Optimal Transport in Structured Domains: Algorithms and Applications

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
OCT 03 2019
LIBRARIES
ARCHIVES

by

David Alvarez Melis

B.S., Instituto Tecnológico Autónonmo de México (2011)
M.S., Courant Institute, New York University (2013)

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

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

© Massachusetts Institute of Technology 2019. All rights reserved.

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
July 19, 2019

Signature redacted

Certified by .....
Tommi S. Jaakkola
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

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

Optimal Transport in Structured Domains: Algorithms and Applications

by

David Alvarez Melis

Submitted to the Department of Electrical Engineering and Computer Science
on July 19, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

Optimal transport provides a powerful mathematical framework for comparing probability distributions, and has found successful application in various problems in machine learning, including point cloud matching, generative modeling, and document comparison. However, some important limitations curtail its broader applicability. In many applications there is often additional structural information that is not captured by the classic formulation of the problem. This information can range from explicit tree and graph-like structure, to global structural invariances. Failure to fully model this structure can hinder—if not preclude—the use of optimal transport-based approaches.

This thesis presents several extensions of the optimal transport problem to incorporate structural information. First, a non-linear generalization of the cost objective based on submodularity is proposed. The resulting formulation provides a flexible framework to model explicit or latent discrete structure in the data and admits efficient optimization. Next, we investigate the issue of geometric invariances when matching embedded representations, for which a general framework for optimal transport in the presence of latent global transformations is developed. Various approaches to solve the resulting optimization problem are proposed and compared. The last part of the thesis addresses the problem of aligning datasets in which the structure is encoded through non-Euclidean manifolds, such as hyperbolic spaces. In response to an unexpected type of invariance that hyperbolic embeddings learned from data exhibit, a novel framework that interweaves optimal transport and hyperbolic nonlinear registration with deep neural networks is proposed.

While these extensions are formulated in general terms, the experimental results presented in this thesis are focused on motivating applications in natural language processing, including unsupervised word translation, sentence similarity, domain adaptation, and ontology alignment.

Thesis Supervisor: Tommi S. Jaakkola
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

First and foremost, I would like to thank my advisor Tommi Jaakkola for sharing his wisdom, experience, and knowledge with me for the past 5 years; for tolerating (and encouraging!) my eclectic interests; and for supporting my mutable—sometimes erratic—ideas and projects. Despite knowing that some directions were worth pursuing more than others—his remarkable insight into what ideas are likely to succeed never ceases to amaze me—he gave me complete freedom to explore and grow as a researcher throughout my PhD. The lessons learned from this process (i.e., how to uncover rough ideas, how to gauge their feasibility, and how to choose which ones to pursue) is without a doubt the most valuable skill I have acquired during my PhD.

I would also like to thank the other members of my thesis committee: Stefanie Jegelka and Justin Solomon. My gratefulness to Stefanie goes far beyond this role. She has been like a second advisor, and without her invaluable mentorship, this thesis would certainly not have been the same. My interactions with her were very intellectually rewarding, and I have her to thank for introducing me to submodularity and various other fascinating topics in discrete optimization. On the other hand, I am very grateful to Justin for his advice and suggestions during this process, which have helped shape the last part of the thesis. But perhaps more importantly, I'm indebted to him because his own PhD work on optimal transport featured prominently in my first contact with this topic, and was a determining factor in getting me interested in it.

Throughout my PhD, various other mentors have left a mark on me as a researcher. Chief among them is Suvrit Sra, whose enthusiasm for research, high standards for theoretical soundness and perpetual quest for open problems, are models that I aspire to. He was always available when I brought him a technical question, or when I needed help in shaping and polishing vague ideas. As if it weren't enough, both him and Stefanie both took me in, invited me to their group meetings and treated me as a member of their group. I will be forever grateful for this. On the other hand, I would like to thank my mentors from the MSR FATE group (Hanna Wallach, Jenn

W. Vaughan and Hal Daumé III) for the opportunity to work with them, for our various philosophical discussions about interpretability and for their career advice. Although predating my PhD stage, various other mentors throughout my academic career have all shaped the way I think, among which Carlos Bosch, Mehryar Mohri, and Ken Church deserve special recognition.

Albeit trite, John Donne's "no man is an island" quote is so on-target for academic research that I cannot avoid succumbing to it. Indeed, I am profoundly indebted to the collaborators with whom I worked on the papers included in this thesis (Tommi, Stefanie and Youssef Mroueh) and on those which were not (Tatsu Hashimoto, Martin Saveski, Chengtao Li, Keyulu Xu, Guang-He Lee, Wengong Jin, and Charlotte Bunne). Besides them, there have been many other colleagues with whom I have interacted in various ways and who have enriched my PhD journey. Worth special recognition are the members of Tommi's group (Andreea Gane, Vikas Garg, John Ingraham, Paresh Malalur, Jonas Muller, David Reshef, Yu Xin), Regina's group (Nate Kushman, Tao Lei, Yuang Zhang), LOGSS and RIIAA (too many to list!).

No acknowledgment section would be complete without a hat-tip to those who besides providing emotional support during the PhD journey, make the journey itself worth it. First, I would like to thank my friends here in Boston: those who have been by my side since orientation (Sirma, Martin, Tal, Thras, Tugce, Lea, Viirj, Valerio, Alex. Prashan) and those who I've met (or re-encountered) along the way (Anna, Jordi, Celiacos++, Silvi, Sebas, Chris). I am also indebted to my friends from NYU (Joana, Kala, Konst & Nilu) and ITAM (Gerardo 1, Gerardo 2, Kawas & Peter) for their continued support and companionship, which endures despite the distance.

There are people for whom the word *friend* falls short, and can only be described as *de facto siblings.* I am fortunate to have found more of these than I could have ever hoped for: Andrés, Baez, David 2, Doogie, Drea, Elisa, Garci, Jarro, Joan, Jiks, Lau, Maris, Mich, Salas, Tat, Tort, Yisos. The mere remembrance of the good times with them was enough to power through the hard ones.

Also making every step of this process worth it was Judith, who as been by my side from early in the PhD, living the ups and downs just as intensely as I have. I

must thank her for supporting me throughout this journey; for inspiring me with her early-morning philosophical questions about the human mind; for collaborating with me in our shared research projects; for helping with many design aspects of my papers; and, ironically, for encouraging me from time to time to take a break from this thesis, which helped me maintain my sanity and rekindle my enthusiasm for research. Moltes gràcies!

Finally, I must thank my family, from Guadalajara to Brussels, and in particular my brother Tomás and my parents Jesús and Chantal for their unconditional, tireless and perpetual support. I am truly grateful to my parents for raising us in a home where asking *how* and *why* of essentially anything was the substance of everyday life. Their hunger for knowledge, love for scholarly curiosity and passion for academic research provided a model that turned out to be irresistible to aspire to. Had it not been for them, I would certainly not be where I am today.

Contents

List of Figures

List of Tables

Glossary of Symbols

$\mathcal{B}_F$	Base polytope of set function F
$\mathbb{D}^d$	Poincare Ball of dimension d
$\mathbb{H}$	Hyperbolic space of an arbitrary dimension
$\odot$	Matrix Hadamard (entry-wise) product
$\otimes$	Matrix Kronecker product
$\oplus$	Matrix Kronecker sum
$\|\cdot\|_p$	Matrix Schatten p-norm
Σ_d	Probability simplex of size d
γ	Continuous coupling matrix between measures
Γ	Discrete coupling matrix between measures
$\Pi(\alpha, \beta)$	Set of couplings between measures α and β
$\mathbf{\Pi}(\mathbf{a}, \mathbf{b})$	Set of couplings between vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbb{1}_{n,m}$	Matrix of size $n \times m$ with all entries equal to 1
$\mathbb{0}_{n,m}$	Matrix of size $n \times m$ with all entries equal to 0
$\text{vec}(\cdot)$	Vectorization (flattening) operator on a matrix

Chapter 1

Introduction

Optimal transport (OT) is a mathematical toolbox for comparing probability distributions tracing back its roots to the 18th century [132], born from the need to find cost-effective schemes to transport coal from mines to factories. Since then, OT has developed into a mature subfield at the intersection of mathematics, statistics, and optimization, boasting elegant theory and applications ranging from economics to computer graphics.

Optimal transport plays dual roles across machine learning applications. First, it provides a well-founded, geometrically driven approach to realizing correspondences between sets of objects such as shapes in different images. Such correspondences can be used for image registration [81] or to interpolate between them [163]. More generally, OT extends to problems such as domain adaptation where we wish to transport a set of labeled source points to the realm of the target task [44, 43]. Second, the transportation problem induces a theoretically well-characterized distance between distributions. This distance is expressed in the form of a transport cost and serves as a natural population difference measure, which can be exploited as a source of feedback in adversarial training [14, 32].

Despite its long history, widespread use of OT in machine learning was somewhat limited until very recently. Arguably, the main obstacle for further adoption had been scalability. The computational complexity of the transportation problem made its use prohibitive for the large-scale problems that abound in machine learning. This issue

has been greatly alleviated by major recent achievements on the optimization side of OT, yielding remarkably more efficient algorithms [45, 3, 70].

A central argument of this thesis is that the most significant limitation for broader use of OT in machine learning now stems from its *applicability*. In particular, in most tasks in machine learning there is additional *structural information* beyond the ground metric that remains uncaptured by the original formulation of the transportation problem. This structure can be *explicit* if the distributions are defined over structured objects, such as trees or graphs, or if there is additional information associated with the support points (e.g., class labels) that induces structure. On the other hand, there might exist structural invariances in the domains of interest that *implicitly* define structure in the problem (e.g., rotational invariance). Very recently, there has been important progress towards modeling structure in OT, particularly in the context of computer graphics [163, 57], though various important open problems and untouched applications remain.

In the remainder of this chapter, we present various tasks in machine learning for which there is such additional information and discuss why the classic formulation of the optimal transport problem is ill-equipped to solve them. These applications will motivate and serve as evaluation for the extensions of optimal transport developed throughout this thesis. We end this chapter by providing an overview of the rest of the thesis and a summary of the contributions presented in each chapter.

1.1 Structure in Machine Learning

Our definition of *structure* in this thesis is purposely broad. On the one hand, we use it in its more traditional meaning to refer to *explicit* structure in the data of interest, such as when these consist of sequences, trees or graphs. But we use this term also to refer to *implicit structure*. For example, labels or other metadata might confer a latent structure to the data, or there might be structural priors on its representation, such as those defined by global invariances.

It is hard to overstate the prevalence and importance of structured objects in

various domains within and around machine learning. Sequences and trees are at the core of natural language processing (NLP); the former as the basic surface-form by which written and spoken language is transmitted (through sequences of characters and words), and the latter in the form of various latent structures underlying it, such as constituency or semantic parse trees. Depending on what a *unit* is defined to be, it could be argued that all of these are examples of settings where the *instances* of the learning problem themselves are structured. But structure can appear in NLP at the dataset-level too. The archetypal examples of this are structured lexical databases like WordNet [129], which are widely used as an additional resource in various downstream tasks in computational linguistics [131, 158, 30]. These databases consist of a collection of words (the nodes) labeled with metadata and edges linking them, such as hypernym or synonym relations.

Structured data are also ubiquitous in bioinformatics and chemistry. For example, evolutionary relationships among organisms are naturally represented in terms of phylogenetic trees. Large-scale protein-protein interaction and gene regulatory networks are now routinely available, and are the primary ingredient for various computational approaches. As for chemistry, it should come as no surprise that molecules are traditionally represented mathematically as graphs, with nodes playing the role of atoms as edges the role of bonds. Yet another type of structured object which occurs in biology—and various other domains—are ontologies, a generic data representation type consisting of entities and relations between them [58]. These relations usually correspond to a hierarchical or graph structure.

But structure can appear in much more subtle ways in machine learning too. For example, whenever *labels* are provided as part of the learning problem, these often induce a latent structure on the examples. Concretely, in the case of classification, discrete labels implicitly define clusters in the feature space. Furthermore, if the labels themselves belong to a hierarchy or can be otherwise represented relationally, then the corresponding feature vectors derive complex structure from these relations too. All of these are instances where the structure is present in *additional information* beyond the feature representation of the data. However, the representation itself can reveal

structure, such as geometric invariants or other group actions.

Structure is thus pervasive in machine learning, often conspicuously present in the data, sometimes lurking silently in the tools we use to represent it. Therefore, it is no surprise that learning with structure has been one of the core problems in this field over the past two decades [94, 107, 170, 47, 103, 48]. Recently, the advent of representation learning has brought about a new set of challenges in this realm. Designing methods to derive meaning from large collections of structured data is at the forefront of modern machine learning research [127, 139, 169, 136, 99, 82]. Conversely, new computational challenges emerge from the need to operate on these representations. The object of study of this thesis is precisely rooted at this intersection of *representation of* and *computation on* structured data. In the next section, we provide various concrete examples of problems that arise from this junction, and which motivate the contributions of this thesis.

1.2 Motivating Applications and the Case for Optimal Transport

Comparing and relating data instances is one of the most fundamental tasks in machine learning, and a key building block of most learning algorithms. This task can be conceptually thought of as consisting of two independent but equally important steps, which require answering the following questions:

i) *how should the data be represented computationally?*

ii) *how should instances thus represented be compared?*

When the data can be naturally represented in Euclidean space (e.g., the features are independent and scalar-valued) neither of these questions warrants much attention: the representation problem is a-priori solved and the comparison can be done meaningfully using the Euclidean metric. However, when the data of interest are of a more complex nature (such as structured objects like trees or graphs) or is to be operated on at

a more abstract level (e.g., through *collections* of items or probability distributions defined over them), these seemingly innocent questions become challenging problems.

Consider for example the problem of automatic sentence similarity assessment. In this task, we seek a method to predict the similarity between any pair of sentences in a given language. A popular approach to addressing question (i) within this task is to use distributional representations of words [127, 139], which capture rich semantic information in dense feature vectors. With this, we can now compare words in terms of their distance in this vector space.[1] However, truly answering question (ii), i.e., comparing the *sentences* themselves, requires a method to lift the notion of word-to-word similarity into one that takes into account the relations, the relative order and the role of these words within their respective sentences.

Another setting where finding correspondences is at the core of the learning problem is domain adaptation. In this paradigm, training labels in the *target* domain of interest are assumed to be minimal or non-existent, while labeled data for a similar —though not identical— *source* domain is available in much larger quantities. The goal is thus to leverage the similarities between the domains to make use of source labeled data to train a model for the target domain. If the representations of the data from the two domains are compatible, an instinctive idea is to find correspondences between the two collections of data, and use these to *transport* the labeled source samples to the target domain. However, doing so only using the geometry of the representations —e.g., their pairwise distances in the feature space— leaves important information behind. Intuitively, one would want to encourage points sharing the same class label to be "moved" together, so as to respect the cluster structure of the domain. In other words, we seek for the transportation of samples to be informed both by the geometry and the underlying label-induced structure of the examples.

Our final motivating application concerns the problem of unsupervised word translation, which involves finding word-to-word correspondences across languages without access to any parallel data, but only monolingual texts. Again, a sensible

[1] However, the question of what metric is most appropriate to use when comparing vector-space models of language has sparked much debate since their early days [160, 116, 84].

starting point is to use word embedding algorithms on the monolingual corpora to produce vector representations of the words in the two languages. The observation that word embeddings across different languages exhibit similar semantic phenomena [127, 84] suggests that these embedded representations might be similar enough to allow for correspondences to be inferred between them. One could, for example, translate words as their closest neighbor among all vectors from the other language. Unfortunately, this naive approach fails – for two separate reasons. First, greedily assigning correspondences often leads to many-to-one mappings (same translation for several source words), but more crucially, direct computation of distances between the embeddings is meaningless because there is no guarantee that the spaces are globally aligned (e.g., all the vectors could be rotated by a constant angle in one of the spaces). Thus, a better approach would involve coherent assignment of words *as a collection* on the one hand, and a notion of similarity that is invariant to such global transformations.

Despite being seemingly unrelated, all the problems mentioned above share three crucial characteristics:

1. They involve a combination of correspondence, similarity, and transportation of collections of samples

2. Any sensible approach to solving them should leverage the geometry of the vector feature representations

3. Besides the feature representation, there is important additional structural information that should inform the proposed approach

Somewhat surprisingly, there is a mathematical toolbox that provides a unified framework for the three tasks described in Point (1), and furthermore, has geometry embedded in its core. As the reader will surely suspect by this point, we are referring to optimal transport. As we have mentioned before, OT is an appealing tool for these tasks because of its strong theoretical foundations, efficient algorithms, and intuitive nature. Indeed, many problems in machine learning and computer graphics that share

Points (1) and (2) have been approached with optimal transport approaches, including sentence similarity [106], domain adaptation [44] and unsupervised word translation [184]. However, addressing Point (3) through the lens of optimal transport remains largely an open problem. Indeed, one of the contributions of this thesis is showing that neglecting structural information in the transportation problem is significantly detrimental to the performance of these approaches. Then, the question we seek to answer is

Can the framework of optimal transport be extended to structured domains?

In the next section, we briefly discuss why the problem posed by this question is challenging and outline *in broad strokes* directions to tackle it. The rest of this thesis is devoted to filling out the details of those strokes.

1.3 Optimal Transport with Structure: a Roadmap

In the previous section, we argued that optimal transport is an appealing approach to tackle problems involving correspondence between collections of objects represented in a vector space. Indeed, an important aspect of optimal transport distances is that they reflect the metric of the underlying space in the transport cost. Yet, in all the motivating applications discussed in the previous section, there is further important structure that remains uncaptured. From a modeling perspective, there are three main components of the transportation problem onto which this additional structure information can be injected:

i) The cost function $c(\cdot, \cdot)$

ii) The representation spaces $\mathcal{X}, \mathcal{Y}$

iii) The marginal constraints (i.e., the set $\Pi(\mathbf{p}, \mathbf{q})$ described in the next chapter)

Each of these approaches might be appropriate for different types of structure and needs. For example, if the representation is fixed (either because it is expensive to generate it, or it is provided as such) and does not already reflect the additional

structural information, modifying the cost function is perhaps the most sensible route. If instead the representation can be chosen, one might choose to proceed with (ii), e.g., certain hierarchies can be naturally modeled through non-Euclidean (hyperbolic) geometries. This, however, necessitates an investigation of whether optimal transport is valid (and efficiently computable) in these alternative representations. Approach (iii) might be appealing if one seeks to fully leverage the OT toolkit (e.g., doing meaningful displacement interpolation, for which marginal constraints enforce structurally-coherent distributions at every step).

In this thesis, we propose extensions of optimal transport using the approaches (i) and (ii), and discuss possible approaches to tackle the last one—i.e., enforcing structured marginal constraints—in the concluding chapter, leaving a detailed exploration of this third angle for future work. More specifically, we propose several extensions of optimal transport to account for the various types of structural information described in Section 1.2. In the next section, we provide a birds-eye view of these extensions and the broader outline of this thesis.

1.4 Overview of this Thesis

This thesis covers a subset of the author's work conducted as part of the PhD requirements. For the sake of producing a coherent and well-structured manuscript, the author opted to focus it on a single line of research. Thus, additional lines of work on structured representation learning [85, 84, 7]; interpretability and robustness [8, 5, 6, 111, 112, 113]; and generative modeling [117, 35] are not included in this thesis, but the interested reader can find their details in the bibliography.

The layout of this thesis is intended to facilitate independent reading of chapters by minimizing the dependencies between them. Except for Chapter 2 which provides background for the rest of the thesis, the other chapters, each of which presents an independent extension on the classic formulation of optimal transport presented in Chapter 2, and can be read for the most part separately, in any order. However, the task tackled in Chapter 5 will be better motivated (and its complexity more clearly

appreciated!) after reading Chapter 4. A brief summary of each chapter is provided below.

CHAPTER 2 sets up the background for the rest of the thesis. It is presented as a review of fundamental concepts in the theory of optimal transport, highlighting key properties and results which form the basis of the concepts presented in all subsequent chapters. While this chapter contains crucial notions that will be referred to throughout the thesis, a reader familiar with basic concepts of optimal transport can safely skip it.

CHAPTER 3 deals with the problem of extending optimal transport to model explicit structure in the form of sequence or tree-shaped data, or latent hierarchies implicitly defined through labels. Compared to the two other principal chapters, this one is perhaps the one requiring the most additional background beyond the basic theory of optimal transport. The notion of submodularity, which plays a crucial role in this chapter, is a deep and complex one, so we present a gentle introduction for the uninitiated reader at the beginning of this chapter. Combined with the background on optimal transport, this provides all the necessary ingredients to define a generalization of optimal transport with structured submodular cost functions. After formalizing the approach and deriving various optimization approaches to solve it, the last section validates this framework on various experimental settings involving domain adaptation, color transfer and sentence similarity. This chapter almost entirely based on Alvarez-Melis et al. [9], with additional background on submodularity, a more detailed derivation of the optimization routines and extended experimental results.

CHAPTER 4 addresses the problem of finding correspondences between embedded representations in the presence of global invariances, such as rotations. The motivating application is unsupervised alignment of word embedding spaces, where orthogonal invariances preclude the direct application of optimal transport. In response to this challenge, a general method to endow the transportation cost with invariance to various types of transformation is developed. At a high level, the resulting framework seeks to simultaneously optimize instance-wise correspondences between the embeddings and global alignment of the spaces. This chapter primarily builds upon Alvarez-Melis

et al. [10], exploring many other optimization approaches beyond the alternating minimization scheme proposed in that work, and includes Alvarez-Melis and Jaakkola [4] as an alternative approach to the same task.

CHAPTER 5 considers the problem of unsupervised alignment of hierarchical data. In this case, we adopt an alternative approach to encode structure: directly through the representation space. In a way, this chapter can be thought of as coupling the previous two: sharing the goal of Chapter 4—unsupervised estimation of correspondences between embedded spaces— but in a setting where the data we seek to match itself has an underlying hierarchical structure, as in Chapter 3. Experimental results in unsupervised WordNet translation and ontology matching are presented. This chapter is based on Alvarez-Melis et al. [11], with additional background on hyperbolic spaces and an extended experimental section.

CHAPTER 6 brings the thesis to a close. It discusses the implications of the contributions of the results of the preceding chapters under a unified view, elaborates on connections between them, and proposes various avenues of future work.

Chapter 2

Optimal Transport

The unifying theme throughout this thesis is to extend the optimal transport problem in various ways, and therefore we must begin by introducing the *classic* version of this problem. The purpose of this chapter is to provide a concise but self-contained introduction to OT and to introduce all the notions upon which the rest of this thesis will build. Naturally, there is a myriad of possible ways such an introduction could be presented, each with its own flavor and unique emphasis on the mathematical, statistical or optimization facets of OT. Here, we focus on computational and geometric aspects of the problem, and we closely follow the notations and conventions of Peyré and Cuturi [140]. Nevertheless, we explicitly state these conventions in section 2.1. The existence and regularity results discussed in Sections 2.6.1 and 2.6.2 are not necessary for understating Chapter 3, but will become relevant for Chapters 4 and 5, respectively.

2.1 Notation and Fundamentals

Sets, Spaces and Groups. Throughout this thesis, we denote sets as X, Y. When these sets correspond to spaces, we use calligraphic variables $\mathcal{X}, \mathcal{Y}$ instead. For a positive integer n, we denote by $[\![n]\!]$ the set of all positive integers up to an including n, i.e., $[\![n]\!] \triangleq \{1, \ldots, n\}$. Finally, $\mathrm{O}(n)$ and $\mathrm{SO}(n)$ are the orthogonal and special orthogonal groups of order n.

Vectors and Matrices. We denote vectors and matrices with bold font (e.g., $\mathbf{x}$, $\mathbf{X}$) and their entries without it (x_i, X_{ij}). We use super-indices $\mathbf{x}^{(i)}$ to enumerate vectors, and subindices $\mathbf{x}_j$ to denote their entries. The operators $\odot$ and $\oslash$ denote entry-wise operations between vectors or matrices, i.e., for matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \odot \mathbf{B}$ is their Hadamard product. On the other hand, $\otimes$ denotes the Kronecker product (the outer product for vectors), and for squared matrices we use $\oplus$ for their Kronecker sum (i.e., $\mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \mathbf{B}$. Note that in particular for vectors $\mathbf{a} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$ we have $[\mathbf{a} \oplus \mathbf{b}]_{ij} = \mathbf{a}_i + \mathbf{b}_j$. For matrices $\mathbf{A}, \mathbf{B}$, we denote by $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i,j} [\mathbf{A}]_{ij} [\mathbf{B}]_{ij}$ their Frobenius inner product.

Throughout this thesis, we will make use of the following matrix norms:

- Nuclear (trace) norm: $\|\mathbf{A}\|_* = \mathrm{Tr}(\sqrt{\mathbf{A}^*\mathbf{A}}) = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\mathbf{A})$
- Spectral norm: $\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$
- Frobenius norm: $\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A}^\top \mathbf{A})} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(\mathbf{A})}$
- Schatten-p norm: $\|\mathbf{A}\|_p = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(\mathbf{A}) \right)^{1/p}$

Note that the Schatten-p norm includes the previous three as particular cases, for the values $p = 1$ (nuclear), $p = 2$ (Frobenius) and $p = \infty$ (spectral).

Finally, $\mathrm{diag}(\cdot)$ denotes the vector obtained by extracting the main diagonal of a square matrix, while for a vector $\mathbf{v} \in \mathbb{R}^n$, $[[\mathbf{v}]]$ is a square $n \times n$ matrix with $\mathbf{v}$ in its main diagonal and zeros elsewhere, i.e., $\mathrm{diag}([[\mathbf{v}]]) = \mathbf{v}$.

Functions and Probability Measures. We denote by $\mathcal{P}(\mathcal{X})$ the set of probability distributions over a metric space $\mathcal{X}$. We use lower-case greek letters for members of this set, e.g., $\alpha \in \mathcal{P}(\mathcal{X})$. We use $\delta_{\mathbf{x}}$ to denote a Dirac point mass supported on $\mathbf{x} \in \mathbb{R}^d$, and thus a discrete distribution $\alpha \in \mathcal{P}(\mathbb{R}^d)$ supported on $\{\mathbf{x}^{(i)}\}_{i=1}^n$ can be expressed as $\alpha = \sum_i \mathbf{a}_i \delta_{\mathbf{x}^{(i)}}$, where $\mathbf{a} \in \Sigma_n$ is a vector of probability weights. For measures α and β, $\alpha \otimes \beta$ is their product measure on $\mathcal{X} \times \mathcal{Y}$, that is, $\int_{\mathcal{X}\times\mathcal{Y}} d(\alpha \otimes \beta)(x, y) = \int_{\mathcal{X}\times\mathcal{Y}} d\alpha(x) d\beta(y)$. For a continuous map $f : \mathcal{X} \to \mathcal{Y}$ we note by $f_\sharp : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ its associated push-forward operator, i.e., for any $\mu \in \mathcal{P}(\mathcal{X})$, $\nu = f_\sharp(\mu)$ is the push-forward measure

satisfying:

$$\int_{\mathcal{Y}} h(y)d\nu(y) = \int_{\mathcal{X}} h(f(x))d\mu(x)\ \forall h \in \mathcal{C}(\mathcal{Y})$$

The image of a function f with domain $\mathcal{X}$ is denoted as $f[\mathcal{X}] = \{f(x) \mid x \in \mathcal{X}\}$. Finally, $\mathcal{C}(\mathcal{X})$ and $\mathcal{C}^\infty(\mathcal{X})$ denote the space of continuous and smooth functions over $\mathcal{X}$, respectively.

2.2 Optimal Assignment and Monge's Problem

Our journey into optimal transport begins with the assignment problem, one of the most fundamental problems in combinatorial optimization. The notions of assignment, matching, and correspondence underpin the theory optimal transport; played an important role in its birth and development; and will feature prominently in this thesis – all of which make this a natural start point for this introductory chapter.

Consider two sets of items of equal size, labeled for convenience with indices $i, j \in [\![n]\!]$, and a cost matrix $[\mathbf{C}_{ij}]_{i\in[\![n]\!],j\in[\![n]\!]}$. The linear assignment problem consists of finding a bijection $\sigma : [\![n]\!] \to [\![n]\!]$ which minimizes the total cost of matching these two sets of items. Formally, the cost objective of this problem is

$$\min \sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}. \tag{2.1}$$

Note that this problem may have several optimal solutions, for example, if the cost matrix is symmetric.

A generalization of this problem is due to Gaspard Monge [132], one of the forefathers of optimal transport. Monge considered two discrete measures:

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}^{(i)}}, \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{y}^{(i)}}, \tag{2.2}$$

over metric spaces $\mathcal{X}, \mathcal{Y}$, and a measurable cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which represents the cost of transporting a unit of mass from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. Monge's problem seeks a transport map $T : \mathcal{X} \to \mathcal{Y}$ that associates each source point $\mathbf{x}^{(i)}$ to

a target point $\mathbf{y}^{(i)}$, and which pushes the mass of α to β at minimal cost. With the notion on pushforward operator at hand, the problem can be succinctly expressed as:

$$\inf_{T} \left\{ \int_{\mathcal{X}} c(\mathbf{x}, T(\mathbf{x})) d\alpha(\mathbf{x}) \mid T_\sharp \alpha = \beta \right\}, \tag{2.3}$$

Note that while T can still be expressed as an assignment $\sigma : [\![n]\!] \to [\![m]\!]$, where $\sigma(i) = j$ iff $T(\mathbf{x}^{(i)}) = \mathbf{y}^{(j)}$, this problem is a generalized version of the assignment problem. Indeed, for the particular case where $n = m$ and the two measures are uniform, Problem (2.3) is none other that (2.1). But, unlike before, the probability vectors are now allowed to be non-uniform and of different size. However, note that this generality comes at a price: the solution on Monge's problem might not exist if the measures are not compatible. This is clearly the case if $n < m$, but also for supports of the same size if the corresponding probability weight vectors are not be compatible (e.g., consider a uniform vector $\mathbf{a} = \frac{1}{n}\mathbb{1}$ and a sparse one $\mathbf{b} = e_1$).

Before ending this section, we note that for discrete measures α, β as introduced above, all relevant geometric information—i.e., the pairwise costs—can be compactly captured the matrix $\mathbf{C}_{ij} = c(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})$, after which all remaining relevant information is purely probabilistic, i.e., contained in their probability weight vectors $\mathbf{a}$ and $\mathbf{b}$. This shifts the focus from transportation between *measures* into transportation between *histograms*, whereby the bins in these histograms have associated costs, expressed in $[\mathbf{C}_{ij}]$. This allows for a definition of the Monge (and subsequent Kantorovich) problem without the need to appeal to the concept and terminology of measures, namely, by defining the problem for a pair of histograms $\mathbf{a} \in \Sigma_n$ and $\mathbf{b} \in \Sigma_m$ with associated bin-to-bin costs $\mathbf{C} \in \mathbb{R}^{n \times m}$ as finding an assignment σ between their bins, which satisfies $\sum_{i \in \sigma^{-1}(j)} \mathbf{a}_i = \mathbf{b}_j$ and which minimizes the sum of costs $\sum_{i=1}^{n} \mathbf{C}_{i,\sigma(i)}$. Owing to this observation, whenever dealing with discrete measures in this thesis we will often interchangeably refer to the problem in terms of the measures α, β or their underlying probability vectors $\mathbf{a}$ and $\mathbf{b}$.

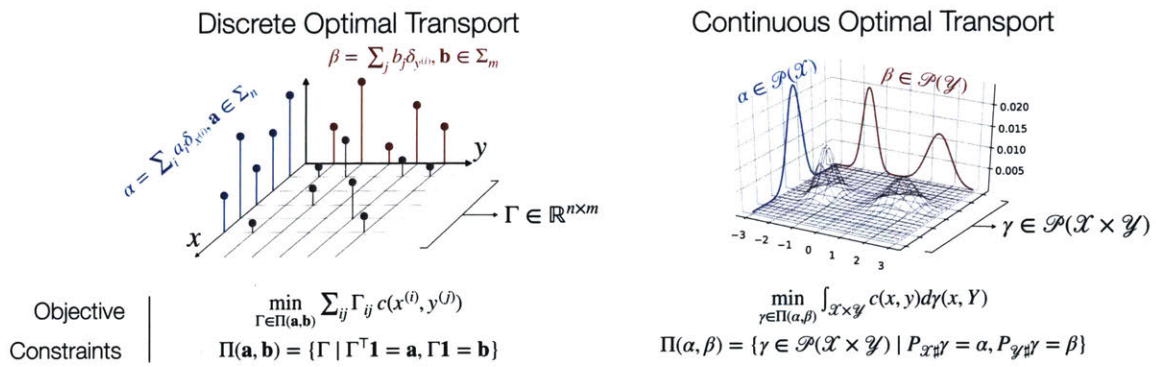


Figure 2-1: The discrete and continuous versions of Kantorovich's formulation of optimal transport in a nutshell. In the former case, the transport coupling (i.e., the optimization variable of interest) is finite and discrete (an n-by-m matrix Γ), while in the latter it is a joint continuous measure γ on $\mathcal{X} \times \mathcal{Y}$.

2.3 Kantorovich Relaxation

The problem proposed by Monge introduced in the previous section captures the intuition behind the transportation problem but has several limitations. Chief among these is the fact that it is often ill-defined, i.e., a solution might not exist. Furthermore, whenever the solution does exist, the combinatorial nature of the problem make it hard to solve it in practice.

Kantorovich's idea [97] was to relax Monge's problem by replacing the deterministic matching by a "fuzzy" or probabilistic correspondence, which allows for transportation of mass from a single source point to various target points (and vice versa). This phenomenon, often referred to as *mass splitting*, can be expressed via a coupling matrix $\Gamma \in \mathbb{R}_+^{n \times m}$ whose (i, j)-th entry describes the amount of mass transported from point (or bin) i to point (or bin) j. Naturally, in order for this coupling to be meaningful, the *exact* mass of the source distribution should be allocated to the target distribution without any surplus, i.e., the row and column sums of this matrix should add up to $\mathbf{a}$ and $\mathbf{b}$, respectively. The set of admissible couplings, also known as the *transportation polytope*, can be succinctly expressed as

$$\Pi(\mathbf{a}, \mathbf{b}) = \{\Gamma \in \mathbb{R}_+^{n \times m} \mid \Gamma \mathbb{1}_n = \mathbf{a},\ \Gamma^\top \mathbb{1}_n = \mathbf{b}\}. \tag{2.4}$$

This set is never empty. Indeed, it is easy to verify that $\mathbf{a} \otimes \mathbf{b} \in \Pi(\mathbf{a}, \mathbf{b})$. Considering, as before, the cost matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$, where $\mathbf{C}_{ij} = c(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})$, Kantorovich's relaxation of the problem reads:

$$\mathrm{OT}_c(\mathbf{a}, \mathbf{b}) \triangleq \min_{\Gamma \in \Pi(\mathbf{a}, \mathbf{b})} \langle \Gamma, \mathbf{C} \rangle. \tag{2.5}$$

where the notation $\mathrm{OT}_c(\mathbf{a}, \mathbf{b})$ makes explicit the dependence of the problem on the cost function c.

As before, the problem can be easily generalized from discrete (i.e. histograms) to continuous measures α and β. In that case, the couplings are now joint distributions over the product space $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, formally:

Definition 2.3.1 (Coupling). *Let $(\mathcal{X}, \alpha)$ and $(\mathcal{Y}, \beta)$ be two probability spaces. A **coupling** of α and β is a pair of random variables (X, Y) (marginally) distributed according to these measures, i.e., $X \sim \alpha$ and $Y \sim \beta$. The law of (X, Y), that is, the measure $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ for which $(X, Y) \sim \gamma$ is called the **coupling measure**, or for brevity, simply **coupling**.*

Hence, the feasible set is now the set of all such couplings between α and β, i.e.,

$$\Pi(\alpha, \beta) \triangleq \{\Gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid P_{\mathcal{X}\sharp}\Gamma = \alpha, P_{\mathcal{Y}\sharp}\Gamma = \beta\}. \tag{2.6}$$

The Kantorovich problem for general probability measures is thus:

$$\mathrm{OT}_c(\alpha, \beta) \triangleq \min_{\gamma \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y). \tag{2.7}$$

The effect of relaxation makes this formulation, unlike Monge's version (2.3), guaranteed to have a solution under very mild assumptions on the cost function [174].

Again, we emphasize that we can immediately recover Problem (2.5) from the generalized Problem (2.7) by taking discrete measures of the form

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}^{(i)}}, \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{y}^{(i)}} \tag{2.8}$$

and imposing the product measure to be of the form $\gamma = \sum_{ij} \Gamma_{ij} \delta_{(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})}$. For the

discrete case, following the discussion at the end of the previous section, we will often abuse the notation and interchangeably denote the problem and set of feasible couplings in terms of the measures themselves or their associated probability vectors, i.e., we write $\Pi(\mathbf{a}, \mathbf{b})$ or $\Pi(\alpha, \beta)$, and $\mathrm{OT}(\mathbf{a}, \mathbf{b})$ or $\mathrm{OT}(\alpha, \beta)$, interchangeably.

Whenever $\mathcal{X}$ is equipped with a metric $d_{\mathcal{X}}$, it is natural to use it as ground cost, e.g., $c(x, y) = d_{\mathcal{X}}(x, y)^p$, with $p \geq 1$. In such case, the transportation cost in Equation (2.7) is called the p-Wasserstein distance, which we denote as $\mathrm{W}_p(\alpha, \beta) \triangleq \mathrm{OT}_{d_{\mathcal{X}}^p}(\alpha, \beta)$. The case $p = 1$ is also known as the Kantorovich-Rubinstein in statistics or the Earth Mover's Distance in computer vision [149]. The Proposition below shows that these are indeed proper distances.

Proposition 2.3.2 (Proof adapted form [175]). *Assume $\mathcal{X} = \mathcal{Y}$, and suppose $(\mathcal{X}, d)$ is a metric space and that $\alpha, \beta \in \mathcal{P}(\mathcal{X})$. Then, $W_p(\alpha, \beta) = \inf_{\gamma \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y) d\gamma(x, y)$ is a metric.*

Proof. The symmetry of W_p is trivially inherited from that of d. On the other hand, since d is a metric, it must satisfy $d(x, y) = 0$ iff $x = y$. Therefore, if $\mathbb{W}_p(\alpha, \beta) = 0$ it can only be that there exists a transportation plan entirely concentrated on the diagonal ($y = x$) in $\mathcal{X} \times \mathcal{Y}$, so that $\beta = \mathrm{Id}_\sharp \alpha = \alpha$. Now consider probability measures $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{P}(\mathcal{X})$. Let (X_1, X_2) be an optimal coupling of α_1 and α_2, with associated measure $\gamma^*_{1,2}$, and analogously for (X_2, X_3) and $\gamma^*_{2,3}$ with respect to α_2 and α_3. By the Gluing Lemma [24], there exists random variables (X'_1, X'_2, X'_3) such that $(X'_1, X'_2) \overset{d}{=} (X_1, X_2)$ and $(X'_2, X'_3) \overset{d}{=} (X_2, X_3)$. Hence, (X'_1, X'_3) is a coupling of α_1 and α_3, which in turn implies:

$$
\begin{aligned}
\mathrm{W}_p(\alpha_1, \alpha_3) &\leq \left(\mathbb{E}[d(X'_1, X'_3)^p]\right)^{1/p} && \text{(optimality of } \mathrm{W}_p\text{)} \\
&\leq \left(\mathbb{E}[d_(X'_1, X'_2)^p + d(X'_2, X'_3)^p]\right)^{1/p} && (d \text{ is a metric}) \\
&\leq \left(\mathbb{E}\, d(X'_1, X'_2)^p\right)^{1/p} + \left(\mathbb{E}\, d(X'_2, X'_3)^p\right)^{1/p} && \text{(Minkowski's inequality)} \\
&= \mathrm{W}_p(\alpha_1, \alpha_2) + \mathrm{W}_p(\alpha_2, \alpha_3) && \text{(optimality of } (X'_1, X'_2)\text{)}
\end{aligned}
$$

So W_p satisfies the triangle inequality. This concludes the proof. □

Kantorovich's problem has an appealing probabilistic interpretation. It is easy to verify that equation (2.7) can be equivalently written as

$$\min_{(X,Y)} \left\{ \mathbb{E}_{(X,Y)}\, c(X,Y) \;\middle|\; X \sim \alpha, Y \sim \beta \right\} \tag{2.9}$$

for random variables X and Y with distributed according to α and β respectively, and with an underlying joint distribution given by $\gamma \in \Pi(\alpha, \beta)$.

2.4 Entropic Regularization

The high computational cost of solving the Kantorovich problem (which we discuss in the next section) has led to various schemes to solve it approximately. One of the most popular such approaches is to add an entropy regularization term to the objective [146, 45]. For this we define the discrete entropy of a coupling as

$$\mathrm{H}(\Gamma) \triangleq -\sum_{i,j} \Gamma_{ij}(\log \Gamma_{ij} - 1) = -\langle \Gamma, \log \Gamma - \mathbb{1}_{n\times m} \rangle, \tag{2.10}$$

and use it to obtain a regularized version of problem (2.5) as follows:

$$\mathrm{OT}_c^{\varepsilon}(\mathbf{a}, \mathbf{b}) \triangleq \min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \langle \Gamma, \mathbf{C} \rangle - \varepsilon\, \mathrm{H}(\Gamma). \tag{2.11}$$

To generalize this to continuous measures we will need an additional concept, the Kullback-Leibler (KL) divergence between couplings, defined as:

$$\mathrm{KL}(\Gamma \parallel \kappa) \triangleq \sum_{i,j} \Gamma_{ij} \log \tfrac{\Gamma_{ij}}{\kappa_{ij}} - \Gamma_{ij} + \kappa_{ij} \tag{2.12}$$

The following Lemma shows that we can equivalently write the entropy regularized problem it terms of this divergence.

Lemma 2.4.1. *The entropy-regularized objective* (2.11) *is equivalent to:*

$$\min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \langle \Gamma, \mathbf{C} \rangle + \varepsilon \mathrm{KL}(\Gamma \parallel \mathbf{a} \otimes \mathbf{b}) \tag{2.13}$$

Proof. With our definitions (2.10) and (2.12) it can be immediately seen that

$$\mathrm{KL}(\Gamma \parallel \mathbf{a}\otimes\mathbf{b}) = \sum_{i,j} \Gamma_{ij} \log \frac{\Gamma_{ij}}{\mathbf{a}_i \mathbf{b}_j} - \Gamma_{ij} + \mathbf{a}_i \mathbf{b}_j = -\,\mathrm{H}(\Gamma) - \sum_{i,j} \Gamma_{ij} \log \mathbf{a}_i \mathbf{b}_j + \mathbf{a}_i \mathbf{b}_j \quad (2.14)$$

For the second of these terms, we have

$$\sum_{i,j} \Gamma_{ij} \log \mathbf{a}_i \mathbf{b}_j = \langle \Gamma, \log(\mathbf{a} \otimes \mathbf{b}) \rangle = \langle \Gamma, \log \mathbf{a} \oplus \log \mathbf{b} \rangle =$$

$$= \langle \Gamma \mathbb{1}_n, \log \mathbf{a} \rangle + \langle \log \mathbf{b}, \Gamma^\top \mathbb{1}_m \rangle = \langle \mathbf{a}, \log \mathbf{a} \rangle + \langle \mathbf{b}, \log \mathbf{b} \rangle,$$

where the log acts element-wise in all cases. Therefore, only the first term in (2.14) depends on Γ, and thus it is clear that Problems (2.11) and (2.13) are equivalent. □

With this, we can naturally extend Problem (2.11) to continuous measures, thus defining an entropy-regularized version of Problem (2.7):

$$\mathrm{OT}_c^\varepsilon(\alpha, \beta) \triangleq \min_{\Gamma \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) + \varepsilon \mathrm{KL}(\Gamma \parallel \alpha \otimes \beta) \quad (2.15)$$

In analogy to the classic Wasserstein distance, when $c(x,y) = \|x-y\|^p$ we denote this as W_p^ε, dropping the p from the notation when it is clear from the context.

The first observation about this regularized formulation of optimal transport is that the objective is now ε-strongly convex, and therefore it has a unique optimal solution. Moreover, the following proposition (whose proof can be found in Peyré and Cuturi [140]) shows that entropy-regularization leads to a well-behaved approximation of the original Kantorovich problem:

Proposition 2.4.2. *Let Γ_ε^* be the unique solution of Problem* (2.11)*. Then*

$$\Gamma_\varepsilon^* \xrightarrow{\varepsilon \to 0} \operatorname*{argmin}_{\Gamma} \{ -\,\mathrm{H}(\Gamma) \mid \Gamma \in \Pi(\mathbf{a}, \mathbf{b}),\ \langle \Gamma, \mathbf{C} \rangle = \mathrm{OT}_c(\mathbf{a}, \mathbf{b}) \}$$

So that, in particular: $\mathrm{OT}_c^\varepsilon(\mathbf{a}, \mathbf{b}) \xrightarrow{\varepsilon \to 0} \mathrm{OT}_c(\mathbf{a}, \mathbf{b})$.

In other words, this result shows that the solution of the regularized problem

converges to a solution of the unregularized one, which in fact is the solution with maximal entropy.

Besides computational advantages—which we discuss in Section 2.5—regularizing the OT problem often leads to better empirical performance in applications where having denser correspondences is beneficial, e.g., when the support points correspond to noisy features [9]. If sparse matchings are nevertheless desired, there exist methods to encourage sparsity and still take advantage of regularized objectives [163, 27].

2.5 Computation

The discrete version of Kantorovich's problem (Eq (2.5)) is a linear program. Practical methods to solve it include Orlin's algorithm and interior-point methods, both of which have $O(n^3 \log n)$ complexity [138]. As discussed before, this is often prohibitive in machine learning applications, which has led to various approximations of the problem, including the celebrated entropy-regularization scheme discussed in Section 2.4.

The regularized version of discrete OT (Eq. (2.11)) is a strictly convex optimization problem. Below we show that its solution has a simple analytic expression.

Proposition 2.5.1 (adapted from [140]). *The solution to* (2.11) *is unique and has the form* $\Gamma^* = [[\mathbf{u}]]\mathbf{K}[[\mathbf{v}]]$, *for* $\mathbf{K} = e^{-\frac{\mathbf{C}}{\lambda}}$ *and some* $\mathbf{u} \in \mathbb{R}_+^n$, $\mathbf{v} \in \mathbb{R}_+^m$.

Proof. The Lagrangian of (2.11) is given by

$$\mathcal{L}(\Gamma, \mathbf{f}, \mathbf{g}) \triangleq \langle \Gamma, \mathbf{C} \rangle - \varepsilon \, \mathrm{H}(\Gamma) - \langle \mathbf{f}, \Gamma \mathbb{1}_m - \mathbf{a} \rangle - \langle \mathbf{g}, \Gamma^\top \mathbb{1}_m - \mathbf{b} \rangle$$

for dual variables $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$. Using basic algebraic manipulations and grouping all terms that do not depend on Γ in ξ, we can rewrite this as

$$\mathcal{L}(\Gamma, \mathbf{f}, \mathbf{g}) = \langle \Gamma, \mathbf{C} \rangle - \varepsilon \langle \Gamma, \Gamma - \mathbb{1}_{n \times m} \rangle - \langle \Gamma, \mathbf{f} \oplus \mathbf{g} \rangle + \xi \tag{2.16}$$

Elementary matrix calculus shows that the first order conditions are given by

$$\frac{\partial \mathcal{L}}{\partial \Gamma} = \mathbf{C} + \varepsilon \log \Gamma - \mathbf{f} \oplus \mathbf{g} = \mathbb{0}_{n \times m}$$

Hence,

$$\Gamma^* = \exp\left\{\tfrac{1}{\varepsilon}\big(\mathbf{f} \oplus \mathbf{g} - \mathbf{C}\big)\right\} = \exp(\tfrac{1}{\varepsilon}(\mathbf{f} \oplus \mathbf{g})) \odot \exp(\tfrac{1}{\varepsilon}(-\mathbf{C})) = [[\exp \tfrac{1}{\varepsilon}\mathbf{f}]] \exp(\tfrac{1}{\varepsilon}(-\mathbf{C}))[[\exp \tfrac{1}{\varepsilon}\mathbf{g}]]$$

where the last equality is a well-known property of the Hadamard product. Therefore, letting $\mathbf{u} \triangleq \exp \frac{1}{\varepsilon}\mathbf{f}$ and $\mathbf{v} \triangleq \exp \frac{1}{\varepsilon}\mathbf{g}$ the result holds. □

The vectors $\mathbf{u}$ and $\mathbf{v}$ of Proposition 2.5.1 can be obtained efficiently via the Sinkhorn-Knopp[1] algorithm, an approach popularized in the machine learning community by Cuturi [45]. With Proposition (2.5.1) in hand, its derivation is simple. Using the form of the solution Γ^*, and knowing that it must satisfy the marginal constraints, we arrive at the following conditions:

$$[[\mathbf{u}]]\mathbf{K}[[\mathbf{v}]]\mathbb{1}_m = \mathbf{a} \implies \mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a} \tag{2.17}$$

$$[[\mathbf{v}]]\mathbf{K}^\top[[\mathbf{u}]]\mathbb{1}_n = \mathbf{b} \implies \mathbf{v} \odot (\mathbf{K}^\top\mathbf{u}) = \mathbf{b} \tag{2.18}$$

Whence the optimal $\mathbf{u}, \mathbf{v}$ can be found via fixed-point iterations, i.e., computing in alternation:

$$\mathbf{u} \leftarrow \mathbf{a} \oslash \mathbf{K}\mathbf{v} \quad \text{and} \quad \mathbf{v} \leftarrow \mathbf{b} \oslash \mathbf{K}^\top\mathbf{u}, \tag{2.19}$$

where we recall that $\oslash$ denotes entry-wise division. Regardless of the initialization, these iterations converge to the same solution Γ^* (albeit with potentially different vectors $\mathbf{u}$, $\mathbf{v}$) [140]. Altschuler et al. [3] showed that after $O(\|\mathbf{C}\|_\infty^3 \log(n)\tau^{-3})$ iterations of Sinkhorn's algorithm, the reconstructed solution $\mathbf{P}$ satisfies $\langle \mathbf{P}, \mathbf{C} \rangle \leq \mathrm{OT}(\mathbf{a}, \mathbf{b}) + \tau$, which in turn implies that this method provides a τ-approximate solution of the unregularized Kantorovich problem in $O(n^2 \log n \tau^{-3})$ time.

[1] Although seemingly first proposed a century ago by Yule [180], this simple algorithm has been rediscovered in many contexts, and is thus known by a myriad of names: *Sinkhorn*, *Sinkhorn-Knopp*, *Bregman Iterations*, *soft-assign*, *iterative proportional fitting procedure*, among others. The first proof of convergence is due to Sinkhorn [161], which is where it gets its most popular name.

2.6 Theoretical Guarantees

2.6.1 Euclidean case

Whenever optimal transport is used primarily with the goal of obtaining correspondences (as opposed to only as a means to compute a cost/distance between distributions), there are various theoretical considerations that become crucial.

The first of such considerations pertains to the nature of the solution, i.e., the optimal coupling γ^* which minimizes the cost (2.7). When the end goal is to transport points from one space to the other, the best-case scenario would be if the optimal γ happens to be a "hard" deterministic mapping. A celebrated result by Brenier [33, 34] shows that this indeed the case for the quadratic cost,[2] i.e., for the 2-Wasserstein distance. Even when solving the problem approximately with entropic regularization (Eq. (2.15)), this result guarantees that the solution found in this way converges to a deterministic mapping as $\varepsilon \to 0$.

Now, assuming now that such a map exists, the next aspect we might be interested in is its smoothness. Intuitively, smoothness of this mapping is desirable since it is more likely to lead to robust matchings in the context of correspondences, even if, again, the argument holds asymptotically for the regularized problem. This, clearly, is a very strong property to require. While not even continuity can be guaranteed in general [13], again for the quadratic-cost things are simpler: if the source and target densities are smooth and the support of the target distribution satisfies suitable convexity assumptions, the optimal map is guaranteed to be smooth too [36, 37].

2.6.2 Riemannian manifold case

Extending the problem beyond Euclidean to more general spaces has been one of the central questions theoretical optimal transport research over the past decades [174]. For obvious reasons, here we focus the discussion on results related to hyperbolic spaces, and more generally, to Riemannian manifolds.

[2]This result holds in more general settings. We refer the reader to [154, 13] for further details.

Let us first note that Problem (2.7) is well-defined for any complete and separable metric space $\mathcal{X}$. Since the arc-length metric of a Riemannian manifold allows for the direct construction of an accompanying metric space $(\mathcal{X}, d_{\mathcal{X}})$, then OT can be defined over those too. However, some of the theoretical results of their Euclidean counterparts do not transfer that easily to the Riemannian case [13]. Nevertheless, the existence and uniqueness of the optimal transportation plan γ^*, which in addition is induced by a transport map T, can be guaranteed with mild regularity conditions on the source distribution α. This was first shown in seminal work by McCann [124]. The result, which acts as a Riemannian analogue of that of Brenier for the Euclidean setting [33], is shown below as presented by Ambrosio and Gigli [13]:

Theorem 2.6.1 (McCann, version of [13]). *Let M be a smooth, compact Riemannian manifold without boundary and $\alpha \in \mathcal{P}(\mathcal{M})$. Then the following are equivalent:*

(i) $\forall \beta \in \mathcal{P}(\mathcal{M})$, there exists a unique optimal $\gamma \in \Pi(\alpha, \beta)$, and this plan is induced by a map T.

(ii) α is regular.

If either (i) or (ii) holds, the optimal T can be written as $x \mapsto \exp_x(-\nabla\phi(x))$ for some c-concave function $\phi : \mathcal{M} \to \mathbb{R}$.

The question of regularity of the optimal map, on the other hand, is much more delicate now than in the Euclidean case [13, 123, 120]. In addition to the suitable convexity assumptions on the support of the target density, a restrictive structural condition, known as the Ma-Trudinger-Wang (MTW) condition [123], needs to be imposed on the cost in order to guarantee continuity of the optimal map.

Unfortunately for the setting of Chapter 5, in the case of Riemannian manifolds the MTW condition for the usual quadratic cost $c = d^2/2$ is so restrictive that it implies that $\mathcal{X}$ has non-negative sectional curvature [120], which rules out hyperbolic spaces. However, a recent sequence of remarkable results Lee and Li [114] and Li [118] prove that for simple variations of the Riemannian metric d on hyperbolic spaces, smoothness is again guaranteed:

Theorem 2.6.2 (Lee and Li, [114]). *Let d be the Riemannian distance function on a manifold of constant sectional curvature -1; then the cost functions $-\cosh \circ d$ and $-\log \circ (1+\cosh) \circ d$ satisfy the strong MTW condition, and the cost functions $\pm \log \circ \cosh \circ d$ satisfy the weak MTW condition.*

Consequently, matching approaches based on optimal transport over hyperbolic spaces —as the one proposed in Chapter 5— are well supported by theory too, with the results presented here guaranteeing the existence of ideal (even if perhaps unachievable) smooth solutions, suggesting that even approximate solutions to OT objectives are likely to yield relatively stable correspondences.

2.7 Optimal Transport as a Learning Loss

Recent work has proposed using Wasserstein distances as differentiable loss functions, particularly in the context of deep generative modeling [14, 71, 153]. When used as a loss function, the entropy-regularized version (Eq. (2.15)) has the undesirable property that $\mathrm{W}_{p,\varepsilon}(\alpha, \alpha) \neq 0$, in addition to having biased sample gradients [21].

In response to this, various recent works Genevay et al. [71], Bellemare et al. [21], and Salimans et al. [153], consider instead the *Sinkhorn Divergence*:

$$\mathrm{SD}^{\varepsilon}(\alpha, \beta) \triangleq \mathrm{W}_p^{\varepsilon}(\alpha, \beta) - \tfrac{1}{2}\big(\mathrm{W}_p^{\varepsilon}(\alpha, \alpha) + \mathrm{W}_p^{\varepsilon}(\beta, \beta)\big). \tag{2.20}$$

Besides being a proper divergence and providing unbiased gradients, this function is convex, smooth and positive-definite [62], and its sample complexity is well characterized [69], all of which make it an appealing loss function.

Chapter 3

Optimal Transport with Structured Costs

This chapter is based on Alvarez-Melis, Jaakkola, and Jegelka [9].

In this chapter, we develop a framework to incorporate structural information directly into the cost objective of the optimal transport problem. This novel formulation opens avenues to a much richer class of (nonlinear) cost functions, allowing us to encode known or desired interactions of mappings, such as grouping constraints, correlations, and explicitly modeling topological information that is present, for instance, in sequences and graphs.

Our main tool for modeling structure is *submodularity* – a fundamental concept from combinatorial optimization, which we review in Section 3.2.1. Submodular functions possess two highly desirable properties for our problem: (1) they naturally encode combinatorial structure, via diminishing returns and as combinatorial rank functions; and (2) their geometry leads to efficient algorithms. Indeed, the tractability of this novel nonlinear formulation of OT arises from the polytopes induced by submodular cost functions.

The resulting combination of the geometries of transportation and submodularity leads to a problem with rich, favorable polyhedral structure and connections to game theory and saddle point optimization. We leverage this structure to solve this *submodular optimal transport* problem via a saddle-point mirror-prox algorithm

involving alternating projections onto the polytope defined by the transportation constraints and the base polytope associated with the submodular cost function. The former can be done efficiently through Sinkhorn iterations, while the latter, can be solved exactly in $O(n \log n)$ time for a suitable class of submodular functions.

Via various applications and experiments, we explore the characteristics of the solutions to this novel transportation problem and demonstrate the efficiency of our algorithms. We show how different submodular functions yield solutions that interpolate between strictly structure-aware transportation plans and structure-agnostic regularized versions of the problem. Besides these synthetic experiments, we evaluate our framework in various real-life applications: domain adaption for digit classification, color transfer, and sentence similarity prediction. In both cases, introducing structure leads to better empirical results.

3.1 Motivation and Applications

A concrete example of the need to include structure arises when applying optimal transport to domain adaptation, where a subset of the source points to be matched have known class labels. In this case, we may desire source points with the same label to be matched coherently to the same compact region of the target space, preserving compact classes, and not be split into disjoint, distant locations [44].

In the context of randomized experiments, when pairing control and treatment units in observational studies of treatment effects it is beneficial to compare treated and control subjects from the same "natural block" (e.g., family, hospital) so as to minimize the difference between unmeasured covariates [143]. In all these examples, the additional structure essentially seeks correlations in the mappings of "similar" source points. Such dependencies, however, cannot be induced by standard formulations of optimal transport whose cost is separable in the mapping variables;[1] they require nonlinear interactions.

[1]The original optimal transport formulation with cost $\sum_{ij} c_{ij}\Gamma_{ij}$ is linear in the mappings Γ_{ij}, Γ_{kl} of separate source locations i, k; the mappings are counted independently.

A different motivation comes from the need to compare probability distributions over combinatorial objects (e.g., sequences, trees or graphs), in a manner that takes into account both the node-level properties (i.e, the *atom-level geometry*) and the relationship between them (*the graph-level topology*). This is a common problem with applications to computing similarity between sentences [106], phylogenetic trees [46] or social networks [23].

3.2 Preliminaries

3.2.1 Submodularity

A set function $F : 2^V \to \mathbb{R}$ over a ground set V of items is called *submodular* if it satisfies *diminishing returns*: for all sets $S \subseteq T \subseteq V$ and all element v in $V \setminus T$, it holds that

$$F(S \cup \{v\}) - F(S) \geq F(T \cup \{v\}) - F(T). \tag{3.1}$$

Equivalently, submodularity can be characterized by the following union-intersection property:

$$\forall S, T \subseteq V \qquad F(S) + F(T) \geq F(S \cup T) + F(S \cap T).$$

The function F is called supermodular if $-F$ is submodular, and modular if it is both sub- and supermodular.

It is easy to see that modular functions are *linear* in the sense that they can always be written as $F(S) = \sum_{e \in S} w(e)$ for some weight function $w : V \to \mathbb{R}$. This result suggests that we can identify modular set functions over a ground set V of n items with vectors in $\mathbb{R}^n$, by labeling (without loss of generality) these items with positive integers (i.e., $V = [\![n]\!]$), and defining $y_S = \sum_{i \in S} \mathbf{y}_i$. With this, every modular function F corresponds to a unique $\mathbf{y}^F \in \mathbb{R}^n$, and we have $F(S) \equiv y_S^F$. We will use this notation even in the more general case where F is not modular.

In motivating the importance of submodularity, Lovász [121] points out that submodular set functions play a similar role in discrete optimization to that of convex functions in continuous optimization. He bases this analogy in the fact that

they both occur naturally in many contexts, are preserved under various operations, have sufficient structure to yield a mathematically beautiful theory and come with efficient minimization. The analogy to convexity is nevertheless delicate and often counter-intuitive, because even though certain properties of submodularity are indeed shared with convexity (e.g., convex relaxations, duality theory), some other are in fact reminiscent of concavity (e.g., diminishing derivative).

As mentioned before, the tractability of submodular functions arises from the polytopes they define, and to which we now turn our attention. Every submodular function has an associated polyhedron (known as the *base polyhedron*), which is given by

$$\mathcal{P}_F \triangleq \{\mathbf{y} \in \mathbb{R}^n \mid \sum_{s \in S} y_s \leq F(S) \text{ for all } S \subseteq V\}.$$

An intuitive way to understand this object is the following. For every subset S of the n axes in $\mathbb{R}^n$, F implicitly defines a hyperplane which separates those points $\mathbf{y} \in \mathbb{R}^n$ for which $y_S \leq F(S)$—call this the set of *compatible* points—from those for which $y_S > F(S)$. Note that there are 2^n of these hyperplanes. The intersection of all the sets of compatible points is the base polyhedron. The *base polytope* of F is the *hyper*-face of this polyhedron for which the inequalities are active, that is,

$$\mathcal{B}_F \triangleq \{\mathbf{y} \in \mathbb{R}^{|V|} \mid y_V = F(V);\ y_S \leq F(S)\ \forall S \subseteq V\} = \{y \in \mathcal{P}_F \mid \sum_{s \in S} y_s = F(S)\}.$$

Base polytopes generalize matroid polytopes (convex hulls of combinatorial "independent sets"), and lead to strong links with convexity.

Another fundamental concept in submodular optimization is the *Lovász extension* of a set function F, which extends its domain from 2^V to $\mathbb{R}^n_+$ [121]. For any $w \in \mathbb{R}^n_+$, order its coordinates so that $w_1 \geq \cdots \geq w_n$ and define $w_{n+1} = 0$ and $S_j = \{i \,|\, w_i \geq w_j\}$. The Lovász extension f of F is

$$f(w) = \sum_{j=1}^{n} (w_j - w_{j+1}) F(S_j). \tag{3.2}$$

Fortunately, when F is submodular its Lovász extension can be equivalently expressed

in a much simpler form, as a support function over the base polytope:

$$f(w) = \max_{\mathbf{x} \in \mathcal{B}_F} \mathbf{w}^\top \mathbf{x}. \tag{3.3}$$

It is readily apparent that, in this case, f is convex. In fact, it was shown by Lovász himself that f is convex if and only if F is submodular [121]. Since we will be dealing with submodular functions throughout this chapter, our treatment of the Lovász extension will always rely on the expression (3.3). In fact, one of the two proposed optimization approaches will crucially rely on this definition of f as the solution of a maximization problem.

The tractability of submodular minimization that we have been alluding to throughout this section rests upon two main ingredients. First, the Lovász extension gives an exact relaxation of its corresponding submodular function, in the sense that the convex hull of all minimizers of the discrete function F is exactly the set of minimizers of the extension [19]. Second is the fact that —despite the exponentially many constraints— linear optimization over the base polytope can be done efficiently. Pioneering work by Edmonds [54] showed that the Lovász extension can be computed efficiently by a form of sorting (thus, in only $O(n \log n)$ time) which is often known as Edmonds' algorithm. In fact, this algorithm is constructive, i.e., it returns the maximizing argument $\mathbf{x}^*$ in Equation (3.3). This, in turn, implies that subgradients of f can be obtained by the same procedure.

Although beyond the scope of this thesis, it is worth mentioning that while unconstrained submodular minimization is tractable and various efficient polynomial algorithms for it exist [78, 115, 38], even slight variations of the problem quickly become intractable. For example, simple constraints like cardinality lower bounds make the problem NP-hard. Similarly, submodular *maximization* is usually NP-hard in most settings, admitting only polynomial time constant approximation algorithms [134, 60]. We refer the interested reader to the surveys by Krause and Golovin [105] and Bach [19] for further details.

3.3 Optimal Transport with Submodular Costs

In the classical formulation of optimal transport (2.5), the total cost $\langle \Gamma, \mathbf{C} \rangle$ is linear in the decision variables Γ. This means each potential pairwise assignment Γ_{ij} (i.e., every pair $(\mathbf{a}_i, \mathbf{b}_j)$) is treated independently. But, in some applications, it is desirable to bias certain points to be mapped *together*, i.e., to introduce dependencies between assignments. In our running example of domain adaptation, we want points from the same class to be transported "together". Intuitively, the joint cost of mapping points from the same class to close-by target points should be lower than splitting them apart, even if the transportation distances are the same.

More generally, we might want to encourage mappings of subspaces to subspaces, or, on the contrary, discourage some combinations of assignments. A flexible framework to express such interactions over discrete choices is via submodular functions [119, 93, 100]. Intuitively, property (3.1) implies that the marginal cost of an additional element decreases as more "compatible" items have already been chosen, and thus it is relatively *cheaper* to select compatible items together (e.g., items from the same group) than non-compatible ones.

To see how submodularity can be leveraged for optimal transport, consider for a moment Monge's formulation (2.3), where we seek a matching of the elements in U and V with minimal cost. Any matching can be expressed as a set of edges $S = \{(u_1, v_1), \ldots, (u_k, v_k)\}$, and its cost as a set function $F : 2^{|U| \times |V|} \to \mathbb{R}^+$. Under this formulation, the classic definition of optimal transport uses a *modular* cost function:

$$F(S) = \sum_{(u,v) \in S} c_{uv},$$

so the cost of the additional match (u, v) is the same, namely c_{uv}, regardless of what assignments have already been made. If we let F be submodular instead, property (3.1) implies that the marginal cost of additional edges decreases as the set of matches grows. The magnitude of decrease depends on S, the new item v, and the choice of F. We will channel this decrease to occur only when the additional "item" (assignment

(u, v)) is compatible with already chosen "items".

3.3.1 Submodular cost functions

The rich class of submodular functions allows various types of structural information (compatibility) to be encoded in the cost function. As an example, recall the local consistency structure induced by class labels in domain adaptation. We may divide the support of the source and target distributions α and β into regions (subsets of samples) $U_k \subset U$ and $V_l \subset V$. These induce a partition of the set of assignments too:

$$E_{kl} := \{(u, v) \mid u \in U_k,\ v \in V_l\}.$$

Now define

$$F(S) := \sum_{kl} F_{kl}(S \cap E_{kl}), \tag{3.4}$$

where each F_{kl} is submodular with reduced support E_{kl}. One possible choice for F_{kl} is

$$F_{kl}(S) = g_{kl}\Big(\sum_{(u,v)\in S\cap E_{kl}} C_{uv}\Big), \tag{3.5}$$

where $C_{ij} \in \mathbb{R}^+$ is the ground metric cost between x_i^s and x_j^t, and $g_{kl} : \mathbb{R} \to \mathbb{R}$ are scalar monotone increasing concave functions whose effect is to dampen the cost of additional edges between the partitions U_l and V_k, thus encouraging edge selections that map most of the mass in U_l to the same V_k. To grant discounts only after a sufficient number of assignments have been chosen from a group, we may use an explicit threshold, e.g.,

$$g_{kl}(x) = \min\{x, \alpha\} + \sqrt{[x - \alpha]_+}. \tag{3.6}$$

We use such functions in the clustered point matching, domain adaptation and sentence similarity experiments in Section 3.6. We may also use subspaces for encoding structure. For example, a smoother grouping of assignments (u, v) could be encoded by stacking feature vectors for u and v into one vector $\phi(u, v)$ and taking $F(S) = \text{rank}(\Phi_S)$,

i.e., the rank of the matrix of features of the selected assignments, or the volume $F(S) = \log\det(\Phi_S^\top \Phi_S)$. This function captures discrete groups if the feature vectors are indicator vectors of groups. Other important examples include hierarchical structures and coverage functions.

3.3.2 Submodular optimal transport

The functions defined above have discrete domains, i.e., they correspond to discrete matchings, but we really seek a formulation like that of original Problem (2.5) with continuous, fractional assignments. The key to obtaining a nonlinear, structured analog of Kantorovich's problem is the convex *Lovász extension* f of the submodular function F. The above intuitions and effects carry over, and we define the *submodular optimal transport* problem as

$$\min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} f(\Gamma) \quad \equiv \quad \min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \max_{\kappa \in \mathcal{B}_F} \langle \Gamma, \kappa \rangle. \tag{3.7}$$

The right hand side follows since the Lovász extension is also the support function of the submodular base polytope. This relaxation has another advantage: while the discrete version is hard to even solve approximately [72], problem (3.7) is a convex optimization problem on Γ.

This new structured optimal transport problem recovers many desirable properties of the original optimal transport formulation. For example, for suitable choices of cost function, the "distance" implied by it is a semi-metric, as we show in the following theorem.

Theorem 3.3.1. *Suppose the ground cost $C(\cdot,\cdot)$ is a metric and that F is a submodular non-decreasing function such that $F(\emptyset) = 0$ and $F(\{(i,j)\}) > 0$ iff $C(x_i, y_j) > 0$. Then $d_F(\alpha,\beta) = \min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} f(\Gamma)$ is a semi-metric.*

Proof. Let $\mathbf{C} \in \mathbb{R}^{n \times m}$ be the cost matrix associated with c, i.e. $\mathbf{C}_{ij} = c(x_i, y_j)$ for $i \in [\![n]\!]$ and $j \in [\![m]\!]$. In addition, let $\mathbf{a}$ and $\mathbf{b}$ be the vectors of probability weights of α and β, respectively, i.e. $\alpha = \sum_i^n \mathbf{a}_i \delta_{\mathbf{x}^{(i)}}$ and $\beta = \sum_j^m \mathbf{b}_j \delta_{\mathbf{y}^{(j)}}$.

Since $c(\cdot,\cdot)$ is a metric, every $\mathbf{C}_{ij}$ is non-negative. Furthermore, since we assume support points are not duplicated, $\mathbf{C}$ has at most n zero entries, and the rest are strictly positive. This, combined with the fact that F is non-decreasing, implies $F(S) \geq 0$ for every $S \subseteq V$, and therefore its Lovász extension must also be non-negative. In particular,

$$d_F(\alpha,\beta) = \min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} f(\Gamma) \geq 0 \qquad \forall \alpha,\beta \tag{3.8}$$

Now, suppose $\alpha = \beta$, and without loss of generality, assume the support points are indexed such that $\mathbf{x}^{(i)} = \mathbf{y}^{(i)}$ for every i. In addition, we must have $\mathbf{a} = \mathbf{b}$, so $\Gamma = [[\mathbf{a}]] \in \Pi(\mathbf{a},\mathbf{b})$. On the other hand, since c is a metric $\mathbf{C}_{ii} = 0$ for every i, which in turn implies that for any $\kappa \in \mathcal{B}_F$ and every i, $\kappa_{ii} \leq F(\{i,i\}) = 0$. By (3.8) and the minimax equilibrium properties, we have

$$0 \leq d_F(\alpha,\beta) = \langle \Gamma^*, \kappa^* \rangle \leq \langle \Gamma, \kappa^* \rangle \qquad \forall \Gamma \in \Pi(\mathbf{a},\mathbf{b})$$

In particular, for $\Gamma = \mathrm{diag}(\mathbf{a})$, we get

$$0 \leq d_F(\alpha,\beta) \leq \sum_i \mathbf{a}_i \kappa^*_{ii} \leq 0$$

So we conclude that $d_F(\alpha,\beta) = 0$. Conversely, let $d_F(\alpha,\beta) = 0$, and suppose, for the sake of contradiction, that $\alpha \neq \beta$. Then, at least one of the following is true:

(i) $\mathbf{a} \neq \mathbf{b}$

(ii) the support points are different, i.e. there is no reordering of indices such that $\mathbf{x}^{(i)} = \mathbf{y}^{(i)}$ for every i.

If (i) is true, $\Pi(\mathbf{a},\mathbf{b})$ cannot be a permutation matrix, so in particular Γ^* has at least $n+1$ positive entries. We can thus find a $\kappa \in \mathcal{B}_F$ which has positive weights in all those entries. In that case, we have $\langle \Gamma^*, \hat{\kappa} \rangle > 0$, a contradiction. Now, if on the other hand (ii) is true, then $\mathbf{C}$ has strictly less than n zero entries. This, by our assumptions on F, means that there exist $\kappa \in \mathcal{B}_F$ with less than n non negative entries. Any such matrix will have $\langle \Gamma^*, \kappa \rangle > 0$, a contradiction.

Finally, the symmetry of $d_F(\alpha,\beta)$ is trivial. □

Problem (3.7) suggests two possible approaches for computing the optimal transport plan Γ^*. The left-hand side is a non-smooth but convex optimization problem, which can be solved via subgradient methods. Alternatively, the minimax form is a *smooth* convex-concave optimization over nonempty, closed and convex sets.[2] Therefore, (3.7) is a convex-concave saddle-point problem [95]. The solutions $z^* := (\Gamma^*, \kappa^*)$ of this problem, i.e., the *saddle points* $\phi := \langle \cdot, \cdot \rangle$ in $\mathcal{Z} := \Pi(\mathbf{a}, \mathbf{b}) \times \mathcal{B}_F$, satisfy

$$\phi(\Gamma^*, \kappa) \leq \phi(\Gamma^*, \kappa^*) \leq \phi(\Gamma, \kappa^*) \quad \forall \Gamma \in \Pi(\mathbf{a}, \mathbf{b}), \kappa \in \mathcal{B}_F$$

This formulation gives rise to a primal-dual pair of convex optimization problems:

$$\text{Opt}(P) = \min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \bar{\phi}(\Gamma), \quad \bar{\phi}(\Gamma) := \sup_{\kappa \in \mathcal{B}_F} \phi(\Gamma, \kappa) \tag{3.9}$$

$$\text{Opt}(D) = \max_{\kappa \in \mathcal{B}_F} \underline{\phi}(\kappa), \quad \underline{\phi}(\kappa) := \sup_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \phi(\Gamma, \kappa) \tag{3.10}$$

If a saddle point (Γ^*, κ^*) exists, it must be a primal-dual optimal pair and $\text{Opt}(P) = \text{Opt}(D)$. The *saddle-point gap* quantifies the accuracy of a candidate solution $(\hat{\Gamma}, \hat{\kappa})$:

$$\Delta_{\text{sp}} = \sup_{\Gamma} \phi(\Gamma, \hat{\kappa}) - \inf_{\kappa} \phi(\hat{\Gamma}, \kappa) = [\bar{\phi}(\Gamma) - \text{Opt}(P)] - [\text{Opt}(D) - \underline{\phi}(\kappa)] \tag{3.11}$$

Since ϕ is continuous and convex-concave, and $\Pi(\mathbf{a}, \mathbf{b}), \mathcal{B}_F$ are convex and bounded, a solution always exists.

Although more involved than the alternative convex optimization approach, this saddle-point formulation results in a smooth objective, which allows for the use of methods with $O(\frac{1}{t})$ convergence rate instead of $O(\frac{1}{\sqrt{t}})$. This, however, comes at the price of a higher cost per iteration. We analyze these opposing effects theoretically in the next section and empirically in Section 3.6. Beyond these computational issues, the saddle-point formulation provides interesting interpretations of the structured optimal transport problem through the lens of minimax optimization and its well-known connections to game theory and robust optimization.

[2] $\Pi(\mathbf{a}, \mathbf{b}), \mathcal{B}_F$, being polytopes, are closed and convex. Note $\Pi(\mathbf{a}, \mathbf{b})$ is always nonempty since $\mathbf{a}\mathbf{b}^\top \in \Pi(\mathbf{a}, \mathbf{b})$, and so is $\mathcal{B}_F$ [19].

3.4 Two Interpretations of the Objective

3.4.1 Games over polytopes

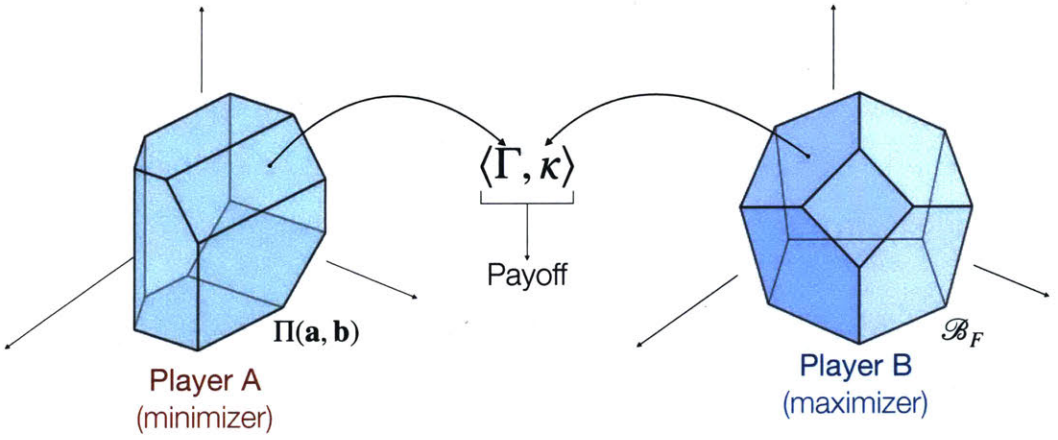


Figure 3-1: Schematic representation of the submodular optimal transport objective from a game-theoretic perspective.

The minimax formulation (3.7) is a *min-max strategy polytope* (MSP) game [80]: a two-player zero-sum game with strategies played over polytopes with payoff function $\langle \Gamma, \kappa \rangle$. In this optimal transport game, Player A (the *minimizer*) chooses a transport plan Γ between α and β, and Player B (the *adversary*) chooses a cost matrix κ from the set of *admissible* costs, i.e., those that lie on the base polytope defined by the submodular cost function F. After this, Player A pays $\langle \Gamma, \kappa \rangle$ to Player B. Since the game is guaranteed to have a Nash equilibrium, there is a pair of transport plan Γ^* and cost matrix κ^* such that Γ^* is optimal for fixed cost κ^* and vice-versa.

The shape and size of the adversary's strategy polytope $\mathcal{B}_F$, an $nm-1$ dimensional set in $\mathbb{R}^{n \times m}$, depends on the characteristics of F. The "more submodular" this function is—i.e., the earlier and sharper the marginal costs decrease—the larger $\mathcal{B}_F$ is. If F is modular, the base polytope collapses to a single point, that is, Player B plays a fixed strategy: a ground cost matrix $\mathbf{C}$. The problem then reduces to $\min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \langle \Gamma, \mathbf{C} \rangle$: the traditional optimal transport problem (2.5).

3.4.2 Worst-case robust optimization

Problem (3.7) can also be viewed in the light of *robust optimization* [22, 25], where uncertain observations are treated in a worst-case scenario. This paradigm is useful when one is interested in avoiding solutions that are highly sensitive to small perturbations or noise in the problem parameters. Through this lens, our formulation of structured optimal transport could then be viewed as a transportation problem with uncertain cost matrix κ, where we aim for a solution that is robust to any fluctuation of costs within the confidence set $\mathcal{B}_F$. In other words, we seek robustness with respect to the uncertainty cost set defined by F's base polytope.

3.5 Optimization

3.5.1 A case for proximal methods

Most popular first-order optimization methods for constrained convex problems fall into one of two categories: conditional gradient and proximal methods. Methods in the former class, like the Frank-Wolfe algorithm, require solving linear minimization oracles (LMO) as a subroutine. In the case of (3.7), this means solving a classic (non-regularized) optimal transport problem in each iteration, which might be prohibitively expensive for the applications of interest.

On the other hand, proximal methods require mirror map computations and projections. The choice of mirror map is crucial for the efficiency of these methods, and it should take into account the geometry of the constraint set. Only if the resulting projections can be easily computed are proximal methods an attractive alternative. As we show below, for appropriately chosen mirror maps this is the case for both constraint sets in problem (3.7).

Over the next three sections, we adapt three popular proximal methods to our context. We can solve the left-hand side of Problem (3.7) using the mirror descent algorithm (MDA). For the minimax formulation, on the other hand, we can use either saddle-point mirror-descent (SP-MD) or saddle-point mirror-prox (SP-MP) [95,

96]. As we will see in Section 3.5.6, these two alternative approaches (convex vs. saddle-point) imply a trade-off between cost-per-iteration and convergence rate. In a way, these are meta-algorithms that rely on lower-level subroutines that compute gradients and projections. We discuss these subroutines in detail in Section 3.5.4. Finally, we put all the pieces together in Section 3.5.6 where we provide pseudo-code for the algorithms and complexity analysis for each.

3.5.2 Mirror descent

For a closed convex set $\mathcal{X}$ (the transportation polytope $\Pi(\mathbf{a}, \mathbf{b})$ in our case) and Lipschitz continuous convex objective f (the Lovász extension), the Mirror Descent Algorithm (MDA) requires the choice of a mirror map $\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma)$. Here, we take this to be the entropy map, i.e., $\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma) \triangleq \mathrm{H}(\Gamma)$. It also requires access to subgradients of the objective.

The MDA consists of iteratively computing:

a) $w_{t+1} \in D$ such that $\nabla\Phi(w_{t+1}) = \nabla\Phi(\Gamma_t) - \eta\kappa_t$, for $\kappa_t \in \partial f(\Gamma_t)$

b) $\Gamma_{t+1} \in \operatorname{argmin}_{\Gamma\in\mathcal{X}} D_\Phi(z, w_{t+1})$

For the choice of entropic mirror map, we have: $\nabla\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma) = \mathbf{1} + \log\Gamma$ (where the logarithm is to be understood element-wise), so the condition in step (a) becomes:

$$\log w^{\Gamma}_{t+1} = \log\Gamma_t - \eta\kappa_t \tag{3.12}$$

Hence,

$$w^{\Gamma}_{t+1} = \Gamma_t \odot e^{\eta\kappa_t},$$

where the product and exponential are, again, element-wise. Step (b) requires projecting w_{t+1} into $\Pi(\mathbf{a}, \mathbf{b})$ according to the Bregman divergence associated with the mirror maps $\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma)$. For the entropy map, this becomes an KL-divergence projection, so we have

$$\Gamma_{t+1} \in \operatorname*{argmin}_{\Gamma} \mathrm{KL}(\Gamma \parallel \Gamma^t \odot e^{\eta\kappa_t}) \tag{3.13}$$

Therefore, it only remains to discuss how to compute subgradients of the Lovász extension and how to project (in the KL-sense) onto the transportation polytope, which we describe in detail in Section 3.5.4.

3.5.3 Saddle-Point mirror descent and mirror-prox

The setting for the Saddle-Point Mirror Descent (SP-MD) and Saddle Point Mirror-Prox (SP-MP) algorithms is the same, which we now introduce. We consider a joint variable $z = (\Gamma, \kappa) \in \mathcal{Z} := \Pi(\mathbf{a}, \mathbf{b}) \times \mathcal{B}_F$ and let $\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma)$ and $\Phi_{\mathcal{B}_F}(\kappa)$ be mirror maps on $\Pi(\mathbf{a}, \mathbf{b})$ and $\mathcal{B}_F$, respectively. Then, the mirror map for z is defined as $\Phi(z)_{\mathcal{Z}} = \Phi_{\mathcal{Z}}(\Gamma, \kappa) = \Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma) + \Phi_{\mathcal{B}_F}(\kappa)$. On the other hand, we assume access to a first-order oracle to obtain subgradients:

$$\partial\Phi_{\mathcal{Z}}(z) = \{\partial_\Gamma[\Phi_{\mathcal{Z}}(\Gamma, \kappa)]\} \times \{\partial_\kappa[-\Phi_{\mathcal{Z}}(\Gamma, \kappa)]\}.$$

Thus, both the gradient computation and projection decouple over κ and Γ, and we can use the projections described in Section 3.5.4. Below, we derive the steps for SP-MD, the (simpler) SP-MD is analogous with a single Sinkhorn/projection step.

The SP-MD algorithm computes at every step:

a) $w_{t+1} \in D$ such that $\nabla\Phi_{\mathcal{Z}}(w_{t+1}) = \nabla\Phi_{\mathcal{Z}}(z_t) - \eta g_t$

b) $z_{t+1} \in \operatorname{argmin}_{z\in\mathcal{Z}} D_\Phi(z, w_{t+1})$

Note that $\partial\Phi_{\mathcal{Z}} = (\nabla\Phi_{\Pi(\mathbf{a},\mathbf{b})}, \nabla\Phi_{\mathcal{B}})$, so (a) amounts to finding $w_{t+1} = (w^\Gamma_{t+1}, w^\kappa_{t+1})$ such that:

$$\nabla\Phi_{\Pi(\mathbf{a},\mathbf{b})}(w^\Gamma_{t+1}) = \nabla\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma_{t+1}) - \eta\kappa_t \tag{3.14}$$

$$\nabla\Phi_{\mathcal{B}}(w^\kappa_{t+1}) = \nabla\Phi_{\mathcal{B}}(\kappa_{t+1}) + \eta\Gamma_t \tag{3.15}$$

At this point, the updates take different forms depending on the mirror maps. For our choice of $\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma) = H(\Gamma)$, we have $\nabla\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma) = \mathbf{1} + \log\Gamma$ (where the

logarithm is to be understood element-wise), so (3.14) becomes:

$$\log w_{t+1}^{\Gamma} = \log \Gamma_t - \eta \kappa_t \tag{3.16}$$

Hence,

$$w_{t+1}^{\Gamma} = \Gamma_t \odot e^{\eta \kappa_t},$$

where the product and exponential are, again, element-wise. On the other hand, for the mirror map $\Phi_{\mathcal{B}}(\kappa) = \frac{1}{2}\|\kappa\|_2^2$, Equation (3.15) becomes

$$w_{t+1}^{\kappa} = \kappa_t + \eta \Gamma_t \tag{3.17}$$

The second step in SP-MD (step (b) above) requires projecting w_{t+1} and thus $(w_{t+1}^{\Gamma}, w_{t+1}^{\kappa})$ into $(\Pi(\mathbf{a}, \mathbf{b}), \mathcal{B}_F)$ according to the Bregman divergences associated with the mirror maps $\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma), \Phi_{\mathcal{B}}(\kappa)$. For the entropy map, this becomes an KL-divergence projection, so we have

$$\Gamma_{t+1} \in \operatorname*{argmin}_{\Gamma} \mathrm{KL}(\Gamma \parallel \Gamma^t \odot e^{\eta \kappa_t}). \tag{3.18}$$

On the other hand, the divergence associated with the ℓ_2 norm map is again an ℓ_2 distance, so

$$\kappa_{t+1} \in \operatorname*{argmin}_{\kappa} \|\kappa - \kappa_t + \eta \Gamma_t\|_2^2. \tag{3.19}$$

3.5.4 Subroutines: projections and subgradients

In this section, we describe in detail the computation of the three main subroutines used by the top-level optimization algorithms described above: subgradients of the Lovász extension, projections onto the transportation polytope and projections onto the submodular base polytope.

Subgradients of f

The subdifferential of f is

$$\partial f(\Gamma) = \operatorname*{argmax}_{\kappa \in \mathcal{B}_F} \langle \kappa, \Gamma \rangle.$$

Thus, a subgradient of f is computed by a linear optimization over the base polytope, which, despite exponentially many constraints, can be solved by a simple sort via Edmonds' greedy algorithm in $O(N \log N)$ time, where $N = n \times m$ is the dimension of Γ.

Let f be the Lovász extension of a submodular function $F : 2^V \to \mathbb{R}$. Then f can be evaluated at $w \in \mathbb{R}^n$ as follows. Let σ be a reordering of the elements of V such that $w_{\sigma_1} \geq w_{\sigma_2} \geq \cdots \geq w_{\sigma_n}$, and define $S_i = \{\sigma_1, \ldots, \sigma_i\}$. Then

$$f(w) = \sum_{i=1}^{n} w_{\sigma_i} \big[F(S_i) - F(S_{i-1}) \big]$$

The computational cost in this procedure is dominated by the sorting, so it maintains an $O(N \log N)$ complexity. Now, recalling that equivalence $f(x) = \max_{y \in \mathcal{B}_F} \langle y, x \rangle$, we note that this same procedure yields the maximizing y, setting $y_{\sigma_i} := F(S_i) - F(S_{i-1})$. It is trivial to verify that indeed y is contained in $\mathcal{B}_F$.

Projections on the transportation polytope

If we use (negative) entropy as the mirror map in $\Pi(\mathbf{a}, \mathbf{b})$, i.e., $\Phi_{\Pi(\mathbf{a},\mathbf{b})}(\Gamma) := H(\Gamma) = \sum_{i,j} \Gamma_{ij} \ln(\Gamma_{ij})$, the projection of a point w onto $\Pi(\mathbf{a}, \mathbf{b})$ is given by the KL-divergence:

$$\hat{\Gamma} = \operatorname*{argmin}_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \mathrm{KL}(\Gamma \parallel w). \tag{3.20}$$

In Section 2.4 we discussed how projection onto $\Pi(\alpha, \beta)$ with the KL divergence corresponds to solving an entropy-regularized optimal transport problem. Therefore, this can be computed by Sinkhorn-Knopp algorithm. Thus, ε-accurate solution of equation (3.20) can be computed in $O(N \log N \varepsilon^{-3})$ time [3], but often much faster empirically [45].

Projections on the base polytope

If we use $\Phi_{\mathcal{B}_F}(\kappa) = \frac{1}{2}\|\kappa\|^2$, the resulting Euclidean projection[3] on the base polytope,

$$\hat{\kappa} = \operatorname*{argmin}_{\kappa \in \mathcal{B}_F} \|\kappa - w\|_2^2 = \operatorname*{argmin}_{\kappa' \in \mathcal{B}_{F-w}} \|\kappa'\|_2^2 + w, \tag{3.21}$$

is equivalent to minimizing the "shifted" submodular function $F(S) - \sum_{i \in S} w_i$ and can be computed, for instance, via the Fujishige-Wolfe minimum norm point (MNP) algorithm [179, 65], via parametric submodular minimization and with recent cutting-plane algorithms [115]. These generic methods are nevertheless computationally very expensive, except for small problems.

However, most of the functions of interest, such as the group functions defined in Section 3.3.1, have additional structure: they are of the form $F(S) = \sum_{i=1}^{k} F_i(S)$ (also called *decomposable*), each F_i with small support or "simple" structure. Here, "simple" means that the minimum norm point problem can be solved fast. For the functions defined in (3.5), and more generally, for certain hierarchical functions [87, 89], coverage functions [168] and graph cuts on lines (equivalent to Total Variation), this can be solved in $O(m \log m)$ time, where m is the support size of the respective F_i. We provide an $O(m \log m)$ algorithm for our cluster functions in the next section.

If the supports of the F_i's are disjoint, then the base polytope is a product of polytopes $\mathcal{B}_{F_i}$, and the projection can be computed for each $\mathcal{B}_{F_i}$ separately in parallel. If the supports overlap, then we can still exploit decomposition structure via randomized coordinate descent [56], operator splitting methods [92, 137] or others [168] for fast optimization.

3.5.5 Fast projections into submodular base polytopes

The problem of computing the point of minimal norm on the base polytope of a submodular function is intimately related to that of minimizing the function itself.

[3]Perhaps surprisingly, the projection onto the base polytope resulting from choosing $\Phi_{\mathcal{B}_F}(\kappa) := H(\kappa)$ instead is also solved by (3.21) [52], and hence we may implement mirror descent with either projection.

The solutions to these two problems are related through the parametric minimization problem

$$S^*_\lambda = \operatorname{argmin} F(S) - \lambda|S|.$$

Let $\mathbf{y}^*$ be the min-norm point in $\mathbf{B}_F$. We can recover the solution to the original submodular function minimization (SFM) problem, $S^* := S^*_{\lambda=0}$ from y^* as $S^* = \{i \mid y^*_i \leq 0\}$. Conversely, we can recover $\mathbf{y}^*$ from the solutions of the parametric problem as

$$\mathbf{y}^*_j = \max\{\lambda \mid j \in S^*_\lambda\}$$

Thus, given a method for minimizing the function $F^\lambda := F(S) - \lambda|S|$, one can obtain the min-norm-point by repeated calls to this oracle and a divide-and-conquer strategy as the one Jegelka et al. [92] use, which runs in $O(n \log n)$ time.

Now, in our case, we are dealing with cluster functions of the form $F_i(S) = g(\sum_{i\in S} w_i)$, and in addition, we are interested in computing projections, rather than the min-norm-point, i.e., we are interested in $\tilde{\kappa} = \operatorname{argmin}_{\kappa\in\mathcal{B}_F} \|\kappa - m\|_2^2$ for some $m \in \mathbb{R}^{n\times m}$. Equivalently, we want to minimize $F_w(S) := F(S) - M(S)$, where M is the modular function implied by the vector m. Thus, the parametric submodular function minimization (SFM) problem we are dealing with is

$$\begin{aligned}
F^\lambda_w &= g(\sum_{i\in S} w_i) + \sum_{i\in S} m_i - \lambda|S| \\
&= g(\sum_{i\in S} w_i) + \sum_{i\in S} (m_i - \lambda) \\
&= \min_{\alpha\in I} c_\alpha + (\alpha \sum_{i\in S} w_i) + \sum_{i\in S} (m_i - \lambda) \\
&= \min_{u\in[0,\sum_{i\in V} w_i]} g(u) + \nabla g(u)\big(\sum_{i\in S} w_i - u\big) + \sum_{i\in S} (m_i - \lambda)
\end{aligned}$$

where we used the fact that any concave function can be written as the pointwise supremum of (potentially infinite) linear functions, parametrized by α, and an interval I where the valid gradients lie. Since the minimization is jointly over S and α, we can

Algorithm 1: Fast SFM for Concave-of-Sum

Input: Submodular set function F.
Output: Optimal set $S^* = \operatorname{argmin} F(S)$

for $i = 1, \ldots, n$ **do**
  $r_i \leftarrow -(m_i + \lambda)/w_i$
$\hat{V} \leftarrow Sort(V)$ // By increasing value of r_i
for $i = 1, \ldots, n$ **do**
  $S_k \leftarrow \{1, \ldots, V(k)\}$
$S^* = \operatorname{argmin}_{S_i} F(S_i)$
return S^*

rewrite the problem as

$$\min_{\alpha} \min_{S} c_\alpha + \alpha \sum_{i \in S} w_i + \sum_{i \in S} (m_i - \lambda) \tag{3.22}$$

As the slope $\alpha = \nabla g(u)$ shrinks, the constant $c_\alpha = g(u) - u\nabla g(u)$ grows. We make the following observations:

1. Equation (3.22) suggests the following strategy: (1) for each α, find the minimizing set S^α. (2) Evaluate the function above for each S^α, and pick the one minimizing $F(S)$.

2. For a fixed α, the optimal S^α is easy to find:

$$S^\alpha - \{i \mid \alpha w_i + m_i + \lambda \leq 0\} = \{i \mid \alpha \leq -(m_i + \lambda)/w_i$$

3. Observation 2 shows that the optimal sets as α shrinks are nested: once an item enters the optimal set, it never leaves.

These observations suggest a simple sorting-based algorithm for finding the minimizer of $F(S)$, shown here as Algorithm 1. It runs in time $O(n \log n + nT)$, where T is the evaluation time of F and n is the size of the ground set of F. We emphasize that this algorithm is only valid for the concave-of-sum functions as defined in Section 3.3.1.

3.5.6 Putting it all together

Convex formulation

The final version of Mirror Descent that we use to solve the convex (left-hand side) formulation of Problem (3.7) is shown here as Algorithm 2. As discussed in Section 3.5.2, the choice of entropy mirror map $\Phi(\Gamma) = H(\Gamma)$ means that every iteration will require a KL-projection onto the base polytope and a subgradient computation, bringing the total cost per iteration to $O(N \log N + N(\log N)\varepsilon^{-3})$. For a non-smooth, not strongly convex function like the Lovász extension, MDA converges with rate $O(\frac{1}{\sqrt{t}})$.

Algorithm 2: Mirror Descent (MDA) for Structured Optimal Transport

Input: Initial coupling Γ_0.
Parameters: Initial step size η_0.
Output: Optimal transportation coupling Γ^*.

while $\Delta > tol$ **do**
  $g_t \leftarrow$ EDMONDS(f, Γ_t)
  $\tilde{\Gamma}_{t+1} \leftarrow$ SINKHORN$(\Gamma_t \circ \exp\{-\eta_t g_t\})$
  $\Gamma_{t+1} \leftarrow [\sum_{s=1}^{t+1} \eta_s]^{-1} \sum_{s=1}^{t+1} \eta_s \tilde{\Gamma}_s$
  $\Delta \leftarrow f(\Gamma_t) - f(\Gamma_{t+1})$
  $t \leftarrow t + 1$
return Γ_t

Saddle-point formulation

The final versions of the SP-MD and SP-MP methods used to solve the minimax formulation of Problem (3.7) are shown here as Algorithm 3 and Algorithm 4, respectively. Compared to MDA and SP-MD, the mirror-prox version enjoys a better convergence rate of $O(\frac{1}{t})$, at the cost of doubling the per-iteration cost, requiring two projections onto each of $\Pi(\mathbf{a}, \mathbf{b})$ and $\mathcal{B}_F$. Using the fast projection method for the cluster-based functions proposed here (Eq. 3.4), the total cost per iteration in either SP-MD and SP-MP is $O(N(\log N)\varepsilon^{-3} + K \log K)$, where K is the size of the largest cluster.

Algorithm 3: Saddle-Point Mirror Descent for Structured Optimal Transport

Input: Initial variables $z_0 = (\Gamma_0, \kappa_0)$.
Parameters: Initial step size η_0.
Output: Optimal transportation coupling Γ^* and cost matrix κ^*.

while $\Delta > tol$ **do**
  $\Gamma_{t+1} \leftarrow$ SINKHORN$(\Gamma_t \circ \exp\{-\eta_t \kappa_t\})$
  $\kappa_{t+1} \leftarrow$ BASEPOLYPROJECT$(\kappa_t + \eta_t \Gamma_t)$
  // Compute saddle point gap of current solution (Eq. (3.11))
  $z_{t+1} \leftarrow [\sum_{s=1}^{t+1} \eta_s]^{-1} \sum_{s=1}^{t+1} \eta_s (\Gamma_s, \kappa_s)$
  $\Delta \leftarrow$ SADDLEGAP(z_t)
  $t \leftarrow t + 1$
return Γ_t, κ_t

Algorithm 4: Saddle-Point Mirror-Prox for Structured Optimal Transport

Input: Initial variables $z_0 = (\Gamma_0, \kappa_0)$.
Parameters: Initial step size η_0.
Output: Optimal transportation coupling Γ^* and cost matrix κ^*

while $\Delta > tol$ **do**
  // Mirror step on true gradient
  $u_{t+1} \leftarrow$ SINKHORN$(\Gamma_t \circ \exp\{-\eta_t \kappa_t\})$
  $v_{t+1} \leftarrow$ BASEPOLYPROJECT$(\kappa_t + \eta_t \Gamma_t)$
  // Mirror step on proxy gradient
  $\Gamma_{t+1} \leftarrow$ SINKHORN$(\Gamma_t \circ \exp\{-\eta_t v_{t+1})$
  $\kappa_{t+1} \leftarrow$ BASEPOLYPROJECT$(kappa_t + \eta_t u_{t+1})$
  // Compute saddle point gap of current solution (Eq. (3.11))
  $z_{t+1} \leftarrow [\sum_{s=1}^{t+1} \eta_s]^{-1} \sum_{s=1}^{t+1} \eta_s (\Gamma_s, \kappa_s)$
  $\Delta \leftarrow$ SADDLEGAP(z_t)
  $t \leftarrow t + 1$
return Γ_t, κ_t

Initialization

A simple choice to initialize the transportation coupling is via $\Gamma_0 = \mathbf{a}\mathbf{b}^\top$. On the other hand, a random corner in the base polytope can be used to initialize κ_0. This can be computed, for example, by evaluating f for a random $w \in \mathbb{R}^{n \times m}$. However, we found that initializing κ as the projection of C onto $\mathcal{B}_F$ often results in faster convergence.

3.6 Experimental Results

Our implementation of the algorithms proposed in the previous section is done on Python. We rely on the Python Optimal Transport library [64] for entropic projections onto the transport polytope. Since we use decomposable submodular cost functions in all experiments, for the projections onto the base polytope required by SP-MP (Alg. 4) we rely on the fast projection method described in Algorithm 1, combined with RCDM [55] when the supports are not disjoint. All experiments were run on a 2.8GHz Intel Core i7 Processor.

3.6.1 Clustered point cloud matching

Synthetic Point Clouds. In our first set of experiments, we seek to understand the characteristics of the transport plans obtained with our structured optimal transport (SOT) framework. For this, we generate two point clouds in $\mathbb{R}^2$ from two distinct 3-Gaussian mixture distributions (20 points each, 60/20/20% class splits). We use the class labels to define a sum-of-clusters function as in (3.5), using square-root thresholding functions (3.6) for varying values of α. The optimal coupling matrices are shown in Figure 3-2. As expected, lower values of α enforce cluster structure more aggressively, while for larger α the cost effectively becomes modular, causing the solution to resemble those of the unstructured OT formulations.

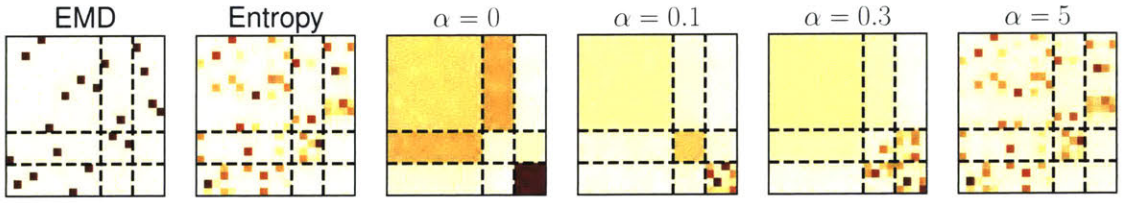


Figure 3-2: Optimal transport plans for clustered point matching obtained with two structure-agnostic formulations (EMD, entropy-regularized) and our submodular approach with varying concavity threshold parameter α (Eqn. (3.6)). Dashed lines show class partitions.

In terms of empirical runtimes (Fig. 3-3), SP-MP generally outperforms both SP-MD and MDA except in the very low sample size regime.

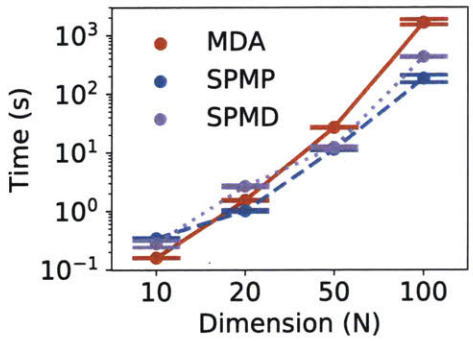


Figure 3-3: Runtimes for alternative optimization methods for the submodular optimal transport problem on the synthetic examples.

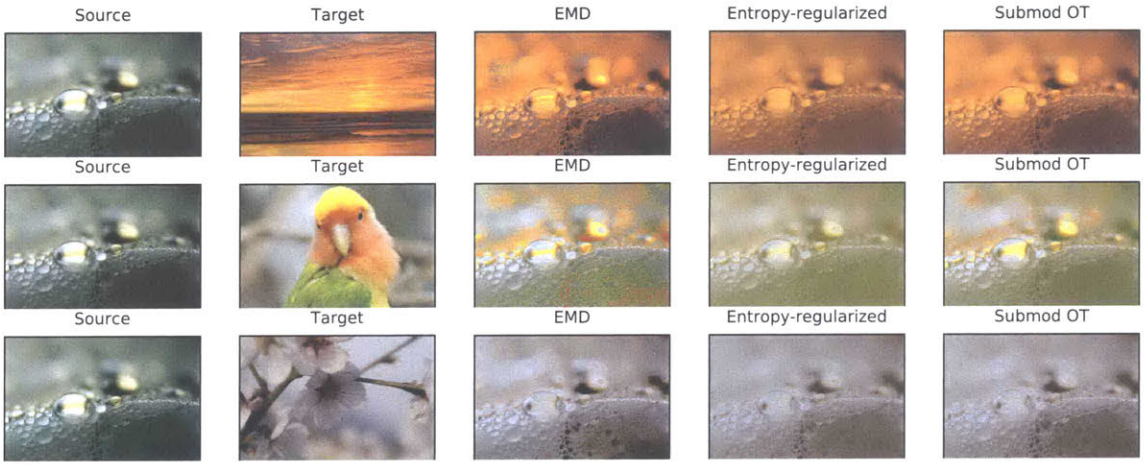


Figure 3-4: Color transfer with various optimal transport methods. The pixels in the source image get their color from the transported pixels in the target image.

Color transfer. An interesting application of this matching with group information is color transfer. Here, we seek to transfer the colors of one image (the *target* color scheme) into another one, the *source*. To do so, we view pixels as points in RGB space, transport them using optimal transport, and assign their color to the matched pixels. Here we define partitions through super-pixels obtained by segmentation [61]. The example in Figure 3-4 shows that including structure in the cost function results in a coloring scheme that is more uniform that the EMD variant and sharper than the entropy-regularized one.

3.6.2 Domain adaptation

Domain adaptation can be naturally cast as a transportation problem. When modeling the source and target distributions via discrete samples, DOT yields an optimal transport plan Γ^* between the two samples, according to which source points can be "transported" to the target domain through the *barycentric* mapping implicitly defined by Γ^* [174, Chapter 7].

In our motivating example of domain adaptation for classification, we wish to incorporate any available class labels on either domain into the cost function, so as to encourage points of the same class to be mapped to the same region of the target space. This is seamlessly attainable with our proposed framework and the cluster functions defined before (3.5). In the experiments below, we partition the source samples according to their class label, but we do not use the target labels (i.e., every target sample forms its own cluster), so as to simulate the harder—and more realistic—unsupervised domain adaptation setting.

We test this adaptation approach on the benchmark USPS and MNIST digit classification datasets. We preprocess the data by normalizing, and downscale MNIST to the 16×16 size of USPS. Here, we simulate an extreme adaptation setting where only 100 samples of each domain are provided, and no target labels are available. We train a 1-NN classifier on the transported samples and use it to predict labels on the test set (10K examples for MNIST, ~2K for USPS).

We compare our method (using (3.5) with (3.6), and a default $\alpha = 0.2$ threshold) against the two class-regularized OT formulations of Courty et al. [44]: one using an ℓ_p–ℓ_1 group-sparsity norm, and the other a Laplacian regularization term. We also compare against the original and entropy-regularized formulations, neither of which uses class labels.

Figure 3-5 shows the optimal couplings obtained with the various formulation of the optimal transport problem. The rows and columns in the couplings are sorted by the corresponding class label of the samples, so an ideal solution would be as close to a block-diagonal matrix as possible, which would entail matching digits coherently

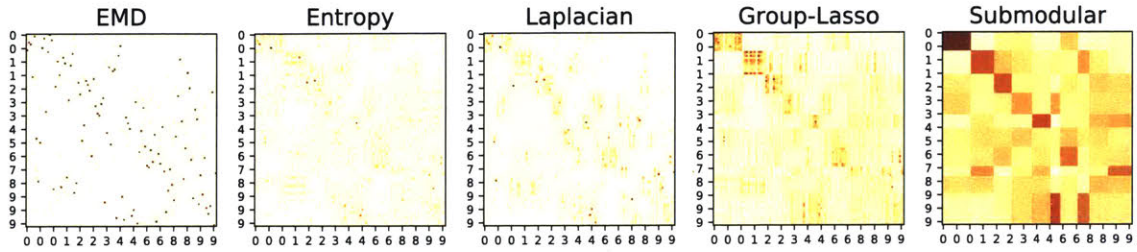


Figure 3-5: Optimal transport plans for the MNIST→USPS adaptation task. Rows and columns are sorted by class.

Method	MNIST→USPS	USPS→MNIST
No adaption	41.20	33.10
EMD	37.72	33.68
Entropy	55.70	43.64
Laplace	54.37	37.73
Group-Lasso	57.12	49.49
Struct-OT	**62.97**	**58.34**

Table 3.1: Results on digit recognition adaptation. The number shown correspond to prediction accuracy (%) on the test set. The "No Adaptation" baseline corresponds to directly applying the source model to the target domain (without transfer).

to the corresponding class on the other domain. Clearly, the two classic solutions of the problem, being oblivious of class labels, hardly reflect this structure. On the other hand, the two regularization schemes of Courty et al. [44] show some subtle block-diagonal form. However, our submodular formulation yields the clearest block diagonal structure of all of these.

Moving beyond a merely qualitative analysis of the coupling matrices, the results in Table 3.1 show that the submodular formulation achieves better accuracy in both directions of adaptation. We emphasize that the target labels are not used when defining the groupings of the submodular function, so this block structure is obtained solely by encouraging source points with the same label to be mapped together. Example source and transported digits are shown in Figure 3-6.

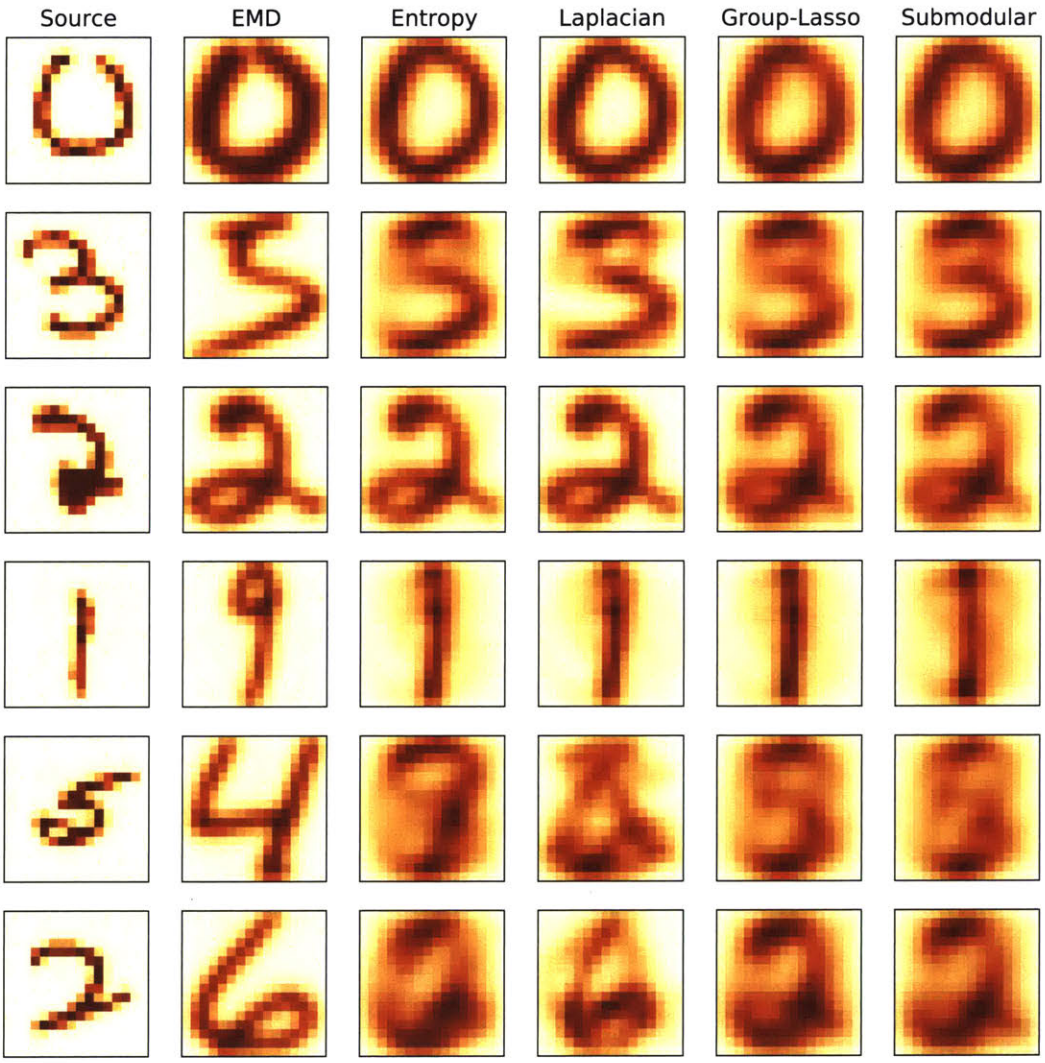


Figure 3-6: Examples from the MNIST→USPS domain adaptation task. The first column is the source image from MNIST, and the remaining columns are the result of transporting the source image into the target domain with the barycentric mapping defined by the various optimal transport plans.

3.6.3 Syntax-aware word mover's distance

The *Word Mover's Distance* (WMD) is an application of optimal transport to natural language processing [106]. It measures dissimilarity between strings (sentences or documents) by computing the cost of "moving" the words from one to the other, using a ground metric of distances between vector-space embeddings of words. The WMD, however, is syntax-agnostic, i.e., it does not take into account word order. That is, the cost of "moving" a word u_i in sentence U to v_j in sentence V depends only on their distance in the embedded space, and not on their relative positions in the two sentences. When using WMD to predict sentence similarity of long sentences with subclauses, this approach can have obvious drawbacks, like transporting words across noun-phrase boundaries.

There are obvious limitations to the WMD's purely semantic bag-of-words approach to sentence similarity, arising from ignoring the relations among words in a sentence. For example, consider the following sentences:

a) *The hotel does not appear in this book*

b) *I will book this hotel*

c) *I will reserve this hotel*

The WMD between (a) and (b) will likely be less that than between (b) and (c), even though the latter two are paraphrases of each other. Although (a) and (b) have strong single-word semantic overlap, the order in which the words occur in these two sentences entails different meanings. As contrived as this example might be, it is a good reminder that syntax and word-meaning go hand-in-hand for assessing semantic similarity at the sentence level.

We can obtain a syntax-aware alternative to WMD with a simple clustered cost function as before, where now each n-gram in a sentence defines a group (i.e., we allow overlaps between the groups). With this, we are encouraging neighboring words in a sentence to be matched to neighboring words in the other. Word-to-word costs are defined as before. We compare this distance against the original WMD in a simple sentence similarity task: the SICK dataset, consisting of pairs of English sentences

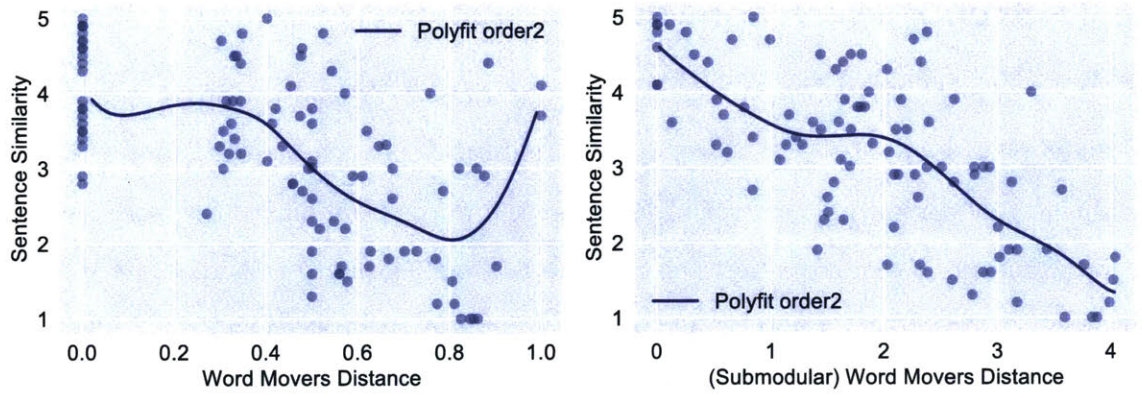


Figure 3-7: Sentence similarity prediction with two classes of optimal transport distances over sentences.

labeled with human-generated similarity scores. We randomly select 100 sentences with at most 10 words from the train and test folds, we compute optimal transport distances between all training pairs, and then fit a non-parametric regression model to predict similarity scores from these distances. At test time, given a pair of sentences, we compute the distance between them and use the regression model to predict their similarity. The distances, gold similarity scores and fitted models are shown in Figure 3-7. The WMD model obtains a mean squared error of 0.67 (Spearman's ρ of .71), while our proposed syntax-aware version has a much better correlation with gold similarity scores (MSE=0.59, $\rho = .75$).

3.6.4 Further illustrative examples

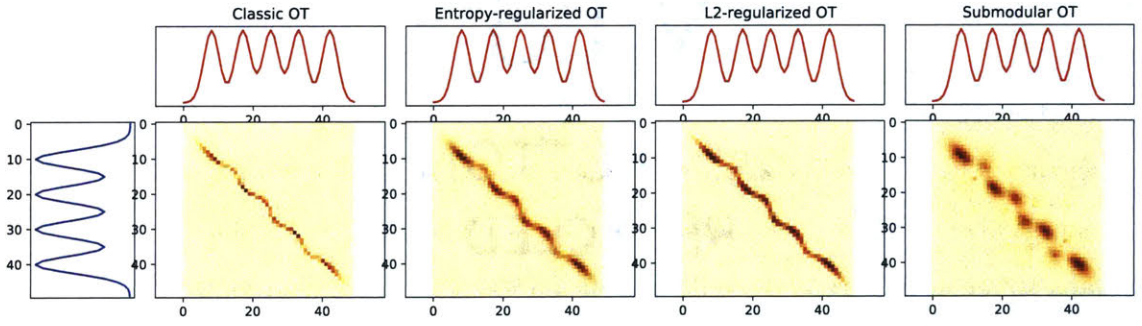


Figure 3-8: Comparison of optimal transportation couplings between two 1-dimensional multimodal densities obtained by various flavors of the Optimal Transport objective. The submodular objective leads to mode preservation.

Besides the settings presented in this work where structure arises from group labels, the framework proposed here allows us to explicitly encourage certain topological aspects of the distributions to be preserved. One such possible constraint for discrete distributions that lie on a low-dimensional manifold is to encourage neighboring points to be matched together. Such type of constraints can substantially alter the resulting transport plans, as shown in Figure 3-9 for a simple two-moons dataset. Here, the SOT solution favors neighborhood preservation over element-wise cost, resulting in a block-structured optimal coupling. On the other hand, the constraints can arise from the topology of the distribution itself, e.g., by encouraging modes of a distribution to be mapped together and thus inheriting this multi-modality in the coupling density (Figure 3-8).

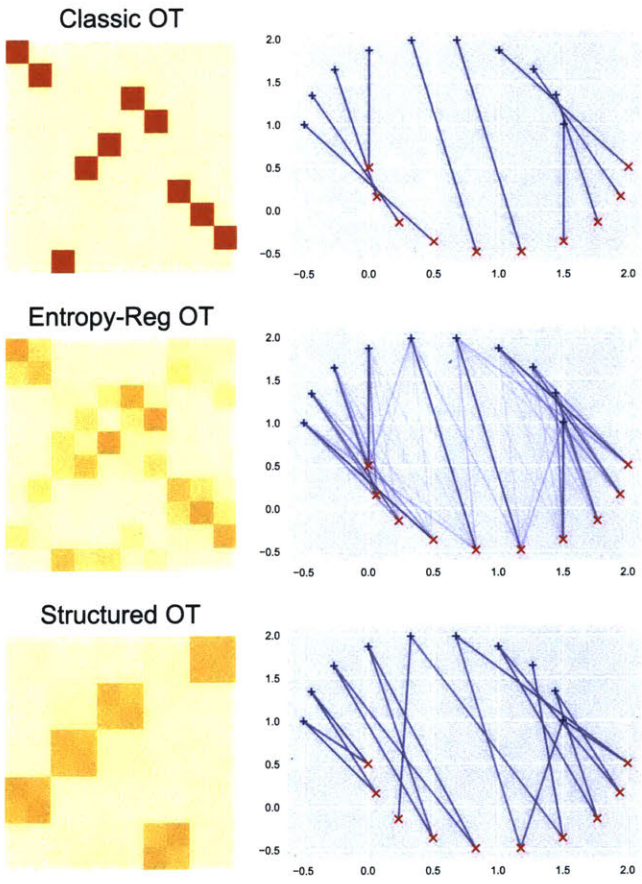


Figure 3-9: Optimal transport plans and matchings for classic and submodular versions of OT on a toy two-moons dataset.

3.7 Discussion and Extensions

In this chapter, we proposed a generic framework for including structural information into optimal transport problems, which are finding a growing range of applications in machine learning. While we demonstrated the utility of the framework via examples in domain adaptation, color transfer, and sentence similarity, our framework can encode a variety of structures beyond these settings, since it allows arbitrary submodular functions. This choice will depend on the specifics of the problem and the efficiency with which the projections can be solved. The overall resulting convex optimization problem is efficiently solvable via mirror descent methods. For very large problems or general submodular functions, approximate or stochastic submodular optimization subroutines (if applicable) may be suitable.

In fact, the flexibility of our framework goes beyond submodularity; any convex function with bounded closed gradient maps would work as f. Here, we explicitly chose submodular functions due to their favorable geometry and resulting tractability, and their ability to encode a wide range of combinatorial structures.

Chapter 4

Optimal Transport with Global Invariances

This chapter is partially based on Alvarez-Melis and Jaakkola [4] and Alvarez-Melis, Jegelka, and Jaakkola [10], with substantial extensions and various new methods proposed.

A key limitation of classic optimal transport is that it implicitly assumes that the two sets of objects in question are represented in the same space, or at least that meaningful pairwise distances between them can be computed. This is not always the case, especially when the objects are represented by learned feature vectors. For example, word embedding algorithms operate at the level of inner products or distances between word vectors, so the representations they produce can be arbitrarily rotated, sometimes even for different runs *of the same algorithm on the same data.* Such global degrees of freedom in the vector representations render direct pairwise distances between objects across the sets meaningless. Indeed, OT focuses on minimizing individual movement of mass, oblivious to global transformations. As a concrete example, consider two identical sets of points where one set is subjected to a global rotation. The optimal transport coupling evaluated between the resulting sets may no longer recover the correct correspondences.

When the global transformation is known or can be easily estimated, it can be incorporated in the computation of pairwise distances, thereby enabling the use

of traditional OT. Unfortunately, only the *type* of underlying transformation (e.g., rotation) is typically known, not the actual realization. In such cases, we would like the optimal transport problem to also find the best latent transformation along with the optimal coupling. In other words, we seek a formulation of OT that remains invariant under global transformations. In this chapter, we introduce such a formulation.

Simultaneous learning of transformations and matchings is a problem the arises in many settings, and for which many tailored solutions have been proposed. We review some of these related approaches in Section 4.2. In contrast to many of these, here we seek a general formulation that combines the power and theoretical footing of optimal transport with a flexible framework to learn cross-space transformations. The resulting problem, which can be understood as enforcing invariance to certain transformations in the OT cost objective, is general enough to subsume various related application-tailored approaches, but sufficiently confined so as to allow for tractable optimization and practical implementation. In particular, we show that modeling invariances through linear operators of bounded Schatten norm leads to a problem that can be solved very efficiently, and which simplifies even further under mild conditions often satisfied in the applications of interest.

The rest of this chapter is structured as follows. As a preamble, in Section 4.1 we discuss applications which motivate the framework of optimal transport with invariances, and in Section 4.3 we introduce two concepts which will play a prominent role throughout this chapter: the orthogonal Procrustes problem, and classic approach for finding correspondences across pair data from different domains; and the Gromov-Wasserstein distance, a recent generalization of the optimal transport problem for the case of incomparable domains.

The main content of this chapter starts in Section 4.4, where we discuss why classic OT is not applicable for unsupervised matching in many settings—including the motivating applications described in Section 4.1—and subsequently introduce the generalization of the problem which accounts for this. Upon defining the problem, we observe that it can be expressed in three equivalent forms, each one leading to a different family of optimization approaches. We then introduce the use of Schatten-norms to

define invariance classes, the last step for instantiating the proposed framework. After this, we turn to optimization. A technical section (which can be safely skipped unless the reader has interest in the details related to the implementation of the optimization methods) describes the building blocks—gradient and projection computation—of these general approaches (§4.6.1), after which we describe three main families of optimization approaches: alternating-minimization (§4.6.2), joint gradient descent (§4.6.3) and single-block gradient descent (§4.6.4). In Section 4.7 we take a step back and propose an alternative approach that can be applied in similar settings (for matching across unregistered spaces) based on the Gromov-Wasserstein distance.

The final part of this chapter presents an empirical evaluation of the proposed framework. We first compare the various approaches to solving the problem proposed here in a controlled setting (§4.8.2). For the rest of the experimental section, we focus on the two methods that yielded better performance in the initial simple tasks: invariant OT via alternating minimization and the Gromov-Wasserstein approach. We then test these methods in the problem of unsupervised word translation, showing that they perform on par with state-of-the-art, at a fraction of the computational cost. We end this chapter with a high-level discussion of practical considerations to choose between the Gromov-Wasserstein and invariant optimal transport approaches to unsupervised embedding alignment (§4.9).

4.1 Motivation and Applications

Finding correspondences across collections of objects represented in a fully unsupervised manner is a challenging problem that arises in many applications within machine learning. These are often used as a preliminary step in a multi-step pipeline, such as domain adaptation or transfer learning.

Cross-domain alignment is of particular importance in natural language processing. Indeed, many key linguistic tasks, within and across languages or domains, including machine translation, rely on learning cross-lingual correspondences between words or other semantic units. While the associated alignment problem could be solved with

access to large amounts of parallel data, broader applicability relies on the ability to do so with largely monolingual data, from Part-of-Speech (POS) tagging [185], dependency parsing [79], to machine translation [109]. The key subtask of bilingual lexical induction, for example, while long-standing as a problem [66, 148, 147], has been actively pursued recently [17, 183, 42].

Current methods for learning cross-domain correspondences at the word level rely on distributed representations of words, building on the observation that monolingual word embeddings exhibit similar geometric properties across languages Mikolov et al. [128]. While most early works assumed some, albeit minimal, amount of parallel data [128, 51, 185], recently fully-unsupervised methods have been shown to perform on par with their supervised counterparts [42, 15]. While successful, the mappings arise from multiple steps of processing, requiring either careful initial guesses or post-mapping refinements, including mitigating the effect of frequent words on neighborhoods. The associated adversarial training schemes can also be challenging to tune properly [15].

The main challenge for finding correspondences across word embedding spaces is the fact that these are not globally *registered*, i.e., there is no guarantee that their overall orientation is the same. As a concrete example, the vectors for *father* and *padre* might play a similar role with respect to other vectors in their corresponding collections, but could be pointing in different directions if one of the spaces is rotated with respect to the other. Indeed, word embeddings are estimated primarily in a relational manner (i.e., with distance-based optimization objectives) to the extent that the algorithms are naturally interpreted as metric recovery methods [84]. As a consequence, previous work has sought to correct this lack of *registration* by finding an orthogonal mapping that best aligns the spaces, but has traditionally assumed access to prior correspondences—seed translations—to do so Mikolov et al. [128] and Zhang et al. [185]. The success of these approaches provides strong evidence that orthogonality is the right notion of invariance across word embedding spaces, and has prompted many recent attempts to learn such a transformation even without access to seed translations [184, 42, 15]. We discuss many such methods in the next section.

4.2 Related Work

The general problem of unsupervised estimation of correspondence between sets of features is well-studied and arises in various fields under different names, such as manifold alignment [176], feature set matching [75] and feature correspondence finding [172]. While a –very long– thesis could be written just on reviewing all such methods, here we focus the discussion on related methods that combine soft correspondences (such as those produced by OT) with explicit space alignment. The only exception to this scope, which we mention due to its historical importance and connection to some of the methods proposed here, is the iterative closest point (ICP) algorithm [40, 26] (and its generalizations, e.g. [152]). ICP is a classic method to align point clouds in low-dimensions (usually 2D or 3D), which alternates between finding (hard) correspondences through nearest-neighbor pairing and finding the best rigid transformation based on those correspondences (i.e., solving an Orthogonal Procrustes problem). While practical, this method is limited by the fact that it computes strict assignments between points, which often leads to poor local solutions.

Perhaps the earliest approach which combines soft-matches with explicit space alignment is by Rangarajan et al. [146], who derive a framework to establish correspondences between shapes that rejects non-homologies (e.g., rotations) based on an entropy-regularized version of the OT problem. The resulting *Softassign Procrustes Algorithm* proceeds iteratively by alternating between estimating optimal rotations and performing Sinkhorn iterations. This approach, however, only considers rotations, and is tailored to 2D data, where rotations can be easily parametrized. Compared to methods that directly solve a Procrustes problem from a few known correspondences [185, 16] or by generating pseudo-matches through an initial unsupervised step [42, 15], optimal transport allows for more flexible correspondence estimation.

More recently, Zhang et al. [184] propose combining OT with Procrustes alignment to find correspondences between word embedding spaces. They initialize their orthogonal mapping using an adversarial training phase, much like Conneau et al. [42], and solve the optimization problem with alternating minimization. Our approach, on

the other hand, does not rely on neural network initialization, instead leveraging a convexity annealing scheme that leads to smooth convergence, with little sensitivity to initialization. A different line of work has investigated from a theoretical perspective the intersection of Procrustes alignment and Wasserstein means of distributions [181].

Concurrently with the work of the author of this thesis on this topic [10], Grave et al. [76] tackle the word embedding alignment task with an approach similar to that of Zhang et al. [184], combining Wasserstein distances (an instance of OT) and Procrustes alignment. Their approach differs from Zhang et al. [184] in how they scale up optimization, by relying on a stochastic Sinkhorn solver [71], and in how they initialize it, by solving a convex relaxation of the original problem.[1]

Although driven by a similar motivation (word embedding alignment) and relying on similar principles (joint optimization of transport coupling and feature mapping) as the work of Zhang et al. [184] and Grave et al. [76], our approach differs from them in several aspects. First, it allows for more general types of invariance classes (characterized as Schatten ℓ_p-norm balls), subsuming orthogonal invariance considered in prior work as a special case. Second, we dispense with the need for any ad-hoc initialization by introducing instead a convexity-annealing approach to optimization. Third, our approach remains robust to the choice of entropy regularization parameter.

A different generalization of the OT problem aimed at overcoming lack of intrinsic correspondence between spaces is the Gromov-Wasserstein (GW) distance [126]. It has been recently applied to various correspondence problems, including shape interpolation [164, 141] and unsupervised word translation [4]. While our framework recovers this distance in certain scenarios (see §4.5.3), it is best understood as a compromise between the classic formulation of OT that requires the spaces to be fully registered, and the GW distance, which completely forgoes explicit computation of distances across spaces, relying instead on comparison of intra-space similarities. Thus, our approach is best suited to tasks where distances across spaces can be computed, but are meaningful only if made invariant to some latent transformation. A further difference with the

[1] Interestingly, this relaxation corresponds to a hybrid of two instances of our framework: optimizing a Frobenius-norm objective (§4.5.3) over orthogonal matrices (§4.5.1).

usual OT and GW distances is that our approach produces, as an intrinsic part of optimization, a global mapping that can be used to map *out-of-sample points* across spaces.

It is worth noting that the problem of unsupervised word translation has a long history —under the name *bilingual lexical induction*— in the computational linguistics literature, going back to Rapp [148] and Fung [66]. The literature on this topic is extensive, and we refer the reader to one of the many existing surveys for a broader panorama [173, 150]. In our discussion above, we have already mentioned above various fully unsupervised methods for this task based on word embeddings [184, 76, 42]. In addition, there exists many minimally-supervised approaches that assume some coarse or limited parallel data. Most of these fall in one of two categories: methods that learn a mapping from one space to the other, e.g., as a least-squares objective (e.g., [128]) or via orthogonal transformations Zhang et al. [185], Smith et al. [162], and Artetxe et al. [17], and methods that find a common space on which to project both sets of embeddings [59, 122].

4.3 Preliminaries

4.3.1 Supervised alignment and the Procrustes problem

Space alignment from paired samples is a classical problem in statistics and linear algebra. In this problem, we assume we are given two sets of *paired* examples, $X = \{\mathbf{x}^{(i)}\}_{i=1}^n$ and $Y = \{\mathbf{y}^{(j)}\}_{j=1}^n$, drawn from potentially distinct feature spaces $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$. Here *paired* means that the elements across the two samples are known to be in correspondence, i.e., $\mathbf{x}^{(i)}$ corresponds to $\mathbf{y}^{(i)}$ and so forth. The problem of finding the best mapping T that maps the target samples to the source ones can be cast as

$$\min_{T \in \mathcal{F}} \|\mathbf{X} - T(\mathbf{Y})\|^2$$

where $\mathcal{F}$ is some class of functions and $\|\cdot\|$ is typically taken to be the Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}$. Naturally, the choice of space $\mathcal{F}$ will determine the difficulty of

finding T as well as the quality of the alignment implied by it.

The classic *Orthogonal Procrustes problem* restricts $\mathcal{F}$ to be rigid (rotation and reflection) transformations–i.e., orthogonal matrices:

$$\min_{\mathbf{P} \in \mathrm{O}(n)} \|\mathbf{X} - \mathbf{P}\mathbf{Y}\|_F^2. \tag{4.1}$$

Despite its simplicity, Procrustes analysis is a powerful tool used in various applications, from statistical shape analysis [73] to market research and others [74]. Its main advantage is that it has a closed-form solution in terms of a singular value decomposition (SVD) [157]. Namely, given an SVD, say $\mathbf{U}\Sigma\mathbf{V}^\top$, of $\mathbf{X}\mathbf{Y}^\top$, the orthogonal matrix minimizing problem (4.1) is $\mathbf{P}^* = \mathbf{U}\mathbf{V}^\top$, which is a direct consequence of a well-known approximation property of the SVD:

Lemma 4.3.1. *If $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$ is an SVD of $\mathbf{A}$, then the solution of the Orthogonal Procrustes is given by:* $\operatorname{argmax}_{\mathbf{P} \in \mathrm{O}(n)} \langle \mathbf{P}, \mathbf{A} \rangle = \mathbf{U}\mathbf{V}^\top$.

Proof. This is a particular case of the more general Lemma 4.5.2, which is proven in Section 4.5. □

We emphasize that the Procrustes problem (4.1) crucially requires the columns of $\mathbf{X}$ and $\mathbf{Y}$ be paired, making it an intrinsically *supervised* approach. Thus, its application to the problem of feature alignment requires either an —ideally small— initial set of true paired examples or a method to generate them.

Besides obvious computational advantage, constraining the mapping between spaces to be orthonormal is justified in the context of word embedding alignment because orthogonal maps preserve angles (and thus distances), which is often the only information used by downstream tasks (e.g., for nearest neighbor search) that rely on word embeddings. [162] further show that orthogonality is required for self-consistency of linear transformations between vector spaces.

Clearly, the Procrustes approach only solves the supervised version of the problem as it requires a known correspondence between the columns of $\mathbf{X}$ and $\mathbf{Y}$. Steps beyond this constraint include using small amounts of parallel data [185] or an unsupervised technique as the initial step to generate pseudo-parallel data [42] before solving for $\mathbf{P}$.

4.3.2 The Gromov-Wasserstein distance

The classic optimal transport requires a distance between vectors *across* the two domains where the measures are defined. As mentioned in the introduction to this chapter, such a metric may not be available, for example, when the sample sets to be matched do not belong to the same metric space (e.g., different dimensions). The Gromov-Wasserstein distance [126] generalizes optimal transport by comparing the metric spaces directly instead of samples across the spaces. In other words, this framework operates on distances between pairs of points calculated within each domain and measures how these distances compare to those in the other domain. Thus, it requires a weaker —but easy to define— notion of *distance between distances*, and operates on pairs of points, turning the problem from a linear to a quadratic one.

Formally, in its discrete version, the problem considers two measure spaces expressed in terms of within-domain similarity matrices $(\mathbf{C}, \mathbf{a})$ and $(\mathbf{C}', \mathbf{b})$ and a loss function defined between *similarity pairs*: $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, where $L(\mathbf{C}_{ik}, \mathbf{C}'_{jl})$ measures the discrepancy between the distances $d(\mathbf{x}^{(i)}, \mathbf{x}^{(k)})$ and $d'(\mathbf{y}^{(j)}, \mathbf{y}^{(l)})$. Typical choices for L are $L(a, b) = \frac{1}{2}(a - b)^2$ or $L(a, b) = \mathrm{KL}(a \parallel b)$. In the view of optimal transport theory, $L(\mathbf{C}_{ik}, \mathbf{C}'_{jl})$ can also be understood as the cost of transporting one unit of mass i to j *and* a unit from k to l.

All the relevant values of $L(\cdot, \cdot)$ can be put in a 4-th order tensor $\mathbf{L} \in \mathbb{R}^{N_1 \times N_1 \times N_2 \times N_2}$, where $\mathbf{L}_{ijkl} = L(\mathbf{C}_{ik}, \mathbf{C}'_{jl})$. As before, we seek a coupling Γ specifying how much mass to transfer between each pair of points from the two spaces. The Gromov-Wasserstein problem is then defined as solving

$$\mathrm{GW}(\mathbf{C}, \mathbf{C}', \mathbf{a}, \mathbf{b}) = \min_{\Gamma \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl} \tag{4.2}$$

It can be analogously defined for continuous measures α, β on metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$: Its continuous version is analogous:

$$\mathrm{GW}(d_{\mathcal{X}}, d_{\mathcal{Y}}, \alpha, \beta) = \min_{\Gamma \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{X} \times \mathcal{Y}} L(d_{\mathcal{X}}(x, x'), d_{\mathcal{X}}(y, y')) d\gamma(x, y) d\gamma(x', y') \tag{4.3}$$

This notion of discrepancy between metric spaces possesses various desirable theoretical properties, including the fact that for a suitable choice of the loss function it is indeed a distance:

Theorem 4.3.2 (Mémoli [126]). *With the choice $L = L_2$, $GW_2^{\frac{1}{2}}$ is a distance on the space of metric measure spaces.*

Proof. We refer the reader to [126] for a complete proof. □

Compared to the classic optimal transport problem (2.5), this version is substantially harder since the objective is now not only non-linear, but non-convex too.[2] In addition, it requires operating on a fourth-order tensor, which would be prohibitive in most settings.

Surprisingly, this problem can be optimized efficiently with first-order methods, whereby each iteration involves solving a traditional optimal transport problem [164, 141]. Furthermore, if the loss function can be written as:

$$L(a, b) = f_1(a) + f_2(b) - h_1(a)h_2(b),$$

which is the case of most common choices of L including $L = L_2$, Solomon et al. [164] show that instead of the $O(N_1^2 N_2^2)$ complexity implied by naive fourth-order tensor product, this computation reduces to $O(N_1^2 N_2 + N1N_2^2)$ cost. Their approach consists of solving Problem (2.5) by projected gradient descent, which yields iterations that involve projecting onto $\Pi(\mathbf{a}, \mathbf{b})$ a pseudo-cost matrix of the form

$$\hat{\mathbf{C}}_\Gamma(\mathbf{C}, \mathbf{C}', \Gamma) = \mathbf{C}_{xy} - h_1(\mathbf{C})\Gamma h_2(\mathbf{C}')^\top \tag{4.4}$$

where

$$\mathbf{C}_{xy} = f_1(\mathbf{C})\mathbf{a}\mathbb{1}_m^\top + \mathbb{1}_n\mathbf{b}^\top f_2(\mathbf{C}')^\top$$

We provide an explicit algorithm for the case $L = L_2$ in Section 4.6.

[2] In fact, the discrete (Monge-type) formulation of the problem is essentially an instance of the well-known (and NP-hard) quadratic assignment problem (QAP).

4.4 Unsupervised Matching with Optimal Transport

Besides providing a principled geometric approach to compare distributions, optimal transport has the advantage of producing, as an intrinsic part of its computation, a realization of the optimal way to match distributions. Indeed, any feasible coupling $\gamma \in \Pi(\alpha, \beta)$ in Problem (2.7) (or in Problems (2.5) or (2.11) analogously) can be interpreted as a "soft" or "multivalued" matching between α and β. Therefore, the optimal γ^* corresponds to the minimum-cost way to match them. In the case where the distributions are discrete (e.g., point clouds) γ^* is a matrix of soft correspondences between them. This (producing a soft-matching as intrinsic part of its computation) is one of the most appealing practical characteristics of optimal transport, and the main reason it has found successful application to a myriad of problems that involve finding correspondences or matches, such as image matching [186, 177], shape interpolation [163], shape registration [108], domain adaptation [44] and music transcription [63].

It is tempting to directly apply OT to the problem of unsupervised embedding alignment. Indeed, the OT toolbox does exactly what we seek intuitively: finding cost-optimal correspondences between two collections (or distributions) based on their geometry (i.e., via the metric) of the underlying space. But note that throughout our discussion of optimal transport in Chapter 2, we made the implicit —yet crucial— assumption that either the two distributions are defined in the same space $\mathcal{X}$, or if they are not, that a cost function between their respective spaces $\mathcal{X}$ and $\mathcal{Y}$ be specified. When the embedding spaces are estimated in a data-driven way, as is usually the case in machine learning, even if these spaces are compatible (e.g., have the same dimensionality) there is no guarantee that the usual metric $d(x, y)$ is meaningful. This could be, for example, because the spaces are defined up to rotations and reflections, creating a class of invariants that the ground metric does not take into account. A natural approach to deal with this lack of registration between the two spaces is to simultaneously find a global transformation that corrects for this *and* an optimal coupling that minimizes the transportation cost between the distributions.

Formally, consider two collections $\{\mathbf{x}^{(i)}\}_{i=1}^n$ and $\{\mathbf{y}^{(j)}\}_{j=1}^m$ and their associated

discrete probability measures α, β. In addition, consider a function class $\mathcal{F} : \mathcal{Y} \to \mathcal{Y}$ that defines the type of invariances present in the data.[3] Naturally, the choice of class $\mathcal{F}$ should be informed by the application domain, and we will discuss many such choices throughout this chapter. With this setting, in addition to the optimal coupling between the two measures (i.e., *local* or point-wise correspondence) we also seek among all transformations in $\mathcal{F}$ that which best aligns the two spaces (i.e, *global* or space-wise correspondence). In the language of optimal transport, this problem can be succinctly written as:

$$\mathrm{OT}_{c|\mathcal{F}}(\mathbf{a}, \mathbf{b}) \triangleq \min_{\substack{\Gamma \in \Pi(\mathbf{a}, \mathbf{b}) \\ f \in \mathcal{F}}} \langle \Gamma, \mathbf{C}(\mathbf{X}, f(\mathbf{Y})) \rangle. \tag{4.5}$$

where $f(\mathbf{Y})$ is a matrix formed by the columns $f(\mathbf{y}^{(i)})$, and as before $\mathbf{C}$ is the matrix of pair-wise costs, i.e. $[\mathbf{C}(\mathbf{X}, f(\mathbf{Y}))]_{ij} = c(\mathbf{x}^{(i)}, f(\mathbf{y}^{(j)}))$, but now we make the dependence on $\mathbf{X}$ and $\mathbf{Y}$ explicit. Naturally, just as we saw in the case of the original formulation of OT in Section 2.3, this objective can formulated for distributions too

$$\mathrm{OT}_{c|\mathcal{F}}(\alpha, \beta) \triangleq \min_{\substack{\gamma \in \Pi(\alpha, \beta) \\ f \in \mathcal{F}}} \int_{\mathcal{X} \times \mathcal{Y}} c(x, f(y)) d\gamma(x, f(y)) \tag{4.6}$$

As before, we can additionally define entropy-regularized versions of these problems too, which we denote by $\mathrm{OT}^{\varepsilon}_{c|\mathcal{F}}(\mathbf{a}, \mathbf{b})$ and $\mathrm{OT}^{\varepsilon}_{c|\mathcal{F}}(\alpha, \beta)$. We refer to all of these collectively as the **invariant optimal transport problem**.

Variations of this problem for particular cases of $\mathcal{X}, d(\cdot, \cdot)$ and $\mathcal{F}$ have been proposed in various contexts, particularly for image registration (e.g., [146, 41]), and more recently, for word embedding alignment [184, 76]. Virtually all these approaches instantiate $\mathcal{F}$ as the class of orthogonal transformations $\mathrm{O}(d)$. In such cases, minimization with respect to f is easy to compute, as it corresponds to an Orthogonal Procrustes problem, which has a closed-form solution [74]. In such cases, solving Problem (4.5) by alternating minimization is a sensible choice.

[3] Throughout this chapter, we assume the transformation f is applied to the target space. Naturally, this could be formulated for $\mathcal{X}$ instead, or —with some additional generalization— on both spaces simultaneously.

To simplify the notation in our derivation of methods below, we introduce the short-hand notation $\mathcal{T}_c(f, \gamma; \alpha, \beta)$ to denote the total transportation cost between α and β (after mapping by f) incurred by coupling γ, i.e.,

$$\mathcal{T}_c(f, \gamma; \alpha, \beta) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} c(x, f(y)) d\gamma(x, f(y)) \tag{4.7}$$

so that, for example, $\mathrm{OT}_c(\alpha, \beta) = \min_{\gamma\in\Pi(\alpha,\beta)} \mathcal{T}(\mathrm{Id}, \gamma; \alpha, \beta)$, and more generally, $\mathrm{OT}_c(\alpha, f_\sharp\beta) = \min_{\gamma\in\Pi(\alpha,\beta)} \mathcal{T}(f, \gamma; \alpha, \beta)$. In addition, we define analogously as before an $\mathcal{F}$-invariant version of this cost:

$$\mathcal{T}_{c|\mathcal{F}}(\gamma; \alpha, \beta) \triangleq \min_{f\in\mathcal{F}} \int_{\mathcal{X}\times\mathcal{Y}} c(x, f(y)) d\gamma(x, f(y)). \tag{4.8}$$

With this notation in hand, we observe that Problem (4.6) can be expressed in three equivalent forms depending on the order of optimization:

$$\min_{\substack{\gamma\in\Pi(\alpha,\beta)\\ f\in\mathcal{F}}} \left[\int_{\mathcal{X}\times\mathcal{Y}} c(x, f(y)) d\gamma(x, f(y)) \right] = \min_{\substack{\gamma\in\Pi(\alpha,\beta)\\ f\in\mathcal{F}}} \mathcal{T}_c(f, \gamma; \alpha, \beta) \tag{4.9}$$

$$\min_{\gamma\in\Pi(\alpha,\beta)} \left[\min_{f\in\mathcal{F}} \int_{\mathcal{X}\times\mathcal{Y}} c(x, f(y)) d\gamma(x, f(y)) \right] = \min_{\gamma\in\Pi(\alpha,\beta)} \mathcal{T}_{c|\mathcal{F}}(\gamma; \alpha, \beta) \tag{4.10}$$

$$\min_{f\in\mathcal{F}} \left[\min_{\gamma\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, f(y)) d\gamma(x, f(y)) \right] = \min_{f\in\mathcal{F}} \mathrm{OT}_c(\alpha, f_\sharp\beta) \tag{4.11}$$

Naturally, we can define entropy-regularized versions of any of these objectives by adding the term $-\varepsilon\,\mathrm{H}(\Gamma)$ (as before, we use superscript ε to denote those). Over the next sections, we will discuss approaches to solve these three versions of the problem.

A note on regularity. Whenever OT is used with the goal of *matching* (i.e., not just *comparison*), guarantees on the solution of the problem are important. A full exposition on such guarantees falls beyond our scope here, so we refer the interested reader to a survey on this [13]. However, from our brief discussion in Section 2.6.1, we recall that for the quadratic cost, the optimal coupling γ^* is guaranteed to exist, be unique, and correspond to a deterministic map (i.e., a "hard" matching).[4]

[4]Note that $\gamma(\alpha, \beta)$ "includes" all maps $T : \mathcal{X} \to \mathcal{X}$, which can be expressed as $\gamma(\cdot, \cdot) = (Id \times T)^\sharp\alpha$.

4.5 Modeling Invariances with Schatten Norms

In this section, we propose a class of transformations that leads to a simple but effective formulation of the invariant OT problem. The class of invariances is defined by linear operators with bounded norm:

$$\mathcal{F}_p \triangleq \{\mathbf{P} \in \mathbb{R}^{d\times d} \mid \|\mathbf{P}\|_p \leq k_p\}, \tag{4.12}$$

where $\|\cdot\|_p$ is the Schatten ℓ_p-norm, that is, $\|\mathbf{P}\|_p = \|\sigma(\mathbf{P})\|_p$ where $\sigma(\mathbf{P})$ is a vector containing the singular values of $\mathbf{P}$. In addition, k_p is a norm- and problem-dependent constant.[5]

This choice of invariance sets follows both modeling and computational motivations. As for the former, Schatten norms allow for immediate interpretation of the elements of $\mathcal{F}_p$ in terms of their spectral properties (Fig. 4-1). For example, choosing $p = 1$ encourages solutions with sparse spectra (e.g., projections, useful when the support of one of the two distributions is known to be contained in a lower-dimensional subspace), while $p = \infty$ instead seeks solutions with uniform spectra (e.g., unitary matrices, thus enforcing invariance to rigid transformations). Intermediate values of p interpolate between these two extremes. Interestingly, the choice $p = 2$ recovers a recent popular generalization of the optimal transport problem motivated by a similar goal: the Gromov-Wasserstein distance [126], as we show in Section 4.5.3. Thus, the proposed Schatten invariance framework offers significant flexibility. In terms of computation, Schatten norms exhibit various desirable properties, such as unitary invariance, sub-multiplicativity, and easy characterization via duality, all of which play an important role in deriving efficient optimization algorithms below.

Here, we formulate the problem for the case where the ground metric c is the squared euclidean distance, i.e, $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$, which is arguably the most common choice in practice. With this choice of ground metric, Lemma 4.6.2 shows that for

[5] In the most common case, k_p would be chosen to ensure the identity mapping is contained in this set.

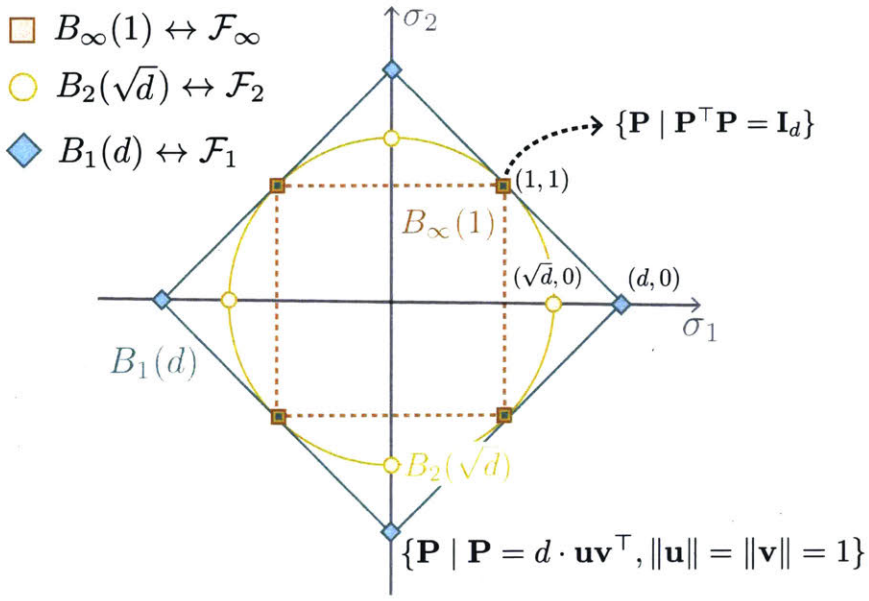


Figure 4-1: **Schatten-norm invariance classes**. The depicted ℓ_p-norm balls in singular value space correspond to matrix invariance classes $\mathcal{F}_p$. The radius is chosen so as to include the identity matrix ($\boldsymbol{\sigma} = [1, 1]$). For linear objectives, solutions when optimizing over $\mathcal{F}_\infty$ and $\mathcal{F}_1$ can be found on the extreme points of their respective constraint spaces: orthogonal matrices for the former and rank-one matrices for the latter.

linear transformations, we can express the transportation cost objective as

$$\mathcal{T}_c(f, \Gamma; \mathbf{a}, \mathbf{b}) = -2\langle \Gamma, \mathbf{X}^\top \mathbf{P} \mathbf{Y} \rangle + \langle \mathbf{x}_{|||}, \mathbf{a} \rangle + \langle \mathbf{P}\mathbf{y}_{|||}, \mathbf{b} \rangle,$$

where we remind the reader that $\mathbf{x}_{|||}$ is the vector with row-wise norms of $\mathbf{X}$, i.e., $\mathbf{x}_{|||} = (\|\mathbf{x}^{(1)}\|^2, \ldots, \|\mathbf{x}^{(n)}\|^2)$, and analogously for $\mathbf{P}\mathbf{y}_{|||}$. Thus, we can equivalently solve the maximization problem

$$\max_{\Gamma \in \Pi(\mathbf{a}, \mathbf{b})} \max_{\mathbf{P} \in \mathcal{F}} 2\langle \Gamma, \mathbf{X}^\top \mathbf{P} \mathbf{Y} \rangle - \langle \mathbf{x}_{|||}, \mathbf{a} \rangle - \langle \mathbf{P}\mathbf{y}_{|||}, \mathbf{b} \rangle. \tag{4.13}$$

This objective has a clear interpretation. The first term, which can be equivalently written as $\langle \mathbf{X}\Gamma, \mathbf{P}\mathbf{Y} \rangle$, measures agreement between $\mathbf{X}\Gamma$, the source points mapped according to the barycentric mapping implied by Γ, and $\mathbf{P}\mathbf{Y}$, the target points mapped according to $\mathbf{P}$. The other two terms, which can be interpreted as empirical expectations $\hat{\mathbb{E}}_{\mathbf{x} \sim \alpha} \|\mathbf{x}\|_2^2$ and $\hat{\mathbb{E}}_{\mathbf{y} \sim \beta} \|\mathbf{P}\mathbf{y}\|_2^2$, act as a counterbalance, normalizing the

objective and preventing artificial maximization of the similarity term by arbitrary scaling of the mapped vectors.

In general, Problem (4.13) is not jointly concave on $\mathbf{P}$ and Γ, but it is concave in either variable if the other one is fixed. This suggests an alternating maximization approach on $\mathbf{P}$ and Γ. Since only the first term depends on Γ, solving for this variable for a fixed $\mathbf{P}$ is a usual OT problem, for which we discuss optimization in Section 4.6.2. On the other hand, for a fixed Γ the problem is a concave maximization over a compact and convex set, which can be solved efficiently with Frank-Wolfe-type algorithms since projecting onto Schatten norm balls is tractable [90].

While the approach we have just described provides a tractable way to solve problem (4.13) in general, we show that under conditions that often hold practice, optimization is much simpler. This simplification relies on eliminating the dependence on $\mathbf{P}$ of the third term in problem (4.13):

Lemma 4.5.1. *Under either of the conditions*

1. $\forall \mathbf{P} \in \mathcal{F}$, $\mathbf{P}$ *is angle-preserving (i.e.,* $\forall \mathbf{x}, \mathbf{y}\ \langle \mathbf{Px}, \mathbf{Py} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle$*), or*

2. $\exists k \geq 0 : \|\mathbf{P}\|_F = k\ \forall \mathbf{P} \in \mathcal{F}$ *and* $\mathbf{Y}$ *is* β*-whitened (i.e.,* $\mathbf{Y}[[\mathbf{b}]]\mathbf{Y}^\top = \mathbf{I}_d$*)*

Problem (4.13) *is equivalent to*

$$\max_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \max_{\mathbf{P} \in \mathcal{F}} \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{P} \rangle. \tag{4.14}$$

Proof. Suppose (1) holds, i.e., $\langle \mathbf{Px}, \mathbf{Py} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Then, in particular $\|\mathbf{Py}\|_2 = \|\mathbf{y}\|_2$ for every $\mathbf{y}^{(j)}$, and therefore:

$$\langle \mathbf{P}\mathbf{y}_{|||}, \mathbf{b} \rangle = \sum_{j=1}^{m} \mathbf{b}_j \|\mathbf{P}\mathbf{y}^{(j)}\|^2 = \sum_{j=1}^{m} \mathbf{b}_j \|\mathbf{y}^{(j)}\|^2$$

and therefore only the first term in (4.13) depends on $\mathbf{P}$ or Γ, from which the conclusion follows. On the other hand, suppose (2) holds, and let $\tilde{\mathbf{Y}} = \mathbf{Y}[[\mathbf{b}^{1/2}]]$, so

that $\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\top = \mathbf{I}_d$. We have:

$$\langle \mathbf{P}\mathbf{y}_{|||}, \mathbf{b} \rangle = \sum_{i=1}^{m} \mathbf{b}_j \|\mathbf{P}\mathbf{y}^{(j)}\|^2 = \sum_{j=1}^{m} \|\mathbf{P}\mathbf{y}^{(j)}\sqrt{\mathbf{b}_j}\|^2$$
$$= \|\mathbf{P}\tilde{\mathbf{Y}}\|_F^2 = \langle \mathbf{P}\tilde{\mathbf{Y}}, \mathbf{P}\tilde{\mathbf{Y}} \rangle = \langle \mathbf{P}, \mathbf{P}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}' \rangle = \|\mathbf{P}\|_F^2 = k^2,$$

that is, $\langle \mathbf{P}\mathbf{y}_{|||}, \mathbf{b} \rangle$ again does not depend on $\mathbf{P}$. This concludes the proof. □

The first condition in Lemma 4.5.1 is reasonable as it guarantees $\mathbf{P}$ preserves geometric relations across spaces. On the other hand, whitening is a common pre-processing step in feature learning [88] and correspondence problems [18].

The inner problem in (4.14) is a generalized version of the orthogonal Procrustes Problem (§4.3). The following generalization of Lemma 4.3.1 shows that this problem too has a closed-form solution when optimizing over Schatten ℓ_p-norm balls.

Lemma 4.5.2. *Let* $\mathbf{M}$ *be a matrix with singular value decomposition given by* $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^\top$ *and let* $\Sigma = \operatorname{diag}(\boldsymbol{\sigma})$*, then*

$$\operatorname*{argmax}_{\mathbf{P}:\|\mathbf{P}\|_p \leq k} \langle \mathbf{P}, \mathbf{M} \rangle = \mathbf{U} \operatorname{diag}(\mathbf{s}) \mathbf{V}^\top \tag{4.15}$$

where $\mathbf{s}$ *is such that* $\|\mathbf{s}\|_p \leq k$ *and* $\mathbf{s}^\top \boldsymbol{\sigma} = k\|\boldsymbol{\sigma}\|_q$*, for* $\|\cdot\|_q$ *the dual norm of* $\|\cdot\|_p$*.*

Proof. Suppose $\mathbf{P}$ is such that $\|\mathbf{P}\|_p \leq k$, and let $\mathbf{U}_\mathbf{P} \operatorname{diag}(\mathbf{s}) \mathbf{V}'_\mathbf{P}$ be its singular value decomposition. This implies that $\|\mathbf{s}\|_p = \|\mathbf{P}\| \leq k$. In addition,

$$\langle \mathbf{P}, \mathbf{M} \rangle = \langle \mathbf{U}^\top \mathbf{P} \mathbf{V}, \Sigma \rangle = \sum_{i=1}^{d} [\mathbf{U}^\top \mathbf{P} \mathbf{V}]_{ii} \sigma_i(\mathbf{M}) = \sum_{i=1}^{d} \mathbf{u}_i^\top \mathbf{P} \mathbf{v}_i \sigma_i(\mathbf{M})$$
$$\leq \sum_{i=1}^{d} s_i \sigma_i(\mathbf{M}) = \langle \mathbf{s}, \boldsymbol{\sigma} \rangle$$

Here, the inequality holds because, by definition of the SVD decomposition, for every i it must hold that $\|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1$ and

$$\mathbf{u}_i^\top \mathbf{P} \mathbf{v}_i \leq \sup_{\substack{\mathbf{u} \perp \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}\} \\ \mathbf{v} \perp \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}\}}} \frac{\mathbf{u}^\top \mathbf{P} \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \leq \sigma_i(\mathbf{P}) = s_i \tag{4.16}$$

Therefore:

$$\sup_{\mathbf{P}:\|\mathbf{P}\|_p\leq k} \langle \mathbf{P}, \mathbf{M}\rangle \leq \sup_{\mathbf{s}:\|\mathbf{s}\|_p\leq k} \langle \mathbf{s}, \boldsymbol{\sigma}\rangle = k \sup_{\mathbf{s}:\|\mathbf{s}\|_p\leq 1} \langle \mathbf{s}, \boldsymbol{\sigma}\rangle = k\|\boldsymbol{\sigma}\|_q$$

where the last equality follows from the definition of dual norm for vectors.

Conversely, take any vector $\mathbf{s}$ with $\|\mathbf{s}\|_p = k$, and define $\tilde{\mathbf{P}}(\mathbf{s}) = \mathbf{U}\operatorname{diag}(\mathbf{s})\mathbf{V}^\top$. Clearly, $\|\tilde{\mathbf{P}}(\mathbf{s})\|_p = k$, so the supremum must satisfy:

$$\begin{aligned}\sup_{\mathbf{P}}\langle \mathbf{P}, \mathbf{M}\rangle &\geq \sup_{\mathbf{s}:\|\mathbf{s}\|_p\leq k} \langle \tilde{\mathbf{P}}(\mathbf{s}), \mathbf{M}\rangle \\ &= \sup_{\mathbf{s}:\|\mathbf{s}\|_p\leq k} \langle \mathbf{U}\operatorname{diag}(\mathbf{s})\mathbf{V}^\top, \mathbf{U}\Sigma\mathbf{V}^\top\rangle = \sup_{\mathbf{s}:\|\mathbf{s}\|_p\leq k} \langle \operatorname{diag}(\mathbf{s}), \Sigma\rangle = k\|\boldsymbol{\sigma}\|_q\end{aligned}$$

Therefore, we conclude that the optimal value of (4.15) is exactly $k\|\boldsymbol{\sigma}\|_q$.

Furthermore, (4.16) holds with equality if and only if $(\mathbf{u}_i, \mathbf{v}_i)$ coincide with the left and right singular vectors of $\mathbf{P}$. Thus, any $\mathbf{P}$ maximizing (4.15) must have the form $\mathbf{P} = \mathbf{U}\operatorname{diag}\mathbf{s}\mathbf{V}'$, with $\|\mathbf{s}\|_p \leq k$ and $\langle \mathbf{s}, \boldsymbol{\sigma}\rangle = k\|\boldsymbol{\sigma}\|_q$, as stated. □

Therefore, the inner problem in (4.14) boils down to maximization of support functions of vector-valued ℓ_p balls, which can be done in closed form for any $p \geq 1$ by choosing $s_i \propto \sigma_i^{q-1}$ [90]. This, in turn, greatly simplifies the alternating optimization approach. For a fixed Γ, we can use Lemma 4.5.2 to obtain a closed-from solution $\mathbf{P}^*$. On the other hand, for a fixed $\mathbf{P}$, optimizing for Γ is a classic discrete optimal transport problem with cost matrix $\tilde{\mathbf{C}} = -\mathbf{X}^\top\mathbf{P}\mathbf{Y}$,[6] which can be solved with off-the-shelf OT algorithms.

Next, we investigate what Lemma 4.5.2 implies for three salient cases, $p = \infty$ and $p = 2$ and $p = 1$.

4.5.1 The case $p = \infty$

The Schatten ℓ_∞-norm is the spectral norm $\|A\|_\infty = \sigma_{\max}(A)$. To guarantee that the identity is contained in $\mathcal{F}_\infty$, we take $k_\infty \triangleq 1$. Note that combining either condition

[6] This is of course equivalent to solving the original problem (4.5), whose cost matrix has a simpler interpretation.

in Lemma 4.5.1 with this implies that $\mathcal{F}_\infty = \mathrm{O}(n)$. Therefore, this choice of norm naturally encodes invariance to rotations and reflections. The dual characterization of Schatten norms implies that

$$\max_{\mathbf{P}\in\mathcal{F}_\infty} \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{P}\rangle = \|\mathbf{X}\Gamma\mathbf{Y}^\top\|_* \tag{4.17}$$

so that (4.14) becomes a single-block problem:

$$\max_{\Gamma\in\Pi(\mathbf{a},\mathbf{b})} \|\mathbf{X}\Gamma\mathbf{Y}^\top\|_* \tag{4.18}$$

Albeit succinct, this alternative representation of the problem is not easier to solve. Despite having eliminated $\mathbf{P}$, the objective is now non-convex with respect to Γ (maximization of a convex function). Nevertheless, this formulation offers an interesting geometric interpretation. When $\mathbf{a}, \mathbf{b}$ are uniform distributions, then $\hat{\mathbf{X}} \triangleq \mathbf{X}\Gamma$ is a matrix whose columns correspond to the those of $\mathbf{X}$ transported to $\mathcal{Y}$ according to the optimal barycentric mapping. Hence, $\hat{\mathbf{X}}\mathbf{Y}^\top$ is the (shifted) cross-covariance matrix of the features in $\mathcal{X}$ and $\mathcal{Y}$ space, i.e., $[\hat{\mathbf{X}}\mathbf{Y}^\top]_{ij} = \mathrm{cov}(\hat{x}_i, y_j)$, and its norm indicates the strength of correlation of these features. Therefore, problem (4.18) essentially seeks a transport coupling that maximizes the correlation of feature dimensions after transportation. We leave exploration of direct techniques to optimize (4.18) for future work. Here instead we rely on the generic alternating minimization scheme described in the previous section.

4.5.2 The case $p = 1$

Recall that the Schatten ℓ_1-norm is the nuclear norm $\|A\|_* = \sum_{i=1}^n \sigma_i(A)$. Therefore, the invariance set of interest is now

$$\mathcal{F}_1 = \left\{\mathbf{P} \;\middle|\; \|\mathbf{P}\|_* = d\right\}, \tag{4.19}$$

which, as before, contains the identity matrix.

Note that adding either condition in Lemma 4.5.1 yields, again, the set of or-

thonormal matrices.[7] Therefore, in the case one wants to rely on Lemma 4.5.2 to solve the problem efficiently, this choice of invariance ends up being equivalent to the $p = \infty$ case described in Section 4.5.1. However, we remark that this equivalence is a consequence of the simplifying assumptions, and that one could still solve this problem with the Frank-Wolfe approach described earlier, in which case the cases $p = \infty$ and $p = 1$ would indeed lead to different solutions.

4.5.3 The case $p = 2$

The Schatten ℓ_2-norm is the Frobenius norm. Since $\|\mathbf{I}_d\|_F = \sqrt{d}$, we take $\mathcal{F}_2 = \left\{\mathbf{P} \;\middle|\; \|\mathbf{P}\|_F = \sqrt{d}\right\}$. As before, we use Schatten norm duality to note that

$$\max_{\mathbf{P}\in\mathcal{F}_2} \langle \mathbf{X}\boldsymbol{\Gamma}\mathbf{Y}^\top, \mathbf{P}\rangle = \sqrt{d}\|\mathbf{X}\boldsymbol{\Gamma}\mathbf{Y}^\top\|_F,$$

whereupon problem (4.14) now becomes

$$\max_{\boldsymbol{\Gamma}\in\Pi(\mathbf{a},\mathbf{b})} \|\mathbf{X}\boldsymbol{\Gamma}\mathbf{Y}^\top\|_F, \tag{4.20}$$

which admits a similar interpretation to Problem (4.18), albeit for a different metric. However, this subtle difference has important consequences, such as the following connection.

Lemma 4.5.3. *Consider the Gromov-Wasserstein problem for discrete measures α and β with probability vectors* $\mathbf{a}$ *and* $\mathbf{b}$*:*

$$\min_{\boldsymbol{\Gamma}\in\Pi(\mathbf{a},\mathbf{b})} \sum_{i,j,k,l} L(\mathbf{C}^x_{ik}, \mathbf{C}^y_{jl})\Gamma_{ij}\Gamma_{kl}, \tag{4.21}$$

where $(\mathbf{C}^x, \mathbf{a})$ *and* $(\mathbf{C}^y, \mathbf{b})$ *are (intra-space) measured similarity matrices and L is a loss function. For the choice of cosine similarity and squared loss $L(a, b) = \frac{1}{2}|a - b|^2$, Problems* (4.21) *and* (4.20) *are equivalent.*

[7] The intersection of the Schatten ℓ_2 and ℓ_∞ norm balls, defined in terms of that of the ℓ_2 and ℓ_∞ vector norm balls, occurs in the extreme points of the latter (see Fig. 4-1).

Proof. For the choice of cosine metric, and assuming without loss of generality that the columns of $\mathbf{X}$ and $\mathbf{Y}$ are normalized, the similarity matrices are given by $\mathbf{C}^x = \mathbf{X}^\top \mathbf{X}$ and $\mathbf{C}^y = \mathbf{Y}^\top \mathbf{Y}$. In addition, let L be the ℓ_2 loss, i.e., $L(a, b) = |a - b|^2$. Then the objective in problem (4.21) becomes:

$$\begin{aligned}\mathcal{L}(\Gamma) &= \sum_{i,j,k,l} \left(\mathbf{C}^x_{ik} - \mathbf{C}^y_{jl}\right)^2 \Gamma_{ij}\Gamma_{kl} \\ &= \sum_{i,j,k,l} \left(\mathbf{C}^x_{ik}\right)^2 \Gamma_{ij}\Gamma_{kl} - 2\sum_{i,j,k,l} \left(\mathbf{C}^x_{ik}\mathbf{C}^y_{jl}\right)\Gamma_{ij}\Gamma_{kl} + \sum_{i,j,k,l} \left(\mathbf{C}^y_{jl}\right)^2 \Gamma_{ij}\Gamma_{kl}\end{aligned}$$

Since $\Gamma \in \Pi(\mathbf{a}, \mathbf{b})$, the first of these terms becomes

$$\sum_{i,k} \left(\mathbf{C}^x_{ik}\right)^2 \sum_{j,l} \Gamma_{ij}\Gamma_{jl} = \sum_{i,k} \left(\mathbf{C}^x_{ik}\right)^2 \mathbf{a}_i \mathbf{a}_k = \mathbf{a}^\top (\mathbf{C}^x)^2 \mathbf{a}$$

where the last equation follows from the definition of the transportation polytope. Crucially, this term does not depend on Γ anymore. Analogously, the last term in $\mathcal{L}(\Gamma)$ does not depend on Γ either, so

$$\operatorname*{argmin}_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \mathcal{L}(\Gamma) = \operatorname*{argmax}_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \sum_{i,j,k,l} \left(\mathbf{C}^x_{ik}\mathbf{C}^y_{jl}\right)\Gamma_{ij}\Gamma_{kl} \tag{4.22}$$

On the other hand, consider problem (4.20). The objective it seeks to maximize is

$$\begin{aligned}\|\mathbf{X}\Gamma\mathbf{Y}^\top\|_F^2 &= \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{X}\Gamma\mathbf{Y}^\top \rangle \\ &= \langle \mathbf{X}^\top\mathbf{X}\Gamma, \Gamma\mathbf{Y}\mathbf{Y}^\top \rangle \\ &= \sum_{i=1}^{n}\sum_{l=1}^{m} \left[\mathbf{X}^\top\mathbf{X}\Gamma\right]_{il} \left[\Gamma\mathbf{Y}^\top\mathbf{Y}\right]_{il} \\ &= \sum_{i=1}^{n}\sum_{l=1}^{m} \left[\mathbf{C}^x\Gamma\right]_{il} \left[\Gamma\mathbf{C}^y\right]_{il} \\ &= \sum_{i=1}^{n}\sum_{l=1}^{m} \left(\sum_{k=1}^{n} \mathbf{C}^x_{ik}\Gamma_{kl}\right)\left(\sum_{j=1}^{m} \Gamma_{ij}\mathbf{C}^y_{jl}\right) = \sum_{i=1}^{n}\sum_{l=1}^{m}\sum_{k=1}^{n}\sum_{j=1}^{m} \mathbf{C}^x_{ik}\Gamma_{kl}\Gamma_{ij}\mathbf{C}^y_{jl}\end{aligned}$$

which is equal to (4.22). Hence, Problems (4.20) and (4.21) are equivalent. □

4.6 Optimization Approaches

At this point, we are ready to discuss specific optimization approaches to the invariant optimal transport problem. In this section, we develop optimization approaches for each of the three equivalent formulations of the problem discussed in Section 4.4. As we will see later on (§4.8.2), the alternating minimization approach turns out to be the most stable and better performing in practice among these, so we discuss that method in considerably more detail than the others. Before developing the algorithms, however, we begin by deriving the necessary optimization ingredients (gradients and projects) that these rely on. Unless the reader has a specific interest in these derivations, the next subsection can be safely ignored or skimmed through.

4.6.1 Ingredients: gradients and projections

All the optimization methods proposed in this section —except, notably, the alternating minimization approach— are gradient-based, so we derive here first- and second-order derivatives of the three equivalent formulations of the invariant-OT problem described in Section 4.4, and discuss projections into the two constraint spaces of interest. In the interest of generality, we provide gradient derivations in the most general form possible. In particular, we will not assume a specific form of the invariance set $\mathcal{F}$ yet —in particular, it is not necessarily defined in terms of Schatten norms as in the previous section— unless otherwise noted.

We begin with two simple lemmas that will greatly simplify the notation throughout the remainder of this section, and which provide useful equivalences exploited by the subsequent results. The remaining four propositions each provide first- and second-order derivatives of the different views of Problem (4.6).

Lemma 4.6.1. *Let* $\Gamma \in \Pi(\mathbf{a}, \mathbf{b})$. *For any vectors* $\mathbf{u} \in \mathbb{R}^n$ *and* $\mathbf{v} \in \mathbb{R}^m$, *we have*

$$\langle \Gamma, \mathbf{u} \oplus \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{a} \rangle + \langle \mathbf{v}, \mathbf{b} \rangle \tag{4.23}$$

where $\oplus$ *is the Kronecker sum as defined in Section 2.1.*

Proof. Recall that $\Gamma \in \Pi(\mathbf{a}, \mathbf{b})$ implies that the row and column sums of Γ agree with $\mathbf{a}$ and $\mathbf{b}$. Hence,

$$\begin{aligned}\langle \Gamma, \mathbf{u} \oplus \mathbf{v} \rangle &= \sum_{ij} \Gamma_{ij} [\mathbf{u} \oplus \mathbf{v}]_{ij} = \sum_{ij} \Gamma_{ij} (\mathbf{u}_i + \mathbf{v}_j) \\ &= \sum_i \mathbf{u}_i \sum_j \Gamma_{ij} + \sum_j \mathbf{v}_j \sum_i \Gamma_{ij} = \sum_i \mathbf{u}_i \mathbf{a}_i + \sum_j \mathbf{v}_j \mathbf{b}_j\end{aligned}$$

for which the conclusion follows. □

Lemma 4.6.2. *For the squared euclidean cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$, the transportation cost can be written as:*

$$\mathcal{T}_c(f, \Gamma; \mathbf{a}, \mathbf{b}) = -2\langle \Gamma, \mathbf{X}^\top f(\mathbf{Y}) \rangle + \langle \Gamma, \mathbf{x}_{|||} \otimes \mathbb{1}_m \rangle + \langle \Gamma, \mathbb{1}_n \otimes f(\mathbf{y})_{|||} \rangle \tag{4.24}$$

where $\mathbf{x}_{|||}$ is the vector with row-wise norms of $\mathbf{X}$, i.e., $\mathbf{x}_{|||} = (\|\mathbf{x}^{(1)}\|^2, \ldots, \|\mathbf{x}^{(n)}\|^2)$, and analogously for $f(\mathbf{y})_{|||}$. Furthermore, for a feasible $\Gamma \in \Pi(\mathbf{a}, \mathbf{b})$, this is equivalent to:

$$\begin{aligned}\mathcal{T}_c(f, \Gamma; \mathbf{a}, \mathbf{b}) &= -2\langle \Gamma, \mathbf{X}^\top f(\mathbf{Y}) \rangle + \mathrm{Tr}(\mathrm{Cov}_{\mathbf{a}}(\mathbf{X})) + \mathrm{Tr}(\mathrm{Cov}_{\mathbf{b}}(f(\mathbf{Y}))) && (4.25) \\ &= -2\langle \Gamma, \mathbf{X}^\top f(\mathbf{Y}) \rangle + \sum_{i=1}^n \mathbf{a}_i \|\mathbf{x}^{(i)}\|_2^2 + \sum_{j=1}^m \mathbf{b}_i \|f(\mathbf{y}^{(j)})\|_2^2 && (4.26)\end{aligned}$$

where $\mathrm{Cov}_{\mathbf{a}}(\mathbf{X}) = \mathbf{X}[[\mathbf{a}]]\mathbf{X}^\top$ and $\mathrm{Cov}_{\mathbf{b}}(f(\mathbf{X})) = f(\mathbf{Y})[[\mathbf{b}]]f(\mathbf{Y})^\top$ are weighted covariance matrices.

Proof. For the squared euclidean cost, the cost matrix has entries

$$[\mathbf{C}(\mathbf{X}, f(\mathbf{Y}))]_{ij} = \|\mathbf{x}^{(i)}\|_2^2 + \|\mathbf{y}^{(j)}\|_2^2 - 2\langle \mathbf{x}^{(i)}, \mathbf{y}^{(j)} \rangle, \tag{4.27}$$

and thus can be expressed as

$$\begin{aligned}\mathbf{C}(\mathbf{X}, f(\mathbf{Y})) &= -2\mathbf{X}^\top f(\mathbf{Y}) + \mathbf{x}_{|||} \oplus f(\mathbf{y})_{|||} && (4.28) \\ &= -2\mathbf{X}^\top f(\mathbf{Y}) + \mathbf{x}_{|||} \otimes \mathbb{1}_m + \mathbb{1}_n \otimes f(\mathbf{y})_{|||} && (4.29)\end{aligned}$$

this yields the first part of the result. Now, for a feasible Γ, we can use Lemma 4.6.1 on step (4.28) to obtain:

$$\begin{aligned}
\langle \Gamma, \mathbf{C}(\mathbf{X}, f(\mathbf{Y})) \rangle &= -2\langle \Gamma, \mathbf{X}^\top f(\mathbf{Y}) \rangle + \langle \mathbf{x}_{|||}, \mathbf{a} \rangle + \langle f(\mathbf{y})_{|||}, \mathbf{b} \rangle \\
&= -2\langle \Gamma, \mathbf{X}^\top f(\mathbf{Y}) \rangle + \sum_{i=1}^{n} \mathbf{a}_i \|\mathbf{x}^{(i)}\|_2^2 + \sum_{j=1}^{m} \mathbf{b}_i \|f(\mathbf{y}^{(j)})\|_2^2 \\
&= -2\langle \Gamma, \mathbf{X}^\top f(\mathbf{Y}) \rangle + \mathrm{Tr}(\mathbf{X}^\top \mathbf{X}[[\mathbf{a}]]) + \mathrm{Tr}(f(\mathbf{Y})^\top f(\mathbf{Y})[[\mathbf{b}]]) \\
&= -2\langle \Gamma, \mathbf{X}^\top f(\mathbf{Y}) \rangle + \mathrm{Tr}(\mathbf{X}[[\mathbf{a}]]\mathbf{X}^\top) + \mathrm{Tr}(f(\mathbf{Y})[[\mathbf{b}]] f(\mathbf{Y})^\top)
\end{aligned}$$

which proves the second part of the statement. □

Remark 4.6.3. *Note that Lemma 4.6.2 implies that the problems* $\min_\Gamma \mathcal{T}_{\|x-y\|^2}(f, \Gamma; \mathbf{a}, \mathbf{b})$ *and* $\min_\Gamma \mathcal{T}_{-\langle x,y \rangle}(f, \Gamma; \mathbf{a}, \mathbf{b})$ *are equivalent.*

The following result provides explicit first- and second-order derivatives for the simultaneous optimization objective $\mathcal{T}(\mathbf{P}, \Gamma; \alpha, \beta)$ for the case where $\mathcal{F}$ consists of linear operators, i.e., $\mathcal{F} = \{f \mid f(\mathbf{y}) = \mathbf{P}\mathbf{y},\ \mathbf{P} \in \mathbb{R}^{d \times d}\}$.

Proposition 4.6.4 (Derivatives of transportation cost with linear transformation)**.** *For the squared euclidean cost* $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ *and* $f(\mathbf{y}) = \mathbf{P}\mathbf{y}$ *we have:*

$$\frac{\partial}{\partial \Gamma} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = \mathbf{C}(\mathbf{X}, \mathbf{P}\mathbf{Y}) \tag{4.30}$$

$$\frac{\partial}{\partial \mathbf{P}} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = 2\mathbf{P}\mathbf{Y}[[\mathbf{b}]]\mathbf{Y}^\top - 2\mathbf{X}\Gamma\mathbf{Y}^\top \tag{4.31}$$

$$\frac{\partial^2}{\partial \Gamma^2} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = \mathbb{0}_{n \times m} \tag{4.32}$$

$$\frac{\partial^2}{\partial \mathbf{P}^2} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = 2\mathbf{Y}[[\mathbf{b}]]\mathbf{Y}^\top \otimes \mathbf{I}_{d \times d} \tag{4.33}$$

Proof. Directly from the definition of $\mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b})$, we see that $\frac{\partial}{\partial \Gamma} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = \mathbf{C}(\mathbf{X}, \mathbf{P}\mathbf{Y})$, and therefore $\frac{\partial^2}{\partial \Gamma^2} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = \mathbb{0}_{n \times m}$.

Now, to compute derivatives with respect to $\mathbf{P}$, we assume Γ is feasible, and plug in the form of $f(\mathbf{Y}) = \mathbf{P}\mathbf{Y}$ in Equation (4.25), yielding:

$$\mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = -2\langle \Gamma, \mathbf{X}^\top \mathbf{P}\mathbf{Y} \rangle + \mathrm{Tr}(\mathrm{Cov}_{\mathbf{a}}(\mathbf{X})) + \mathrm{Tr}(\mathrm{Cov}_{\mathbf{b}}(\mathbf{P}\mathbf{Y})) \tag{4.34}$$

The derivative with respect to $\mathbf{P}$ of the first of these terms is immediate (by rewriting as $-2\langle \mathbf{P}, \mathbf{X}\Gamma\mathbf{Y}^\top \rangle$), while the second one does not depend on $\mathbf{P}$. For the third one we have

$$\tfrac{\partial}{\partial \mathbf{P}} \operatorname{Tr}(\operatorname{Cov}_{\mathbf{b}}(\mathbf{PY})) = \tfrac{\partial}{\partial \mathbf{P}} \operatorname{Tr}(\mathbf{P} \operatorname{Cov}_{\mathbf{b}}(\mathbf{Y})\mathbf{P}) = \mathbf{P}[\operatorname{Cov}_{\mathbf{b}}(\mathbf{Y}) + \operatorname{Cov}_{\mathbf{b}}(\mathbf{Y})^\top] = 2\mathbf{P} \operatorname{Cov}_{\mathbf{b}}(\mathbf{Y})$$

which yields the desired result. Finally, the second derivative with respect to $\mathbf{P}$ can be obtained immediately from this. □

Remark 4.6.5. *Note that if* $\mathbf{P}$ *is orthogonal,* $\operatorname{Tr}(\operatorname{Cov}_b(\mathbf{PY})) = \operatorname{Tr}(\operatorname{Cov}_b(\mathbf{Y}))$ *(equivalently,* $f(\mathbf{y})_{\|\|} = \mathbf{y}_{\|\|}$*), so in that case in the proof of Proposition 4.6.4 in Eq.* (4.34) *only the first term depends on* $\mathbf{P}$*, yielding instead* $\frac{\partial}{\partial \mathbf{P}} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = -2\mathbf{X}\Gamma\mathbf{Y}^\top$*. On the other hand, if* $\mathbf{Y}$ *is* ν*-whitened (i.e.,* $\mathbf{Y}[[\nu]]\mathbf{Y}^\top = \mathbf{I}_{d\times d}$*) we end up with* $2\mathbf{P}^\top - 2\mathbf{Y}\Gamma^\top\mathbf{X}^\top$*.*

We will often be interested in solving the entropy-regularized version of the transportation instead, for which we derive an analogous result below.

Proposition 4.6.6 (Derivatives of entropy-regularized transportation cost with linear transformation)**.** *For the squared euclidean cost* $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$*, we have:*

$$\tfrac{\partial}{\partial \Gamma} \mathcal{T}_c^\varepsilon(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = \mathbf{C}(\mathbf{X}, \mathbf{PY}) + \varepsilon \log \Gamma \tag{4.35}$$

$$\tfrac{\partial}{\partial \mathbf{P}} \mathcal{T}_c^\varepsilon(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = 2\mathbf{PY}[[\mathbf{b}]]\mathbf{Y}^\top - 2\mathbf{X}\Gamma\mathbf{Y}^\top \tag{4.36}$$

$$\tfrac{\partial^2}{\partial \Gamma^2} \mathcal{T}_c^\varepsilon(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = \varepsilon[[\operatorname{vec}(\Gamma)^{-1}]] \tag{4.37}$$

$$\tfrac{\partial^2}{\partial \mathbf{P}^2} \mathcal{T}_c^\varepsilon(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b}) = 2\mathbf{Y}[[\mathbf{b}]]\mathbf{Y}^\top \otimes \mathbf{I}_{d\times d} \tag{4.38}$$

Proof. Starting from Lemma 4.6.4 we only need to compute derivatives of $H(\Gamma)$. Note that:

$$\frac{\partial H(\Gamma)}{\partial \Gamma_{ij}} = \frac{\partial}{\partial \Gamma_{ij}} \left(-\Gamma_{ij}(\log \Gamma_{ij} - 1)\right) = -\log \Gamma_{ij} - 1 + 1 = -\log \Gamma_{ij} \tag{4.39}$$

so $\frac{\partial}{\partial \Gamma} - \varepsilon \operatorname{H}(\Gamma) = \varepsilon \log \Gamma$, where the log is to be understood element-wise. This yields

Equation (4.35). Taking derivatives again we get

$$\frac{\partial H(\Gamma)}{\partial \Gamma_{kl} \partial \Gamma_{ij}} = \begin{cases} -\Gamma_{ij}^{-1} & \text{if } i = k, j = l \\ 0 & \text{otherwise} \end{cases}, \tag{4.40}$$

which can be expressed in matrix form as $[[\mathrm{vec}(\Gamma)^{-1}]]$, resulting in Equation (4.37). Since the entropy regularization term is only a function of Γ and not of $\mathbf{P}$, the derivatives with respect to the latter are the same as in Proposition 4.6.4. □

We now turn our attention to the partial (single-block) objectives (4.10) and (4.11). Clearly, for these we only need to provide derivatives with respect one (the only) optimization variable: either Γ or $\mathbf{P}$. For the sake of conciseness, we now combine the derivatives for the unregularized and entropy-regularized versions of the problems each into one a single result.

Proposition 4.6.7 (Derivatives of $\mathcal{T}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b})$ and $\mathcal{T}^{\varepsilon}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b})$ with linear transformation). *For the squared euclidean cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$, we have:*

$$\frac{\partial}{\partial \Gamma} \mathcal{T}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b}) = \mathbf{C}(\mathbf{X}, \mathbf{P}^* \mathbf{Y}) \tag{4.41}$$

$$\frac{\partial}{\partial \Gamma} \mathcal{T}^{\varepsilon}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b}) = \mathbf{C}(\mathbf{X}, \mathbf{P}^*_{\varepsilon} \mathbf{Y}) + \varepsilon \log \Gamma \tag{4.42}$$

$$\frac{\partial^2}{\partial \Gamma^2} \mathcal{T}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b}) = \mathbb{0}_{n \times m} \tag{4.43}$$

$$\frac{\partial^2}{\partial \Gamma^2} \mathcal{T}^{\varepsilon}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b}) = \varepsilon [[\mathrm{vec}(\Gamma)^{-1}]] \tag{4.44}$$

where $\mathbf{P}^* = \mathrm{argmin}_{\mathbf{P} \in \mathcal{F}} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b})$ *and* $\mathbf{P}^*_{\varepsilon} = \mathrm{argmin}_{\mathbf{P} \in \mathcal{F}} \mathcal{T}^{\varepsilon}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b})$ *respectively.*

Proof. Since both $\mathcal{T}_{c|\mathcal{F}}(\gamma; \mathbf{a}, \mathbf{b})$ and $\mathcal{T}^{\varepsilon}_{c|\mathcal{F}}(\gamma; \mathbf{a}, \mathbf{b})$ are continuously differentiable, we can use the Envelope Theorem to obtain their derivatives through their inner optimization problem. Thus, we only need to plug-in the optimal values of $\mathbf{P}$ under these objectives in the derivatives for Γ computed in Propositions 4.6.4 and 4.6.6, which immediately yields the stated identities. □

Proposition 4.6.8 (Derivatives of $\mathrm{OT}_c(\alpha, f_\sharp\beta)$ and $\mathrm{OT}_c^\varepsilon(\alpha, f_\sharp\beta)$ with linear transformation). *For the squared euclidean cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$, we have:*

$$\frac{\partial}{\partial \mathbf{P}} \mathrm{OT}_c(\alpha, \mathbf{P}_\sharp\beta) = \frac{\partial}{\partial \mathbf{P}} \mathrm{OT}_c^\varepsilon(\alpha, \mathbf{P}_\sharp\beta) = 2\mathbf{P}\mathbf{Y}[[\mathbf{b}]]\mathbf{Y}^\top - 2\mathbf{X}\Gamma^*\mathbf{Y}^\top \tag{4.45}$$

$$\frac{\partial^2}{\partial \mathbf{P}^2} \mathrm{OT}_c(\alpha, \mathbf{P}_\sharp\beta) = \frac{\partial^2}{\partial \mathbf{P}^2} \mathrm{OT}_c^\varepsilon(\alpha, \mathbf{P}_\sharp\beta) = 2\mathbf{Y}[[\mathbf{b}]]\mathbf{Y}^\top \otimes \mathbf{I}_{d\times d} \tag{4.46}$$

where $\Gamma^* = \mathrm{argmin}_{\Gamma\in\Pi(\mathbf{a},\mathbf{b})} \mathcal{T}_c(\mathbf{P}, \Gamma; \mathbf{a}, \mathbf{b})$.

Proof. This is again a direct application of the Envelope Theorem combined with Lemma 4.6.4. □

4.6.2 Alternating minimization

Recall that whenever the conditions of Lemma 4.5.1 satisfied, Lemma 4.3.1 guarantees that the inner problem has a closed-form solution. In that case, we propose to solve problem (4.14) with alternating maximization on Γ and $\mathbf{P}$. Naturally, this falls within the scope of the first of the three equivalent objectives, namely the joint-optimization:

$$\min_{\substack{\gamma\in\Pi(\alpha,\beta)\\ f\in\mathcal{F}_p}} \mathcal{T}_c(f, \gamma; \alpha, \beta)$$

where we are optimizing $\mathcal{T}_c(f, \gamma; \alpha, \beta)$ by block coordinate descent on f and γ. Both for computational efficiency and performance reasons, we usually solve the entropy-regularized objective $\mathcal{T}_c^\varepsilon(f, \gamma; \alpha, \beta)$ instead, which under the aforementioned simplifying conditions would correspond to solving the problem:

$$\max_{\Gamma\in\Pi(\mathbf{a},\mathbf{b})} \max_{\mathbf{P}\in\mathcal{F}} \langle \Gamma, \mathbf{X}^\top\mathbf{P}\mathbf{Y}\rangle + \varepsilon\,\mathrm{H}(\Gamma). \tag{4.47}$$

For a fixed Γ, Lemma 4.5.2 shows a closed-from solution $\mathbf{P}^*$ at the cost of an $d \times d$ SVD, i.e., $O(d^3)$. For a fixed $\mathbf{P}$, the problem in Γ is clearly a traditional optimal transport problem, which can be solved with the methods discussed in Section 2.5. For example, for the entropy-regularized objective, the optimal Γ^* can be obtained in $O(N^3 \log N)$ with the Sinkhorn-Knopp algorithm and its variants. Note that the

Sinkhorn algorithm can be applied even to the original (non-regularized) problem (4.14) by relying on inexact alternating minimization methods that allow for approximate solution of intermediate steps [53, 130]. Besides providing an alternative algorithmic approach, this observation could be used to prove convergence rates for problem (4.14). Here, we instead focus on optimizing the regularized formulation (4.47).

Alternating optimization methods for non-convex objectives are known to be sensitive to initialization [91, 83]. Indeed, a key component of fully unsupervised approaches to feature alignment is finding good quality initializations. For example, for unsupervised word embedding alignment, state-of-the-art methods rely on additional—often heuristic—steps to generate good initial solutions, such as adversarially-trained neural networks [42, 184, 183], which themselves are often very sensitive to initialization, sometimes failing completely on the same problem for different random restarts [18].

Neither Problem (4.14) nor (4.47) is jointly concave in Γ and $\mathbf{P}$, thus facing in principle a similar challenge in terms of sensitivity to initialization. However, in (4.47) the strength of the entropic regularization controls the extent of non-concavity: strong regularization leads to a more concave objective, while $\varepsilon \to 0$ leads to an increasingly more non-concave objective. We propose to leverage this observation to alleviate sensitivity to initialization by using an annealing scheme on the regularization term.[8] Starting from a large value of ε, we decay this value in each iteration by setting $\varepsilon_t = \zeta \times \varepsilon_{t-1}$ with $\zeta \in (0, 1)$, until a minimum value $\underline{\varepsilon}$ is reached. We stop the method when the value of the objective converges. The advantage of this annealing approach is that it avoids ad-hoc initialization and eliminates the need for hyper-parameter tuning on ε, since any sufficiently large choice of ε_0 achieves the same objective. In *all* our experiments, we use the same parameter values $\varepsilon_0 = 1$ and decay $\zeta = 0.95$.

The method described so far, summarized as Algorithm 5 here, is used in our first set of experiments. Direct application of this method leads to high-quality solutions for small and mid-sized problems. However, scaling up to very large sets of points—e.g., hundreds of thousands of word embeddings in the word translation application—can

[8] As was recently brought to the author's attention, similar annealing schemes have been used in other assignment and optimal transport settings [104, 156]

be prohibitive.

We address this issue by dividing the problem into two phases. In the first stage, we solve a smaller problem (by taking a subsample of k points on each domain thus leading to smaller Γ and faster OT solution, but $\mathbf{P}$ of same size). Once the first phase reaches convergence, we use the solution $\mathbf{P}^*$ of the first stage to initialize the full-size problem. Note that while this might resemble other approaches that also consider a reduced set of points in their initialization step [42, 76], a crucial difference is that here we rely on the same optimization problem (4.47) in both stages, although with different problem sizes.

We experimented with various choices of parameter k, and observed that the algorithm is remarkably robust to the choice of this parameter. We conjecture that the ordering in which word embeddings are provided (higher-frequency words first, in every language) helps ensure that the solution of the initial problem of reduced size is consistent with the full-size problem.[9]

While the end performance is consistent regardless of the choice of sub-sample size k, there is naturally a trade-off in run time of the two stages. While solving a smaller initial problem is obviously faster, we observed that in such cases the second stage required more iterations to converge, suggesting that the initial $\mathbf{P}^*$ fed into the second stage was of lower quality (further from the optimal for the full-size problem). In the results presented in Section 4.8.4, we take k as large as possible while keeping the time-per-iteration reasonable: $k = 5000$.

Note that this strategy of *bootstrapping* solutions of smaller problems can be applied repeatedly, to increasingly grow the problem size over multiple stages. While we did not require to do so in our experiments, it might be an appealing approach for solving extremely large problems.

[9]This, in fact, points to an issue mostly ignored in previous work on this task: the order of the word embeddings *leaks* important —albeit noisy— correspondence information, which various methods presented as 'fully-unsupervised' seem to rely on one way or another, yet rarely acknowledge it.

Algorithm 5: Alternating Minimization for OT with Schatten Invariances

Input: Data matrices and histograms $(\mathbf{X}, \mathbf{a})$, $(\mathbf{Y}, \mathbf{b})$

Parameters: Order of invariance p and radius k_p;
initial and final entropy regularization ε_0 and $\underline{\varepsilon}$;
entropy regularization decay rate η.

Output: Transport coupling $\Gamma \in \mathbb{R}_+^{n \times m}$ and global mapping $\mathbf{P} \in \mathrm{O}(d)$

/* Initialize feasible transformation in $\mathcal{F}_I$ */

$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top \leftarrow \mathrm{SVD}(\textsc{RandomMatrix}(d \times d))$

$\boldsymbol{\sigma} \leftarrow \mathrm{diag}(\mathbf{\Sigma})$

$\mathbf{s} \leftarrow k_p \cdot \boldsymbol{\sigma} / \|\boldsymbol{\sigma}\|_p$

$\mathbf{P} = \mathbf{U}\,\mathrm{diag}(\mathbf{s})\mathbf{V}^\top$

$\varepsilon \leftarrow \varepsilon_0$

while *not converged* **do**

  /* Compute distances w.r.t. current mapping $\mathbf{P}$ */

  $\mathbf{C_P} \leftarrow \textsc{PairwiseDistances}(\mathbf{X}, \mathbf{PY})$

  /* Γ-Step via Sinkhorn-Iterations */

  $\mathbf{v} \leftarrow \mathbb{1}$

  $\mathbf{K} \leftarrow \exp\{-\mathbf{C_P}/\varepsilon\}$

  while *not converged* **do**

    $\mathbf{u} \leftarrow \mathbf{a} \oslash \mathbf{K}\mathbf{v}$

    $\mathbf{v} \leftarrow \mathbf{b} \oslash \mathbf{K}^\top \mathbf{u}$

  $\Gamma \leftarrow \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$

  /* $\mathbf{P}$-Step via Generalized Procrustes */

  $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top \leftarrow \mathrm{SVD}(\mathbf{X}\mathbf{\Gamma}\mathbf{Y}^\top)$

  $\boldsymbol{\sigma} \leftarrow \mathrm{diag}(\mathbf{\Sigma})$

  $q \leftarrow \frac{p}{p-1}$

  $\mathbf{s} \leftarrow k_p \cdot \boldsymbol{\sigma}^{q-1} / \|\boldsymbol{\sigma}^{q-1}\|_p$

  $\mathbf{P} = \mathbf{U}\,\mathrm{diag}(\mathbf{s})\mathbf{V}^\top$

  /* Anneal entropy regularization */

  $\varepsilon \leftarrow \max\{\varepsilon * \eta, \underline{\varepsilon}\}$

return $\Gamma, \mathbf{P}$

4.6.3 Joint gradient descent

In the previous section, we proposed solving the joint objective $\mathcal{T}_c(f, \Gamma; \mathbf{a}, \mathbf{b})$ via alternating minimization. Instead of solving these subproblems to completion, an alternative approach would be to take gradient steps simultaneously on the two optimization blocks (i.e., on Γ and f). Naturally, this being a constrained optimization problem on the constraint set $\mathcal{Z} = \Pi(\mathbf{a}, \mathbf{b}) \times \mathcal{F}_p$, the gradients would need to be project back to these sets. However, given the very different geometries of these constraint sets (i.e., the *correct* notion of projection on the former uses the KL metric, while for the latter euclidean projections are more meaningful), here we advocate against this approach and do not consider it further. In the next two sections, we instead explore gradient descent schemes on each of the two blocks of variables independently.

4.6.4 Single-block gradient descent

Recall from the equivalent expressions (4.10) and (4.11) of the invariant OT objective that we can "fold" the minimization with respect to one of the variables into the loss. In this case, evaluating the loss (and computing gradients, as discussed in Section 4.6.1) entails solving an optimization problem on the other variable.

Descent on f

Let us first consider Problem (4.11), where minimization over the transportation polytope is "folded" into the objective, yielding a minimization over transformations in $\mathcal{F}$ of the optimal transport cost, i.e., $\min_{f \in \mathcal{F}} \mathrm{OT}_c(\alpha, f_\sharp \beta)$. In Section 2.7 we discussed the use of optimal transport distances as a learning loss. Recall that to avoid the problem of biased gradients the entropy-regularized problem OT_ε, it is common practice to use the "normalized" Sinkhorn Divergence (2.20) instead. We can trivially define a generalized version of this divergence that includes the transformation f in its computation:

$$\mathrm{SD}^\varepsilon(\alpha, f_\sharp \beta) \triangleq \mathrm{W}^\varepsilon(\alpha, f_\sharp \beta) - \tfrac{1}{2}\big(\mathrm{W}^\varepsilon(\alpha, \alpha) + \mathrm{W}^\varepsilon(f_\sharp \beta, f_\sharp \beta)\big). \tag{4.48}$$

With this, our final optimization objective in this setting takes the form:

$$\min_{f \in \mathcal{F}} \mathrm{SD}^{\varepsilon}(\alpha, f_\sharp \beta) \tag{4.49}$$

In Proposition 4.6.8 we derived gradients and Hessians for $\mathrm{OT}(\alpha, f_\sharp \beta)$, which trivially yield their counterparts for $\mathrm{SD}^{\varepsilon}(\alpha, f_\sharp \beta)$. Therefore, we can solve this as a constrained optimization problem. Depending on the invariance class, we propose different approaches. If $\mathcal{F}$ is a well-behaved manifold over which projections can be computed efficiently, such as the Stiefel manifold, we can use manifold optimization to solve this problem [182, 178]. In our experiments we use the `pymanopt` toolbox for optimization over the Stiefel manifold.

Alternatively, if one wishes to maintain a more general class of invariances $\mathcal{F}$, the problem can be solved by parametrizing this class with deep neural networks. Since the objective is fully differentiable, it can be used to learn end-to-end the parameters of this neural network with backpropagation and stochastic gradient descent.

In either of these approaches, there are two possibilities for computing gradients. We can use the explicit gradients derived in Section 4.6.1, or leverage modern packages for automatic differentiation.

Descent on Γ

Now, we revisit instead Problem (4.10), where the optimization over the invariance set has been folded into the cost objective, yielding a cost-invariant optimal transport problem, i.e., $\min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \mathcal{T}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b})$. In this case, the problem can be solved via constrained optimization approaches, such as conditional or projected gradient methods. We use a projected gradient approach, similar to that of Peyré et al. [141]. Using the KL projection and a step size τ, the iterations are computed as

$$\Gamma \leftarrow \mathrm{Proj}^{\mathrm{KL}}_{\Pi(\mathbf{a},\mathbf{b})}\left(\Gamma \odot e^{-\tau(\nabla \mathcal{T}_{c|\mathcal{F}}(\Gamma;\mathbf{a},\mathbf{b}))}\right) \tag{4.50}$$

As for the single-block descent on f, gradients can be computed explicitly (using Proposition 4.6.7) or by means of automatic differentiation. The stepsize is a crucial

hyperparameter in gradient descent methods, and here its relation with the regularization parameter ε is particularly important. In fact, for a very specific choice of stepsize, project gradient descent on $\mathcal{T}_{c|\mathcal{F}}(\Gamma; \mathbf{a}, \mathbf{b})$ becomes equivalent to the alternating minimization approach of the previous section, as shown in the following result.

Proposition 4.6.9. *For the choice of stepsize $\tau = 1/\varepsilon$, iteration (4.50) becomes:*

$$\mathbf{P}^* \leftarrow \operatorname*{argmin}_{\mathbf{P} \in \mathcal{F}_p} \langle \Gamma, \mathbf{C}(\mathbf{X}, \mathbf{P}\mathbf{Y}) \rangle$$
$$\Gamma \leftarrow \textsc{Sinkhorn}(\mathbf{C}(\mathbf{X}, \mathbf{P}^*\mathbf{Y}), \mathbf{a}, \mathbf{b}, \varepsilon)$$

Proof. As discussed in Section 2.4, the projection onto $\Pi(\mathbf{a}, \mathbf{b})$ according to the KL divergence leads to a regularized OT problem:

$$\operatorname{Proj}^{\mathrm{KL}}_{\Pi(\mathbf{a},\mathbf{b})}(\mathbf{M}) \triangleq \operatorname*{argmin}_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \mathrm{KL}(\Gamma \parallel \mathbf{M}) = \operatorname*{argmin}_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \mathrm{OT}^{\varepsilon}_{-\varepsilon \log(\mathbf{M})}(\mathbf{a}, \mathbf{b})$$

Taking log of the expression in iteration (4.50), and plugging in the gradient value computed in Proposition 4.6.7, we get

$$\begin{aligned}
\log\big(\Gamma \odot e^{-\tau(\nabla \mathcal{T}_{c|\mathcal{F}}(\gamma;\alpha,\beta))}\big) &= \log \Gamma - \tau\big(\nabla \mathcal{T}_{c|\mathcal{F}}(\gamma; \alpha, \beta)\big) \\
&= \log \Gamma - \tau\big(\mathbf{C}(\mathbf{X}, \mathbf{P}^*\mathbf{Y}) + \varepsilon \log \Gamma\big) \\
&= -\tau \mathbf{C}(\mathbf{X}, \mathbf{P}^*\mathbf{Y}) + (1 - \varepsilon\tau) \log \Gamma = -\frac{1}{\varepsilon}\mathbf{C}(\mathbf{X}, \mathbf{P}^*\mathbf{Y})
\end{aligned}$$

Therefore, the OT problem related to the KL projection has precisely $\mathbf{C}(\mathbf{X}, \mathbf{P}^*\mathbf{Y})$ as a cost (after cancellation of signs and ε), so the updated Γ can be computed with the Sinkhorn algorithm on this cost matrix as stated. □

As Peyré et al. [141] note for their own approach the iterations (4.50) are guaranteed to converge for small enough stepsize τ. The simplifying choice $\tau = 1/\varepsilon$ in Proposition 4.6.9 is usually too large for this guarantee to hold. In our experiments, we observed monotonously decreasing objectives value and good empirical performance for most reasonable stepsize values and schemes, including —unsurprisingly, given the success of the alternating minimization approach— for $\tau = 1/\varepsilon$.

4.7 An Alternative Gromov-Wasserstein Approach

As discussed in Section 4.3.2, there exists a recent generalization of the optimal transport problem that allows defining the problem over *incomparable* spaces, i.e., when a metric between them is not available. The Gromov-Wasserstein distance [126] generalizes optimal transport by comparing the metric spaces directly instead of samples across the spaces. In other words, this framework operates on distances between pairs of points calculated within each domain and measures how these distances compare to those in the other domain. Thus, it requires a weaker but easy to define notion of *distance between distances*, and operates on pairs of points, turning the problem from a linear to a quadratic one. In Figure 4-2 we show an intuitive description of why the Gromov-Wasserstein distance is well suited to our motivating problem of unsupervised word embedding alignment.

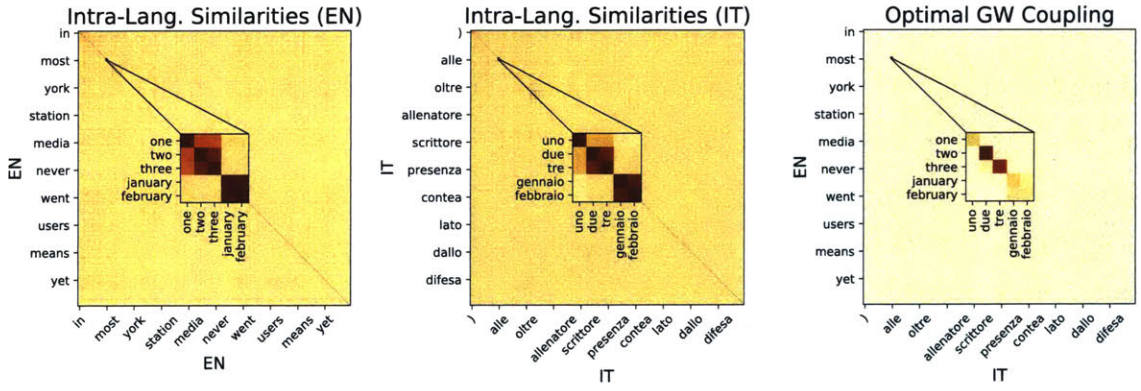


Figure 4-2: The Gromov-Wasserstein distance is well suited for the task of cross-lingual alignment because it relies on *relational* rather than *positional* similarities to infer correspondences across domains. Computing it requires two intra-domain similarity (or equivalently cost) matrices (**left** & **center**), and it produces an optimal coupling of source and target points with minimal discrepancy cost (**right**).

Compared to the classic optimal transport problem, computing the Gromov-Wasserstein distance is substantially harder since the objective is now not only non-linear, but non-convex too.[10] In addition, it requires operating on a fourth-order tensor, which would be prohibitive in most settings. Surprisingly, this problem can be

[10]In fact, the discrete (Monge-type) formulation of the problem is essentially an instance of the well-known (and NP-hard) quadratic assignment problem (QAP).

optimized efficiently with first-order methods, whereby each iteration involves solving a traditional optimal transport problem [164, 141].

Recall that the Gromov-Wasserstein problem corresponds to a proper distance (§4.3.2). Therefore, when applied to the problem of unsupervised word embedding alignment, this approach yields a fascinating accompanying notion: the *Gromov-Wasserstein distance between languages*, a measure of semantic discrepancy purely based on the relational characterization of their word embeddings. Owing to Theorem 4.3.2, such values can be interpreted as distances, so that, e.g., the triangle inequality holds among them. In Section 4.8.5 we compare various languages in terms of their GW-distance.

While the pure Gromov-Wasserstein approach leads to high-quality solutions, it is best suited to small-to-moderate problems,[11] since its optimization becomes prohibitive for very large problems. For such settings, we propose a two-step approach in which we first match a subset of the samples via the optimal coupling, after which we learn an orthogonal mapping through a modified Procrustes problem. Formally, suppose we solve problem (4.2) for a reduced matrices $\mathbf{X}_{1:k}$ and $\mathbf{Y}_{i:k}$ consisting of the first columns k of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and let Γ^* be the optimal coupling. We seek an orthogonal matrix that best recovers the barycentric mapping implied by Γ^*. Namely, we seek to find $\mathbf{P}$ which solves:

$$\min_{\mathbf{P} \in \mathrm{O}(n)} \|\mathbf{X}\Gamma^* - \mathbf{P}\mathbf{Y}\|_2^2 \tag{4.51}$$

It is easy to show that this is equivalent to a Procrustes problem (c.f. (4.1)), so it has a closed-form solution in terms of a singular value decomposition. Namely, the solution to (4.51) is $\mathbf{P}^* = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U}\Sigma\mathbf{V}^* = \mathbf{X}_{1:m}\Gamma^*\mathbf{Y}_{1:m}^\top$. After obtaining this projection, we can immediately map the rest of the embeddings via $\hat{\mathbf{y}}^{(j)} = \mathbf{P}^*\mathbf{y}^{(j)}$.

We end this section by discussing parameter and configuration choices. To leverage the fast algorithm of Peyré et al. [141], we always use the L_2 distance as the loss function L between cost matrices. On the other hand, we observed throughout our

[11] As shown in the experimental section, we are able to run problems of size in the order of $|V_s| \approx 10^5 \approx |V_t|$ on a single machine **without** relying on GPU computation.

Algorithm 6: Gromov-Wasserstein Computation for Embedding Alignment

Input: Source and target embeddings $\mathbf{X} \in \mathbb{R}^{d \times n}$, $\mathbf{Y} \in \mathbb{R}^{d \times m}$, and their corresponding probability vectors $\mathbf{a} \in \Sigma_n, \mathbf{b} \in \Sigma_m$.

Parameters: Regularization strength ε.

Output: Transport coupling $\Gamma \in \mathbb{R}_+^{n \times m}$ and global mapping $\mathbf{P} \in \mathrm{O}(d)$

/* Compute intra-domain similarities */

$\mathbf{C}_s \leftarrow \textsc{PairwiseDistances}(\mathbf{X}, \mathbf{X})$

$\mathbf{C}_t \leftarrow \textsc{PairwiseDistances}(\mathbf{Y}, \mathbf{Y})$

$\mathbf{C}_{st} \leftarrow \mathbf{C}_s^2 \mathbf{a} \mathbb{1}_m^\top + \mathbb{1}_n \mathbf{b} (\mathbf{C}_t^2)^\top$

while *not converged* **do**

  /* Compute pseudo-cost matrix (Eq. (4.4)) */

  $\hat{\mathbf{C}}_\Gamma \leftarrow \mathbf{C}_{st} - 2\mathbf{C}_s \Gamma \mathbf{C}_t^\top$

  /* Sinkhorn-Iterations */

  $\mathbf{u} \leftarrow \mathbb{1}$

  $\mathbf{K} \leftarrow \exp\{-\hat{\mathbf{C}}_\Gamma / \varepsilon\}$

  while *not converged* **do**

    $\mathbf{u} \leftarrow \mathbf{a} \oslash \mathbf{K}\mathbf{v}$

    $\mathbf{v} \leftarrow \mathbf{b} \oslash \mathbf{K}^\top \mathbf{u}$

  $\Gamma \leftarrow \mathrm{diag}(\mathbf{u}) \mathbf{K} \, \mathrm{diag}(\mathbf{v})$

/* [Optional] Learn explicit transformation to extrapolate */

$\mathbf{U}, \Sigma, \mathbf{V}^\top \leftarrow \mathrm{SVD}(\mathbf{X} \Gamma \mathbf{Y}^\top)$

$\mathbf{P} = \mathbf{U}\mathbf{V}^\top$

return $\Gamma, \mathbf{P}$

experiments that the choice of cosine distance as the metric in both spaces consistently leads to better results, which agrees with common wisdom on computing distances between word embeddings. This leaves us with a single hyper-parameter to control: the entropy regularization term ε. By applying any sensible normalization on the cost matrices (e.g., dividing by the mean or median value), we are able to almost entirely eliminate sensitivity to that parameter. In practice, we use a simple scheme in all experiments: we first try the same fixed value ($\varepsilon = 5 \times 10^{-5}$), and if the regularization proves too small (by leading to floating-point errors), we instead use $\varepsilon = 1 \times 10^{-4}$. We never had to go beyond these two values in all our experiments.

We emphasize that at no point we use train (let alone test) supervision available with many datasets—model selection is done solely in terms of the unsupervised objective. Pseudocode for the full method (with $L = L_2$ and cosine similarity) is shown here as Algorithm 6.

4.8 Experiments

Through this experimental evaluation we seek to: (i) validate the framework of optimal transport with invariances and compare the various optimization methods discussed before in a controlled setting (§4.8.2), (ii) understand their optimization dynamics (§4.8.3), (iii) evaluate their performance on benchmark cross-lingual word embedding tasks (§4.8.4), and (iv) qualitatively investigate the notion of distance-between-languages that the Gromov Wasserstein approach provides (§4.8.5). Due to the computational burden of the word embedding alignment task, we select from among the optimization approaches proposed in Section 4.6 the best performing one, and compare that one against third-party state-of-the-art methods.

When evaluating our proposed approaches in the context of word embedding alignment, rather than focusing solely on prediction accuracy, we seek to demonstrate that these offer a fast, principled, and robust alternative to alternative methods for these unsupervised alignment tasks. Here again, as was the case throughout this chapter, our focal point is the Schatten invariance framework solved by alternating optimization, so most of our experiments revolve around this method.

4.8.1 Evaluation tasks and methods

Datasets For the first set of experiments we use synthetic datasets consisting of simple 2D and 3D point clouds, in which one of the two clouds is obtained by applying a (known) transformation on the other. The second —and main— part of the experimental evaluation revolves around the task of unsupervised word embedding alignment. For this, we consider two standard benchmark tasks for cross-lingual embeddings. First, we consider the dataset of Conneau et al. [42], which consists of word embeddings trained with FASTTEXT [28] on Wikipedia and parallel dictionaries for 110 language pairs. Here, we focus on the language pairs for which they report results: English (EN) from/to Spanish (ES), Italian (IT), French (FR), German (DE), Russian (RU) and simplified Chinese (ZH). We also experiment with the —substantially

harder[12]—dataset of [51], which has been extensively compared against in previous work. It consists of embeddings and dictionaries in four pairs of languages; EN from/to ES, IT, DE, and FI (Finnish).

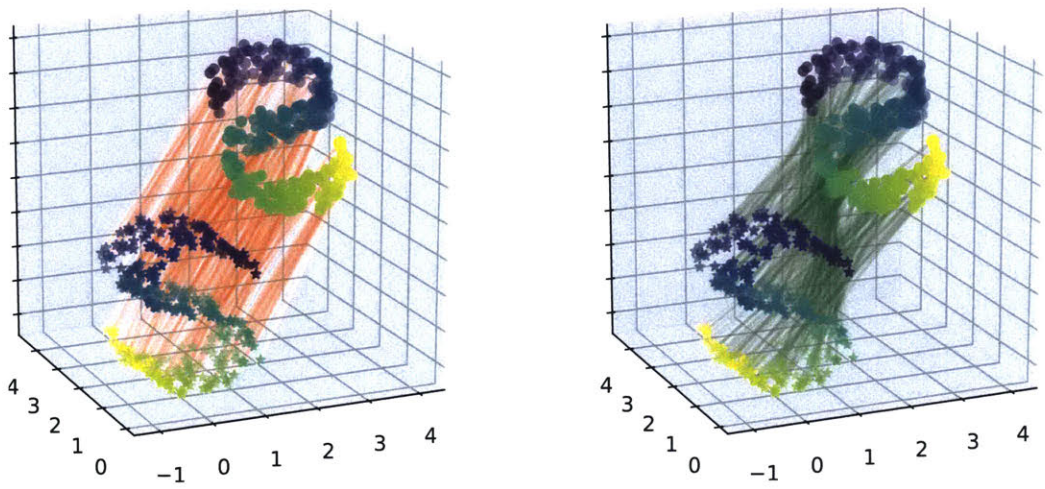
Methods In the first part of the experiments we compare our methods mostly against classic (invariance-agnostic) versions of the optimal transport problem, in addition to the Iterative Closest Points (ICP) method [40, 26]. For the word translation task, we follow Conneau et al. [42] and consider a simple but strong baseline consisting of solving a Procrustes problem directly using the available cross-lingual embedding pairs (we refer to this method simply as PROCRUSTES), emphasizing that this baseline is not fully unsupervised. In addition, we compare against the fully-unsupervised methods of Zhang et al. [183] (ADV), Artetxe et al. [15] (SELF-LEARN), Conneau et al. [42] (MUSE) and Grave et al. [76] (WASSERSTEIN). The code for the last of these was not available, so we report results from their paper (which excludes EN–IT), and omit runtime. As first proposed by Conneau et al. [42], we use Cross-domain Similarity Local Scaling (CSLS) whenever a nearest-neighbor search is required, which has been shown to improve upon naive nearest-neighbor retrieval in various works.

4.8.2 Recovery and noise robustness on synthetic datasets

We first test our approach in a controlled setting with known underlying invariance. We generate a point cloud in $\mathbb{R}^2$ or $\mathbb{R}^3$ (the *source*), and then apply a transformation $\mathbf{P}$ sampled randomly from one of the families $\mathcal{F}_p$ to generate a *target* point cloud. The goal is thus to recover the true correspondences between source and target points. We generate a discrete matching ψ from a coupling Γ as $\psi(i) = \text{argmax}_j \Gamma_{ij}$, and compute its accuracy with respect to the known true point-wise correspondences. Throughout this section we run all methods with 10 random restarts and keep the best performing one in terms of the optimization objective (not the accuracy itself, so as to simulate a truly unsupervised scenario with no known correspondences for validation).

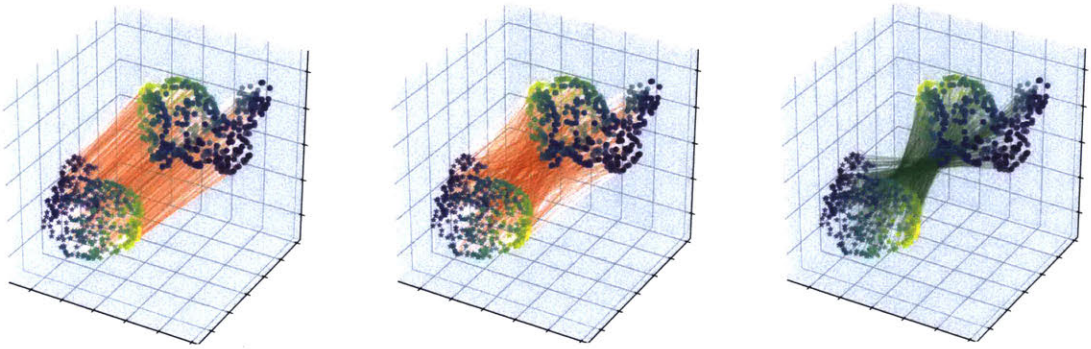
An instance of this dataset and the corresponding solutions found using the classic

[12]We discuss the difference in hardness of these two benchmark datasets in Section 4.8.4.



(A) Classic OT (B) ℓ_∞-invariant OT

Figure 4-3: Optimal couplings for the synthetic point cloud dataset with underlying orthogonal ($\mathcal{F}_\infty$) invariance; green (red) edges denote correct (incorrect) matchings.



(A) Classic OT (B) Iterative Closest Points (C) ℓ_∞-invariant OT

Figure 4-4: Optimal couplings for the synthetic point cloud dataset with underlying orthogonal ($\mathcal{F}_\infty$) invariance; green (red) edges denote correct (incorrect) matchings.

and invariant versions of OT are shown in Figure 4-3. As expected, when the true latent transformation is orthogonal, endowing OT with ℓ_∞ invariance allows it to recover the correct matching between the point clouds, while the classic (invariance-agnostic) formulation does not, greedily matching based on proximity instead. We observed that most of the proposed optimization approaches to solve the invariant OT problem perform almost indistinguishably well in these simple datasets where perfect recovery is possible. Surprisingly, even the *unconstrained* version of the **P**-descent method was able to recover the correct mapping and matchings in all of these settings. We show another instance in Figure 4-4. Additional results can be found in Appendix A.

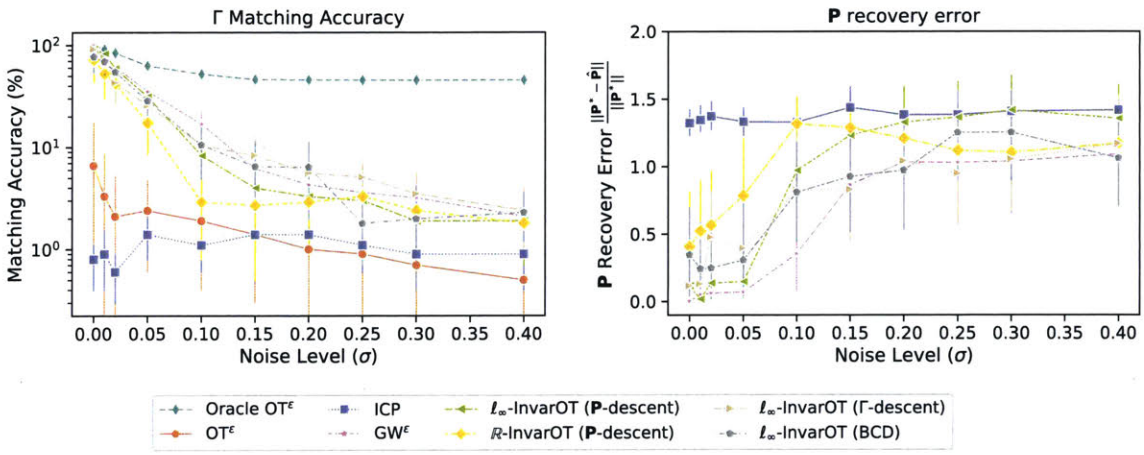
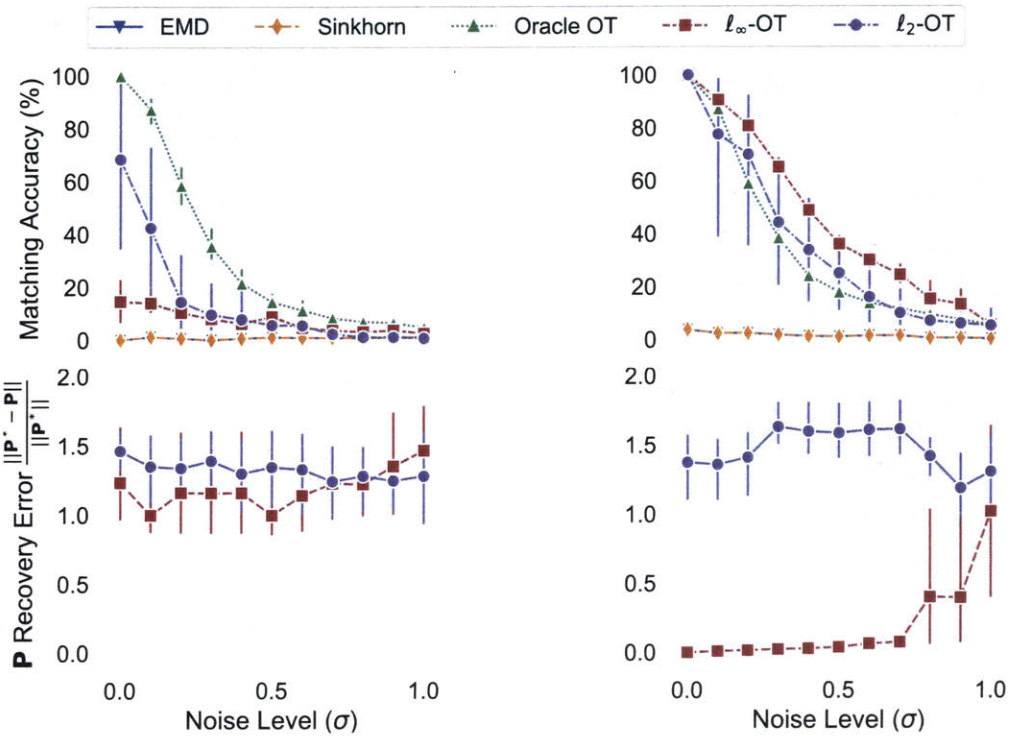


Figure 4-5: Noise-robustness comparison for the various optimization approaches of the proposed invariant OT method and baselines. These results corresponds to the same BUNNY dataset as before, with a latent orthogonal transformation and various levels of noise. **Left**: matching accuracy for the computed coupling $\hat{\Gamma}$. **Bottom**: error in recovering $\mathbf{P}$. The plots show mean values and one s.d. error bars over 5 repetitions.

Few real applications, however, would involve such a simple scenario with perfect cross-domains correspondence. Instead, the two collections of interest are likely to exhibit some level of noise. Thus, even if the true latent invariance is of the prescribed class (Schatten-bounded norm in our assumptions), it might be hard to find it because no transformation would yield perfect correspondence. To investigate the robustness of our methods with respect to noise, we simulate noisy correspondences as follows. As before, we generate point clouds with two types of invariances ($\mathcal{F}_2$ and $\mathcal{F}_\infty$), but now add a Gaussian noise term with variance σ to the target points.

In the first instance, we consider a dataset with a latent orthogonal transformation, and compare all of our optimization methods and the following baselines: iterative closest points (ICP), classic OT (2.5), the entropy-regularized formulation OT^ϵ (i.e., Problem (2.11)) solved via the SINKHORN algorithm, and an ORACLE which solves a entropy-regularized problem *without the transformation applied*, i.e. only adding noise. Figure 4-5 shows the matching accuracy (mean and one standard deviation over 5 repetitions), where again each method is run with 10 random restarts. Notably, the Gromov-Wasserstein method performs very well for small noise levels, while the Invariant OT varieties solved with block-coordinate descent and $\mathbf{P}$-descent are



(A) True transform in $\mathcal{F}_2$ (B) True transform in $\mathcal{F}_\infty$

Figure 4-6: Robustness with respect to noise when matching under latent Schatten invariances of two types. **Top**: matching accuracy for the computed coupling $\hat{\Gamma}$. **Bottom**: error in recovering $\mathbf{P}$. The plots show mean values and one s.d. error bars over 5 repetitions.

consistently among the best for all levels of noise. Note also that (unsurprisingly) the invariance-agnostic entropic OT and (somewhat surprisingly) ICP are by far the two worst performing methods in this dataset.

Next, we investigate the effect of misspecification of the invariance class $\mathcal{F}_p$, again at different levels of noise.[13] We now compare two versions of the invariant OT objective with different invariances (the ℓ_2 and ℓ_∞ Schatten norm cases, i.e., invariance to Frobenius and orthogonal transformations, respectively), in addition to the same baselines as before. As before, Figure 4-6 (top) shows the matching accuracy (mean and one standard deviation over 5 repetitions). As expected, ℓ_∞-OT is better than ℓ_2-OT at recovering the correspondences when the true transformation is orthogonal,

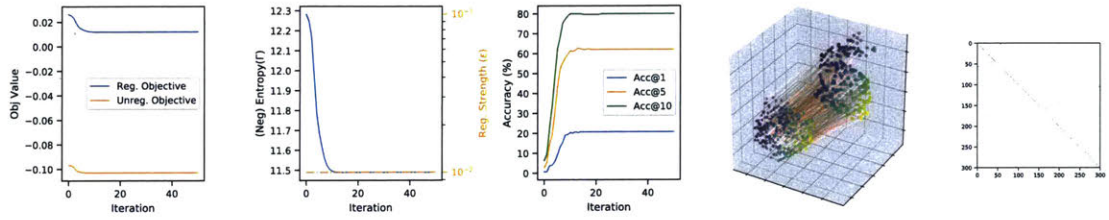
[13]This time, we use a slightly different noise scheme, where now the noise is added per entry, i.e., using $\hat{x}_j^{(i)} \leftarrow \hat{x}_j^{(i)} + N(0, \sigma)$. This leads to lower relative distortion for the same level of σ compared to the multivariate Gaussian noise approached used before.

and vice versa; and there is a loss of accuracy caused by the estimation of $\mathbf{P}$ in the ℓ_2 case, as shown by the gap between our methods and ORACLE OT. But surprisingly, both invariance methods outperform the oracle in the ℓ_∞ case, which we attribute to the added freedom of choosing $\mathbf{P}$ to overcome noise, combined with the ease of optimizing over $\mathcal{F}_\infty$ compared to $\mathcal{F}_2$. This hypothesis is supported by the overall higher error in recovering $\mathbf{P}$ in the latter case (Fig. 4-6, bottom).

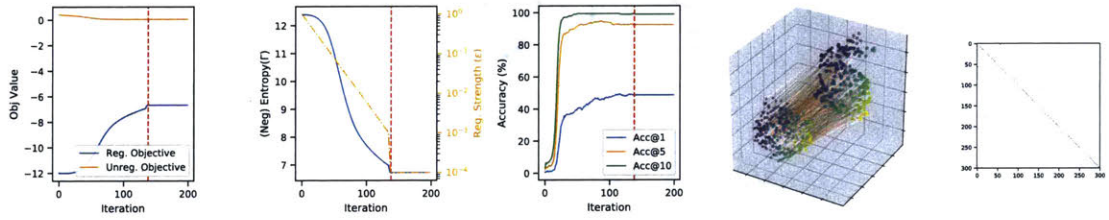
4.8.3 Optimization dynamics

Figure 4-7 shows the best-of-then restarts run for the various optimization approaches to the invariant OT problem and the entropic Gromov-Wasserstein alternative in the same bunny point cloud used before, now with noise added at level $\sigma = 0.1$. There are various interesting phenomena to discuss here. First, it is clear from the couplings and optimal matches that the amount of noise in this instance makes the alignment task quite challenging. In terms of optimization dynamics, all methods shown there exhibit a mostly smooth and monotonous objective drop, which —crucially— correlates strongly with the matching accuracy (middle column). Since such validation curves would not be available during training in a truly unsupervised setting, it is important that the (unsupervised) objective be a good predictor of the metric of interest, since model selection, random restarting and early stopping would all need to be done on the former.

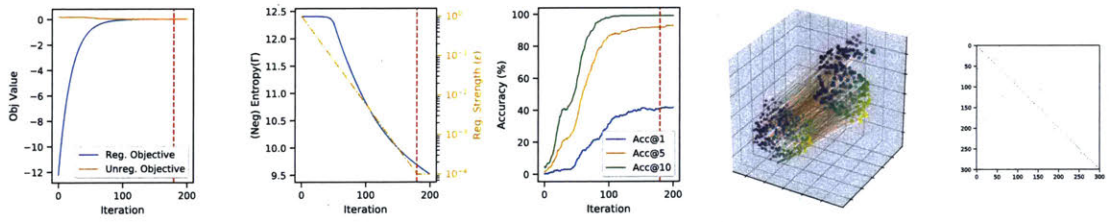
Also worth pointing out is the fact that the optimization approaches that treat Γ as one of the optimization blocks (namely, the block coordinate descent and Γ-descent methods – rows (B) and (C) in Figure 4-7) they allow for smoothly annealing of the entropy regularization strength, which leads to smooth and gradual sharpening of the optimal coupling (second column from the left). For the same reason (high entropy of Γ^* in initial iterates), the difference between regularized and unregularized objectives (first column) is more pronounced than for all the other methods. On the flip side, these methods take a far larger number of iterations to reach the same accuracy than the others (again, a consequence of the entropy annealing scheme, which causes initial couplings to be too dense for accurate matching inference).



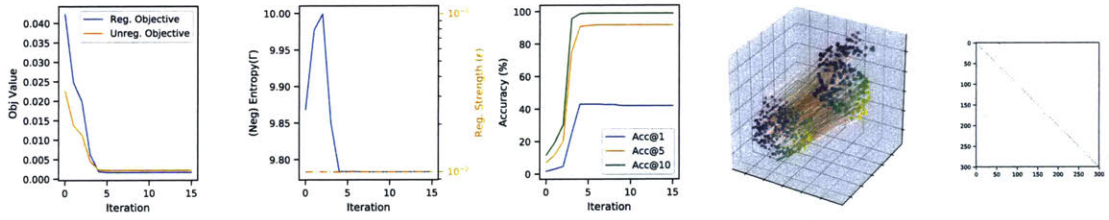
(A) (Entropic) Gromov-Wasserstein alignment



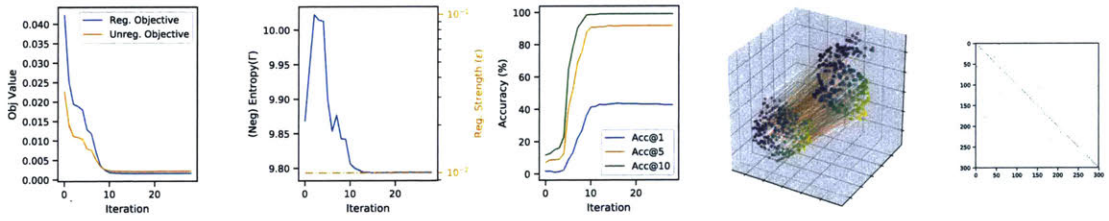
(B) Block Coordinate Descent on Γ and $\mathbf{P}$



(C) Single-block descent on Γ via Projected Gradient Descent on $\Pi(\mathbf{a}, \mathbf{b})$



(D) Single-block descent on $\mathbf{P}$ via Riemannian conjugate gradient on $\mathrm{O}(n)$



(E) Single-block descent on $\mathbf{P}$ via unconstrained conjugate gradient on $\mathbb{R}^d$

Figure 4-7: Training dynamics for the various invariant OT approaches on a simple 3D shape matching task with underlying $\mathcal{F}_\infty$ invariance and added noise ($\sigma = 0.1$). Shown here is the best-of-ten restart for each model. The first three panes show objective values, entropy regularization and matching accuracy; the right-most two show the optimal coupling represented as pairwise matches and the transportation coupling Γ^*. Vertical red dashed lines indicate entropy decay was frozen at that iteration.

4.8.4 Unsupervised word translation

As discussed in the introduction to this chapter, unsupervised word translation is an ideal testbed for optimal transport with invariances. Most recent fully unsupervised methods cast the problem as feature alignment between sets of word embeddings, motivated by the observation that these possess similar geometry across languages [128]. Though their relational structure is similar, the absolute position of these vectors is irrelevant. Indeed, word embedding algorithms are naturally interpreted as metric recovery methods [84], making these vectors intrinsically invariant to angle (or distance) preserving transformations. This observation suggests inducing invariance to orthogonal transformations, as described in Section 4.5.1.

Most current unsupervised methods circumvent this issue by resorting to ad-hoc normalization, joint re-embedding, or by estimating a complex mapping between the two spaces with adversarial training. These methods require careful initialization and post-mapping refinements, such as mitigating the effect of frequent words on neighborhoods, and are often hard to tune properly [15].

In our optimal-transport based approaches, the optimal transport coupling Γ^* provides an explicit (soft) matching between source and target samples, which for the problem of interest can be interpreted as a probabilistic translation: for every pair of words $(w_{src}^{(i)}, w_{trg}^{(j)})$, Γ_{ij}^* provides a likelihood that these two words are translations of each other. This itself is enough to translate, and we show in the experiments section that Γ^* by itself, without any further post-processing, provides high-quality translations. This stands in sharp contrast to mapping-based methods, which rely on nearest-neighbor computation to infer translations, and thus become prone to hub-word effects which have to be mitigated with heuristic post-processing techniques such as Inverted Softmax [162] and Cross-Domain Similarity Scaling (CSLS) [42]. The transportation coupling Γ, being normalized *by construction*, requires no such artifacts.

When solving this task via the Gromov-Wasserstein distance, we rely on the scheme described in Section 4.7 to scale up to the very large problem size that the vocabularies in this task imply. We point out that this two-step procedure resembles that of

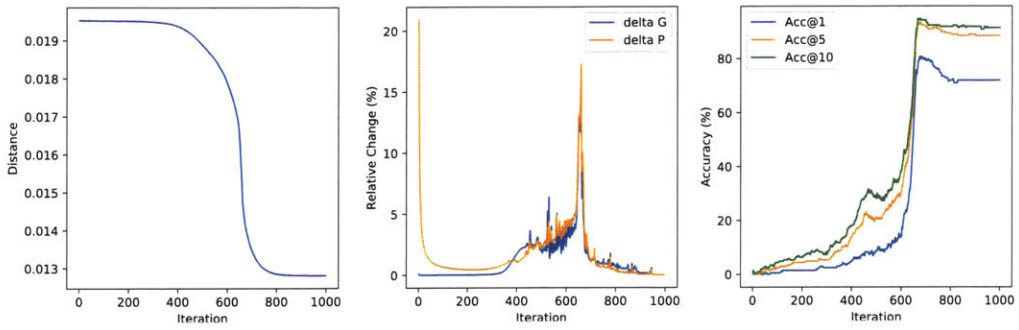


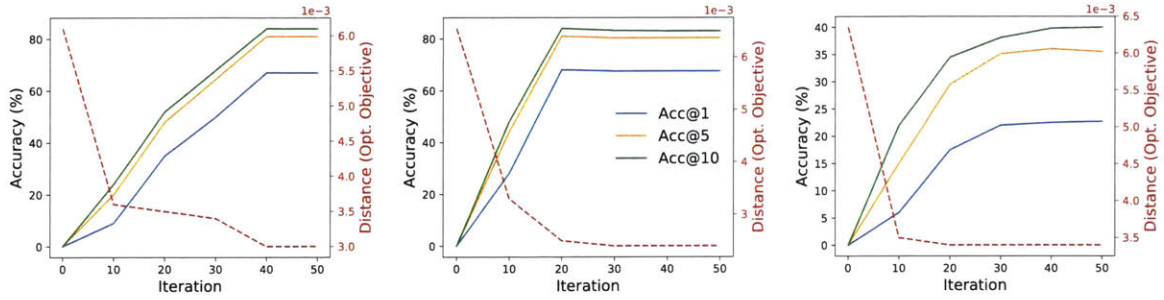
Figure 4-8: **Training dynamics for the ℓ_∞-Invariant-OT approach on the word translation task**. Left to right: objective, change in $\mathbf{P}$ and Γ from previous iteration, and translation accuracy, on a typical EN→IT run with a 5K vocabulary.

Conneau et al. [42]. Both ultimately produce an orthogonal mapping obtained by solving a Procrustes problem, but they differ in the way they produce pseudo-matches to allow for such second-step: while their approach relies on an adversarially-learned transformation, we use an explicit optimization problem.

Finally, we note that whenever word frequency counts are available, those would be used for $\mathbf{a}$ and $\mathbf{b}$. If they are not, but words are sorted according to frequency (as they often are in popular off-the-shelf embedding formats), one can estimate rank-probabilities such as Zipf power laws, which are known to accurately model multiple languages [142]. In order to provide a fair comparison to previous work, throughout our experiments we use uniform distributions to avoid providing our method with additional information not available to others.

Optimization details and overall dynamics

Recall that in Section 4.7 we proposed a two-step approach when using Gromov-Wasserstein for large tasks like this one. Since running Algorithm 6 for the full set of embeddings is infeasible (due to memory limitations), one must decide what fraction of the embeddings to use during optimization. In our experiments, we use the largest possible size allowed by memory constraints, which was found to be $K = 20,000$ for the personal computer we used. The other —more interesting— optimization choice involves the entropy regularization parameter ε. As we have discussed at various points throughout this thesis, large regularization values lead to denser optimal coupling



(A) EN →FR, $\varepsilon = 5 \cdot 10^{-4}$ (B) EN →FR, $\varepsilon = 10^{-4}$ (C) EN →RU, $\varepsilon = 10^{-4}$

Figure 4-9: **Training dynamics for the Gromov-Wasserstein approach on the word translation task**. The algorithm provably makes progress in each iteration, and the objective (red dashed line) closely follows the metric of interest (translation accuracy, not available during training). More related languages (e.g., EN →FR) lead to faster optimization, while more distant pairs yield slower learning curves (EN →RU). These results are obtained when running on 15K word vocabularies.

Γ^*, while less regularization leads to sparser solutions, at the cost of a harder (more non-convex) optimization problem.

Figure 4-8 shows a typical run of our algorithm for ℓ_∞-OT, exhibiting a common pattern: little progress at the beginning (during which $\mathbf{P}$ is being aggressively adjusted), followed by a steep decline in the objective (during which both $\mathbf{P}$ and Γ are increasingly modified in each step), after which convergence is reached. Note how the value of the optimization objective (left) and the accuracy in the translation task (right) are strongly correlated, particularly when compared against adversarial networks [42]. As mentioned previously, this is crucial because accuracy (shown here for expository purposes) *is not available* during the actual task, so model selection and early stopping are made based solely on the unsupervised objective. In addition, note that except for a small adjustment at the end of training, our method does not risk degradation by over-training, as is often the case for adversarial training alternatives.

The training dynamics for the Gromov-Wasserstein approach are similar, although convergence is faster at the beginning of training (Figure 4-9). As expected, larger values of ε lead to smoother improvements with faster runtime-per-iteration, at a price of some drop in performance. In addition, we found that computing GW distances between closer languages (such as EN and FR) leads to faster convergence than for more

			EN-ES		EN-FR		EN-DE		EN-IT		EN-RU	
	Supervision	Time	→	←	→	←	→	←	→	←	→	←
PROCRUSTES	5K words	3	77.6	77.2	74.9	75.9	68.4	67.7	73.9	73.8	47.2	58.2
PROCRUSTES + CSLS	5K words	3	81.2	82.3	81.2	82.2	73.6	71.9	76.3	75.5	51.7	63.7
MUSE	None	643	75.7	79.7	77.8	71.2	70.1	66.4	72.4	71.2	37.1	48.1
MUSE + REFINE	None	957	81.7	83.3	82.3	82.1	74.0	72.2	77.4	76.1	44.0	59.1
WASSERSTEIN + CSLS	None	–	82.8	84.1	82.6	82.9	75.4	73.3	–	–	43.7	59.1
SELF-LEARN + CSLS	None	476	82.3	84.7	82.3	83.6	75.1	74.3	79.2	79.8	48.9	65.9
ℓ_∞-INVAROT + CSLS	None	70	81.3	81.8	82.9	81.6	73.8	71.1	77.7	77.7	41.7	55.4
GROMOV-WASS ($\varepsilon = 10^{-4}$)	None	70	78.3	79.5	79.3	78.3	69.6	66.9	75.3	74.1	26.1	35.4
GROMOV-WASS ($\varepsilon = 10^{-5}$)	None	37	81.7	80.4	81.3	78.9	71.9	72.8	78.9	75.2	45.1	43.7

Table 4.1: Performance (P@1) of unsupervised and minimally-supervised methods on the MUSE translation task [42]. The time column shows the average runtime in minutes of an instance (one language pair) of the method in this task on the same (CPU) machine. We report results for our methods *without* relying on the iterative refinement step of Conneau et al. [42], so it is more appropriately compared to their MUSE + CSLS version.

distant ones (such as EN and RU, in Fig. 4-9). As with the Invariant OT approach, GW exhibits three desirable optimization properties that set both of these methods apart from other unsupervised alignment approaches, particularly adversarial-training ones: (i) the objective decreases monotonically (ii) its value closely follows the true metric of interest (translation, which naturally is not available during training) and (iii) there is no risk of degradation due to *overtraining*, as is the case for adversarial-based methods trained with stochastic gradient descent [42].

Quantitative results

We report the results on the dataset of Conneau et al. [42] in Table 4.1. The strikingly high performance of all methods on this task belies the hardness of the general problem of unsupervised cross-lingual alignment. Indeed, as pointed out by Artetxe et al. [15], the FASTTEXT embeddings provided in this task are trained on very large and highly comparable—across languages—corpora (Wikipedia), and focuses on closely related pairs of languages. Nevertheless, we carry out experiments here to have a broad evaluation of our approach in both *easier* and *harder* settings. The results in Table 4.1 show that our optimal transport-based methods perform on par with state-of-the-art approaches tailored to this task, at a fraction of the computational cost.

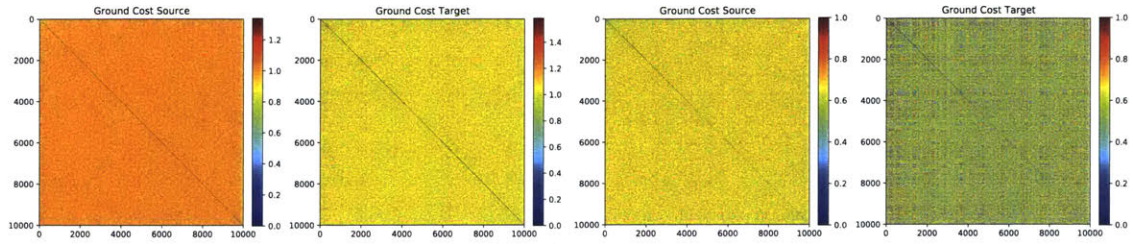


Figure 4-10: Word embeddings trained on non-comparable corpora can lead to uneven distributions of pairwise distances as shown here for the EN-FI pair of [51] (left-most two plots). Normalizing the cost matrices leads to better optimization and improved performance (right-most plots).

	EN-IT		EN-DE		EN-FI		EN-ES	
	P@1	Time	P@1	Time	P@1	Time	P@1	Time
[183]†	0	46.6	0	46.0	0.07	44.9	0.07	43.0
MUSE [42]†	45.40	46.1	47.27	45.4	1.62	44.4	36.20	45.3
SELF-LEARN [15]†	48.53	8.9	**48.47**	7.3	**33.50**	12.9	**37.60**	9.1
G-W	44.4	35.2	37.83	36.7	6.8	15.6	12.5	18.4
G-W + NORMALIZE	**49.21**	36	46.5	33.2	18.3	42.1	**37.60**	38.2

Table 4.2: Results of unsupervised methods on the dataset of Dinu et al. [51] with runtimes in minutes. Those marked with † are from [15]. Runtimes are not directly comparable since they rely on GPU computation but here we do not.

Next, we present results on the more challenging dataset of [51] in Table 4.2. Part of what makes this dataset hard is the wide discrepancy between word distance across languages, which translates into uneven distance matrices (Figure 4-10), and in turn leads to poor results for G-W. To account for this, previous work has relied on an initial whitening step on the embeddings. In our case, it suffices to normalize the pairwise similarity matrices to the same range to obtain substantially better results. While we have observed that careful choice of the regularization parameter ε can obviate the need for this step, we opt for the normalization approach since it allows us to optimize without having to tune ε. We compare our method (with and without normalization) against alternative approaches in Table 4.2. Note that we report the runtimes of Artetxe et al. [15] as-is, which are obtained by running on a Titan XP GPU, while our runtimes are, as before, obtained purely by CPU computation.

4.8.5 The Gromov-Wasserstein cross-language distance

As mentioned earlier, Theorem 4.3.2 implies that the optimal value of the Gromov-Wasserstein problem can be legitimately interpreted as a distance between languages, or more explicitly, between their word embedding spaces. This distributional notion of distance is completely determined by pair-wise geometric relations between these vectors. In Figure 4-11 we show the values $\mathrm{GW}(\mathbf{C}_s, \mathbf{C}_t, \mathbf{a}, \mathbf{b})$ computed on the FASTTEXT word embeddings of Conneau et al. [42] corresponding to the most frequent 2000 words in each language.

Overall, these distances conform to our intuitions: the cluster of romance languages exhibits some of the shortest distances, while classical Chinese (ZH) has the overall largest discrepancy with all other languages. But somewhat surprisingly, Russian is relatively close to the romance languages in this metric. We conjecture that this could be due to Russian's rich morphology (a trait shared by romance languages but not English). Furthermore, both Russian and Spanish are pro-drop languages [86] and share syntactic phenomena, such as dative subjects [133, 125] and differential object marking [31], which might explain why ES is closest to RU overall.

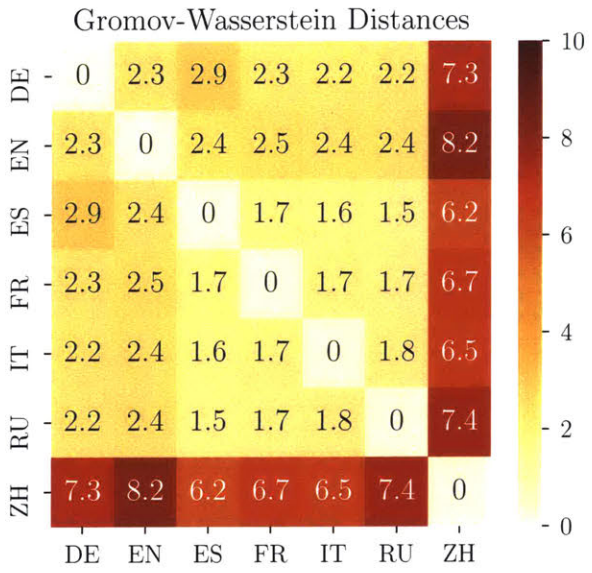


Figure 4-11: Pairwise language Gromov-Wasserstein distances obtained as the minimal transportation cost (4.2) between word embedding similarity matrices. Values scaled by 10^2 for easy visualization.

On the other hand, English appears remarkably isolated from all languages, equally distant from its Germanic (DE) and Romance (FR) cousins. Indeed, other aspects of the data (such as corpus size) might be underlying these observations.

4.9 Discussion and Extensions

In this chapter we introduced a general formulation of optimal transport that accounts for global invariances in the underlying feature spaces, unifying various existing approaches to deal with such invariances. The problem allows for very efficient algorithms in two cases often found in practice. The experiments show that this framework provides a fast, principled and robust alternative to state-of-the-art methods for unsupervised word translation, delivering comparable performance. These results suggest that OT with invariances is a viable alternative to adversarial methods that infer correspondences from complex, often underdetermined, neural network maps.

On the other hand, we showed that the Gromov-Wasserstein distance is well-suited for finding correspondences across unaligned spaces, as it performs a relational comparison of vectors across domains rather than comparing the vectors directly. In this case too, the resulting optimization objective is concise and can be optimized efficiently. The experimental results show that the resulting alignment framework is fast, stable and robust, yielding near state-of-the-art performance at a computational cost that is orders of magnitude lower than that of alternative fully unsupervised methods.

A natural question is which of the two approaches, namely optimal transport with invariances or Gromov-Wasserstein alignment, should be preferred for any given application. The answer depends on both the nature of the invariance sets and the size of the problem. As for the former, the Gromov-Wasserstein approach provides a very flexible framework with minimal assumptions on the type of invariances faced. As long as the two embedding spaces exhibit a sufficiently similar geometry (and that this geometry be faithfully captured by the metrics defined on them), the GW-alignment is likely to be successful. The invariance OT approach, on the other hand, requires specifying an invariance function class, and as expected and confirmed in our experiments, choosing the wrong class often—though, surprisingly, not always—leads to poor results.

The second aspect to consider when deciding between these two approaches is of a

computational nature. An obvious limitation of the Gromov-Wasserstein approach compared to the general invariance approach is its computational complexity, both in terms of memory and time. The word embedding task, although certainly large, is far from atypical in machine learning applications, and without a doubt smaller that many problems in computer vision. Even in this setting, we were forced to resort to an ad-hoc two-step scheme to scale up the Gromov-Wasserstein matching approach. Stochastic optimization would be an obvious approach to avoid in this secondary step. Naturally, stochastic variants of the entropy regularized OT problem would also benefit the invariant OT approaches.

In summary, if the problem of interest is small or there is no prior information or reasonable guess on the type of invariance faced, the Gromov-Wasserstein approach is an obvious choice. On the other hand, if the problem is large or the *class* of invariance—not the specific *instance* of course—is known, then the more constrained and scalable approach offered by the optimal transport generalization proposed here is a sensible choice.

Chapter 5

Optimal Transport over Hyperbolic Riemannian Manifolds

This chapter is based on Alvarez-Melis, Mroueh, and Jaakkola [11].

In this chapter, we extend the framework of unsupervised embedding alignment presented in the previous chapter to the setting of hyperbolic embeddings. The motivation for this extension is the problem of unsupervised alignment of hierarchical data such as ontologies or lexical databases. This is a problem that appears across areas, from natural language processing to bioinformatics, and is typically solved by appeal to outside knowledge bases and label-textual similarity.

Optimal transport allows us to approach this problem from a purely geometric perspective: given only a vector-space representation of the items in the two hierarchies, we seek to infer correspondences across them. In keeping with the premise of this thesis, we propose to use optimal transport to infer the correspondences, as we did in the previous chapter. However, this setting differs in the fact that the data itself has structure (namely, hierarchical structure). But here, as opposed to Chapter 3, we model the structure of the data *through the representation space itself*, rather than through the cost function as we did in that case. For this, we build upon a recent work that shows the advantage of embedding hierarchical structures in hyperbolic (rather than Euclidean) spaces [135, 67, 49]. Thus, we seek to combine the approach

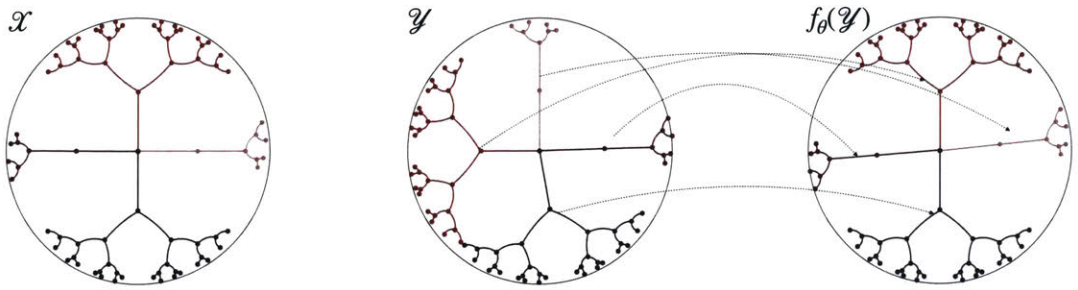


Figure 5-1: Schematic representation of the *branch permutability* phenomenon in hyperbolic embeddings. As we will see later in this chapter, besides rotational invariance hyperbolic embeddings can exhibit and additional (non-linear) type of invariance: while the geometry of individual branches of the hierarchy is approximately preserved across embedding instances, their relative positions with respect to one another might not. In Section 5.4 we propose a framework to overcome this invariance, which at a high-level reorders the branches of one of the embedding spaces ($\mathcal{Y}$ in this case) in such a way that allows for direct comparison with the other space ($\mathcal{X}$).

of Chapter 4 with this recent trend in machine learning, by extending the former to non-Euclidean settings, and using them to find correspondences between datasets by relying solely on their geometric structure, as captured by their hyperbolic-embedded representations. We will see that in this setting too there are global invariances that need to be accounted for. However, we will see that they are much more complex that the Schatten-norm invariances tackled before. Thus, in this Chapter we are simultaneously facing *explicit* structure in the data and *implicit* structure induced by invariance in the representations, interweaving ideas presented in all the previous chapters.

Our focus in this Chapter —and in fact throughout this thesis— is on the *matching* aspect of the problem, so we assume the embeddings of the hierarchies are already provided. After introducing the necessary background on hyperbolic embeddings in Section 5.2, we begin the analysis with a set of negative results. We show that state-of-the-art methods for unsupervised (Euclidean) embedding alignment, including many of those introduced in Chapter 4, perform remarkably poorly when used on hyperbolic embeddings, even after modifying them to account for this geometry. The cause of this failure lies in a type of invariance —not exhibited by Euclidean embeddings— which we refer to as *branch permutability*. At a high level, this phenomenon is characterized

by a lack of consistent ordering of branches in the representations of a dataset across different runs of the embedding algorithm (Fig 5-1), and is akin to the node order invariance in trees.

In response to this challenge, we further generalize our approach by learning a flexible nonlinear registration function between the spaces with a hyperbolic neural network [68]. This nonlinear map is complex enough to register one of the hyperbolic spaces (Fig 5-3b), and is learned by minimizing an optimal transport problem over hyperbolic space, which provides both a gradient signal for training and a pointwise (soft) matching between the embedded entities. The resulting method is capable of aligning embeddings in spite of severe branch permutability, which we demonstrate with applications in WordNet translation and biological ontology matching.

5.1 Motivation and Applications

Hierarchical structures are among the most common types of structured data in various domains, such as natural language processing and bioinformatics. For example, structured lexical databases like WordNet [129] are widely used in computational linguistics as an additional resource in various downstream tasks [131, 158, 30]. On the other hand, ontologies are often used to store and organize relational data.

Building such datasets is expensive and requires expert knowledge, so there is great interest in methods to merge, extend and extrapolate across these structures. A fundamental ingredient in all of these tasks is *matching*[1] different datasets, i.e., finding correspondences between their entities. For example, the problem of ontology alignment is an active area of research, with important implications for integrating heterogeneous resources, across domains or languages [166]. We refer the reader to Euzenat and Shvaiko [58] for a thorough survey on the state of this problem. On the other hand, there is a long line of work focusing on automatic WordNet construction that seeks to leverage existing large WordNets (usually, English) to automatically build WordNets in other low-resource languages [110, 155, 144, 98].

[1] We interchangeably use *matching* and *alignment* to refer to this task.

Euzenat and Shvaiko [58] recognize three dimensions for similarity in ontology matching: semantic, syntactic and external. A similar argument can be made for other types of hierarchical structures. Most current methods for aligning such types of data rely on a combination of these three, i.e., in addition to the relations between entities they exploit lexical similarity and external knowledge. For example, automatic WordNet construction methods often rely on access to machine translation systems [144], and state-of-the-art ontology matching systems commonly assume access to a large external knowledge base. Unsurprisingly, these methods perform poorly when no such additional resources are available [159]. Thus, effective fully-unsupervised alignment of hierarchical datasets remains largely an open problem.

5.2 Preliminaries

A fundamental question when dealing with any type of symbolic data is how to represent it. As the advent of representation learning has proven, finding the right feature representation is as—and often more—important than the algorithm used on it. Naturally, the goal of such representations is to capture relevant properties of the data. For our problem, this is particularly important. Since our goal is to find correspondences between datasets based purely on their relational structure, it is crucial that the representation capture the semantics of these relations as precisely as possible.

Traditional representation learning methods embed symbolic objects into low-dimensional Euclidean spaces. These approaches have proven very successful for embedding large-scale co-occurrence statistics, like linguistic corpora for word embeddings [127, 139]. However, recent work has shown that data for which semantics are given in the form of hierarchical structures is best represented in hyperbolic spaces, i.e., Riemannian manifolds with negative curvature [39, 136, 67]. Among the arguments in favor of these spaces is the fact that any tree can be embedded into finite hyperbolic spaces with arbitrary precision [77]. This stands in stark contrast with Euclidean spaces, for which the dependence on dimension grows exponentially. In practice, this

means that very low-dimensional hyperbolic embeddings often perform on-par or above their high-dimensional Euclidean counterparts in various downstream tasks [136, 67, 171]. This too is an appealing argument in our application, as we are interested in matching very large datasets, making computational efficiency crucial.

Working with hyperbolic geometry requires a model to represent it and operate on it. Recent computational approaches to hyperbolic embeddings have mostly focused on the Poincaré Disk (or, in higher dimensions, *Ball*) model. This model is defined by the manifold $\mathbb{D}^d = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$, equipped with the metric tensor $g_x^{\mathbb{D}} = \lambda_x^2 g^E$, where $\lambda_x := 1/(1 - \|x\|_2^2)$ is the *conformal factor* and g^E is the Euclidean metric tensor. With this, $(\mathbb{D}^d, g_x^{\mathbb{D}})$ has a Riemannian manifold structure, with the induced Riemannian distance given by:

$$d_{\mathbb{D}}(\mathbf{u}, \mathbf{v}) = \operatorname{arcosh}\left(1 + 2\frac{\|\mathbf{u} - \mathbf{v}\|^2}{(1 - \|\mathbf{u}\|^2)(1 - \|\mathbf{v}\|^2)}\right). \tag{5.1}$$

From this, the norm on the Poincaré Ball can be derived as

$$\|\mathbf{u}\|_{\mathbb{D}} = d_{\mathbb{D}}(0, \mathbf{u}) = 2\operatorname{arctanh}(\|\mathbf{u}\|). \tag{5.2}$$

It can be seen from this expression that the magnitude of points in the Poincaré Ball tends to infinity towards its boundary. This phenomenon intuitively illustrates the tree-like structure of hyperbolic space: starting from the origin, the space becomes increasingly—in fact, exponentially more—densely packed towards the boundaries, akin to how the width of a tree grows exponentially with its depth.

Hyperbolic embedding methods find representations in the Poincaré Ball by constrained optimization (i.e., by imposing $\|\mathbf{x}\| < 1$) of a loss function that is often problem-dependent. For datasets in the form of entailment relations $\mathcal{D} = \{(u, v)\}$, where $(u, v) \in \mathcal{D}$ means that u is a subconcept of v, Nickel and Kiela [136] propose to minimize the following soft-ranking loss:

$$\mathcal{L}(\Theta) = \sum_{(u,v)\in\mathcal{D}} \log \frac{e^{-d(\mathbf{u},\mathbf{v})}}{\sum_{\mathbf{v}'\in\mathcal{N}(u)} e^{-d(\mathbf{u},\mathbf{v}')}}, \tag{5.3}$$

where $\Theta = \{\mathbf{u}\}$ are the embeddings and $\mathcal{N}(u) = \{v \mid (u, v) \notin \mathcal{D}\}$ a set of negative examples for u.

Transformations in the Poincaré Ball will play a prominent role in the development of our approach in Section 5.4, so we discuss them briefly here. Since the Poincaré Ball is bounded, any meaningful operation on it must map $\mathbb{D}^d$ onto itself. Furthermore, for registration we are primarily interested in isometric transformations on the disk, i.e., we seek analogues of Euclidean vector translation, rotation, and reflection. In this model, translations are given by Möbius addition, defined as

$$\mathbf{u} \oplus \mathbf{v} \triangleq \frac{(1 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{v}\|_2^2)\mathbf{u} + (1 - \|\mathbf{u}\|_2^2)\mathbf{v}}{1 + 2\langle \mathbf{u}, \mathbf{v} \rangle + \|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2}. \tag{5.4}$$

This definition conforms to our intuition of translation, e.g., if the origin of the disk is translated to $\mathbf{v}$, then $\mathbf{x}$ is translated to $\mathbf{v} \oplus \mathbf{x}$. Note that this addition is neither commutative nor associative. More generally, it can be shown that all isometries in the Poincaré Ball have the form $T(\mathbf{x}) = \mathbf{P}(\mathbf{v} \oplus \mathbf{x})$, where $\mathbf{v} \in \mathbb{D}^d$ and $\mathbf{P} \in \mathrm{SO}(d)$, i.e., it is an orientation-preserving isometry in $\mathbb{R}^d$.

Two other fundamental concepts in Riemannian geometry are the logarithmic $\log_{\mathbf{p}}(\cdot)$ and exponential $\exp_{\mathbf{p}}(\cdot)$ maps on a Riemannian manifold $\mathcal{M}$. These are operators that map between the manifold and its tangent space $T_{\mathbf{p}}\mathcal{M}$ at a given point $\mathbf{p}$. For the Poincaré Ball, these maps can be succinctly expressed in terms of the operations defined above as follows:

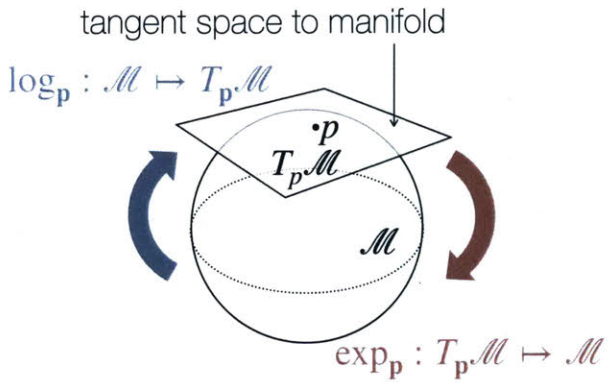


Figure 5-2: Visualization of exponential and logarithmic maps on a Riemannian manifold.

$$\exp_{\mathbf{p}}(\mathbf{u}) = \mathbf{p} \oplus \left(\tanh(\tfrac{1}{2}\lambda_{\mathbf{p}}\|\mathbf{u}\|)\frac{\mathbf{u}}{\|\mathbf{u}\|}\right), \qquad \log_{\mathbf{p}}(\mathbf{v}) = \tfrac{2}{\lambda_{\mathbf{p}}} \operatorname{arctanh}\left(\|(-\mathbf{p}) \oplus \mathbf{v}\|\right) \frac{(-\mathbf{p}) \oplus \mathbf{v}}{\|(-\mathbf{p}) \oplus \mathbf{v}\|}.$$

While this is far from a complete introduction to hyperbolic geometry, the concepts introduced so far suffice for the purposes of this thesis.

5.3 Wasserstein Matching of Hyperbolic Spaces

In Chapter 4 we studied in detail how optimal transport distances can be used to find correspondences between two embedding spaces in a fully unsupervised manner. However, all the methods we mentioned in that chapter have been applied exclusively to Euclidean settings. One might be hopeful that naive application of those approaches on hyperbolic embeddings without further modifications might simply work, but — unsurprisingly— it does not (see Table 5.4). Indeed, ignoring the special geometry of these spaces leads to poor alignment. Thus, we now investigate how to adapt such a framework to non-Euclidean settings.

The first fundamental question towards this goal that one should ask is whether optimal transport extends to more general Riemannian manifolds (i.e., beyond Euclidean space). The answer is mostly positive. Again, an in-depth treatment of this question falls outside the scope of this thesis, but we appeal to the review of guarantees for OT on Riemannian Manifolds presented in Chapter 2 (§2.6.2). We recall that for hyperbolic spaces, under mild regularity assumptions, it can be shown that: (i) OT is well-defined [174], (ii) its solution is guaranteed to exist, be unique and be induced by a transport map [124]; and (iii) this map is not guaranteed to be smooth for the usual cost $d_{\mathbb{D}}(x, y)^2$, but it is for variations of it (e.g., $-\cosh \circ d_{\mathbb{D}}$) [114]. This set of theoretical results support the use of Wasserstein distances for finding correspondences in the hyperbolic setting of interest. Furthermore, Theorem 2.6.2 provides various Riemannian cost functions with strong theoretical foundations and potential for better empirical performance.

The second questions towards generalizing Problem (4.6) to hyperbolic spaces that should be resolved involves the transformation $f \in \mathcal{F}$. First, we note that using orthogonal matrices as in the Euclidean case is still valid because, as discussed in Section 5.2, these map the unit disk into itself. Therefore, we can now solve a generalized (hyperbolic) version of the Orthogonal Procrustes problem as before. However, this approach performs surprisingly bad in practice too (see results for HyperOT+Orthogonal $\mathbf{P}$ in Table 5.4).

To understand the cause of this surprising failure, we recall that orthogonality was a natural choice of invariance for embedding spaces that we assumed might differ by a rigid transformation, but were otherwise compatible. This was a natural assumption for shapes and word vector representations generated with popular word embedding algorithms. However, Poincaré embeddings exhibit another, more complex, type of invariance, which to the best of our knowledge has not been reported before. It is a *branch permutability invariance*, whereby the relative positions of branches in the hierarchy might change abruptly across different runs of the embedding algorithm, even for the exact same data and hyperparameters.

This phenomenon is shown for a simple hierarchy embedded in the Poincaré Disk in the first row of Figure 5-3. The two embedded spaces, $\mathcal{X}$ and $\mathcal{Y}$, which were obtained with the same algorithm on the same data with different random initializations, show overall similar branch structure. For example, the branches for the *ruminant* and *canine* families (shown in green and blue in the plot) have almost identical shape. However, the reader will readily notice that the space $\mathcal{Y}$ is not just simply a rotation of $\mathcal{X}$. Instead, it is clear that the *ordering* of the branches is different in these two embeddings: while the two aforementioned branches are continuous in space $\mathcal{X}$, there are other branches in between them on both directions in space $\mathcal{Y}$.

Naturally, actual discrete trees are invariant to node ordering, but *a priori* it is not obvious why this property would be inherited by the embedded space generated with optimization objective (5.3), where non-ancestrally-related nodes do indeed interact (as negative pairs) in the objective. We conjecture that the cause of this invariance is the use of negative *sampling* for normalization in that loss function, which has the effect of putting emphasis on preserving distance between entities that are ancestrally related in the hierarchy, at the cost of down-weighting distances between unrelated entities. A formal explanation of this phenomenon is left for future work. Here, instead, we develop a framework to account for and correct these invariances while simultaneously aligning the two embeddings.

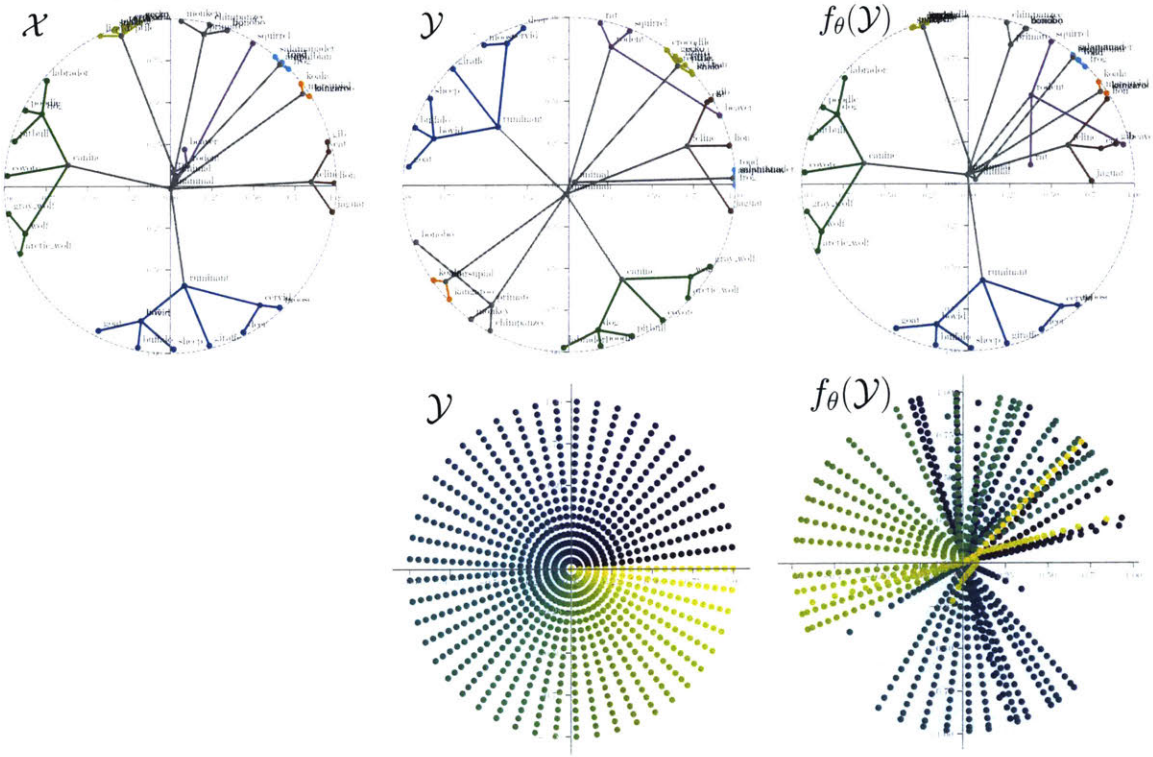


Figure 5-3: Overcoming branch invariance on a simple dataset. The two embeddings $\mathcal{X}$, $\mathcal{Y}$ were produced by the method of Nickel and Kiela [135] on the same simple hierarchy, using the same hyperparameters but different random seeds. After registration by the mapping f_θ learned by our method, the transformed embedding space $f_\theta(\mathcal{Y})$ closely resembles the target space $\mathcal{X}$, facilitating correspondence estimation. The density plots in the second row show that f_θ is highly nonlinear and approximately radial.

5.4 A Deep Invariant Correspondence Approach

The failure of the baseline Euclidean alignment methods (and their hyperbolic versions) discussed in the previous section, combined with the underlying branch permutability invariance responsible for it, make it clear that the space of registration transformations $\mathcal{F}$ in Problem (4.6) has to be generalized not only beyond orthogonality but beyond linearity too. Ideally we would want to search for f among all continuous mappings between $\mathcal{Y}$ and $\mathcal{X}$, i.e, letting $\mathcal{F} = \{f : \mathcal{Y} \to \mathcal{X} \mid f \in \mathcal{C}(\mathcal{Y})\}$. To make this search computationally tractable, we can instead approximate this function class with deep neural networks f_θ parametrized by $\theta \in \Theta$.

While an alternating minimization approach is still possible, solving a minimization for θ to completion in each iteration is undesirable. Instead, we can reverse the order of optimization and rewrite our objective as

$$\min_{f_\theta \in \mathcal{F}} \min_{\gamma \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} d(x, f(y)) d\gamma(x, f(y)) = \min_{f_\theta \in \mathcal{F}} \mathrm{W}_\varepsilon(\alpha, f_{\theta\sharp}\beta). \tag{5.5}$$

Since $\mathrm{W}^\varepsilon(\alpha, f_{\theta\sharp}\beta)$ is differentiable with respect to θ, we can use gradient-descent based methods to optimize it.

Here again, as we did in Chapter 4 for the euclidean version of this problem, we use the "normalized" Sinkhorn Divergence (2.20) instead, which we show here again for convenience:

$$\mathrm{SD}^\varepsilon(\alpha, \beta) \triangleq \mathrm{W}^\varepsilon(\alpha, \beta) - \tfrac{1}{2}\big(\mathrm{W}^\varepsilon(\alpha, \alpha) + \mathrm{W}^\varepsilon(\beta, \beta)\big).$$

Thus, replacing this loss in Problem (5.5), and taking $\mathcal{X} = \mathcal{Y} = \mathbb{D}^d$ to be the hyperbolic space of dimension d, we arrive at the final version of our optimization problem:

$$\min_{\theta : f_\theta[\mathbb{D}^d] \subseteq \mathbb{D}^d} \mathrm{SD}^\varepsilon(\alpha, f_\theta^\# \beta), \tag{5.6}$$

where we recall that $f_\theta[\cdot]$ denotes the image of a set under f_θ.

The last remaining piece of the puzzle is that we need to construct a class of neural networks that parametrizes $\mathcal{F} := \{f_\theta \mid f_\theta[\mathbb{D}^d] \subseteq \mathbb{D}^d\}$, i.e., functions that map $\mathbb{D}^d$ onto

itself. In recent work, Ganea et al. [67] propose a class of hyperbolic neural networks that do exactly this. As they point out, the basic operations in hyperbolic space that we introduced in §5.2 suffice to define analogues of various differentiable building blocks of traditional neural networks. For example, a hyperbolic linear layer can be defined as

$$f_{\text{HYPLIN}}(\mathbf{x}; \mathbf{W}, \mathbf{b}) \triangleq (\mathbf{W} \otimes \mathbf{x}) \oplus \mathbf{b} = \exp_0(\mathbf{W} \log_0(\mathbf{x})) \oplus \mathbf{b}.$$

Analogously, a layer applying a nonlinearity $\sigma(\cdot)$ in the hyperbolic sense can be defined as $\sigma_{\mathbf{D}}(\mathbf{x}) \triangleq \exp_0(\sigma \log_0(\mathbf{x}))$. Here, we also consider *Möbius Transformation layers*, $f_{\text{Möbius}}(\mathbf{x}) = \mathbf{P}(\mathbf{v} \oplus \mathbf{x})$, with $\mathbf{P} \in \text{SO}(d)$ and $\mathbf{v} \in \mathbb{D}^d$.

With these building blocks, we can parametrize highly nonlinear functions $f_\theta : \mathbb{D}^n \to \mathbb{D}^n$ as a sequence of such hyperbolic layers, e.g., $\mathbf{h}^{(i)} = \sigma_{\mathbb{D}}(\mathbf{W}^{(i)} \otimes \mathbf{h}^{(i-1)} \oplus \mathbf{b}^{(i)})$ for the hyperlinear layers, and $\mathbf{h}^{(i)} = \sigma_{\mathbb{D}}(\mathbf{P}(\mathbf{v} \oplus \mathbf{x}))$ for our Möbius layers. Note that for the hyperbolic linear layer—but crucially, not for the Möbius layer—the intermediate hidden states $\mathbf{h}^{(i)}$ need not live in the same dimensional space as the input and output, i.e., using rectangular weight matrices $\mathbf{W}$ we can map intermediate states to Poincaré balls of different dimensionality.

5.4.1 Optimization

Evaluation of the loss function (5.6) is itself an optimization problem, i.e., it requires solving regularized optimal transport. We back-propagate through this objective [71], using the `geomloss` toolbox for efficiency. For the outer-level optimization, we rely on Riemannian gradient descent [182, 178]. We found that the adaptive methods of Bécigneul and Ganea [20] worked best, particularly RADAM. It is worth noting that for the HYPERLINEAR layers only the bias term is constrained (on the Poincaré Ball), while for our MÖBIUS layers the weight matrix is constrained too (in the Stiefel manifold), so in this case we optimize over the product of the two manifolds. Additional details on optimization are provided in the experimental section (§5.5.2).

5.4.2 Avoiding poor local minima

It is easy to see that the loss function (5.6) is highly non-convex with respect to θ, a consequence of both the objective itself and the nature of hyperbolic neural networks [67]. As a result, initialization is likely to play a crucial role in solving this problem, since it is very hard to overcome a poor initial local minimum. In our experiments, we observe that even layer-wise random initialization of weights and biases proves futile. As a solution, we propose three pre-training initialization schemes, that roughly ensure (in different ways) that f does not initially "collapse" the space $\mathcal{Y}$:

- CROSSMAP. Initialize f_θ to approximately match the target points to the source points in a random permuted order: $\min_\theta \sum_{i=1}^n d_{\mathbb{D}}(\mathbf{x}_{\sigma(i)}, f_\theta(\mathbf{y}_i))$ for a some permutation $\sigma(i)$.

- IDENTITY. Initialize f_θ to approximate the identity: $\min_\theta \sum_{i=1}^n d_{\mathbb{D}}(\mathbf{y}_i, f_\theta(\mathbf{y}_i))$

- PROCRUSTES [35]. Initialize f_θ to be approximately end-to-end orthogonal: $\min_\theta \sum_{i=1}^n d_{\mathbb{D}}(f(\mathbf{y}_i), \mathbf{P}\mathbf{y}_i)$, where $\mathbf{P} = \operatorname{argmin}_{\mathbf{P}\in O(n)} \|\mathbf{X} - \mathbf{P}\mathbf{Y}\|_2^2$.

The intuition behind these three schemes is as follows. All of them seek an initial transformation that approximately matches a set of reference points, with the choice of reference being different for each of them. The first one takes as reference a random permutation of the source points, which considering the final goal is indeed to find a mapping between these two collections (albeit with potentially different pair-wise correspondences), is a sensible first approach. The second scheme instead uses the same collection $\mathbf{y}$ as reference, so the mapping that minimizes this discrepancy is the identity. Finally, the third of these schemes, inspired by work by Bunne et al. [35], seeks an initial transformation that as close as possible to the solution of the orthogonal Procrustes problem between the two collections.

Finally, we again (as in Chapter 4) use an annealing scheme on the entropy-regularization parameter ε. Starting from an aggressive regularization (large ε_0), we gradually decrease it with a fixed decay rate $\varepsilon_t = \xi \cdot \varepsilon_{t-1}$. In all our experiments we use $\xi = 0.99$.

5.5 Experiments

5.5.1 Datasets and methods

Datasets

For our first set of experiments, we extract subsets of WordNet [129] in five languages. For this, we consider only nouns and compute their transitive closure according to hypernym relations. Then, for each collection we generate embeddings in the Poincaré Ball of dimension 10 using the `PoincareEmbeddings` toolkit[2] (the official implementation of the method of Nickel and Kiela [135]) with default parameters. To generate the parallel WordNet datasets, we use the `nltk` interface to WordNet, and proceed as follows. In the English WordNet, we first filter out all words except nouns, and generate their transitive closure. For each of the remaining synsets, we query for lemmas in each of the four other languages (ES, FR, IT, CA), for which `nltk` provides multilingual support in WordNet. These tuples of lemmas form our ground-truth translations, which are eventually split into a validation set of size 5000, leaving all the other pairs for test data (approximately 1500 for each language pairs). Note that the validation is for visualization purposes only, and all model selection is done in a purely unsupervised way based on the training objective. After the multilingual synset vocabularies have been extracted, we ensure their transitive closures are complete and write all the relations in these closures to a file, which will be used as an input to the `PoincareEmbeddings` toolkit.

For the second set of experiments, we consider two subtasks of the OAEI 2018 ontology matching challenge [2]: ANATOMY, which consists of two ontologies; and BIODIV, consisting of four. Further details about the OAEI datasets can found on the project's website.[3]. Again, we use the `PoincareEmbeddings` codebase to embed them in 10-dimensional space.

Additional details on all the datasets are provided in Table 5.1.

[2] `https://github.com/facebookresearch/poincare-embeddings`
[3] `http://oaei.ontologymatching.org/2018/`

	WordNet				Anatomy		Biodiv			
	EN	ES	FR	CA	Human	Mouse	FLOPO	PTO	ENVO	SWEET
Entities	8206	8206	8206	8206	3298	2737	360	1456	6461	4365
Relations	47938	47938	47938	47938	18556	7364	472	11283	73881	30101
Embedding Method	[136]	[136]	[136]	[136]	[136]	[136]	[136]	[136]	[136]	[136]
Embedding Size	10	10	10	10	10	10	10	10	10	10

Table 5.1: Dataset statistics. EN: English, ES: Spanish, FR: French, CA: Catalan. HUMAN: NCI Thesaurus of human anatomy, MOUSE: Adult Mouse Anatomy, FLOPO: Flora Phenotype Ontology, PTO: Plant Trait Ontology, ENVO: the Environment Ontology, SWEET: the Semantic Web for Earth and Environment Technology Ontology.

Methods

We first compare ablated versions of our HYPERBOLIC-OT model on a monolingual (EN) WordNet self-recovery experiment. The configuration details for these ablated models are shown in Table 5.3, where dashed lines indicate a parameter being the same as in the FULL MODEL. Then, we compare against three off-the-shelf state-of-the-art unsupervised word embedding alignment models: MUSE [42], SELF-LEARN [15] and INVAROT, our approach from Chapter 4, all run with default settings.

5.5.2 Optimization details

Each forward pass of the loss function (5.5) requires solving three regularized OT problems. As mentioned several times in this thesis, this can be done to completion with the Sinkhorn-Knopp algorithm in $O(N^2 \log N\varepsilon^{-3})$ time [3], although practical implementations often run a fixed number of iterations. We rely on the `geomloss`[4] package for efficient differentiation through Sinkhorn and on the `geoopt`[5] package for Riemannian optimization.

We run our method for a fixed number of outer iterations (200 in all our experiments), which given the decay strategy on the entropic regularization, ensures that ε ranges from 1×10^1 to 1×10^{-2}. All experiments were run on a single machine with a 32-core Intel Xeon CPU @3.20 GHz, leveraging GPU computation whenever possible. The total runtime of our method on these experiments ranges from 1 to 20 minutes.

[4] `https://www.kernel-operations.io/geomloss/`
[5] `https://geoopt.readthedocs.io/en/latest/`

Model	Metric	Cost	Pre-train	ε-anneal	Depth	Hid. dim	Layers	$\sigma(\cdot)$	Opt	LR
FULL	Poinc.	$d_{\mathbb{D}}$	XMAP	$10^{-1}\rightarrow^{-2}10$		20	HYPLIN	elu	RADAM	10^{-3}
SMALL	–	–	–	–	2	10	–	–	–	–
EUCLIDEAN	Eucl.	–	–	–	–	–	–	–	ADAM	–
RELU	–	–	–	–	–	–	–	relu		–
RSGD	–	–	–	–	–	–	–	–	RSGD	10^{-2}
MÖBIUS	–	–	–	–	–	10	MÖBIUS	–	–	–
cosh COST	–	$-\cosh \circ d_{\mathbb{D}}$		–	–	–	–	–	–	–
NO-PRE	–	–	None	–	–	–	–	–	–	–

Table 5.2: Ablated model configurations for the monolingual EN→EN WordNet task.

5.5.3 Evaluation metrics

All the baseline methods return transformed embeddings. Using these, we retrieve nearest neighbors, and following the literature, we report precision at different levels, i.e., Prec@$k = \alpha$ if the true match is within the top k retrieved candidate matches for α percent of the test examples.

5.5.4 Multilingual wordnet alignment

We first investigate the impact on performance of the various components of our model in a controlled setting, where the correspondences between the two datasets are perfect and unambiguous. For this, we embed the same hierarchy (the EN part of our WordNet dataset) twice, using the same algorithm with the same hyperparameters, but different random seeds. We then evaluate the extent to which our method can recover the correspondences. Starting from our FULL MODEL, we remove and/or replace various components and evaluate the performance again. The exact configuration of the ablated models is shown in Table 5.2.

	P@1	P@10
FULL MODEL	22.2	88.8
small	8.7	38.0
euclidean	3.1	13.0
elu→relu	6.9	37.6
RADAM→RSGD	14.7	69.5
MÖBIUS layers	11.9	54.3
cost: $-\cosh \circ d$	16.6	70.2
no pretrain	0.1	0.2

Table 5.3: Ablation on EN→EN WordNet.

	Time	EN-ES		EN-IT		EN-FR		EN-CA	
	(min)	→	←	→	←	→	←	→	←
Baseline Euclidean Methods									
MUSE [42]	42	0.60	1.20	0.06	0.45	0.06	2.22	0.09	0.87
SELF-LEARN [15]	3	0.31	1.40	0.46	0.46	0.55	0.40	0.25	0.34
INVAR-OT (Chapter §4)	4	0.25	0.25	2.15	0.60	0.38	2.14	0.51	7.65
Proposed Hyperbolic Methods									
HYPER-OT+$\mathbf{P} \in \mathrm{O}(d)$	13	4.51	5.29	11.2	0.47	5.32	4.68	7.21	5.55
HYPER-OT+NN f_θ	21	43.9	56.8	48.0	60.1	54.3	57.5	38.4	57.4

Table 5.4: Results on the multilingual wordnet matching task. The numbers indicate precision@10 for pair-wise language matching in both directions. All baseline models use Euclidean metrics to compare embeddings.

The results in Table 5.3 suggest that the most crucial components are the use of the appropriate Poincaré metric and the pretraining step. The moderate performance of MÖBIUS is likely due to the constraint on dimensionality that this type of layer has (§5.4). We next consider a realistic task of inferring correspondences across the multilingual WordNet embeddings. Naturally, in this case there might not be perfect correspondences across the entities in different languages. As before, we report Precision@10 and compare against baseline models in Table 5.4.

5.5.5 Ontology matching

Finally, we test our method on the OAEI tasks. The results (Table 5.5) show that again our method decidedly outperforms the two off-the-shelf Euclidean methods, but now the overall performance of all methods is remarkably lower, which suggests the correspondences between these domains are more subtle and/or noisy that those in the WordNet task.

	Anatomy		Biodiv			
	H→M	M→H	F→P	P→F	S→E	S→E
MUSE	0.12	0.00	3.23	0.00	0.00	0.00
SELF-LEARN	0.00	0.00	4.00	0.00	0.01	0.02
HYP-OT	7.89	4.22	10.12	8.73	3.45	9.66

Table 5.5: Ontology matching results.

5.6 Discussion and Extensions

The framework for hierarchical structure matching proposed in this chapter admits various extensions, some of which are immediate. We focused on the particular case of the Poincaré Ball, but since most of the components of our approach —the optimization, nonlinear registration, optimal transport— generalize to other Riemannian manifolds, our framework would too. As long as optimizing over a given manifold is tractable, our framework would enable computing correspondences across instances of it.

On the other hand, we purposely adopted the challenging setting where no additional information is assumed. This setting is relevant both for extreme practical cases and to stress-test the limits of unsupervised learning in this context. However, our method would likely benefit from incorporating any additional available information as state-of-the-art methods for ontology matching do. In our framework, this information could for example be injected intro the transport cost objective.

Chapter 6

Conclusion

This thesis presents a collection of extensions of the optimal transport toolbox that account for various types of additional structure that the original formulation of the problem does not. We showed how this structure arises naturally in various machine learning applications; how it can be modeled within the framework of optimal transport; and how the resulting problems lead to better empirical solutions over various other alternative approaches, including the (structure-oblivious) original formulations. In the rest of this concluding chapter we discuss extensions and limitations of the methods proposed in this thesis.

There are many possible extensions of the methods developed here, some of them immediate, some of them requiring overcoming various challenges. As pointed out in the discussion of Chapter 3, our approach relied on submodularity because of its tractability and well-understood properties, which made it a very appealing toolbox to model structure. However, the framework that was eventually developed is flexible enough that most of it carries beyond submodularity; any non-linear convex function that suitably models some structure of interest and has bounded closed gradient maps would work as a replacement for the role of the soft matching cost function that the Lovász extension played in the one proposed here. On the other hand, in Chapter 4 we focused on invariances defined by transformations with bounded Schatten norm, but, as discussed at the time, any other invariance could be dealt with in a similar manner with appropriate constrained optimization routines. As for Chapter 5,

a natural extension would be to also optimize over the embedding representation too—i.e., not treat it as an immutable given—to actively encourage them to facilitate correspondence inference.

A limitation that is common to the three families of methods presented in this thesis is computational complexity. Of course, a type of *no free lunch theorem* applies here: methods that model structure will almost certainly be more costly than those that do not. In the context of optimal transport, any generalization that adds complexity to the classic formulation—be it to the cost function, the ground spaces, or the marginal constraints—is likely to result in more challenging optimization. The experimental results presented throughout this thesis show that this price is almost always offset by significant gains in performance on the task of interest. Needless to say, this by no means implies that there is nothing to do in this regard. This thesis had an emphasis on the modeling and application sides of the problem, not the computational efficiency aspects; as a consequence, there is plenty of room for improvement in this area.

Given that solving instances of the usual entropy-regularized optimal transport problem is at the backbone of all our methods, any improvement in computational efficiency of the methods to solve this problem would immediately translate into better efficiency of ours. There is plenty of recent work proposing fast algorithms for OT, many of them relying on stochastic optimization [70, 1, 167] and *sliced* approximations [29, 102, 50, 145, 101], which could be used to speed and scale up our regularized OT subroutines.

The last extension we discuss here, and undoubtedly the most challenging one, pertains to the structure *modeling* itself. When laying out a roadmap on approaches to injecting structure into the optimal transport problem (§1.3), we mentioned three generic components of the problem where this could be done: the cost function, the representation spaces, and the constraints. In Chapters 3 and 4 respectively we investigated the first two in-depth. The third one remains open.

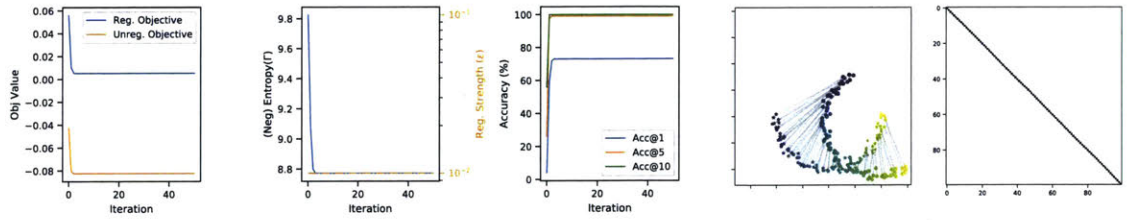
Modeling structure through the marginal distributions is appealing because it holds the promise of a more natural—and powerful—formulation of optimal transport in structured domains. Consider the setting where the objects of interest have explicit

structure, such as sentences or trees. While the structure of an instance (say, sentence) can be simulated through the cost objective (e.g., as we do in Chapter 3), without modifying the marginal distributions to actually be defined over sentences, we will not be able to fully leverage the OT toolkit (e.g., doing meaningful displacement interpolation), as the spaces themselves will not distinguish between low-probability (e.g., ungrammatical sentences) and high probability configurations. Therefore, in order to fully leverage the toolkit of optimal transport (e.g., barycentric mappings and interpolation) in a manner that is consistent with the underlying structure, it will likely be necessary to rely on modeling structure *probabilistically* through the marginal distributions instead. In Appendix B we outline some preliminary ideas on how this problem could be tackled.

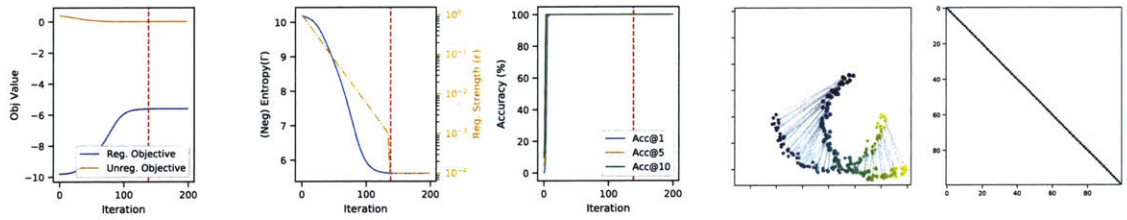
Finally, a general parting thought befitting a thesis on an interdisciplinary topic must be offered. The author, who takes unparalleled delight in ideas that make unexpected connections between seemingly disparate topics or fields, is optimistic that fascinating challenges and opportunities for future work at the intersection of optimization, representation geometry and statistics abound, and will continue to do so for the foreseeable future. Natural language processing is the perfect *playground* on which to tackle such challenges because it is intimately related to all three of these domains. While certainly not the only application domain for which this is true, it stands out for the *style of play* it offers, being a discipline where the power of computation and the elegance of statistics; the sometimes-contradicting sometimes-harmonious natures of discrete and continuous optimization; and the beauty and mysteries of mathematics and human language are all exquisitely intertwined.

Appendix A

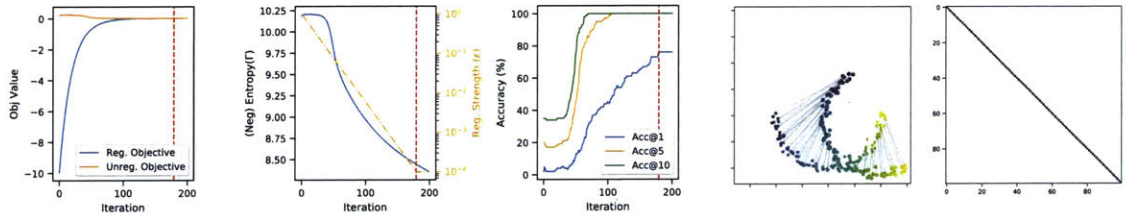
Additional Experimental Results for Invariant OT



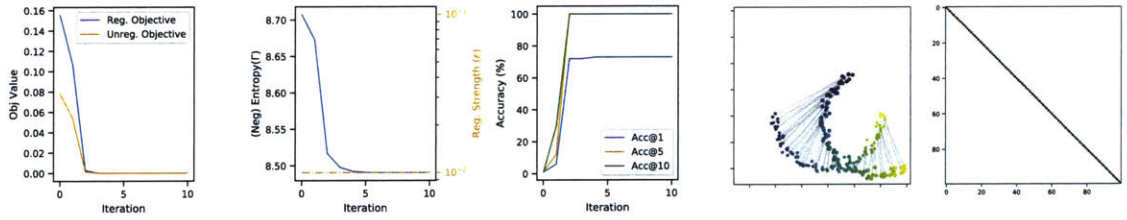
(A) (Entropic) Gromov-Wasserstein alignment.



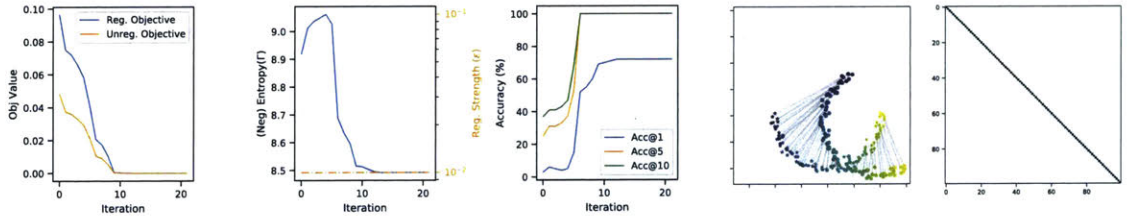
(B) Alternating Minimization on Γ and $\mathbf{P}$



(C) Single-block descent on Γ via Projected Gradient Descent on $\Pi(\mathbf{a}, \mathbf{b})$

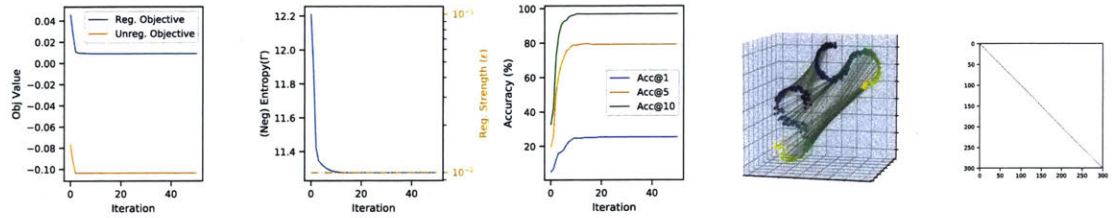


(D) Single-block descent on $\mathbf{P}$ via Riemannian conjugate gradient on $\mathrm{O}(n)$

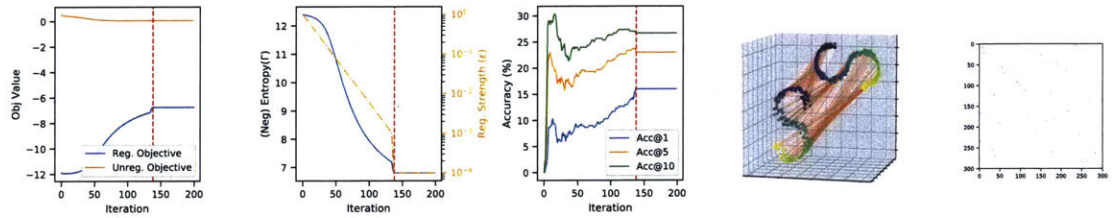


(E) Single-block descent on $\mathbf{P}$ via unconstrained conjugate gradient on $\mathbb{R}^d$

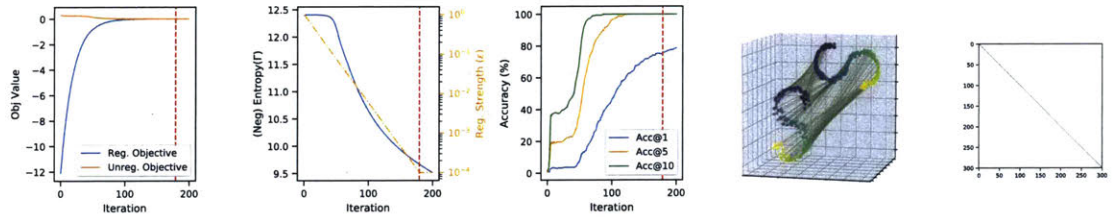
Figure A-1: Training dynamics for the various invariant OT approaches on a simple noiseless 2D moons point cloud dataset with underlying $\mathcal{F}_\infty$ invariance. Shown here is the best-of-ten restart for each model. The first three panes show objective values, entropy regularization and matching accuracy; the right-most two show the optimal coupling represented as pairwise matches and the transportation coupling Γ^*. Vertical red dashed lines indicate entropy decay was frozen at that iteration.



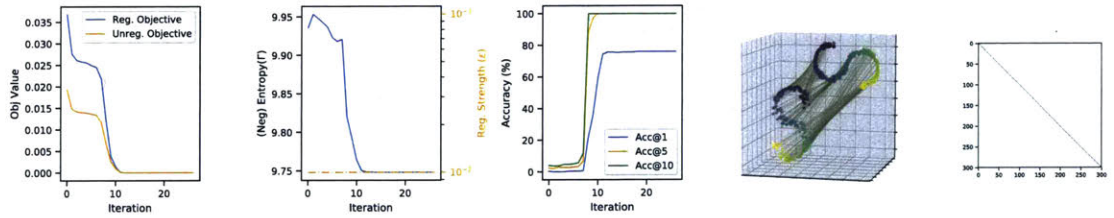
(A) (Entropic) Gromov-Wasserstein alignment



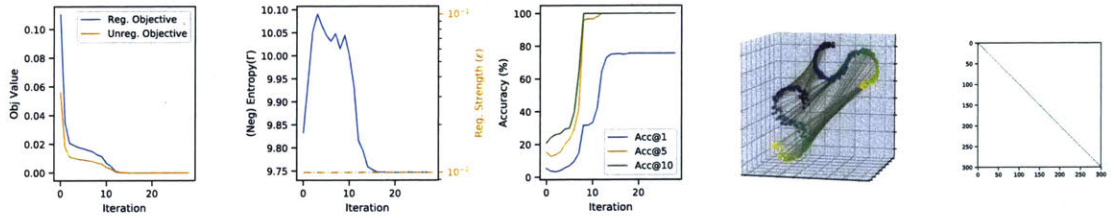
(B) Alternating Minimization on Γ and $\mathbf{P}$



(C) Single-block descent on Γ via Projected Gradient Descent on $\Pi(\mathbf{a}, \mathbf{b})$



(D) Single-block descent on $\mathbf{P}$ via Riemannian conjugate gradient on $\mathrm{O}(n)$



(E) Single-block descent on $\mathbf{P}$ via unconstrained conjugate gradient on $\mathbb{R}^d$

Figure A-2: Training dynamics for the various invariant OT approaches on a simple noiseless 3D s-shaped point cloud dataset with underlying $\mathcal{F}_\infty$ invariance. Shown here is the best-of-ten restart for each model. The first three panes show objective values, entropy regularization and matching accuracy; the right-most two show the optimal coupling represented as pairwise matches and the transportation coupling Γ^*. Vertical red dashed lines indicate entropy decay was frozen at that iteration.

Appendix B

Towards Optimal Transport with Structured Marginals

For distributions defined over structured objects, a fully rigorous treatment of couplings between them should model the marginal distributions α and β in the transportation problem as such. That is, the spaces $\mathcal{X}$ and $\mathcal{Y}$ should themselves be spaces of structured objects. In the case of sentences, this would translate to operating on distributions $\alpha \in P(\mathcal{X})$ and $\beta \in P(\mathcal{Y})$ over sentences, a departure from most current optimal-transport based approaches for language, which still operate on distributions over words.

However, naively operating on such structured distributions might cause a combinatorial explosion in the complexity of the problem. For example, for sentences of length k defined over a vocabulary of n words, the support of the marginal distributions q and q would be in the order of n^k. Consequently, the coupling between them would have size n^{2k}, making any direct application of OT prohibitive. Instead, we propose to enforce structure in the marginals through modular, hierarchical constraints.

As a concrete example, suppose $\mathcal{X}$ and $\mathcal{Y}$ are spaces of sequences and that we have a notion of distance between elements of these two spaces (e.g., based on a sequence kernel), $d(x, y)$. Furthermore, assume the distributions over them are first-order Markov models. That is, given probability measures $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and $\beta \in \mathcal{M}_+^1(\mathcal{Y})$,

we assume they admit decompositions

$$\alpha(x) = p(x_n \mid x_{n-1})p(x_{n-1} \mid x_{n-2}) \dots p(x_1) \tag{B.1}$$

$$\beta(x) = q(x_n \mid x_{n-1})q(x_{n-1} \mid x_{n-2}) \dots q(x_1) \tag{B.2}$$

The mass conservation constraints can in principle still be expressed as before:

$$\mathcal{U}(\alpha, \beta) = \left\{\gamma \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \,:\, P_{\mathcal{X}\sharp}\gamma = \alpha,\ P_{\mathcal{Y}\sharp}\gamma = \beta\right\} \tag{B.3}$$

but now the pushforward constraints $P_{\mathcal{X}\sharp}\gamma = \alpha$ have combinatorial complexity, and might be prohibitive to compute or even express. Instead, we can leverage (B.1) and (B.2) to define a set of n marginal constraints on the conditional probabilities, implicitly imposing on the coupling γ an analogous block structure decomposition.

The resulting decomposable description of the constraints would not only lead to a simplified definition of the transportation problem in the structured case, but to tractable algorithms to solve it too. Furthermore, it would be necessary to analyze the implications of such structure for the dual problem, where such decompositions often lead (after suitable relaxations) to a decoupling of variables and therefore efficient algorithms. For this, we might be to take inspiration from dual decomposition methods [165, 151] or other decomposition methods from the information geometry literature [12].

Bibliography

[1] B. K. Abid and R. Gower. "Greedy stochastic algorithms for entropy-regularized optimal transport problems". In: *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. Playa Blanca, Lanzarote, Canary Islands: PMLR, 2018, pp. 1505–1512.

[2] A. Algergawy et al. "Results of the Ontology Alignment Evaluation Initiative 2018". In: *CEUR Workshop Proceedings*. CEUR-WS: Workshop proceedings. 2018, pp. 76–116.

[3] J. Altschuler, J. Weed, and P. Rigollet. "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems*. 2017, pp. 1961–1971. arXiv: 1705.09634.

[4] D. Alvarez-Melis and T. S. Jaakkola. "Gromov-Wasserstein Alignment of Word Embedding Spaces". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 1881–1890.

[5] D. Alvarez-Melis and T. S. Jaakkola. "On the Robustness of Interpretability Methods". In: *ICML Workshop on Human Interpretability in Machine Learning*. 2018. arXiv: 1806.08049.

[6] D. Alvarez-Melis and T. S. Jaakkola. "Towards Robust Interpretability with Self-explaining Neural Networks". In: *Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc, 2018, pp. 7775–7784. arXiv: 1806.07538.

[7] D. Alvarez-Melis and T. S. Jaakkola. "Tree-structured decoding with doubly-recurrent neural networks". In: *International Conference on Learning Representations*. Ed. by S. B. Garnett, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. 2017.

[8] D. Alvarez-Melis and T. S. Jaakkola. "A causal framework for explaining the predictions of black-box sequence-to-sequence models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 412–421.

[9] D. Alvarez-Melis, T. S. Jaakkola, and S. Jegelka. "Structured Optimal Transport". In: *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. PMLR, 2018, pp. 1771–1780.

[10] D. Alvarez-Melis, S. Jegelka, and T. S. Jaakkola. "Towards Optimal Transport with Global Invariances". In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. PMLR, 2019, pp. 1870–1879.

[11] D. Alvarez-Melis, Y. Mroueh, and T. S. Jaakkola. "Unsupervised Hierarchy Matching with Optimal Transport over Hyperbolic spaces". In: *In Submission*. 2019.

[12] S.-I. Amari. "Information geometry on hierarchy of probability distributions". In: *IEEE Transactions on Information Theory* 47.5 (2001), pp. 1701–1711.

[13] L. Ambrosio and N. Gigli. "A User's Guide to Optimal Transport". In: *Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009, Editors: Benedetto Piccoli, Michel Rascle*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–155.

[14] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. PMLR, 2017, pp. 214–223.

[15] M. Artetxe, G. Labaka, and E. Agirre. "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018, pp. 789–798.

[16] M. Artetxe, G. Labaka, and E. Agirre. "Learning bilingual word embeddings with (almost) no bilingual data". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2017, pp. 451–462.

[17] M. Artetxe, G. Labaka, and E. Agirre. "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* 2016, pp. 2289–2294.

[18] M. Artetxe, G. Labaka, E. Agirre, and K. Cho. "Unsupervised Neural Machine Translation". In: *International Conference on Learning Representations.* 2018.

[19] F. Bach. "Learning with submodular functions: A convex optimization perspective". In: *Foundations and Trends in Machine Learning* 6.2-3 (2013), pp. 145–373.

[20] G. Bécigneul and O.-E. Ganea. "Riemannian Adaptive Optimization Methods". In: *International Conference on Learning Representations.* 2019.

[21] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. "The cramer distance as a solution to biased wasserstein gradients". In: *arXiv preprint arXiv:1705.10743* (2017).

[22] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust optimization.* Princeton University Press, 2009, p. 542. arXiv: 1011.1669.

[23] N. Bennacer, C. N. Jipmo, A. Penta, and G. Quercini. "Matching user profiles across social networks". In: *International Conference on Advanced Information Systems Engineering.* Springer. 2014, pp. 424–438.

[24] I. Berkes and W. Philipp. "An almost sure invariance principle for the empirical distribution function of mixing random variables". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* (1977).

[25] D. Bertsimas, D. B. Brown, and C. Caramanis. "Theory and Applications of Robust Optimization". In: *SIAM Review* 53.3 (2011), pp. 464–501. arXiv: 1010.5445.

[26] P. J. Besl and N. D. McKay. "A Method for Registration of 3-D Shapes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1992).

[27] M. Blondel, V. Seguy, and A. Rolet. "Smooth and Sparse Optimal Transport". In: *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics.* Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. Playa Blanca, Lanzarote, Canary Islands: PMLR, 2018, pp. 880–889.

[28] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[29] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. "Sliced and Radon Wasserstein Barycenters of Measures". In: *Journal of Mathematical Imaging and Vision* (2014).

[30] A. Bordes, X. Glorot, J. Weston, and Y. Bengio. "Joint learning of words and meaning representations for open-text semantic parsing". In: *Artificial Intelligence and Statistics*. 2012, pp. 127–135.

[31] G. Bossong. "Differential object marking in Romance and beyond". In: *New analyses in Romance linguistics* (1991), pp. 143–170.

[32] O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf. *From optimal transport to generative modeling: the VEGAN cookbook*. Tech. rep. 2017. arXiv: 1705.07642.

[33] Y. Brenier. "Décomposition polaire et réarrangement monotone des champs de vecteurs". In: *CR Acad. Sci. Paris Sér. I Math* 305.19 (1987), pp. 805–808.

[34] Y. Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* (1991).

[35] C. Bunne, D. Alvarez-Melis, S. Jegelka, and A. Krause. "Learning Generative Models Across Incomparable Spaces". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. PMLR, 2019, pp. 851–861.

[36] L. A. Caffarelli. "The regularity of mappings with a convex potential". In: *Journal of the American Mathematical Society* 5.1 (1992), pp. 99–104.

[37] L. A. Caffarelli. "Boundary regularity of maps with convex potentials". In: *Communications on Pure and Applied Mathematics* (1992).

[38] D. Chakrabarty, Y. T. Lee, A. Sidford, and S. C.-w. Wong. "Subquadratic submodular function minimization". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2017, pp. 1220–1231.

[39] B. P. Chamberlain, J. Clough, and M. P. Deisenroth. "Neural embeddings of graphs in hyperbolic space". In: *arXiv preprint arXiv:1705.10359* (2017).

[40] Y. Chen and G. Medioni. "Object modelling by registration of multiple range images". In: *Image and Vision Computing* (1992).

[41] S. Cohen and L. Guibas. "The Earth Mover's Distance under transformation sets". In: *Proceedings of the 7th IEEE International Conference on Computer Vision*. 1999.

[42] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. "Word Translation Without Parallel Data". In: *International Conference on Learning Representations*. 2018, pp. 1–13. arXiv: 1710.04087.

[43] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. "Joint Distribution Optimal Transportation for Domain Adaptation". In: *Neural Information Processing Systems*. 2017. arXiv: 1705.08848.

[44] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. "Optimal Transport for Domain Adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865. arXiv: 1507.00504.

[45] M. Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2292–2300. arXiv: 1306.0895.

[46] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. "On distances between phylogenetic trees". In: *SODA*. Vol. 97. 1997, pp. 427–436.

[47] H. C. Daumé III. "Practical Structured Learning Techniques for Natural Language Processing". PhD Thesis. University of Southern Californoa, 2006.

[48] H. Daumé, J. Langford, and D. Marcu. "Search-based structured prediction". In: *Machine learning* 75.3 (2009), pp. 297–325.

[49] C. De Sa, A. Gu, C. Ré, and F. Sala. "Representation tradeoffs for hyperbolic embeddings". In: *Proceedings of the 35th International Conference on Machine Learning* 80 (2018). Ed. by J. Dy and A. Krause, pp. 4460–4469.

[50] I. Deshpande, Z. Zhang, and A. Schwing. "Generative Modeling Using the Sliced Wasserstein Distance". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2018.

[51] G. Dinu, A. Lazaridou, and M. Baroni. "Improving zero-shot learning by mitigating the hubness problem". In: *International Conference on Learning Representations (Workshop Track)* (2015).

[52] J. Djolonga and A. Krause. "Scalable Variational Inference in Log-supermodular Models". In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. PMLR, 2015, pp. 1804–1813. arXiv: 1502.06531.

[53] J. Eckstein and W. Yao. "Approximate ADMM algorithms derived from Lagrangian splitting". In: *Computational Optimization and Applications* 68.2 (2017), pp. 363–405.

[54] J. Edmonds. "Submodular functions, matroids, and certain polyhedra". In: *Combinatorial Structures and Their Applications* (1970), pp. 69–87.

[55] A. Ene and H. L. Nguyen. "Random Coordinate Descent Methods for Minimizing Decomposable Submodular Functions". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 2015, pp. 787–795. arXiv: 1502.02643.

[56] A. Ene, H. L. Nguyen, and L. A. Végh. "Decomposable Submodular Function Minimization: Discrete and Continuous". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 2870–2880. arXiv: 1703.01830.

[57] M. Essid and J. Solomon. "Quadratically regularized Optimal Transport on Graphs". In: *SIAM Journal on Scientific Computing* 40 (2018), A1961–A1986. arXiv: 1704.08200.

[58] J. Euzenat and P. Shvaiko. *Ontology matching*. 2nd. Heidelberg (DE): Springer-Verlag, 2013.

[59] M. Faruqui and C. Dyer. "Improving vector space word representations using multilingual correlation". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 2014, pp. 462–471.

[60] M. Feldman, J. (Naor, and R. Schwartz. "A Unified Continuous Greedy Algorithm for Submodular Maximization". In: *IEEE Symposium on Foundations of Computer Science (FOCS)*. 2011.

[61] P. F. Felzenszwalb and D. P. Huttenlocher. "Efficient graph-based image segmentation". In: *International Journal of Computer Vision* 59.2 (2004), pp. 167–181.

[62] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré. "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *Artificial Intelligence and Statistics*. Vol. 89. PMLR, 2019, pp. 2681–2690.

[63] R. Flamary, C. Févotte, N. Courty, and V. Emiya. "Optimal spectral transportation with application to music transcription". In: *Advances in Neural Information Processing Systems*. 2016.

[64] R. Flamary and N. Courty. *POT: Python Optimal Transport library*. Tech. rep. 2017.

[65] S. Fujishige, T. Hayashi, and S. Isotani. "The Minimum-Norm-Point Algorithm Applied to Submodular Function Minimization and Linear Programming". In: *RIMS preprint 1571* (2006), pp. 1–19.

[66] P. Fung. "Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus". In: *Third Workshop on Very Large Corpora*. 1995.

[67] O.-E. Ganea, G. Bécigneul, and T. Hofmann. "Hyperbolic Entailment Cones for Learning Hierarchical Embeddings". In: *International Conference on Machine Learning* (2018).

[68] O. Ganea, G. Bécigneul, and T. Hofmann. "Hyperbolic neural networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 5345–5355.

[69] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. "Sample complexity of Sinkhorn divergences". In: *Artificial Intelligence and Statistics* 89 (2019). Ed. by K. Chaudhuri and M. Sugiyama, pp. 1574–1583.

[70] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. "Stochastic optimization for large-scale optimal transport". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3432–3440.

[71] A. Genevay, G. Peyre, and M. Cuturi. "Learning Generative Models with Sinkhorn Divergences". In: *International Conference on Artificial Intelligence and Statistics*. Vol. 84. PMLR, 2018, pp. 1608–1617.

[72] G. Goel, C. Karande, P. Tripathi, and L. Wang. "Approximability of combinatorial problems with multi-agent submodular cost functions". In: *Foundations of Computer Science, 2009*. IEEE. 2009, pp. 755–764.

[73] C. Goodall. "Procrustes Methods in the Statistical Analysis of Shape". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 53.2 (1991), pp. 285–339.

[74] J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*. Vol. 30. Oxford University Press on Demand, 2004.

[75] K. Grauman and T. Darrell. "The pyramid match kernel: Discriminative classification with sets of image features". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2005.

[76] E. Grave, A. Joulin, and Q. Berthet. "Unsupervised Alignment of Embeddings with Wasserstein Procrustes". In: *Artificial Intelligence and Statistics* 89 (2019). Ed. by K. Chaudhuri and M. Sugiyama, pp. 1880–1890. arXiv: 1805.11222.

[77] M. Gromov. "Hyperbolic groups". In: *Essays in group theory*. Springer, 1987, pp. 75–263.

[78] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*. Vol. 2. Springer Science & Business Media, 2012.

[79] J. Guo, W. Che, D. Yarowsky, H. Wang, and T. Liu. "Cross-lingual dependency parsing based on distributed representations". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1. 2015, pp. 1234–1244.

[80] S. Gupta, M. X. Goemans, and P. Jaillet. "Solving Combinatorial Games using Products, Projections and Lexicographically Optimal Bases". In: *CoRR* abs/1603.0 (2016).

[81] S. Haker and A. Tannenbaum. "Optimal mass transport and image registration". In: *Proceedings - IEEE Workshop on Variational and Level Set Methods in Computer Vision, VLSM 2001*. 2001, pp. 29–36.

[82] W. L. Hamilton, R. Ying, and J. Leskovec. "Representation learning on graphs: Methods and applications". In: *IEEE Data Engineering Bulletin* (2017).

[83] M. Hardt. "Understanding alternating minimization for matrix completion". In: *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE. 2014, pp. 651–660.

[84] T. B. Hashimoto, D. Alvarez-Melis, and T. S. Jaakkola. "Word Embeddings as Metric Recovery in Semantic Spaces". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 273–286.

[85] T. B. Hashimoto, D. Alvarez-Melis, and T. S. Jaakkola. "Word, graph and manifold embedding from Markov processes". In: *NIPS Workshop on Nonparametric Methods for Large Scale Representation Learning*. 2015, pp. 1–6. arXiv: 1509.05808.

[86] M. Haspelmath. "The European linguistic area: Standard Average European". In: *Language typology and language universals: An international handbook*. Vol. 2. de Gruyter, 2001, pp. 1492–1510.

[87] D. S. Hochbaum and S.-P. Hong. "About strongly polynomial time algorithms for quadratic optimization over submodular constraints". In: *Mathematical Programming* (1995), pp. 269–309.

[88] A. Hyvärinen and E. Oja. "Independent component analysis: algorithms and applications". In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[89] S. Iwata and N. Zuiki. "A network flow approach to cost allocation for rooted trees". In: *Networks* 44 (2004), pp. 297–301.

[90] M. Jaggi. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization". In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. PMLR, 2013, pp. 427–435.

[91] P. Jain, P. Netrapalli, and S. Sanghavi. "Low-rank matrix completion using alternating minimization". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM. 2013, pp. 665–674.

[92] S. Jegelka, F. Bach, and S. Sra. "Reflection methods for user-friendly submodular optimization". In: *Advances in Neural Information Processing Systems*. 2013, pp. 1313–1321. arXiv: 1311.4296v1.

[93] S. Jegelka and J. Bilmes. "Submodularity beyond submodular energies: Coupling edges in graph cuts". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2011, pp. 1897–1904.

[94] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. "An introduction to variational methods for graphical models". In: *Machine learning* 37.2 (1999), pp. 183–233.

[95] A. Juditsky and A. Nemirovski. "First order methods for nonsmooth convex large-scale optimization, I: general purpose methods". In: *Optimization For Machine Learning*. Ed. by S. Sra, S. Nowozin, and S. J. Wright. MIT Press, 2011. Chap. 5, pp. 121–148.

[96] A. Juditsky and A. Nemirovski. "First order methods for nonsmooth convex large-scale optimization, II: utilizing problem structure". In: *Optimization For Machine Learning*. Ed. by S. Sra, S. Nowozin, and S. J. Wright. MIT Press, 2011. Chap. 6, pp. 149–183.

[97] L. Kantorovitch. "On the Translocation of Masses". In: *Dokl. Akad. Nauk SSSR* 37.7-8 (1942), pp. 227–229.

[98] M. Khodak, A. Risteski, C. Fellbaum, and S. Arora. "Automated wordnet construction using word embeddings". In: *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*. 2017, pp. 12–23.

[99] T. N. Kipf and M. Welling. "Semi-supervised classification with graph convolutional networks". In: *International Conference on Learning Representations (ICLR)* (2017).

[100] P. Kohli, A. Osokin, and S. Jegelka. "A Principled Deep Random Field Model for Image Segmentation". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1971–1978.

[101] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. "Optimal Mass Transport: Signal processing and machine-learning applications". In: *IEEE Signal Processing Magazine* (2017).

[102] S. Kolouri, Y. Zou, and G. K. Rohde. "Sliced Wasserstein kernels for probability distributions". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2016.

[103] T. Koo, A. Globerson, X. Carreras Pérez, and M. Collins. "Structured prediction models via the matrix-tree theorem". In: *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).* 2007, pp. 141–150.

[104] J. J. Kosowsky and A. L. Yuille. "The invisible hand algorithm: Solving the assignment problem with statistical physics". In: *Neural Networks* 7.3 (1994), pp. 477–490.

[105] A. Krause and D. Golovin. *Submodular function maximization.* 2014.

[106] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. "From Word Embeddings To Document Distances". In: *Proceedings of The 32nd International Conference on Machine Learning* 37 (2015), pp. 957–966.

[107] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning.* 2001.

[108] R. Lai and H. Zhao. "Multiscale Nonrigid Point Cloud Registration Using Rotation-Invariant Sliced-Wasserstein Distance via Laplace–Beltrami Eigenmap". In: *SIAM Journal on Imaging Sciences* (2017).

[109] G. Lample, L. Denoyer, and M. Ranzato. "Unsupervised Machine Translation Using Monolingual Corpora Only". In: *International Conference on Learning Representations* (2018).

[110] C. Lee, G. Leer, and S. JungYun. "Automatic WordNet mapping using word sense disambiguation". In: *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* 2000.

[111] G.-H. Lee, D. Alvarez-Melis, and T. S. Jaakkola. "Game-theoretic Interpretability for Temporal Modeling". In: *Fainess Acountability and Transparency in Machine Learning (FAT/ML)*. 2018.

[112] G.-H. Lee, D. Alvarez-Melis, and T. S. Jaakkola. "Towards Robust, Locally Linear Deep Networks". In: *International Conference on Learning Representations*. 2019.

[113] G.-H. Lee, W. Jin, D. Alvarez-Melis, and T. Jaakkola. "Functional Transparency for Structured Data: a Game-Theoretic Approach". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 3723–3733.

[114] P. Lee and J. Li. "New Examples Satisfying Ma-Trudinger-Wang Conditions". In: *SIAM Journal on Mathematical Analysis* 44.1 (2012), pp. 61–73.

[115] Y. T. Lee, A. Sidford, and S. C.-w. Wong. "A faster cutting plane method and its implications for combinatorial and convex optimization". In: *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE. 2015, pp. 1049–1065.

[116] O. Levy and Y. Goldberg. "Linguistic Regularities in Sparse and Explicit Word Representations". In: *Proceedings of the Eighteenth Conference on Computational Language Learning*. 2014, pp. 171–180.

[117] C. Li, D. Alvarez-Melis, K. Xu, S. Jegelka, and S. Sra. "Distributional Adversarial Networks". In: *Internat. Conference on Learning Representations (ICLR) - Workshop Track* (2018). arXiv: 1706.09549.

[118] J. Li. "Smooth Optimal Transport on Hyperbolic Space". MS Thesis. University of Toronto, 2009.

[119] H. Lin and J. Bilmes. "Optimal Selection of Limited Vocabulary Speech Corpora". In: *Twelfth Annual Conference of the International Speech Communication Association*. 2011, pp. 1489–1492.

[120] G. Loeper. "On the regularity of solutions of optimal transportation problems". In: *Acta Mathematica* 202.2 (2009), pp. 241–283.

[121] L. Lovász. "Submodular functions and convexity". In: *Mathematical programming – The State of the Art*. Ed. by A. Bachem, M. Grötschel, and B. Korte. Springer-Verlag Berlin Heidelberg, 1983, pp. 235–257.

[122] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. "Deep multilingual correlation for improved word embeddings". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 250–256.

[123] X. N. Ma, N. S. Trudinger, and X. J. Wang. "Regularity of potential functions of the optimal transportation problem". In: *Archive for Rational Mechanics and Analysis* (2005).

[124] R. J. McCann. "Polar factorization of maps on Riemannian manifolds". In: *Geometric and Functional Analysis* (2001).

[125] C. Melis, M. Flores, and A. Holvoet. "On the historical expansion of non-canonically marked 'subjects' in Spanish". In: *The diachronic Typology of Non-Canonical Subjects, Amsterdam/Philadelphia, Benjamins* (2013), pp. 163–184.

[126] F. Mémoli. "Gromov–Wasserstein distances and the metric approach to object matching". In: *Foundations of computational mathematics* 11.4 (2011), pp. 417–487.

[127] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems (NIPS)* (2013). arXiv: 1310.4546.

[128] T. Mikolov, Q. V. Le, and I. Sutskever. *Exploiting Similarities among Languages for Machine Translation*. Tech. rep. 2013. arXiv: 1309.4168.

[129] G. A. Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* (1995).

[130] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro. "Decentralized quadratically approximated alternating direction method of multipliers". In: *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*. IEEE. 2015, pp. 795–799.

[131] D. Moldovan and V. Rus. "Logic form transformation of wordnet and its applicability to question answering". In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 2001.

[132] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[133] J. Moore and D. M. Perlmutter. "What does it take to be a dative subject?" In: *Natural Language and Linguistic Theory* 18.2 (2000), pp. 373–416.

[134] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. "An analysis of approximations for maximizing submodular set functions – I". In: *Mathematical Programming* 14.1 (1978), pp. 265–294.

[135] M. Nickel and D. Kiela. "Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry". In: *Proceedings of the 35th International Conference on Machine Learning* 80 (2018). Ed. by J. Dy and A. Krause, pp. 3779–3788.

[136] M. Nickel and D. Kiela. "Poincaré embeddings for learning hierarchical representations". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 6338–6347.

[137] R. Nishihara, S. Jegelka, and M. I. Jordan. "On the convergence rate of decomposable submodular function minimization". In: *Advances in Neural Information Processing Systems*. 2014, pp. 640–648.

[138] O. Pele and M. Werman. "Fast and robust earth mover's distances". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. 2009, pp. 460–467.

[139] J. Pennington, R. Socher, and C. Manning. "Glove: Global Vectors for Word Representation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[140] G. Peyré and M. Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11 (2019), pp. 355–607. arXiv: 1803.00567.

[141] G. Peyré, M. Cuturi, and J. Solomon. "Gromov-Wasserstein averaging of kernel and distance matrices". In: *International Conference on Machine Learning*. 2016, pp. 2664–2672.

[142] S. T. Piantadosi. "Zipf's word frequency law in natural language: A critical review and future directions". In: *Psychonomic Bulletin & Review* 21.5 (2014), pp. 1112–1130.

[143] S. D. Pimentel, R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons". In: *Journal of the American Statistical Association* 110.510 (2015), pp. 515–527.

[144] Q. Pradet, G. De Chalendar, and J. Baguenier-Desormeaux. "WoNeF, an improved, expanded and evaluated automatic French translation of WordNet". In: *Proceedings of the Seventh Global Wordnet Conference*. 2014, pp. 32–39.

[145] J. Rabin, G. Peyré, J. Delon, and M. Bernot. "Wasserstein barycenter and its application to texture mixing". In: *International Conference on Scale Space and Variational Methods in Computer Vision*. 2011.

[146] A. Rangarajan, H. Chui, and F. L. Bookstein. "The Softassign Procrustes Matching Algorithm". In: *Lecture Notes in Computer Science* 1230 (1997), pp. 29–42.

[147] R. Rapp. "Automatic identification of word translations from unrelated English and German corpora". In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics. 1999, pp. 519–526.

[148] R. Rapp. "Identifying word translations in non-parallel texts". In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 1995, pp. 320–322.

[149] Y. Rubner, C. Tomasi, and L. J. Guibas. "Earth mover's distance as a metric for image retrieval". In: *International Journal of Computer Vision* (2000).

[150] S. Ruder, I. Vulić, and A. Søgaard. "A survey of cross-lingual embedding models". In: *arXiv preprint arXiv:1706.04902* (2017).

[151] A. M. Rush, D. Sontag, M. Collins, and T. Jaakkola. "On dual decomposition and linear programming relaxations for natural language processing". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processingnlp*. 2010, pp. 1–11.

[152] S. Rusinkiewicz and M. Levoy. "Efficient variants of the ICP algorithm." In: *3DIM*. Vol. 1. 2001, pp. 145–152.

[153] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. "Improving GANs using optimal transport". In: *International Conference on Learning Representations*. 2018.

[154] F. Santambrogio. "Introduction to Optimal Transport Theory". In: *arXiv preprint arXiv:1009.3856* (2010), pp. 1–16. arXiv: 1009.3856.

[155] M. Saveski and I. Trajkovski. "Automatic construction of wordnets by using machine translation and language modeling". In: *Proceedings of the Seventh Language Technologies Conference*. 2010.

[156] B. Schmitzer. "Stabilized sparse scaling algorithms for entropy regularized transport problems". In: *SIAM Journal on Scientific Computing* 41.3 (2019), A1443–A1481.

[157] P. H. Schönemann. "A generalized solution of the orthogonal procrustes problem". In: *Psychometrika* 31.1 (1966), pp. 1–10.

[158] L. Shi and R. Mihalcea. "Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing". In: *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. Springer. 2005, pp. 100–111.

[159] P. Shvaiko and J. Euzenat. *Ontology matching: State of the art and future challenges.* 2013.

[160] A. Singhal. "Modern information retrieval: A brief overview". In: *IEEE Data Eng. Bull.* (2001).

[161] R. Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *Annals of Mathematical Statististics* 35 (1964), pp. 876–879.

[162] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. "Offline bilingual word vectors, orthogonal transformations and the inverted softmax". In: *International Conference on Learning Representations* (2017).

[163] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. "Convolutional Wasserstein Distances : Efficient Optimal Transportation on Geometric Domains". In: *ACM Transactions of Graphics (TOG)* 34 (2015), p. 66.

[164] J. Solomon, G. Peyré, V. G. Kim, and S. Sra. "Entropic metric alignment for correspondence problems". In: *ACM Transactions on Graphics (TOG)* 35.4 (2016), p. 72.

[165] D. Sontag, A. Globerson, and T. Jaakkola. "Introduction to Dual Decomposition for Inference". In: *Optimization for Machine Learning* (2010).

[166] D. Spohr, L. Hollink, and P. Cimiano. "A machine learning approach to multilingual and cross-lingual ontology matching". In: *International Semantic Web Conference*. Springer. 2011, pp. 665–680.

[167] M. Staib, S. Claici, J. M. Solomon, and S. Jegelka. "Parallel Streaming Wasserstein Barycenters". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 2644–2655.

[168] P. Stobbe and A. Krause. "Efficient Minimization of Decomposable Submodular Functions". In: *Advances in Neural Information Processing Systems*. 2010, pp. 2208–2216.

[169] K. S. Tai, R. Socher, and C. D. Manning. "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.* 2015, pp. 1556–1566. arXiv: 1503.0075.

[170] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin. "Learning structured prediction models: A large margin approach". In: *Proceedings of the 22nd international conference on Machine learning.* ACM. 2005, pp. 896–903.

[171] Y. Tay, L. A. Tuan, and S. C. Hui. "Hyperbolic representation learning for fast and efficient neural question answering". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.* ACM. 2018, pp. 583–591.

[172] L. Torresani, V. Kolmogorov, and C. Rother. "Feature correspondence via graph matching: Models and global optimization". In: *European Conference on Computer Vision.* Springer. 2008, pp. 596–609.

[173] S. Upadhyay, M. Faruqui, C. Dyer, and D. Roth. "Cross-lingual models of word embeddings: An empirical comparison". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016), pp. 1661–1670. arXiv: 1604.00425.

[174] C. Villani. *Optimal Transport: Old and New.* Vol. 338. Springer Science & Business Media, 2008.

[175] C. Villani. *Topics in Optimal Transportation.* American Mathematical Society, 2003.

[176] C. Wang and S. Mahadevan. "Manifold alignment without correspondence". In: *IJCAI International Joint Conference on Artificial Intelligence.* 2009, pp. 1273–1278.

[177] W. Wang, D. Slepcev, S. Basu, J. A. Ozolek, and G. K. Rohde. "A linear optimal transportation framework for quantifying and visualizing variations in sets of images". In: *International Journal of Computer Vision* (2013).

[178] B. Wilson and M. Leimeister. "Gradient descent in hyperbolic space". In: *arXiv preprint arXiv:1805.08207* (2018). arXiv: 1805.08207.

[179] P. Wolfe. "Finding the nearest point in A polytope". In: *Mathematical Programming* 11.1 (1976), pp. 128–149.

[180] G. U. Yule. "On the Methods of Measuring Association Between Two Attributes". In: *Journal of the Royal Statistical Society* 75.6 (1912), pp. 579–652.

[181] Y. Zemel and V. M. Panaretos. "Fréchet means and Procrustes analysis in Wasserstein space". In: *Bernoulli* 25.2 (2019), pp. 932–976.

[182] H. Zhang, S. J. Reddi, and S. Sra. "Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4592–4600.

[183] M. Zhang, Y. Liu, H. Luan, and M. Sun. "Adversarial training for unsupervised bilingual lexicon induction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2017, pp. 1959–1970.

[184] M. Zhang, Y. Liu, H. Luan, and M. Sun. "Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 1934–1945.

[185] Y. Zhang, D. Gaddy, R. Barzilay, and T. Jaakkola. "Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016, pp. 1307–1317.

[186] L. Zhu, Y. Yang, S. Haker, and A. Tannenbaum. "An Image Morphing Technique Based on Optimal Mass Preserving Mapping". In: *IEEE Transactions on Image Processing* 16.6 (2007), pp. 1481–1495.