SPATIAL SIGNAL PROCESSING VIA OPTIMAL TRANSPORT: APPLICATIONS TO TIME SERIES ANALYSIS AND MEASURES ON GRAPHS

A dissertation submitted by

David Charles Gentile

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Mathematics

TUFTS UNIVERSITY

MAY 2026

© Copyright 2026 by David Gentile

Advisor: James M. Murphy

Abstract

The optimal transportation problem endows the space of probability measures with a metric structure which encodes the geometry of the underlying space upon which the measures are defined. This makes it an apt tool for signal processing problems where the geometric information of the domain should not be discarded. This distance between probability measures is known as the Wasserstein metric. We explore several applications of the Wasserstein geometry to signal processing, first for segmentation and clustering of time series data, where a sliding window approach is employed to detect change points in the distribution of the stochastic process generating the observations. Then we consider a novel signal processing paradigm for data which may be represented as a probability measure on the nodes of a graph, and implement a barycentric coding model, which allows for the synthesis and analysis of measures with respect to a given dictionary of reference measures.

Acknowledgments

I owe a boundless debt to every person who helped me get to this point.

I thank Thomas Smith, and Mary Farrell, who taught me that mathematics could be both interesting and fun, and Matthew Aumiller, who taught me to seek beauty in the world, stay weird, and embrace failure.

I thank Mike Siff, who taught me that a good computer scientist is both lazy and impatient, words of wisdom which I will always heed, and who graciously agreed to officiate my wedding, which meant more to Hannah and I than we could ever possibly convey. I thank Philip Ording, who taught me to be patient when my understanding was slow, and I thank Melissa Frazier, who told me to stop being a lazy child and exercise my mind to its fullest potential. I needed to hear that, and maybe still do.

I thank Sarah Dennis, Olif Olson, Cole Cuthbert, Brooke Prakash, Isabella Romano, and Hannah Ford, for their company and friendship and many long nights in the Sarah Lawrence College Library. Without their influence and camaraderie, I probably never would have made it to graduate school.

I thank Thaís De Faria, Daniel Dougherty, Ben McDowell, Annie Price, Jess Lohrman, Adam and Stacy York, and all the members of the DWJ Lab, especially Luca Zocchi and Willow Davis, who welcomed me in Oregon when I was lost in the world.

I thank Merek Johnson, David Cohen, Brendan Mallery, Chandler Smith, Martin Buck, and Kecheng Li, for the many wonderful conversations, about mathematics and otherwise, over the years in 520H, and I will miss all of you terribly as we splinter off into

the future.

I thank Joshua Huang, who was an invaluable collaborator on the change point detection work described in Chapter 3; this work would not have gotten over the finish line without him.

I thank my colleagues at Analysis Group, particularly Stefano Parravano, Upamanu Pathare, and Esteban Léon, as well as my colleagues at the Naval Research Laboratory, particularly Roger M. Oba, Laurie T. Fialkowski, for their excellent mentorship and advice.

I thank my parents, Tom and Linda, and my brother Geof, as well as my uncle Jack Gillis, and I thank my oldest friends in the world, Pierce Keough and Nate Guevin. You have all sustained me from my youngest and most fragile days. I also thank Shane Frisco, whose friendship these last years has meant more to me than they could possibly know.

I thank the members of my committee, Kasso Okoudjou, Abiy Tasissa, and Maria Cameron, for agreeing to take the time to participate in the culmination of time as a graduate student, as well as all the professors who have taken the time to teach me something interesting or chat over tea during my time at Tufts University. Above all, I must give my most profound and especial thanks to my advisor James M. Murphy, who did not need to agree to take me on as a fourth year student adrift and in need of a new path but did anyways. James, thank you for your time, your patience, and your excellent music recommendations. It's been a real pleasure to have worked with you, and I hope that we continue to collaborate in the future.

Finally, I thank my beautiful, funny, one-of-a-kind wife, Dr. Hannah J. Bates, and our two wonderful kittens Ezra and Kali. None of this would be possible without you, and the limitless support you have always provided to me without batting an eye. I love you and I always will.

The work in this dissertation was partially supported by the Camille and Henry Dreyfus Foundation, and National Science Foundation grants DMS-2309519, DMS-2318894.

"What does it mean to be invincible? Invincible ... it's merely a word."

Contents

1 Introduction **2**

2 Background **10**

2.1 Common Objects 11

2.2 Transport in a Euclidean Domain 14

2.2.1 Overview 14

2.2.2 Static Formulation, Kantorovich Problem, and Brenier's Theorem 15

2.2.3 Dynamic Formulation, Benamou-Brenier Theorem 16

2.2.4 Differential Calculus in the Wasserstein Space 19

2.2.5 Barycenters of Probability Measures and the Barycentric Coding Model 21

2.3 Computational Optimal Transport 24

2.3.1 Special Case: Transport on the Real Line 24

2.3.2 Entropic Regularization of the Optimal Transport Problem and Sinkhorn's Algorithm 25

2.4 Optimal Transport on Discrete Spaces 27

2.5 Intrinsic Gradient Descent 33

3 Wasserstein-Based Identification of Metastable States in Time Series Data via Change Point Detection and Segment Clustering **36**

3.1 Introduction 37

3.1.1 Motivation . . . 37
3.1.2 Contributions . . . 41
3.1.3 Organization of the chapter . . . 41
3.2 Background . . . 42
3.2.1 Previous Work on Moving Window Change Point Detection and Relation to the Literature . . . 42
3.2.2 Clustering Techniques Based on Similarity Matrices . . . 47
3.3 Algorithms . . . 49
3.4 Methodology & Numerical Results . . . 52
3.4.1 Generated Data Sets . . . 53
3.4.2 Molecular Dynamics Trajectories . . . 61
3.4.3 Underwater Acoustics . . . 62
3.5 Discussion & Conclusion . . . 66

4 Regularized W_2 Barycenters on Discrete Metric Spaces 68
4.1 Introduction . . . 69
4.1.1 Summary of Contributions . . . 70
4.2 Computing W_2 Barycenters via Entropic Regularization . . . 70
4.3 Methodology & Numerical Results . . . 75
4.4 Discussion . . . 77
4.5 Acknowledgements . . . 79

5 Static and Dynamic Approaches to Computing Barycenters of Probability Measures on Graphs 80
5.1 Introduction . . . 82
5.1.1 Structure of the chapter . . . 84
5.2 Background . . . 84
5.2.1 Relation to the Literature . . . 84

5.3 Algorithms 87
5.3.1 Overview 87
5.3.2 Computation of Optimal Transport on Discrete Spaces 88
5.3.3 Approximating the Exponential Map with the Continuity Equation 92
5.3.4 Approximation of the Exponential Map's Inverse 95
5.3.5 Hyperparameters for the Synthesis and Analysis Algorithms . . . 96
5.3.6 Barycentric Coding Model via Entropic Regularization 97
5.4 Experiments 99
5.4.1 Comparison: Recovering Geodesic Curves via Wasserstein Gradient Descent 100
5.4.2 Comparison: Barycenters Obtained by Gradient Descent vs by Entropic Regularization 104
5.4.3 Effect of Different Initializations for the Descent Scheme 106
5.4.4 Recovery of Coordinates and Effect of Hyperparameters on Recovery Quality 108
5.4.5 Empirical Convergence 110

6 Conclusions and Closing Thoughts **114**
6.1 Future Directions 115
6.2 Postscript, or Making Meaning 118

List of Tables

3.1 Precision and recall scores for the toy data set. 53

3.2 Precision and recall scores for the underwater acoustics data set 65

5.1 Table of hyperparameters associated to each algorithm 97

List of Figures

1.1 Schematic of the generic optimal transport problem 5

1.2 A simple example of a time series change point 7

1.3 A graph on 16 vertices representing the 4 dimensional hypercube. 9

2.1 Different adjacency matrices can generate identical distance matrices . . 12

2.2 Metric derivative of a curve in a manifold 20

2.3 A barycenter of 3 points on a generic manifold 22

2.4 Vectors in the tangent plane determining the analysis system 23

2.5 Effect of entropic regularization on optimal transport plans 26

3.1 Change point detection pipeline 48

3.2 Detecting change points in an idealized setting. 54

3.3 Clustering a toy trajectory with ADPC 55

3.4 Toy Laplace-based Trajectory Segmentation and Clustering 55

3.5 Comparison of segment-based clustering with K-means 55

3.6 Empirical distributions of the clusters in the toy data 56

3.7 Empirical distributions of the clusters in the toy data set learned by K-means. 56

3.8 Detected change points in a Langevin trajectory. 56

3.9 Segment based clustering of a Langevin trajectory. 57

3.10 Langevin Trajectory Segmentation and Clustering 57

3.11 Empirical distributions of identified clusters in a Langevin trajectory. . . 57

3.12 Clustering of the Langevin trajectory via K-means. 58
3.13 Empirical distributions for Langvein data, segment-based clustering. . . . 58
3.14 Empirical distributions for Langevin data clustered via K-means. 58
3.15 Detected change points for a particle in a Prinz potential. 59
3.16 Segment-based clustering of a Prinz trajectory. 59
3.17 Empirical distributions for Prinz data, segment-based clustering. 60
3.18 Empirical distributions for Prinz data clustered via K-means. 60
3.19 Structure of Ace-Val-NMe, annotated with backbone dihedral angles. . . 61
3.20 VDP dihedral angles labeled via segment-based clustering. 63
3.21 Valine Dipeptide Clustering 63
3.22 Comparison of ADP and VDP dihedral angles. 64
3.23 Detected change points in underwater acoustics data. 66

4.1 Comparison of the transport and L^2 interpolation of densities. 73
4.2 Graph metrics relative to a fixed node. 75
4.3 Interpolation of two measures using the shortest-path metric. 77
4.4 Interpolation of two measures using the diffusion distance. 78
4.5 Barycenter of several measures using the diffusion distance. 78

5.1 Schematic of the barycentric coding model. 83
5.2 Schematic of the Chambolle-Pock routine for approximating geodesics. . 93
5.3 Barycenter of 4 measures on the hypercube. 100
5.4 Barycenter of 4 measures on a spatial geography graph. 101
5.5 Computation of barycenters on a 2 node graph 102
5.6 Distribution of norm differences computed via Chambolle-Pock and descent. 103
5.7 Distribution of norm differences computed via CP and descent. 104
5.8 Barycenters of 3 measures computed via static and dynamic methods. . . 107
5.9 Effect of initialization on descent scheme 109

5.10 Distribution of observed relative errors in recovery of coordinates. 110

5.11 Recovery of coordinates for synthesized barycenters. 111

5.12 Convergence trendlines of descent scheme. 112

Examining Committee

Chair of committee:	James M. Murphy Associate Professor Department of Mathematics Tufts University
Internal committee member:	Kasso Okoudjou Robinson Professor and Department Chair Department of Mathematics Tufts University
Internal committee member:	Abiy Tasissa Assistant Professor Department of Mathematics Tufts University
External committee member:	Maria Cameron Professor Department of Mathematics University of Maryland

SPATIAL SIGNAL PROCESSING VIA OPTIMAL TRANSPORT: APPLICATIONS TO TIME SERIES ANALYSIS AND MEASURES ON GRAPHS

Chapter 1

Introduction

"It is customary to preface a work with an explanation of the author's aim, why he wrote the book, and the relationship in which he believes it to stand to other earlier or contemporary treatises on the same subject."

— G.W.F. Hegel[1]

[1]Translation A.V. Miller [HMF77]

It is difficult to find the right place to begin introducing the work which has occupied me in the past three years; from this vantage, the initial point of departure seems so far off in the distance. Worse yet, the place to which I've come is no high mountain of knowledge with vistas and crisp air. It is more of a lean-to in the forest, a place where I must shelter for a moment and assess the situation, and a road which stretches off further yet going who knows where. If the mission of this missive is to answer the questions "what have you done, and why?" then perhaps it will be simpler to begin with the former question and address what was done, in the hope that after discussion of "what" that "why" will become apparent, for author and reader both.

The contribution of this dissertation consists of two main projects unified around a broad theme: signal processing via optimal transport. The first project concerns applications of optimal transport to time series analysis via segmentation; given a time series, we use optimal transportation to identify change points, and then identify patterns in the overall data by clustering the much smaller number of segments in the transport geometry. The second project concerns the development of a barycentric coding model for probability measures supported on graphs. This latter project proceeded in two stages. Firstly, we developed a barycentric coding model by leveraging entropic optimal transportation algorithms in combination with cost matrices that encode some aspect of the connectivity information intrinsic to a given graph. This approach is successful, but the method requires a caveat: proceeding in this way demands the practitioner choose for themselves what cost metric meaningfully encodes the relevant graph. The approach based on entropic regularization can be thought of as a "static" way of tackling the problem, because the algorithm employed searches for an optimal *transport plan*, which describes how much mass to move from one place to another. After experimenting with the "static" point of view, we opted to investigate an approach which depends on a "fluid dynamics" inspired perspective, which traces its roots back to the study of gradient flows in the space of measures supported on Markov chains. By contrast with the first transport plan-based

"static" implementation, this approach to identifying barycenters adopts the "fluid dynamics" perspective on transport by using a discretized continuity equation to update its iterates. While more difficult from a computational and implementation stand point, this second approach has the benefit that it does not require the practitioner to make a choice about how to best encode the edge information on the graph: it acts directly on the (symmetric) random walk representation of a given graph, which is a well-defined object.

But this is an introduction, and already I am rushing ahead of myself to get the conclusions out before the principals have been introduced. So, what is optimal transport? The problem was first communicated by Monge in 18^{th} century France [Mon81], and I will roughly summarize its physical motivation as this: given an initial configuration of mass (say, a pile of raw materials) and a target configuration of mass (a completed structure), can we find a mapping which transforms the former into the latter in a way which minimizes total work done? This intuitive problem in applications, represented visually in Figure 1.1, turns out to open the door to a rich mathematical theory, which spans from convex analysis to partial differential equations to statistical mechanics and beyond. For an overview of the topic, the reader will be hard-pressed to find a more thorough explanation than Villani's [Vil09]. The magnitude of that tome is commensurate with the intricacies and beauty of its subject, and it is neither appropriate nor feasible to try condensing all of that information down into a pithy introduction for this dissertation. In Chapter 2, I will endeavor to give an overview of the language and results necessary to follow and understand this text.

In short, optimal transport provides a toolbox for analyzing probability measures in a manner which does not sacrifice information about the geometry of the underlying domain, while simultaneously supplying the set of probability measures with a geometric structure of its own, known as *the Wasserstein geometry*; in so many words, the distance between two measures is the total cost it takes to transport one to the other

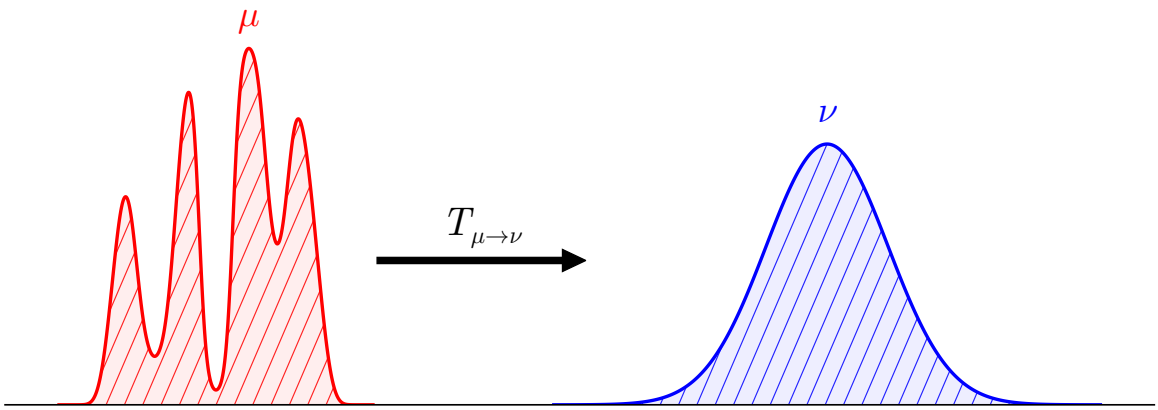


Figure 1.1: At a high level, the optimal transport problem is about transforming one pile of mass into another one, here represented by probability distributions μ and ν. In certain circumstances one can find a mapping $T_{\mu\to\nu}$ which accomplishes this, something we will address in Chapter 2.

— this contrasts with, e.g., the L^2 difference between densities of two absolutely continuous distributions, or the Kullback-Leibler divergence between them. In the world of applied mathematics, this has made optimal transport a popular computational tool and particularly attracted attention from the statistics and machine learning communities [GM00, ACB17, GPC18, CTV19, Cut13, PC19, Rig25]. In recent years, this has opened up new paradigms for signal processing problems [SHB$^+$18, MAMT23, MMA25, WMAM25, CRT20]. The barycentric coding model for probability measures [BPC16, WJT$^+$22, GHL24, MAMT23, GM25, WMAM25, MMA25, JCP25] is a recently developed tool for the synthesis and analysis of probability measure-valued data. The central conceit of the model is that the theory of barycenters in the L^2-Kantorovich metric (also known as the 2-Wasserstein metric) on probability measures supported on $\mathbb{R}^d$ [AC10] provides a synthesis mechanism for expressing the weighted average of several reference measures, provided that the weights are simplicial, i.e., that they are non-negative and sum to unity. Simultaneously, an inverse problem for this weighted synthesis of measures can also be established and solved, meaning that for a given target measure and

a family of reference measures, it is possible to identify an appropriate set of weights so that (re)-synthesizing the references according to these found weights results in the target measure.

In Chapter 3, our primary interest is in leveraging the Wasserstein metric on probability measures to facilitate change point detection and, together with classic techniques in data clustering, to produce compact descriptions of large and complicated time series data sets, specifically trajectories emerging from molecular dynamics simulations, by studying applications of optimal transportation to the *change point detection* problem. Change point detection (CPD) is the problem of identifying a segmentation of a time series which partitions the data into a small number of internally coherent sequences. An example of a change point in a time series is shown in Figure 1.2. The CPD problem is an essential task in the analysis of time series data, and many different methods, covering a wide breadth of approaches, exist to tackle the problem [AC17]. As a statistical tool, it finds applications in finance [KPLS22, Hab21, PP17], climate modeling [RCW$^+$07, BCS12, GLR13], speech analysis [TR08, CRBV13, Gup15], edge detection [TDL03, HHRR20, HT88], and health monitoring [LWH18, AHD25, WLJ24]. The task of a such an algorithm is to produce a segmentation of a time series by identifying the specific observations at which a significant qualitative change occurs — for example, the moments where blood pressure becomes excessively elevated, or when anomalous trades are made in the stock market. Our transport-based segmentation approach is an inversion of a pattern observed in some CPD pipelines [AM07, ZLWX08, RMB$^+$10], wherein clustering is performed first directly on the data and used to identify segmentation; here we instead use change point detection to facilitate the labeling of the data by first identifying a segmentation and then clustering the segments, treating them as a reduced representation of the original information.

Graph theory is a topic in discrete mathematics. A graph, the central object of study, consists of a collection of discrete vertices V and edges connecting the vertices $E \subset V \times V$.

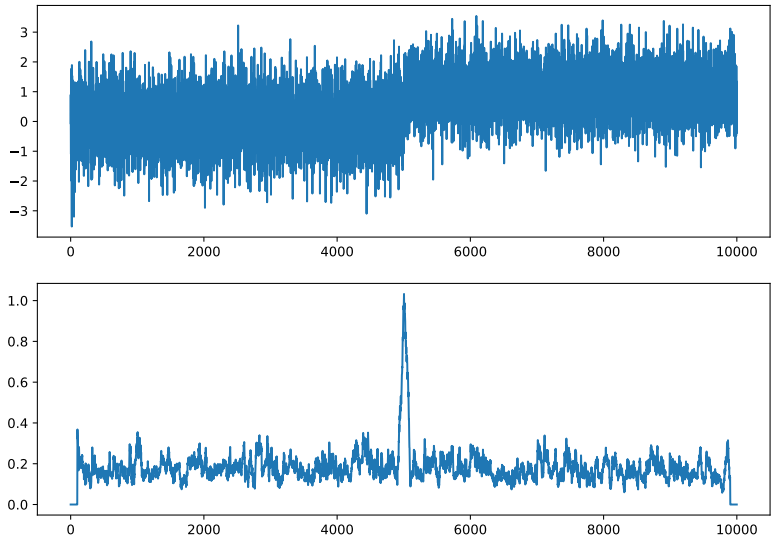


Figure 1.2: The simplest possible instance of a change point: an abrupt change in the law occurs at time t=5000. The change in the law of the process is reflected in the Wasserstein metric derivative. Top: an artificial time series, with first 5000 samples drawn from a $\mathcal{N}(0, 0.85)$ distribution and the second 5000 samples drawn from a $\mathcal{N}(1, 0.75)$ distribution. Bottom: approximated metric Wasserstein derivative of the timeseries, with extremum appearing at $t = 5000$, corresponding to the abrupt change in the distribution.

An example of such an object is shown in Figure 1.3: the extremal points of the hypercube form the vertex set, and are connected by edges if they lie in the same coordinate plane. Applications and problems centering on graphs have long attracted the attention of mathematicians, dating as far back as Euler's celebrated consideration of the seven bridges of Königsberg [Eul41]. Problems formulated in the graph setting have ranged from classical combinatorial optimization considerations, such as k-coloring problems [Bro41] and max-flow/min-cut algorithms [FJF56], to modern machine learning tools such as spectral clustering [NJW01] and graph convolutional neural networks [WCPZ22]. In recent years, signal processing on graphs has become a problem of growing interest in the applied mathematics community because graphs and discrete metric spaces arise naturally as data structures in domains such as sensor networks, opinion dissemination, and measurements of community segregation [SMD+19b, RCR21, DMW23], where the signal domain is best represented as some discrete, potentially irregular space. The essence of a graph is in its connectivity information — i.e., which nodes are connected to each other

— and it therefore is natural to look for signal processing methods which incorporate that connectivity information. Study of graph connectivity finds application in several subdomains of machine learning, such as clustering methods [NJW01, RW13, NWJH16] and supervised/semisupervised learning problems [SYXK23, CC23, HHRR20].

In Chapters 4 & 5, we present two distinct numerical implementations of the barycentric coding model for measures supported on graphs, i.e., we provide algorithms and code for the synthesis and analysis of measures with respect to a given dictionary of reference measures. Synthesizing a barycenter of a given family of reference measures, called a dictionary, can be thought of as approximating the center of mass in a certain geometry on the simplex — in order to make this approximation numerically, we perform intrinsic gradient descent using the Riemannian structure on the probability simplex which is induced by a transport-like metric for graph-supported measures; first the intrinsic gradient is estimated by approximating the geodesic curves between the current descent iterate and the dictionary elements, and then the iterate is pushed forward via a discretization of a graph-adapted continuity equation. A measure can also be analyzed relative to a given dictionary by solving a certain quadratic program — it is known in the Euclidean case that the coefficients of this quadratic program are related to the inverse of the exponential map, and hence can also be obtained by approximating the geodesics between target and reference measures [WJT$^+$22].

While these projects and their respective chapters each have distinctive flavors and different computational challenges, they share a common motif: the application of calculus in a transport metric, in particular an interest in critical points, objects first encountered in an introductory calculus course. In calculus, we teach freshman that if you want to optimize a function, you should take its derivative, set that equal to 0, and then perform some kind of second-order analysis to confirm whether the critical point is a maximum or a minimum, which we can potentially avoid if some other nice analytical property, like convexity, is in play. Here we carry on the spirit of freshman calculus: in order to look

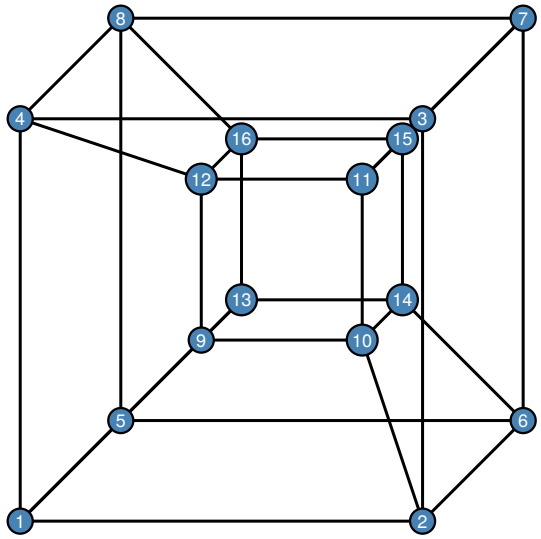


Figure 1.3: A graph on 16 vertices representing the 4 dimensional hypercube.

for change points in noisy data, we look for local maximizers of the Wasserstein metric derivative; in order to look for barycenters, which we characterize as minimizers of a certain variance functional, we look for zeros of the functional's derivative vector field.

Chapter 2

Background

"If you haven't understood the א, how can you teach others the ב?"

— Yeshua of Nazareth[1]

[1]Translation Guy Davenport & Benjamin Urrutia

Chapter Overview

We recap the basic language and results needed to engage with the remainder of the dissertation. We recall basic definitions from graph theory, probability theory, and optimal transport. Graph metrics such as the shortest path and the diffusion distances are recalled. Static and dynamic formulations of the optimal transportation problem are discussed, as well as the celebrated entropic regularization of the transportation cost, commonly used in applications and numerical approaches. The barycentric coding model for measures is introduced. Riemannian gradient descent is also discussed.

2.1 Common Objects

We start by recalling some common definitions from the theories of graphs, analysis, and probability.

A (finite) *graph* (with its shortest path metric) is a quintessential example of a discrete metric space. Specifically, a graph is a pair $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the vertex set and $E \subset V \times V$ is the edge set, encoding the information of which vertices connect to each other. For a fixed pair of vertices $u, v \in V$, a sequence of vertices $v_1, v_2, \ldots, v_k$ is called a uv walk if $v_1 = u, v_k = v$, and $(v_i, v_{i+1}) \in E$ for $i = 1, 2, \ldots, k-1$. By defining

$$d(u, v) := \inf \left\{ \sum_i \mathbf{1}_{(v_i, v_{i+1}) \in E} : (v_1, \ldots, v_k) \text{ is a } uv \text{ walk} \right\},$$

we obtain the shortest path metric on the graph G. The *degree* of a node $x \in V$ is the number of edges $y \in V$ such that $(x, y) \in E$), and is denoted $\deg(x)$.

For the remainder, it will be useful to represent graphs via Markov transition kernels. Note that we can always pass back and forth between Markov transition kernels

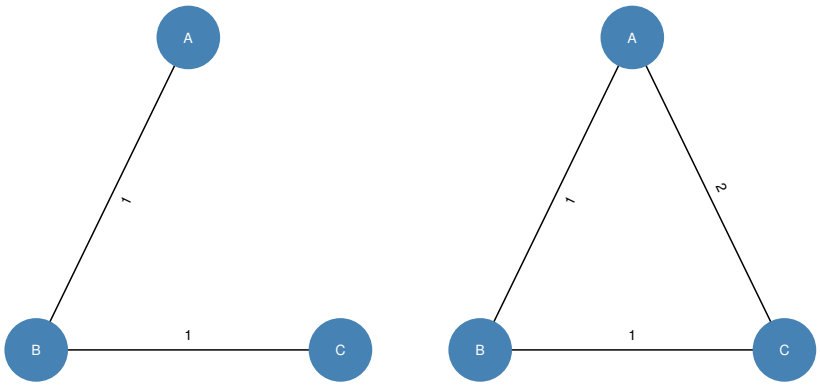


Figure 2.1: The shortest path metric is not always enough information to recover the edge information in a graph. Left: a chain graph on 3 vertices, with all edges weighted 1; right: a complete graph on 3 vertices, with edge (A, C) weighted 2. Both graphs have the same shortest-path pairwise distance matrix, but different adjacency matrices.

and graphs. To construct a Markov transition kernel representing an arbitrary simple, connected graph $G = (V, E)$, one takes

$$\pi(x) = \frac{\deg(x)}{|E|}; \qquad Q(x,y) = \begin{cases} \frac{1}{|E|\pi(x)} & (x,y) \in E \\ 0 & \text{else} \end{cases}.$$

We also note that this construction can be extended to graphs equipped with symmetric weight matrices $[\omega_{x,y}]_{(x,y)\in E}$ by setting $\deg_\omega(x) := \sum_{y\in V} \omega(x,y)$. Then the corresponding Markov kernel is defined by

$$\pi(x) = \frac{\deg_\omega(x)}{\sum_{y\in V} \deg_\omega(y)}; \qquad Q(x,y) = \begin{cases} \frac{\omega(x,y)}{\deg_\omega(x)} & (x,y) \in E \\ 0 & \text{else} \end{cases}.$$

This representation will be useful both because it serves as part of the basis for the theory of transport on discrete metric spaces, and also because it opens up a new class of metrics on graphs. Because the shortest path metric does not always capture all the connectivity information of a graph, other metrics have been constructed and considered

— in particular, the parameterized class of metrics known as the diffusion distances [CL06, CLL$^+$05].

Definition 2.1.1. The squared *diffusion distance* between nodes $x, y \in V$ with time parameter t is given by

$$d_t^2(x,y) := \sum_{w \in V} \frac{(p(w,t|x) - p(w,t|y))^2}{\pi(w)},$$

where $p(w,t|x)$ is the probability of transitioning from node x to node w after t steps.

The diffusion distance for a given time t defines a metric on the vertices that measures the connectivity between a pair $x, y \in V$ — points connected by many short paths will be closer together in the diffusion distance [CLL$^+$05, CL06]. Diffusion distances can be understood as a type of non-linear dimensionality reduction, and we refer the reader interested in this aspect of diffusion distances to the papers [MM19, TdSL00, RS00, BN03]. Connectivity-encoding metrics on graphs are an area of active interest, and other such metrics, such as the Fermat distances [BRS12, TLMM24] might also be employed in this setting, which we leave to future work. Later on in Chapters 4 & 5, we will use the diffusion distances to construct cost matrices in the context of computing entropically regularized W_2 barycenters on discrete spaces.

Definition 2.1.2. Take p to be a positive integer. The *p-simplex*, denoted Δ^{p-1} is the set of vectors in $\mathbb{R}^p$ with non-negative components which sum to 1, in other words

$$\Delta^{p-1} = \left\{ \lambda = (\lambda_1, \ldots, \lambda_p) : \lambda_i \geq 0, \sum_i \lambda_i = 1 \right\}.$$

We denote the set of non-negative measures ν with total mass of 1 defined on a measurable space Ω by $\mathcal{P}(\Omega)$. If $\Omega = \mathbb{R}^d$, then we denote by $\mathcal{P}_2(\Omega)$ the set of probability measures ν such that $\int_\Omega \|x\|^2 \, \mathrm{d}\nu(x) < \infty$. If there exists a density function $\rho(x) : \Omega \to \mathbb{R}$ such that for all Lebesgue measurable $A \subset \Omega$ it holds that $\nu(A) = \int_A \rho(x) \, \mathrm{d}x$, then ν is

said to be absolutely continuous, and belongs to the set $\mathcal{P}_{2,ac}(\Omega)$.

Definition 2.1.3. Let $\mathcal{P}(\Omega)$ be the set of Borel probability measures on Ω and let $\mu, \nu \in \mathcal{P}(\Omega)$. A probability measure $\gamma \in \mathcal{P}(\Omega \times \Omega)$ is said to be a *coupling* of the measures μ and ν if for all Borel measurable sets A, the following identities hold:

$$\gamma(A \times \Omega) = \mu(A); \quad \gamma(\Omega \times A) = \nu(A).$$

The set of all couplings of μ and ν is denoted $\Pi(\mu, \nu)$.

Note that $\Pi(\mu, \nu)$ is always non-empty because the product measure $\mu \otimes \nu$ is itself a coupling. In the context of finite discrete measures, couplings may be represented as matrices satisfying linear constraints. If $T : \mathbb{R}^d \to \mathbb{R}^d$ is a measurable mapping, and $\gamma \in \Pi(\mu, \nu)$ has the form $(\mathrm{Id} \times T)\#\mu$, where $\#$ is the measure push-forward operator, then we say that the plan γ is induced by mapping T, and we call T a *transport map* taking μ to ν.

2.2 Transport in a Euclidean Domain

2.2.1 Overview

Conceptually, the *optimal transportation problem* is this: suppose we have a pile of mass that needs to be moved from one location to a new one, and that transporting a unit of mass from point x to point y has associated cost $c(x, y)$ — how can the mass be transported in a cost minimizing way, and what is the minimum cost associated to the optimal method? As it happens, this intuitive optimization problem is the genesis for a geometry of probability measures encoding the structure of the underlying space on which those measures are defined.

The essential things about optimal transport that we need to understand are: that optimal transport endows the space of probability measures with a geometry; that geodesic

curves in the geometry of transport obey the continuity equation and in particular minimize its action; that barycenters are defined in terms of a geometric minimization problem.

2.2.2 Static Formulation, Kantorovich Problem, and Brenier's Theorem

Let us recall the basic building blocks for the L^2-Kantorovich geometry of probability measures. The theory of optimal transportation in its modern formulation is centrally concerned with the problem of solving a certain linear program with constraints defined in terms of measures. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$, let $c(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$. Then the optimal transport cost for μ and ν with cost c is given by

$$OT(\mu, \nu) := \inf \left\{ \int_{\Omega \times \Omega} c(x, y) \, \mathrm{d}\gamma : \gamma \in \Pi(\mu, \nu) \right\}. \tag{2.1}$$

The minimization problem in (2.2.2) is referred to as the Kantorovich problem. In general, the cost function $c(x, y)$ can take many forms, but the case which has proven most well-studied has turned out to be when $c(x, y) = \frac{1}{2} d^2(x, y)$. For the case of $\Omega = \mathbb{R}^d$, we denote $d(x, y)$ by $\|x - y\|$; norms in other Hilbert spaces are differentiated by subscript, e.g. $\| \cdot \|_H$. Taking $\Omega = \mathbb{R}^d$ is the setting of Brenier's theorem [Bre91], one of the foundational results of the modern field, which we reproduce here.

Theorem 2.2.1 ([Bre91], Theorem 1.3)**.** *Let μ, ν be probability measures on $\mathbb{R}^d$ satisfying $\int_{\mathbb{R}^n} \|x\|^2 \, \mathrm{d}\mu(x), \int_{\mathbb{R}^n} \|y\|^2 \, \mathrm{d}\nu(y) < \infty$. Let $c(x, y) = \frac{1}{2}\|x - y\|^2$, and suppose that μ gives no mass to $(d-1)$ surfaces of class C^2. Then there exists a μ-a.e. unique optimal transport map T from μ to ν, and a convex function ϕ such that $T = \nabla \phi$.*

The convex function ϕ is known as the *Brenier potential* and T is referred to as the *optimal transport (OT) map*. It can be shown that in this case, and under the further stipulation that space of probability measures is restricted to the set $\mathcal{P}_2(\Omega)$ of those

measures with finite second moment, then (2.2.2) defines a metric on $\mathcal{P}_2(\Omega)$, known as the 2-Wasserstein or L^2-Kantorovich metric and denoted by W_2 [San15].

Note also some other foundational work in the modern field of transportation: in [SK87] the cyclical monotonicity characterization of optimal plans is established; in [McC97], what is now known as the *McCann interpolant* was introduced, which describes the dynamic transportation of one distribution to another, achieved through pushing the initial distribution forward by the linear interpolation of the optimal transport map and the identity function; in [GM96], existence and uniqueness for the optimal transport problem are established for a broad class of cost functions both convex and concave.

The Kantorovich problem can be thought of as a static formulation of the optimal transportation problem — the optimal coupling γ^* describes the measure of mass moving from some set A to some set B, but *a priori* gives no information about the path the mass takes when moving between the sets. As we will emphasize below, the dynamical formulation of transportation will be of the utmost importance for our work.

2.2.3 Dynamic Formulation, Benamou-Brenier Theorem

It is also a remarkable fact that the W_2 metric is formally similar to the kinds of metrics which arise in Riemannian geometry. This fact is expressed by the Benamou-Brenier formula reproduced below. In order to provide necessary context for this formula, first recall the *continuity equation with temporal boundary conditions*. A function $\rho : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ and a time-dependent vector field $V : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ are said to satisfy the continuity equation with temporal boundary data (ρ_0, ρ_1) if the following holds:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho V) = 0; \quad \rho(0,x) = \rho_0(x), \quad \rho(1,x) = \rho_1(x). \tag{2.2}$$

On a physical level, this equation is asserting that a quantity of mass is evolving continuously from distribution ρ_0 to distribution ρ_1. The Benamou-Brenier formula connects

this fluid dynamics problem to the optimal transportation problem by asserting that the optimal transport cost for measures ρ_0 and ρ_1 with ground cost given by $\|\cdot - \cdot\|^2$ is exactly the minimum energy required to move the mass continuously from one configuration to another.

Theorem 2.2.2 ([BB00], Proposition 1.1)**.** *Given absolutely continuous probability measures on $\mathbb{R}^d$ with finite second moment ρ_0, ρ_1, the square of the L^2-Kantorovich distance satisfies*

$$W_2^2(\rho_0, \rho_1) = \inf \left\{ \int_0^1 \int_{\mathbb{R}^d} \|V(t,x)\|^2 \rho(t,x) \, \mathrm{d}x \, \mathrm{d}t \right\},$$

where the infimum is taken over all pairs (ρ, V) satisfying the continuity equation with boundary data ρ_0, ρ_1.

To see how this can be interpreted as Riemannian metric, notice that the integrand might as well be written

$$\begin{aligned} W_2^2(\rho_0, \rho_1) &= \inf \left\{ \int_0^1 \int_\Omega \rho(t,x) \|V(t,x)\|^2 \, \mathrm{d}x \, \mathrm{d}t \right\} \\ &= \inf \left\{ \int_0^1 \|V(t,x)\|^2_{L^2(\rho(t,x))} \, \mathrm{d}t \right\}. \end{aligned} \tag{2.3}$$

The RHS can now be read as the minimization of an energy functional, with the minimization occurring over the space of all absolutely continuous curves with endpoints ρ_0 and ρ_1. This point will be important in the development of a transport-like metric for probability measures supported on graphs. We want to emphasize here that the Benamou-Brenier formula gives a *dynamic* interpretation of the optimal transportation problem — the minimizer of (2.3) will be a curve in the space of measures $\rho(t)$, which can be interpreted as tracking the instantaneous distribution of mass at each time $t \in [0,1]$ as the distribution evolves from ρ_0 to ρ_1. This is in contrast to the static formulation given by the Kantorovich problem.

The energy minimizing curve may also be connected to the static plans and maps

formulation of Kantorovich via the McCann interpolation. In [McC97], it is shown that if T^* is the optimal transport map pushing ρ_0 to ρ_1, then the interpolation, defined by the expression

$$\rho_t := ((1-t)\,\mathrm{Id} + tT^*)\#\rho_0,$$

is energy minimizing over the interval $[0,1]$.

With this interpretation of the L^2-Kantorovich metric in mind, a natural question to ask is the following: is there a corresponding interpretation of the notion of a tangent space at a point? The answer is yes, and, following [AGS05], we give the following definition:

$$T_\nu \mathcal{P}_{2,ac}(\mathbb{R}^n) := \{\beta(\nabla\phi - \mathrm{Id}) : \beta > 0, \phi \in C_c^\infty(\mathbb{R}^n), \phi \text{ convex } \},$$

where the closure is in $L^2(\nu)$. If u, v are two tangent vectors in $T_\nu \mathcal{P}_{2,ac}(\mathbb{R}^n)$, then their inner product is defined by multiplying them together and integrating against the measure ν, $\langle u, v\rangle_\nu := \int_{\mathbb{R}^d} \langle u, v\rangle \,\mathrm{d}\nu(x)$. The existence of this formal Riemannian interpretation will be important because it opens the door to applying first-order optimization methods for minimizing functionals defined on $(\mathcal{P}_{2,ac}(\Omega), W_2)$.

In full generality, absolute continuity of the boundary measures is not necessary for (2.2.2) to hold. In [AGS05], a rigorous investigation is undertaken to show that in a setting as general as Ω being a separable Hilbert space, there is a direct correspondence between absolutely continuous curves $\nu_t : (a,b) \to \mathcal{P}(\Omega)$ and solutions *in the sense of distributions* to the continuity equation. Taking $\Omega = \mathbb{R}^d$, a curve ν_t of Borel probability measures and a time-dependent vector field v_t are said to satisfy the continuity equation in the sense of distributions if the following holds for all test functions $\varphi \in C_c^\infty(\mathbb{R}^d \times (a,b))$:

$$\int_a^b \int_{\mathbb{R}^d} \left(\partial_t \varphi(x,t) + \langle v_t(x), \nabla_x \varphi(x,t)\rangle\right) \,\mathrm{d}\nu_t(x)\,\mathrm{d}t = 0.$$

This weak formulation allows for a class of admissible solutions which are lacking in

the regularity that would normally be required to solve the equation (2.2) in the classical sense. This will be important later on when finding weak solutions to a modification of (2.2) is of use for approximating the discretized transport distance for measures on graphs (see Section 5.3.2).

2.2.4 Differential Calculus in the Wasserstein Space

We will now introduce some definitions and results from the theory of Wasserstein gradient flows. Gradient flows are one of the most celebrated results associated to the theory of modern transport, particularly the results of Otto in [JKO98, Ott01], and the theory of gradient flows in general metric spaces developed by Ambrosio, Gigli, and Savare [AGS05].

Metric Derivatives

In Chapter 3, we employ the scalar *Wasserstein metric derivative*, defined (following [AGS05]) as follows: let $\gamma : [0,1] \to \mathcal{P}_2(\Omega)$, let $t_0 \in (0,1)$. We say that γ has metric derivative $|\dot{\gamma}|$ at t_0 if the following limit exists:

$$|\dot{\gamma}| = \lim_{t \to t_0} \frac{W_2(\gamma(t), \gamma(t_0))}{t - t_0}. \tag{2.4}$$

Metric derivatives may be defined in any metric space for which a notion of absolute continuity is available. In [Amb90], it was shown that the limit in (2.4) exists under very general conditions on the measures $\gamma(t)$, namely that $\gamma(t)$ be absolutely continuous in time with respect to the Lebesgue measure on $[0,1]$ for almost-all t.

The intuition for the metric derivative is that it represents the instantaneous speed of the "particle" in the space of probability distributions, which we will think of as the law of some stochastic process. The geometry of the definition is shown in Figure 2.2. The metric derivative is the starting point for the much-celebrated theory of gradient flows

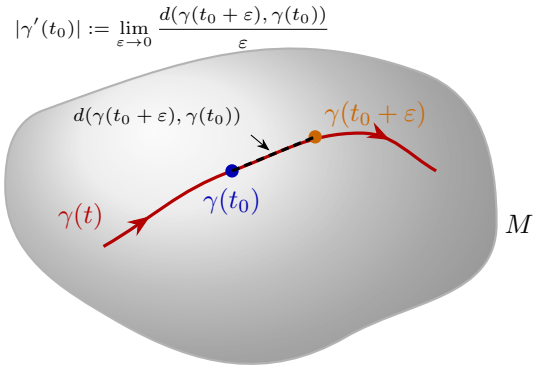


Figure 2.2: A curve on a manifold; the metric derivative is defined in terms of a limit of a difference quotient where the numerator is given by the difference between nearby points.

on metric spaces [AGS05], which could in the future serve as a good basis from which to expand the change point detection methodology.

Wasserstein Gradients

A *gradient flow* in Euclidean space is a partial differential equation in time and space defined by a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$, together with Cauchy boundary data, expressed by

$$\begin{cases} x'(t) = -\nabla F(x(t)), \quad t > 0, \\ x(0) = x_0. \end{cases} \tag{2.5}$$

The idea is simple enough: if $x(t)$ solves (2.5), then x is a curve of maximal slope and moves towards minimizers of F, evolving in the direction of $-\nabla F$ at each time t. Of particular interest is the time discretization of such a scheme: fix a $\tau > 0$ and define the iterates

$$x_{k+1}^\tau = \operatorname*{argmin}_x \frac{1}{2\tau} \|x - x_k^\tau\|^2 + F(x). \tag{2.6}$$

This is known as a minimizing movements scheme, and can be directly related to the implicit Euler scheme for solving (2.5).

It is the remarkable result of [JKO98] that if one replaces $\|\cdot - \cdot\|$ in (2.6) by W_2

and considers the discretized gradient flow in the space $\mathcal{P}_{2,ac}(\Omega)$ with $\Omega = \mathbb{R}^d$ of the Boltzmann entropy , defined by

$$H(\rho) = \int_\Omega \rho(x) \log(\rho(x)) \, \mathrm{d}x, \tag{2.7}$$

one finds that the solution to this gradient flow problem exactly coincides with heat flow.

Theorem 2.2.3 ([JKO98], Theorem 5)**.** *Let $\rho_0 \in \mathcal{P}_{2,ac}(\Omega)$ satisfy $H(\rho_0) < \infty$, and for given $\tau > 0$, let $\{\rho_k^\tau\}$ solve* (2.6)*. Define the interpolation $\rho^\tau : (0,\infty) \times \mathbb{R}^d \to [0,\infty)$ by*

$$\rho^\tau(t) = \rho_k^\tau(t) \quad t \in [kh, (k+1)h), \quad k = 0, 1, 2, \ldots.$$

Then as $\tau \downarrow 0$,

$$\rho^\tau(t) \rightharpoonup \rho(t) \quad \textit{weakly in } L^1(\mathbb{R}^d) \forall t \in (0,\infty),$$

where $\rho \in C^\infty((0,\infty) \times \mathbb{R}^d)$ is the unique solution of

$$\frac{\partial \rho}{\partial t} = \Delta \rho.$$

Otto's result is fundamental to the modern study of evolution equations, and we recall it here because it will serve as motivation for the discussion of transportation on a graph.

2.2.5 Barycenters of Probability Measures and the Barycentric Coding Model

Now with a geometry of probability distributions in hand, we would like to consider extending the notion of convex combinations of points to this setting. An example of weighted center of mass on a non-Euclidean domain is shown in Figure 2.3. In a Euclidean space, given points $\{x_i\}_{i=1}^p$ and $\lambda \in \Delta^{p-1}$, the convex combination $\sum \lambda_i x_i$ minimizes the function $\sum \lambda_i \|x_i - x\|^2$. Replacing $\|\cdot - \cdot\|^2$ by $d^2(\cdot,\cdot)$ suggests how convex combinations

might be reasonably defined in generic metric spaces lacking linear structure, such as $(\mathcal{P}_2(\Omega), W_2)$. In [AC10], the minimization problem

$$\inf_{\nu} \left\{ J(\nu) := \sum_{i=1}^{p} \lambda_i W_2^2(\nu_i, \nu) \right\} \tag{2.8}$$

is studied, wherein the coefficients $(\lambda_1, \ldots, \lambda_p)$ are real numbers such that $\lambda_i \geq 0$ for each $i = 1, \ldots, p$, and $\sum \lambda_i = 1$, and where the ν_i are probability measures with finite second moment defined on $\mathbb{R}^n$. The fixed measures $\{\nu_i\}_{i=1}^p$ are referred to as the *reference measures*, and the coefficients $\{\lambda_i\}_{i=1}^p$ are the *weights*. We refer to the functional J as the *variance functional*. Minimizers ν^* of (2.8) are referred to as *barycenters* of the reference measures $\{\nu_i\}_{i=1}^p$ with weights $\{\lambda_i\}_{i=1}^p$, and we write $\nu^* = \text{Bary}(\{\nu_i\}, \lambda)$. Recall that for a Hilbert space H and convex function $\varphi : H \to \mathbb{R}$, the *Fenchel conjugate* of φ is denoted φ^* and defined by the expression:

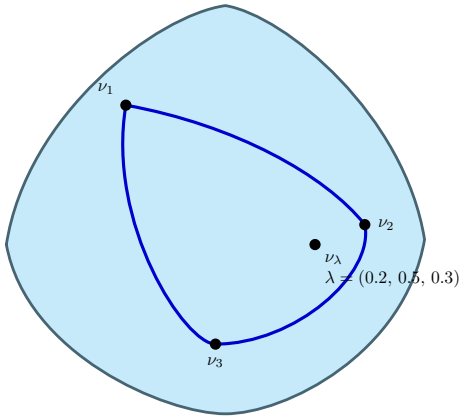


Figure 2.3: Weighted center of mass on a non-Euclidean space; ν_λ is the synthesis of the points ν_1, ν_2, ν_3 with weights $\lambda = (0.2, 0.5, 0.3)$

$$\varphi^*(y) := \sup_x \langle x, y \rangle_H - f(x).$$

A major result of [AC10] is the following:

Theorem 2.2.4 ([AC10], Proposition 3.8)**.** *Let $\mathcal{M}(\mathbb{R}^d)$ be the space of bounded Radon measures on $\mathbb{R}^d$, let $\mathcal{M}_+^1(\mathbb{R}^d)$ be the set of Radon probability measures on $\mathbb{R}^d$, and let*

$$X' = \left\{ \mu \in \mathcal{M}(\mathbb{R}^d) : (1 + \|x\|^2)\mu \in \mathcal{M}(\mathbb{R}^d) \right\}.$$

Assume that ν_i vanishes on sets of Hausdorff dimension $\leq d - 1$ for every $i = 1, \ldots, p$,

and let $\nu \in X' \cap \mathcal{M}_+^1(\mathbb{R}^n)$. *Then the following conditions are equivalent:*

1. ν *solves* (2.8).

2. *There exist convex potentials* ψ_i *such that* $\nabla\psi_i$ *is Brenier's map transporting* ν_i *to* ν, *and a constant* C *such that* $\sum_{i=1}^{p} \lambda_i \psi_i^*(y) \leq C + \frac{\|y\|^2}{2}, \forall y \in \mathbb{R}^n$, *with equality* ν*-a.e. Furthermore, the potentials* ψ_i^* *are differentiable on* $\mathrm{Supp}(\nu)$ *and satisfy* $\sum_{i=1}^{p} \lambda_i \nabla\psi_i^* = \mathrm{Id}$.

The last piece of background from the theory of transport in the Euclidean setting is the *barycentric coding model for measures*. Fix a family of reference measures $\{\nu_i\}_{i=1}^p \subset \mathcal{P}_{2,ac}(\mathbb{R}^n)$, and let ν_λ be the minimizer of (2.8) for a given $\lambda \in \Delta^{p-1}$. Then we denote by $\mathrm{Bary}(\{\nu_i\}_{i=1}^p) = \{\nu_\lambda : \lambda \in \Delta^{p-1}\}$. The barycentric coding model consists of two problems: *synthesis*, the problem of finding ν_λ for given $\lambda \in \Delta^{p-1}$, and *analysis*, the problem of finding λ for given ν_λ. The content of Theorem 2.2.4 is that the synthesis problem is well-defined, provided certain regularity assumptions hold. In [WJT+22], it is shown under mild regularity assumptions on the family of reference measures that the analysis problem can also be solved.

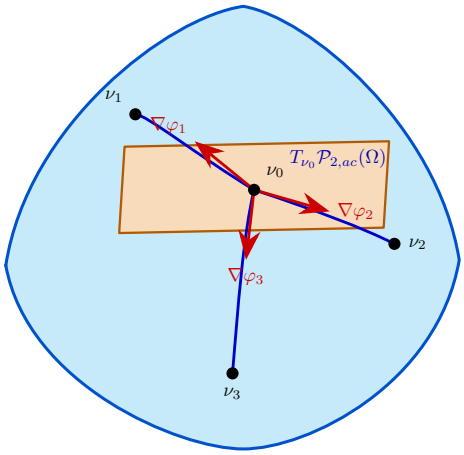


Figure 2.4: To solve the analysis problem, we compute the tangent vectors associated to the unit time geodesic curves connecting the target measure to each reference measure, then form the associated Gram matrix.

Theorem 2.2.5 ([WJT+22], Proposition 1)**.** *Let* $\{\nu_i\}_{i=1}^p$ *be a collection of suitably regular probability measures on* Ω. *Then* $\nu_0 \in \mathrm{Bary}(\{\nu_i\}_{i=1}^p)$ *if and only if*

$$\min_{\lambda \in \Delta^{p-1}} \lambda^T A \lambda = 0, \tag{2.9}$$

where, letting T_i be the optimal transport map from μ_0 to μ_i, $A \in \mathbb{R}^{p \times p}$ is defined by

$$A_{ij} = \langle T_i - \mathrm{Id}, T_j - \mathrm{Id} \rangle_{L^2(\nu_0)}. \tag{2.10}$$

Furthermore, if the minimum value of $\lambda^T A \lambda$ is 0, and λ^* is an optimal argument, then $\nu_0 = \nu_{\lambda^*}$.

Thus, if we can approximate the matrix A, and solve the quadratic program (2.9), we can solve the analysis problem for the barycentric coding model; the geometry of this result is illustrated in Figure 2.4. We note that this assumes perfect access to the probability measures under consideration; estimation of transport maps via i.i.d. samples from the underlying measures enables estimation of the optimal λ in this statistical setting.

2.3 Computational Optimal Transport

2.3.1 Special Case: Transport on the Real Line

We now address a special case with particular relevance to the work will be described in Chapter 3. In the case where measures are supported on the real line and absolutely continuous with respect to the Lebesgue measure, the following convenient identity holds:

$$W_2^2(\mu, \nu) = \int_0^1 |F_\mu^{-1}(x) - F_\nu^{-1}(x)|^2 \, \mathrm{d}x, \tag{2.11}$$

where F_μ^{-1} is the generalized inverse cumulative distribution function, defined by

$$F_\mu^{-1}(x) = \inf\{t : \mu([-\infty, t)) \geq x\}.$$

The reason this formula is so convenient when working with one-dimensional empirical measures is that it reduces the problem of computing an optimal transport cost to the

problem of sorting two lists, which is a much easier problem than solving a general linear program. Following [PC19], for two empirical measures $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and $\hat{\nu} = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$, which we assume to be sorted so that $x_i \leq x_{i+1}$, $y_i \leq y_{i+1}$ for each i, we have that

$$W_2^2(\hat{\mu}, \hat{\nu}) = \sum_{i=1}^{n} |x_i - y_i|^2.$$

The case when the measures have a different number of atoms is not meaningfully more difficult using (2.11), since the pseudo-inverse of an empirical CDF is easily obtained by sorting the observed samples. We will also be interested in working with 1D transport costs on the unit circle, details for which can be found in [DSS10]. All transport algorithms used for the change point detection application work outlined in Chapter 3 are those implemented in [FCG+21].

2.3.2 Entropic Regularization of the Optimal Transport Problem and Sinkhorn's Algorithm

Up to this point, reference has been made frequently to regularity of measures — but in applications, typically one works with empirical measures based on some set of observations. In the case of finitely supported measures the Kantorovich problem may be optimized via standard techniques in convex programming. However, linear programs exhibit poor scaling in the size of that support ($\mathcal{O}(n^3 \log n)$) [PW09]. In [Cut13], it was shown that transport distances of empirical measures, i.e., normalized superpositions of Dirac point masses, could be approximated more quickly by appending a regularization term. Let $\hat{\mu} = \sum_{i=1}^{n}\alpha_i\delta_{x_i}, \hat{\nu} = \sum_{i=1}^{m}\beta_i\delta_{y_i}$, be finitely supported measures, fix $\varepsilon_{reg} > 0$, define the pairwise distance matrix $C = [C_{i,j}]_{i=1,j=1}^{n,m}$ via

$$C_{i,j} = \|x_i - y_j\|^2,$$

and for $\gamma \in \Pi(\hat{\mu}, \hat{\nu})$ define the entropy of γ by

$$H(\gamma) := -\sum_{i,j} \gamma_{i,j}(\log \gamma_{i,j} - 1).$$

Then the entropic optimal transport cost for $\hat{\mu}$ and $\hat{\nu}$ is defined via

$$OT_{\varepsilon_{reg}}(\hat{\mu}, \hat{\nu}) := \inf \left\{ \langle \gamma, C \rangle + \varepsilon_{reg} H(\gamma) \right\}. \tag{2.12}$$

In the limit as $\varepsilon_{reg} \to 0^+$, we have $OT_{\varepsilon_{reg}}(\hat{\mu}, \hat{\nu}) \to OT(\hat{\mu}, \hat{\nu})$. The effect of the regularization is demonstrated in Figure 2.5 — the optimal plan in the unregularized case is a pure permutation matrix, assigning mass to move from one specific point to another in a cost-minimizing manner, while the presence of the entropy term in 2.12 results in a diffusive effect in the optimal plan. Computing the entropic optimal transport cost is accomplished via the Sinkhorn algorithm [SK67, Sin67], which converges to an optimal solution of (2.12) in an amount of time which is roughly quadratic in the number of samples [Cut13]. For this reason, entropic regularization has become a popular computational technique in the optimal transport community [GPC18, MNW19, CTV19, PNW21,

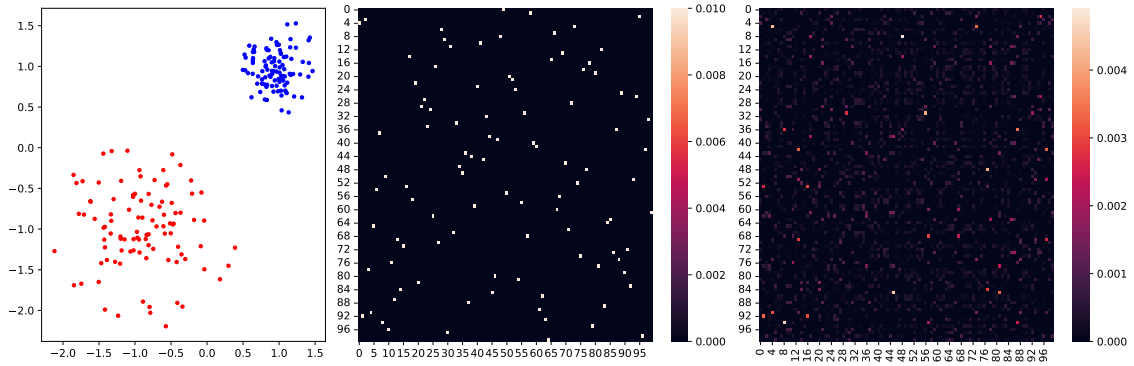


Figure 2.5: Comparison between a classic optimal transport plan and an entropically regularized one. Left: two samples of 100 points each in the plane (normally distributed with different means and variances). Center: optimal assignment between points to minimize transport cost, obtained via computation of a linear program. Right: entropic transport plan for minimizing entropic transport cost. Notice the difference in sparsity between the two plans: entropic transport results in a diffusive effect.

MWMA23, MMA25, Wer24, MMS25].

2.4 Optimal Transport on Discrete Spaces

Definitions and Calculus for Discrete Metric Spaces

We now describe how the theory of transportation differs when the underlying metric space is discrete. A *(finite) discrete metric space* is a pair $(\mathcal{X}, d)$ where $\mathcal{X} = \{1, 2, \ldots, n\}$, with $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ satisfying the classic properties of a metric (symmetry, positive definiteness, and the triangle inequality).

A probability measures on a graph always has a simple structure, namely it is a non-negative function $p : V \to \mathbb{R}^+$ such that $\sum p_i = 1$. It is easy to see that the space of functions on the vertex set of G is identified with $\mathbb{R}^{|V|}$ and hence that the set of probability measures is the $|V|$-simplex.

In [Maa11] a natural question is posed: suppose that Q is an irreducible, reversible, Markov transition kernel on a finite set $\mathcal{X}$, and let $H(t)$ be the heat semigroup on $\mathcal{X}$, defined as $H(t) := e^{t\Delta_X}$, where $\Delta_{\mathcal{X}}$ is the associated graph Laplacian (see 5.3.2 for details on graph Laplacians) — can this heat flow [2] be identified as a gradient flow in the space of probability densities on $\mathcal{X}$?

The answer turns out to be yes, and the functional defining the gradient flow is even the standard entropy, but the geometry defining the gradient flow is *not* the L^2-Kantorovich geometry. To understand why, we need to revisit the static approach to the optimal transprot problem. In the Euclidean setting, the approach to the optimal transport problem based on the Kantorovich formulation and transportation plans, as opposed to transportation maps, arose in part because it was easier to treat analytically. While the problem remains tractable when the underlying domain of the measures is discrete, certain technical issues arise that make it unsuitable as a metric on the probability

[2] The connection between this semigroup and heatflow in Euclidean space will be clarified below when we discuss discrete calculus.

simplex. In particular, one can show that if $\mathcal{X}_2 = (\{0,1\},\{(0,1)\})$ is the simple two point graph, with two vertices connected by an undirected edge, and W_2 is the classic L^2-Kantorovich distance, then as metric spaces $(\mathcal{P}(\mathcal{X}_2), W_2) \simeq ([0,1], \sqrt{|x-y|})$. This is problematic, since the space on the RHS admits no non-trivial geodesics; in particular, every geodesic curve is 2-Hölder continuous, and hence constant. Thus there can be no theory of gradient flows or of variance minimizing measures [Maa11]. To see this, first note that the set $\mathcal{P}(\mathcal{X}_2)$ is easily characterized: $\nu \in \mathcal{P}(\mathcal{X}_2)$ if and only if it can be written $\nu = \nu_\alpha = (\alpha, 1-\alpha), \alpha \in [0,1]$. Now, because γ must be countably additive, if we fix two measures $\nu_\alpha, \nu_\beta \in \mathcal{P}(\mathcal{X}_2)$, we can similarly characterize the set of couplings $\Pi(\nu_\alpha, \nu_\beta)$ as the set of 2×2 matrices γ with appropriate row and column sums by writing $\gamma(\{(x,y)\}) = \gamma_{xy}$ for $(x,y) \in \mathcal{X}_2 \times \mathcal{X}_2$. This in turn leads us to an underdetermined linear system which any coupling γ must satisfy:

$$\begin{aligned}
\gamma_{00} + \gamma_{01} &= \alpha \\
\gamma_{10} + \gamma_{11} &= 1-\alpha \\
\gamma_{00} + \gamma_{10} &= \beta \\
\gamma_{01} + \gamma_{11} &= 1-\beta
\end{aligned}$$

Solving this system we find that any coupling $\gamma \in \Pi(\nu_\alpha, \nu_\beta)$ has the form

$$\gamma = \begin{bmatrix} \alpha+\beta-1+\gamma_{11} & 1-\beta-\gamma_{11} \\ 1-\alpha-\gamma_{11} & \gamma_{11} \end{bmatrix}, \tag{2.13}$$

noting that the upper bound $\gamma_{11} \leq \min\{1-\alpha, 1-\beta\}$ is obtained as an easy consequence of (2.4). Together with the simple metric structure of $\mathcal{X}_2$, this allows us to directly solve (2.2.2). We have, recalling the identity

$$\min\{a,b\} = \frac{a+b}{2} - \frac{|a-b|}{2},$$

$$\begin{aligned}\inf\left\{\int_{\mathcal{X}_2\times\mathcal{X}_2} d^2(x,y)\,\mathrm{d}\gamma(x,y) : \Pi(\nu_\alpha,\nu_\beta)\right\} &= \inf\{\gamma_{01}+\gamma_{10} : \gamma\in\Pi(\nu_\alpha,\nu_\beta)\}\\ &= \inf\{1-\alpha+1-\beta-2(\gamma_{11})\}\\ &= 1-\alpha+1-\beta-2(\min\{1-\alpha,1-\beta\})\\ &= |\beta-\alpha|,\end{aligned}$$

We conclude then that for $\Omega=\mathcal{X}_2$, we have

$$W_2(\nu_\alpha,\nu_\beta)=\sqrt{|\beta-\alpha|}$$

The final piece of the exercise is to recall a fact about arbitrary geodesic curves from metric geometry. Let (X,d) be a metric space, fix $a,b\in X$, and let $x(t):[0,1]\to X$ be a geodesic with $x(0)=a$, and $x(1)=b$. Then, for any $s,t\in[0,1]$, we have

$$d(x(s),x(t))=|t-s|d(a,b).$$

So now suppose that $\nu(t)$ is a geodesic in the L^2-Kantorovich geometry. Then we have

$$|\nu(t)-\nu(s)|^{1/2}=W_2(\nu(s),\nu(t))=|t-s|W_2(\nu_\alpha,\nu_\beta)=|t-s|\cdot|\beta-\alpha|, \tag{2.14}$$

which directly implies that the mapping $\nu:[0,1]\to[0,1]$ is 2-Hölder continuous, and therefore constant. Since $\nu(t)$ is an arbitrary geodesic, this shows that all geodesics in the L^2-Kantorovich geometry of $\mathcal{X}_2$ are trivial: they are single points.

To remedy this issue, a construction based on the Benamou-Brenier formula was proposed in [Maa11]. The key to the construction is adapting the continuity equation to the setting of a discrete metric space, defining an action functional for curves in the $|V|$-simplex, and declaring the squared distance between two measures ρ_A and ρ_B to be the minimal action of a curve $\rho(t)$ where $\rho(0)=\rho_A$, and $\rho(1)=\rho_B$, where the

minimization occurs over the space of all curves satisfying the adapted continuity equation with prescribed temporal boundary conditions.

The construction in [Maa11] is phrased in terms of Markov kernels on finite sets. Let $\mathcal{X}$ be a finite set, let $Q : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be the transition rate matrix of a continuous time irreducible Markov chain on $\mathcal{X}$, and let π be the (unique) stationary distribution of the process, satisfying $\pi : \mathcal{X} \to (0, 1), \sum_i \pi(i) = 1$. Assume also that Q and π satisfy a detailed balance equation, meaning that $\pi(i)Q(i, j) = \pi(j)Q(j, i)$ for all $i, j \in \mathcal{X}$. The set of probability measures *with respect to* π is denoted

$$\mathcal{P}(\mathcal{X}) := \left\{ \rho : \mathcal{X} \to \mathbb{R}_0^+ : \sum_{i=1}^{|V|} \rho_i \pi_i = 1 \right\}.$$

In order to describe the adaptation of the continuity equation to the discrete settings, we first need to recall the tools of discrete calculus. Let $\mathcal{X} = (V, E)$ and d be the shortest path metric. A function on (the nodes of) $\mathcal{X}$ is identified with a vector $f \in \mathbb{R}^{|V|}$. A vector field on (the edges of) $\mathcal{X}$ is a function $m \in \mathbb{R}^{|E|}$. The spaces $\mathbb{R}^{|V|}$ and $\mathbb{R}^{|E|}$ are equipped with the respective inner products

$$\langle \varphi, \psi \rangle_\pi = \sum_{x \in V} \varphi(x)\psi(x)\pi(x); \quad \langle \Phi, \Psi \rangle_Q = \frac{1}{2} \sum_{x,y \in V} \Phi(x, y)\Psi(x, y)Q(x, y)\pi(x).$$

The graph gradient of a function f is a vector field ∇_f, and is defined by

$$\nabla_{\mathcal{X}} f(x, y) = f(x) - f(y).$$

The graph divergence of a vector field m is a function $\mathrm{div}_{\mathcal{X}}\, m$, defined by

$$\mathrm{div}_{\mathcal{X}}\, m(x) = \frac{1}{2} \sum_{y \in V} (m(y, x) - m(x, y))Q(x, y).$$

The graph Laplacian of a function f is the composition of the graph gradient and graph

divergence, and is denoted $\Delta_{\mathcal{X}} f$. The operator $\Delta_{\mathcal{X}}$ admits the explicit form

$$\Delta_{\mathcal{X}} f(x) = \sum_{y \in \mathcal{X}} Q(x,y)(f(y) - f(x)) = ((Q - D)f)(x),$$

where $D := \operatorname{diag}\left(\sum_{y \in \mathcal{X}} Q(x,y)\right)_{x \in \mathcal{X}}$. The graph Laplacian has been an object of intense study and finds applications in network cluster identification, vertex dimensionality reduction, and Fourier analysis on graphs; for details, we refer the reader to the surveys [Mer95, Mer94, SMD+19b]. Furthermore, the graph Laplacian can be used to define a heat flow: given an initial condition ρ_0, we can consider the differential equation

$$\frac{\partial \rho}{\partial t} = \Delta_{\mathcal{X}} \rho,$$

which admits the formal solution

$$\rho(t) = e^{t\Delta_{\mathcal{X}}} \rho_0.$$

In the context of graph calculus, there is no way to multiply a vector field by a function and obtain another vector field. Therefore, if the continuity equation (2.2) is to be made legible in the context of a discrete metric space, some choice must be made about how to "scale" a vector field. To that end, an *admissible mean* $\theta(\cdot,\cdot) : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ is chosen. For technical reasons, there are certain properties that θ will need to satisfy: it must be concave, lower semi-continuous, 1-homogeneous, symmetric, and C^2 on $(0,\infty)^2$. The choices most often considered in the literature are the logarithmic mean and the geometric mean, respectively given by

$$\theta_{\log}(s,t) = \frac{s-t}{\log(s) - \log(t)}; \quad \theta_{geom}(s,t) = \sqrt{st}.$$

The ultimate result of [Maa11] is that choosing $\theta = \theta_{\log}$ recovers Otto's result connecting

heat flow to the gradient flow of entropy. However, in our implementation, we prefer θ_{geom} simply because it is easier to work with on a practical level. Once θ is chosen, for any measure $\rho \in \mathcal{P}(\mathcal{X})$, define the vector field $\theta(\rho)(i,j) := \theta(\rho(i), \rho(j))$. Then the set of solutions to the *discrete continuity equation* is defined by

$$\mathcal{CE}(\rho_A, \rho_B) := \left\{ (\rho, m) : \frac{\partial \rho}{\partial t} + \mathrm{div}_{\mathcal{X}}(\theta(\rho) \odot m) = 0,\ \rho(0) = \rho_A,\ \rho(1) = \rho_B \right\},$$

where $\odot$ means componentwise multiplication. Then, if $\rho : [0,1] \to \mathbb{R}^{\mathcal{X}}$ and $m : [0,1] \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, the action of the pair (ρ, m) is defined to be

$$\mathcal{A}(\rho, m) := \frac{1}{2} \int_0^1 \sum_{i,j \in \mathcal{X}} \frac{m(t,i,j)^2}{\theta(\rho(t,i), \rho(t,j))} Q(i,j)\pi(i) \mathrm{d}t.$$

Taking $\mathcal{I}_{\mathcal{CE}(\rho_A,\rho_B)}(\rho, m)$ to be 0 for $(\rho, m) \in \mathcal{CE}(\rho_A, \rho_B)$ and $+\infty$ otherwise, the energy of the pair (ρ, m) is defined to be

$$\mathcal{E}(\rho, m) = \mathcal{A}(\rho, m) + \mathcal{I}_{\mathcal{CE}(\rho_A,\rho_B)}(\rho, m),$$

and finally the metric $\mathcal{W}$ is defined by

$$\mathcal{W}(\rho_A, \rho_B) := \inf \left\{ \sqrt{\mathcal{E}(\rho, m)} : (\rho, m) \in \mathcal{CE}(\rho_A, \rho_B) \right\}.$$

For full details of the construction and proof that it does indeed define a proper Riemannian structure on the interior of the simplex, we refer the reader to [Maa11]. Two things which are important to note concerning the discrete transport geometry. Firstly, the tangent space has an interpretation analogous to the one given in the Euclidean setting, which is to say that vectors in the tangent space at a point are in correspondence with the set of graph gradients of functions on the graph,

$$T_\nu \mathcal{P}(\mathcal{X}) = \left\{ \nabla\varphi \in \mathbb{R}^{|V|\times|V|} : \varphi \in \mathbb{R}^{|V|} \right\}.$$

Second, that the metric tensor has an explicit expression in terms of the admissible mean θ — this is key, because it allows us to form the Gram matrix needed for solving the analysis problem. We have that for $\nabla\varphi, \nabla\psi \in T_\nu\mathcal{P}(\mathcal{X})$

$$\langle \nabla\varphi, \nabla\psi \rangle_\nu = \frac{1}{2} \sum_{x,y \in V} \theta(\nu(x), \nu(y)) \nabla\varphi(x,y) \nabla\psi(x,y).$$

Lastly, and perhaps most important from a theoretical perspective, is that as in the case with the W_2 metric for probability measures on $\mathbb{R}^d$, $\mathcal{W}^2$ is convex with respect to linear interpolation.

Theorem 2.4.1 ([EM12], Proposition 2.11)**.** *For $\tau \in [0,1]$ and $i, j = 0, 1$, take $\rho_i^j \in \mathcal{P}(\mathcal{X})$ and set $\rho_i^\tau := (1-\tau)\rho_i^0 + \tau\rho_i^1$. Then*

$$\mathcal{W}^2(\rho_0^\tau, \rho_1^\tau) \leq (1-\tau)\mathcal{W}^2(\rho_0^0, \rho_1^0) + \tau\mathcal{W}^2(\rho_0^1, \rho_1^1).$$

Corollary 2.4.2. *Fix $\{\nu_i\}_{i=1}^p \subset \mathcal{P}(\mathcal{X})$, let $\lambda \in \Delta^{p-1}$, and define the variance functional $\mathcal{J} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ via*

$$\mathcal{J}[\nu] = \sum_{i=1}^p \frac{\lambda_i}{2} \mathcal{W}^2(\nu, \nu_i). \tag{2.15}$$

Then there exists $\nu^* \in \mathcal{P}(\mathcal{X})$ which minimizes $\mathcal{J}$.

Proof. Convexity of the functional $\mathcal{J}$ implies that is continuous in the Euclidean geometry and since $\mathcal{J}$ is defined over a compact domain, there therefore exists $\nu^* \in \mathcal{P}(\mathcal{X}) \cong \Delta^{|V|-1}$ minimizing $\mathcal{J}$ by the extreme value theorem. □

2.5 Intrinsic Gradient Descent

In order to compute minimizers to the variance functional, we take a perspective on the problem originating in Riemannian optimization. One of the key tools we will need for this approach is the *exponential map*, a classical object in Riemannian geometry [Tu17].

Let $(\mathcal{M}, g)$ be a Riemannian manifold, let $p \in \mathcal{M}$, and let $T_p\mathcal{M}$ be the tangent space to $\mathcal{M}$. The exponential map at p is a function

$$\exp_p : \mathcal{U} \subset T_p\mathcal{M} \to \mathcal{M},$$

which maps tangent vectors at p to points in $\mathcal{M}$ based on the geodesics emanating from p. If $\gamma_{p,v} : [0, 1] \to \mathcal{M}$ is a geodesic curve such that $\gamma(0) = p, \gamma'(0) = v$, then the exponential map is defined by the relation

$$\exp_p(v) = \gamma_{p,v}(1).$$

Or in physical terms, it is the point in $\mathcal{M}$ reached after unit time by departing from p in the direction v. Typically the exponential map is difficult to work with, because it is defined in terms of an ordinary differential equation and the Levi-Civita connection associated to the metric g. Nevertheless, the exponential map is the key ingredient for what we shall refer to as *intrinsic gradient descent*.

Now the classical gradient descent scheme in the Euclidean setting is defined by the following relation: if F is a smooth function and we wish to find a (perhaps local) minimizer of F, we compute the iterates

$$x_{k+1} = x_k - t_k \nabla F(x_k),$$

for some positive sequence $\{t_k\}$. Lacking a linear structure, this equation becomes meaningless on a Riemannian manifold, but the essential idea of walking "in the direction of steepest descent" remains legible, provided the exponential map is well-defined. In that case, we write

$$x_{k+1} = \exp_{x_k} \left(-t_k \nabla_{\mathcal{M}} F|_{x_k} \right),$$

where $\nabla_{\mathcal{M}}$ is the Riemannian gradient, i.e., the unique vector in $T_{x_k}\mathcal{M}$ which is dual to the differential $\mathrm{d}F$. The intrinsic gradient of the functional (2.15) is actually straightforward

enough to compute, if the point at which the gradient is to be evaluated can be joined to each of the reference points by a unique length-minimizing geodesic and if one is willing to leave the computation in terms of the exponential mapping [Afs11] (in fact, a similar result holds in the classical L^2-Kantorovich geometry [AGS05]). One has, for fixed $\{x_i\}_{i=1}^p \subset \mathcal{M}$, with $\lambda \in \Delta^{p-1}$

$$\nabla_x \left\{ \sum_{i=1}^{p} \frac{\lambda_i}{2} d^2(x, x_i) \right\} = -\sum_{i=1}^{p} \lambda_i \exp_x^{-1}(x_i),$$

and therefore, what we will need to compute is

$$\nu_{k+1} = \exp_{\nu_k} \left(-t_k \left(\sum_{i=1}^{p} \lambda_i \exp_{\nu_k}^{-1}(\mu_i) \right) \right).$$

Chapter 3

Wasserstein-Based Identification of Metastable States in Time Series Data via Change Point Detection and Segment Clustering

"But I know a change is gonna come."

— Sam Cooke

Chapter Overview

Change point detection for time series analysis is a difficult and important problem in applied statistics, for which a variety of approaches have been developed in the past several decades. Here[1], the Wasserstein metric is employed as a tool for change-point identification in multi-dimensional time series data in order to identify clusters in time series in an unsupervised way. We leverage the simplicity of the optimal transport cost in the 1-dimensional setting to quickly identify both a segmentation (family of change points for a trajectory) and a clustering for the data when the number of segments is much smaller than the number of data points, making no parametric assumptions about the particular distributions involved. Our change point detection method scales linearly in the size of the data and in the dimension of the samples. We test our approach on idealized synthetic data trajectories, as well as real world trajectories coming from the domain of molecular dynamics simulations and underwater acoustics. We find that segmenting these time series via change points obtained by estimating the Wasserstein metric derivative and then clustering the identified segments as measures with similarity measured by the Wasserstein metric, successfully identifies metastable states in the law of the processes.

3.1 Introduction

3.1.1 Motivation

Change point detection (CPD) is a powerful tool for the study of time series data because although a given trajectory may be large in the sense that the number of observations is large, containing potentially millions of samples, it may also be redundant if the time

[1]This chapter is based on a paper co-authored with Joshua Huang and James M. Murphy, [GHM26].

series tends to oscillate between few metastable states wherein the variance between observations is low. Therefore, if both the number of metastable states and the rate at which the trajectory passes between them is relatively small, significant reduction in the number of pairwise comparisons needed to apply a distance based clustering technique can be achieved by viewing it as a sequence of empirical distributions representing segments of the trajectory. How one searches for these change points is therefore a key problem for this approach.

The motivating problem for this chapter arises from studying the trajectory data of molecular dynamics simulations (MDS). A common assumption when studying these trajectories is that the configuration space of a given particle in a fixed setting can be partitioned into regions which represent particle states which are "metastable" and transitions between them; the metastable regions of the space are those where the particle tends to remain for extended periods of time, e.g., the basins of an energy well, while the transitions are the complement of the metastable sets. A natural question follows from this assumption, namely, how can we construct reduced-order models of a particle's behavior based on its trajectory data? One approach to the problem is *segment-based clustering*, a technique explored in [DML21]. By looking for change points in the time series data, a very large number of data points is reduced to a relatively small number of segments, leveraging the time sequence aspect of the data to reduce the number of pairwise distances needed to compute a similarity matrix, which is then used to identify a clustering of those segments via the clustering by density peaks algorithm [RL14, WQH$^+$24]. The individual data points then inherit the label of their parent segment.

One of the CPD methods which originally inspired this work is BarT [DLM23], which is a transport-based modification of the SIMPLE algorithm [FDM$^+$15], both of which depend on maximum likelihood estimation [Was04] and consequently must have a statistical prior about the distributions [FDM$^+$15] [DLM23]. Both of these are methods that take an optimization-oriented view of the CPD problem, where a loss function is

established which is based on the log-likelihood of observing the trajectory for a given set of change points, together with a regularization term which modulates how many change points the algorithms produce by penalizing each change point, with a lower penalty for simultaneously detected changes in several observables. These optimization procedures are extremely expensive, meaning that finding the optimal penalizing parameter for the regularization term is impractical. By changing our statistical view from a parametric to a non-parametric one, we give up theoretical guarantees about optimality of our change points, but gain a method which can analyze a trajectory much more quickly.

Optimal transport provides a toolbox which is particularly well-suited to data which has a spatial aspect. In particular, it allows us to quantify the difference between two probability measures in a way which encodes the underlying geometry of the space on which the measures are supported. This is important if we are trying to discern the points in time at which the trajectory leaves or enters distinct regions of a configuration space for extended periods via observed data. Note also that this is an advantage that transport has over other commonly employed ways of comparing distributions when the supports of the distributions differ — for example, the Kullback-Liebler divergence is not even defined for two measures with disjoint support [Was04]. We apply this transport toolbox to the CPD problem by viewing a segmentation of a time series as a sequence of empirical probability measures. Using a moving window, we scan along the time series, measuring the distance between neighboring segments using the optimal transport cost, then identify change points as those times where the distance between the empirical measures representing two neighboring segments is unusually large.

We conceive of our time series as being a realization of some stochastic process with hidden law. If the law of our stochastic process evolves in a continuous way, then we can consider it to be a parameterized curve in the space of measures, i.e., a particle trajectory in something resembling a manifold. If this particle trajectory is concentrated in a small region of the manifold for some period of time, then the stochastic process

is in a metastable state. If the particle is moving between metastable states it is in a transition state. After segmenting the trajectory by identifying change points and transitions between metastable states we will produce a reduced-order description of the trajectory by grouping together empirically observed metastable segments, using classical unsupervised clustering methods.

Following our analogy to particle trajectories in a manifold, our heuristic for metastable state detection will be the following: the particle is not moving (and hence metastable) if the speed of the trajectory is relatively small. Our approach is data-driven, and in general we cannot hope to have access to, e.g., an analytic stochastic differential equation governing the trajectory, so we must approximate the differential of the curve based on available data.

To this end, we will apply the 2-Wasserstein metric as a non-parametric statistical tool for the analysis of noisy time series data. Formally, our hypothesis is this: suppose we are observing a stochastic process X_t on a compact interval $[0, T]$, and that there is some time $t_0 \in (0, T)$ where a change in the law of the process occurs. Then, if $\hat{\mu}_{t_0^-}$ is the empirical distribution of the process generated by the samples $\{X_t\}_{t \leq t_0}$ and $\hat{\mu}_{t_0^+}$ is the empirical distribution generated by the samples $\{X_t\}_{t_0 \geq t}$, i.e.,

$$\hat{\mu}_{t_0^-} = \sum_{t=0}^{t_0} \frac{1}{t_0} \delta_{X_t}, \quad \hat{\mu}_{t_0^+} = \sum_{t=t_0+1}^{T} \frac{1}{T - t_0} \delta_{X_t},$$

then we expect that t_0 should solve

$$\operatorname*{argmax}_{0 \leq t \leq T} W_2^2(\hat{\mu}_{t^-}, \hat{\mu}_{t^+}).$$

Naturally, we do not expect to see only one change point in a given trajectory — rather than looking for a global maximizer of the Wasserstein distance, we will examine local maximizers. To find such maximizers, we employ a sliding window approach, fixing a window size w, and compute for each $t = w, \ldots, T - w$ the Wasserstein distance between

the empirical measures defined by the previous and the next w samples. Then, a cut-off is applied to identify outliers where the distance between consecutive windows is unusually large.

Once we have identified a family of change points, we have determined a segmentation of pairwise disjoint contiguous subsets of our trajectory. We would now like to classify the segments in an unsupervised manner. To do this, we once again interpret the segments as empirical measures, whose pairwise similarity can be measured via the 2-Wasserstein metric. The clustering method we employ is clustering by advanced density peaks (ADPC) [RL14, WQH$^+$24]. ADPC is advantageous because it is entirely unsupervised, determining a number of cluster centers automatically via a statistical criterion.

3.1.2 Contributions

In this chapter, we present a non-parametric change-point detection algorithm, and combine it with an unsupervised clustering approach to identify metastable states in time series data. We find that in idealized synthetic data sets, molecular dynamics simulation data, and underwater acoustics data that our approach is able to identify coherent clusters in the space of probability distributions, thereby providing a compact model of the data. Our CPD approach scales linearly in the size of the data, and is thus well-adapted to large-scale data sets, and, provided that the data does not oscillate between states too rapidly, has the advantage of greatly reducing the number of pairwise distances that need to be computed for the cluster identification.

3.1.3 Organization of the chapter

In Section 2, we outline the necessary background to understand both the change point detection method and the clustering algorithms employed. In Section 3, we describe the algorithm we used to obtain our numerical results. In Section 4, we present the numerical results of our experiments, demonstrating efficacy of our approach in a variety

easily understood low-dimensional examples. In Section 5, we conclude with a discussion of our results, and propose some future directions for expanding and extending the work presented herein.

3.2 Background

3.2.1 Previous Work on Moving Window Change Point Detection and Relation to the Literature

CPD encompasses a vast collection of methods from the field of applied statistics and for a comprehensive overview of the field, we refer the reader to the surveys [AC17, RCW$^+$07, NHZ16, TOV20] as well as the textbook [BN$^+$93]. Generically speaking, CPD methods have several qualities which are used to categorize them, which we outline below.

CPD methods may be online or offline. Online methods are those which run concurrently with the process about which they are trying to make inferences, whereas offline methods take an after-the-fact approach to the problem of identifying changes in the trajectory, examining all the data at once [Pag54, Pic85]. Celebrated online approaches include [AM07, JSB$^+$05, KCHP01], and have seen a rise in popularity as hardware improvements make real-time change detection more feasible. Our method is an offline approach, because we use statistical quantities based on the entire trajectory, retrospectively segmenting the time series and then looking for patterns in the data.

Another distinction to make when differentiating CPD methods is whether or not they are driven by some parametric assumption on the underlying data. Parametric approaches to the CPD problem are among some of the earliest, dating back to the Wald's sequential probability ratio test [Wal45]. Parametric and non-parametric approaches offer distinct advantages and trade-offs; the former work by making an explicit assumption on the underlying nature of the distribution (e.g., a process is stationary and Gaussian), while models in the latter category make no such assumptions.

Parametric models may incur higher computational costs during the process of optimizing model selection. For example, SIMPLE [FDM+15], one of the algorithms that served as initial inspiration for our work, is a parametric CPD scheme which optimizes a loss function which must be derived on some statistical prior, e.g. the data is drawn from a family of distributions which is Gaussian, Laplace, exponential, and so on. Non-parametric approaches by contrast avoid model selection issues, and but tend to require access to larger amounts of data at inference time [AC17]. Tools previously used in non-parametric CPD include but are not limited to the Kolmogorov-Smirnov test [Pic85, PYWR21, Pat07], graphs [CZ15, CC23], and, of particular interest to us, optimal transport, where approaches based on distribution-free methods [Che23, CMHA20, FCS+21], kernel-based methods [WGX22], and entropic regularization of the Wasserstein distance have been studied [WMMA24]. Our approach is similar to the one that appears in [CAH+20, Che23], with some key differences — in particular, our use of the metric derivative rather than the Wasserstein two-sample test. The Wasserstein two-sample test [dCMR99, RTC17] is a distribution free non-parametric test which is applicable to 1D distributions on the real line and measures whether or not the monotone rearrangement of one distribution into another differs from the identify function in a statistically significant way; if there is no such significant difference, the null hypothesis (both samples are drawn from the same distribution) is accepted. This differs our approach, which we will describe below, in that we work directly with the Wasserstein distances between neighboring distributions and look for outliers in that data. Additionally, we are interested in data sets which take values in a non-Euclidean domains, such as the 3-torus. Recently Wasserstein two-sample tests have been studied on the flat torus [GGCN23], and this may provide an avenue for future extension of our work.

As a last categorical distinction among approaches, CPD methods are either supervised or unsupervised. The regime of the former is one in which change points are labeled ahead of time, and the goal of training is to develop a model which predicts well labels

excluded from the training data. On the other hand, unsupervised CPD aims to identify structures in the data without knowledge of a ground truth — this is the regime under which our method operates. One common iteration of the unsupervised approach is to cluster the data into a family of states, and look for the points where the cluster label changes. Classic examples of this approach include the sliding window and bottom-up (SWAB) algorithm [KCHP01] and the minimum description length (MDL) algorithm [RKLE11]. Our approach reverses this paradigm, and attempts to identify suitable labels for the data based on the identified change points.

One challenge of moving window schemes, and which is acute to our approach in particular, is the choice of the window size w. Being an essentially statistical estimator of the instantaneous distribution, it is naturally preferable for our windows to have a large number of samples, and thus be better estimators of the true distribution. On the other hand, as a practical matter, it isn't hard to take w too large — supposing for a moment that there were some set of true change points, if the average (true) segment length is much smaller than the window size, then a typical window will contain many segments, and thus our approach will have no hope of detecting the change points in the large w regime. Therefore, our algorithm leaves w as a tunable parameter, which the user can adjust as fits their use case.

The primary predecessors which inspired our work are the SIMPLE and BarT algorithms, both of which were developed for specific application to molecular dynamics trajectory data [DLM23], with the latter being an extension of the former based on optimal transport and the Wasserstein distance. SIMPLE works for a process $\{\{Y_{j,t}\}_{t=1}^{T}\}_{j=1}^{J}$in J dimensions by first choosing a log-likelihood function $\hat{l}$ representing the plausibility of observing the given data conditioning on a set of change points $\{\tau\}_{i=1}^{K}$, and proceeds by minimizing by means of a dynamic programming algorithm

$$\sum_{j=1}^{J}\sum_{i=0}^{K}\hat{l}(Y_{j,\tau_{j,i}+1},\ldots,Y_{j,\tau_{j,i+1}}) - \lambda\sum_{i=1}^{K}q(S_i)$$

where S_i is a subset of $\{1, \ldots, J\}$ corresponding to the dimensions which are changing at time τ_i, and $q : 2^{\{1,\ldots,J\}} \to \mathbb{R}$ is a strictly increasing penalty function satisfying $q(A \cup B) \leq q(A) + q(B)$ whenever A and B are non-empty disjoint subsets, e.g. in the original implementation of SIMPLE, the authors take $q(\cdot) = |\cdot|^\alpha$ for $\alpha \in (0, 1]$. This encourages the algorithm to choose change points where many dimensions are changing simultaneously — any change point whatsoever incurs a penalty, but points which exhibit changes only in a few dimensions of the data set are penalized more heavily.

BarT takes a point of view based on optimal transport (see Section 2.2 for details on optimal transport), adapting the loss function $\hat{l}$ to consider the possibility that a point belongs to a barycentric transition between metastable states. This adaptation is made by taking $\hat{l}(\cdot) = \max\{L_{ms}(\cdot), L_{tr}(\cdot)\}$, where L_{ms} and L_{tr} are, respectively, the log-likelihood functions corresponding to a metastable and a transition region. Explicitly, the algorithm supposes that metastable states are Laplace distributed, and takes (μ_j, σ_j) to be the parameters of the j-th segment $\{X_t\}_{t=\tau_j}^{\tau_{j+1}}$. Then one has

$$L_{ms}(\{X_t\}_{t=\tau_j}^{\tau_{j+1}}) = \log \prod_{t=\tau_j}^{\tau_{j+1}} \frac{1}{2\sigma_j} \exp\left(-\frac{|X_t - \mu_j|}{\sigma_j}\right),$$

$$L_{tr}(\{X_t\}_{t=\tau_j}^{\tau_{j+1}}) = \log \prod_{t=\tau_j}^{\tau_{j+1}} \frac{1}{2\sigma_{j,t}} \exp\left(-\frac{|X_t - \mu_{j,t}|}{\sigma_{j,t}}\right),$$

wherein $(\mu_{j,t}, \sigma_{j,t})$ correspond to the parameters of the Wasserstein geodesic between the $(j-1)$-th and $(j+1)$-th segments. These parameters are easily computed because the Wasserstein geodesic of two Laplace distributions γ_1 and γ_2 is itself Laplace, and the parameters of the (unit time length) interpolant are directly estimable from the estimated parameters of γ_1 and γ_2 [DLM23]. Our work is inspired by this transport-driven approach, but differs essentially in that it forsakes an explicit loss function, which is expensive to optimize, in favor of working with the geometric structure that the Wasserstein metric gives to the set of empirical distributions.

The problem of change point detection and segment clustering for time series data was also considered in [CAH$^+$20], although our work takes a somewhat different approach to the problem for 1D timeseries, and further expands the approach to multi-dimensional non-Euclidean timeseries data, both those taking on vector values and those satisfying a manifold hypothesis e.g., we will apply our methods to molecular dynamics simulation data which can be represented as a timeseries on the torus.

The Wasserstein distance metrizes the space of probability distributions in a way which encodes the geometry of the underlying space, making it an apt choice for the analysis of physical trajectories, and is amenable to working with Dirac clouds as estimates of the relevant measures. This allows us to quantify the difference between two neighboring windows of points and approximate the metric derivative of the trajectory in Wasserstein space. This information allows us to predict whether or not the law of the process is metastable or in the midst of transition between metastable states.

At a high level, what is essential to our work is the following. Let $\{X_t\}_{t=0}^T$ be a stochastic process defined on some metric space Ω; in practice we consider our data as being a realization of the process [Øks10], and from the given data we will try to draw inferences about the particular realization. For each $t \in [0, T]$, we can denote by γ_t the distribution of the random variable X_t. We will assume that this parameterized family of measures lives in $\mathcal{P}_2(\Omega)$—the set of probability measures on Ω with finite second moment —then we can consider $\mathcal{P}_2(\Omega)$ as an infinite-dimensional Riemannian manifold, at least in a formal sense by interpreting the Wasserstein metric as coming from a Riemannian structure. If the law of our stochastic process X_t evolves in a continuous way, then we can consider the parameterized family of measures γ_t to be a curve, i.e. a particle trajectory, in this manifold. If this particle trajectory is concentrated in a small region of the manifold for some period of time, then the stochastic process X_t is in a *metastable state*. If the particle is moving between metastable states it is in a *transition state*. *A priori*, we generally do not know what the metastable states of a process will be — our goal is

thus to leverage the metric structure of $\mathcal{P}_2(\Omega)$, given by the 2-Wasserstein distance, to identify stretches of time (hereafter, "segments") when the particle is metastable. Then, using classical clustering methods, we will produce a reduced-order description of the trajectory by grouping together empirically observed metastable segments.

3.2.2 Clustering Techniques Based on Similarity Matrices

Once we have split a trajectory into a sequence of segments, our next task is to group them together in a coherent way. In other words, we want to identify clusters in the reduced data set. Each segment is represented by an empirical distribution, and continuing with the theme established by our CPD method, we will quantify the difference between segments with the 2-Wasserstein metric. Because each element in the Wasserstein space is a measure, there is a restriction on the family of clustering methods which are suited to the task of labeling the segments — for example, while a Wasserstein based K-means approach has been investigated in the literature, theoretical results on its scaling remain in development and empirical work suggests the scaling may be poor [ZCY22]. To deal with this difficulty, we will employ a clustering technique which works just from the pairwise distances of the segment data. Conversely, working with the pairwise distance matrix further highlights one of the reasons why it might be desirable to segment a time series. *In toto*, our data sets may contain millions of individual data points, making the computation of the pairwise distance matrix prohibitive at the level of points — however, if the data can be broken into a number of segments which is much smaller than the number of raw data points, that computational burden is eased.

Letting S be the collection of segments in the time series and $\hat{\nu}_i$ be the empirical measure associated to each segment, we form the similarity matrix via the rule

$$A = \left[\exp\left(-\frac{1}{2\sigma}W_2^2(\hat{\nu}_i, \hat{\nu}_j)\right)\right]_{i,j=1}^{|S|}$$

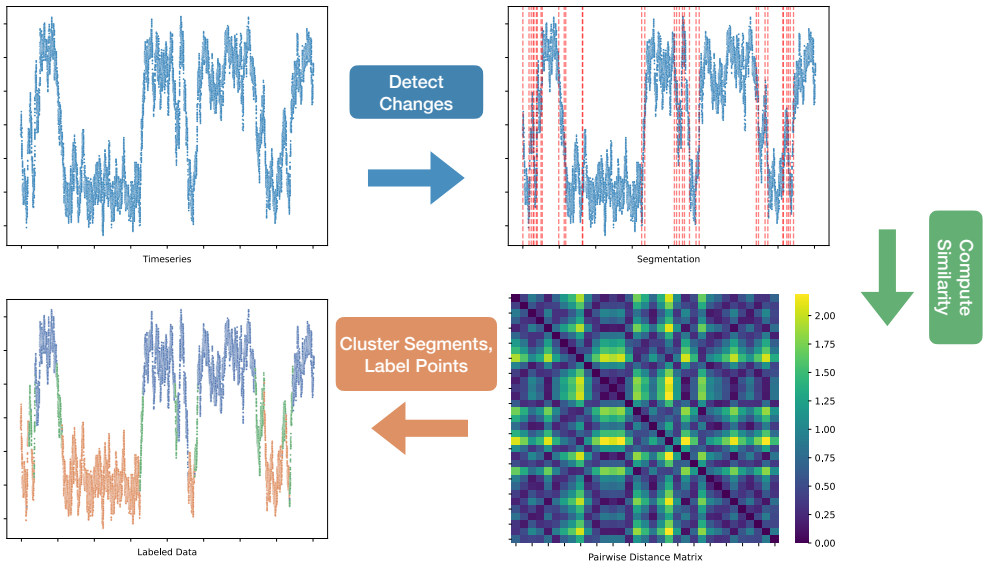


Figure 3.1: Pipeline of the procedure: first, the time series is split into a family of non-overlapping segments, then a pairwise similarity matrix is computed based on the Wasserstein distance, and finally a clustering algorithm is applied to the similarity matrix to learn labels for the segments and by extension the points of the timeseries.

and thus just the formation of the similarity matrix requires $\mathcal{O}(|S|^2)$ operations. In this work, we always take $\sigma = 1$.

Clustering by advanced density peaks is a method that aims to identify the topography of a multimodal dataset [RL14, dFLR21]. This scheme is driven by several heuristics, and has the advantage of being entirely unsupervised. Firstly, “putative centers" of the data set are identified by looking for those points such that the local density of points around the centers is maximal. Once a family of centers is identified, this induces a Voronoi partitioning of the ambient space. The borders of this partition are then examined to identify saddle points in the topography of the data set, which in turn allows for the identification of peaks (modes) in the data. All of these steps are driven via statistical methods, i.e. suitable estimators on quantities such as local density. The two main advantages of using ADPC in our setting is that it is both dimension agnostic, in the sense that it will attempt to estimate the intrinsic dimension of the collection of segments implied by the distance matrix on its own, and that it does not require as an

input a guess on the number of clusters present in the data. This makes our pipeline amenable to complicated trajectories which cannot be "eye-balled", either because the data is essentially high-dimensional or because the modes of the clusters are too close together to be easily differentiated by human judgment.

3.3 Algorithms

Beyond the computational aspects of 1D transport costs as described in chapter 2, our CPD also depends upon two hyperparameters, w and q. The window size w specifies the sample size of the empirical distributions we use at each point in time to approximate the law of the process, and q is a real number between 0 and 1 that specifies a quantile cutoff with respect to the distribution of approximated Wasserstein metric derivatives, indicating that a data point may potentially be labeled as a change point. In essence, w controls the number of samples we demand at any given time to approximate the instantaneous change in distribution of the time series, while q controls how sensitive the algorithm is to the changes in the time series. For example, if q is taken to be small, then more points will register as candidates, and conversely if q is taken close to 1, few candidates will be identified.

Our approach is this: for a fixed data set $\{X_t\}_{t=1}^{T}$ and parameters w, q we approximate the metric Wasserstein derivative via

$$\left|\hat{\dot{\gamma}}\right| = W_2(\mu_t^-, \mu_t^+),$$

where $\mu_t^- = \{X_t\}_{t-w}^{t}$ $\mu_t^+ = \{X_t\}_{t}^{t+w}$. This approximation gives us a sense of the speed at which the law of the process is moving through the space of probability distributions. Then, with that information in hand, we look for the regions where we see outlier acceleration of the distribution.

To give a better sense of what exactly it is that we are looking for, consider the ideal

scenario in a physical sense: suppose that we are looking at a particle trajectory in the plane over a time interval $[0, 1]$. At some time t_0, the particle begins to move uniformly from its initial position to its terminal position, which it reaches at time t_1, after which it ceases to move. The speed of the particle would then be graphically represented by a square wave, i.e, the characteristic function $\mathbf{1}_{[t_0,t_1)}(t)$.

Algorithm 1 Change Point Detection Algorithm (ComputeChangePoints)

Input: Time series data X_t, total number of data points T, window size w, quantile threshold q
Output: Set of change point times

Step 1: *Compute derivative magnitude array*
Initialize $\left|\hat{\dot{\gamma}}\right| \leftarrow$ empty array of size $(T - 2w)$
for $t = w$ **to** $T - w$ **do**
    Compute $|\hat{\dot{\gamma}}|[t] \leftarrow$ estimate of $|\dot{\gamma}(t)|$ from X_t
end for

Step 2: *Identify candidate times based on quantile threshold*
Compute $\tau_q \leftarrow q$-th quantile of $|\hat{\dot{\gamma}}|$
$\mathcal{C} \leftarrow \{t : |\hat{\dot{\gamma}}|[t] \geq \tau_q,\ t = w, w + 1, \ldots, T - w\}$ ▷ Candidate set

Step 3: *Filter for inflection points in contiguous segments*
Partition $\mathcal{C}$ into contiguous subsequences $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$
Initialize $\mathcal{I} \leftarrow \emptyset$ ▷ Inflection point set
for each segment $\mathcal{S}_i$ **do**
    Identify inflection points within segment $\mathcal{S}_i$
    Add detected inflection points to $\mathcal{I}$
end for
return $\mathcal{I}$

Implicit in our overall change point detection scheme is a filtering choice: *a priori*, if we set a quantile cutoff q, then simply returning the points above that threshold will return $(1 - q)\%$ of the data set as change point candidates, and many of these candidate points will be contiguous with each other, as a consequence of the continuity of the trajectory. In other words, we will get out of the first two steps of the scheme a collection of change *regions*, from which we now want to extract a fine-grained selection of change *points*.

Finally, we extend our approach to multi-dimensional data simply by working component-

wise. By this we mean that we treat each component of the time series as being a process all its own, and perform for each dimension the segmentation and clustering procedure. Formally, we can consider our pipeline as a transformation $F = L \circ S$ where S maps the input time series $\{\vec{X}_t\} \subset \mathbb{R}^D$ to a corresponding set of segments $\mathcal{S} = \{S_i\}_{i=1}^k$, and where $L : S_i \mapsto l_i$ maps segments to labels, so that $F : \{\{X_{t,d}\}_{t=1}^T\}_{d=1}^D \mapsto \{\{l_{t,d}\}_{t=1}^T\}_{d=1}^D$. Then an overall state is assigned to $\vec{X}_{t_0}$ by concatenating the labels $(l_{t_0,1} l_{t_0,2} \dots l_{t_0,d})$.

Algorithm 2 Change Point Detection Subroutine: Identifying Inflection Points

Input: Metric derivative data $\{D_t\}_{t=1}^T$, quantile threshold q
Output: Inflection points of metric derivative data
Step 1: *Identify continuous subsequences in* $|\hat{\dot{\gamma}}|$
$\mathcal{C} \leftarrow \{t : D_t > \textbf{quantile}(D, q)\}$
$\mathcal{S} \leftarrow \{\{D_t\}_{t=t_i}^{t_{i+1}} : \{t_i, t_i + 1, \dots, t_i + k\} \subset \mathcal{C}$ is a maximum length subsequence$\}$
Step 2: *Report extrema of the subsequences*
$\mathcal{I} \leftarrow \{\}$
for each subsequence $S \in \mathcal{S}$ **do**
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{\operatorname{argmin} S, \operatorname{argmax} S\}$
end for
return $\mathcal{I}$

Algorithm 3 Metastable State Identification Procedure (IdentifyStates)

Input: time series data $\{\{X_{t,i}\}_{t=1}^T\}_{i=1}^d$, window size $w \in \mathbb{N}$, quantile threshold $q \in (0, 1)$, clustering method Cluster
Output: Point labels $\{l_t\}_{t=1}^T$
for $i = 1 : D$ **do**
  Step 1: *Identify segmentation of time series*
  $\mathcal{C} \leftarrow \text{ComputeChangePoints}(X_t, w, q)$ ▷ Identify change points
  $\mathcal{S} \leftarrow \{\{X_t\}_{t=t_i}^{t_{i+1}} : t_i \in \mathcal{C}\}$ ▷ Partition data into segments
  Step 2: *Compute pairwise* W_2 *similarity of segments*
  $D \leftarrow \left[\exp\left(-\frac{1}{2} W_2^2(S_i, S_j)\right) \text{ for } S_i, S_j \in \mathcal{S}\right]$ ▷ Compute similarity matrix
  Step 3: *Cluster segments based on similarity matrix*
  $L \leftarrow \text{Cluster}(D)$
  Step 4: *Label points according to the label of their parent segment*
  $l_i \leftarrow [L(S(X_{t,i})) \text{ for } t = 1 : T]$
end for
PointLabels $\leftarrow [l_1(t) l_2(t) \dots l_d(t) \text{ for } t = 1 : T]$ ▷ Concatenate labels to make final assignment
return PointLabels

We summarize the above with the following proposition:

Proposition 3.3.1. *For fixed w and q, for inputs $\{\vec{X}_t\}_{t=1}^T \subset \mathbb{R}^D$, the runtime complexity of Algorithm 1 is $\mathcal{O}(TD)$. Fix $w \in \{1, \dots \lfloor T/2 \rfloor\}$, $q \in (0,1)$, and assume that the number of segments N scales proportionally to the length of the input data $X = \{X_t\}_{t=1}^T$.*

Proof. The claim follows directly from the construction of the algorithm: for each $t = w, \dots, T-w$, we perform two sorting operations, which are $\mathcal{O}(w \log w)$, but $\mathcal{O}(1)$ with respect to the size and dimension of the input data. Likewise, the subroutine identifying the inflection points of the identified segments is no more complex than sorting the empirical distribution of each segment. Since the number of segments is assumed to grow linearly in the number of data points, we amortize by considering the expected length of each segment, which will be T/N. Performing $N \propto T$ sorting jobs on a list of size T/N has complexity $\mathcal{O}\left(N \frac{T}{N} \log\left(\frac{T}{N}\right)\right) = \mathcal{O}(T)$. Because this procedure is carried out over each of the D dimensions of the data, the algorithm has total asymptotic time complexity $\mathcal{O}(TD)$. □

3.4 Methodology & Numerical Results

For each of the particular time series we examine going forward, we will apply our method and examine the point labels so produced. To keep things visualizable, we will look at timeseries which takes values in $\mathbb{R}$, and the 3-torus. While almost all the data sets are computer generated, some are more artificial than others — the "toy" data set, for example, was constructed explicitly to serve as a testing ground for our procedure [2]. The molecular dynamics simulations were retrieved from the deeptime Python library [HSH+21] and the GROMACS library for chemical simulations [AMS+15]. Our collection of data sets includes one physically generated trajectory, courtesy of the Naval Research Laboratory [TFS16].

[2] Our code and data are available at https://github.com/dcgentile/mfpy

3.4.1 Generated Data Sets

Toy Data Set

Our first data set comes from a family of artificially constructed trajectories meant to give our procedure an idealized setting in which to work. We fix two means $m_1, m_2 \in \mathbb{R}$ and a variance $\sigma > 0$ to determine two separated Laplace distributions μ_1, μ_2, and we fix a transition length L, a total number of samples T and collection of change points $\{t_i\} \subset \{1 \dots T\}$. The trajectory is then generated by sampling $X_t \sim \mu_1$ for $t = 0, \dots, t_1$, the Wasserstein interpolant of $X_t \sim [\mu_1, \mu_2]_t$ for $t = t_1 + 1, \dots, t_2$, and then $X_t \sim \mu_2$ for $t = t_2 + 1, \dots, t_3$, repeating the pattern until time T.

Toy Laplacian Precision & Recall			
# True CPS	# Detected CPS	Precision	Recall
38	38	0.89	0.89

Table 3.1: Table of precision and recall scores for the toy data set. Precision and recall are averaged over tolerances of $tol = 0, \dots, 100$.

We find that our procedure is able to identify the change points of the trajectory and accurately classify the resulting segments, and we summarize the output in relation to the ground truth data in table 3.1. Because of the high variance in the timeseries, we do not expect our algorithm to output a set of change points which is identical to the ground truth, but rather a set where each candidate change point is close to a ground truth element. For this reason, we average the precision and recall scores over a range of tolerances.

Langevin Trajectory

Next, we investigate the application of our algorithm to a sample Langevin trajectory. The energy landscape governing the particle trajectory has two distinct wells, centered at $+1$ and -1 with a barrier between the two which we can think of as regulating how difficult it is for the particle to transition from one state to the other. The trajectories

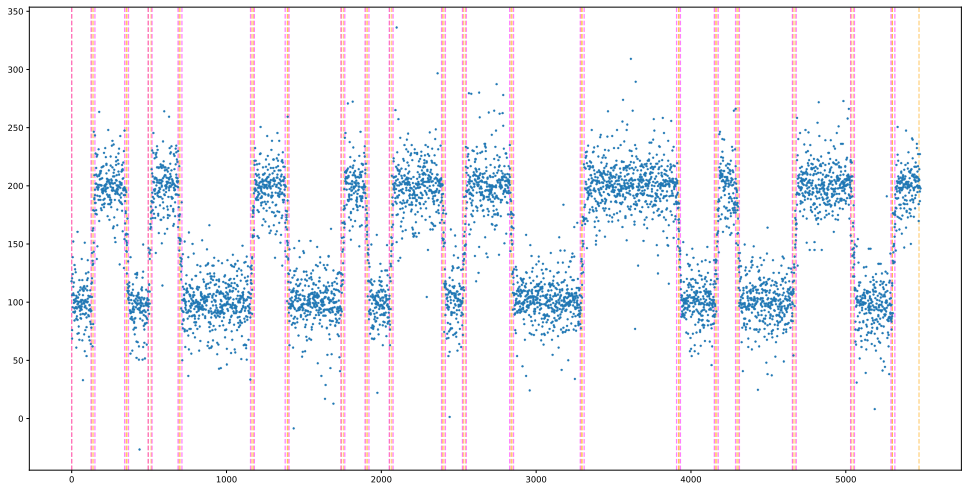


Figure 3.2: Detecting change points in an idealized setting. We generate a synthetic data set by sampling from a fixed Laplace distribution with mean 100 and variance 20, and one with mean 200 and variance 20 in an alternating pattern. We simulate transitions between the two states by sampling from the Wasserstein geodesic (McCann interpolant) between the two states; change points are superimposed on the trajectory as red dashed lines.

were simulated using OpenMM [ESC+17] with particle mass set to 1amu, temperature set to 500K, and damping coefficient of 10ps^{-1}.

We observe that after segmenting the trajectory with our algorithm and clustering via ADPC, three distinct states are identified, with one corresponding to each well, and the third capturing the transition between the two metastable states.

Particle in a Prinz Potential

Increasing the complexity of our example now, we apply our method to a particle trajectory in a Prinz potential. The trajectory is qualitatively similar to the previous Prinz trajectory — there are four wells in the energy landscape that the particle can move between, with the height of the wall between each neighboring pair of wells differing somewhat. The trajectory is generated by an integrator of Euler-Maruyama type, with

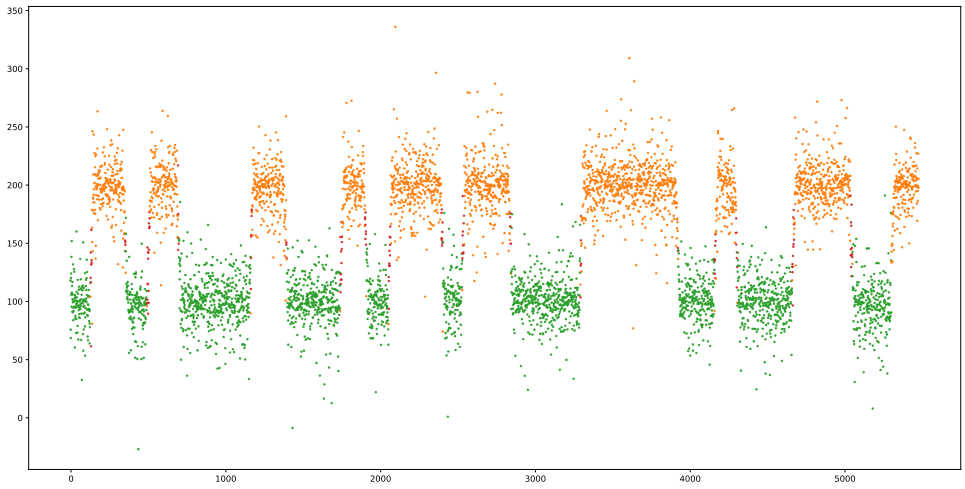


Figure 3.3: Segment cluster labels are learned in an unsupervised manner using the clustering by advanced density peaks algorithm. In this case, the clustering scheme identifies both metastable states and outputs a separate label for the transition regions. Parameters used to obtain this clustering were $w = 25, q = 0.95$.

Figure 3.4: Toy Laplace-based Trajectory Segmentation and Clustering

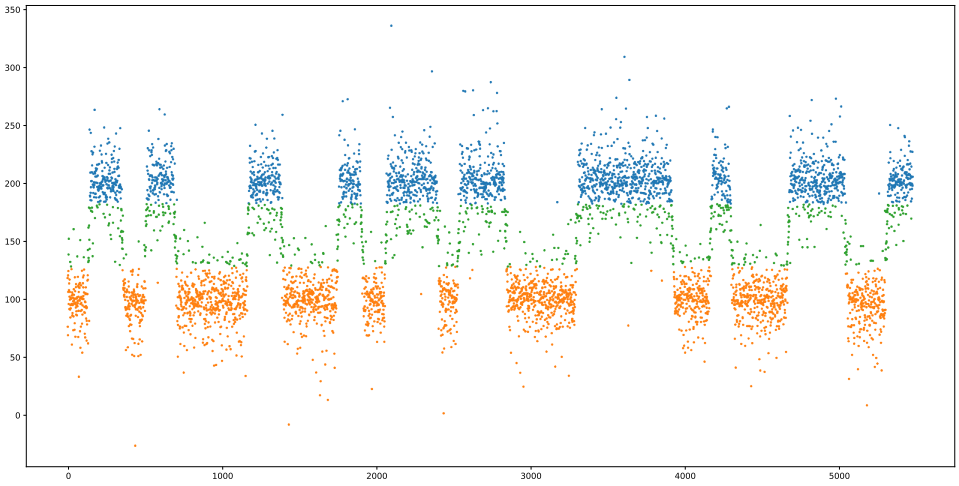


Figure 3.5: To compare our results against standard “off-the-shelf" clustering methods, we cluster our data set with K-means. The resultant clustering fails to reflect the variance of the true underlying distributions and exhibits hard dividing hyperplanes.

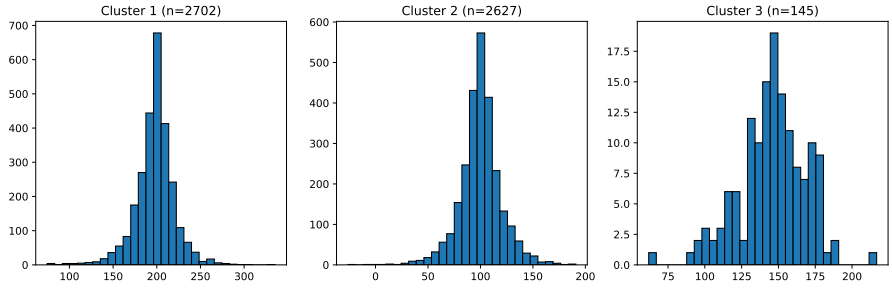


Figure 3.6: Empirical distributions of the clusters in the toy data set learned by segment-based clustering with ADPC.

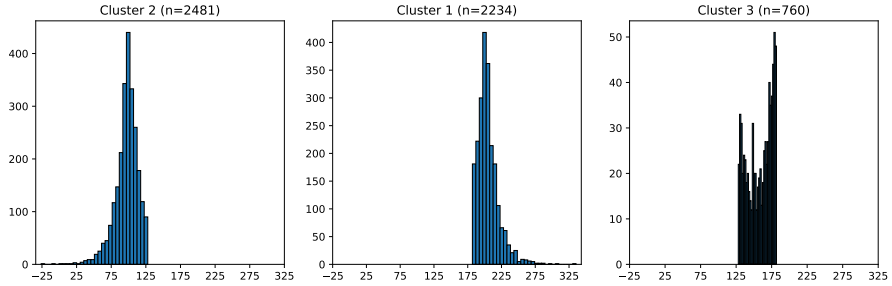


Figure 3.7: Empirical distributions of the clusters in the toy data set learned by K-means.

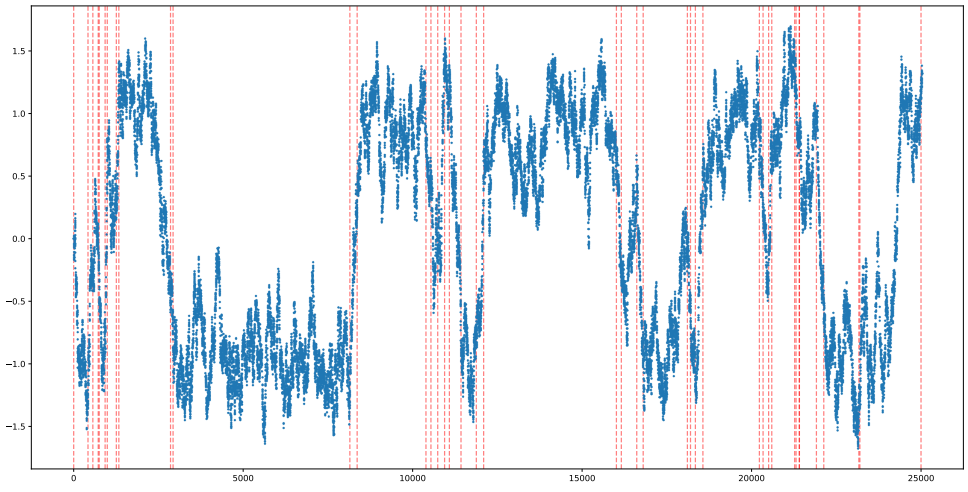


Figure 3.8: The first 25,000 steps in a Langevin trajectory annotated with detected change points.

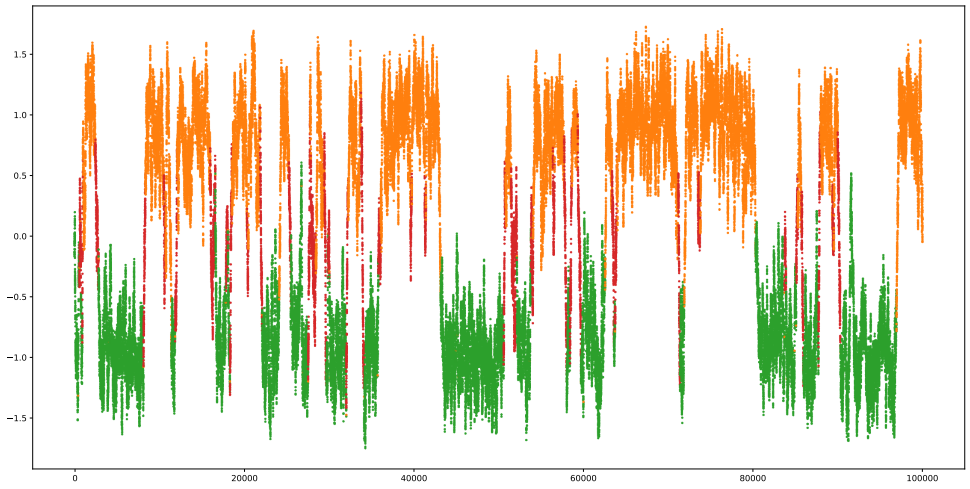


Figure 3.9: Segment based clustering of a Langevin trajectory using advanced density peaks clustering. Parameters used to obtain this clustering are $w = 215, q = 0.86$.

Figure 3.10: Langevin Trajectory Segmentation and Clustering

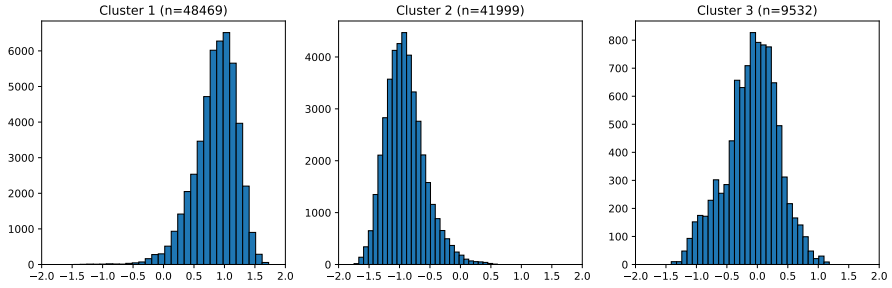


Figure 3.11: Empirical distributions of clusters in the Langevin trajectory data set identified by segment-based clustering with advanced density peaks clustering

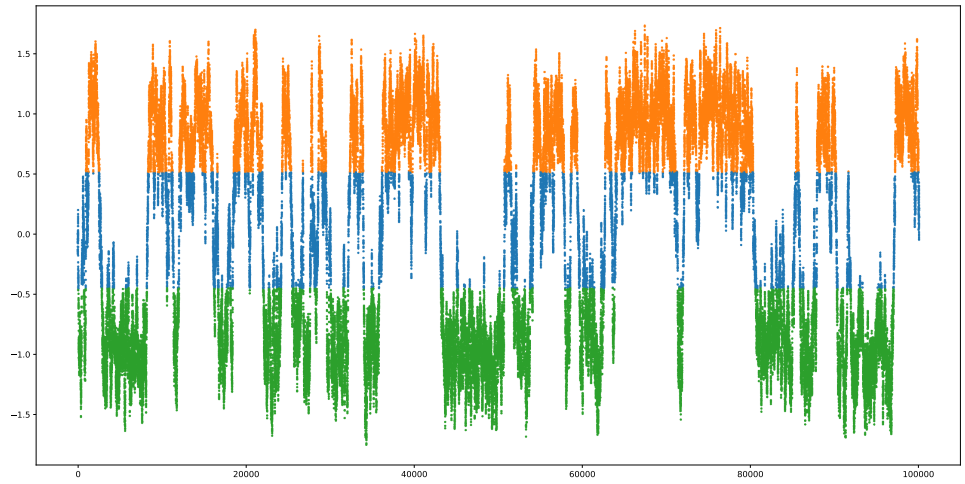


Figure 3.12: To compare our results against standard "off-the-shelf" clustering methods, we cluster our data set with K-means. The resultant clustering fails to reflect the variance of the true underlying distributions and exhibits hard dividing hyperplanes.

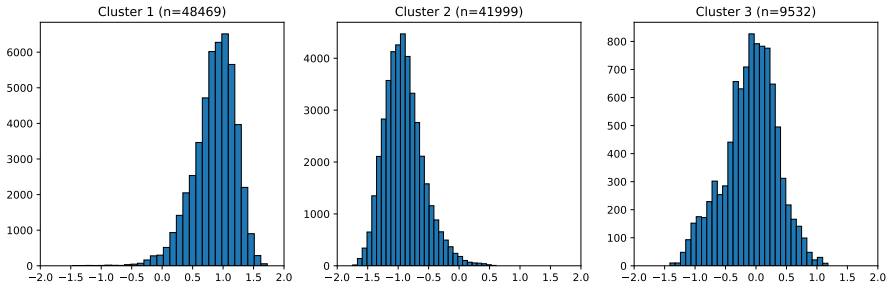


Figure 3.13: Empirical distributions of the clusters in the Langevin data set learned by segment-based clustering with ADPC.

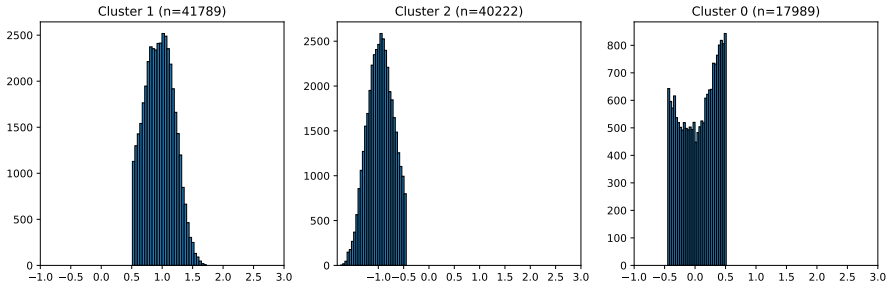


Figure 3.14: Empirical distributions of the clusters in the Langevin data set learned by K-means.

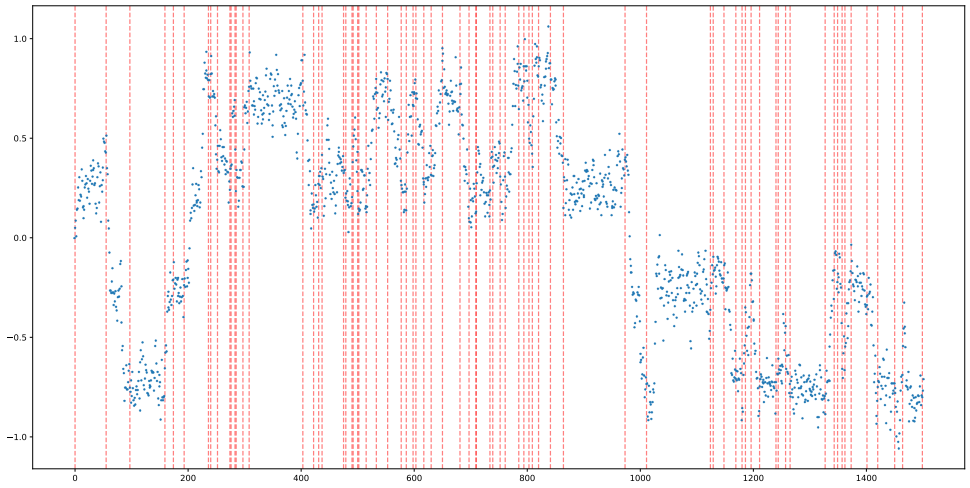


Figure 3.15: Trajectory of a particle in a Prinz potential annotated with detected change points.

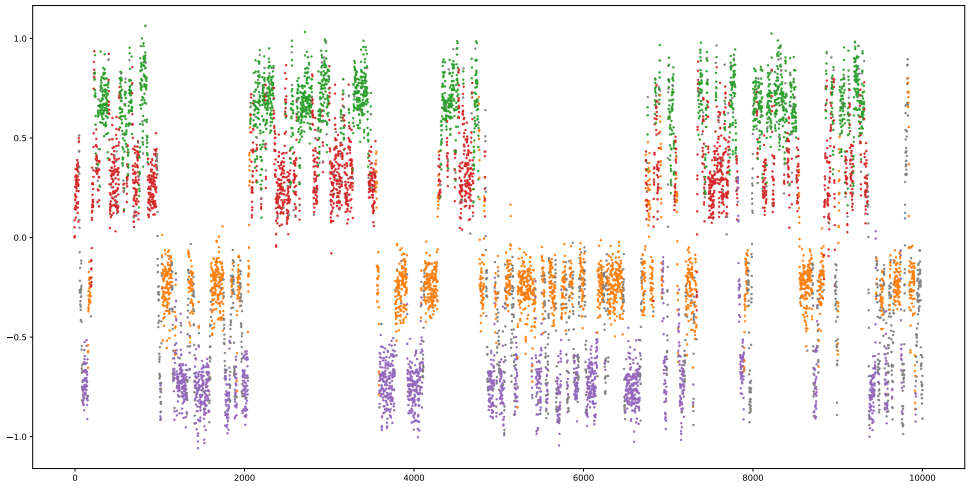


Figure 3.16: Trajectory of a particle in a Prinz potential clustered via segment-based clustering with a advanced density peaks based clustering. The top four clusters by population are colored, while all points outside of these clusters appear in grey. Parameters used to obtain this clustering were $w = 16, q = 0.5$.

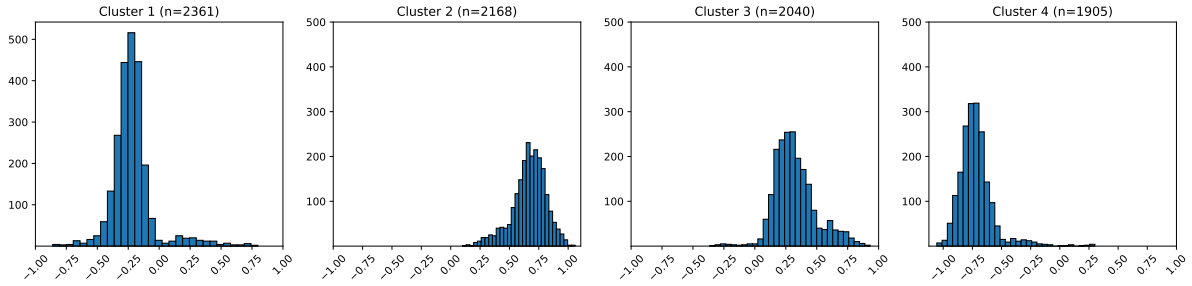


Figure 3.17: Empirical distribution of the top four clusters in the Prinz potential data set identified via segment-based clustering using advanced density peaks based clustering.

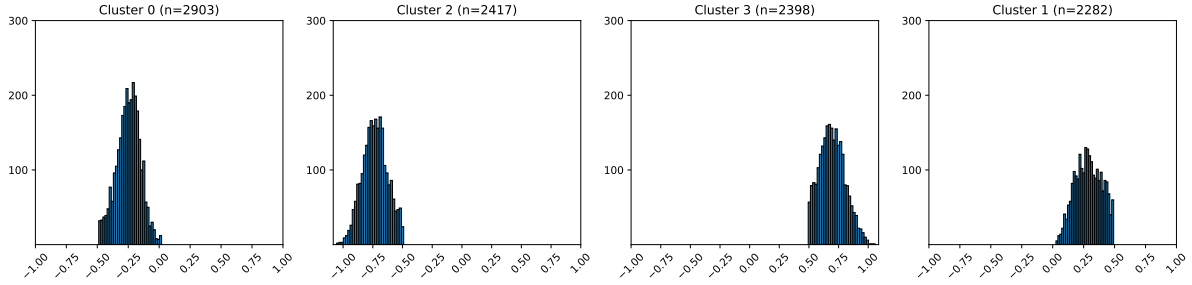


Figure 3.18: Empirical distributions of the clusters in the Prinz data set learned by K-means.

potential

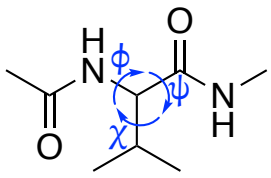
$$V(x) = 4\left(x^8 + 0.8e^{-80x^2} + 0.2e^{-80(x-0.5)^2} + 0.5e^{-40(x+0.5)^2}\right),$$

and driven by the equation

$$x_{t+1} = x_t - \frac{h\nabla V(x_t)}{md} + \sqrt{\frac{2hkT}{md}}\eta_t$$

where $\eta_t \sim \mathcal{N}(0,1)$ [HSH+21].

After segmenting the trajectory with our algorithm and clustering the segments via ADPC, we find that five states are identified, one corresponding to each distinct well in the landscape, and an additional one that seems to correspond to the transition region between the two wells with the lowest barrier to transition.



2-acetamido-*N*,3-dimethylbutanamide
Chemical Formula: $C_8H_{16}N_2O_2$

Figure 3.19: Structure of Ace-Val-NMe, annotated with backbone dihedral angles.

3.4.2 Molecular Dynamics Trajectories

We now turn our attention to simulated molecular trajectories for Ace-Val-NMe (valine dipeptide, or VDP) which were generated with the GROMACS molecular simulations package [AMS+15]. Typically, VDP trajectores are modeled via Markov State Modeling techniques [BPN14], where the data is clustered using a technique like PCCA++ [RW13], and then a reversible transition matrix is estimated based on the cluster identifications. Such trajectories are well-studied in the computational chemistry community and are commonly used for benchmarking algorithms in that domain [DML21]. Modern machine learning techniques are also employed to model these trajectories, wherein an optimal basis of nonlinear functions to describe the model is found via deep learning [MPWN18], [WN20]. Although these trajectories are simulated with several dozen spatial dimensions (VDP contains 28 atoms and hence in principle requires 84 dimensions to express the coordinates of each atomic position), corresponding to the atoms which constitute the molecule, it is well-known that the dimension of a VDP trajectory can be reduced down to $d = 3$, by tracking the so-called "dihedral" angles, known as ϕ, ψ, and χ, which are periodic and hence the trajectory can be embedded onto the 3-torus.

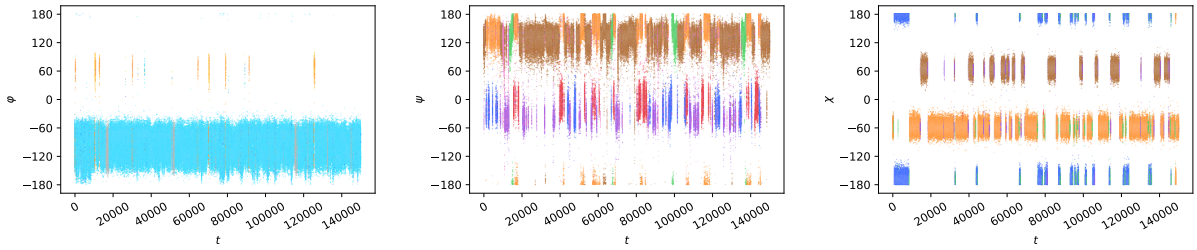
Consequently, this means that our approach must now be applied to a setting which is non-Euclidean. This does not in principle impose a change to our approach, but it does mean that we can no longer rely on (2.11) for a computational boost, and instead we must deal with transport costs on the circle [HKM22]. Our segmentation method allows ADPC to identify 9 distinct clusters which are clearly separated.

Another data set of interest, and one which is, like VDP, commonly used for benchmarking MDS algorithms, is Ace-Ala-NMe (alanine dipeptide, or ADP). Unlike our previous data sets, ADP provides difficulties for our approach because of the rapid way it transitions between its metastable states, resulting in an undersampling problem. By this we mean that among the collection of metastable states of interest, some appear infrequently and for short bursts of time, and even when trying to identify change points by hand for small portions of the trajectory, we find that segments contain sometimes as few as a dozen points. As discussed earlier, this would imply that it is necessary to take a very small window size w to avoid windows which contain multiple distinct states for a given angle, but of course choosing a w small means that the segments we see tend to be poor empirical representations of any cumulative distribution function.

Thankfully, VDP is a somewhat more cooperative MD trajectory, as we can observe in the relative stability of each of its angular trajectories. The 9 most signficantly populated clusters identified by our pipeline are clearly separated and account for approximately 82% of the total data, with the remainder belonging to still smaller clusters. Because VDP does not experience the slow and frequent transitions that ADP does, the cluster separation obtained for the VDP data set is much cleaner and is also readily apparent when examined at the componentwise level. Looking at the individual components of the VDP data, we see that the pipeline is identifying relatively subtle differences between states which are consistent in their appearance but with overlapping support. This is most pronounced in the ψ angle, where the algorithm identifies 3 distinctive clusters supported in the vicinity of $[-60, 60]$ and 3 more distinctive clusters supported in the vicinity of $[60, 210]$.

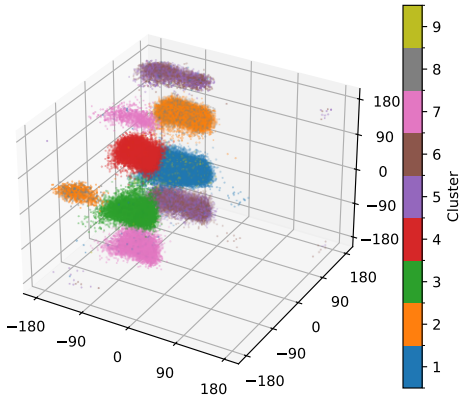
3.4.3 Underwater Acoustics

Our real-world data set comes from an experiment conducted by the Naval Research Laboratory [TFS16]. For this experiment, two underwater speakers were towed behind

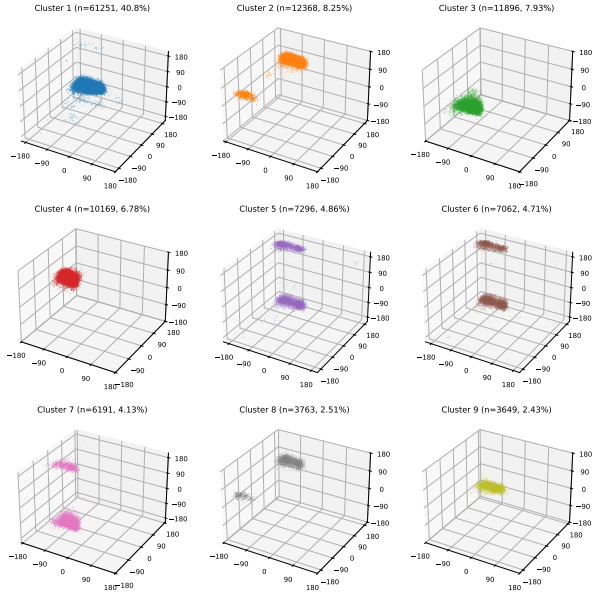


(a) ϕ angle (b) ψ angle (c) χ angle

Figure 3.20: VDP dihedral angles labeled via segment-based clustering with circular optimal transport. Each angle is put through the process of CPD and segment clustering. Then clusters are combined componentwise to produce a final labeling.



(a) Final label assignments for valine dipeptide trajectory clustered with segment-based clustering with advanced density peaks. The parameters used for each angle are $w = 50, q = 0.95$



(b) Valine dipeptide trajectory clustered with segment-based clustering with advanced density peaks. N.B.: percentages do not sum to 1, since only the top 9 most populous clusters are accounted for. Together these clusters account for 82.43% of the data set.

Figure 3.21: Valine Dipeptide Clustering

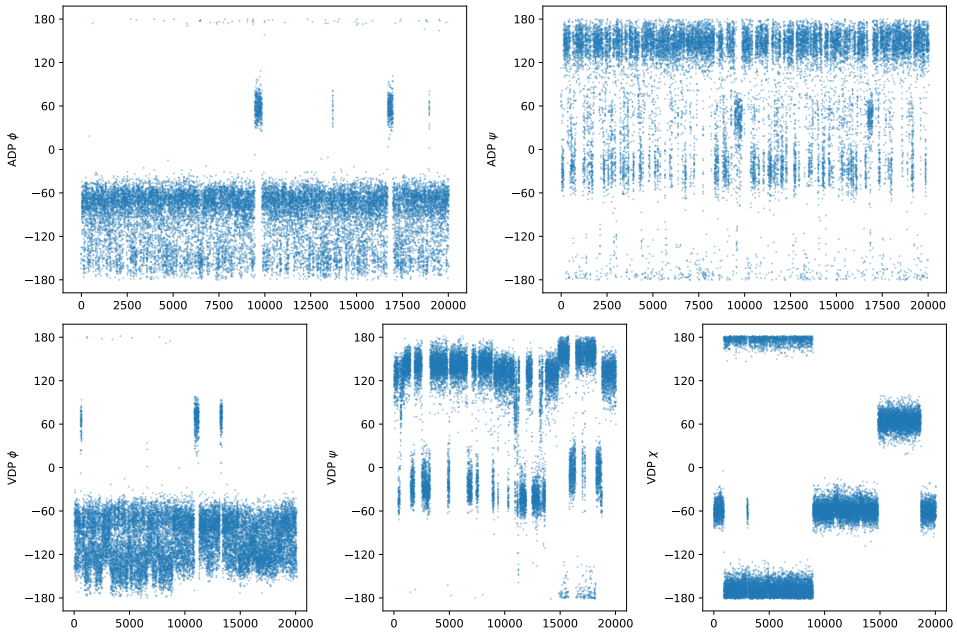


Figure 3.22: Comparison between the componentwise angles of ADP and VDP. Our algorithm struggles to handle ADP, as it bounces back and forth between states very rapidly, with a large amount of overlap between states, something which is particularly pronounced in the ψ angle of ADP (top right). VDP is a more tractable data set for us because of the cleaner and longer-lived changes.

a vessel emitting four distinct signals, each repeated twice, over the course of 9 seconds. Three of these samples were linear frequency modulations (LFM), wherein the signal starts at one frequency and increases linearly to another terminal frequency, and one was a continuous waveform.

We find that a modification of our algorithm applied to the spectrogram of the data is able to identify a suitable selection of change points for the samples. The simplification of our approach in this setting occurs because computing the spectrogram of a signal such as this naturally results in a sequentially ordered set of power spectral densities. This differs in character from the reduction that occurs in mapping the high-dimensional atomic coordinate representation of VDP onto the 3-torus. In that case, there is a one-to-one correspondence between samples represented in full atomic coordinates and samples represented by dihedral backbone angles. In the case of our underwater acoustics experiment, a different sort of compression occurs. Rather than computing the Wasserstein distances

between the sampled voltages read by an analog-to-digital converter, we instead use a short-time Fourier transform (STFT) to examine the change in the frequency domain content overtime. The spectrogram obtained by computing the STFT of the raw voltage data, using a Tukey window with a shape parameter of 0.25 [VGO$^+$20], is an array where the columns represent the distribution of frequency content in the signal over some window of time, while the rows represent the concentration of frequency content for a certain band of frequencies over time. The resolution of the spectrogram is a function of the number of samples per window used in the STFT; in this case, we use $n = 512$. Overall this results in a significant reduction in the number of Wasserstein distances computed, since rather than sliding across every sample of the raw data, we now simple compute the Wasserstein distance between contiguous columns of the spectrogram. In 3.4.3, the raw number of samples in the time domain is $421,848$, but the number of power densities to compare after computing the spectrogram is just 941. Because ground truth data is available, we also produce a table of precision and recall scores for our algorithms output as we did for the artificially generated toy-data set, with the slight modification that the tolerance now scales with number of samples per Fourier window, which is here taken to be $n = 512$. The sampling rate of the data is 11718 samples per second, thus the corresponding tolerance range can be interpreted as detecting a change within roughly $0.1 - 1.0$ seconds.

Underwater Acoustics Data Precision & Recall					
# True CPS	# Detected CPS	Precision	Recall	Max Prec.	Max Recall
31	36	0.513	0.597	0.694	0.806

Table 3.2: Table of precision and recall scores for the underwater acoustics data set. Precision and recall are averaged over tolerances of $tol = 1536, \ldots, 12800$

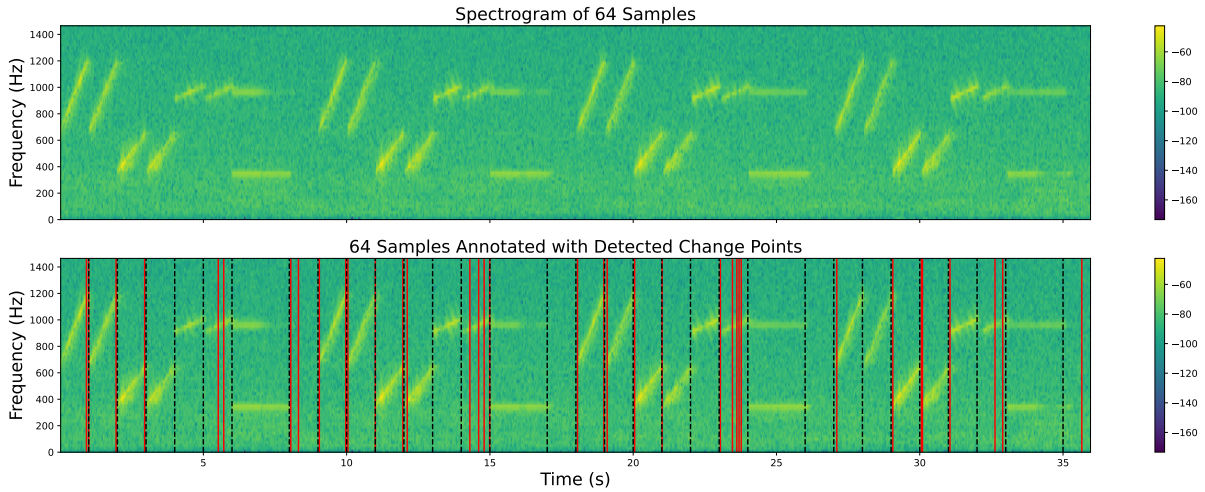


Figure 3.23: Underwater Acoustics Data annotated with detected change points. The parameters for this modified version of the algorithm are a windowsize of $n = 512$ for the short-time Fourier transform and $q = 0.96$.

3.5 Discussion & Conclusion

In this work, we have investigated the use of the Wasserstein metric as a tool for identifying change points in time series data and for clustering large data sets by lifting the problem into the space of distributions so that points can be considered as a part of a larger segment, thereby reducing the number of pairwise computations needed to form the similarity matrix. Our work extends previous considerations of the approach by removing parametric assumptions on the statistical structure of the data, and expanding the space of relevant data sets amenable to transport-based change point detection to include multi-dimensional data. We circumvent the curse of dimensionality by performing our change point detection and clustering on each component of the data set.

In the future, this work could be extended by taking an approach to the clustering aspect of the problem which is more harmonious with the transport driven framework for the segmentation. In particular, one could leverage the geometry of Wasserstein barycenters to implement a K-means type algorithm, or potentially adapt a variation of density peaks-based clustering into the Wasserstein space. This approach would likely involve entropic regularization of the Wasserstein distance, due to the computational complexity

involved in computing Wasserstein barycenters. Also of interest is the optimization of the parameter choices w and q. Elsewhere in the literature the problem of choosing an optimal window size for CPD algorithms based on the Kolmogorov-Smirnov test have been considered, and similar pursuits might be considered in the context of Wasserstein-based algorithms.

Furthermore, it is of interest to harmonize this approach with the Markov state modeling approach which is prominent in the computational chemistry community.

Acknowledgments: The authors gratefully acknowledge support from NSF DMS-2309519 and NSF DMS-2318894. We also extend our thanks to our colleagues at the Naval Research Laboratory Roger M. Oba and Laurie T. Fialkowski for helpful conversations and supply of the underwater acoustics dataset, as well as to Dr. Hannah J. Bates for her knowledge of chemical drawing software and willingness to help create figures.

Chapter 4

Regularized W_2 Barycenters on Discrete Metric Spaces

"He began then, bewilderingly, to talk about something called entropy."

— Thomas Pynchon, *The Crying of Lot 49*

Chapter Overview

This chapter represents a first pass at implementing a barycentric coding model for measures supported on graphs, and it is largely superseded by Chapter 5[1]. We apply entropically regularized L^2-Kantorovich geometry to the setting of a discrete metric space, where the cost functions are given by different choices of graph metric, which capture varying levels of connectivity information. We find that despite the degeneracy of the Monge-Kantorovich geometry in the case of discrete metric spaces, regularized transport leads to geometric objects (e.g., geodesics and more generally barycenters) which convincingly resemble what one would expect from a transport-like geometry on the space of probability distributions defined for a network. We consider both synthesis (combining measures with respect to the regularized L^2-Kantorovich geometry) and analysis (decomposing signals with respect to fixed reference measures) on real geography data, demonstrating the utility of our approach. Our code is available on GitHub[2].

4.1 Introduction

At this point in the development of data-driven scientific methods, the significance of network-structured data hardly bears remarking upon. Graphs arise naturally in models of social networks, discrete resource distribution, and segregation patterns [SMD+19b], and there is growing interest in signal processing methods when the underlying space lacks a continuous structure [SMD+19a]. In the past three decades, the topic of optimal transportation has become increasingly popular due (at least in part) to its myriad connections with a variety of mathematical fields (e.g., analysis, probability, PDE)[Vil09, FG21, ABS21], and its applications for problems in machine learning (ML) via, for ex-

[1] This chapter is based on a paper co-authored with James M. Murphy, [GM25]

[2] https://github.com/dcgentile/fixed-support-barycenters

ample, Wasserstein GANs [ACB17] or the barycentric coding model (BCM) [WJT+22]. While optimal transportation has proven to be a font of ideas giving rise to analytic techniques which have lead to developments in these fields and many others, applications of these ideas to the network setting have proven somewhat tricky, often relying either on restrictive assumptions about the nature of the networks in question [Sol18], or requiring abstract machinery leading to cumbersome numerical methods [ERSS20] and hence with as-of-yet limited use.

4.1.1 Summary of Contributions

Here, we apply some work originally developed by [BPC16] together with a family of graph metrics which in some sense encode connectivity to implement a version of the barycentric coding model for probability measures, a scheme for encoding probability measures as weighted geometric averages of a given family of "reference" measures in the L^2-Kantorovich geometry, and we find that we are reliably able both to synthesize and analyze measures using these tools in a way which resembles the spirit of transport.

4.2 Computing W_2 Barycenters via Entropic Regularization

We start by quickly recapping the relevant material from Chapter 2. Letting (Ω, d) be some metric space, $c : \Omega \times \Omega \to \mathbb{R}^+$ be a cost function, and $\mu, \nu \in \mathcal{P}(\Omega)$ be probability measures on Ω , the Kantorovich problem is to find a *coupling* γ of μ and ν (i.e., probability distributions on the product space $\Omega \times \Omega$ having marginals given by μ and ν, the set of which is denoted $\Pi(\mu, \nu)$) which solves

$$OT(\mu, \nu) := \inf \left\{ \int_{\Omega \times \Omega} c(x, y) \ \mathrm{d}\gamma : \gamma \in \Pi(\mu, \nu) \right\}.$$

When $c = d^p$, for some $p \geq 1$ the p^{th} root of the value of the Kantorovich problem is known as the *L^p-Kantorovich* metric and is denoted by W_p. The L^2-Kantorovich metric, in particular, induces a rich geometric structure in the space of probability measures, and of particular interest are geodesic curves and gradient flows in this geometry [JKO98]. Specifically, it offers a method of comparing probability distributions in a way that respects the underlying geometry of the base space. For example, if two probability distributions have disjoint support, the L^2 distance between them will be invariant, no matter where the (disjoint) supports lie; on the other hand, the Kullback-Leibler divergence will be infinite, and yield essentially no information at all. Because the L^2-Kantorovich geometry preserves geometric information about the space on which the distributions are supported, it is a natural tool of interest for studying the dynamics of distributions on, for example, a fixed geographical region.

It is an unfortunate fact, however, that the Kantorovich problem leads to a degenerate geometry when the underlying metric space is discrete. The geometry so induced is one in which there are no non-trivial geodesic curves. While this problem can be tamed by approaching from a frame of reference beginning with the celebrated Benamou-Brenier formula, and suitable choices and assumptions will lead to the analagous results one might hope to see [Maa11], the computational tools required for working in this regime are totally distinct from the popular methods we will describe below, and are much harder to implement[ERSS20].

The Kantorovich problem is a linear program, and when the distributions have finite support, standard convex optimization techniques may be employed to identify optimal couplings. Although polynomial in time, algorithms for solving such linear programs have poor scaling, roughly cubic in the support of the measures; however, a convex regularization of the problem, obtained by simultaneously computing the Kullback-Leibler divergence of a given coupling and the product coupling, converts this linear program into a strongly convex one. Consequently, a host of new tools become available from the

realm of (strongly) convex analysis, and this results in a dramatic speed-up, shaving off an entire order of magnitude in compute time [Cut13]. If we fix $\varepsilon_{reg} > 0$ and suppose that $\mu = \sum_{i=1}^{N} \alpha_i \delta_{x_i}, \nu = \sum_{j=1}^{M} \beta_j \delta_{y_j}$ and take C to be a matrix whose entries are given by $C_{ij} = d^2(x_i, y_j)$, one writes for the entropy-regularized transport problem,

$$OT_{\varepsilon_{reg}}(\mu, \nu) := \inf \{\langle \gamma, C \rangle + \varepsilon_{reg} H(\gamma)\}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product on matrices and

$$H(\gamma) := -\sum_{i,j} \gamma_{i,j}(\log \gamma_{i,j} - 1).$$

In the limit of $\varepsilon_{reg} \to 0^+$, we recover the optimal transport cost associated with the measures μ and ν [BCC$^+$15]. Without regularization, the optimal plan is simply a permutation matrix, assigning one point in the first distribution to another in the second; with regularization, diffusion is introduced from the entropic term. Now, given p reference measures $\{\nu_i\}_{i=1}^{p} \subset \mathcal{P}(\Omega)$ and a point in the p-simplex $\lambda \in \Delta^{p-1}$, a *barycenter* in the L^2-Kantorovich metric is a minimizer of the variance functional

$$J_\lambda(\nu) = \sum_{i=1}^{p} \frac{\lambda_i}{2} W_2^2(\nu, \nu_i).$$

The theory of W_2 barycenters was first explicated in [AC10], wherein the authors demonstrate existence and uniqueness for the case of distributions which are supported on $\Omega = \mathbb{R}^d$, and it supplies us with a notion of geometric mean with respect to the L^2-Kantorovich geometry. In that paper the authors also show that for the case of $p = 2$ W_2 barycenters coincide exactly with the McCann interpolation of measures [McC97], i.e., for $\lambda = (t, 1-t)$, we recover the unit speed geodesic between the two reference measures in the L^2-Kantorovich geometry. Thus W_2 barycenters give us a notion of measure interpolation which respects the underlying geometry of the domain, as illustrated

in figure 4.1 — with the L^2-Kantorovich geometry, there is a "continuous" movement of mass from one position to the other, whereas in the L^2 geometry the mass is "teleported" between its initial and terminal configurations. The computation of W_2 barycenters is

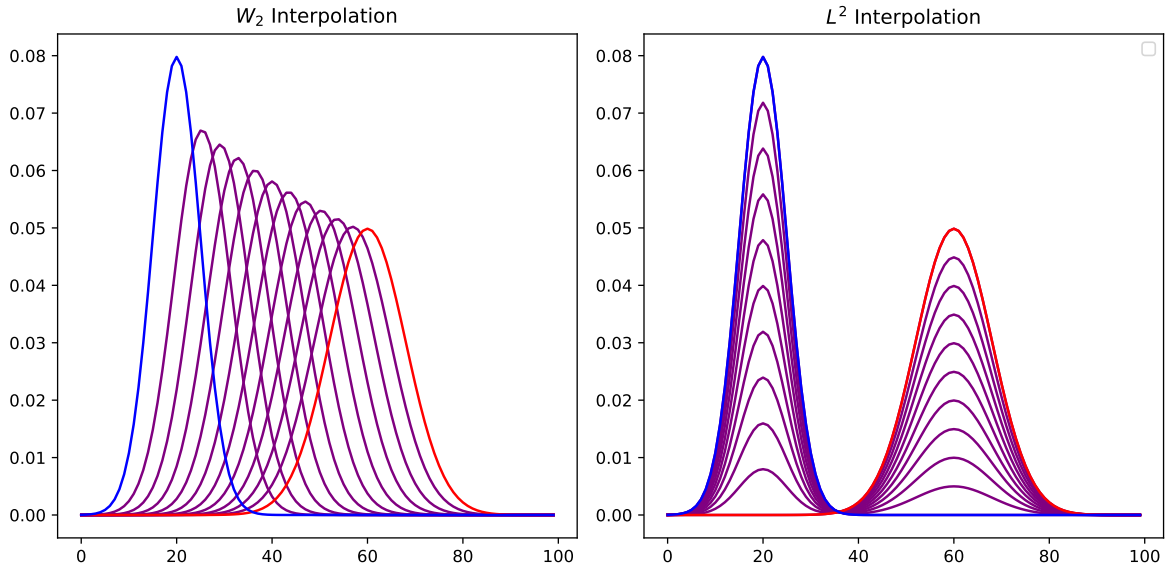


Figure 4.1: *Left:* McCann Interpolation of Gaussian measures. *Right:* L^2 Interpolation of Gaussian measures

in general an NP-hard problem [AB22], but once again entropic regularization makes the computation of regularized barycenters feasible, particularly the realization that they may be computed via Bregman projections [BCC+15, BPC16]. Finally, the *barycentric coding model* for measures in the L^2-Kantorovich geometry is the problem of identifying a suitable "barycentric basis set" of measures such that

$$\text{Bary}(\{\nu_i\}_{i=1}^p) = \left\{\nu : \nu \text{ solves } \inf\{J_\lambda\} \text{ for some } \lambda \in \Delta^{p-1}\right\}$$

adequately represents a dataset comprised of measures. For the BCM, there are two essential operations which are necessary: the so-called "synthesis" and "analysis" problems. For the former, we mean the problem of computing the barycenter ν^* of a given family of reference measures $\{\nu_i\}_{i=1}^p$ for a given weight vector $\lambda \in \Delta^{p-1}$, and for the latter we mean finding, for a given family of reference measures and a target measures ν, a weight

vector λ^* such that ν is the barycenter of the ν_i with weights λ^*. One can think of this as a starting point for a kind of principal component analysis in the L^2-Kantorovich geometry — although there is no linear structure on the space, W_2 barycenters give a way to encode a family of measures as the "barycentric combinations" of a finite family of measures. The problem of identifying suitable basis measures is also of interest, but not the primary concern of this work.

Since the essence of the L^2-Kantorovich geometry follows from the metric structure of the underlying space, we should take care to specify what the precise metrics of interest are on our networks. For the remainder, let $\mathcal{X} = (V, E)$ be a graph with vertex set V and edge set $E \subset V \times V$, and let $\omega : E \to \mathbb{R}^+$ be a weighting of the edges. The *shortest path metric* on $\mathcal{X}$ will be defined to be

$$d(u, v) := \inf \left\{ \sum_k \omega_{i_k, i_{k+1}} \right\}$$

where the infimum is taken over all uv walks. If $|V| < +\infty$, then the shortest path between any two nodes can be computed in $\mathcal{O}(|V|^2)$ via Dijkstra's algorithm [CCPS97]. An alternative metric on the nodes is given by the so-called *diffusion distances*, a family of metrics parameterized by $t > 0$, and which intuitively measures the likelihood that a random walker on the network traverses from u to v in time t [CL06, CLL$^+$05]. Fix $t > 0$, let Q be the natural Markov transition matrix induced by the structure of the graph, that is,

$$Q_{ij} = \begin{cases} \deg(i)^{-1} & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

and let π be the steady state distribution, i.e., the eigenvector of Q with unit eigenvalue, and let $p(y, t|x)$ be the probability of a random walker landing on node y t steps after landing on node x. Then the (squared) diffusion distance between nodes $u, v \in V$ at time

t is given by

$$d_t^2(u,v) = \sum_w \frac{(p(w,t|u) - p(w,t|v))^2}{\pi(w)}.$$

As we see in figure 4.2, as $t \to \infty$, the diffusion distance between a given node and its neighbors decreases in a way reminiscent of heat dispersion on the graph.

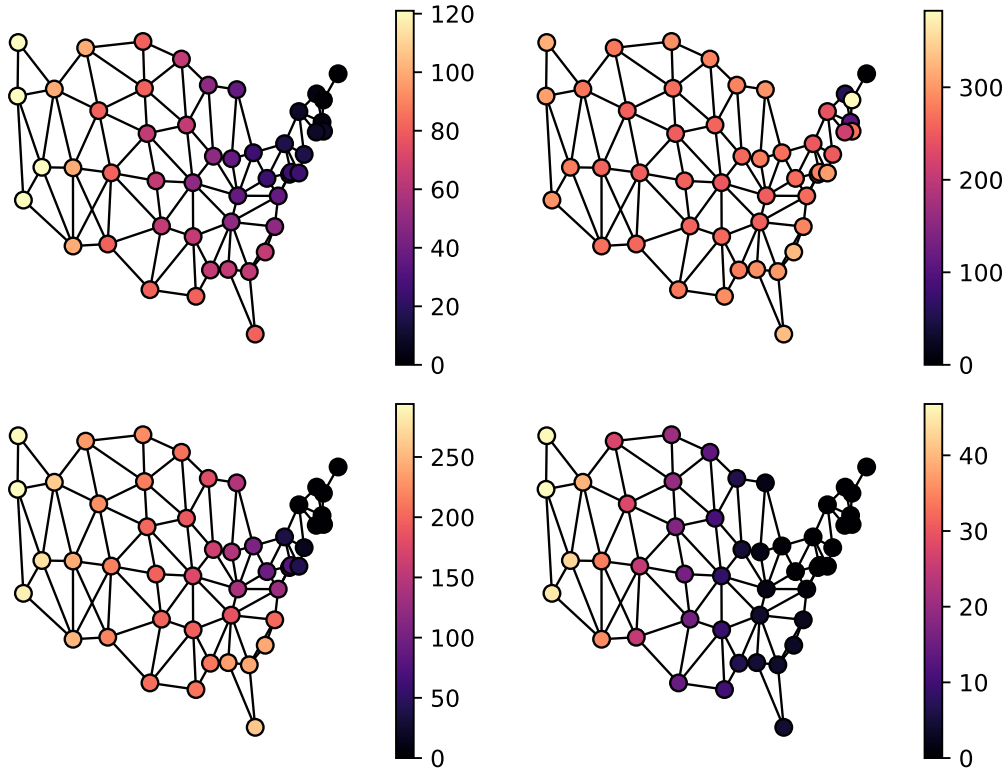


Figure 4.2: Color intensity represents the cost of moving from one fixed node (Maine). Upper left: squared shortest path metric; Upper right: squared diffusion metric at $t = 8$; Lower left: squared diffusion metric at $t = 32$; Lower right: squared diffusion metric at $t = 1024$

4.3 Methodology & Numerical Results

Following [BPC16], we use gradient descent to compute barycenters of measures defined on a fixed-support, with the cost matrix realized by 1) the shortest path metric induced by the connectivity of the graph and 2) the diffusion distances between the nodes for various values of t. In other words, fixing a graph, a family of reference measures

$\{\nu_i\}_{i=1}^p$, and a prescribed set of barycentric coordinates $\lambda = (\lambda_1, \ldots, \lambda_p)$, we compute the measure ν^* which minimizes J_λ. With ν^* in hand, we then attempt to recover the prescribed coordinates λ. The network is defined by the 48 contiguous states of the U.S.[3], and demonstrates irregular connectivity (i.e., the degree of the vertices is not constant). Correctness is assured in our computations because the algorithm laid out in [BPC16] simultaneously computes an approximation of the regularized barycenter and the optimal barycentric coordinates for barycentric representation over the dictionary set, and we can check directly that the recovered coordinates of the output barycenter match the input coordinates.

Now, let $\hat{\nu}$ be the barycenter with coordinates $\hat{\lambda}$ be a step along our sequence of iterates in the gradient descent scheme. In minimizing the "barycentric loss" of the coordinates $\hat{\lambda}$ during the process of coordinate recovery (that is, the difference between $\hat{\nu}$ and the target barycenter ν^*), a choice of loss function must be made. Suitable choices are the L^1 and L^2 norms, the Kullback-Leibler divergence, and the total variation between the measures — here we choose the L^2 norm for the loss.

In figures 4.3, 4.4, and 4.5, we solve both the analysis and synthesis problems for some chosen measures on the network. In figures 4.3 and 4.4, we have two measures which are concentrated in disparate geographical regions, and using this transport-like interpolation, transport mass from one locale to the other. Interpreting this geodesic as having barycentric coordinates $(t, 1-t)$, we check our ability to recover the coordinates for $t = 1/3, 2/3$ and find that we do so with modest error. In figure 4.5 we append an additional reference measure, concentrated away from either of the other two, and synthesize a barycenter as an uneven transport-like interpolation of the three. As with the geodesic, the algorithm accurately recovers the input coordinates.

We observe natural-looking barycenters arising, particularly for the case $p = 2$, corresponding to geodesic curves in the space of probability measures, and that the barycentric

[3]https://hub.arcgis.com/datasets/usdot::states/

coordinates of these synthesized measures are accurately recovered, with degeneracy occurring in the limit as $\varepsilon_{reg} \to 0^+$. In figures 4.3 and 4.4 we see that the geodesics induced

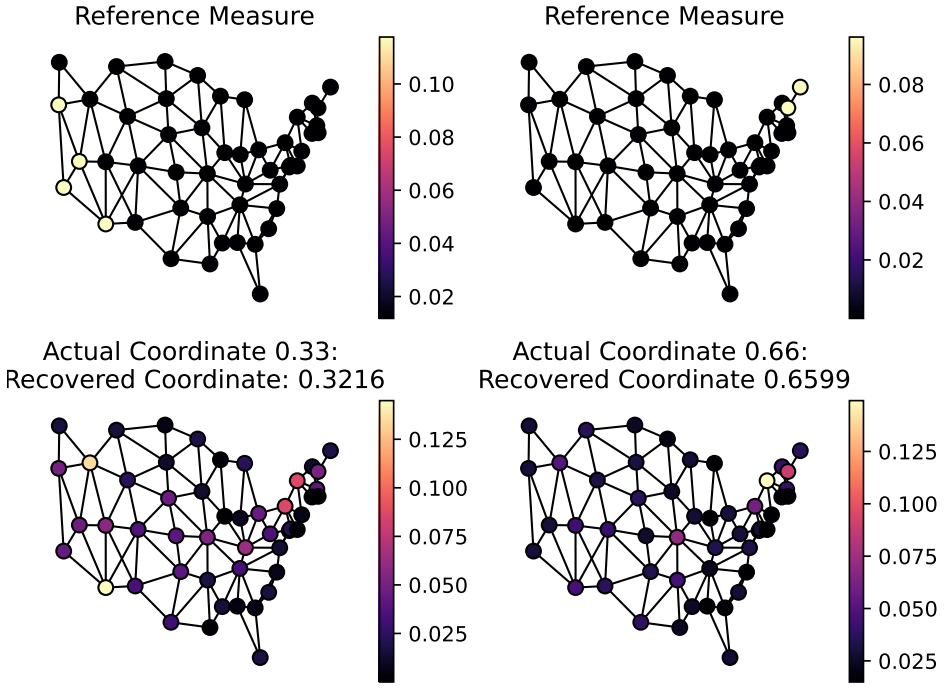


Figure 4.3: Transport-like interpolation of measures on graph, with cost given by squared shortest-path distance, $\varepsilon_{reg} = 0.125$; Upper left, right: reference measures, lower right: computed geodesic with weights $(1/3, 2/3)$; lower right: computed geodesic with weights $(2/3, 1/3)$

by this transport-like geometry inherit the diffusivity of the ground metric. It is interesting that strictly speaking there is no direct access to edge/connectivity information in these metrics, yet we still obtain a motion that has a "continuous" appearance.

4.4 Discussion

Our experiments demonstrate that despite the degeneracy of the L^2-Kantorovich geometry in the graph setting, a measure interpolation that respects the geometry and connectivity of a graph can still be achieved via a combination of entropy-regularized optimal transportation and choice of suitable ground cost, and indeed that the BCM may be de-

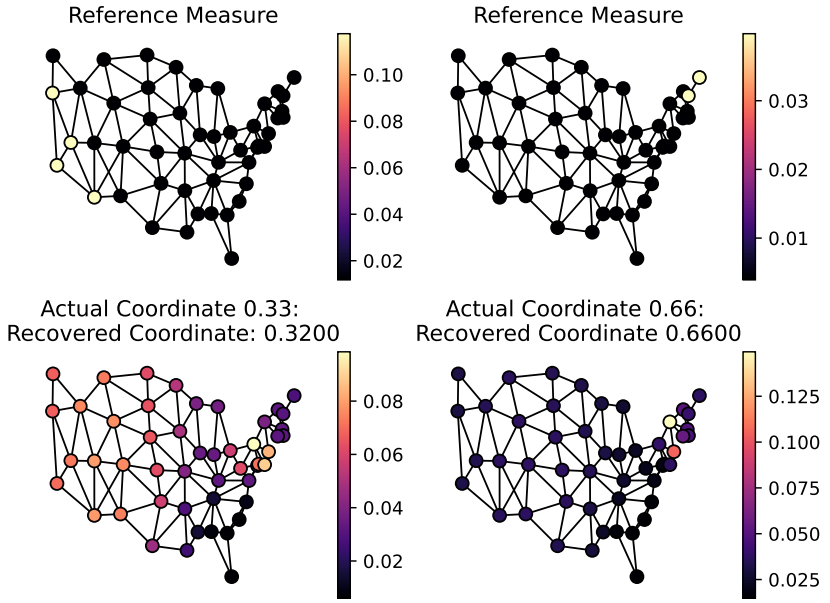


Figure 4.4: Transport-like interpolation of measures on graph, with cost given by squared diffusion distance, $t = 16, \varepsilon_{reg} = 0.125$; Upper left, right: reference measures, lower right: computed geodesic with weights $(1/3, 2/3)$; lower right: computed geodesic with weights $(2/3, 1/3)$

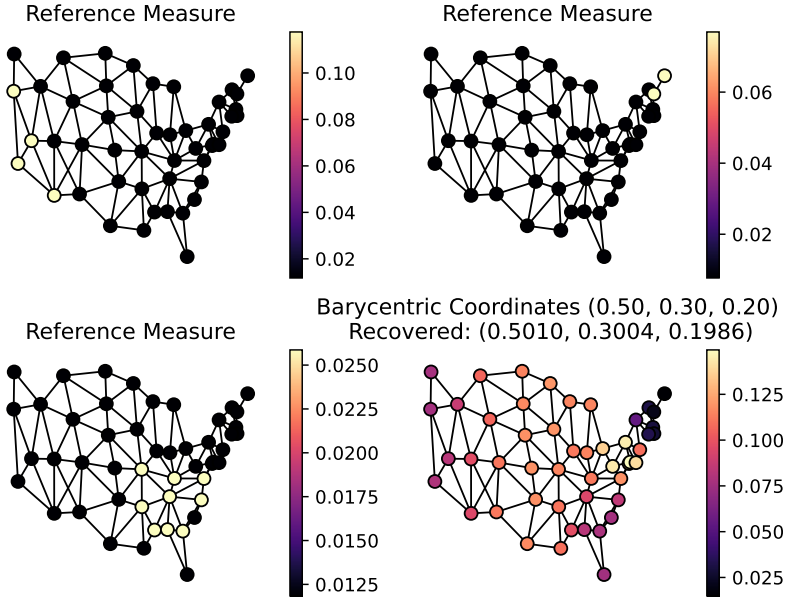


Figure 4.5: Transport-like barycenter of measures on graph, with cost given by squared diffusion distance, $t = 16, \varepsilon_{reg} = 0.125$; Upper left, upper right, lower left: reference measures, lower right: computed barycenter with weights $0.5, 0.3, 0.2)$, and recovered coordinates (0.5010, 0.3004, 0.1986)

ployed both for measure synthesis and decomposition in this setting. This is a promising baseline for the development of transport-based signal processing tools on networks. As a matter of future work, it is of interest to develop an algorithm that can learn an optimal set of dictionary atoms for measure representation in the barycentric coding model. It is also of interest to investigate any possible relationship between barycenters obtained via regularization of the Kantorovich problem and those that respect the formal Riemannian structure on the probability simplex induced by the Benamou-Brenier approach to network transportation, where the transport-like geometry so developed depends directly on the graph structure.

4.5 Acknowledgements

The authors acknowledge generous support from NSF-DMS 2318894.

Chapter 5

Static and Dynamic Approaches to Computing Barycenters of Probability Measures on Graphs

"Everything that matters is finite"

— billy woods

Chapter Overview

The optimal transportation problem defines a geometry of probability measures which leads to a definition for weighted averages (barycenters) of measures, finding application in the machine learning and computer vision communities as a signal processing tool. Here[1], we implement a barycentric coding model for measures which are supported on a graph, a context in which the classical optimal transport geometry becomes degenerate, by leveraging a Riemannian structure on the simplex induced by a dynamic formulation of the optimal transport problem.

We approximate the exponential mapping associated to the Riemannian structure, as well as its inverse, by utilizing past approaches which compute action minimizing curves in order to numerically approximate transport distances for measures supported on discrete spaces. Intrinsic gradient descent is then used to synthesize barycenters, wherein gradients of a variance functional are computed by approximating geodesic curves between the current iterate and the reference measures; iterates are then pushed forward via a discretization of the continuity equation. Analysis of measures with respect to given dictionary of references is performed by solving a quadratic program formed by computing geodesics between target and reference measures. We compare our novel approach to one based on entropic regularization of the static formulation of the optimal transport problem where the graph structure is encoded via graph distance functions, we present numerical experiments validating our approach[2], and we conclude that intrinsic gradient descent on the probability simplex provides a coherent framework for the synthesis and analysis of measures supported on graphs.

[1] Contents of this chapter are based on a paper co-authored with James M. Murphy, [GM26].

[2] Library and code for reproducing experiments is available at github.com/dcgentile/GraphTransportation.jl

5.1 Introduction

The purpose of this chapter is to propose and implement a barycentric coding model for data which can be represented as probability measures on a graph [3] . The change of setting from a continuous domain to a discrete one poses some interesting technical challenges. The L^2-Kantorovich metric becomes degenerate when the underlying metric space is discrete, and immediately renders the aforementioned work on barycenters of measures inapplicable [Maa11]. Nevertheless, since the early 2010s, transport metrics for probability measures on graphs have been constructed and studied [Maa11, CHLZ12], and it has been shown that these discrete transport metrics do indeed bear real relation to the classically studied L^2-Kantorovich metric for measures supported on $\mathbb{R}^d$ — in particular, it is known that, for a sequence of suitably constructed graphs & appropriately chosen admissible mean, the discrete transport metric converges in the sense of Gromov-Hausdorff to the classical L^2-Kantorovich distance on $\mathbb{R}^d$, as the graphs become finer and finer approximations of Euclidean domains [GKM20, GM13, EM12].

In order to achieve our implementation, we leverage the Riemannian structure of transport on discrete metric spaces, along with previously developed numerical routines for approximating geodesics in such a setting [ERSS20]. The momentum information associated to discretized energy minimizing geodesics is interpreted as an oracle for the inverse of the exponential mapping. Then, barycenters, characterized as minimizers of a certain variance functional which is convex in the Euclidean geometry, are synthesized using an intrinsic gradient descent approach, where movement on the simplex is approximated by a continuity equation-inspired heuristic, while analysis of barycenters is performed by solving a finite dimensional quadratic program formed by approximating the inverse of the exponential mapping based at the target of the analysis. We find that our algorithms are consistent in the sense that synthesizing a barycenter of given refer-

[3] In this work, all measures considered are probability measures, potentially with respect to a fixed reference measure, and having total mass 1, although we may occasionally drop the modifier "probability".

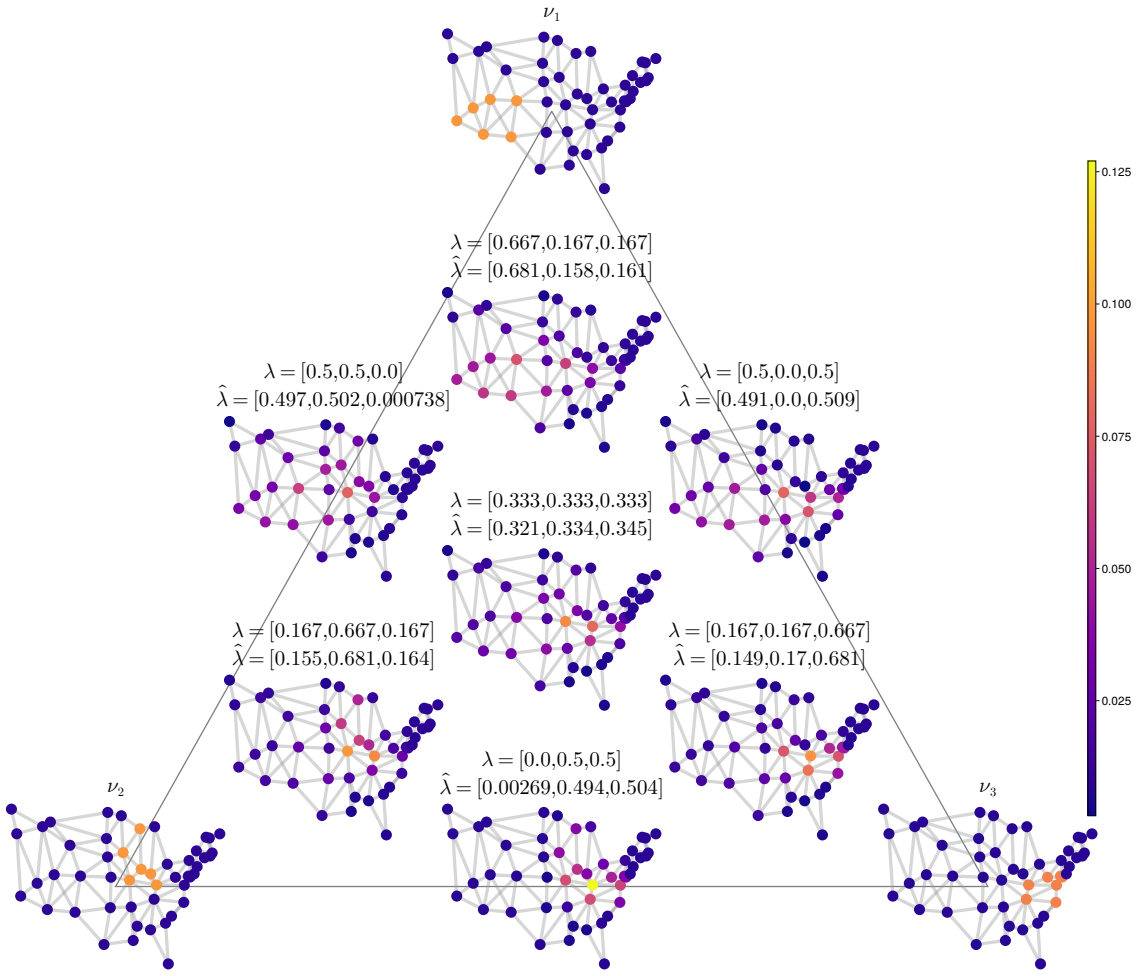


Figure 5.1: A schematic of the barycentric coding model for measures supported on a graph. At the extremal points of the simplex, which correspond to the vertices of the triangle, are the reference measures; away from the extremal points are *barycentric interpolations* of the reference measures for certain weights λ. The coordinates recovered by solving the analysis problem are displayed and labeled with $\hat{\lambda}$, to 3 significant figures.

ence measures and coordinates and then analyzing the output measure with respect to the reference measures leads to accurate recovery of the given coordinates.

Our experiments indicate that studying the theoretical convergence properties of algorithm may be a fruitful avenue for future research, as we observe consistent convergence of our synthesis routine. Concrete directions in this area include giving characterizing how strongly convex the graph-transport distance $\mathcal{W}$ is, identifying $L > 0$ such that the variance functional can be said to be L-smooth, or if the variance functional satisfies a Polyak-Łojasiewicz inequality. Any of these characterizations would yield information that could lead to theoretical guarantees about the convergence of our descent scheme.

5.1.1 Structure of the chapter

The structure of this chapter is as follows. In Section 2, we provide the essential background information concerning graphs, optimal transport, transport on discrete metric spaces, and Riemannian optimization. In Section 3, we explicate the algorithms used to implement the barycentric coding model on graphs. In Section 4, we present our numerical results. We conclude with a discussion in Section 5, and outline future directions for improving this work both from theoretical and practical points of view.

5.2 Background

5.2.1 Relation to the Literature

Let us briefly recap the development of barycenters in the L^2-Kantorovich geometry (also known as the W_2 geometry) and optimal transport methods for measures on graphs, as well as highlight some alternatives to the one which serves as our focal point. For the granular details of optimal transport, we refer the reader to any of the excellent textbooks on the subject, such as [ABS21, FG21, Vil09, San15]; in Section 2.2 we recall the most essential definitions and results for understanding of this work.

The particular subdomain of optimal transport which will be of interest to us in this work concerns W_2 *barycenters*, a construction first introduced in [AC10], wherein the authors demonstrate that under mild regularity assumptions on a collection of measures supported on $\mathbb{R}^d$, a unique weighted center of mass can be defined for the measures. In [KP17], it was shown that the construction can be extended to the non-Euclidean case, where measures are supported on a Riemannian manifold. W_2 barycenters have since attracted much attention in the optimal transport community. Computationally, it was shown in [BCC$^+$15] that entropic regularization (see Section 2.3.2 for discussion of entropic regularization) allows barycenters of discrete measures to be approximated quickly via Bregman projection methods, and in [ÁEdCM16] a fixed-point method is applied to obtain rapid convergence to barycenters when all the measures under consideration are Gaussian. Outside of special cases however, W_2 barycenters are generally hard to compute [AB22]. W_2 barycentric coordinates were introduced for histogram data in [BPC16]. A similar regression scheme exists when the measures are absolutely continuous [WJT$^+$22]. The statistics of W_2 barycenters have also been of interest [ZP19, PZ20]. Recently stochastic gradient descent in the W_2 metric has been applied to approximate W_2 barycenters [BFRT25]. This method is similar to ours in spirit, but the distinction between continuous and discrete domains makes them essentially different in their implementation and background. Recently a unified framework for studying first-order optimization methods in the Wasserstein space was presented in [LBD25], yielding results on necessary and sufficient optimality conditions for constrained optimization problems in the W_2 geometry, in particular justifying "set the gradient to zero" approaches.

In [JCG20], the smoothing bias introduced by the reference measure used to define the entropic regularization of the W_2 geometry was studied resulting in the definition of the debiased Sinkhorn barycenter, as a compromise preserving the usefulness of fast Sinkhorn iterations for barycenter computations while eliminating the smoothing introduced by the regularization. The debiased Sinkhorn divergence was further explored in the context of

the barycentric coding model in [MMA25], where the synthesis and analysis problems are characterized as solutions to a fixed point of the weighted average of certain regularized displacement maps and solutions to finite-dimensional convex quadratic programs respectively. Linearization of the barycentric coding model (LBCM), where the problem is lifted into the tangent space at a fixed reference measure, was considered in [MDC20] as well as [WMAM25], with particular attention paid to the case of compatible measures; additionally, a suitable family of reference measures is identified such that the LBCM is capable of expressing all probability measures on the closed unit interval. Regularity of entropy-regularized barycenters was explored in [CEK21], where an elliptic system of Monge-Ampére equations is employed to establish existence and uniqueness, bounds on moments and the Fisher information of the barycenters, and to perform asymptotic analysis, obtaining a central limit theorem and law of large numbers for barycenters of empirical measures.

Also of special interest to us is recent research on optimal transportation of measures which are supported on graphs, instead of $\mathbb{R}^d$. The earliest work in this direction are the approaches developed in parallel by [Maa11] and [CHLZ12], wherein a Riemannian structure is constructed for the simplex, inspired by the fluid dynamics formulation of the classical optimal transport problem. This approach has been expanded on in recent years. For example, it has been used to define a notion of discrete Ricci curvature for Markov chains [EM12]. The so-called Gromov-Hausdorff scaling limits of this construction, where the graph is interpreted as a coarse representation of some continuous metric space, have been investigated and found, under certain technical assumptions, to be consistent in the limit with the classical L^2-Kantorovich metric [GM13, GKM20]. Attention has been given to the underlying Hamilton-Jacobi dynamics of the geodesics in this discrete transport geometry [EMW19, GLM19], and many classical objects from Riemannian geometry, such as the Christoffel symbols and the Riemannian curvature tensor have been computed [Li22]. We also point out that other approaches to the problem of continuous flows in

the space of measures on graphs have been developed [SRGB16], as well as approaches inspired by the Beckmann problem [RKMC25], the notion of circuit resistance [RWC24], and vector-valued transport for measures supported on spaces which are products of Euclidean spaces and graphs [CTN25].

Finally, let us highlight a few papers in the domain of Riemannian optimization. Riemmanian optimization is a broad field of active interest to the applied mathematics community. The particular problem of finding the Riemannian center of mass via intrinsic gradient descent was considered in [Afs11], and further investigation into the algorithm's convergence properties was made in [ATV13].

5.3 Algorithms

5.3.1 Overview

We now present our approach to realizing the barycentric coding model for measures on discrete metric spaces. Synthesis is accomplished by minimizing the discrete variance functional (2.15) via intrinsic gradient descent in the $\mathcal{W}$ metric. To be able to actually implement this minimization scheme we need to be able to do two things: evaluate the exponential map in order to push our iterates forward, and we also need to be able compute the intrinsic gradient of the functional at point. Analysis turns out to be significantly easier, because all we need to do in this case is form the Gram matrix (2.10) and minimize it with an off-the-shelf optimization library, such as CVX [GB14, GB08]. Forming the Gram matrix requires approximating the inverse of the exponential map at the target measures ν for each of the reference measures. This can be accomplished as a direct by-product of approximating the distance between the target and reference measures using the algorithm outlined in [ERSS20], as we will explain below.

5.3.2 Computation of Optimal Transport on Discrete Spaces

Having discussed the discretized transport metric on $\mathcal{P}(\mathcal{X})$, a natural question to ask is whether or not that distance can actually be computed in practice for a given graph and pair of measures. The answer to this is that while analytic computations of the distance are out of reach in general, at the least an approximation of the distance can be computed numerically. We recap here some essential definitions, constructions, and results from [ERSS20].

The insight in [ERSS20] is to proceed via Galerkin discretization in time of the discrete continuity equation. By first constructing an appropriate discrete analog to the graph-transport metric, it is shown via a Γ-convergence technique that this discretization of the distance function converges to the true distance function in the limit of small step size. To that end, the following definitions are needed. Divide the unit interval $[0,1]$ into N subintervals $I_i = [t_i, t_{i+1})$ with uniform step size $h = \frac{1}{N}, t_i = ih$ and define the following sets:

$$V^1_{n,h} := \left\{\psi_h \in C^0([0,1], \mathbb{R}^{\mathcal{X}} : \psi_h|_{I_i}) \text{ is affine } \forall i = 0, \dots, N-1\right\},$$

$$V^0_{n,h} := \left\{\psi_h : [0,1] \to \mathbb{R}^{\mathcal{X}} : \psi_h|_{I_i} \text{ is constant } \forall i = 0, \dots, N-1\right\},$$

$$V^0_{e,h} := \left\{\psi_h : [0,1] \to \mathbb{R}^{\mathcal{X}\times\mathcal{X}} : \psi_h|_{I_i} \text{ is constant } \forall i = 0, \dots, N-1\right\}.$$

We can think of these sets as representing, respectively: piecewise linear curves in the space of functions on $\mathcal{X}$, their piecewise constant weak derivatives, and piecewise constant curves in the space of vector fields on $\mathcal{X}$.

Definition 5.3.1. The set of solutions to the *discretized continuity equation* for given

boundary values $\rho_A, \rho_B \in \mathbb{R}^{\mathcal{X}}$ is given by

$$\mathcal{CE}_h(\rho_A, \rho_B) = \Big\{ (\rho_h, m_h) \in V^1_{n,h} \times V^0_{e,h} : \\ h \sum_{i=0}^{N-1} \left\langle \frac{\rho_h(t_{i+1}, \cdot) - \rho_h(t_i, \cdot)}{h} + \operatorname{div}_{\mathcal{X}} m_h(t_i, \cdot), \varphi_h(t_i, \cdot) \right\rangle_{\pi} = 0 \ \forall \varphi_h \in V^0_{n,h} \Big\}.$$

An example of such a weak solution is shown at a high level in Figure 5.2. The *Galerkin discretization of the metric* $\mathcal{W}$ is defined by the expression

$$\mathcal{W}_h^2(\rho_A, \rho_B) = \inf \left\{ \mathcal{E}_h(\rho, m) := \mathcal{A}(\rho, m) + \mathcal{I}_{\mathcal{CE}_h(\rho_A, \rho_B)}(\rho, m) \right\}.$$

The major analytic result of [ERSS20] is reproduced below, and in words asserts that taking $h \to 0$, one recovers the discrete transport metric $\mathcal{W}$.

Theorem 5.3.2 ([ERSS20], Theorem 3.6)**.** *Let ρ_A, ρ_B be fixed temporal boundary conditions. Then, the sequence of functionals $(\mathcal{E}_h)_h$ Γ-converges for $h \to 0^+$ to the functional $\mathcal{E}$ with respect to the weak* topology in $\mathcal{M}([0,1], \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X} \times \mathcal{X}})$.*

An appropriate inner product can be defined so that the set $H = V^1_{n,h} \times (V^0_{e,h})^4 \times (V^0_{n,h})^2$ is a Hilbert space. Then, well-chosen slack variables are introduced so that the problem of computing $\mathcal{I}_{\mathcal{CE}_h(\rho_A, \rho_B)}$ can be decomposed into several low-dimensional projection problems. A typical element of H looks like $v = (\rho, m, \theta, \rho_+, \rho_-, \bar{\rho}, q)$; here ρ, m are in some sense the actual objects of interest; in brief, we have ρ and m representing the rectification of the geodesic curve and its momentum, while the additional components $\theta, \rho_+, \rho_-, \bar{\rho}$, and q are the slack variables used for decomposing the projection problem. The components θ, ρ_+, and ρ_- are vector fields on the edge set, translating the mass on the nodes into mass averaged on the edges, while q and $\bar{\rho}$ encode the average mass on the nodes over a time interval. The Chambolle-Pock routine is carried out until the convergence threshold

$$\int_0^1 \|\rho_{k+1} - \rho_k\|_{\pi}^2 \, \mathrm{d}t < 10^{-10}.$$

The Chambolle-Pock algorithm, first introduced by its eponymous authors in [CP11], is a primal-dual algorithm for solving non-smooth convex optimization problems, a form into which the discrete transportation geodesic approximation problem may be cast. Specifically, if $F, G : H \to \mathbb{R}$ are convex functions on a Hilbert space, the Chambolle-Pock algorithm approximates the solution to

$$\min_{x \in H} F(x) + G(x).$$

It belongs to a broad class of extra-gradient methods [Kor76], such as the well-known Douglas-Rachford technique [LM79, CKCH23]. In order to apply the Chambolle-Pock routine to this problem, a set of slack variables is introduced to decouple certain non-linearities arising from the non-linear form of the action functional. Given convex functions $F, G : H \to \mathbb{R}$, the Chambolle-Pock routine requires that we be able to compute the proximal mappings of the G and the Fenchel conjugate of F (denoted F^*). Recall that for an arbitrary function f, its proximal mapping in H is defined by

$$\mathrm{prox}_f(x) := \operatorname*{argmin}_{y \in H} \frac{1}{2}\|x - y\|_H^2 + f(y).$$

Note that if $A \subset H$, and $\mathcal{I}_A : H \to \mathbb{R}$ is defined so that $\mathcal{I}_A(x) = 0$ for $x \in A$ and $+\infty$ otherwise, then

$$\mathrm{prox}_{\mathcal{I}_A}(x) = \mathrm{proj}_A(x).$$

This observation allows for the algorithm given in [ERSS20] to be expressed as the composition of many low-dimensional projection problems.

It is essential for our approach that this method of approximating the discretized $\mathcal{W}$ metric proceeds by approximating geodesic curves connecting measures in $\mathcal{P}(\mathcal{X})$. Because the optimizers output by Algorithm (4) include the momentum information associated to the geodesic curves, we are therefore able to approximate the intrinsic gradient of the discrete variance functional (2.15) at a point ν by approximating the geodesics between

Algorithm 4 Generic Chambolle-Pock Routine

Input: Convex function F, G; $\tau, \sigma > 0$ satisfying $\tau\sigma < 1$, $\lambda \in (0, 1]$; initialization $x^{(0)}, y^{(0)}, \bar{x}^{(0)} = x^{(0)}$
Output: x^* minimizing $F(x) + G(x)$
  while not converged **do**
    $y^{(k+1)} = \text{prox}_{\sigma F^*}(y^{(k)} + \sigma\bar{x}^{(k)})$
    $x^{(k+1)} = \text{prox}_{\tau G}(x^{(k)} - \tau\bar{y}^{(k+1)})$
    $\bar{x}^{(k+1)} = x^{(k+1)} + \lambda(x^{(k+1)} - x^{(k)})$
  end while
  return $x^{(k)}$

ν and each reference measure and summing the associated momenta, weighted according to the prescribed coordinates $\lambda \in \Delta^{p-1}$.

In [ERSS20], an implicit Euler scheme in the spirit of the construction given in [JKO98] and expanded upon in [AGS05], is constructed to compute gradient flows of the entropy for the discretized metric $\mathcal{W}_h$. It's classical formulation for a generic functional F is given by

$$\rho_k := \text{argmin}\left\{\frac{1}{2\tau}W_2^2(\rho, \rho_{k-1}) + F(\rho)\right\}.$$

This is sometimes referred to as a "minimizing movements" scheme, since the next iterate must minimize the sum of the objective functional and the W_2 distance from the current iterate. In [ERSS20], this works because the entropy functional, defined by the equation

$$H(\rho) = \sum_{x\in V} \rho(x)\log(\rho(x))\pi(x),$$

admits a Fenchel conjugate whose proximal mapping may be quickly approximated via the Moreau identity and a Newton scheme, and H^* can be folded neatly into the Chambolle-Pock objective function after certain modifications to the projection problems (see Section 6 of [ERSS20] for full details). The implicit scheme is desirable from an analytic perspective because it is well-studied and is known to generalize the notion of a gradient flow to a generic metric space lacking linear structure — however, it is not obvious how one

would implement it numerically, which restricts its use in the practitioner's setting. In an ideal world, a similar approach might be used for computing the gradient flow associated to (2.15), but computing the proximal mapping of (2.15) is itself as complicated as minimizing the functional itself to begin with. Therefore, the implicit Euler approach is not appropriate to the task, and so we instead proceed by an explicit approach.

5.3.3 Approximating the Exponential Map with the Continuity Equation

First, we outline the general steps of the algorithm. In words, to synthesize a barycenter for a given dictionary with prescribed weights, we fix some initialization measure ν_0, and then for $i = 1, 2, \ldots$, we approximate the $\mathcal{W}$ gradient at iterate ν_i and update the iterate using a heuristic based on the continuity equation. We stop iterating once the iterates converge in the $\|\cdot\|_\pi$ norm, corresponding to a zero of the gradient vector field $\nabla_{\mathcal{W}}\mathcal{J}$. In order to carry out the update step of the intrinsic gradient descent scheme, it is necessary to have a way of "stepping forward" on the surface of the manifold. Generically the ideal situation would be to have some expression of the exponential map in terms of local coordinates, but this is rarely the case. Here we propose an approximation of the exponential mapping on $\mathcal{P}(\mathcal{X})$ by infinitesimal reasoning, leveraging the structure of the continuity equation. This is a formal argument at best: all geodesics in this manifold obey the continuity equation, but not all solutions to the continuity equation are geodesic curves. Nevertheless, our idea is that provided we do not move too far away from the initial point, we will move forward on the manifold in a way which is roughly distance minimizing.

If (ρ, m) is a pair which satisfies the continuity equation, then we have

$$\lim_{\varepsilon \to 0} \frac{\rho(t+\varepsilon) - \rho(t)}{\varepsilon} = -\operatorname{div}_{\mathcal{X}}(m(t)).$$

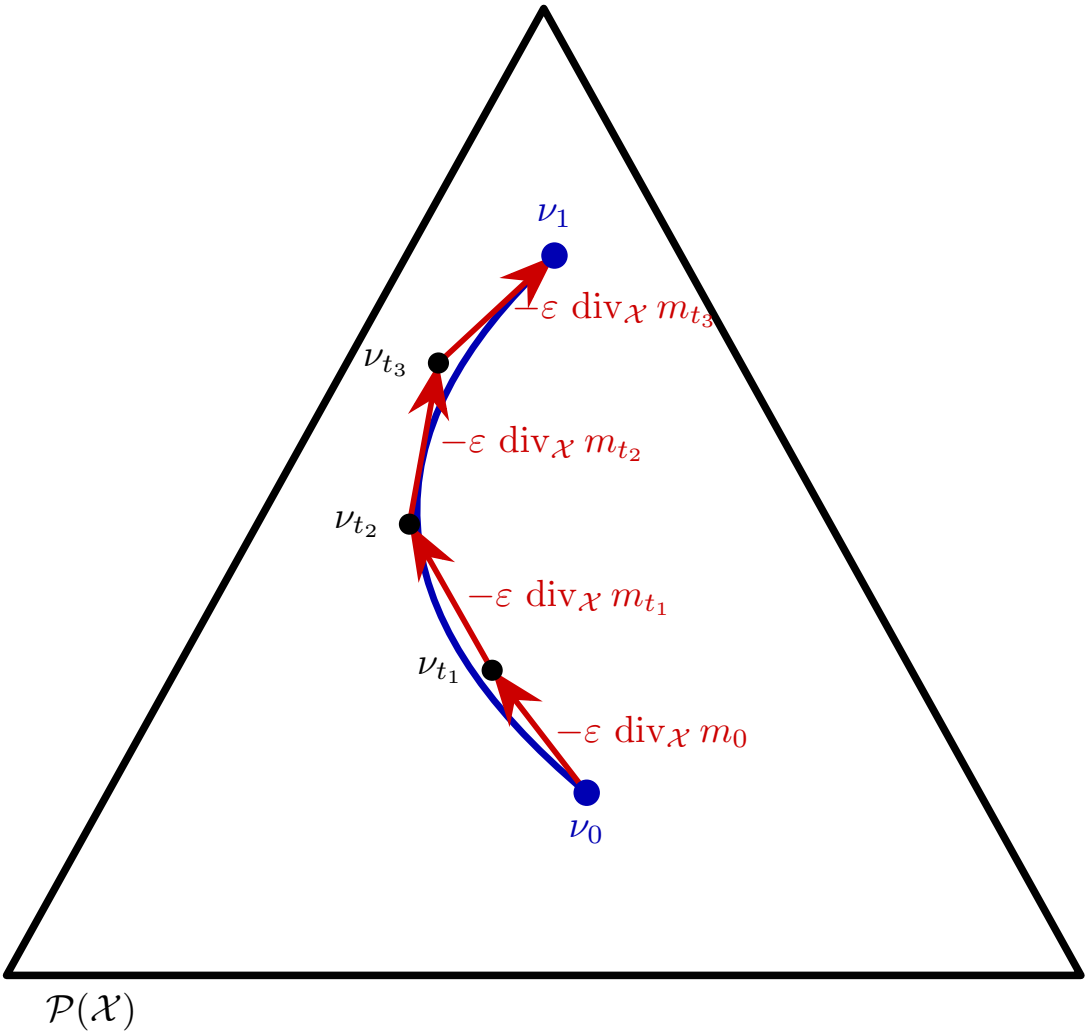


Figure 5.2: The Chambolle-Pock routine is used to approximate energy minimizing curves in $\mathcal{P}(\mathcal{X})$, ultimately producing a rectified curve encoded as a sequence of points and a collection of momentum vector fields. We leverage these vector fields to approximate the gradient of (2.15).

Rearranging, taking $\varepsilon > 0$ and evaluating at $t = 0$ we then write

$$\rho(\varepsilon) \approx \rho(0) - \varepsilon \operatorname{div}_{\mathcal{X}}(m(0)). \tag{5.1}$$

We will use (5.1) as our approximation of the exponential map. Applying this to (2.15), we obtain our update step:

$$\nu_{k+1} = \nu_k + t_k \operatorname{div}_{\mathcal{X}} \left(\sum_{i=1}^{p} \lambda_i m_{\nu,\nu_i}(0) \right).$$

Algorithm 5 Synthesis of Barycenters via Intrinsic Gradient Descent in the $\mathcal{W}$ Metric

Input: Markov chain (Q, π), reference measures $\{\nu_i\}_{i=1}^p$, weights $\lambda \in \Delta^{p-1}$, convergence criterion
Output: Barycenter $\nu = \mathrm{Bary}(\{\nu_i\}_{i=}^p), \lambda)$
  $\nu \leftarrow \nu_1$ ▷ Initialize with a measure which is trivially in $\mathrm{Bary}(\{\nu_i\}_{i=1}^p)$
  while not converged **do**
    $m \leftarrow 0 \in \mathbb{R}^{|V|\times|V|}$
    for $r = 1, \ldots, p$ **do**
      $m \leftarrow m + \lambda_r m_{\nu,\nu_r}(0)$ ▷ Compute the initial momentum of the geodesic $[\nu, \nu_r]$ and sum
    end for
    $\nu \leftarrow \nu + \varepsilon \operatorname{div}_{\mathcal{X}} m$ ▷ Update via continuity equation heuristic
  end while
  return ν

A natural concern here is whether or not (5.1) remains in the admissible set $\mathcal{P}(\mathcal{X})$. We show that this is almost the case with the following lemma.

Lemma 5.3.3. *Let $\nabla\varphi \in T_\nu \mathcal{P}(\mathcal{X})$. Then*

$$\sum_{x \in V} \operatorname{div}_{\mathcal{X}} \nabla\varphi(x)\pi(x) = 0.$$

Proof. The claim follows from the reversibility of the pair (Q, π). By definition of the

graph gradient, $\nabla\varphi$ is a skew-symmetric matrix. Therefore, we have

$$\begin{aligned}\sum_{x\in V} \operatorname{div}_{\mathcal{X}} \nabla\varphi(x)\pi(x) &= \sum_{x\in V} \frac{1}{2} \sum_{y\in V} Q(x,y)(\nabla\varphi(y,x) - \nabla\varphi(x,y))\pi(x), \\ &= \sum_{x,y\in V} \pi(x)Q(x,y)\nabla\varphi(y,x).\end{aligned}$$

Therefore we may interpret the graph divergence of $\nabla\varphi$ as an elementwise sum of a certain matrix $\hat{Q} \odot \nabla\varphi$, where $\hat{Q}$ is the matrix $\hat{Q} = [\pi(x)Q(x,y)]_{x,y\in\mathcal{X}}$ and $\odot$ is the Hadamard product for matrices (i.e., elementwise multiplication). By reversibility of the pair (Q,π), the matrix $\hat{Q}$ is symmetric, so its Hadamard product with $\nabla\varphi$ is skew-symmetric. Observing that the elementwise sum of any skew-symmetric matrix is 0 finishes the proof. □

Corollary 5.3.4. *Let $\nu \in \mathcal{P}(\mathcal{X})$. Then $\sum_{x\in V} (\nu(x) + \varepsilon \operatorname{div}_{\mathcal{X}} \nabla\varphi(x))\pi(x) = 1$.*

In other words, while it could in principle be possible to leave the probability simplex by violating the non-negativity constraint, the updates will remain on the hyperplane $\{\langle\rho, 1\rangle = 1\}$.

5.3.4 Approximation of the Exponential Map's Inverse

Recalling that the exponential mapping of tangent vector $v \in T_p\mathcal{M}$ is defined to be $\gamma_{p,v}(1)$, where $\gamma_{p,v}$ is a geodesic curve starting at p with initial velocity v and time domain $[0,1]$, it follows then that if we had an oracle which could produce for arbitrary $p, q \in \mathcal{M}$ a geodesic γ, then the inverse of the exponential mapping at p could be computed by evaluating $\frac{d}{dt}\gamma(t)|_{t=0}$. Fortunately this is precisely one of the outputs of Algorithm (4). Forming the linear system associated to the quadratic program in (2.9) is therefore as simple as approximating the geodesic between the target measures ν and the dictionary of reference measures $\{\nu_i\}$, computing the pairwise inner products of their initial momentum vectors in the space $T_\nu\mathcal{P}(\mathcal{X})$ and invoking an off-the-shelf quadratic programming solver.

Algorithm 6 Analysis of Barycenters via Quadratic Programming

Input: Markov chain (Q, π), target measure ν, reference measures $\{\nu_i\}_{i=1}^p$
Output: Coordinates λ^* such that $\nu = \text{Bary}(\{\nu_i\}_{i=}^p, \lambda^*)$

```
A ← 0 ∈ ℝ^{p×p}
for i = 1, ..., p do
    ∇φ_i ← m_{ν,ν_i}(0)          ▷ Compute the initial momentum of the geodesic [ν, ν_r]
end for
for i, j = 1, ..., p do
    A_{i,j} ← (1/2) Σ_{x,y∈𝒳} θ(ν(x), ν(y)) ∇φ_i(x, y) ∇φ_j(x, y)          ▷ Form the Gram matrix
end for
λ* ← min{λ^T A λ : λ ∈ Δ^{p−1}}          ▷ Solve via CVX, etc.
return λ*
```

5.3.5 Hyperparameters for the Synthesis and Analysis Algorithms

Several parameters need to be set when running these algorithms which are divorced from the relevant measures and weights themselves; Table 5.1 summarizes which hyperparameters are involved in which subroutines. First, there are two hyperparameters associated to the underlying Chambolle-Pock routine computation of action minimizing geodesics: the convergence tolerance, δ_g and the number of time steps in each geodesic, N. For the synthesis computation, there is also the descent step size ε, and the convergence threshold for the descent scheme δ_b.

The parameters δ_g and N control the quality of the approximated geodesics in the $\mathcal{W}$ geometry. As $\delta_g \to 0$ and $N \to \infty$, the outputs of the Chambolle-Pock routine become finer and finer, but at the cost of computational complexity — note that the subroutines (proximal map computations) have time complexity growing quadratically in N (see [ERSS20] for full details), and that in general Chambolle-Pock exhibits a rate of convergence $\mathcal{O}(1/k)$ in the number of iterations.

The descent parameter δ_b governs the threshold for convergence of the synthesis scheme, while ε controls the size of the step taken after each gradient computation. In principle, intrinsic gradient descent, under certain assumptions on the underlying manifold, exhibits $\mathcal{O}(1/k)$ rate of convergence, and an optimal descent step size can be chosen

for the constant step size paradigm as detailed in [ATV13], but such a choice requires knowledge of the injectivity radius of $(\Delta^{p-1}, \mathcal{W})$ and an upper bound on the eigenvalues of the Riemannian Hessian of the variance functional $D^2\mathcal{J}$. At present, these quantities are unstudied, and we leave investigation of them to future work; here we simply remark that we find that taking $\varepsilon \in (0.1, 0.5)$ generally results in reasonable rates of convergence. In recent years, interest has been taken in accelerating the rate of convergence for intrinsic gradient descent [AOBL21], taking inspiration from techniques well-known in Euclidean optimization [Pol87], and this may also be a useful direction in which to expand the present work.

Routine	δ_g	N	δ_b	ε
Chambolle-Pock	✓	✓		
Synthesis	✓	✓	✓	✓
Analysis	✓	✓		

Table 5.1: Table of hyperparameters associated to each algorithm

5.3.6 Barycentric Coding Model via Entropic Regularization

In [BPC16], an algorithm is proposed for performing transport-based histogram regression for measures on fixed support. For a given family of histograms $\{p_s\}_{s=1}^S$, consider the mapping

$$P : \lambda \mapsto \operatorname{argmin} \sum_s \lambda_s OT_{\varepsilon_{reg}}(p, p_s),$$

and seek solutions to the optimization problem

$$\operatorname{argmin} \mathcal{E}(\lambda) := \mathcal{L}(P(\lambda), q),$$

for prescribed reference histogram q, where $\mathcal{L}$ is some common and easily differentiated loss function (e.g., the squared L^2 difference, or the Kullback-Liebler divergence). While it is expensive to compute the Jacobian $[\partial P(\lambda)]^T$ needed for the naive descent scheme,

they show that replacing it by a running estimate $[\partial^{(L)} P(\lambda)]^T$ obtained after L Sinkhorn iterations makes the problem computationally tractable.

In [GM25], that approach was leveraged to implement an entropically regularized barycentric coding model for measures supported on graphs, using diffusion distances as ground metrics encoding the connectivity information of the graph. There are trade-offs associated to this approach to barycenters of measures on graphs. Firstly, there is the question of how to choose the cost matrix C. As alluded to previously, there is not an isomorphism between pairwise distance matrices and graph structures — it is perfectly possible to have two identical cost matrices which represent different network structures. Secondly, the Bregman projections algorithm, which depends on the entropic regularization, both in regularizing parameter ε_{reg} and choice of reference measure (typically taken to be the tensor product of the reference measures) is sensitive to the choice of regularization parameter ε_{reg}, with instability of the computations occurring in the limit as $\varepsilon_{reg} \to 0$. Strictly speaking, one could choose to forgo the regularization of the transport cost and try solving the linear program directly, but this is not without issues of its own. One such issue is that solving linear programs is expensive, but another issue is that even if we are happy to pay the cost of solving an expensive linear program, a fundamental problem remains in that the geometry induced by the Kantorovich problem does not admit non-trivial geodesics [Maa11].

In practice this trade-off may be worth it, since synthesizing barycenters via the Sinkhorn algorithm is much faster than the gradient descent approach outlined above. However, it is worth emphasizing that the introduction of the entropic regularization term introduces bias and in fact computation of entropy-regularized transport cost can be interpreted as performing a maximum-likelihood estimation for Gaussian deconvolution [RW18] — in short, it causes a loss in geometric information which we see realized as a kind of "blurring" of the support of the measures across the entire domain.

Summarizing then, we are comparing two approaches which could be taken to the

problem of synthesizing and analyzing barycenters of measures on graphs: one is the "static" approach born of the Kantorovich problem, and the other is the "dynamic" approach born of the transport metric proposed in [Maa11]. The former offers the following trade-off: if we are willing to make a non-canonical choice about how to encode the connectivity of a given graph and choose a workable loss function, and as well introduce the entropic regularization, then a combination of Bregman projections, the Sinkhorn-Knopp algorithm, and gradient descent can achieve system for synthesizing and analyzing barycenters of measures on graphs. The latter requires us on the other hand to set certain tolerances for convergence, but otherwise allows us to construct a barycentric coding model using a canonical representation of the graph via its Markov chain random walk transition matrix.

5.4 Experiments

Here we present some experimental results based on our numerical implementation of the outlined scheme above and highlight our major observations. Broadly speaking, we find that our intrinsic descent scheme is able to synthesize and analyze barycenters of measures on graphs. In Figures 5.3 and 5.4 we demonstrate the general capability of our program. Figure 5.3 features the barycenter of 4 measures on a graph with 16 nodes, and for both of the weighted and unweighted edge cases. Figure 5.4 features two barycenters of measures defined on a spatial geography graph where each vertex corresponds a district for the Massachusetts state house, with edges indicating that two districts share a border (in total, there are 156 vertices in the graph and 404 edges); the two barycenters share the same references and prescribed coordinates, but differ in terms of their overall convergence threshold δ_b.

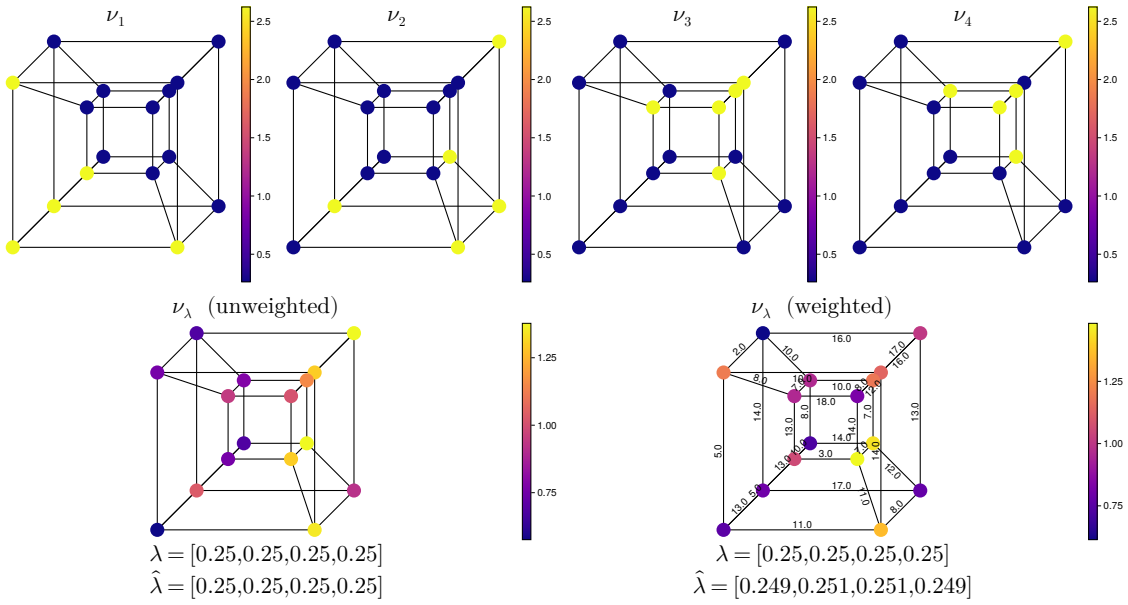


Figure 5.3: Barycenter of 4 measures on the graph formed by the edges and vertices of a hypercube. Bottom left: edges of the hypercube are uniformly weighted with $\omega_{i,j} \equiv 1$; bottom right: edges are randomly weighted (weights are symmetric to preserve the reversibility condition of the Markov chain).

5.4.1 Comparison: Recovering Geodesic Curves via Wasserstein Gradient Descent

The first test we can conduct to verify our algorithm's outputs is to consider the case of 2 reference measures. As pointed out in [AC10], if we fix reference measures ν_0, ν_1 and denote by $\nu(t) : [0,1] \to \mathcal{P}(\mathcal{X})$ the geodesic connecting ν_0 to ν_1, with $\nu(0) = \nu_0$, then a consequence of the triangle inequality is that

$$\nu(t) = \operatorname*{argmin}_{\nu} \left\{ (1-t)\mathcal{W}^2(\nu_0, \nu) + t\mathcal{W}^2(\nu_1, \nu) \right\}.$$

This implies that computing a barycenter of two measures requires no additional bells and whistles beyond a mechanism for computing the geodesics between them.

We will proceed in two distinct ways for this procedure. First, we note that in the case of a two-point space $\mathcal{X}_2 = \{a, b\}$, there exists a simple ODE governing the geodesic

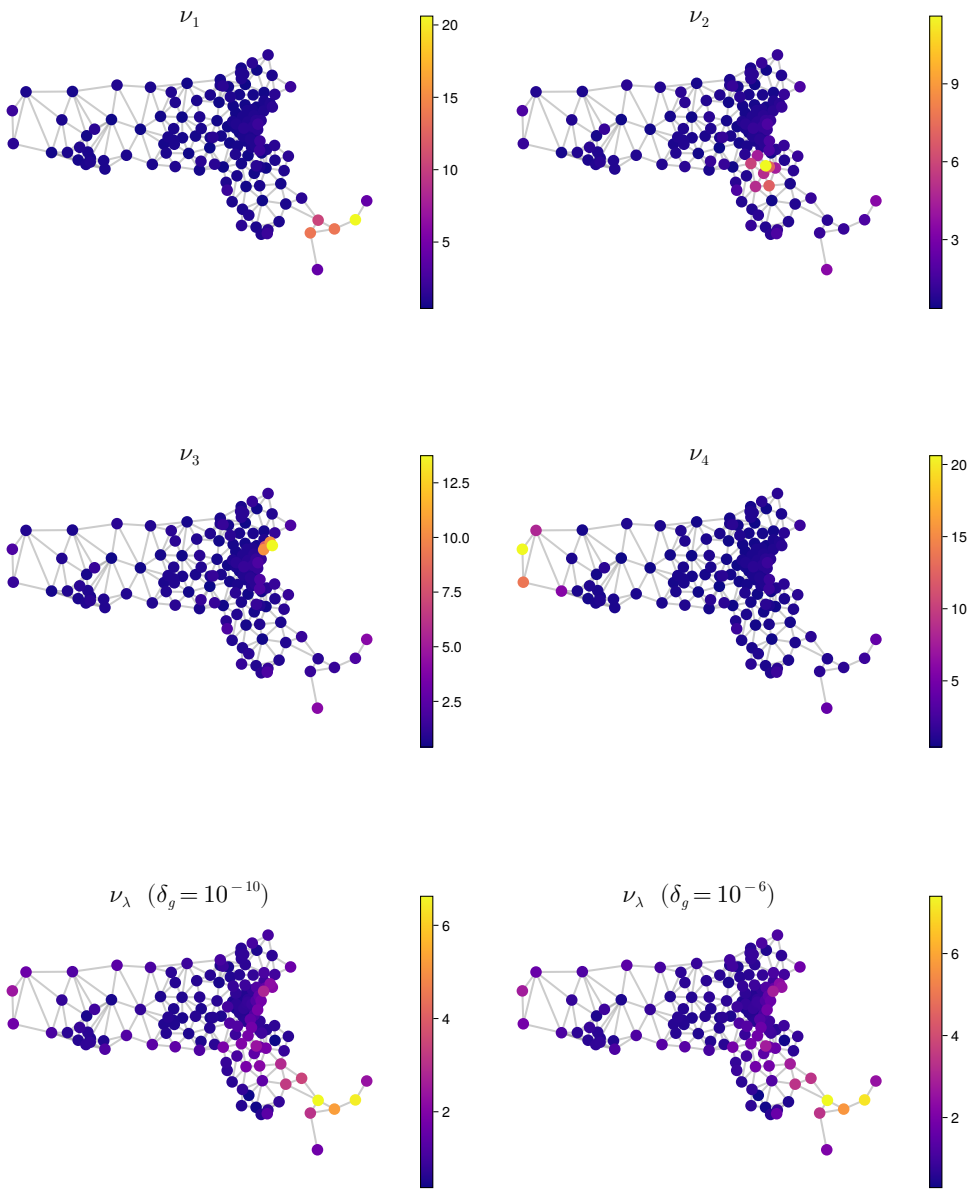
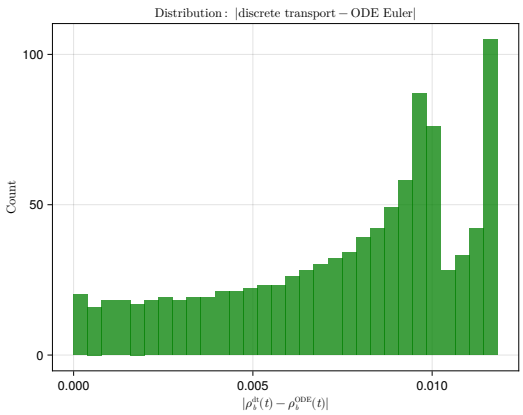
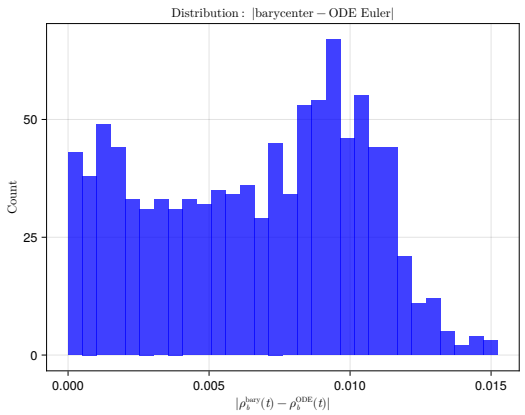


Figure 5.4: Barycenter of 4 measures on the graph formed by the spatial geography node constructed from the Massachusetts state house districts. We randomly generate 4 geographically concentrated reference measures and compute the barycenter with weights $\lambda = [0.4, 0.3, 0.2, 0.1]$, for convergence tolerances $\delta_b = 10^{-6}, 10^{-10}$, then recover the coordinates. We observe that more rigorous convergence tolerance in the Chambolle-Pock routine leads to superior coordinate recovery in the analysis problem. Data retrieved from https://www.mass.gov/info-details/massgis-data-massachusetts-house-legislative-districts-2021.



(a) Distribution of deviation between Chambolle-Pock and Euler methods for approximating mass on node b across geodesic between Dirac masses on a two point graph.



(b) Distribution of deviation between intrinsic descent and Euler methods for approximating mass on node b across geodesic between Dirac masses on a two point graph.

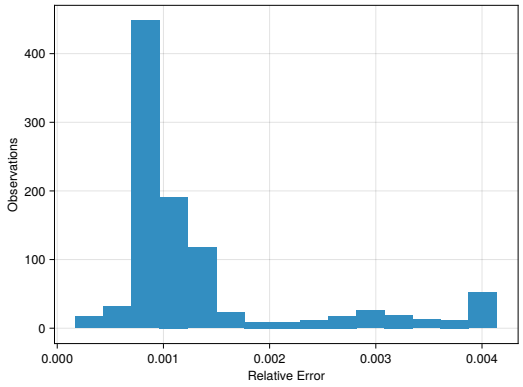
Figure 5.5: Computation of barycenters on a 2 node graph
Comparison of barycenters on a 2 node graph computed by solving an ODE vs computed by intrinsic gradient descent.

curve between any given pair of measures $\nu_0, \nu_1 \in \mathcal{P}(\mathcal{X}_2)$. We can solve the ODE via an explicit Euler method, and compare the geodesic predicted by the solution of the ODE to the barycenters produced by our descent method; see Figure 5.5 for the distribution of results.

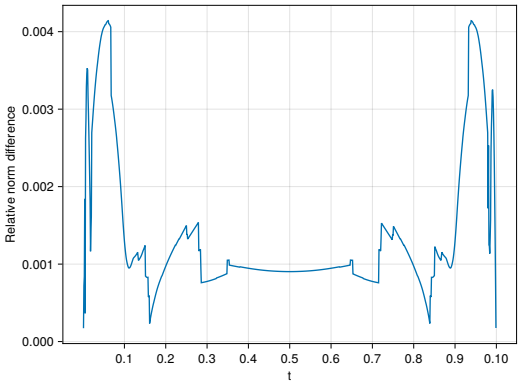
For the other approach to this experiment, we can on the one hand compute the geodesic $\nu(t)$ between ν_0, ν_1 using the Chambolle-Pock algorithm, evaluate that geodesic at different values of $t = t_0, t_1, \ldots$, and see how those evaluations compare to the barycenters obtained by, on the other hand, applying our descent method to the reference measures $R = \{\nu_0, \nu_1\}$ with weights $\lambda = (1 - t_0, t_0), \ldots, (1 - t_N, t_N)$. We perform two versions of this approach to the experiment. First, we fix a single pair of measures $\{\nu_0, \nu_1\}$ on a graph G, compute the points on the geodesic $\{\nu_{t_i}\}_{i=0}^{N}$ between them with $N = 100$ time steps, and then for each $i = 1, \ldots, N-1$, compute the barycenter $\operatorname{argmin}\{(1 - i/N)\mathcal{W}^2(\nu, \nu_0) + (i/N)\mathcal{W}^2(\nu, \nu_1)\}$. We treat the measures ν_{t_i} output by the Chambolle-Pock computation of the geodesic curve between ν_0 and ν_1 as the ground truth and report the relative error of the barycenters computed via intrinsic gradient descent;

the results of this experiment are plotted in Figures 5.6a and 5.6b, both as an ensemble distribution and as a function of the interpolation parameter t. Then we perform a variation on the experiment, keeping G fixed but now letting the measures and coordinates be randomly generated each time, see Figure 5.7. In order to generate the random measures on the triangle, we sample from the uniform distribution on the interval $[0, 1]$ three times, and normalize by dividing through by the sum of values, and then componentwise by the steady state vector π.

We also hasten to point out that taking the measures along the computed geodesic as ground truth it not necessarily optimal. Note that the object being optimized by the Chambolle-Pock routine is a representation of the entire geodesic. While in the limit of $N \to \infty$ we would expect perfect correspondence between points on the geodesic and barycenters of ν_0 and ν_1, in the case of finite step size we must anticipate some difference between the true barycenters and the measures along the approximated curve. Nevertheless, the small relative error in our experiments indicates that the intrinsic descent scheme does converge towards a variance minimizing measure.



(a) Distribution of 1000 computed relative norm differences between (a) the barycenter of two fixed measures ν_0, ν_1 on a triangle graph and (b) the corresponding point on the geodesic between ν_0 and ν_1.



(b) 1000 computed relative norm differences between (a) the barycenter of two fixed measures ν_0, ν_1 on a triangle graph and (b) the corresponding point on the geodesic between ν_0 and ν_1 as a function of the interpolation parameter t.

Figure 5.6: Comparison barycenters of two measures computed as geodesics via Chambolle-Pock and via intrinsic gradient descent for a fixed geodesic curve.

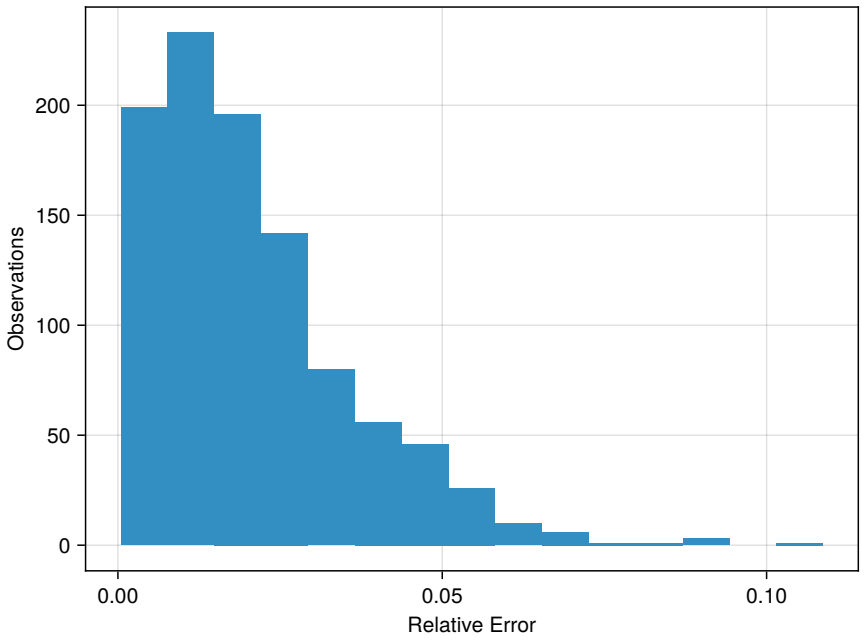


Figure 5.7: Distribution of 1000 computed relative norm differences between (a) the barycenter ν_b of two randomly generated measures ν_0^i, ν_1^i on a triangle graph and (b) the corresponding point on the geodesic between ν_0^i and ν_1^i, ν_g. When computing the relative error we assume that the point on the geodesic is a genuine representative of the barycenter, and therefore compute $\|\nu_g - \nu_b\|_\pi / \|\nu_g\|_\pi$

5.4.2 Comparison: Barycenters Obtained by Gradient Descent vs by Entropic Regularization

Here we compare qualitatively the barycenters obtained by our dynamic approach to those obtained via the entropically regularized static approach. We will compare our algorithm's outputs to those obtained when the cost matrix used in formulating the entropic regularization problem is (1) the squared shortest path metric on the graph and (2) the squared diffusion distance between graph vertices for various values $t > 0$.

For our first comparison between the methods, we fix a grid graph on 49 vertices, generate 3 random measures and coordinates $\lambda = (0.5, 0.3, 0.2)$, and then compute the barycenter ν_λ (1) by intrinsic gradient descent, (2) by entropic regularization with the ground metric given by the squared shortest path metric and (3) by entropic regularization with the squared diffusion distance at time $t = 12$, shown in Figure 5.8a. For our

second comparison, we repeat the process outlined above on a graph representing the 48 contiguous U.S. States + D.C, shown in Figure 5.8b, this time taking $t = 11$ for the diffusion distance. Diffusion times in these experiments are selected by setting t according to the diameter of the graph, i.e., the maximal path length among all shortest paths between pairs of vertices.

In order to generate the measures which have some semblance of geographic concentration we perform the following procedure. First, we pick an arbitrary node $v \in V$, and let $N(v)$ be the set of vertices connected to v, together with v. Then we generate a weighting vector, for a set weight parameter $w > 0$, and assign mass 1 to $u \notin N(v)$ and mass w to $u \in N(v)$. Finally we normalize the resulting vector so that it lies in $\mathcal{P}(\mathcal{X})$, which is done by dividing componentwise by the steady state vector π and the total unnormalized mass $|N(v)^c| + w|N(v)|$. The parameter w roughly controls how close the measures are to the boundary of $\mathcal{P}(\mathcal{X})$. In these specific experiments, we take $w = 100$.

Some commonalities jump out from these experiments. For brevity, let us refer to the barycenter computed with Algorithm 5 as ν_{dyn}, the one computed via entropic regularization with cost given by the squared shortest path metric as $\nu_{st,sp}$, and the one computed via entropic regularization with cost given by the squared diffusion distance as $\nu_{st,diff}$.

First, we look at the results for the grid graph. Qualitatively, all three barycenters have supported concentrated in the lower left corner of the graph, but none of them are actually similar to each other. ν_{dyn} has support which extends along the off-diagonal of the grid, while $\nu_{st,sp}$ has support mostly concentrated below the main diagonal and parallel to it, and $\nu_{st,diff}$ is significantly more concentrated than either of the others. In terms of recovery error, $\nu_{st,sp}$ dominates with a relative error of approximately 2.43×10^{-7}, while ν_{dyn} and $\nu_{st,diff}$ both admit relative recovery errors of approximately 0.006.

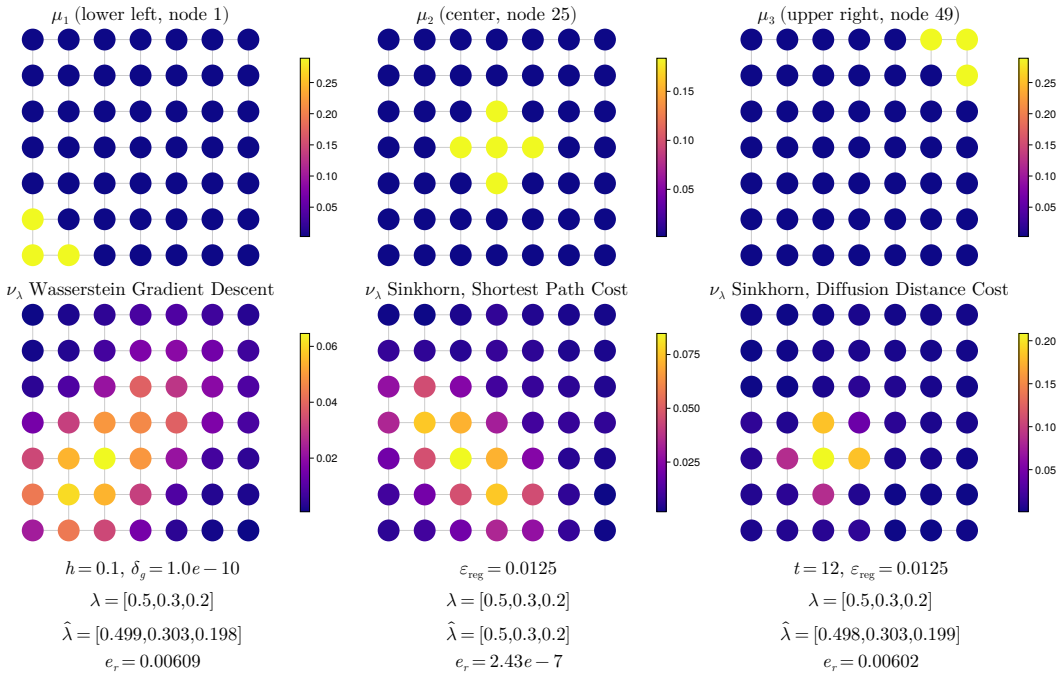
Turning our attention to the U.S. states graph, we see that $\nu_{st,diff}$ significantly improves its relative recovery error to roughly match the relative error obtained for $\nu_{st,sp}$, the former coming in at 6.02×10^{-8} and the latter at 5.24×10^{-8}. Recovery is not much

changed for ν_{dyn} on the graph, still at 0.006. Also of note in regards to the U.S. states graph, the barycenters computed via the static approaches are much more visually similar to each other than they are in the case of the grid.

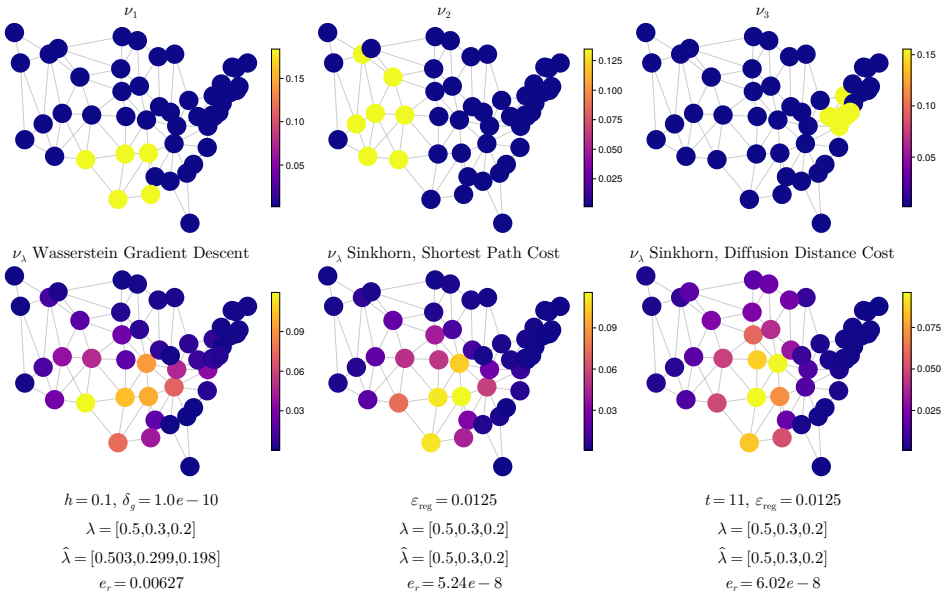
These experiments highlight the trade-offs between approaching the barycentric coding model from either the static or dynamic point of view. On the one hand, results can be obtained quickly using the static formulation with entropic regularization, but at the cost of having to make a good choice about what ground metric to use, while on the other hand, consistent results can be obtained without having to make such a choice, at the cost of significantly more compute time when using the dynamic formulation.

5.4.3 Effect of Different Initializations for the Descent Scheme

Lacking guarantees about the geodesic convexity of (2.15), or its strict convexity with respect to the Euclidean geometry, a natural question to ask is whether or not the choice of initialization ν_0 in the descent scheme (5.1) has a meaningful effect on the convergence behavior of the algorithm, and whether or not it becomes stuck in local minima. To test this, we generate a collection of arbitrary measures on a graph of 25 vertices, as well as a spatial geography data graph representing the 48 contiguous states of the U.S., plus D.C., and initiate the descent from each of the reference measures. We make this choice because, as prescribed in [ATV13], it is generally best practice to choose an initialization we know to be within the geodesic hull of the dictionary of reference measures. Results are displayed in Figures 5.9a and 5.9b. Once the algorithm converges for each choice of initialization, we compare the distance between the output barycenters in the $\|\cdot\|_\pi$ norm. Barycenters are considered to be converged at a threshold of $\delta_b = 10^{-10}$; for the experiment on the grid, the maximal norm difference between computed barycenters was observed to be 10^{-9}, while on the graph of the contiguous states the maximal observed norm difference was 8×10^{-10}, which indicates that the choice of initializing measure for the descent scheme may not have a meaningful impact on the final output of the



(a) Barycenter of three locally concentrated measures on a grid graph on 49 vertices.



(b) Barycenter of three geographically concentrated measures on a graph representing the 48 contiguous states + DC.

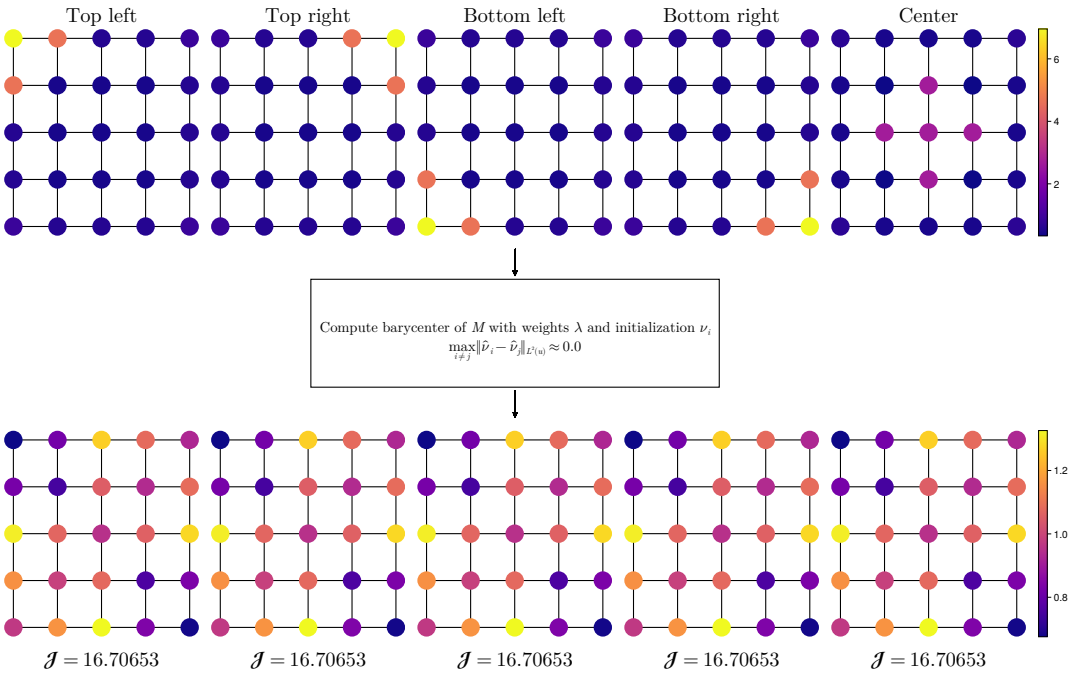
Figure 5.8: Barycenters of three geographically concentrated measures on two different graphs Top left, middle, right: reference measures; bottom left: computed via intrinsic gradient descent and the dynamic formulation of discrete transport; bottom center: computed via entropic regularization with cost given by the squared shortest path metric; bottom right: computed via entropically regularized transport, with the cost matrix given by the squared diffusion distance with t given by the diameter of the graph.

algorithm. This also suggests a modest computational optimization to make: initiating descent from the dictionary atom $\nu_{i^*} \in \{\nu_i\}_{i=1}^p$, where

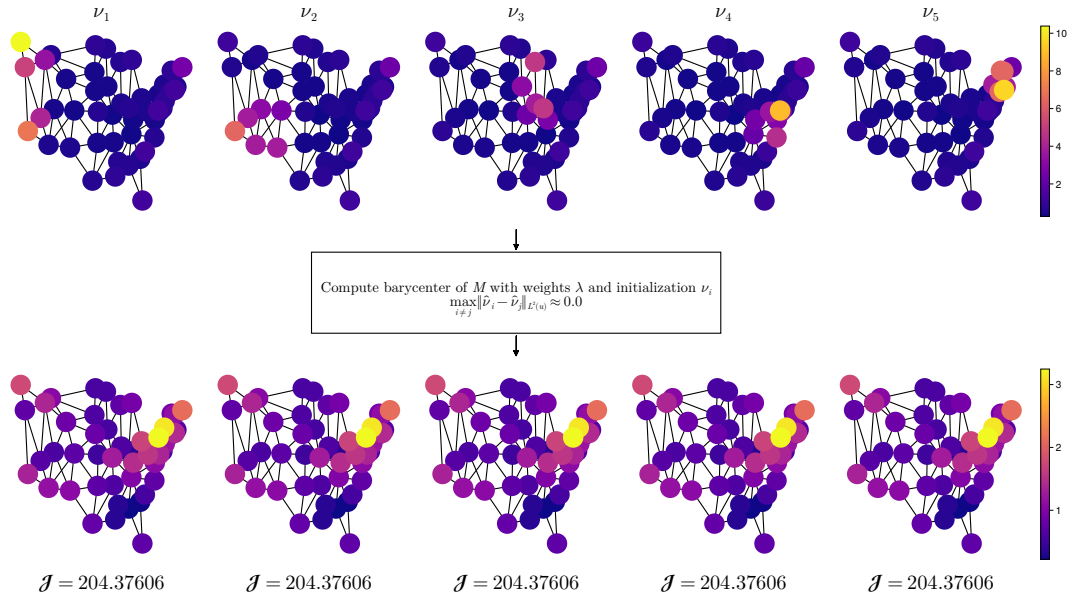
$$i^* = \operatorname*{argmin}_i \{\mathcal{J}[\nu_i]\}.$$

5.4.4 Recovery of Coordinates and Effect of Hyperparameters on Recovery Quality

Perhaps the single most desirable property this tool could have is consistency. It should be the case that if we fix $\lambda \in \Delta^{p-1}$ and a family of reference measures $R = \{\nu_i\}_{i=1}^p$, then compute their corresponding barycenter ν^*, and finally analyze the computed ν^* with respect to the family of reference measures R to obtain coordinates λ^*, then it should be the case that $\lambda \approx \lambda^*$. We test the consistency of our algorithm in two ways. For the first experiment we fixed a grid graph on 16 nodes, generated 3 random measures, a random set of coordinates $\lambda \in \Delta^2$, and computed the relative error of the recovered coordinates λ^*, taking $\delta_g = 10^{-9}, N = 10, \varepsilon = 0.5, \delta_b = 10^{-10}$. We find that a large majority of experiments result in < 0.01 relative error, and refer to 5.10 for the full distribution of relative error in coordinate recovery. For the second experiment, we modulate the hyperparameters associated to the Chambolle-Pock routine to test the effect of geodesic quality on coordinate recovery. For this experiment, we fix a single graph on 64 nodes, three measures ν_1, ν_2, ν_3 and set of coordinates $\lambda \in \Delta^2$, then vary the hyperparameters δ_g and N to test their effect on coordinate recovery. Our results are optimistic: even taking very coarse parameters, we see the relative error stays under 0.01 (see 5.11 for details). This indicates that the computational cost of synthesizing barycenters might be significantly reduced by a judicious choice of hyperparameters.



(a) A barycenter of measures on a grid graph for different choices of initialization. The maximum observed difference in $\|\cdot\|_\pi$ between the computed barycenters was 1×10^{-9}, the parameters used to compute the barycenters were $N = 16, \delta_g = 10^{-6}, \varepsilon = 0.25, \delta_b = 10^{-10}$.



(b) A barycenter of measures on the States + D.C. graph for different choices of initialization. The maximum observed difference in $\|\cdot\|_\pi$ between the computed barycenters was 8×10^{-10}, the parameters used to compute the barycenters were $N = 5, \delta_g = 10^{-6}, \varepsilon = 0.25, \delta_b = 10^{-10}$.

Figure 5.9: A barycenter of several reference measures computed with differing initializations for the descent scheme.

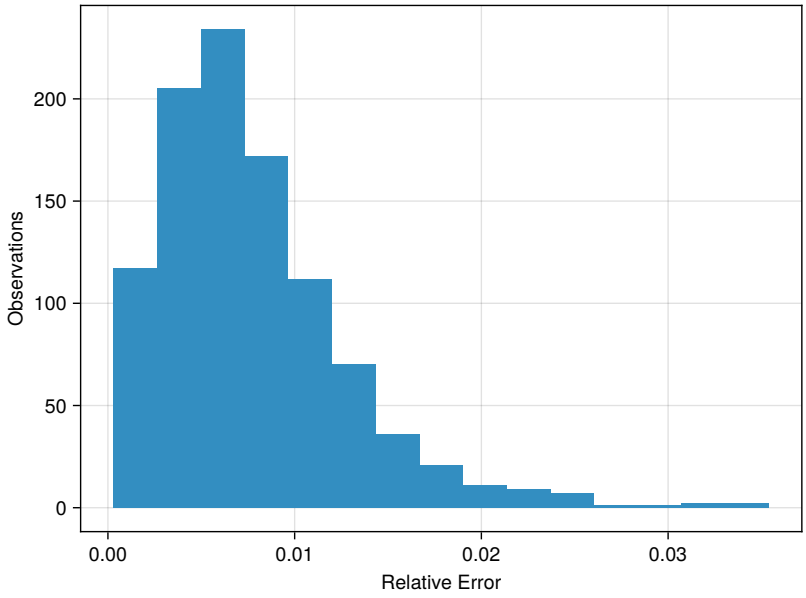


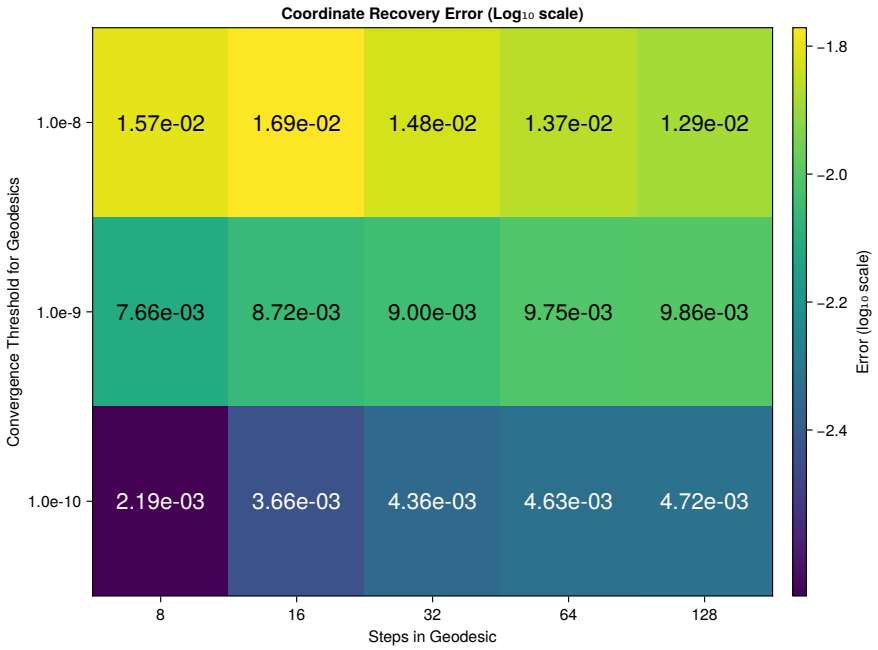
Figure 5.10: Distribution of 1000 observed relative errors in recovery of coordinates from a barycenter synthesized via intrinsic gradient descent for a grid graph on 16 nodes, 3 random measures, a random set of coordinates $\lambda \in \Delta^2$, with hyperparameters $\delta_g = 10^{-9}, N = 10, \varepsilon = 0.5, \delta_b = 10^{-10}$.

5.4.5 Empirical Convergence

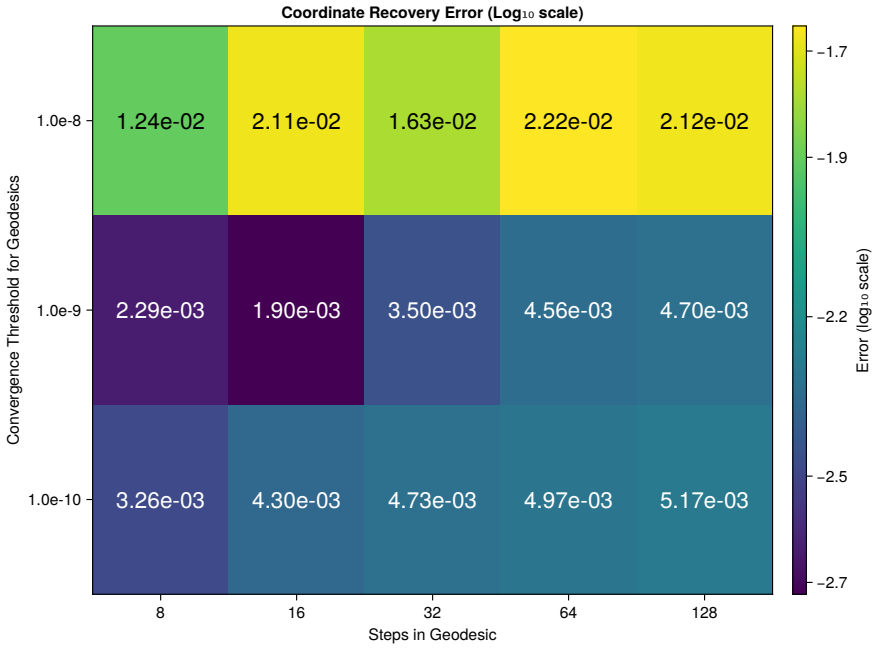
Lacking analytical information about either the injectivity radius or the sectional curvatures of the space $\mathcal{P}(\mathcal{X})$, we cannot make guarantees about the rate of convergence for the intrinsic gradient descent scheme [ATV13], which is unfortunate. However, we are able to study the rate of convergence for our algorithm empirically. For this experiment, we fixed a grid graph on 16 vertices, and for each run of the experiment, generated 3 random measures and a random set of barycentric coordinates λ; then we tracked the difference in norm between iterates as the descent scheme proceeded. In other words, we track the quantity

$$\|\nu_{k+1} - \nu_k\|_\pi = h\| \operatorname{div}_\mathcal{X} \nabla J|_{\nu_k}\|_\pi.$$

We repeated this procedure 100 times, with a convergence threshold $\delta_b = 10^{-10}$, and plotted the results. We observe clear linear convergence behavior in figure 5.12.



(a) Recovery on a graph defined by the 48 contiguous states + D.C.



(b) Recovery on a grid graph with 64 vertices.

Figure 5.11: Error matrices for coordinate recovery experiment. Three geographically concentrated measures ν_1, ν_2, ν_3 are generated, along with randomly chosen coordinates λ. The barycenter ν_λ is computed via intrinsic gradient descent, and then analyzed by solving the associated quadratic program. The relative error of the recovered coordinates is recorded as a function of the convergence threshold for the geodesic computations and the number of steps in each geodesic. Convergence threshold for the barycenters is set to $\delta_g = 10^{-8}$, and stepsize is taken to be $\varepsilon = 0.05$.

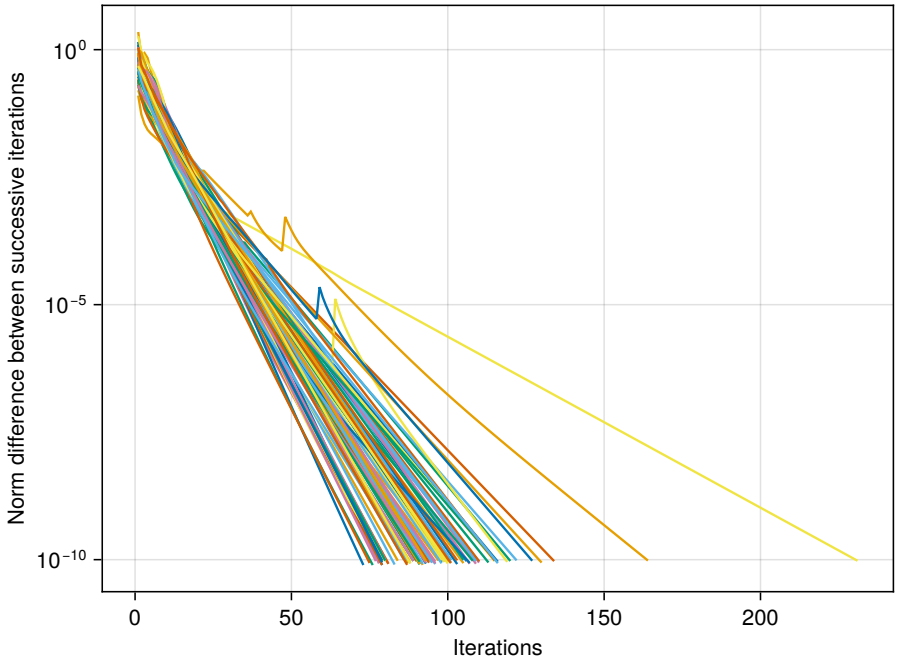


Figure 5.12: Log-linear scale of the absolute ℓ^2 norm difference between consecutive iterates and the number of iterations, or equivalently, the norm of the $\mathcal{W}$ gradient at each iteration. Hyperparameters for this experiment: $\varepsilon = 0.25$, $\delta_b = 10^{-10}$, $N = 10$, $\delta_g = 10^{-9}$.

Acknowledgments: This work was partially supported by the National Science Foundation through grants DMS-2309519 and DMS-2318894.

Chapter 6

Conclusions and Closing Thoughts

"... there are no hard and fast rules; you just win your battles."

— Isshin Ashina

6.1 Future Directions

We have seen now some nice applications of optimal transport to signal processing problems where the domain has some interesting spatial features. We have studied the change point detection problem and developed a barycentric coding model for measures supported on graphs, both using the entropically regularized transport cost via Bregman projections as well as by using intrinsic gradient descent in transport-like metric based on the continuity equation. We have made our code available for all of these experiments and hope that other interested scholars will find use of it in some way. So now, I would like to use the remaining space to address some ideas which appeared in the course of the development of this dissertation but were not afforded the time to be fully investigated.

As mentioned in Chapter 3, there are certain tradeoffs to be made when choosing the w and q parameters in the change point detection algorithm. For instance, taking w too small may result in a metric derivative approximation which is not faithful to the underlying data, while taking w too large may prevent the method from capturing states which are persistent over the course of the time series but whose individual instances do not last long enough to be captured by the large window. A potential remedy for this would be the introduction of a score function: by producing a model of the series, and cross-validating its predictive quality against a hidden portion of the data set, we could identify an optimal parameter set and relieve ourselves of the need to tune w and q by hand.

Interesting theoretical questions concerning analysis and computation of barycenters of measures supported on graphs abound. The overarching question that requires an answer is this: *under what conditions can the gradient descent iterates defined by* (5.1) *be guaranteed to converge?* Progress has been made on this question in the case of transport on a Euclidean domain, although it is not clear that those results have natural analogs in the discrete setting [CMRS20] — investigating such a potential parallel is therefore a natural next step for building up the theory of synthesizing barycenters of measures

on graphs. Another approach of interest is to approach from the Riemannian point of view: estimate the sectional curvatures of the probability simplex under the transport metric [Li22], and invoke relevant results from the general theory of Riemannian gradient descent [ATV13].

Here we will list some specific suggestions about further study barycenters of measures on graphs.

1. Although Lemma 5.3.3 shows that the gradient descent iterates in the synthesis algorithm remain signed measures of unit mass with respect to the stationary distribution π, ultimately we would like to find sufficient conditions guaranteeing that the iterates are always non-negative probability measures.

2. While it is clear that barycenters exist by virtue of the convexity of the variance functional with respect to linear interpolation, it is *not* clear that the variance functional is itself geodesically convex. In the case of measures supported on a Euclidean domain, the answer to the corresponding question is known, and in fact it is known to not be geodesically convex — nevertheless, gradient descent in the W_2 metric has been shown to be a legitimate tool for approximation of W_2 barycenters under suitable conditions [CMRS20]. It is of great interest to decide if variance functional is geodesically convex in our discrete setting.

3. Can we make meaningful improvements to the algorithm described in [ERSS20]?

4. Given the popularity of entropic regularization in the computational approximation of Euclidean transport, can such a regularization be found for discrete transport distances?

5. The step-size in the intrinsic gradient descent scheme proposed in this paper is taken as a constant for simplicity, but adaptive step sizes often have computational benefits in achieving faster rates of convergence [ATV13, CP11] — could such a scheme be appropriate and beneficial in this setting?

6. Alternatively, changing nothing at all about the way that we are currently approaching the problem of synthesizing and analyzing barycenters, there are already hyperparameters which have an unclear effect on the quality of our final outputs, namely the number of steps used to compute discrete geodesics, the convergence tolerance for the Chambolle-Pock routine, and the convergence tolerance for the intrinsic gradient descent scheme. It is entirely conceivable that a high tuning any of these knobs e.g., taking the number of steps per geodesic N to be large, will exhibit diminishing returns.

7. It has been shown in [EMW19] that the geodesics of the $\mathcal{W}$ metric may exhibit non-locality if certain topological properties are not satisfied. It would be of great interest to see if similar results can be extended to $\mathcal{W}$ barycenters — i.e., if $\{\nu_i\}_i$ is a collection of measures supported on a subset $A \subset \mathcal{X}$, when can it be guaranteed that the barycenter $\text{Bary}(\{\nu_i\}_i)$ is also supported entirely in A?

8. We have totally avoided consideration of the boundary of the simplex in our discussion, and it is known that geodesic curves in the $\mathcal{W}$ geometry may intersect the boundary of $\mathcal{P}(\mathcal{X})$ even when both end points lie in the interior of $\mathcal{P}(\mathcal{X})$ [GLM19].

9. We may also be able to leverage a "warm start" to enhance the speed of our algorithm. Fix a family of reference measures $\{\nu_i\}_{i=1}^p$ and let $\nu^{(k)}$ be the k-th iterate of the descent scheme for the synthesis of $\text{Bary}(\{\nu_i\}, \lambda)$ for some arbitrary λ. Denote by $\rho_i^{(k)}$ the (rectified) geodesic curves computed by the Chambolle-Pock routine which connect $\nu^{(k)}$ to ν_i. Presuming that $\|\nu^{(k+1)} - \nu^{(k)}\|_\pi$ is small, it seems reasonable to conclude that $\|\rho_i^{(k+1)} - \rho_i^{(k)}\|_H$ will also be small. In that case, converged primal and dual variables of the Chambolle-Pock routine obtained on the k-th iteration of the synthesis descent scheme may prove to be suitable initializations for the $k+1$-th iteration, and yield faster convergence for our algorithm.

6.2 Postscript, or Making Meaning

I've now said what I have to say about the mathematical contents of this dissertation, and in closing I find myself returning to the question I waved off in the introduction: why do all this? I have spent the better part of the last 7 years at Tufts University, and I have produced a dissertation "in partial fulfillment of the requirements for the degree". Being an applied mathematics dissertation I might clear my throat and mutter something about the importance of data analysis in computational sciences, but that would amount to dissembling in my view and even be dishonest. That sort of answer might gesture at a question like "why should a practitioner compute barycenters of measures on graphs via intrinsic gradient descent", but that isn't the question that hangs over a journey like this. The question is "why should I as a graduate student have spent all this time trying to compute barycenters in this way?" In other words, how do I make meaning of this experience, beyond the simple material consequences of its end? Mathematics is good for a great many things, but I have never found it to be much help in the development of spiritual understanding. And for that, I turn now to Hegel.

In Chapter B., "Self-Consciousness" of his *Phenomenology of Spirit*, Hegel relates his allegory of the lord and the bondsman[1]. He begins with the assertion that self-consciousnesses is impossible in isolation, that "[it] exists only in being acknowledged", and that therefore our human spirit cannot be separated from its relation to other persons. In Kojève's reading of the allegory, History begins when the need for mutual recognition leads to conflict. There are surely entire dissertations in their own right to be written just about this passage in the *Phenomenology*, but to my mind it makes 3 essential points: (1) recognition is necessary for self-realization, (2) the domination of other persons for the sake of achieving recognition is a futile endeavor, and (3) the fruits of labor transmute struggle into material reality. Two self-conscious beings, desiring recognition, risk their

[1]My understanding and interpretation is informed by and indebted to Alexandre Kojève's *Lectures* [KQBN69], as well as the translation and summary of the original by A.V. Miller.

lives and engage in a conflict which itself transforms the self-conscious beings engaged in it, and renders them as the lord and the bondsman, the dominator and dominated. The struggle cannot be resolved by death — for death does not preserve the struggle. The nature and outcome of the conflict, and the relationship between lord and bondsman in turn affects how each principal is able to *make meaning*; in reducing another to a lower caste of humanity, the lord ceases to see his interlocutor as essentially self-conscious, and because he deprives himself of a peer, he destroys the source of recognition which he desired in the first place. Simultaneously, the labor imposed on the bondsman becomes a material outlet preserving his struggle, contrasting with the fleeting nature of domination.

The pursuit of graduate studies is, by construction, a struggle for recognition. After years of labor, the submission of this thesis "in partial fulfillment of the requirements for the degree" affirms me as a formal member of the scientific community. And yet, here at the terminus, I cannot help but think to myself: "doctor" is just an honorific. I am the same person before and after the recognition, per se, but of course that person is shaped by the struggle and labor from which the formal recognition stems.

So: why? While it is true that on some level, this has all been pursued because I found it beautiful and fun, at least some of the time, at the heart of the matter, it is this object in front of the reader's eyes that motivates the whole endeavor: this path has afforded me the chance to realize my self through labor. I am not sure that I have managed to achieve "beautiful mathematics" in the sense that, say, G.H. Hardy was thinking of in his *Apology*, or even in the way I might have dreamed of achieving as a wide-eyed undergraduate. In beginning this journey, I had preconceptions about what it would mean to pursue mathematics and to engage in scientific work, and likely all of them were wrong to lesser or greater extent. But, and this perhaps is the most important thing, I have certainly *worked*; this dissertation is nothing if not a testament to that which I have built, the transmutation of my personal experience into an independent object. And it is by the virtue of this work, this culmination of labor, separate from the imposition of

any preconception about what it was supposed to look like according to the mind's eye of a novice, that I can affirm myself, in and for myself.

I will close with a final reflection. Mathematics has shaped my life. When I was a child, although I did not know it at the time, mathematical work brought me to the edge of the world, when my father's job took our family to the island of Kwajalein Atoll in Micronesia — it's about as far from everything as it's possible to get without leaving the planet. As an undergraduate, mathematics took me to Budapest, where I became entranced by differential geometry, and combinatorial optimization. Now as a graduate student, mathematics has continued to take me to wondrous places. Last year, it was to Vienna, where I was able to share some of my research at the Sampling Theory and Applications conference poster symposium. This year, it was to Kobe, where I had the privilege of learning about exascale computation on the largest supercomputer in Asia, as of writing. I have forged deep friendships, and I have shared in fascinating conversation and adventure; in short, I have been unbelievably fortunate, although at times I have done my best to forget that fact, and I am grateful.

Bibliography

[AB22] Jason M. Altschuler and Enric Boix-Adserà. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

[ABS21] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on Optimal Transport*. Springer International Publishing, Cham, 2021.

[AC10] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2010.

[AC17] Samaneh Aminikhanghahi and Diane J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[ÁEdCM16] Pedro C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[Afs11] Bijan Afsari. Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(02):655–673, 2011.

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birhaäuser Basel, Basel, 2005.

[AHD25] Abu Al Hassan and Phong Ba Dao. Bridging data and diagnostics: A systematic review and case study on integrating trend monitoring and change point detection for wind turbines. *Energies*, 18(19):5166, 2025.

[AM07] Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprent arXiv:0710.3742*, 2007.

[Amb90] Luigi Ambrosio. Metric space valued functions of bounded variation. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, 17(3):439–478, 1990.

[AMS+15] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015.

[AOBL21] Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. Momentum improves optimization on Riemannian manifolds. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 1351–1359. PMLR, 2021.

[ATV13] Bijan Afsari, Roberto Tron, and René Vidal. On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013.

[BB00] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[BCC+15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[BCS12] Claudie Beaulieu, Jie Chen, and Jorge L. Sarmiento. Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1962):1228–1249, 2012.

[BFRT25] Julio Backhoff, Joaquin Fontbona, Gonzalo Rios, and Felipe Tobar. Stochastic gradient descent for barycenters in Wasserstein space. *Journal of Applied Probability*, 62(1):15–43, 2025.

[BN+93] Michele Basseville, Igor V Nikiforov, et al. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, Englewood Cliffs, 1993.

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[BPC16] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):1–10, 2016.

[BPN14] Gregory R. Bowman, Vijay S. Pande, and Frank Noé, editors. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer Netherlands, Dordrecht, 2014.

[Bre91] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

[Bro41] R. L. Brooks. On colouring the nodes of a network. *Mathematical Proceedings of the Cambridge Philosophical Society*, 37(2):194–197, 1941.

[BRS12] Avleen S. Bijral, Nathan Ratliff, and Nathan Srebro. Semi-supervised learning with density based distances. *arXiv preprint arXiv:1202.3702*, 2012.

[CAH+20] Kevin C. Cheng, Shuchin Aeron, Michael C. Hughes, Erika Hussey, and Eric L. Miller. Optimal transport based change point detection and time series segment clustering. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038, 2020.

[CC23] Hao Chen and Lynna Chu. Graph-based change-point analysis. *Annual Review of Statistics and Its Application*, 10(Volume 10, 2023):475–499, 2023.

[CCPS97] William J. Cook, William H. Cunningham, William R. Pulleyblank, and Alexander Schrijver. *Combinatorial Optimization*. John Wiley & Sons, Inc., 1997.

[CEK21] Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. Entropic-Wasserstein barycenters: PDE characterization, regularity and CLT. *SIAM Journal on Mathematical Analysis*, 53(5):5880–5914, 2021.

[Che23] Kevin C. Cheng. *Optimal Transport-Based Models and Methods for the Analysis of Probabilistic Time-Series*. PhD thesis, Tufts University, Medford, MA, 2023.

[CHLZ12] Shui-Nee Chow, Wen Huang, Yao Li, and Haomin Zhou. Fokker–Planck equations for a free energy functional or Markov process on a graph. *Archive for Rational Mechanics and Analysis*, 203(3):969–1008, 2012.

[CKCH23] Laurent Condat, Daichi Kitahara, Andrés Contreras, and Akira Hirabayashi. Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists. *SIAM Review*, 65(2):375–435, 2023.

[CL06] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[CLL+05] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005.

[CMHA20] Kevin C. Cheng, Eric L. Miller, Michael C. Hughes, and Shuchin Aeron. On matched filtering for statistical change point detection. *IEEE Open Journal of Signal Processing*, 1:159–176, 2020.

[CMRS20] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

[CP11] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[CRBV13] Roman Cmejla, Jan Rusz, Petr Bergl, and Jan Vokral. Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders. *Speech Communication*, 55(1):178–189, 2013.

[CRT20] Elsa Cazelles, Arnaud Robert, and Felipe Tobar. The Wasserstein-Fourier distance for stationary time series. *IEEE Transactions on Signal Processing*, 69:709–721, 2020.

[CTN25] Katy Craig, Nicolás García Trillos, and Đorđe Nikolić. Vector valued optimal transport: from dynamic to static formulations. *arXiv preprint arXiv:2505.03670*, 2025.

[CTV19] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[Cut13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[CZ15] Hao Chen and Nancy Zhang. Graph-based change-point detection. *The Annals of Statistics*, 43(1):139–176, 2015.

[dCMR99] Eustasio del Barrio, Juan A. Cuesta-Albertos, Carlos Matrán, and Jesús M. Rodríguez-Rodríguez. Tests of goodness of fit based on the L^2-Wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

[dFLR21] Maria d'Errico, Elena Facco, Alessandro Laio, and Alex Rodriguez. Automatic topography of high-dimensional data sets by non-parametric density peak clustering. *Information Sciences*, 560:476–492, 2021.

[DLM23] Jovan Damjanovic, Yu-Shan Lin, and James M. Murphy. Modeling changes in molecular dynamics time series as Wasserstein barycentric interpolations. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, pages 1–7, 2023.

[DML21] Jovan Damjanovic, James M. Murphy, and Yu-Shan Lin. CATBOSS: Cluster analysis of trajectories based on segment splitting. *Journal of Chemical Information and Modeling*, 61(10):5066–5081, 2021.

[DMW23] Moon Duchin, James M. Murphy, and Thomas Weighill. Measuring segregation via analysis on graphs. *SIAM Journal on Matrix Analysis and Applications*, 44(1):80–105, 2023.

[DSS10] Julie Delon, Julien Salomon, and Andrei Sobolevski. Fast transport optimization for Monge costs on the circle. *SIAM Journal on Applied Mathematics*, 70(7):2239–2258, 2010.

[EM12] Matthias Erbar and Jan Maas. Ricci curvature of finite Markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, 206(3):997–1038, 2012.

[EMW19] Matthias Erbar, Jan Maas, and Melchior Wirth. On the geometry of geodesics in discrete optimal transport. *Calculus of Variations and Partial Differential Equations*, 58(1):19, 2019.

[ERSS20] Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon. Computation of optimal transport on discrete metric measure spaces. *Numerische Mathematik*, 144(1):157–200, 2020.

[ESC+17] Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, 2017.

[Eul41] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. In *Commentarii Academiae scientiarum imperialis Petropolitanae*. Petropolis, Typis Academiae, 1741.

[FCG+21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[FCS+21] Kamil Faber, Roberto Corizzo, Bartlomiej Sniezynski, Michael Baron, and Nathalie Japkowicz. WATCH: Wasserstein change point detection for high-dimensional time series data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4450–4459, 2021.

[FDM+15] Zhou Fan, Ron O Dror, Thomas J Mildorf, Stefano Piana, and David E Shaw. Identifying localized changes in large systems: Change-point detection for biomolecular simulations. *Proceedings of the National Academy of Sciences*, 112(24):7454–7459, 2015.

[FG21] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. EMS Press, Berlin, 2021.

[FJF56] Lester R Ford Jr and Delbert. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[GB08] Michael C. Grant and Stephen P. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, pages 95–110, London, 2008. Springer London.

[GB14] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, 2014.

[GGCN23] Javier González-Delgado, Alberto González-Sanz, Juan Cortés, and Pierre Neuvial. Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology. *Electronic Journal of Statistics*, 17(1):1547–1586, 2023.

[GHL24] Florian Gunsilius, Meng Hsuan Hsieh, and Myung Jin Lee. Tangential Wasserstein Projections. *Journal of Machine Learning Research*, 25(69):1–41, 2024.

[GHM26] David Gentile, Joshua Huang, and James M. Murphy. Wasserstein-based identification of metastable states in time series data via change point detection and segment clustering. *arXiv preprint arXiv:2603.01989*, 2026.

[GKM20] Peter Gladbach, Eva Kopfer, and Jan Maas. Scaling Limits of Discrete Optimal Transport. *SIAM Journal on Mathematical Analysis*, 52(3):2759–2802, 2020.

[GLM19] Wilfrid Gangbo, Wuchen Li, and Chenchen Mou. Geodesics of minimal length in the set of probability measures on graphs. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:78, 2019.

[GLR13] Colin Gallagher, Robert Lund, and Michael Robbins. Changepoint detection in climate time series with long-term trends. *Journal of Climate*, 26(14):4994–5006, 2013.

[GM96] Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[GM00] Wilfrid Gangbo and Robert J. McCann. Shape recognition via Wasserstein distance. *Quarterly of Applied Mathematics*, 58(4):705–737, 2000.

[GM13] Nicola Gigli and Jan Maas. Gromov–Hausdorff convergence of discrete transportation metrics. *SIAM Journal on Mathematical Analysis*, 45(2):879–899, 2013.

[GM25] David Gentile and James M. Murphy. Regularized 2-Wasserstein barycenters on discrete metric spaces. In *2025 International Conference on Sampling Theory and Applications (SampTA)*, pages 1–5. IEEE, 2025.

[GM26] David Gentile and James M. Murphy. Static and dynamic approaches to computing barycenters of probability measures on graphs. *arXiv preprint arXiv:2603.26940*, 2026.

[GPC18] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

[Gup15] Vishwa Gupta. Speaker change point detection using deep neural nets. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4420–4424, 2015.

[Hab21] Reza Habibi. Bayesian online change point detection in finance. *Financial Internet Quarterly*, 17(4):27–33, 2021.

[HHRR20] Shenyang Huang, Yasmeen Hitti, Guillaume Rabusseau, and Reihaneh Rabbany. Laplacian change point detection for dynamic graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 349–358, New York, NY, 2020. Association for Computing Machinery.

[HKM22] Shayan Hundrieser, Marcel Klatt, and Axel Munk. The Statistics of Circular Optimal Transport. In *Directional Statistics for Innovative Applications: A Bicentennial Tribute to Florence Nightingale*, pages 57–82. Springer Nature, Singapore, 2022.

[HMF77] Georg Wilhelm Friedrich Hegel, Arnold V. Miller, and J. N. Findlay. *Phenomenology of Spirit*. Clarendon Press, Oxford, 1977.

[HSH+21] Moritz Hoffmann, Martin Scherer, Tim Hempel, Andreas Mardt, Brian de Silva, Brooke E Husic, Stefan Klus, Hao Wu, Nathan Kutz, Steven L Brunton, and Frank Noé. Deeptime: A Python library for machine learning dynamical models from time series data. *Machine Learning: Science and Technology*, 3(1):015009, 2021.

[HT88] Jun S Huang and Dong H Tseng. Statistical theory of edge detection. *Computer Vision, Graphics, and Image Processing*, 43(3):337–346, 1988.

[JCG20] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased Sinkhorn barycenters. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4692–4701. PMLR, 2020.

[JCP25] Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *Foundations of Computational Mathematics*, pages 1–52, 2025.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[JSB+05] Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108, 2005.

[KCHP01] E. Keogh, S. Chu, D. Hart, and M. Pazzani. An online algorithm for segmenting time series. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 289–296, 2001.

[Kor76] GM Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i matematicheskie metody*, 12(4):747–756, 1976.

[KP17] Young-Heon Kim and Brendan Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017.

[KPLS22] Kyungwon Kim, Ji Hwan Park, Minhyuk Lee, and Jae Wook Song. Unsupervised change point detection and trend prediction for financial time-series using a new CUSUM-based approach. *IEEE Access*, 10:34690–34705, 2022.

[KQBN69] Alexandre Kojève, Raymond Queneau, Allan Bloom, and James H. Nichols. *Introduction to the Reading of Hegel : Lectures on the Phenomenology of Spirit*. Basic Books, New York, 1969.

[LBD25] Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-order conditions for optimization in the Wasserstein space. *SIAM Journal on Mathematics of Data Science*, 7(1):274–300, 2025.

[Li22] Wuchen Li. Transport information geometry: Riemannian calculus on probability simplex. *Information Geometry*, 5(1):161–207, 2022.

[LM79] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[LWH18] Siqi Liu, Adam Wright, and Milos Hauskrecht. Change-point detection method for clinical decision support system rule monitoring. *Artificial Intelligence in Medicine*, 91:49–56, 2018.

[Maa11] Jan Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, 2011.

[MAMT23] Marshall Mueller, Shuchin Aeron, James M. Murphy, and Abiy Tasissa. Geometrically Regularized Wasserstein Dictionary Learning. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, pages 384–403. PMLR, 2023.

[McC97] Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[MDC20] Quentin Mérigot, Alex Delalande, and Frederic Chazal. Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 3186–3196. PMLR, 2020.

[Mer94] Russell Merris. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 197:143–176, 1994.

[Mer95] Russell Merris. A survey of graph laplacians. *Linear and Multilinear Algebra*, 39(1-2):19–31, 1995.

[MM19] Mauro Maggioni and James M. Murphy. Learning by unsupervised nonlinear diffusion. *Journal of Machine Learning Research*, 20(160):1–56, 2019.

[MMA25] Brendan Mallery, James M. Murphy, and Shuchin Aeron. Synthesis and analysis of data as probability measures with entropy-regularized optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 2584–2592. PMLR, 2025.

[MMS25] Eduardo Fernandes Montesuma, Fred Maurice Ngolè Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1161–1180, 2025.

[MNW19] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[Mon81] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

[MPWN18] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):5, 2018.

[MWMA23] Shoaib Bin Masud, Matthew Werenski, James M. Murphy, and Shuchin Aeron. Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling. *Journal of Machine Learning Research*, 24(160):1–65, 2023.

[NHZ16] Yue S. Niu, Ning Hao, and Heping Zhang. Multiple change-point detection: A selective overview. *Statistical Science*, 31(4):611–623, 2016.

[NJW01] Andrew Ng, Michael Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

[NWJH16] Feiping Nie, Xiaoqian Wang, Michael Jordan, and Heng Huang. The constrained laplacian rank algorithm for graph-based clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1):1969–1976, 2016.

[Øks10] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media, 2010.

[Ott01] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[Pag54] Ewan. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954.

[Pat07] Jan Peter Patist. Optimal window change detection. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 557–562, 2007.

[PC19] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Now Foundations and Trends, 2019.

[Pic85] Dominique Picard. Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4):841–867, 1985.

[PNW21] Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

[Pol87] Boris T Polyak. *Introduction to Optimization*. Optimization Software, Inc., New York, New York, 1987.

[PP17] Andrey Pepelyshev and Alexey Polunchenko. Real-time financial surveillance via quickest change-point detection methods. *Statistics and Its Interface*, 10(1):93–106, 2017.

[PW09] Ofir Pele and Michael Werman. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, Kyoto, 2009.

[PYWR21] Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric change point detection and localization. *Electronic Journal of Statistics*, 15:1154–1201, 2021.

[PZ20] Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer International Publishing, Cham, 2020.

[RCR21] Luana Ruiz, Luiz F. O. Chamon, and Alejandro Ribeiro. Graphon signal processing. *IEEE Transactions on Signal Processing*, 69:4961–4976, 2021.

[RCW+07] Jaxk Reeves, Jien Chen, Xiaolan L. Wang, Robert Lund, and Qi Qi Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

[Rig25] Philippe Rigollet. The mean-field dynamics of transformers. *arXiv preprint arXiv:2512.01868*, 2025.

[RKLE11] Thanawin Rakthanmanon, Eamonn J. Keogh, Stefano Lonardi, and Scott Evans. Time series epenthesis: Clustering time series streams requires ignoring some data. In *2011 IEEE 11th International Conference on Data Mining*, pages 547–556, 2011.

[RKMC25] Sawyer Jack Robertson, Dhruv Kohli, Gal Mishne, and Alexander Cloninger. On a generalization of wasserstein distance and the beckmann problem to connection graphs. *SIAM Journal on Scientific Computing*, 47(5):A2774–A2800, 2025.

[RL14] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[RMB+10] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions Sensor Networks*, 6(2):1–27, 2010.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[RTC17] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[RW13] Susanna Röblitz and Marcus Weber. Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Advances in Data Analysis and Classification*, 7(2):147–179, 2013.

[RW18] Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus. Mathématique*, 356(11-12):1228–1235, 2018.

[RWC24] Sawyer Robertson, Zhengchao Wan, and Alexander Cloninger. All you need is resistance: On the equivalence of effective resistance and certain optimal transport problems on graphs. *arXiv preprint arXiv:2404.15261*, 2024.

[San15] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, Cham, 2015.

[SHB+18] Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngolè, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[Sin67] Richard Sinkhorn. Diagonal equivalence to matrices with brescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

[SK67] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[SK87] C. S. Smith and M. Knott. Note on the optimal transportation of distributions. *Journal of Optimization Theory and Applications*, 52(2):323–329, 1987.

[SMD+19a] Ljubisa Stankovic, Danilo Mandic, Milos Dakovic, Milos Brajovic, Bruno Scalzo, and Anthony G. Constantinides. Graph signal processing – Part II: Processing and analyzing signals on graphs. *arXiv preprint arXiv:1909.10325*, 2019.

[SMD+19b] Ljubisa Stankovic, Danilo Mandic, Milos Dakovic, Milos Brajovic, Bruno Scalzo, and Tony Constantinides. Graph signal processing – Part I: Graphs, graph spectra, and spectral clustering. *arXiv preprint arXiv:1907.03467*, 2019.

[Sol18] Justin Solomon. Optimal transport on discrete domains. *arXiv preprint arXiv:1801.07745*, 2018.

[SRGB16] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Continuous-flow graph transportation distances. *arXiv preprint arXiv:1603.06927*, 2016.

[SYXK23] Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8174–8194, 2023.

[TDL03] Jean-Yves Tourneret, Michel Doisy, and Marc Lavielle. Bayesian off-line detection of multiple change-points corrupted by multiplicative noise: Application to SAR image edge detection. *Signal Processing*, 83(9):1871–1887, 2003.

[TdSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[TFS16] Altan Turgut, Laurie T. Fialkowski, and Jeffrey A. Schindall. Measured depth-dependence of waveguide invariant in shallow water with a summer

profile. *The Journal of the Acoustical Society of America*, 139(6):EL184–EL189, 2016.

[TLMM24] Nicolás García Trillos, Anna Little, Daniel McKenzie, and James M. Murphy. Fermat distances: Metric approximation, spectral convergence, and clustering algorithms. *Journal of Machine Learning Research*, 25(1):1–65, 2024.

[TOV20] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[TR08] Rasool Tahmasbi and Sadegh Rezaei. Change point detection in GARCH models for voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):1038–1046, 2008.

[Tu17] Loring W. Tu. *Differential Geometry: Connections, Curvature, and Characteristic Classes*. Springer International Publishing, 2017.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[Vil09] Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, Heidelberg, 2009.

[Wal45] Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[Was04] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, New York, 2004.

[WCPZ22] Lingfei Wu, Peng Cui, Jian Pei, and Liang Zhao, editors. *Graph Neural Networks: Foundations, Frontiers, and Applications*. Springer Nature Singapore, Singapore, 2022.

[Wer24] Matthew Werenski. *Optimal Transport Maps: Theory and Applications*. PhD thesis, Tufts University, Medford, MA, 2024.

[WGX22] Jie Wang, Rui Gao, and Yao Xie. Two-sample test with kernel projected wasserstein distance. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 8022–8055. PMLR, 2022.

[WJT+22] Matthew Werenski, Ruijie Jiang, Abiy Tasissa, Shuchin Aeron, and James M. Murphy. Measure estimation in the barycentric coding model. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23781–23803. PMLR, 2022.

[WLJ24] Rui Wu, Chao Liu, and Dongxiang Jiang. Unsupervised Bayesian change-point detection approach for reliable prognostics and health management of complex mechanical systems. *Reliability Engineering & System Safety*, 245:110037, 2024.

[WMAM25] Matthew Werenski, Brendan Mallery, Shuchin Aeron, and James M. Murphy. Linearized Wasserstein barycenters: Synthesis, analysis, representational capacity, and applications. In *The 28th International Conference on Artificial Intelligence and Statistics*, pages 258:2584–2592, 2025.

[WMMA24] Matthew Werenski, Shoaib Bin Masud, James M. Murphy, and Shuchin Aeron. On rank energy statistics via optimal transport: Continuity, convergence, and change point detection. *IEEE Transactions on Information Theory*, 70(7):5168–5190, 2024.

[WN20] Hao Wu and Frank Noé. Variational approach for learning Markov processes from time series data. *Journal of Nonlinear Science*, 30(1):23–66, 2020.

[WQH+24] Yizhang Wang, Jiaxin Qian, Muhammad Hassan, Xinyu Zhang, Tao Zhang, Chao Yang, Xingxing Zhou, and Fengjin Jia. Density peak clustering algorithms: A review on the decade 2014–2023. *Expert Systems with Applications*, 238:121860, 2024.

[ZCY22] Yubo Zhuang, Xiaohui Chen, and Yun Yang. Wasserstein K-means for clustering probability distributions. *Advances in Neural Information Processing Systems*, 35:11382–11395, 2022.

[ZLWX08] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 247–256, New York, New York, 2008.

[ZP19] Yoav Zemel and Victor M. Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976, 2019.