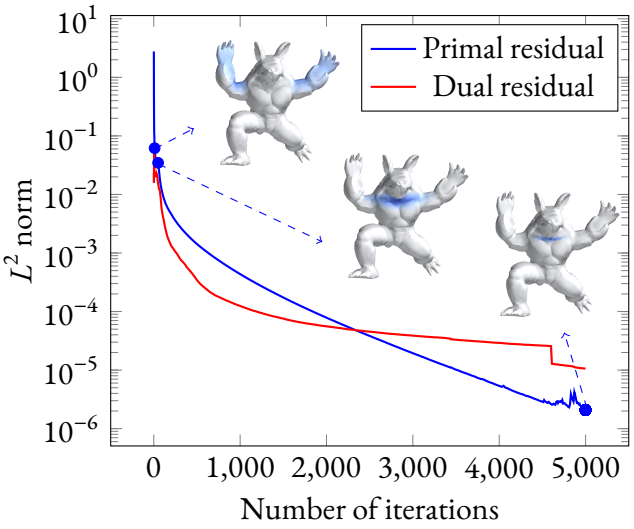


Figure 12-4: Amplitude of the primal and dual residual (Boyd et al., 2011, Section 3.3) in L^2 norm. The distributions $\bar{\mu}^0$ and $\bar{\mu}^1$ are delta functions located on respectively the right and left hand of the armadillo. We also show the midpoint $\mu^{1/2}$ for different numbers of iterations (10,50 and 5000). After a few hundred iterations, there is no visible difference in $\mu^{1/2}$. There is a jump in the value of the dual residual at around 4600 iterations. It is due to a change in the value of the parameter r, which is updated according to the heuristic rule presented in Section 3.4.1 of Boyd et al. (2011).

The typical computation is the following: We enter the data $\bar{\mu}^0, \bar{\mu}^1$ and compute a solution of the discrete problem (12.16). Then, we plot the evolution over time of μ, which approximates the geodesic in the Riemannian metric described in Subsection 12.3.4. As a byproduct of the optimisation process, we also obtain the optimal momentum $\mathrm{m} = \mu \mathrm{v}$, which can be also plotted, see Figure 12-1. The code used to conduct all our experiments is available at `https://github.com/HugoLav/DynamicalOTSurfaces`.

As the colour map is sometimes normalised independently for different time instants on the same interpolation curve, let us underscore this fact: For every example, we have checked numerically that the densities are always nonnegative and that mass is always preserved over time.

12.5.1 Convergence of the ADMM iterations

For fixed boundary data $\bar{\mu}^0$ and $\bar{\mu}^1$, we plot in Figure 12-4 the primal and dual error defined by Boyd et al. (Boyd et al., 2011), as a function of the number of iterations of the ADMM scheme. We usually need on the order of a few thousand iterations to satisfy our convergence criteria, this number being dependent

Table 12.1: Timing data for various meshes and boundary data from the figures (numbers listed in the table). N denotes the number of time discretisation points and α is the value of the congestion regularisation parameter (see Section 12.5.4). For the ADMM method, the number of iterations and timing are given. Iterations were stopped once an error of 10^{-4} was reached for the L^2 norm of both the primal and dual residual. One can see that the time per iteration depends on the size of the mesh and the temporal grid, but the number of iteration is quite insensitive to these parameters and rather depends on the boundary conditions and the regularisation parameter. For the CVX implementation of the optimisation problem, the solver time and the total time (includes CVX pre-processing) are given. Standard precision settings were used, but are hard to interpret absolutely due to unknown algebraic rearrangement of the problem. * denotes that CVX reported a failure in this case. Results obtained on an 8-core 3.60GHz Intel i7 processor with 32GB RAM.

Mesh	Figure	N	$\lvert V\rvert$	$\lvert T\rvert$	α	ADMM Iters.	ADMM Time (s)	CVX Time (s)
Punctured sphere	10	13	1020	2024	0.02	546	16	27
Punctured sphere	10	31	1020	2024	0.02	547	47	122
Hand	8	13	1515	3026	0.02	846	47	47
Hand	8	31	1515	3026	0.02	858	97	191
Armadillo	7	31	5002	10000	0	929	332	882
Armadillo	7	63	5002	10000	0	808	649	3970
Armadillo	7	31	5002	10000	1	308	116	1054
Face	2	31	5002	10000	0.1	415	155	1944
Airplane	9	31	3772	7540	0.1	535	144	831
Planar square	3	31	11838	23242	0	565	473	11082

of the boundary data $\tilde{\mu}^0, \tilde{\mu}^1$ (the more diffuse, the fewer iterations are needed).

Because our objective functional is scaled according to the geometry of the mesh (i.e. scalar products are weighted by the areas of the triangles and the number of time steps), the number of iterations needed does not depend on the size of the resolution of the mesh nor the number of discretisation points in time, but the computation time needed per iteration does. Typical values of the timings are given in Table 12.1, they are of the order of 1 second per ADMM iterations for meshes with a few thousand vertices.

12.5.2 CVX implementation

Since the optimisation problem in Equation (12.25) is a convex cone problem, we have also used a straightforward implementation in CVX (Grant & Boyd, 2014, 2008), with Mosek as a solver ApS (2017). This approach is provided as a simpler alternative to the ADMM implementation, and has comparable performance on small meshes with standard precision settings (fewer than 1000 vertices). In general, it is difficult to compare the error thresholds across the two implementations due to algebraic rearrangements

performed by CVX. See Table 12.1.

12.5.3 Convergence with discretisation in space and time

As indicated in Section 12.3, it is not known theoretically whether our discrete distance converges to the true Wasserstein distance when the mesh is refined. This is also the case as far as the time discretisation is concerned; one could likely adapt the method of proof of Erbar et al. (Erbar et al., 2020), but doing so is out of the scope of this article.

In Figure 12-5, however, we present some experiments indicating that convergence under space and time refinement is likely to be true. These were conducted in the simplest case: translation of a given density on a flat space. For this problem, the ground truth is known, and for a flat space it is clear what it means to refine the mesh: We have use a regular triangle mesh with an increasing number of points per side. The error was evaluated at time $t = 1/2$ between the computed geodesic and the ground truth. As a measure of error, as the distributions are compactly supported, we use a total variation norm (in other words the L^1 norm between the densities) rather than the Kullback–Leibler divergence. As expected, we observe a decrease in error as the temporal and spatial meshes are refined.

12.5.4 Congestion and regularisation

In optimal transport there is no price paid for highly-congested densities. Imagine the probability distributions as an assembly of particles moving along the surface. Along a geodesic in Wasserstein space, each particle evolves in time by following a geodesic on the surface—but does not feel the presence of its neighbours.

Now imagine, due to the particular structure of the triangle mesh, there is a small shortcut in terms of geodesic distance through which all geodesics tend to concentrate. This is likely to appear near a hyperbolic vertex (Polthier & Schmies, 2006). Then all the particles have the incentive to take this shortcut, resulting in densely-populated zones, as they are not prone to congestion. As an example, see the first row of Figure 12-6 in which, to go from the left to the right of the armadillo, all the particles go through only two paths, leaving the rest of the mesh without any mass.

This effect, although visually unpleasant, would be observed on a smooth surface $\mathscr{M}$ as soon as

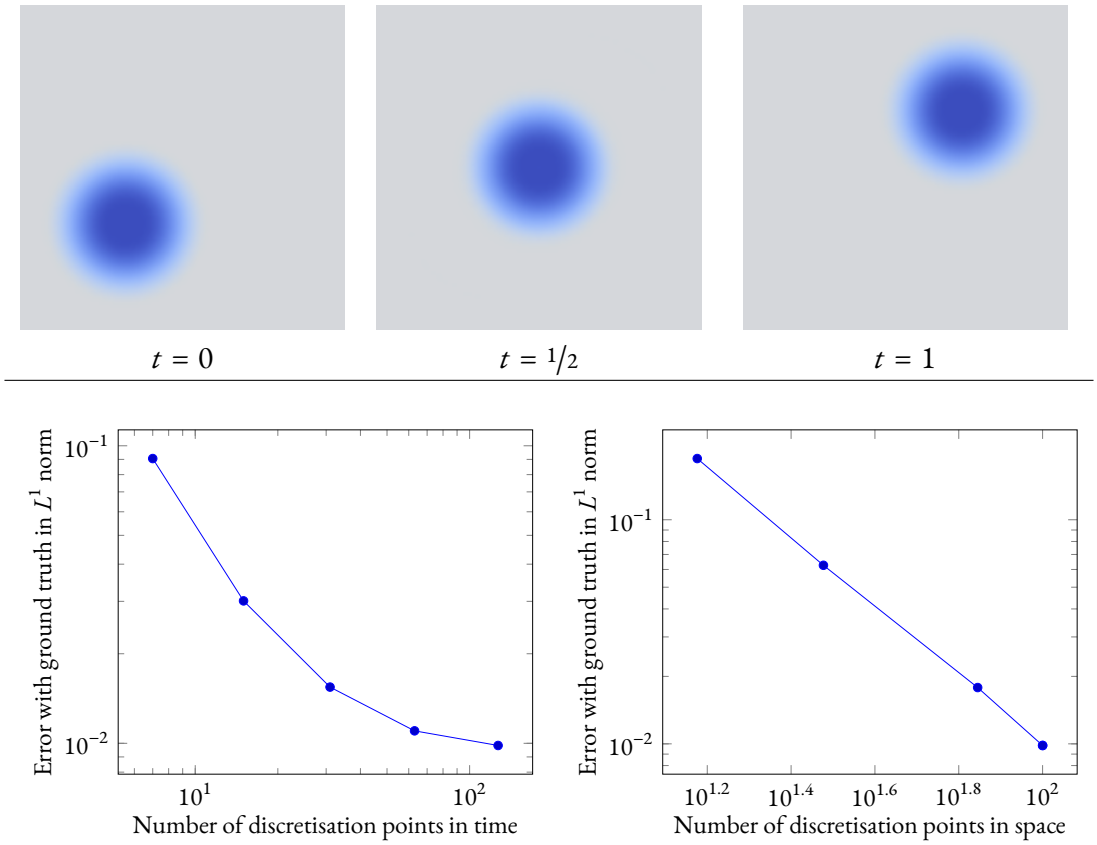


Figure 12-5: Top row: the test case. The inputs, i.e. probability distributions at times $t = 0$ and $t = 1$, correspond to the same density translated two different ways. Optimal transport predicts that at time $t = 1/2$ we should observe the same density again, but translated to the midpoint between the two inputs; this gives us ground truth we can use to verify our algorithm's output. Bottom row: convergence plots. On the left: error, measured in L^1 norm, where the mesh is fixed (regular triangle mesh with 100 points per side of the square) and the number N of discretisation points in time varies. On the right: error, measured in L^1 norm, where the number of discretization points in time is fixed (127 points) and the mesh is a regular triangle mesh whose number of points per side varies and is plotted on the x-axis.

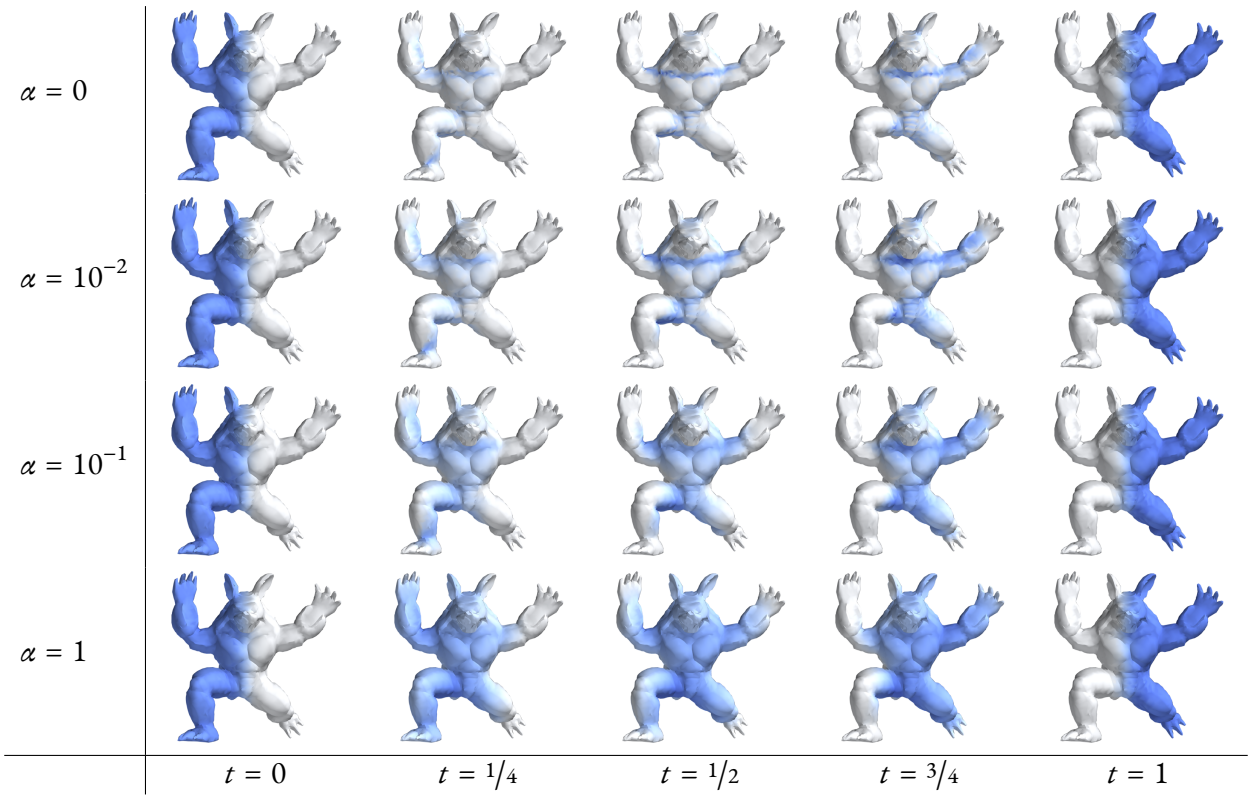


Figure 12-6: Effect of the regularising parameter α penalising congestion. On each row, the interpolation between the same boundary data (distributions located on the right and on the left of the armadillo) is shown. Different rows correspond to different to different values of α. The colour in each image is normalised independently from the others, explaining the change in intensity. Mass is always preserved along the interpolation.

geodesics concentrate in some regions. A way to remove this artefact is to penalise congestion; we can do so with little modification to the algorithm.

We penalise the densities by their L^2 norms: The choice of the exponent 2 is important, as it preserves the quadratic structure of the optimisation problem. Namely, we add to the Lagrangian (12.20) the term

$$\frac{\alpha\tau}{2}\sum_{t\in\mathscr{G}^c_{\text{time}}}\sum_{v\in V}|v||\mu_v^t|^2 = \sup_{\lambda}\sum_{t\in\mathscr{G}^c_{\text{time}}}\sum_{v\in V}\tau|v|\left(\lambda_v^t\mu_v^t - \frac{1}{2\alpha}(\lambda_v^t)^2\right), \tag{12.24}$$

where the parameter α tunes the scale of the congestion effect and $\lambda \in \mathbb{R}^{\mathscr{G}^c_{\text{time}}\times V}$ corresponds to the dual variable associated to the congestion constraint.

Using the notation from Section 12.4.2, one can write the problem as maximising

$$\max_{\hat{A}(\varphi,\lambda)=q} F(\varphi,\lambda) + C(q), \tag{12.25}$$

but this time

$$F(\varphi,\lambda) = (12.17) - \frac{1}{2\alpha}\sum_{t\in\mathscr{G}^c_{\text{time}}}\sum_{v\in V}\tau|v|(\lambda_v^t)^2 \tag{12.26}$$

and $\hat{A}(\varphi,\lambda) = (-\lambda, 0) - A(\varphi)$. Then one runs exactly the same algorithm, with a straightforward adaptation of the update formulas.

After regularisation, the interpolation is no longer a geodesic. For instance, the interpolation between two instances of the same probability distribution is not constant in time, because the L^2 norm potentially can be reduced by diffusing outward in the intermediate time steps. On the other hand, undesirable sharp features and oscillations can be removed, as seen in Figure 12-6. Note that regardless of the level of regularisation, the interpolating curves are still valued in $\mathscr{P}(S)$, i.e. mass is still preserved along the interpolation.

The tuning of the parameter α allows our method to be robust to noisy mesh inputs, as shown in Figure 12-7. Noisy meshes have more local variation in curvature, leading to a higher tendency for congested trajectories, but this can be tamed via greater regularisation.

Recall that the dynamical formulation of optimal transport can be interpreted as the least-action principle for a pressureless gas. The effect of the penalisation of congested densities can be seen, from

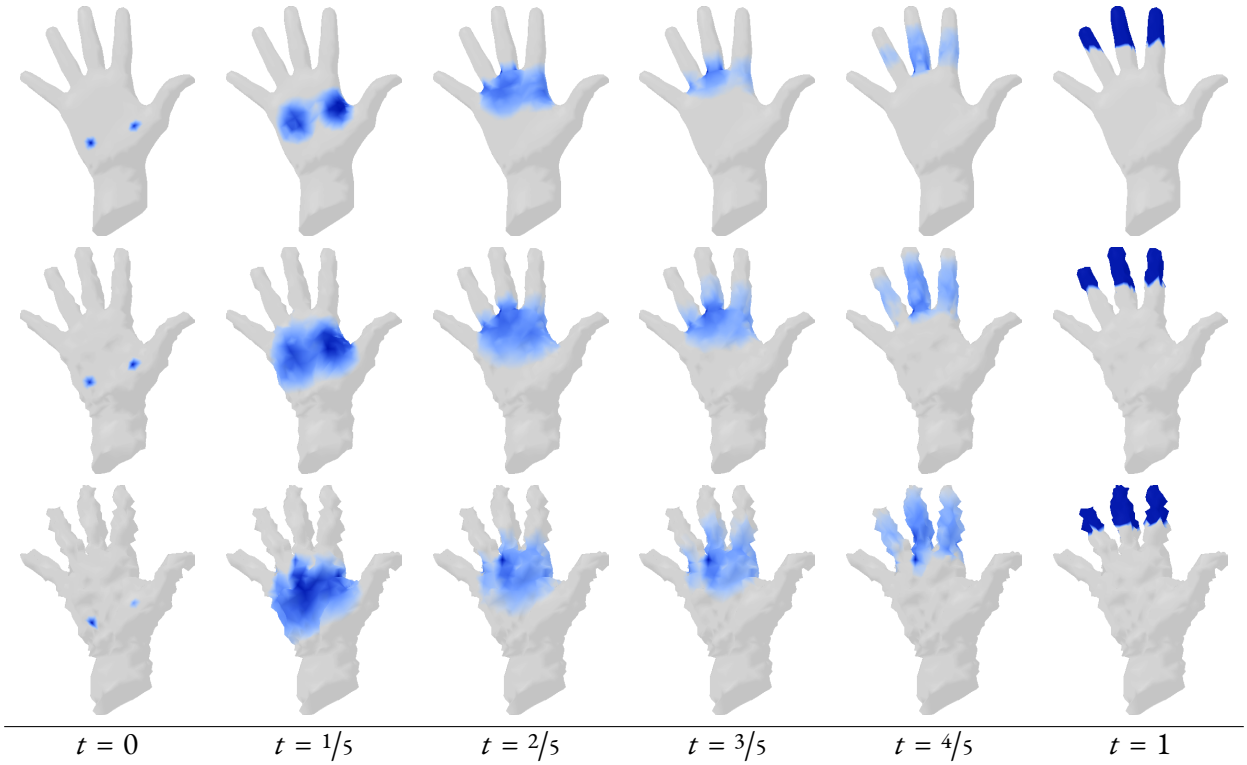


Figure 12-7: Robustness to noisy meshes, after adjusting the parameter α. Top row: original mesh, $\alpha = 0.02$; middle row: noisy mesh, $\alpha = 0.1$; bottom row: very noisy mesh, $\alpha = 0.2$. The bounding boxes of the meshes were of side length ~1.5. Noisy mesh vertices were obtained by uniformly random perturbation, in the normal direction, of magnitudes up to 0.02 and 0.04, for the middle and bottom row, respectively.

the modelling point of view, as adding a pressure force: the trajectories of the moving particles are no longer geodesics, they are bent by the pressure forces. The congestion term can also be see as an instance of variational mean field games, for which the augmented Lagrangian approach has been applied for flat spaces with grid discretisation (Benamou et al., 2017).

12.5.5 Intrinsic geometry

To illustrate the fact that the discrete Wasserstein metric is really associated to the geometric structure of the mesh, we perform the following experiment. We design a mesh where the right part is much coarser than the left one, and we let the density evolve. As one can see in Figure 12-8, the jump in coarseness does not affect the density and does not produce any numerical artefact.

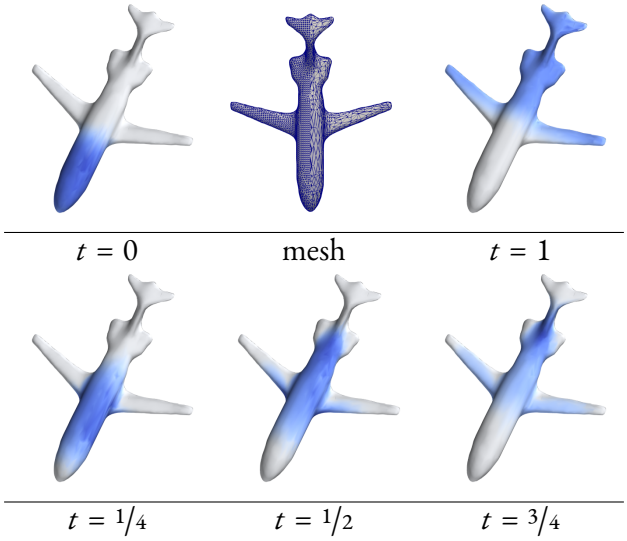


Figure 12-8: Top row: mesh and initial/final probability distributions. Notice the difference of coarseness in the mesh. Bottom row: interpolation shown at different times where no effect of the difference in coarseness is seen. We have used the regularisation described in Subsection 12.5.4 with $\alpha = 0.1$.

12.5.6 Arbitrary topologies

The discrete formulation that we have chosen applies without change to meshes with boundary and those of non-spherical topology. This is illustrated in Figure 12-9 with two meshes topologically equivalent to a disc and a torus.

In the first example, the interpolating distribution stays near the boundary, approximately following the geodesic between the means of the endpoint distributions. In the second example, one can see the initial distribution splitting to travel both ways to the other side of a handle, before merging again to achieve the final distribution.

12.5.7 Comparison to convolutional method

Solomon et al. (Solomon et al., 2015) provide a convolutional method for approximating the Wasserstein geodesic between two distributions supported on triangle meshes. Their approach solves a regularised optimal transport barycenter problem using a modified Sinkhorn algorithm, with a heat kernel taking the place of explicitly-calculated pairwise distances between vertices. As a result, their method blurs the

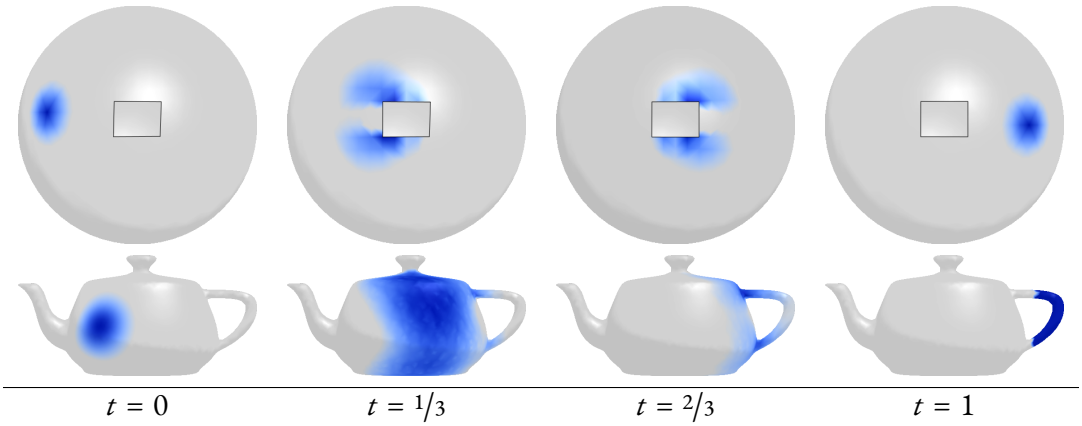


Figure 12-9: Our formulation easily handles non-spherical topologies. In the top row is a punctured sphere, and in the bottom row is a genus-1 teapot mesh. These interpolations were generated with $\alpha = 0.02$ and $\alpha = 0.2$, respectively.

input distributions, and the interpolated distributions are typically of higher entropy than the endpoints. This is combated with a nonconvex projection method that attempts to lower the entropy of intermediate distributions to an approximated bound.

In comparing our methods, we found that Solomon et al. (2015) also tends to produce interpolating distributions that do not travel with constant speed. This effect can be seen in Figures 12-10 and 12-11, where their interpolating distributions remain mostly stationary for times near $t = 0$ and $t = 1$, but move with high speed for times near $t = 1/2$. Loosening the entropy bound in the nonconvex step helps somewhat, but the problem persists regardless. Most likely this effect is due to the fact that the entropy reduction step of their algorithm is not geometry-aware but rather simply sharpens the regularised interpolant.

Our method does not suffer from this issue, and the spread of our interpolating distributions is comparable or better in both cases. Furthermore, unless the regulariser α is large, our interpolating distributions tend to diffuse only in the direction of the geodesics along which particles are travelling, which better mimics the behaviour of Wasserstein geodesics; this diffusion is reduced by adding more time steps to our interpolation problem.

Our formulation also has comparable runtimes to the convolutional method of Solomon et al. (Solomon et al., 2015). For instance, the implementation of the convolutional method provided by the authors of that paper took 57 and 141 seconds to converge, on the punctured sphere (1020 vertices) and teapot (3900

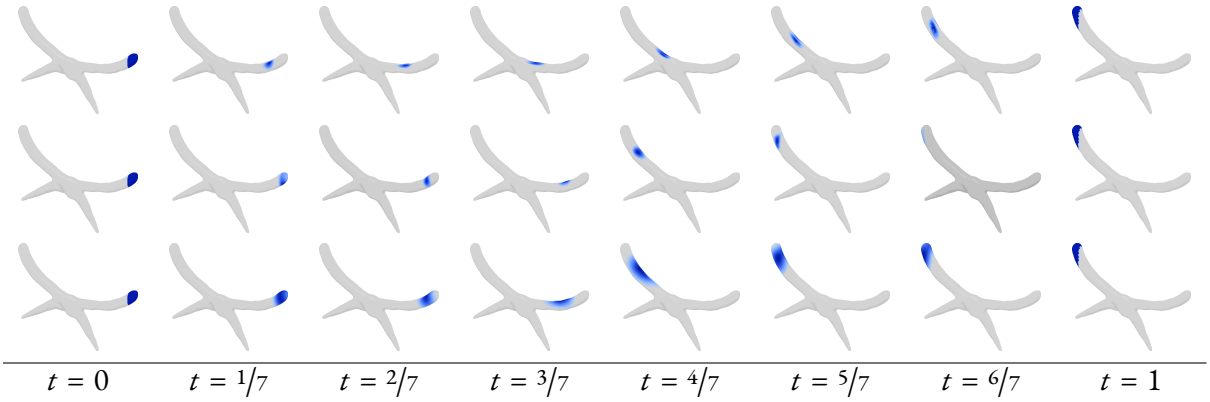


Figure 12-10: Constant-speed interpolation. Indicator distributions on handle ends of a pliers mesh are interpolated. Top row: our method, calculated with $\alpha = 0.001$; middle row: method of Solomon et al. (Solomon et al., 2015), calculated with entropy bounded by that of the endpoint distributions; bottom row: method of Solomon et al. Solomon et al. (2015), calculated with no entropy bound. As can be seen, the method of Solomon et al. Solomon et al. (2015) stays mostly stationary except for the middle frames.

vertices), respectively, for 13 time steps. This is to be put in comparison with the timings provided in Subsection 12.5.2.

The comparisons in this section were computed on a 3.60GHz Intel i7-7700 processor with 32GB of RAM. For the convolutional method, the heat kernel was used to diffuse to $t = 0.0015$ with 10 implicit Euler steps.

12.6 Discussion and conclusion

Although techniques using entropic regularisation or semi-discrete optimal transport can interpolate between distributions on a discrete surface, they do not provide a Riemannian structure and are subject to practical limitations that restrict the scenarios to which they can be applied. Using an intrinsic formulation of dynamical transport, we can realise the theoretical and practical potential of optimal transport on discrete domains enabled by the Riemannian structure on the space of probability distributions, the so-called Otto calculus. Our technique can be phrased in familiar language from discrete differential geometry and is implementable using standard tools in that domain. The key ingredients, namely first- and second-order operators in geometry processing (gradient, divergence, Laplacian) as well as SOCP optimisation, remain in the realm of what is already widely used.

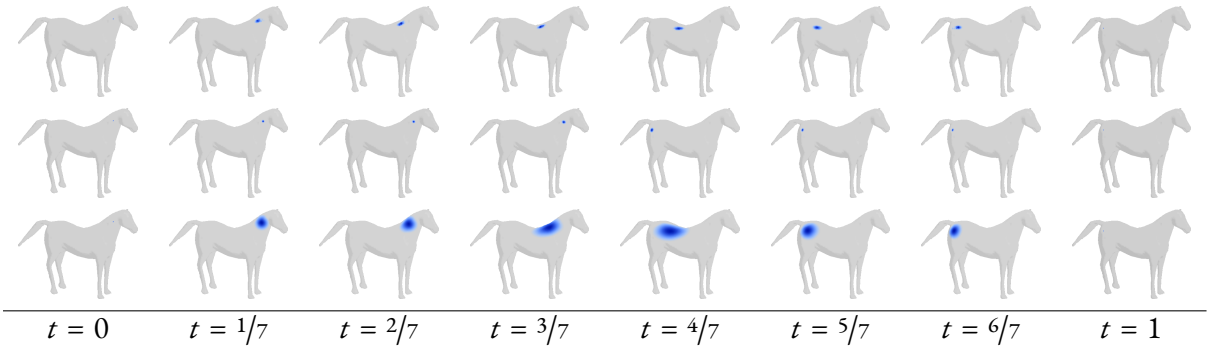


Figure 12-11: Constant-speed interpolation. Delta distributions on a horse mesh are interpolated. Top row: our method, calculated with $\alpha = 0.01$; middle row: method of Solomon et al. (Solomon et al., 2015), calculated with entropy bounded by that of endpoint distributions; bottom row: method of Solomon et al. Solomon et al. (2015), calculated with no entropy bound. For the middle row, the motion is even more concentrated in the middle frames. As seen in the bottom row, exclusion of the entropy bound helps somewhat, but the result still is mostly stationary, save for the middle frames.

We have demonstrated the power of our model by showing how it can handle a variety of geometries and peaked distributions, while introducing little diffusion. Mass may concentrate to yield a visually inelegant result, but this behaviour is at the core of optimal transport theory and expected: No price is paid for mass congestion, and hence any concentration of geodesics will result in a concentration of mass. Nevertheless, as we have shown, one can easily modify the optimisation problem to penalise congested densities, leading to smoother interpolants with a controllable level of diffusion. Unlike entropically-regularised transport, however, our optimisation problem does not degenerate as the coefficient in front of the regulariser vanishes.

Beyond evaluation of transport distances, our framework extends to support other tasks involving transport terms. We can reliably compute harmonic mappings valued in this discrete Wasserstein space, and the JKO integrator based on our discrete Wasserstein distance exhibits expected qualitative behaviour.

The main drawback of our approach remains its scalability. The bottleneck of the computations is the solution of a linear system whose number of unknowns is the product of the number of discretisation points in time and the size of the mesh. This is an extremely structured linear system on a product manifold, for which specialised matrix inversion techniques may exist. In any event, with the current bottleneck our method can handle meshes with few thousand vertices but is not currently practical for larger meshes.

As one of the first structure-preserving discretisations of transport on meshes, our work also suggests several exciting avenues for future research. Many theoretical properties of our discrete Wasserstein distance remain to be explored. For instance, while we have shown that our formulation is a true Riemannian distance, one could verify the extent to which a wealth of other theoretical properties of transport are preserved. Recent work has proven convergence of our transport over meshes to the true transport in the limit of mesh refinement Lavenant (2019). From a practical perspective, a natural next step is to accelerate the optimisation procedure as much as possible; a faster solver for the convex optimisation problem would clearly benefit our method.

Part IV

Discussion and Conclusions

Wherein we look to the past to inform the future. What comes next? How do we improve on what has come before? What inspiration can we draw from the work we have done so far?

The theme of this thesis has been that structure–either inherent in a problem, or encoded in its solution–leads to faster, better, and more interpretable approaches to various problems in machine learning. We have exploited this structure through the tool of optimal transport, but the list of unanswered questions is almost inexhaustible.

We have seen that quantization computed under a Wasserstein distance provides a remarkably effective way of approximating large datasets, but the question of just how effective it is remains unanswered. The asymptotic rates of convergence known in the literature do not address the low sample regime (Weed et al., 2019; Kloeckner, 2012). For a solution to this problem, we can draw inspiration from recent results for kernel herding (Lacoste-Julien et al., 2015; Chen et al., 2010) where faster convergence rates are found for distributions in certain families.

What are the conditions a distribution has to satisfy for us to break the curse of dimensionality and the $O\left(n^{-\frac{1}{d}}\right)$ convergence rate. Theoretical results imply that no single algorithm can yield a better rate for any given distribution (Kloeckner, 2012). Can we say more about the problem of estimating the distance $W_p(\mu, \nu)$ by $W_p(\mu_n, \nu_n)$? This second problem certainly seems easier. If we can compute the distance exactly, then we can choose two points from μ and ν that are exactly that distance apart to achieve exact recovery of the distance. However, this approach requires computing the distance. Is there a simple algorithm that we can use to select weighted point sets from μ and ν for which the approximation scales better with dimension?

From an algorithms point of view, the approach we have presented scales poorly in the number of points of the approximation, and the dimension of the distribution. Can we improve upon this, or, if not, is there a fast approximation or sketch of the distance we can use to make computation easier? If we approximate by a parametric measure, is there a way to precondition the gradient descent to ensure faster convergence? These ideas echo work done by Li & Montúfar (2018) and Andoni et al. (2009), but applying these approaches to real data poses challenges.

The other viewpoint we have taken throughout the thesis is that hierarchical structure often makes transportation problems simpler. Beyond what we have investigated, there are questions that remain

unanswered. In Chapter 8 we saw how understanding documents via their topics turns a slow and difficult to interpret approach into a fast and understandable method. However, to discover topics in the documents, we have relied on latent Dirichlet allocation. Can we use optimal transport to learn the topics as part of the optimisation process?

In Chapter 9, we explored a problem where invariance to a group action lead to issues when computing expectations. The hierarchy arose naturally when taking orbits of samples under the group, and we saw how optimal transport leads to very efficient algorithms for tackling label switching. What other problems exhibit group invariance, and do their solutions also allow for a similar approach?

What connects these problems and algorithms together is optimal transport. The theory of optimal transport is flexible. Distributional data appears everywhere, from machine learning, to computer graphics, from natural language processing to Bayesian inference, and a fundamental task among these fields is comparing two distributions. That optimal transport is suited to this task is not surprising; that we can use insights from these fields to improve performance and efficiency of transport algorithms is important, however.

None of the problems we have tackled are new, and yet viewing them through the lens of optimal transport lends new perspectives. It is easy to shoehorn a Wasserstein distance wherever a Kullback-Leibler divergence used to be, but the elegance of the transport problem is shown best when it is used as more than just another distance between measures.

A

Supplementary Material: Hierarchical Optimal Transport for Document Representation

A.1 Metric properties

HOIT is a metric in the lifted topic space since W_p is a metric on distributions.

Proof. We can additionally prove that if we can exactly write $d^i = \sum_{k=1}^{|T|} \bar{d}^i_k t_k$ and if $t_i \neq t_j$ for $i \neq j$, then *HOIT* is a metric in document space.

Positivity, symmetry, and the triangle inequality follow from properties of W_2. We prove that if $HOIT(d^i, d^j) = 0$, then $d^i = d^j$. From the definition of *HOIT*,

$$HOIT(d^i, d^j) = W_2\left(\sum_{k=1}^{|T|} \bar{d}^i_k \delta_{t_k}, \sum_{l=1}^{|T|} \bar{d}^j_l \delta_{t_l}\right).$$

If $HOIT(d^i, d^j) = 0$, then if the transport plan is positive at $T_{k,l}$, it must hold that $W_p(t_k, t_l) = 0$. Since W_p is a metric on probability distributions, this implies $t_k = t_l$. As we assumed that topics are distinct, and that documents are uniquely represented as linear combinations of topics we have $d^i = d^j$. □

A.2 HOTT/WMD/RWMD relation

Following the discussion in Section 4 of the main text, we relate HOTT and RWMD to WMD empirically in terms of Mantel correlation and a Frobenius norm. The results are in Table A.1. While it is unsurprising that RWMD is more strongly correlated with WMD (HOTT is neither a lower nor an upper bound), we note that HOTT is on average a better approximation to WMD than RWMD.

Table A.1: Relation between the metrics. For each dataset, we compute distance matrices using exact WMD, RWMD, and HOTT from a few randomly-selected documents. We report results of a Mantel correlation test between WMD/HOTT and WMD/RWMD and the difference between cost matrices under a Frobenius norm.

	Mantel		l_2	
Dataset	HOTT	RWMD	HOTT	RWMD
OHSUMED	0.57	0.87	55	104
20NEWS	0.62	0.90	90	99
AMAZON	0.49	0.84	70	65
REUTERS	0.72	0.91	130	151
BBCSPORT	0.76	0.92	28	90
CLASSIC	0.43	0.89	157	69
Avg	0.60	0.89	88	96

A.3 Additional experimental results

In the main text, we used W_1 distance and did not do any vocabulary reduction, following the experimental setup of Kusner et al. (2015a). W_2 distance has intuitive geometric properties and is equipped with a variety of theoretical characterizations (Villani, 2008); one intuition for the difference between W_1 and W_2 comes from an analogy to the differences between l_1 and l_2 regularization. On the other hand, stemming is a common vocabulary reduction technique to improve quality of topic models. Stemming attempts to merge terms which differ only in their ending, i.e. "cat" and "cats". As stemming sometimes produces words not available in the *GloVe* embeddings (Pennington et al., 2014), to embed a stemmed word we take the average embeddings of the words mapped to it. We used *SnowballStemmer* available

from the *nltk* Python package.

Figures A-1 and A-2 demonstrate results with W_1 and stemming; Figures A-3 and A-4 with W_2 and no stemming; Figures A-5 and A-6 with W_2 and stemming. In all settings HOTT and HOFTT are the best on average. Interestingly, using W_2 degrades performance of RWMD and WMD-T20, while our methods perform equally well with W_1 and W_2. Stemming tends to improve performance of nBOW, therefore aggregated results appear worse. Stemming also negatively effects RWMD and WMD-T20, while appears to have no effect on HOTT and HOFTT. For example, in the case of W_2 with stemming (Figures A-5 and A-6), RWMD is no longer superior to baselines LDA (Blei et al., 2003) and Cosine, while our methods maintain good performance. We conclude that our methods are more robust to the choice of text processing techniques and specifics of the Wasserstein distance.

In Figure A-7 we present additional t-SNE (van der Maaten & Hinton, 2008) visualization results.

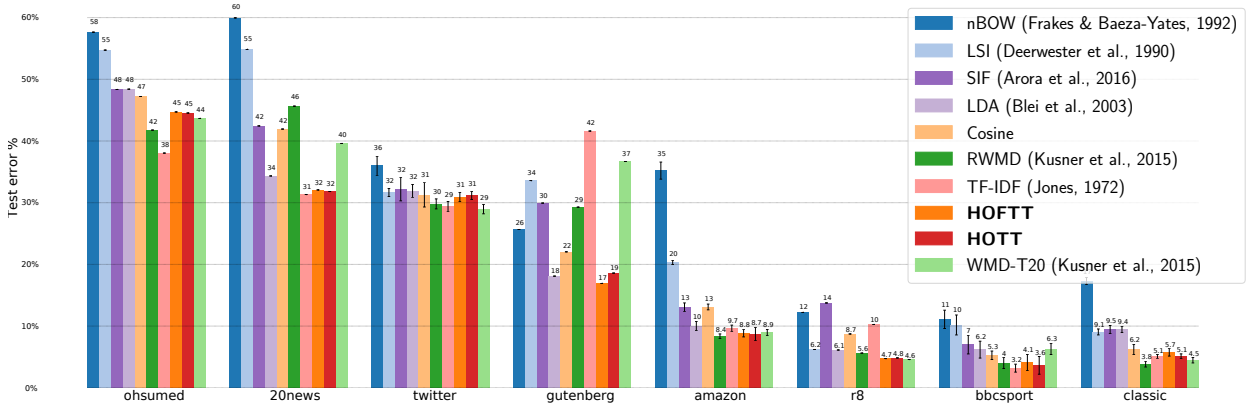


Figure A-1: W_1 and stemming: k-NN classification performance across datasets

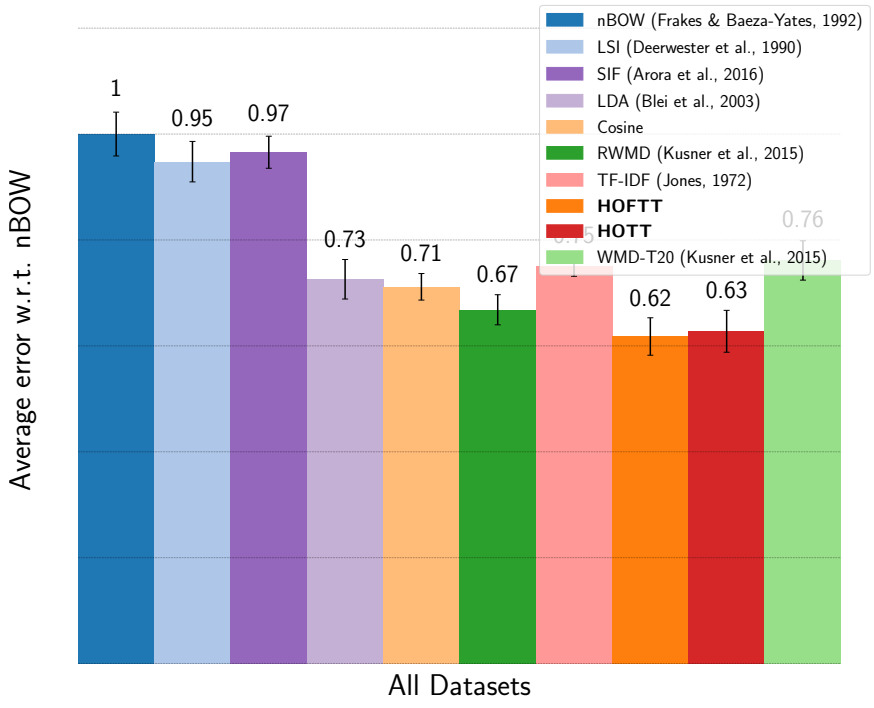


Figure A-2: W_1 and stemming: k-NN classification performance normalized by nBOW

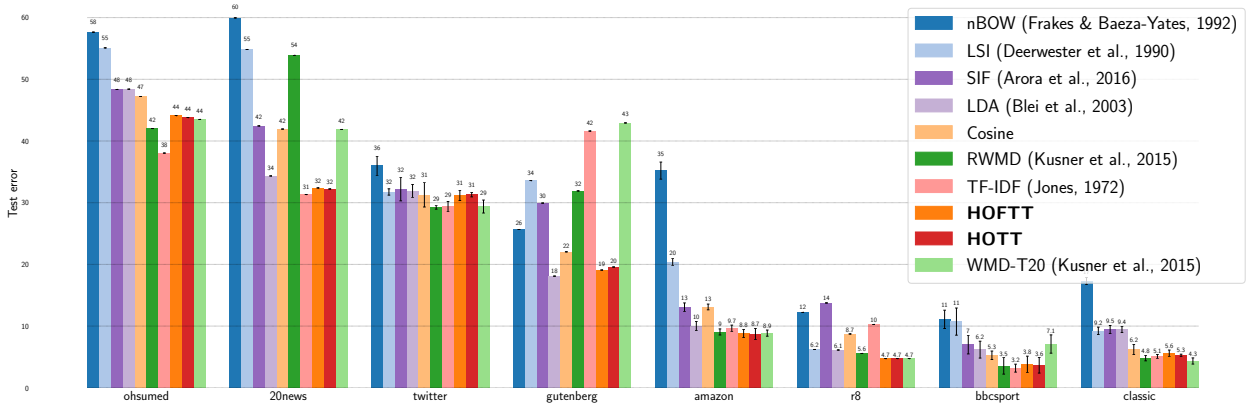


Figure A-3: W_2 without stemming: k-NN classification performance across datasets

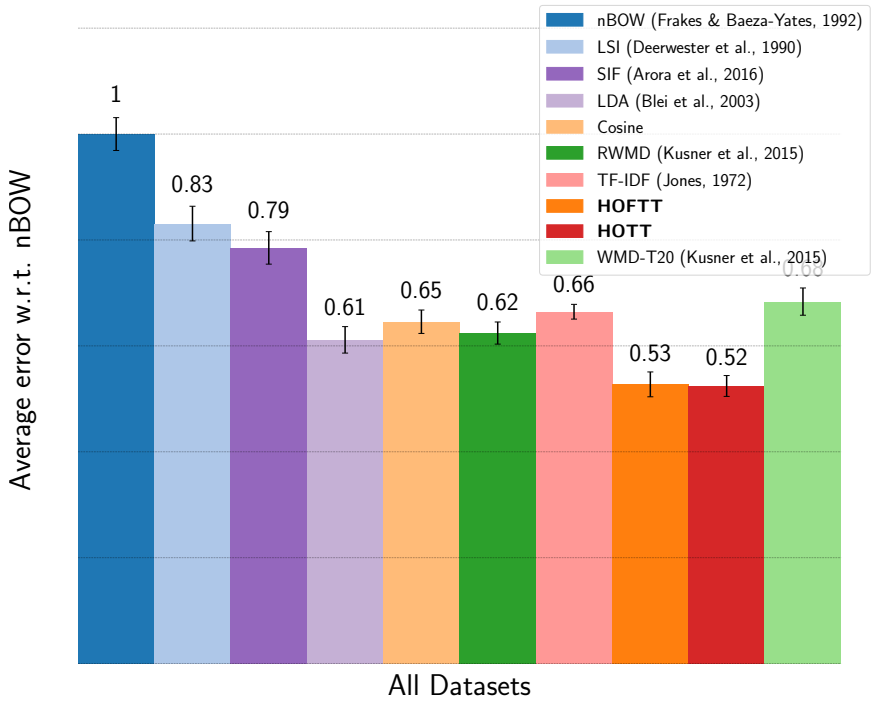


Figure A-4: W_2 without stemming: aggregated k-NN classification performance normalized by nBOW

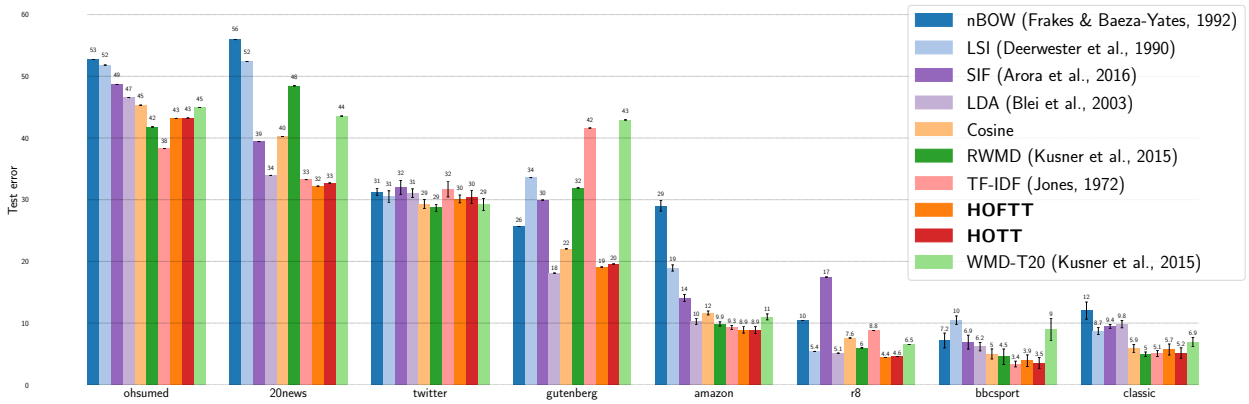


Figure A-5: W_2 and stemming: k-NN classification performance across datasets

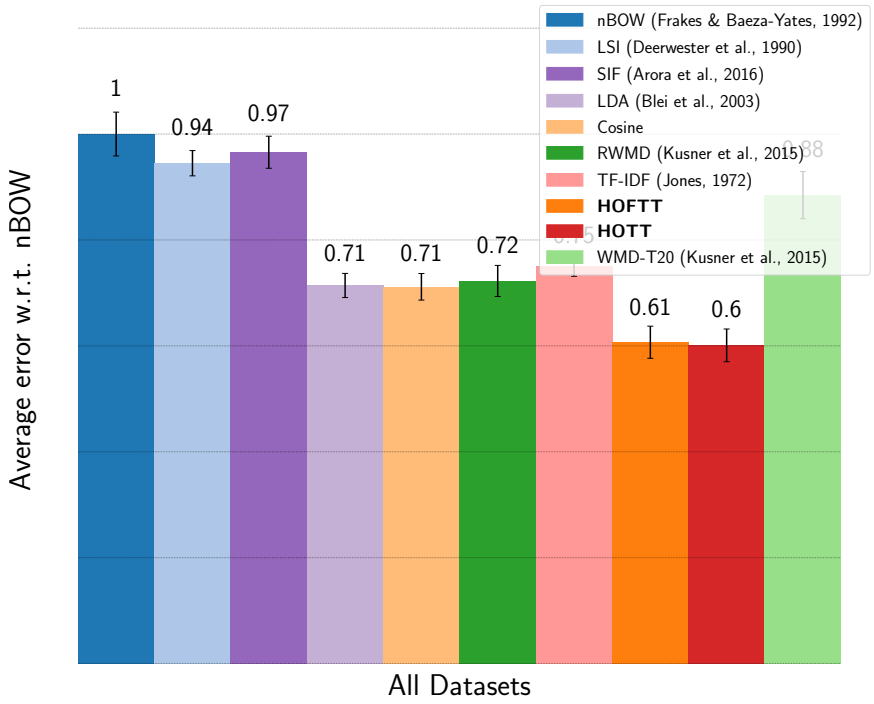


Figure A-6: W_2 and stemming: aggregated k-NN classification performance normalized by nBOW

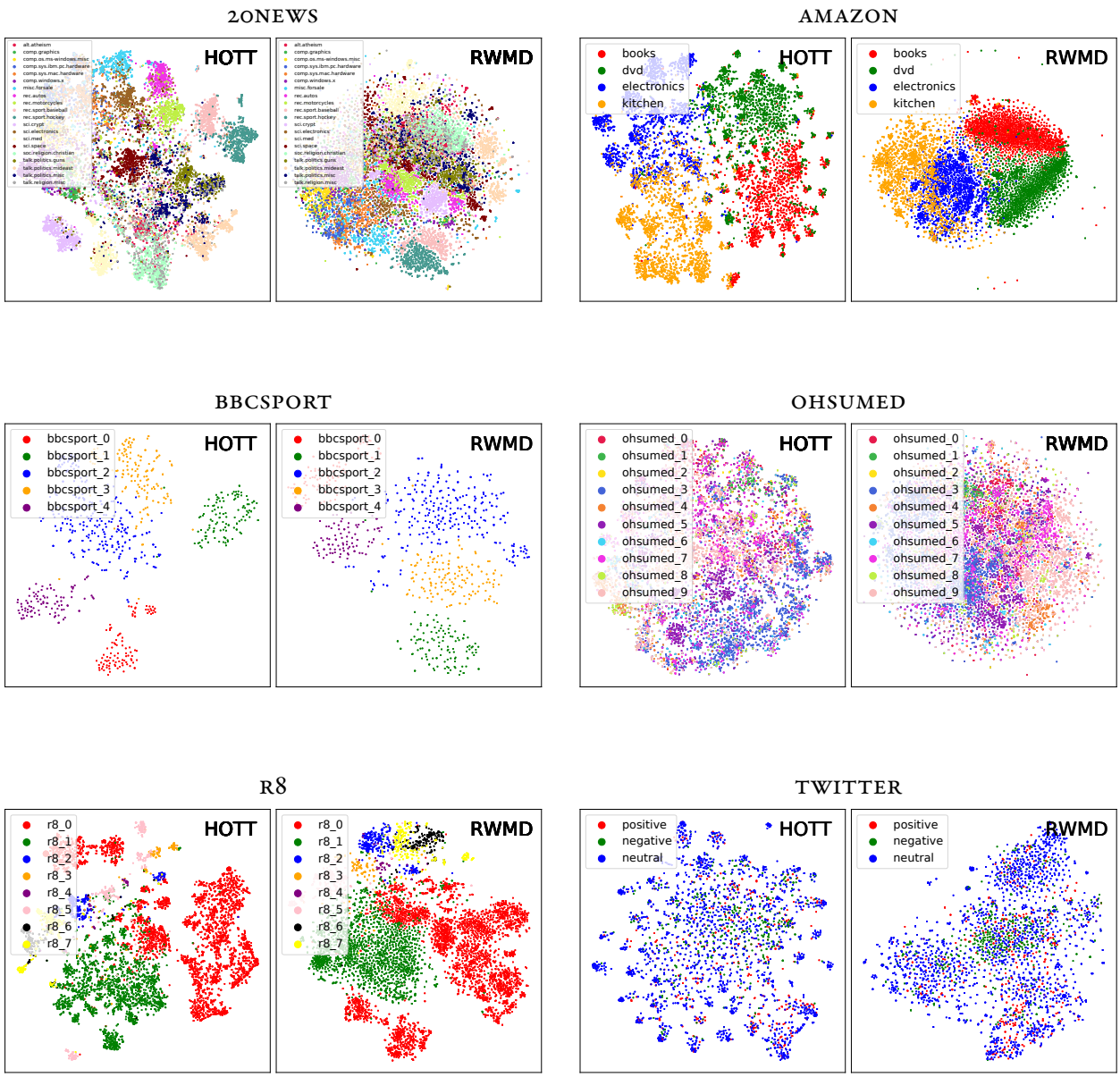


Figure A-7: These are the additional t-SNE results on all other datasets, except GUTENBERG, which is excluded due to its high number of classes (142). These images show that clusters based on our metric better align with the labels, corresponding to a better metric than RWMD. Both methods perform poorly on TWITTER, a difficult dataset for topic modelling.

B

Supplementary Material: Alleviating Label Switching via Optimal Trasport

B.1 Optimal Transport

B.1.1 Proof of Theorem 1

We first recall the definition of sequential compactness and Prokhorov's theorem, which relates it to tightness of measures:

Definition B.1 (Sequential compactness). *A space X is called* sequentially compact *if every sequence of points x_n has a convergent subsequence converging to a point in X.*

Theorem B.1 (Prokhorov's theorem). *A collection $C \subset P_2(X)$ of probability measures is tight if and only if C is sequentially compact in $P_2(X)$, equipped with the topology of weak convergence.*

Now, note that the barycenter objective is bounded below by 0 and is finite, so we may pick out a minimizing sequence μ_n of $B(\mu)$. Prokhorov's theorem allows us to extract a subsequence μ_{n_k} that converges to a minimizer $\mu \in P_2(X)$ and the theorem is proved.

B.1.2 Tightness from Uniform Second Moment Bound

We argue here for a sufficient condition for tightness claimed in the text:

Lemma. *If a collection of measures $\mathscr{C} \subset P_2(X)$ has a uniform second moment bound (about any reference point $x_0 \in X$), i.e.,*

$$\int_X d^2(x_0, x)\, d\nu(x) < M$$

for some $M > 0$ and all $\nu \in \mathscr{C}$, then $\mathscr{C}$ is tight.

Proof. For any $\nu \in \mathscr{C}$ we have the following inequalities:

$$\nu\{x \mid d(x, x_0) > R\} = \int_{d(x,x_0)>R} \mathrm{d}\nu \leq \frac{1}{R^2} \int_{d(x,x_0)>R} d(x, x_0)^2 \mathrm{d}\nu(x) \leq \frac{M}{R^2}.$$

The last term converges to 0 as $R \to \infty$, and the set $\{x \mid d(x, x_0) \leq R\}$ is compact, so tightness follows. □

B.1.3 Mean-only Mixture Models

Here we note some facts about mixture models, where the K components are evenly weighted and identical with only one parameter each in $\mathbb{R}^d$. An example would be the simple case of a Gaussian mixture model with fixed equal covariance across each component, and a remaining unspecified mean parameter $p_i \in \mathbb{R}^d$.

In this instance, we are taking the quotient of $(\mathbb{R}^d)^K$ by an action of S_K which simply permutes the K factors of the product. Let us begin by investigating the case where $d = 1$. In this instance, we note that the sum of the scalar means $\sum_i p_i$ remains fixed under the action of the group. In fact, the action of the group splits into a trivial action on the 1-dimensional fixed subspace $F_K := \{(p_1, \dots, p_k) \mid p_i \text{ all equal}\}$, and an action on $F_K^\perp$ which permutes the vertices of an embedded regular $(K-1)$-simplex about the origin. Namely, one may take the simplex in $F_K^\perp$ with vertices that consist of the point $(K-1, -1, -1, \dots, -1)$ and its orbit. Figure B-1 illustrates the concrete example of three means: $\mathbb{R}^3/S_3$. It shows $F_3^\perp$, an embedded 2-simplex, and the action of S_3 on this space and simplex. Section B.2.2 proves that the quotient space $\mathbb{R}^K/S_K$ is a convex, easily described set, and discusses the consequences for label switching.

The splitting mentioned above is the decomposition into irreducible components. For $d > 1$, the action of S_K is diagonal and acts on the d components of the means p_i in parallel. It preserves the scalar

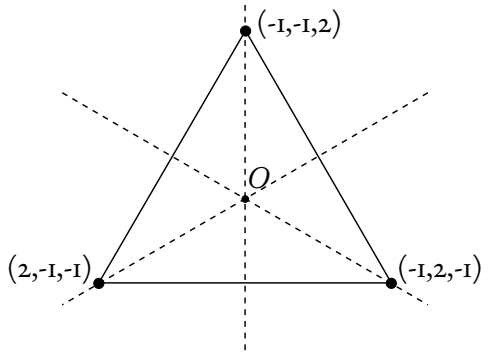


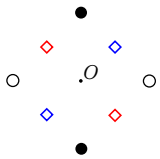
Figure B-1: A schematic illustrating the nontrivial part of the action of S_3 on $\mathbb{R}^3$. It acts on $F_3^\perp$ and the embedded 2-simplex shown via reflection over the dashed lines. One can see that reflection over these lines correspond to swapping of pairs of means, generating S_3 as a group.

sum of these components over each dimension and we obtain the following splitting for the general case:

$$(\mathbb{R}^d)^K = \bigoplus_{j=1}^{d} \left(F_K \oplus F_K^\perp\right) \cong \mathbb{R}^d \oplus \left(\mathbb{R}^{K-1}\right)^d. \tag{B.1}$$

The action on the first $\mathbb{R}^d$ component is trivial, while the second component has the diagonal action permuting the vertices of an embedded regular $(K-1)$-simplex for each $\mathbb{R}^{K-1}$. The simple example of two means in $\mathbb{R}^2$ ($d = K = 2$) is discussed and illustrated in the next section (B.1.4), and also serves to provide a counterexample to barycenter uniqueness. For $d > 1$, the quotient $(\mathbb{R}^d)^K/S_K$ lacks the simple convexity of the $d = 1$ case, as described in Section B.2.3.

B.1.4 Counterexample to uniqueness



Take $d = K = 2$ from the scenario above, which might correspond to our mixture model consisting of two Gaussians in $\mathbb{R}^2$ with equal weights and fixed variance. Only the means $(x, y; z, w) \in (\mathbb{R}^2)^2$ are taken as parameters, and the action of S_2 swaps the means: $(x, y; z, w) \mapsto (z, w; x, y)$. This action splits into a trivial action on $\mathrm{Span}\{(1, 0; 1, 0), (0, 1; 0, 1)\}$ and an antipodal action ($v \mapsto -v$) on $\mathrm{Span}\{(1, 0; -1, 0), (0, 1; 0, -1)\}$, where these are the first and second components in Eq. (B.1). Recall that the 1-simplex is just an interval and the action of S_2 merely flips the endpoints, so the antipodal action arises as the diagonal action of this flip.

The inset figure illustrates a simple schematic counterexample in the second span. The two distributions to be averaged are evenly supported on the black and white dots, invariant under reflection through the center origin O. Two candidate barycenters are those evenly supported on the red and blue diamonds, and in fact, any convex combination of these two are a barycenter. This corresponds to averaging a mixture with means at $(1, 0)$ and $(-1, 0)$ and another with means at $(0, 1)$ and $(0, -1)$. Two sensible averages are a pair of means at $(0.5, 0.5)$ and $(-0.5, -0.5)$, or a pair of means at $(0.5, -0.5)$ and $(-0.5, 0.5)$.

Note that the previous example requires a high degree of symmetry for the input distributions, and uniqueness is recovered if either of the distributions are absolutely continuous. Section B.2.3 further characterizes the geometry of the quotient space for $d = K = 2$, and how it leads to non-unique barycenters.

B.2 Optimal Transport with Group Invariances

B.2.1 Proof of Lemma 4

Consider an arbitrary point $z_0 \in X/G$, and we will show that a minimizer of $z \to \mathbb{E}_{\delta_x \sim \Omega_*}\left[d(x,z)^2\right]$ lies in a closed ball about z_0. As the function is continuous and this is a compact set, existence of a minimizer results.

By the triangle inequality, we have $d(x,z) \geq d(x,z_0) - d(z,z_0)$. Thus, we have:

$$\begin{aligned}
\mathbb{E}_{\delta_x \sim \Omega_*}\left[d(x,z)^2\right] &= \int_{X/G} d(x,z)^2 \, \mathrm{d}\Omega_*(\delta_x) \\
&\geq \int_{X/G} \left(d(x,z_0) - d(z,z_0)\right)^2 \, \mathrm{d}\Omega_*(\delta_x) \\
&= \left(\int_{X/G} d(x,z_0)^2 \, \mathrm{d}\Omega_*(\delta_x)\right) + d(z,z_0)^2 - 2d(z,z_0) \int_{X/G} d(x,z_0) \, \mathrm{d}\Omega_*(\delta_x).
\end{aligned}$$

The last two terms are quadratic in $d(z,z_0)$. Given an arbitrary positive constant $M > 0$, some simple algebra shows that:

$$d(z,z_0) > \frac{c + \sqrt{c^2 + 4M}}{2} \implies d(z,z_0)^2 - c\,d(z,z_0) > M$$

where $c = 2\int_{X/G} d(x, z_0)\, \mathrm{d}\Omega_*(\delta_x)$. The finiteness of this integral follows from the fact that Ω_* has finite second moment, implying finite first moment. Thus, if we set M to a realized value of $\mathbb{E}_{\delta_x \sim \Omega_*}\left[d(x, z)^2\right]$, we see that a minimizer lies in the ball of radius $\frac{c+\sqrt{c^2+4M}}{2}$ about z_0. Taking z outside this ball implies:

$$\begin{aligned}\mathbb{E}_{\delta_x \sim \Omega_*}\left[d(x, z)^2\right] &\geq \left(\int_{X/G} d(x, z_0)^2\, \mathrm{d}\Omega_*(\delta_x)\right) + d(z, z_0)^2 - 2d(z, z_0)\int_{X/G} d(x, z_0)\, \mathrm{d}\Omega_*(\delta_x). \\ &\geq d(z, z_0)^2 - 2d(z, z_0)\int_{X/G} d(x, z_0)\, \mathrm{d}\Omega_*(\delta_x) > M.\end{aligned}$$

B.2.2 Proof of Theorem 3

We recall the minimization problem in (5) of the paper for a sample $\mathrm{q} = (q_1, \ldots, q_K)$ and a current barycenter estimate $\mathrm{p} = (p_1, \ldots, p_K)$ (with a squared distance objective for simplicity of expression):

$$\min_{\sigma \in S_K} d^2_{\mathbb{R}^K}((p_1, \ldots, p_K), (q_{\sigma(1)}, \ldots, q_{\sigma(K)})) = \min_{\sigma \in S_K} \sum_{i=1}^{K} \|p_i - q_{\sigma(i)}\|^2. \tag{B.2}$$

Here, we invoke the monotonicity of transport in 1D (see e.g. Santambrogio (2015), Chapter 2) to see that we should simply order q in the same way that p is. That is to say: assuming $p_1 < p_2 < \ldots < p_K$ (WLOG), then the optimal σ is such that $q_{\sigma(1)} < q_{\sigma(2)} < \ldots < q_{\sigma(K)}$.

The above argument also shows that we have a very concrete realization:

$$\mathrm{UConf}_K(\mathbb{R}) \cong \{(u_1, \ldots, u_K) \in \mathrm{Conf}_K(\mathbb{R}) \mid u_1 < \ldots < u_K\} \subset \mathbb{R}^K.$$

As this is an open convex set, we have uniqueness of the single-point barycenter of Theorem 2 from the paper under mild conditions on the posterior. Namely, consider that $\Omega_* \in P_2(P_2(X))$ descends to a measure $\Omega_\downarrow \in P_2(X)$, and we will need to assume that $\Omega_\downarrow$ is absolutely continuous (as you might expect). With this, Kim & Pass (2017) give us the desired result.

Furthermore, we have guaranteed convergence of stochastic gradient descent (our algoritheorem) in this setting, as $\mathbb{E}[W_2^2(\cdot, \nu)]$ is 1-strongly convex and the domain is convex. The next section shows us that we may not leverage such simple structure for $d > 1$.

B.2.3 Positive Curvature of Mean-Only Models

Section B.1.4 shows us that in the case of $d = K = 2$:

$$\mathrm{UConf}_2\left(\mathbb{R}^2\right) \cong \mathbb{R}^2 \times C^* \qquad \text{where} \qquad C^* = (\mathbb{R}^2 \backslash \{(0,0)\})/\{v \sim -v\}.$$

C^* is isometric to an infinite metric cone (2-dimensional) with cone angle π and cone point excised. It is this positive curvature which gives rise to the counterexample presented.

More generally, B.1.3 showed us that in these mean-only models there is a diagonal action on a subspace isometric to $(\mathbb{R}^{K-1})^d$. In all of these cases, under the action of S_K, the solid angle measure of a sphere about the origin will be divided by $K!$ when quotiented, producing a point of positive curvature, and leading to highly symmetric counterexamples with non-uniqueness of barycenters.

Bibliography

Agarwal, P. K., Har-Peled, S., and Varadarajan, K. R. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

Agueh, M. and Carlier, G. Barycenters in the Wasserstein Space. *SIAM J. Math. Anal.*, 43(2):904–924, January 2011. ISSN 0036-1410. doi: 10.1137/100805741.

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 1961–1971, 2017.

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Alvarez-Melis, D. and Jaakkola, T. S. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 1881–1890, 2018.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

Anderes, E., Borgwardt, S., and Miller, J. Discrete Wasserstein barycenters: Optimal transport for discrete data. *Math Meth Oper Res*, 84(2):389–409, October 2016. ISSN 1432-2994, 1432-5217. doi: 10.1007/s00186-016-0549-x.

Andoni, A., Ba, K. D., Indyk, P., and Woodruff, D. P. Efficient sketches for earth-mover distance, with applications. In *50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009, October 25-27, 2009, Atlanta, Georgia, USA*, pp. 324–330, 2009. doi: 10.1109/FOCS.2009.25.

ApS, M. *The MOSEK optimization toolbox for MATLAB manual. Version 8.0.0.53.*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 214–223, 2017.

Arora, S., Liang, Y., and Ma, T. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017.

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Atasu, K. and Mittelholzer, T. Linear-complexity data-parallel earth mover's distance approximations. In *International Conference on Machine Learning*, pp. 364–373, 2019.

Aurenhammer, F. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

Aurenhammer, F., Hoffmann, F., and Aronov, B. Minkowski-type theorems and least-squares partitioning. In *Proceedings of the Eighth Annual Symposium on Computational Geometry*, pp. 350–357. ACM, 1992.

Aurenhammer, F., Hoffmann, F., and Aronov, B. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

Azencot, O., Vantzos, O., and Ben-Chen, M. Advection-based function matching on surfaces. In *Computer Graphics Forum*, volume 35, pp. 55–64. Wiley Online Library, 2016.

Bachem, O., Lucic, M., and Krause, A. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

Bachem, O., Lucic, M., and Krause, A. Scalable k -means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pp. 1119–1127, 2018a. doi: 10.1145/3219819.3219973.

Bachem, O., Lucic, M., and Lattanzi, S. One-shot coresets: The case of k-clustering. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pp. 784–792, 2018b.

Bandeira, A. S., Charikar, M., Singer, A., and Zhu, A. Multireference alignment using semidefinite programming. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pp. 459–470, 2014.

Bandeira, A. S., Blum-Smith, B., Kileel, J., Perry, A., Weed, J., and Wein, A. S. Estimation under group actions: recovering orbits from invariants. *arXiv preprint arXiv:1712.10163*, 2017.

Baykal, C., Liebenwein, L., and Schwarting, W. Training support vector machines using coresets. *CoRR*, abs/1708.03835, 2017.

Baykal, C., Liebenwein, L., Gilitschenski, I., Feldman, D., and Rus, D. Data-dependent coresets for compressing neural networks with applications to generalization bounds. *CoRR*, abs/1804.05345, 2018.

Benamou, J., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman Projections for Regularized Transportation Problems. *SIAM J. Sci. Comput.*, 37(2):A1111–A1138, January 2015a. ISSN 1064-8275. doi: 10.1137/141000439.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Benamou, J.-D. and Carlier, G. Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1): 1–26, 2015.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015b.

Benamou, J.-D., Carlier, G., and Santambrogio, F. Variational mean field games. In *Active Particles, Volume 1*, pp. 141–171. Springer, 2017.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

Bonneel, N., Van De Panne, M., Paris, S., and Heidrich, W. Displacement interpolation using Lagrangian mass transport. In *ACM Transactions on Graphics (TOG)*, volume 30, pp. 158. ACM, 2011.

Bottou, L. and Bengio, Y. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems*, pp. 585–592, 1995.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Brancolini, A., Buttazzo, G., Santambrogio, F., and Stepanov, E. Long-term planning versus short-term planning in the asymptotical location problem. *ESAIM: Control, Optimisation and Calculus of Variations*, 15(3):509–524, 2009.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Brenner, S. and Scott, R. *The Mathematical Theory of Finite Element Methods*, volume 15. Springer Science & Business Media, 2007.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, pp. 1727–1735, 2013.

Bryant, M. and Sudderth, E. B. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pp. 2699–2707, 2012.

Cachopo, A. M. d. J. C. et al. Improving methods for single-label text categorization. *Instituto Superior Técnico, Portugal*, 2007.

Campbell, T. and Broderick, T. Bayesian coreset construction via greedy iterative geodesic ascent. *CoRR*, abs/1802.01737, 2018.

Campbell, T. and Broderick, T. Automated scalable Bayesian inference via Hilbert coresets. *Journal of Machine Learning Research*, 20(15):1–38, 2019.

Cañas, G. D. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pp. 2501–2509, 2012.

Carlier, G., Oberman, A., and Oudet, E. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: M2AN*, 49(6):1621–1642, November 2015. ISSN 0764-583X, 1290-3841. doi: 10.1051/m2an/2015033.

Carlier, G., Chernozhukov, V., Galichon, A., et al. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Carrière, M., Cuturi, M., and Oudot, S. Sliced Wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 664–673, 2017.

Casella, G. and George, E. I. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Celeux, G., Hurn, M., and Robert, C. P. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. Stein points. In *International Conference on Machine Learning*, pp. 844–853, 2018.

Chen, Y., Welling, M., and Smola, A. J. Super-samples from kernel herding. In *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, pp. 109–116, 2010.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 3040–3050, 2018.

Chow, S.-N., Dieci, L., Li, W., and Zhou, H. Entropy dissipation semi-discretization schemes for fokker–planck equations. *Journal of Dynamics and Differential Equations*, 31(2):765–792, 2019.

Claici, S., Chien, E., and Solomon, J. Stochastic Wasserstein barycenters. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, abs/1802.05757, 2018.

Claici, S., Genevay, A., and Solomon, J. Wasserstein measure coresets, 2020.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal Transport for Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1–1, 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2615921.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Cuturi, M. and Doucet, A. Fast computation of Wasserstein barycenters. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 685–693, 2014.

Cuturi, M. and Peyré, G. A Smoothed Dual Approach for Variational Wasserstein Problems. *SIAM J. Imaging Sci.*, 9(1):320–343, January 2016. doi: 10.1137/15M1032600.

de Goes, F., Cohen-Steiner, D., Alliez, P., and Desbrun, M. An optimal transport approach to robust reconstruction and simplification of 2d shapes. In *Computer Graphics Forum*, volume 30, pp. 1593–1602. Wiley Online Library, 2011.

De Goes, F., Breeden, K., Ostromoukhov, V., and Desbrun, M. Blue noise through optimal transport. *ACM Transactions on Graphics (TOG)*, 31(6):171, 2012.

de Goes, F., Memari, P., Mullen, P., and Desbrun, M. Weighted triangulations for geometry processing. *ACM Trans. Graph.*, 33(3):28:1–28:13, June 2014.

de Goes, F., Desbrun, M., and Tong, Y. Vector field processing on triangle meshes. In *SIGGRAPH Asia 2015 Courses*, pp. 17. ACM, 2015a.

de Goes, F., Wallez, C., Huang, J., Pavlov, D., and Desbrun, M. Power particles: an incompressible fluid solver based on power diagrams. *ACM Trans. Graph.*, 34(4):50–1, 2015b.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391, Sep 01 1990.

Diebolt, J. and Robert, C. P. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375, 1994.

Digne, J., Cohen-Steiner, D., Alliez, P., De Goes, F., and Desbrun, M. Feature-preserving surface reconstruction and simplification from defect-laden point sets. *Journal of Mathematical Imaging and Vision*, 48(2):369–382, 2014.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. *Physics Letters B*, 195 (2):216–222, 1987.

Edmonds, J. and Karp, R. M. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)*, 19(2):248–264, 1972.

Erbar, M., Rumpf, M., Schmitzer, B., and Simon, S. Computation of optimal transport on discrete metric measure spaces. *Numerische Mathematik*, 144(1):157–200, 2020.

Fadell, E. R. and Husseini, S. Y. *Geometry and topology of configuration spaces*. Springer Science & Business Media, 2012.

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pp. 569–578, 2011. doi: 10.1145/1993636.1993712.

Feldman, D., Schmidt, M., and Sohler, C. Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013*, pp. 1434–1453. SIAM, 2013.

Frakes, W. B. and Baeza-Yates, R. *Information retrieval: Data structures & algorithms*, volume 331. prentice Hall Englewood Cliffs, NJ, 1992.

Gallouët, T. O. and Mérigot, Q. A lagrangian scheme à la brenier for the incompressible euler equations. *Foundations of Computational Mathematics*, pp. 1–31, 2017.

Gangbo, W. and McCann, R. J. The geometry of optimal transportation. *Acta Mathematica*, 177(2): 113–161, 1996.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3440–3448. Curran Associates, Inc., 2016.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pp. 1574–1583, 2019.

Gigli, N. and Maas, J. Gromov–Hausdorff convergence of discrete transportation metrics. *SIAM Journal on Mathematical Analysis*, 45(2):879–899, 2013.

Givens, C. R., Shortt, R. M., et al. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

Gladbach, P., Kopfer, E., and Maas, J. Scaling limits of discrete optimal transport. *arXiv:1809.01092*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Grant, M. and Boyd, S. Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H. (eds.), *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008.

Grant, M. and Boyd, S. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

Griffiths, T. L. and Steyvers, M. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

Guennebaud, G., Jacob, B., et al. Eigen v3. http://eigen.tuxfamily.org, 2010.

Guittet, K. On the time-continuous mass transport problem and its approximation by augmented Lagrangian techniques. *SIAM Journal on Numerical Analysis*, 41(1):382–399, 2003.

Gurobi Optimization, L. Gurobi optimizer reference manual, 2018.

Har-Peled, S. and Mazumdar, S. On coresets for k-means and k-median clustering. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing, STOC 2004*, pp. 291–300. ACM, 2004.

Heeren, B., Rumpf, M., Wardetzky, M., and Wirth, B. Time-discrete geodesics in the space of shells. In *Computer Graphics Forum*, volume 31, pp. 1755–1764. Wiley Online Library, 2012.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013.

Huang, G., Guo, C., Kusner, M. J., Sun, Y., Sha, F., and Weinberger, K. Q. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pp. 4862–4870, 2016.

Hug, R., Papadakis, N., and Maitre, E. On the convergence of augmented Lagrangian method for optimal transport between nonnegative densities. *Journal of Mathematical Analysis and Applications*, pp. 123811, 2020.

Huggins, J. H., Campbell, T., and Broderick, T. Coresets for scalable Bayesian logistic regression. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4080–4088, 2016.

Jasra, A., Holmes, C. C., and Stephens, D. A. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pp. 50–67, 2005.

Jost, J. *Riemannian geometry and geometric analysis*, volume 42005. Springer, 2008.

Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pp. 199–201, 1942.

Keller, F., Muller, E., and Bohm, K. Hics: High contrast subspaces for density-based outlier ranking. In *2012 IEEE 28th international conference on data engineering*, pp. 1037–1048. IEEE, 2012.

Kim, Y.-H. and Pass, B. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, February 2017. ISSN 0001-8708. doi: 10.1016/j.aim.2016.11.026.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Kitagawa, J., Mérigot, Q., and Thibert, B. Convergence of a newton algorithm for semi-discrete optimal transport. *arXiv preprint arXiv:1603.05579*, 2016.

Kitagawa, J., Mérigot, Q., and Thibert, B. Convergence of a Newton algorithm for semi-discrete optimal transport. *Journal of the European Math Society (JEMS)*, March 2018.

Klein, M. A primal method for minimal cost flows with applications to the assignment and transportation problems. *Management Science*, 14(3):205–220, 1967.

Kloeckner, B. Approximation by finitely supported measures. *ESAIM Control Optim. Calc. Var.*, 18(2): 343–359, 2012. ISSN 1292-8119.

Kobayashi, S. Isometries of riemannian manifolds. In *Transformation Groups in Differential Geometry*, pp. 39–76. Springer, 1995.

Kuhn, H. W. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *International Conference on Machine Learning*, pp. 957–966, 2015a.

Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 957–966, 2015b.

Lacoste-Julien, S., Lindsten, F., and Bach, F. R. Sequential kernel herding: Frank–Wolfe optimization for particle filtering. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015.

Langberg, M. and Schulman, L. J. Universal epsilon-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pp. 598–607, 2010. doi: 10.1137/1.9781611973075.50.

Lavenant, H. Unconditional convergence for discretizations of dynamical optimal transport. *arXiv preprint arXiv:1909.08790*, 2019.

Lavenant, H., Claici, S., Chien, E., and Solomon, J. Dynamical optimal transport on discrete surfaces. *ACM Trans. Graph.*, 37(6):250:1–250:16, 2018. doi: 10.1145/3272127.3275064.

Lévy, B. A Numerical Algorithm for L2 Semi-Discrete Optimal Transport in 3D. *ESAIM Math. Model. Numer. Anal.*, 49(6):1693–1715, November 2015. ISSN 0764-583X, 1290-3841. doi: 10.1051/m2an/2015055.

Lévy, B. and Schwindt, E. L. Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148, 2018.

Li, W. and Montúfar, G. Natural gradient via optimal transport. *Information Geometry*, 1(2):181–214, 2018.

Li, W., Ryu, E. K., Osher, S., Yin, W., and Gangbo, W. A parallel method for earth mover's distance. *Journal of Scientific Computing*, 75(1):182–197, 2018.

Li, X., Hu, Z., and Wu, F. A note on the convergence of the mean shift. *Pattern Recognition*, 40(6):1756–1762, 2007.

Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, Mar 1982.

Lott, J. and Villani, C. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169(3):903–991, 2009. ISSN 0003486X.

Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

Lyon, R. J., Stappers, B., Cooper, S., Brooke, J., and Knowles, J. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1):1104–1123, 2016.

Maas, J. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

Marin, J.-M., Mengersen, K., and Robert, C. P. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, 25:459–507, 2005.

McCann, R. J. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. Finite mixture models. *Annual Review of Statistics and its Application*, 6:355–378, 2019.

Mérigot, Q. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pp. 1583–1592. Wiley Online Library, 2011.

Mérigot, Q. and Mirebeau, J.-M. Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport. *SIAM Journal on Numerical Analysis*, 54(6):3465–3492, 2016.

Mérigot, Q., Meyron, J., and Thibert, B. An algorithm for optimal transport between a simplex soup and a point cloud. *SIAM Journal on Imaging Sciences*, 11(2):1363–1389, 2018.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Monteiller, P., Claici, S., Chien, E., Mirzazadeh, F., Solomon, J. M., and Yurochkin, M. Alleviating label switching with optimal transport. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 13612–13622, 2019.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Munteanu, A. and Schwiegelshohn, C. Coresets—Methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intelligenz (KI)*, 32(1):37–53, 2018.

Muzellec, B. and Cuturi, M. Generalizing point embeddings using the wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems*, pp. 10237–10248, 2018.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Nesterov, Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pp. 372–376, 1983.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108. Association for Computational Linguistics, 2010.

Orlin, J. B. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 2001.

Otto, F. and Villani, C. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Panozzo, D., Baran, I., Diamanti, O., and Sorkine-Hornung, O. Weighted averages on surfaces. *ACM Transactions on Graphics (TOG)*, 32(4):60, 2013.

Papadakis, N., Peyré, G., and Oudet, E. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

Papastamoulis, P. label.switching: An r package for dealing with the label switching problem in mcmc outputs. *Journal of Statistical Software*, 2016. doi: 10.18637/jss.v069.c01.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Peyré, G. and Cuturi, M. *Computational Optimal Transport*. Submitted, 2018.

Peyre, R. Comparison between W2 distance and H1 norm, and localization of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(4):1489–1501, 2018.

Phillips, J. M. and Tai, W. M. Near-optimal coresets of kernel density estimates. In *34th International Symposium on Computational Geometry, SoCG 2018, June 11-14, 2018, Budapest, Hungary*, pp. 66:1–66:13, 2018. doi: 10.4230/LIPIcs.SoCG.2018.66.

Pinkall, U. and Polthier, K. Computing discrete minimal surfaces and their conjugates. *Experimental Mathematics*, 2(1):15–36, 1993.

Pollard, D. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2): 199–205, 1982.

Polthier, K. and Schmies, M. *Straightest geodesics on polyhedral surfaces*. ACM, 2006.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein Barycenter and Its Application to Texture Mixing. In *Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, Berlin, Heidelberg, May 2011. doi: 10.1007/978-3-642-24785-9_37.

Reddi, S. J., Póczos, B., and Smola, A. Communication efficient coresets for empirical loss minimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pp. 752–761, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Rigollet, P. and Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.

Sanders, N. J. Sanders-twitter sentiment corpus. *Sanders Analytics LLC*, 2011.

Santambrogio, F. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-20827-5 978-3-319-20828-2. doi: 10.1007/978-3-319-20828-2.

Schmitz, M. A., Heitz, M., Bonneel, N., Ngole, F., Coeurjolly, D., Cuturi, M., Peyré, G., and Starck, J.-L. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Schmitzer, B. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, 2016.

Solomon, J. *Optimal Transport on Discrete Domains*. AMS Short Course on Discrete Differential Geometry, 2018.

Solomon, J., Rustamov, R., Guibas, L., and Butscher, A. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 33(4):67, 2014.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains. *ACM Trans Graph*, 34(4):66:1–66:11, July 2015. ISSN 0730-0301. doi: 10.1145/2766963.

Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

Srivastava, S., Cevher, V., Dinh, Q., and Dunson, D. WASP: Scalable Bayes via barycenters of subset posteriors. In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 912–920, San Diego, California, USA, 09–12 May 2015a. PMLR.

Srivastava, S., Cevher, V., Tran-Dinh, Q., and Dunson, D. B. WASP: scalable bayes via barycenters of subset posteriors. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015b.

Staib, M., Claici, S., Solomon, J. M., and Jegelka, S. Parallel streaming Wasserstein barycenters. In *Advances in Neural Information Processing Systems, NIPS 2017*, pp. 2644–2655, 2017.

Stephens, M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

The CGAL Project. *CGAL User and Reference Manual*. CGAL Editorial Board, 5.0.2 edition, 2020.

Trillos, N. G. Gromov-Hausdorff limit of Wasserstein spaces on point clouds. *arXiv:1702.03464*, 2017.

Tsang, I. W., Kwok, J. T., and Cheung, P. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

Uzilov, A. V., Keegan, J. M., and Mathews, D. H. Detection of non-coding rnas on the basis of predicted secondary structure formation free energy change. *BMC bioinformatics*, 7(1):173, 2006.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Villani, C. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Science & Business Media, Berlin, 2008. ISBN 978-3-540-71049-3. OCLC: ocn244421231.

Wan, X. A novel document similarity measure based on earth mover's distance. *Information Sciences*, 177 (18):3718 – 3730, 2007. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2007.02.045.

Wang, C., Paisley, J., and Blei, D. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 752–760, 2011.

Weed, J., Bach, F., et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1151–1158, 2010.

Wu, L., Yen, I. E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.-Y., Ravikumar, P., and Witbrock, M. J. Word mover's embedding: From word2vec to document embedding. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4524–4534, 2018.

Wu, X. and Li, H. Topic mover's distance based document classification. In *Communication Technology (ICCT), 2017 IEEE 17th International Conference on*, pp. 1998–2002. IEEE, 2017.

Xu, H., Wang, W., Liu, W., and Carin, L. Distilled wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems*, pp. 1716–1725, 2018.

Ye, J., Wu, P., Wang, J. Z., and Li, J. Fast Discrete Distribution Clustering Using Wasserstein Barycenter With Sparse Support. *IEEE Trans. Signal Process.*, 65(9):2317–2332, May 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2659647.

Yeh, I. and Lien, C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.*, 36(2):2473–2480, 2009. doi: 10.1016/j.eswa.2007.12.020.

Yurochkin, M. and Nguyen, X. Geometric Dirichlet Means Algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2016.

Yurochkin, M., Guha, A., and Nguyen, X. Conic Scan-and-Cover algorithms for nonparametric topic modeling. In *Advances in Neural Information Processing Systems*, pp. 3881–3890, 2017.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K. H., Hoang, T. N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 7252–7261, 2019a.

Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., and Solomon, J. M. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 1599–1609, 2019b.

Yurochkin, M., Guha, A., Sun, Y., and Nguyen, X. Dirichlet simplex nest and geometric inference. In *International Conference on Machine Learning*, pp. 7262–7271, 2019c.

Zwart, J. P., van der Heiden, R., Gelsema, S., and Groen, F. Fast translation invariant classification of hrr range profiles in a zero phase representation. *IEE Proceedings-Radar, Sonar and Navigation*, 150(6): 411–418, 2003.