Optimal transport for transfer learning across spaces

Quang Huy Tran

COLLEGE MATHSTIC
DOCTORAL BRETAGNE
BRETAGNE OCEANE

Université Bretagne Sud

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ BRETAGNE SUD

ÉCOLE DOCTORALE N° 644
Mathématiques et Sciences et Technologies de l'Information et de la Communication en Bretagne Océane
Spécialité : *Informatique*

Par

Quang Huy TRAN

Optimal transport for transfer learning across spaces

Thèse présentée et soutenue à Vannes, le 15 Mai 2024
Unité de recherche : Institut de Recherche en Informatique et Systèmes Aléatoires - IRISA

Rapporteurs avant soutenance :

Facundo MEMOLI — Professor, Ohio State University
François-Xavier VIALARD — Professeur, Université Gustave Eiffel (LIGM)

Composition du Jury :

Président :	Prénom NOM	Fonction et établissement d'exercice *(à préciser après la soutenance)*
Examinateurs :	Laetitia CHAPEL	Professeure, Institut Agro Rennes-Angers
	Alain DURMUS	Professeur, CMAP, Ecole Polytechnique
	Gabriel PEYRE	Directeur de recherche CNRS, DMA, École Normale Supérieure
Dir. de thèse :	Nicolas COURTY	Professeur des universités, Université Bretagne Sud
Co-dir. de thèse :	Karim LOUNICI	Professeur, CMAP, Ecole Polytechnique
	Rémi FLAMARY	Professeur, CMAP, Ecole Polytechnique

Invité(s) :

Prénom NOM — Fonction et établissement d'exercice

Table of Contents

Notations

Measure theory

$\mathcal{P}(\mathcal{X})$	Space of probability measures on a space X
$\mathcal{M}^+(X)$	Space of finite nonnegative Borel measures on a space X
$U(\mu, \nu)$	Set of admissible couplings, whose marginals are μ and ν
$\mathcal{X}$	All weighted objects, including weighted (metric) space, metric-measure space, measure network, weighted matrix, measure hypernetwork, sample-feature space, are written in italic
$\#$	Push-forward operator
$\pi_{\#1}, \pi_{\#2}$	First and second marginal distributions of measures π, respectively *i.e.*, if $\pi \in \mathcal{M}^+(X \times Y)$, then $\pi_{\#1}(x) = \int_Y d\pi(x, y)$ and $\pi_{\#2}(y) = \int_X d\pi(x, y)$
$\mu \otimes \nu$	Product measure between two measures μ and ν
$\rightharpoonup$	Weak convergence
$m(\mu)$	Mass of measure μ
Δ_n	Set of histograms of n bins, *i.e.*, $\Delta_n := \{p \in \mathbb{R}^n_{>0} : \sum_i p_i = 1\}$

Linear algebra

X	Matrix in discrete setting, or space in continuous setting
$\otimes$	Tensor-matrix multiplication: given a 4D-tensor L and a matrix P, the matrix $L \otimes P$ is defined by $(L \otimes P)_{ij} = \sum_{k,l} L_{ijkl} P_{kl}$
$\oplus$	Sum defined by $(f \oplus g)(x, y) = f(x) + g(y)$
$\odot$	Element-wise multiplication
$\langle \cdot, \cdot \rangle$	Scalar product
$[n]$	Set of the first n positive integers, *i.e.*, $[n] := \{1, ..., n\}$
1_d	d-dimensional vector of ones in $\mathbb{R}^d$

Acronym

OT, UOT	Balanced and Unbalanced Optimal Transport
GW, UGW	Balanced and Unbalanced Gromov-Wasserstein
FGW, FUGW	Fused Balanced and Unbalanced Gromov-Wasserstein
COOT, UCOOT	Balanced and Unbalanced Co-Optimal Transport
AGW	Augmented Gromov-Wasserstein
MMOT	Multi-marginal Optimal Transport
KL	Kullback-Leibler divergence

Chapter 1

Introduction

1.1 Why and how optimal transport?

In the recent years, the unprecedented versatility of optimal transport (OT) has gone far beyond the original formulation of Monge (1781) on the least effort problem, and the seminal work of Kantorovich (1942). From a high-level perspective, we can conceptually describe the OT as a principled approach to compare **weighted objects** (*i.e.*, sets equipped with certain measures), namely graphs (Nikolentzos, Meladianos, and Vazirgiannis, 2017), texts (Kusner et al., 2015), images (Arjovsky, Chintala, and Bottou, 2017), persistence diagrams (Edelsbrunner, Letscher, and Zomorodian, 2002), or tabular data (Redko et al., 2020), while providing a way to align their elements in many situations.

1.1.1 Optimal transport for probability measures

The most fundamental application of OT, featured by the Wasserstein distance, is on the comparison of probability measures. Needless to say, this task is ubiquitous in statistical learning. A classic example is the maximum likelihood estimation, which is asymptotically equivalent to finding an empirical model "closest" to the true one, in terms of Kullback-Leibler divergence.

Giving the long history of development of statistics and probability theory, there are countless choices of divergences existing in the literature [1]: Kullback-Leibler divergence, total variation and Euclidean distance to name a few. But what distinguishes Wasserstein distance from them in practice? First, as opposed to many popular divergences, it allows to compare probability measures with non-overlapped supports. This is because it is not based on bin-by-bin comparison, but rather on all pairwise relations across the supports captured by the distance function. As a result, the Wasserstein distance can characterize the weak convergence, which proves to be

1. For a comprehensive and up-to-date taxonomy of divergences, see https://franknielsen.github.io/Divergence/Poster-Distances.pdf.

particularly useful, for example, in training generative adversarial networks (Arjovsky, Chintala, and Bottou, 2017). Second, as for many OT-based divergences, the resulting OT plan contains meaningful information on the correspondances between samples, and has found important applications, namely in domain adaptation (Courty et al., 2016) and genomics (Schiebinger et al., 2019). Broadly speaking, the OT plan can 1) be used to estimate the barycentric mapping (Courty et al., 2016; Ferradans et al., 2014), which can be seen as the projection of the source data onto the target domain, or 2) provide the label predictions on the target domain in the classification problem, without the need of training a classifier (Redko et al., 2019).

The success of Wasserstein distance, or more generally OT, can also be found in computer graphics (Bonneel and Digne, 2023; Bonneel, Peyré, and Cuturi, 2016; Solomon et al., 2015), dictionary learning (Rolet, Cuturi, and Peyré, 2016), supervised machine learning (Frogner et al., 2015) and natural language processing (Kusner et al., 2015), to name a few.

1.1.2 Optimal transport across spaces and beyond probability measures

By definition, Wasserstein distance requires that the probability measures must live in a common probability space. This is equivalent to comparing two probability spaces whose supports lie in the same underlying (typically metric) space. In other words, we implicitly assume that 1) it is always feasible to compute the inter-domain distance, which also happens to be the only ingredient we need to calculate the Wasserstein distance, and 2) the measures must have unit mass, due to the marginal constraints. To what extent can we relax these assumptions?

Can we handle finite nonnegative measures? Yes, for example, by replacing the hard marginal constraints by soft penalties. This gives rise to the celebrated *Unbalanced Optimal Transport*, first proposed by Benamou (2003). However, in practice, a much more popular alternative is due to the works of Frogner et al. (2015) and Liero, Mielke, and Savaré (2018).

Can the supports lie in different (also known as, incomparable) spaces? Yes, the comparison between two incomparable spaces is still feasible, though indirectly. A natural approach is to project them onto a sufficiently rich common space, so that it is possible to calculate the Wasserstein distance between their embeddings. This leads to a whole family of distances originated from the Gromov-Hausdorff distance. The most famous member in this class is the Gromov-Wasserstein (GW) distance (Mémoli, 2007, 2011b). We also note that direct comparison is still possible, for example, in the discrete setting. More precisely, Redko et al. (2020) take a radically different perspective to exploit the input data, by considering the pairwise differences between the coordinates of samples across spaces and learning simultaneously the sample and feature alignments. The corresponding distance, called Co-Optimal transport, provides a principled approach to compare weighted matrices.

How can we integrate extra information into OT? A natural solution is to optimize a linear combination of the OT's objective function and a term which takes into account the prior knowledge. This simple strategy usually works well in practice and has been used, namely to compare the weighted labelled graphs (Vayer et al., 2019a), or to incorporate the mutual information when the inter and intra-domain distances are informative (Chuang, Jegelka, and Alvarez-Melis, 2023), or in semi-supervised domain adaptation, where the alignments between some labelled source and target samples are available (Courty et al., 2016; Gu et al., 2022).

1.2 Thesis outlines and contributions

My broad research interest lie in the intersection of research directions addressed by the aforementioned questions, with major focus on the incomparable spaces. In particular, this thesis titled *Optimal transport for transfer learning across domains* has several main objectives.

1. Given the discrete nature of Co-Optimal transport, it is natural to study its continuous extension, which may serve as the first step towards further analysis on the numerical, and potentially, statistical properties. Given the close connection between COOT and GW distance, understanding one may shed light on understanding the other. This objective is addressed in Chapters 3 and 4.
2. Merging techniques from different branches in OT allows to have the best of many worlds, thus can provide efficient solutions to problems arised from real-world applications. This strategy has been successfully used, for example in the unbalanced GW (Séjourné, Vialard, and Peyré, 2021b) and the fused GW (Vayer et al., 2019a). This objective is addressed in Chapters 4 to 6.
3. It is desirable to further unlock the potential applicability of COOT on across-domain matching. From a practical perspective, COOT brings two distinct advantages: the access to the feature correspondances and a novel approach to exploit the data. To give an example, COOT offers unique opportunity to perform genomic feature alignments in single-cell multi-omics, which do not exist for GW or other OT-based divergences. This objective is addressed in Chapters 4 and 6.

Our contributions are in line with these objectives and summarized in the next paragraphs. Last but not least, during his PhD, the author has also contributed to the open-source packages, namely Python Optimal Transport, available at `https://pythonot.github.io/`, and Fused Unbalanced Gromov-Wasserstein, available at `https://github.com/alexisthual/fugw`.

Chapter 2 : Technical background on optimal transport The purpose of this chapter is to provide relevant mathematical and numerical background on OT. In particular, we focus on the intuition and motivation of the core starting points of this thesis, notably the balanced

and unbalanced OT, and the GW distance. Second, we spend much effort on their algorithmic and implementation details, which are not always discussed in the literature. We also provide a review on some variations of OT, coming mostly from the machine learning community. This may serve as a supplement to the excellent reference on computational OT of Peyré and Cuturi (2019) and to the informative tutorial of Séjourné, Peyré, and Vialard (2022) on how to used unbalanced transport in machine learning.

Chapter 3: Contributions to Co-Optimal Transport In this chapter, we present two contributions to the study of COOT. The first one is based on our unpublished working paper on continuous COOT in November 2020 and bears similarity with two concurrent works. The first one is the hypergraph COOT (Chowdhury et al., 2023) published in December 2021. Their work and ours are based on the same mathematical framework of Chowdhury and Mémoli (2019), thus result in the same metric property. Apart from that, they pursue different research objectives, where COOT is used to explore the categorical properties of the space of measure hypernetworks. By contrast, we consider continuous COOT as the first step towards the analysis of entropic approximation, unbalanced extension and sample complexity.

Our study on the entropic COOT also shares some resemblance to the approximation error of entropic GW in the paper of Zhang et al. (2022a) published in December 2022. Their analysis and ours use the block approximation technique (Carlier et al., 2017) to quantify the approximation error. In particular, our result can be immediately extended to the GW setting. However, we consider different assumptions on the measure networks, which result in different upper bound of the approximation error.

The second contribution is based on (Tran et al., 2021) and published in NeuRIPS Workshop in Optimal Transport and Machine Learning (OTML 2021). The motivating idea is that, COOT can be reformulated as a multi-marginal OT problem under the additional factorization constraint. We generalize this observation and consider the factored multi-marginal OT problem and its factorization relaxation via Kullback-Leibler divergence. Such relaxation not only enjoys nice interpolation properties, but also can be easily approximated, thanks to the Difference-of-Convex algorithm. Despite high computational cost incurred by the cost tensor, it can provide decent estimation of the COOT and GW distance.

- Factored couplings in multi-marginal optimal transport via difference of convex programming. **Quang Huy Tran**, Hicham Janati, Ievgen Redko, Rémi Flamary and Nicolas Courty. *NeurIPS Workshop on Optimal Transport and Machine Learning*, 2021.

Chapter 4: Unbalanced Co-Optimal Transport We present the unbalanced extension of COOT in the continuous setting. We show that our formulation is a well-defined optimization problem. More importantly, it guaratees the provable robustness to outliers, which is not the case

for COOT. As a byproduct, this result can be extended immediately to GW and unbalanced GW. The proposed method also shows favorable performance in unsupervised heterogeneous domain adaptation and single-cell multi-omics tasks. For the latter, in the unbalanced scenarios, the feature coupling allows to efficiently recover the alignments between genes, when the across-domain cells have different number of genomic features.

This chapter is based on (Tran et al., 2023) and has been accepted at the AAAI Conference on Artificial Intelligence (AAAI 2023). The implementation of Unbalanced COOT will be integrated into the next release of Python Optimal Transport package (Flamary et al., 2021).

- Unbalanced Co-Optimal Transport. **Quang Huy Tran**, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci and Ritambhara Singh. *AAAI Conference on Artificial Intelligence*, 2023.

Chapter 5: Fused Unbalanced Gromov-Wasserstein We present an OT-based method for inter-subject alignment on the brain data, denoted as Fused Unbalanced Gromov-Wasserstein (FUGW). It allows to align cortical surfaces based on the similarity of their functional signatures in response to a variety of stimulation settings, while penalizing large deformations of individual topographic organization. We demonstrate that FUGW is well-suited for whole-brain landmark-free alignment. The unbalanced feature allows to deal with the fact that functional areas vary in size across subjects. Our results show that FUGW matching significantly increases between-subject correlation of activity for independent functional data, and leads to more precise mapping at the group level. The proposed method also allows to learn better barycenters (also known as brain templates), comparing to other anatomical alignment approaches.

This chapter is based on (Thual et al., 2022) and has been accepted at the Neural Information Processing Systems (NeurIPS 2022). It is a collaborative work with MIND team at INRIA Saclay. The main contribution of the author is on the formulation of the mathematical framework and the implementation of the algorithms.

- Aligning individual brains with Fused Unbalanced Gromov-Wasserstein. Alexis Thual*, **Quang Huy Tran***, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene and Bertrand Thirion. *Neural Information Processing Systems (NeurIPS)*, 2022.

Chapter 6: Breaking isometric ties and introducing priors in Gromov-Wasserstein distance It is known that GW distance is invariant under isometric transformations. However, not all of them are born equal. For example, digits 6 and 9 are isomorphic in GW sense, but they clearly represent different labels. We propose a simple, yet efficient variation of GW distance, called *Augmented Gromov-Wasserstein* (AGW) divergence, which partially addresses the above isssue. More precisely, AGW learns simultaneously the sample and feature couplings by linearly combining the objective functions of GW distance and COOT. We show that such combination

results in much less isometries than GW distance. More importantly, the strong performance of AGW in our experiments indicates that these invariants appear to be relevant and meaningful. Furthermore, the information on the feature correspondances also proves to be particularly useful in the single-cell multi-omics integration tasks.

This chapter is based on (Demetci et al., 2024) and has been accepted at the International Conference on Artificial Intelligence and Statistics (AISTATS 2024). It is a collaborative work with Ievgen Redko (Huawei), Pinar Demetci (Broad Institute) and Ritambhara Singh (Brown University). The main contribution of the author is on the theoritical analysis of the proposed method. In particular, despite its simplicity, the study of the isometries induced by AGW requires very different techniques to those of GW distance.

- Breaking isometric ties and introducing priors in Gromov-Wasserstein distance. Pinar Demetci, **Quang Huy Tran**, Ievgen Redko and Ritambhara Singh. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Chapter 2

Technical background on optimal transport

In this chapter, we provide relevant technical background to three problems: balanced optimal transport (OT), unbalanced OT and Gromov-Wasserstein distance. The general structure for each topic includes the motivation, the theoritical and numerical aspects. In particular, we focus on the numerics, which are not usually discussed in the OT literature. By contrast, we only briefly present the theory since it has already been well-studied and can be found in many prior works, which will be precised during the discussion of each topic.

We start with the balanced OT, which compares probability measures whose supports live in the same underlying space. Then, we study two important extensions of this problem. The first one is based on the relaxation of the hard marginal constraints, which results in the unbalanced OT. The second generalization considers the situation where the supports of the probability measures lie in incomparable spaces. This leads to the Gromov-Wasserstein distance, whose origin comes from the Gromov-Hausdorff distance adapted to the Wasserstein distance.

2.1 From Wasserstein distance

2.1.1 Balanced Optimal Transport

We present two viewpoints on the motivation of the OT problem: the Monge problem serves as the historical starting point and the Hausdorff distance later provides the intuition of the Gromov-Wasserstein distance. Then, we introduce some important theoritical properties of the Wasserstein distance, which are widely used in machine learning applications. Finally, we discuss in details the algorithmic and practical aspects of the entropic approximation of the OT distance.

Motivation

Relation with Monge's problem The original OT problem was first formulated by Monge (1781). From a mathematical viewpoint, it aims at transporting all the mass from one probability distribution to the other, so that the displacement cost is minimized. Typically, this cost is measured by a distance function which takes supports of the probability measures as inputs. Transporting from one probability measure μ to the other one ν is equivalent to finding a map T such that $\nu = T_{\#}\mu$. We follow Santambrogio (2015) and define the Monge's problem as follows.

Definition 2.1.1 (Push-forward measure)**.** *Let X, Y be two measurable spaces. Given a probability measure μ on X and a measurable map $T : X \to Y$, we call $T_{\#}\mu \in \mathcal{P}(Y)$ the **push-forward** measure of μ by T, defined by $T_{\#}\mu(E) = \mu(T^{-1}(E))$, for every $E \subset Y$. Equivalently, for every measurable bounded function $\varphi : Y \to \mathbb{R}$, we have $\int_Y \varphi \, dT_{\#}\mu = \int_X \varphi \circ T \, d\mu$. We also say T is a **transport map** from μ to ν.*

Definition 2.1.2 (Monge's problem)**.** *Let X, Y be two complete and separable metric spaces. Given two probability measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ and a measurable cost function $c : X \times Y \to \mathbb{R} \cup \{\infty\}$, we define the Monge's problem as*

$$MOT(\mu, \nu) = \inf_{T \in \mathcal{T}(\mu,\nu)} \int_X c(x, T(x)) \, d\mu(x), \tag{2.1}$$

where $\mathcal{T}(\mu, \nu) := \{T : X \to Y \text{ measurable such that } T_{\#}\mu = \nu\}$ is the set of transport maps from μ to ν. Despite the natural interpretation, the Monge's formulation has some major drawbacks. First, its objective function is nonconvex, thus brings much difficulty to the theoretical analysis and numerical optimization. Second, the transport map may not exist. For example, if the supports of μ and ν are finite such that $|\text{supp}(\mu)| > |\text{supp}(\nu)|$, then the set $\mathcal{T}(\mu, \nu)$ is empty because any transport map must be surjective. Even when it exists, there is no guarantee that the infimum can be attained. Last but not least, the Monge's problem is asymmetric, in the sense that $\text{MOT}(\mu, \nu) \neq \text{MOT}(\nu, \mu)$.

Instead of enforcing one-to-one relation, we can allow one-to-many alignment, meaning that

the mass transported by a source point can be splitted to various target points. Formally, we consider the set of admissible couplings (or transport plans) defined as

$$U(\mu,\nu) := \{\pi \in \mathcal{P}(X \times Y) : \pi_{\#1} = \mu, \pi_{\#2} = \nu\}, \tag{2.2}$$

where $\pi_{\#1} = \int_Y d\pi(\cdot, y)$ and $\pi_{\#2} = \int_X d\pi(x, \cdot)$ are the marginal distributions of the measure π. Clearly, $\mathcal{T}(\mu,\nu) \subset U(\mu,\nu)$, since for any transport map T (if exists), we have $(\mathrm{Id}, T)_{\#\mu} \in U(\mu,\nu)$. Now, we are ready to define the relaxation of the Monge problem, known as the *Kantorovich's problem* (Kantorovich, 1942).

Definition 2.1.3 (Kantorovich's problem)**.** *Let X, Y be two compact metric spaces. Given $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ and $c : X \times Y \to \mathbb{R} \cup \{\infty\}$, the Kantorovich's problem is the following optimization problem*

$$OT(\mu,\nu) = \inf_{\pi \in U(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\pi(x,y). \tag{2.3}$$

Throughout this thesis, we refer Problem (2.3) to as the OT problem. When $c(x,y) = d^p(x,y)$, for $p \geq 1$ and d is the (common) metric on the metric spaces X and Y, we obtain the famous Wasserstein distance of order p (Villani, 2003) defined as

$$W_p^p(\mu,\nu) = \inf_{\pi \in U(\mu,\nu)} \int_{X\times Y} d^p(x,y)\, d\pi(x,y). \tag{2.4}$$

Since $\mathcal{T}(\mu,\nu) \subset U(\mu,\nu)$, we have $\mathrm{MOT}(\mu,\nu) \geq \mathrm{OT}(\mu,\nu)$. Equality may hold, for example in the cases of the celebrated Brenier's theorem (Brenier, 1987) and the Birkhoff-von-Neumann theorem (Birkhoff, 1946) for continuous and discrete measures, respectively.

Relation with Hausdorff distance So far, we have seen the derivation of the OT from the Monge's problem. Now, we present another approach based on the Hausdorff distance. Most materials are inspired from (Mémoli, 2011b).

Given a compact metric space (Z,d), we denote $\mathcal{C}(Z)$ the collection of all compact subsets of Z. The Hausdorff distance between $X, Y \in \mathcal{C}(Z)$ is defined as

$$d_H^{(Z,d)}(X,Y) := \max\left\{\sup_{x\in X} d(x,Y), \sup_{y\in Y} d(y,X)\right\}, \tag{2.5}$$

where the distance between a point to a subset of a metric space is defined by $d(x,Y) := \inf_{y\in Y} d(x,y)$. It is known that $d_H^{(Z,d)}$ is a proper metric on $\mathcal{C}(Z)$ (see for example, Proposition 7.3.3 in (Burago, Burago, and Ivanov, 2001)).

Definition 2.1.4 (Correspondance)**.** *Given two non-empty sets X and Y, a subset $R \subset X \times Y$*

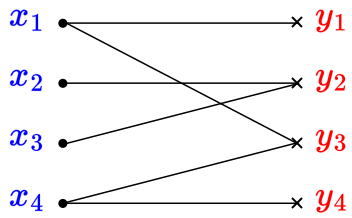

Figure 2.1 – Example of correspondance R between two sets $X = \{x_1, x_2, x_3, x_4\}$ and $Y = \{y_1, y_2, y_3, y_4\}$. Here, $R = \{(x_1, y_1), (x_1, y_3), (x_2, y_2), (x_3, y_2), (x_4, y_3), (x_4, y_4)\}$.

is a correspondance between X and Y if and only if

- *For every $x \in X$, there exists $y \in Y$ such that $(x, y) \in R$.*
- *For every $y \in Y$, there exists $x \in X$ such that $(x, y) \in R$.*

An example of correspondance is illustrated in Figure 2.1. When X and Y are finite with cardinals m and n, respectively, then all correspondances can be constructed as follows: choose a matrix $M \in \{0, 1\}^{m \times n}$ such that every row and column contains at least a value 1, then the correspondance can be defined as $R := \{(x_i, y_j) \in X \times Y : M_{ij} = 1\}$. In particular, if X and Y are disjoint, then R corresponds to a bipartite graph in which every edge has uniform weight of one. In this case, there is an intimate relation between the correspondance and the transportation plan in OT [1]. First, both describe the alignments between X and Y. Second, the transport plan can also be seen as a flow in a bipartite graph, where the source nodes must be connected to all target nodes via weighted edges. Denote $\mathcal{R}(X, Y)$ the collection of all correspondances between X and Y, then by Proposition 2.1 in (Mémoli, 2011b), we have

$$d_H^{(Z,d)}(X, Y) = \inf_{R \in \mathcal{R}(X,Y)} \sup_{(x,y) \in R} d(x, y) = \inf_{R \in \mathcal{R}(X,Y)} ||d||_{L^\infty(R)}. \tag{2.6}$$

Suppose that we equip each compact subset in $\mathcal{C}(Z)$ with a Borel probability measure and consider the collection of such "weighted spaces" $\mathcal{C}_w(Z) := \{(X, \mu_X) : \text{supp}(\mu_X) = X \text{ and } X \in \mathcal{C}(Z)\}$. Given the similarity discussed above, one can replace the correspondance by the admissible coupling and obtain the Wasserstein distance

$$W_{Z,\infty}\big((X, \mu_X), (Y, \mu_Y)\big) := \inf_{\pi \in U(\mu_X, \mu_Y)} \sup_{(x,y) \in R(\pi)} d(x, y) = \inf_{\pi \in U(\mu_X, \mu_Y)} ||d||_{L^\infty(R(\pi))}, \tag{2.7}$$

where $R(\pi) := \text{supp}(\pi)$ the support of the measure π. Note that, by Lemma 2.2 in (Mémoli, 2011b), for any $\pi \in U(\mu_X, \mu_Y)$, we have $R(\pi) \in \mathcal{R}(\text{supp}(\mu_X), \text{supp}(\mu_Y))$, thus $d_H^{(Z,d)} \leq W_{Z,\infty}$.

1. More discussion on the bipartite-graph viewpoint of OT can be found in Chapter 8 in (Brualdi, 2006) or Chapter 3.4 in (Peyré and Cuturi, 2019).

By replacing the supremum norm with the L^p-norm, we recover the p-Wasserstein distance

$$W_{Z,p}((X,\mu_X),(Y,\mu_Y)) = \inf_{\pi \in U(\mu_X,\mu_Y)} ||d||_{L^p(X\times Y,\pi)}. \tag{2.8}$$

To conclude, Diagram 2.2 summarizes the process of transforming the Hausdorff distance into the p-Wasserstein distance, when the compact metric space is equipped with a probability measure. As we will see in Section 2.2, this process is particularly useful when extending to the setting of metric measure space.

$$\boxed{d_H^Z = \inf_R ||d||_{L^\infty(R)}} \xrightarrow{W} \boxed{W_{Z,\infty} = \inf_\pi ||d||_{L^\infty(R(\pi))}} \xrightarrow{L^p} \boxed{W_{Z,p} = \inf_\pi ||d||_{L^p(\pi)}}$$

Figure 2.2 – "W" indicates the *Wassersteinization* process defined by replacing the optimization over the correspondances by over the admissible couplings. "L^p" indicates the *L^p-ization* process defined by replacing the supremum norm by the L^p-norm.

Theory

Since the seminal work of Kantorovich (1942), the theory of OT has been profoundly developped in the last decades. Different theoretical aspects with different level of generality (from compact metric space to Polish space) are covered in various excellent references, (Ambrosio, Gigli, and Savaré, 2005; Santambrogio, 2015; Villani, 2003, 2009), to name a few. This list is by no means exhaustive or representative. In this thesis, we only present some basic and useful properties of OT, which have much impact in machine learning.

Existence of solution, metric and weak convergence properties The existence of minimizer of the Kantorovich problem is guaranteed, for example when the cost is lower semi-continuous and bounded below (Theorem 4.1 in (Villani, 2009)). Note that, apart from the classic choice of cost function $c = d^p$ as in the Wasserstein distance, there are other alternatives, for example, the Bregman divergence (Guo et al., 2021), or even the Wasserstein distance (Huizing, Cantini, and Peyré, 2022).

The Wasserstein distance defines a metric on the space of probability measures with finite moments of order $p \geq 1$ (Theorem 7.3 in (Villani, 2003)) and characterizes the weak convergence: for any $p \geq 1$, we have $\mu_n \rightharpoonup \mu$ if and only if $W_p(\mu_n, \mu) \to 0$ (Theorem 7.12 in (Villani, 2003)). This topological property also holds for the integral probability metric (Müller, 1997), but not for other popular statistical divergences, namely the Kullback-Leibler divergence, the total variation, or the Hellinger distance [2].

2. More detailed comparison amongst divergences can be found in (Gibbs and Su, 2002).

Duality theory Given the convexity of the OT problem, another very powerful property is the duality theorem, which asserts that the strong duality holds. More precisely, if X, Y are compact metric spaces and the cost c is continuous, then by Theorem 1.46 in (Santambrogio, 2015), one has

$$\mathrm{OT}(\mu,\nu) = \sup_{\substack{(f,g)\in\mathcal{C}_b(X)\times\mathcal{C}_b(Y)\\ f\oplus g\leq c}} \int_X f d\mu + \int_Y g d\nu, \tag{2.9}$$

where $\mathcal{C}_b(X), \mathcal{C}_b(Y)$ denote the space of bounded continuous functions on X, Y, respectively. Strong duality still holds in a much more general setting (see Theorem 5.10 in (Villani, 2009)). The theory of duality plays a crucial role, not only in theoritical study, but also in practical applications. In particular, it is at the heart of some exact solvers for the discrete OT problem (see Section 3 in (Peyré and Cuturi, 2019)). In the case of Wasserstein distance, the duality theory allows to deduce other reformulations of the primal problem, which have recently attracted much interest in machine learning. We now discuss two particular important applications.

2-Wasserstein distance Under mild assumptions on the probability measures, thanks to the duality and Brenier's theorems (Brenier, 1987), Theorem 2.9 in (Villani, 2003) states that the 2-Wasserstein distance can be rewritten as the optimization of convex functions. This result has been used to estimate the Wasserstein distance (Chartrand and Wohlberg, 2009; Korotin et al., 2019; Makkuva et al., 2020; Taghvaei and Jalali, 2019), where the functional is parametrized by an input convex neural network (Amos, Xu, and Kolter, 2017).

1-Wasserstein distance An important application of the duality theory is the 1-Wasserstein distance (also known as *Earth mover's distance*). Its dual problem [3] is a reformulation of the primal problem as the maximization over all 1-Lipschitz functionals, which can be parametrized by neural networks. In practice, the 1-Wasserstein distance has been successfully used in the training of generative adversarial networks (GAN) (Goodfellow et al., 2014), thanks to the seminal work of Arjovsky, Chintala, and Bottou (2017) on Wasserstein GAN (WGAN). There have been various extensions and improvements of WGAN, for example smoothed WGAN (Sanjabi et al., 2018), WGAN-GP (Gulrajani et al., 2017), WGAN-LP (Petzka et al., 2018), Sobolev-GAN (Mroueh et al., 2017). It is also the motivation for other OT-based GAN methods, namely Sinkhorn divergence-GAN (Genevay, Peyre, and Cuturi, 2018), OT-GAN (Salimans et al., 2018). Interestingly, WGAN also finds its connections with the Minkowski and Alexandrov problems in convex geometry (Lei et al., 2019), and with the soft-margin formulation of Support Vector Machine (Jolicoeur-Martineau and Mitliagkas, 2019).

3. Also known as the *Kantorovich-Rubinstein duality*, see Remark 6.5 in (Villani, 2009). More discussion on the 1-Wasserstein distance can be found in Chapter 6 in (Peyré and Cuturi, 2019).

Approximation and algorithm

In discrete setting, the OT formulation is a linear program. An example of exact solver [4] is the interior point method (Orlin, 1988). It requires the complexity of $O(n^3 \log n)$, which is computationally prohibitive in many applications. There exist some approaches to overcome this limitation [5]. For example, the mini-batch approach (Fatras et al., 2020; Sommerfeld et al., 2019) considers multiple resamplings of the data and uses the average of distances computed in the mini-batches as an estimation of the true OT distance. The "sliced" method (Bonneel et al., 2015; Rabin et al., 2012) develops an alternative metric called *Sliced Wasserstein distance*. It relies on the fact that computing the 1-D Wasserstein distance boils down to sorting the point values, which operates with a complexity of $O(n \log n)$. Using the framework from statistical physics, Koehl, Delarue, and Orland (2019) approximate the OT distance by the free energy evaluated at sufficiently small temperature. From a more mathematical viewpoint, this idea is based on the interesting relation between the minimum and the logsumexp operation, where informally, one has that: $\min f(x) = \lim_{\beta \to +\infty} -\frac{1}{\beta} \log \int e^{-\beta f(x)} dx$.

Entropic OT problem In this thesis, we focus on the line of works on the discrete entropic approximation [6] defined by: for $\varepsilon > 0$,

$$\mathrm{OT}_\varepsilon(\mu, \nu) = \inf_{P \in U(\mu,\nu)} \langle C, P \rangle + \varepsilon \, \mathrm{KL}(P | \mu \otimes \nu). \tag{2.10}$$

In the OT literature, the entropic regularization can sometimes be referred to the case where the regularizer is the negative entropy defined by $H(P) = \sum_{i,j} P_{ij}(\log P_{ij} - 1)$. However, since $\mathrm{KL}(P|\mu \otimes \nu) = H(P) - H(\mu) - H(\nu)$, for any $P \in U(\mu, \nu)$, the two regularized problems are equivalent, up to a constant. Interestingly, the entropic OT is an even older problem than the unregularized one. It was already studied in the early 30's by Schrödinger (1932), under the name *static Schrodinger bridge problem*. The entropic OT problem has attracted much attention to the machine learning community thanks to the seminal work of Cuturi (2013).

Note that, the entropic regularization introduces bias, in the sense that $\mathrm{OT}_\varepsilon(\mu, \mu) \neq 0$, for any $\varepsilon > 0$. This issue can be easily overcome by considering the Sinkhorn divergence [7] (Feydy et al., 2019; Ramdas, Trillos, and Cuturi, 2017) defined by

$$\mathrm{SD}_\varepsilon(\mu, \nu) = \mathrm{OT}_\varepsilon(\mu, \nu) - \frac{1}{2} \left[\mathrm{OT}_\varepsilon(\mu, \mu) + \mathrm{OT}_\varepsilon(\nu, \nu) \right]. \tag{2.11}$$

4. See also Chapter 3 in (Peyré and Cuturi, 2019) for a more detailed discussion on classic algorithms and Section 2.1 in (Pele and Werman, 2009) for an overview on computational complexity

5. See Section 3 in (Bonneel and Digne, 2023) for an overview on the numerical solvers.

6. For a mathematical introduction in the general setting, interested readers may consult the lecture note of Nutz (2022).

7. This divergence is implemented very efficiently in the Geomloss package (Feydy et al., 2019).

Not only being unbiased, the Sinkhorn divergence also enjoys favorable topological, statistical and interpolation properties (Feydy et al., 2019; Genevay et al., 2019). In particular, if the ground cost is the squared l_2-norm, then it is a good proxy for the squared 2-Wasserstein distance. More precisely, under mild conditions on μ and ν, one has that $|\mathrm{SD}_\varepsilon(\mu,\nu) - W_2^2(\mu,\nu)| = O(\varepsilon^2)$ (Chizat et al., 2020).

Properties The entropic OT has many interesting properties. First, it is a good and scalable approximation of OT distance, where the approximation error can be quantified in some settings (Genevay et al., 2019; Luise et al., 2018). Moreover, the convergence behaviors of minimum and minimizer are also well understood (Carlier et al., 2017; Léonard, 2012) and the convergence rate exists in some practical situations (Cominetti and Martín, 1994; Genevay et al., 2019; Weed, 2018). Second, since entropic OT is a convex problem, strong duality holds and one has that

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \sup_{\substack{f\in\mathbb{R}^m \\ g\in\mathbb{R}^n}} \langle f,\mu\rangle + \langle g,\nu\rangle - \varepsilon \left\langle \exp\left(\frac{f\oplus g - C}{\varepsilon}\right), \mu\otimes\nu \right\rangle. \tag{2.12}$$

Arguably, this is the most important premise in the study of the entropic approximation, which allows to develop and analyze many Sinkhorn-based algorithms. In particular, Genevay et al. (2019) use the dual problem to show that entropic OT has better sample complexity than unregularized OT, though it fixes the bottleneck in the sample size by introducing a new one in the regularization. This issue can be mitigated, for example when the probability measures are sub-Gaussian (Mena and Niles-Weed, 2019). Entropic OT is also known for its metric property (Sanjabi et al., 2018), joint convexity and smoothness with respect to the inputs (Luise et al., 2018).

Algorithm To solve the dual problem (2.12), we are interested in the deterministic approach, where the most popular solver is the Sinkhorn algorithm (Sinkhorn and Knopp, 1967) [8]. Denote $K = e^{-C/\varepsilon}$ the kernel matrix. Then, from the first-order optimality conditions of the dual problem, the iterates read

$$u^{(t+1)} = \frac{\mu}{Kv^{(t)}} \text{ and } v^{(t+1)} = \frac{\nu}{K^T u^{(t+1)}}. \tag{2.13}$$

This algorithm is very easy to implement and available in the Python Optimal Transport (Flamary et al., 2021) and OTT-JAX (Cuturi et al., 2022) packages. It is well known that it converges globally to the optimal dual vectors of the dual problem (Sinkhorn and Knopp, 1967), at a linear rate in variation semi-norm (Franklin and Lorenz, 1989). The OT plan then can be retrieved

8. Also known as *matrix scaling algorithm*. For the historical perspective, see Remark 4.5 in (Peyré and Cuturi, 2019).

from the optimal solution (u, v) by $P = (\mu \otimes \nu) \odot (u \otimes v) \odot K$, where $\odot$ is the element-wise multiplication.

Discussion on Sinkhorn algorithm The Sinkhorn algorithm has two key advantages. First, since the iterates only require vector-matrix multiplication, they can be seamlessly parallerized. One can further speed up the computation, for example, by exploiting the structure of the kernel matrix [9]. Second, it is enough to store only the dual vectors (which costs $O(n)$ in memory) because the OT plan (which costs $O(n^2)$ in memory) can be easily queried on demand.

However, the Sinkhorn algorithm suffers from the small regularization. First, if the number of iterations is not sufficiently large, the algorithm may fail to converge and the resulting transport plan can be inadmissible. To circumvent this issue, one can apply the rounding algorithm (Altschuler, Weed, and Rigollet, 2017), which adjusts the coordinates of the coupling matrix until the marginal constraints are met. While being admissible, the rounded transport plan may not be the solution of the entropic OT problem.

Second, when ε is close to 0, the quick saturation of the kernel $\exp^{-C/\varepsilon}$ to the zero matrix may result in the numerical error triggered by the division by zero. Moreover, the entries of the scaling vectors u, v may become very large, thus cause the numerical instability. To avoid these issues, the implementation in the log-domain is usually recommended at the cost of slowing down the algorithm, due to the logsumexp operation. More precisely, the log-domain iterates read

1. $f^{(t+1)} = \log \mu - \log \sum_j \exp\left(g_j^{(t)} - \frac{C_{\cdot,j}}{\varepsilon}\right)$.
2. $g^{(t+1)} = \log \nu - \log \sum_i \exp\left(f_i^{(t+1)} - \frac{C_{i,\cdot}}{\varepsilon}\right)$.

However, even for small regularization, one can still implement with matrix-vector multiplication by additionally employing the redundant parametrization trick (Chizat et al., 2018a; Schmitzer, 2019) to control the magnitude of the scaling vectors u, v.

Third, the initialization of dual vectors is important. Under the absence of prior knowledge, the most popular practice is to initialize with the zero vectors, which usually leads to poor convergence behavior for very small regularization. One can use the ε-scaling scheme (Schmitzer, 2019) [10], whose idea is simple: we fix a decreasing sequence of regularizations converging to ε, then successively use the solution of the previous problem to initialize the next one. However, this can be very costly due to the amount of subproblems, while not necessarily ensuring the convergence to the true minimizer of the entropic OT problem. Recently, Thornton and Cuturi (2023) leverage the fact that the dual vectors in unregularized OT problem can admit closed-form expressions for 1D-OT and Gaussian measures, and propose to use them as initialization. They empirically show that this practice can result in much less iterations, thus significantly speed up

9. See Section 4.3 in (Peyré and Cuturi, 2019) and the excellent thesis of Feydy (2020) for more implementation details.

10. This strategy is very efficiently exploited and implemented in the Geomloss package (Feydy et al., 2019)

Regularizer	Reference
Convex regularizer	(Marino and Gerolin, 2020)
Strongly convex regularizer (negative entropy, square 2-norm, group lasso)	(Dessein, Papadakis, and Rouas, 2016) (Blondel, Seguy, and Rolet, 2018)
Class-based regularizer (group lasso, Laplacian)	(Courty et al., 2016)
Sparse-promoting regularizer (group lasso, weighted 1-norm)	(Lindbäck, Wang, and Johansson, 2023)
Sparsity-constraint regularizer	(Liu, Puigcerver, and Blondel, 2022)
Csiszár divergence	(Terjék and González-Sánchez, 2022)
Square 2-norm	(Roberts et al., 2017) (Blondel, Seguy, and Rolet, 2018) (Lorenz, Manns, and Meyer, 2021)
Tsallis q-entropy	(Muzellec et al., 2017)
Deformed q-entropy	(Bao and Sakaue, 2022)

Table 2.1 – Classes of regularizers in regularized OT problem.

and improve the convergence of the Sinkhorn algorithm.

Summary We conclude the discussion on balanced OT with some remarks. First, despite the prohibitive theoretical complexity, efficient implementation of exact solver for unregularized OT exists (Flamary et al., 2021), and can work well on datasets of size up to $O(10^4)$.

Second, the recent improvements on the Sinkhorn algorithm, namely the over-relaxation method (Lehmann et al., 2020; Thibault et al., 2021), the screening strategy (Alaya et al., 2019), or the greedy approach (Altschuler, Weed, and Rigollet, 2017; Kostic, Salzo, and Pontil, 2021; Lin et al., 2020), usually focus on careful manipulation of the iterates. There are also gradient-based solvers for the entropic OT, notably the adaptive primal-dual accelerated mirror descent (Dvurechensky, Gasnikov, and Kroshnin, 2018; Lin, Ho, and Jordan, 2022), or the stochastic gradient descent (Abid and Gower, 2018; Genevay et al., 2016; Seguy et al., 2018).

Last, while entropy is the most popular and well-studied regularizer, there exist other alternatives summarized in Table 2.1, which allow to integrate the prior knowledge on the data, or the structure of the transport matrix.

2.1.2 Unbalanced Optimal Transport

Formulation and properties

Recall that the balanced OT enforces hard constraints on the marginal distributions of the OT plan. These constraints lead to two main disadvantages. First, imbalanced datasets where samples are re-weighted cannot be accurately compared. Second, mass transportation must be

exhaustive in the sense that, outliers, if any, must be matched regardless of the cost they induce.

To circumvent these limitations, a straightforward solution is to control the difference between the marginal distributions of the transportation plan and the data by some discrepancy measure. This gives rise to the unbalanced OT (UOT), which was first proposed by Benamou (2003). The theoretical and numerical aspects of this relaxation have been studied extensively (Chizat et al., 2018a,b; Liero, Mielke, and Savaré, 2018; Pham et al., 2020) and have been gaining increasing attention in the machine learning community, with wide-range applications, namely in domain adaptation (Fatras et al., 2021), generative adversarial networks (Balaji, Chellappa, and Feizi, 2020; Yang and Uhler, 2019), dynamic tracking (Lee, Bertrand, and Rozell, 2019), crowd counting (Ma et al., 2021), neuroscience (Bazeille et al., 2019a; Janati et al., 2019) or modeling cell developmental trajectories (Schiebinger et al., 2019). Unbalanced OT and its variants are usually sought for their known robustness to outliers (Balaji, Chellappa, and Feizi, 2020; Fatras et al., 2021; Le et al., 2021; Mukherjee et al., 2021; Nietert, Goldfeld, and Cummings, 2022).

Formulation To define the UOT problem, let us start with the Csiszár divergence (Csiszár, 1963). Given an entropy function $\varphi : \mathbb{R}_{>0} \to [0, \infty]$ (*i.e.*, it is convex, positive and lower semi-continuous such that $\varphi(1) = 0$), we define the recession constant $\varphi'_\infty \in \mathbb{R} \cup \{\infty\}$ as

$$\varphi'_\infty = \lim_{x \to \infty} \frac{\varphi(x)}{x}. \tag{2.14}$$

Denote $\mathcal{M}^+(\mathcal{S})$ the set of finite nonnegative Radon measures on a Hausdorff topological space S. The Csiszár divergence (or φ-divergence) between two measures μ and ν in $\mathcal{M}^+(\mathcal{S})$ is defined as

$$D_\varphi(\mu|\nu) = \int_{\mathcal{S}} \varphi\Big(\frac{d\mu}{d\nu}\Big) d\nu + \varphi'_\infty \int_{\mathcal{S}} d\mu^\perp, \tag{2.15}$$

where, by Lebesgue decomposition, we have $\mu = \frac{d\mu}{d\nu}\nu + \mu^\perp$. It is jointly convex, positive and weakly lower-semicontinuous (Liero, Mielke, and Savaré, 2018). The analysis of some Csiszár divergences can be found in (Séjourné et al., 2019). Here, we are mostly interested in two following special cases:

- The Kullback-Leibler (KL) divergence defined by

$$\mathrm{KL}(\mu|\nu) = \begin{cases} \int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} - \int d\mu + \int d\nu, & \text{if } \mu \ll \nu \\ \infty, & \text{otherwise.} \end{cases} \tag{2.16}$$

  corresponds to $\varphi(x) = x \log x - x + 1$.

- The indicator divergence $\iota_=(\mu|\nu)$ is equal to 0 if $\mu = \nu$ and $+\infty$ otherwise. Its entropy function is $\varphi = \iota_{\{1\}}$, where the indicator function ι_C of a convex set C is defined by $\iota_C(x)$ is equal to 0 if $x \in C$ and $+\infty$ otherwise. This divergence results in the balanced OT.

Given a proper lower semicontinuous cost function $c : X \times Y \to [0, \infty]$ and for $\rho_1, \rho_2 \geq 0$, the UOT cost between two measures $\mu \in \mathcal{M}^+(X)$ and $\nu \in \mathcal{M}^+(Y)$ is defined by

$$\mathrm{UOT}_\rho(\mu, \nu) = \inf_{\pi \in \mathcal{M}^+(X \times Y)} \int c \, d\pi + \rho_1 D_{\varphi_1}(\pi_{\#1}|\mu) + \rho_2 D_{\varphi_2}(\pi_{\#2}|\nu). \tag{2.17}$$

Properties Under mild conditions of the ground cost and entropy functions, the UOT problem always admits a solution (Theorem 3.3 in (Liero, Mielke, and Savaré, 2018)). Similar to the balanced case, UOT also enjoys the metric properties, duality theory and dynamic formulation, to name a few. For a more complete treatment on the theory of UOT, readers are invited to see the seminal works of Chizat et al. (2018b) and Liero, Mielke, and Savaré (2018), and the reference therein. The case where $D_{\varphi_1} = D_{\varphi_2} = \mathrm{KL}$ is also of our special interest, since it allows to show that the UOT is robust to outliers.

Proposition 2.1.1 (Generalization of Lemma 1 in (Fatras et al., 2021))**.** *Let μ, μ_o be two nonnegative finite Borel measures with disjoint compact supports E, E_o, respectively. For $\alpha \in [0, 1]$, denote $\widetilde{\mu} = \alpha\mu + (1-\alpha)\mu_o$ the noisy measure. Then, for any nonnegative finite Borel measures ν with compact support F, we have*

$$OT(\widetilde{\mu}, \nu) \geq (1-\alpha) \inf_{(x,y) \in E_o \times F} c(x, y), \tag{2.18}$$

whereas

$$UOT_\rho(\widetilde{\mu}, \nu) \leq \alpha \, UOT_\rho(\mu, \nu) + (1-\alpha) M \left(1 - \exp\left(-\frac{K}{M}\right)\right), \tag{2.19}$$

where $K = \int c(x, y) \, d\mu_o(x) d\nu(y)$ and $M = \rho_1 m(\mu_o) + \rho_2 m(\nu)$. Here, $m(\mu)$ denotes the mass of measure μ.

Proof. For any $\pi \in U(\widetilde{\mu}, \nu)$, we have

$$\int_{(E \cup E_o) \times Y} c(x, y) d\pi(x, y) \geq \int_{E_o \times F} c(x, y) d\pi(x, y) \geq \min_{(x,y) \in E_o \times F} c(x, y) \int_{E_o \times F} d\pi(x, y) \tag{2.20}$$

$$= \min_{(x,y) \in E_o \times F} c(x, y) \int_{E_o} d\pi_{\#1}(x) = (1-\alpha) \min_{(x,y) \in E_o \times F} c(x, y). \tag{2.21}$$

Taking the infimum over $U(\widetilde{\mu}, \nu)$, we obtain the upper bound of OT. The lower bound on UOT follows immediately from the proof of Lemma 1.1 in (Fatras et al., 2021). ■

Inequality (2.19) indicates that the behavior of the outliers is always controlled in UOT. Thanks to the marginal relaxation, if the outlier is too impactful, its impact will quickly saturate to 0, meaning that it receives no mass from other points. By contrast, this is not the case for balanced OT, since every point (including outliers) must be aligned regardless of its cost, due to

the marginal constraints. The right-hand side of Inequality (2.18) can be made arbitrarily large by taking outliers "far" enough from the clean data.

Optimization and algorithm

In general, given any differentiable divergence and regularizer, since Problem (2.41) is a bound-constrained minimization problem, it can be solved with the L-BFGS-B algorithm (Byrd et al., 1995; Zhu et al., 1997). When $D_{\varphi_1} = D_{\varphi_2} = \mathrm{KL}$, the unregularized UOT can be approximated with entropic regularization. The main advantages of this approach include well-understood convergence analysis (Chizat et al., 2018a), statistical properties (Séjourné et al., 2019), GPU-friendly implementation, strong convexity and smoothness coming from the dual problem, making it a suitable training loss for neural networks.

It is important to note that, in the OT literature, the entropic regularization can be referred to two different regularizers: the negative entropy (Chizat et al., 2018a; Frogner et al., 2015) and the KL divergence (Séjourné et al., 2019). We cover both cases by considering a more general formulation: given the hyperparameters $\rho_1, \rho_2 \geq 0, \varepsilon > 0$, the cost matrix $C \in \mathbb{R}^{m\times n}$ and the positive measures $\mu \in \mathbb{R}^m_{>0}, \nu \in \mathbb{R}^n_{>0}, \gamma \in \mathbb{R}^{m\times n}_{>0}$, we want to solve the problem

$$\mathrm{UOT}_{\varepsilon,\rho} = \min_{P\in\mathbb{R}^{m\times n}_{\geq 0}} \langle C, P\rangle + \rho_1 \mathrm{KL}(P_{\#1}|\mu) + \rho_2 \mathrm{KL}(P_{\#2}|\nu) + \varepsilon \mathrm{KL}(P|\gamma). \tag{2.22}$$

Denote $m(Q) = \sum_{i,j} Q_{ij}$ the mass of measure Q. Observe that

$$\langle C, P\rangle + \varepsilon \mathrm{KL}(P|\gamma) = \varepsilon \left[\sum_{i,j} P_{ij}\log P_{ij} - \sum_{i,j} P_{ij}\left(\log \gamma_{ij} - \frac{C_{ij}}{\varepsilon}\right) - \sum_{i,j} P_{ij} + \sum_{i,j}\gamma_{ij}\right] \tag{2.23}$$

$$= \varepsilon\ \mathrm{KL}\left(P\middle|\exp\left(-\frac{C-\log\gamma}{\varepsilon}\right)\right) + \varepsilon\left[m(\gamma) - m(\gamma\odot e^{-C/\varepsilon})\right], \tag{2.24}$$

where $\odot$ is the element-wise multiplication between two matrices and the exponential operation is also element-wise. Now, the ambiguity of the entropic regularizer is naturally handled in Problem (2.22). More precisely, the KL divergence corresponds to $\gamma = \mu\otimes\nu$, whereas the negative entropy corresponds to $\gamma = 1_{m\times n}$, since $\mathrm{KL}(P|1_{m\times n}) = \langle P, \log P\rangle - m(P) + mn = H(P) + mn$.

An interesting property of Problem (2.22), which does not exist in the balanced OT or for other Csiszár divergences, is that the minimum can be expressed as a **linear** function of minimizer. This is a special case of the following result.

Definition 2.1.5 (Bregman divergence)**.** *Suppose $\varphi : E \to \mathbb{R}$ is a strictly convex and continuously differentiable function, where $E = dom(\varphi)$ is a closed convex set in $\mathbb{R}^d$. Then, the Bregman divergence $D_\varphi : E\times E \to \mathbb{R}_{\geq 0}$ is defined as $D_\varphi(x,y) = \varphi(x) - \varphi(y) - \langle\nabla\varphi(y), x-y\rangle$.*

Corollary 2.1.1. *Let $D_{\varphi_1}, D_{\varphi_2}$ and D_φ be three Bregman divergences whose domains are in $\mathbb{R}^m, \mathbb{R}^n$ and $\mathbb{R}^{m\times n}$, respectively. Denote $dom = \{P \in dom(\varphi) : P_{\#1} \in dom(\varphi_1), P_{\#2} \in dom(\varphi_2)\}$. Given two matrices $C \in \mathbb{R}^{m\times n}, \gamma \in dom(\varphi)$ and two vectors $\mu_1 \in dom(\varphi_1), \mu_2 \in dom(\varphi_2)$, consider the following regularized UOT problem: for $\rho_1, \rho_2, \varepsilon \geq 0$,*

$$UOT_{\varepsilon,\rho} = \inf_{P\in E} \langle C, P\rangle + \rho_1 D_{\varphi_1}(P_{\#1}|\mu_1) + \rho_2 D_{\varphi_2}(P_{\#2}|\mu_2) + \varepsilon D_\varphi(P|\gamma), \tag{2.25}$$

where the constraint set $E \subset dom$ satisfies: for any $P \in E$ and $t \in \mathbb{R}$, if $tP \in dom$, then $tP \in E$. Suppose Problem (2.25) admits a solution P^*, then we have

$$\begin{aligned} UOT_{\varepsilon,\rho} = &\sum_{k=1}^{2} \rho_k \Big[\langle \nabla\varphi_k(\mu_k), \mu_k\rangle - \varphi_k(\mu_k)\Big] + \varepsilon\Big[\langle \nabla\varphi(\gamma), \gamma\rangle - \varphi(\gamma)\Big] \\ &+ \sum_{k=1}^{2} \rho_k \quad \varphi_k(P^*_{\#k}) - \frac{\partial\varphi_k(tP^*_{\#k})}{\partial t}\Big|_{t=1}\Big) + \varepsilon \quad \varphi(P^*) - \frac{\partial\varphi(tP^*)}{\partial t}\Big|_{t=1}\Big). \end{aligned} \tag{2.26}$$

We are interested in two following Bregman divergences.

1. If D_ψ is the squared l_2-norm, then $\psi(tP) = t^2\psi(P)$. So, $\langle \nabla\psi(q), q\rangle - \psi(q) = \psi(q)$ and $(p) - \frac{\partial\psi(tp)}{\partial t}\Big|_{t=1} = -\psi(p)$.
2. If D_ψ is the KL divergence [11], then $\psi(tP) = t\ \psi(P) + t\log t\ m(P) - t + 1$. So, $\langle \nabla\psi(q), q\rangle - (q) = m(q) - 1$ and $\psi(p) - \frac{\partial\psi(tp)}{\partial t}\Big|_{t=1} = 1 - m(p)$.

We note that Corollary 2.1.1 is a generalization of Lemma 4 in (Pham et al., 2020). In particular, in case of Problem (2.22), we have

$$\mathrm{UOT}_{\varepsilon,\rho} = \rho_1 m(\mu) + \rho_2 m(\nu) + \varepsilon m(\gamma) - (\rho_1 + \rho_2 + \varepsilon)\ m(P^*). \tag{2.27}$$

This relation proves to be particularly helpful when studying the computational complexity of Sinkhorn algorithm (Pham et al., 2020). Given the flexible structure of the set $\mathcal{C}$, Corollary 2.1.1 has a broad range of applications, including the unregularized UOT (*i.e.*, $\varepsilon = 0$) and other unbalanced divergences, namely the squared l_2-regularized UOT (2.37) (see Corollary 2.1.3), the unbalanced Gromov-Wasserstein (see Equation (2.84)) and the unbalanced Co-Optimal Transport (see Equation (4.2)). Even more interestingly, these examples share a common trait, that is, the minimum is a polynomial of the minimiser, whose coefficients depend only on the input measures (*i.e.*, μ, ν, γ) and hyperparameters (*i.e.*, $\rho_1, \rho_2, \varepsilon$), but not the cost C. We provide the justification of these claims in Appendix 8.1.1.

Now, we discuss some existing approaches to solve Problem (2.22).

11. KL divergence is also the only divergence which belongs to both Csiszár and Bregman families (Jiao et al., 2014).

Algorithm 1 Sinkhorn algorithm for Problem (2.22).

Input: cost matrix $C \in \mathbb{R}^{m\times n}$, measures $\mu \in \mathbb{R}^m_{>0}, \nu \in \mathbb{R}^n_{>0}, \gamma \in \mathbb{R}^{m\times n}_{>0}$, regularization $\varepsilon > 0$, relaxation parameters $\rho_1, \rho_2 > 0$.
for $t = 1, \dots, T$ **do**
  $f_i^{(t+1)} \leftarrow \frac{\rho_1}{\rho_1+\varepsilon} \overline{\mathrm{Smin}}_\varepsilon^{\mu_i,\gamma_{i,\cdot}}(C_{i,\cdot} - g^{(t)})$, for $i \in [m]$.
  $g_j^{(t+1)} \leftarrow \frac{\rho_2}{\rho_2+\varepsilon} \overline{\mathrm{Smin}}_\varepsilon^{\nu_j,\gamma_{\cdot,j}}(C_{\cdot,j} - f^{(t+1)})$, for $j \in [n]$.
end for
Output: pair of dual vectors $(f^{(T)}, g^{(T)})$.

Sinkhorn algorithm The Sinkhorn algorithm can be extended easily to the entropic UOT. [12] The (strict) convexity of the primal problem ensures that duality holds, thanks to the Fenchel-Rockafellar duality theorem. In particular, The dual problem of Problem (2.22) is equivalent to

$$\sup_{\substack{f\in\mathbb{R}^m\\ g\in\mathbb{R}^n}} -\rho_1 \left\langle \exp\left(-\frac{f}{\rho_1}\right), \mu \right\rangle - \rho_2 \left\langle \exp\left(-\frac{g}{\rho_2}\right), \nu \right\rangle - \varepsilon \left\langle \exp\left(\frac{f \oplus g - C}{\varepsilon}\right), \gamma \right\rangle. \tag{2.28}$$

Similar to the balanced case, the OT plan can be retrieved from the optimal solution (f^*, g^*) by

$$P^* = \gamma \odot \exp\left(\frac{f^* \oplus g^* - C}{\varepsilon}\right). \tag{2.29}$$

Note that, when $\rho_1, \rho_2 \to \infty$, we have $-\rho_1 \left\langle \exp\left(-\frac{f}{\rho_1}\right), \mu \right\rangle \to \langle f, \mu\rangle$ and $-\rho_2 \left\langle \exp\left(-\frac{g}{\rho_2}\right), \nu \right\rangle \to \langle g, \mu\rangle$. So, we recover the dual problem of the entropic balanced OT. When $\varepsilon \to 0$, we have $\varepsilon \left\langle \exp\left(\frac{f\oplus g - C}{\varepsilon}\right), \gamma \right\rangle$ converges to the constraint $f \oplus g \le C$, which corresponds to the dual problem of the unregularized UOT problem.

Similar to the balanced case, the log-domain Sinkhorn algorithm 1 applied to the dual problem (2.28) consists in alternatively updating the dual vectors. Here, the generalized softmin operator is defined by $\overline{\mathrm{Smin}}_\varepsilon^{\nu_j,\gamma_{\cdot,j}}(f) = \varepsilon\left(\log \nu_j - \log\langle \gamma_{\cdot,j}, e^{-f/\varepsilon}\rangle\right)$, for $j \in [n]$. In particular, when $\gamma = \mu \otimes \nu$, we recover the softmin operator (Séjourné et al., 2019) defined by $\mathrm{Smin}_\varepsilon^\mu(f) = -\varepsilon \log\langle \mu, e^{-f/\varepsilon}\rangle$. More generally, the Sinkhorn algorithm is applicable to all Csiszár divergences, as long as the regularizer is the KL divergence (see Definition 3 in (Séjourné et al., 2019)). We leave the convergence analysis to the discussion of the next method. In practice, the Sinkhorn algorithm also suffers from slow convergence for small regularization and can be accelerated using similar workarounds in the balanced setting.

12. When the regularizer is the negative entropy, the Sinkhorn algorithm is also called *scaling algorithm* (Chizat et al., 2018a).

Translation Invariant Sinkhorn (TI-Sinkhorn) Recall that, in the entropic balanced OT, due to the linear structure in the objective function of the dual problem, if (f, g) is the optimal solution, then so is $(f + \lambda, g - \lambda)$, for every $\lambda \in \mathbb{R}$. However, this translation invariant (TI) property no longer holds for the unbalanced counterpart. As observed by Séjourné, Vialard, and Peyré (2021a), this is problematic because the convergence of dual vectors depends on the translation (since the iterates do), thus is sensitive to initialization. Moreover, it becomes very slow when the regularization is too small relative to the relaxation ($\varepsilon \ll \rho$). This issue can be mitigated by considering the optimization problem

$$H_\varepsilon(\overline{f}, \overline{g}) := \sup_{\lambda \in \mathbb{R}} F_\varepsilon(\overline{f} + \lambda, \overline{g} - \lambda), \tag{2.30}$$

where F_ε is the objective function of the dual problem (2.28). By Proposition 2 in (Séjourné, Vialard, and Peyré, 2021a), the optimality happens when

$$\lambda^*(\overline{f}, \overline{g}) = \frac{\rho_1 \rho_2}{\rho_1 + \rho_2} \log \frac{\langle \mu, e^{-\overline{f}/\rho_1} \rangle}{\langle \nu, e^{-\overline{g}/\rho_2} \rangle}. \tag{2.31}$$

By construction, H_ε is TI. Not only H_ε and F_ε have the same maximum, but also the their maximizers are related by the equation $(f, g) = (\overline{f} + \lambda^*(\overline{f}, \overline{g}), \overline{g} - \lambda^*(\overline{f}, \overline{g}))$. The details of the TI-Sinkhorn algorithm applied to Problem (2.22) can be found in Algorithm 2.

To analyze the convergence of TI-Sinkhorn, we introduce two norms: the supremum norm $||f||_\infty = \max_i |f_i|$ and the Hilbert pseudo-norm $||f||_* = \frac{1}{2}(\max_i f_i - \min_i f_i)$. Denote $\kappa_\varepsilon(\mu)$ the contraction rate of the softmin operator $\mathrm{Smin}_\varepsilon^\mu$ for the norm $||f||_*$, meaning that

$$||\mathrm{Smin}_\varepsilon^\mu(C - f_1) - \mathrm{Smin}_\varepsilon^\mu(C - f_2)||_* \leq \kappa_\varepsilon(\mu)\ ||f_1 - f_2||_*. \tag{2.32}$$

The convergence of Sinkhorn and TI-Sinkhorn algorithms when $\gamma = \mu \otimes \nu$ is summarized in the following result.

Proposition 2.1.2 (Theorem 1 in (Séjourné, Vialard, and Peyré, 2021a))**.** *Denote (f^*, g^*) the optimal solution of the dual problem (2.28). Given initialization f_0, suppose $(f^{(t)}, g^{(t)}), (\overline{f}^{(t)}, \overline{g}^{(t)})$ are the iterates of the Sinkhorn and TI-Sinkhorn algorithms at iteration t, respectively. Denote $\kappa = \left(1 + \frac{\varepsilon}{\rho_1}\right)^{-1} \left(1 + \frac{\varepsilon}{\rho_2}\right)^{-1}$ and $\overline{\kappa} = \kappa_\varepsilon(\mu)\kappa_\varepsilon(\nu)\kappa$. Then, for Sinkhorn algorithm, we have*

$$||f^{(t)} - f^*||_\infty + ||g^{(t)} - g^*||_\infty \leq 2\kappa^t ||f^{(0)} - f^*||_*. \tag{2.33}$$

whereas for TI-Sinkhorn algorithm,

$$||\overline{f}^{(t)} - f^*||_\infty + ||\overline{g}^{(t)} - g^*||_\infty \leq 2\overline{\kappa}^t ||f^{(0)} - f^*||_*. \tag{2.34}$$

Algorithm 2 TI-Sinkhorn algorithm for Problem (2.22).

Input: cost matrix $C \in \mathbb{R}^{m\times n}$, measures $\mu \in \mathbb{R}^m_{>0}, \nu \in \mathbb{R}^n_{>0}, \gamma \in \mathbb{R}^{m\times n}_{>0}$, regularization $\varepsilon > 0$, relaxation parameters $\rho_1, \rho_2 > 0$.
Calculate $k_{ij} = \frac{\varepsilon}{\varepsilon+\rho_i}\frac{\rho_j}{\rho_1+\rho_2}$, for $i, j \in \{1, 2\}$, and $\xi_{ij} = \frac{k_{ij}}{1-k_{ij}}$, for $i \neq j$.
for $t = 1, \dots, T$ **do**
  $f_i \quad \frac{\rho_1}{\rho_1+\varepsilon}\overline{\mathrm{Smin}}^{\mu_i,\gamma_{i,\cdot}}_{\varepsilon}(C_{i,\cdot} - g^{(t)}) - k_{11}\,\mathrm{Smin}^{\nu}_{\rho_2}(g^{(t)})$, for $i \in [m]$.
  $f^{(t+1)} \leftarrow f + \xi_{12}\,\mathrm{Smin}^{\mu}_{\rho_1}(f)$.
  $g_j \quad \frac{\rho_2}{\rho_2+\varepsilon}\overline{\mathrm{Smin}}^{\nu_j,\gamma_{\cdot,j}}_{\varepsilon}(C_{\cdot,j} - f^{(t+1)}) - k_{22}\,\mathrm{Smin}^{\mu}_{\rho_1}(f^{(t+1)})$, for $j \in [n]$.
  $g^{(t+1)} \leftarrow g + \xi_{21}\,\mathrm{Smin}^{\nu}_{\rho_2}(g)$.
end for
Calculate $\lambda^* = \lambda^*(f^{(T)}, g^{(T)})$ using Equation (2.31).
Output: pair of dual vectors $(f^{(T)} + \lambda^*, g^{(T)} - \lambda^*)$.

Algorithm 3 Variant of TI-Sinkhorn algorithm for Problem (2.22).

Input: cost matrix $C \in \mathbb{R}^{m\times n}$, measures $\mu \in \mathbb{R}^m_{>0}, \nu \in \mathbb{R}^n_{>0}, \gamma \in \mathbb{R}^{m\times n}_{>0}$, regularization $\varepsilon > 0$, relaxation parameters $\rho_1, \rho_2 > 0$.
for $t = 1, \dots, T$ **do**
  $f_i^{(t+1)} \quad \frac{\rho_1}{\rho_1+\varepsilon}\overline{\mathrm{Smin}}^{\mu_i,\gamma_{i,\cdot}}_{\varepsilon}(C_{i,\cdot} - g^{(t)}) - \lambda^{(t)}$, for $i \in [m]$.
  $g_j^{(t+1)} \quad \frac{\rho_2}{\rho_2+\varepsilon}\overline{\mathrm{Smin}}^{\nu_j,\gamma_{\cdot,j}}_{\varepsilon}(C_{\cdot,j} - f^{(t+1)}) + \lambda^{(t)}$, for $j \in [n]$.
  $\lambda^{(t+1)} = \lambda^*(f^{(t+1)}, g^{(t+1)})$ using Equation (2.31).
end for
Output: pair of dual vectors $(f^{(T)}, g^{(T)})$.

In other words, TI-Sinkhorn improves the convergence rate of Sinkhorn by a factor of $\kappa_\varepsilon(\mu)\kappa_\varepsilon(\nu)$. However, when $\varepsilon \ll \rho$, not only $\kappa \approx 1$, but also empirically $\kappa_\varepsilon(\mu)\kappa_\varepsilon(\nu) \approx 1$. For this reason, despite the acceleration over Sinkhorn, TI-Sinkhorn still suffers from small regularization and remains slow in such situation.

Séjourné, Vialard, and Peyré (2021a) also propose a variant of TI-Sinkhorn, whose details can be found in Algorithm 3. The idea is to directly apply alternative optimization scheme to the problem $\sup_{f,g,\lambda} F_\varepsilon(f + \lambda, g - \lambda)$, which guarantees the convergence to a stationary point (P.Tseng, 2001). In practice, we observe that both TI algorithms work comparatively well.

Majorization-minimization algorithm Interestingly, Problem (2.22) can also be reformulated as a nonnegative penalized linear regression problem (Chapel et al., 2021), whose objective function comprises of a linear term and a KL divergence as penalization. Moreover, since the KL divergence is a Bregman divergence, one can apply the majorization-minimization (MM) approach (see, for example (Hunter and Lange, 2004; Sun, Babu, and Palomar, 2017)) and obtain a closed-form update of the transport plan, without the need of invoking the dual vectors to reconstruct the coupling, as in the Sinkhorn-based methods. Following (Chapel et al., 2021), the

MM iterate of Problem (2.22) reads

$$P^{(t+1)} = \left[\left(\frac{\mu}{P^{(t)}_{\#1}}\right)^{\lambda_1} \otimes \left(\frac{\nu}{P^{(t)}_{\#2}}\right)^{\lambda_2}\right] \odot (P^{(t)})^{\lambda_1+\lambda_2} \odot \gamma^r \odot \exp\left(-\frac{C}{\rho}\right) \tag{2.35}$$

$$= \frac{(P^{(t)})^{\lambda_1+\lambda_2}}{(P^{(t)}_{\#1})^{\lambda_1} \otimes (P^{(t)}_{\#2})^{\lambda_2}} \odot \left(\mu^{\lambda_1} \otimes \nu^{\lambda_2}\right) \odot \gamma^r \odot \exp\left(-\frac{C}{\rho}\right), \tag{2.36}$$

where $\rho = \rho_1 + \rho_2 + \varepsilon$ and $\lambda_i = \frac{\rho_i}{\rho}$ and $r = \frac{\varepsilon}{\rho}$. Here, the division and exponential operations are element-wise. We also remark that, the unregularized UOT ($\varepsilon = 0$) is naturally handled by MM algorithm, whereas the Sinkhorn-based algorithms are only applicable to the regularized UOT. However, unlike its competitors, the MM algorithm does not work in the balanced ($\rho_1 = \rho_2 = \infty$) and semi-relaxed (either $\rho_1 = \infty$ or $\rho_2 = \infty$) OT settings.

Another example of the Bregman divergence is the squared l_2-norm, where $D_\varphi(p, q) := \frac{||p-q||^2}{2}$, with $\varphi(x) = \frac{||x||^2}{2}$. Interestingly, the MM iterates can also be computed explicitly in this case.

Corollary 2.1.2 (Generalization of Equation 7 in (Chapel et al., 2021))**.** *For $\mu \in \mathbb{R}^m, \nu \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}^{m\times n}$, consider the problem*

$$UOT_{\varepsilon,\rho} = \min_{P \in \mathbb{R}^{m\times n}_{\geq 0}} \langle C, P\rangle + \rho_1 \frac{||P_{\#1} - \mu||^2}{2} + \rho_2 \frac{||P_{\#2} - \nu||^2}{2} + \varepsilon \frac{||P - \gamma||^2}{2}. \tag{2.37}$$

Then, the update reads

$$P^{(t+1)} = \max\left(0, (\rho_1\mu) \oplus (\rho_2\nu) + \varepsilon\gamma - C\right) \odot \frac{P^{(t)}}{(\rho_1 P^{(t)}_{\#1}) \oplus (\rho_2 P^{(t)}_{\#2}) + \varepsilon P^{(t)}}, \tag{2.38}$$

where the max and division operations are element-wise.

There are two interesting features of the squared l_2-norm. First, thanks to the max operation, the iterate matrix can be made sparse from the very first iteration, which further accelerates the matrix operations. Second, thanks to Corollary 2.1.1, the minimum of Problem (2.37) can be expressed as a quadratic function of its minimizer.

Corollary 2.1.3. *Denote P^* the solution of Problem (2.37). Then*

$$UOT_{\varepsilon,\rho} = \frac{1}{2}\Big(\rho_1||\mu||^2 + \rho_2||\nu||^2 + \varepsilon||\gamma||^2\Big) - \frac{1}{2}\Big(\rho_1||P^*_{\#1}||^2\rangle + \rho_2||P^*_{\#2}||^2\rangle + \varepsilon||P^*||^2\rangle\Big). \tag{2.39}$$

To conclude, MM is a very appealing alternative to the Sinkhorn-based algorithms, especially when the regularization is too small. However, when the relaxation is too large relative to the regularization and the magnitude of the cost matrix, MM may converge very slowly if the initialization is not carefully chosen. To see this, suppose that $\rho_1 \gg ||C||_\infty$ and $\rho_1 \gg \varepsilon$. Then,

$\exp(-C/\lambda) \approx 1_{m\times n}$ and $r \approx 0$. So,

$$P^{(t+1)} \approx \left[\left(\frac{\mu}{P^{(t)}_{\#1}}\right)^{\lambda_1} \otimes \left(\frac{\nu}{P^{(t)}_{\#2}}\right)^{\lambda_2}\right] \odot P^{(t)}. \tag{2.40}$$

If one initializes with the solution of the balanced OT problem, *i.e.*, $P^{(0)}_{\#1} = \mu$ and $P^{(0)}_{\#2} = \nu$, then $P^{(1)} \approx P^{(0)}$. By induction, we obtain $P^{(t+1)} \approx P^{(t)}$, meaning that the convergence is very slow.

Regarding the case of squared l_2-norm, one needs to carefully choose the hyperparameters, so that the matrix $\max\left(0, (\rho_1\mu) \oplus (\rho_2\nu) + \varepsilon\gamma - C\right)$ is neither too dense (thus fails to achieve sparsity), nor too sparse (thus may incur division error). For this reason, the tuning may be quite troublesome.

Discussion We can consider a more general UOT formulation: given a cost matrix $C \in \mathbb{R}^{m\times n}$, the positive measures $\mu \in \mathbb{R}^m_{>0}, \nu \in \mathbb{R}^n_{>0}$ and $\gamma \in \mathbb{R}^{m\times n}_{>0}$, the parameters $\rho_1, \rho_2, \varepsilon \geq 0$, we want to solve

$$\min_{P\in\mathbb{R}^{m\times n}_{\geq 0}} \langle C, P\rangle + \rho_1 D(P_{\#1}|\mu) + \rho_2 D(P_{\#2}|\nu) + \varepsilon R(P), \tag{2.41}$$

where D is a certain divergence (for example, Csiszár or Bregman divergence) and R is a regularizer. Table 2.2 summarizes some choices of divergences and regularizers, with the corresponding related works.

Divergence	**Regularizer**	**Reference**
Csiszár divergence	N/A	(Liero, Mielke, and Savaré, 2018) (Chizat et al., 2018b)
Csiszár divergence	Negative entropy	(Frogner et al., 2015) (Chizat et al., 2018a) (Lee, Bertrand, and Rozell, 2019)
Csiszár divergence	KL divergence	(Séjourné et al., 2019)
KL divergence	Square 2-norm	(Nguyen et al., 2022)
Maximum mean discrepancy	N/A	(Manupriya, Nath, and Jawanpuria, 2023)
Bregman divergence	Bregman divergence	(Chapel et al., 2021)
Smooth divergence	N/A	(Blondel, Seguy, and Rolet, 2018)
Convex conjugate of co-finite Bregman function	N/A	(Sonthalia and Gilbert, 2020)

Table 2.2 – Summary of choices of divergences and regularizers for Problem (2.41). The case N/A (not available) corresponds to the unregularized problem (*i.e.*, $\varepsilon = 0$).

2.2 To Gromov-Wasserstein distance and beyond

2.2.1 Problem statement

When the source and target data do not live in the same underlying space, it is no longer possible to compute the inter-space distance. Such situations are ubiquitous in practice. For example, images in source and target domains may have different resolutions. Even if they are vectorized by the last layers of neural networks, it is not unusual to have the embeddings of different dimensions. Even worse, in graph, such vectorial embedding of the node does not even exist (though it can be learned), and only within-domain similarity matrix is available.

In this section, we present a metric which can be used to compare incomparable spaces, termed **Gromov-Wasserstein** (GW) distance introduced by Mémoli (2007, 2011b) [13]. It allows to handle a more general object called *metric-measure* space, instead of just probability measure, as in the Wasserstein distance. The GW distance is not only theoretically sound (Mémoli, 2011b; Sturm, 2012), but also computationally tractable. For this reason, it has been used extensively in practice, for example in graph (Chowdhury and Needham, 2021; Vayer et al., 2019a; Vincent-Cuaz et al., 2021; Xu, Luo, and Carin, 2019; Xu et al., 2019), computational biology (Demetci et al., 2020), comparing kernel matrices (Peyré, Cuturi, and Solomon, 2016), correspondance alignment (Solomon et al., 2016), heterogeneous domain adaptation (Yan et al., 2018), or machine translation (Alvarez-Melis and Jaakkola, 2018), clustering and data visualization (Ryner and Karlsson, 2022).

While GW distance is of major interest in this thesis, as a byproduct, we also discuss other related metrics on the space of incomparable spaces. This is beneficial, not only for the understanding of the underlying motivation, but also for a better grasp of a bigger picture on this research direction.

2.2.2 Motivation

A very natural idea to compare incomparable spaces is to appropriately transform them, so that their resulting embeddings are comparable. Typically, one can project one onto the other (known as *extrinsic* comparison), or project both onto a sufficiently rich, common metric space (known as *intrinsic* comparison). The choice of transformation is also of crucial importance. For distance-based methods and applications, especially when working with images or 3D-objects, the projection should preserve as much geometric information amongst objects as possible. Moreover, it should be able to handle the usual invariants exhibited in the data, for example, rotation, translation, reflection. To this extent, we are interested in the class of distance-preserve

13. Historically, the name "Gromov-Wasserstein distance" has several meanings. In the terrific book of Villani (2009), it is a synonym for the L_p-transportation distance. In (Mémoli, 2011a), Facundo Mémoli used it to refer to two other distances, including the now-standard one, which is popularized by the work of the same author (Mémoli, 2011b).

transformations, which will show great interest in the study of metrics in the space of incomparable spaces.

Definition 2.2.1 (Metric measure space (Gromov, 1999)). *A metric measure space (mm-space) $\mathcal{X}$ is a triplet (X, d_X, μ_X), where*

- *(X, d_X) is a compact metric space.*
- *μ_X is a Borel probability measure and has full support, that is $supp(\mu_X) = X$.*

In other words, we focus on essentially the same objects as in the Wasserstein distance, except that the metric is no longer shared across spaces but must be specified for each individual space. As we will see, the relation between Wasserstein and Hausdorff distances presented in Section 2.1.1 becomes very handy as it gives us a natural extension to the setting of mm-space. First, we introduce the Gromov-Hausdorff (GH) distance proposed by Gromov (1981). While we are mostly inspired by Chapter 4 in (Mémoli, 2011b), we perform a more thorough literature review, including the most recent development and applications of the GH-based distances in the OT and machine learning literature.

Definition 2.2.2 (Admissible distance). *Given two compact metric spaces (X, d_X) and (Y, d_Y), we denote $\mathcal{D}(d_X, d_Y)$ the set of all pseudo-metrics on the disjoint union $X \cup Y$, such that for any $d \in \mathcal{D}(d_X, d_Y)$, we have $d = d_X$ on X^2, and $d = d_Y$ on Y^2. Any metric in $\mathcal{D}(d_X, d_Y)$ is called an admissible distance on $X \cup Y$.*

The GH distance between two arbitrary compact metric spaces (X, d_X) and (Y, d_Y) (not necessarily living in the same underlying space) is defined by

$$\mathrm{GH}((X, d_X), (Y, d_Y)) := \inf_{d \in \mathcal{D}(d_X, d_Y)} d_H^{(X \cup Y, d)}(X, Y), \tag{2.42}$$

The idea of the GH distance is to, first, construct a "union" metric space as a common underlying space, in which the Hausdorff distance between two component metric spaces can be calculated. Then, amongst these admissible "union" metric spaces, identify the one which corresponds to the smallest Hausdorff distance. It is well known that the GH distance defines a metric on the space of compact metric spaces, up to isometry (Burago, Burago, and Ivanov, 2001; Gromov, 1999). Interestingly, it also admits many equivalent reformulations and we will see that each of them gives rise to different distances in the space of mm-spaces, including the aforementioned GW distance. Let us first introduce the following useful notions.

Definition 2.2.3 (Pullback). *Given four sets X_1, X_2, Y_1 and Y_2, and three functions $c_X : X_1 \times X_2 \to \mathbb{R}, \varphi_1 : Y_1 \to X_1$ and $\varphi_2 : Y_2 \to X_2$, the pullback of c_X by (φ_1, φ_2) is the function $(\varphi_1, \varphi_2)^* c_X : Y_1 \times Y_2 \to \mathbb{R}$ defined by $(\varphi_1, \varphi_2)^* c_X(y_1, y_2) := c_X(\varphi_1(y_1), \varphi_2(y_2))$. If $\varphi_1 = \varphi_2 = \varphi$ (then $Y_1 = Y_2$ and $X_1 = X_2$), then we simply write $\varphi^* c_X := (\varphi, \varphi)^* c_X$.*

Definition 2.2.4 (Isometric embedding)**.** *Given two metric spaces (X, d_X) and (Y, d_Y), the map $\varphi : X \to Y$ is an isometric embedding if and only if $d_X = \varphi^* d_Y$ on X^2. An isometry is a surjective isometric embedding (i.e., $\varphi(X) = Y$).*

Clearly, any isometric embedding is necessarily injective, so an isometry is bijective. Moreover, it can be shown that $\varphi^* d_Y$ defines a metric if and only if φ is injective. By relaxing the strict preservation of distance, where $d_X = O(\varphi^* d_Y)$, one obtains the *approximate embedding* (Matousek, 2013). This object has already been studied since the 80's, whose notable examples include the famous Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984) and Bourgain's embedding theorem (Bourgain, 1985).

Now, we will briefly discuss some distances betweeen mm-spaces, whose origins come from different reformulations of GH distance.

L_p-transportation distance By combining Formulation (2.42) and Equation (2.6), we have,

$$\mathrm{GH}((X, d_X), (Y, d_Y)) = \inf_{d,R} \sup_{(x,y)\in R} d(x, y) = \inf_{d,R} ||d||_{L^\infty(R)}, \tag{2.43}$$

where the infimum is taken over all $R \in \mathcal{R}(X, Y)$ and $d \in \mathcal{D}(d_X, d_Y)$. Now, we can immediately extend Formulation (2.43) to the mm-space setting, using the Wassersteinization and L^p-ization processes in Figure 2.2. This results in the L_p-transportation distance (Sturm, 2006): for $1 \leq p < \infty$,

$$D_p(\mathcal{X}, \mathcal{Y}) := \inf_{d,\pi} \left(\int_{X\times Y} d(x, y)^p d\pi(x, y) \right)^{1/p} = \inf_{d,\pi} ||d||_{L^p(\pi)}, \tag{2.44}$$

and for $p = \infty$,

$$D_\infty(\mathcal{X}, \mathcal{Y}) := \inf_{d,\pi} \sup_{(x,y)\in R(\pi)} d(x, y) = \inf_{d,\pi} ||d||_{L^\infty(R(\pi))}, \tag{2.45}$$

where the infimum is taken over all $\pi \in U(\mu_X, \mu_Y)$ and $d \in \mathcal{D}(d_X, d_Y)$. This distance can also be obtained from a more popular definition of the GH distance, also due to Gromov (1999)

$$\mathrm{GH}((X, d_X), (Y, d_Y)) = \inf_{Z,f,g} d_H^Z(f(X), g(Y)), \tag{2.46}$$

where the infimum is taken over all metric spaces (Z, d) and isometric embeddings $f : X \to Z$ and $g : Y \to Z$. It can be shown that Formulations (2.42) and (2.46) are equivalent (see Appendix 8.1.2). Intuitively, Formulation (2.46) implies that the GH distance can be alternatively obtained by projecting the two original metric spaces onto a common metric one, where the Hausdorff distance between the two embeddings can be computed. Moreover, each projection map must

also preserve the within-space distance.

Thanks to Lemma 3.3 in (Sturm, 2006), we can adapt the idea of Formulation (2.46) to the setting of mm-space by replacing the Hausdorff distance with the Wasserstein distance, and obtain the equivalent form of the L_p-transportation distance

$$D_p(\mathcal{X}, \mathcal{Y}) = \inf_{Z,f,g} W_{Z,p}((f(X), f_{\#\mu_X}), (g(Y), g_{\#\mu_Y})), \tag{2.47}$$

for $1 \leq p < \infty$ and

$$D_\infty(\mathcal{X}, \mathcal{Y}) = \inf_{Z,f,g} W_{Z,\infty}((f(X), f_{\#\mu_X}), (g(Y), g_{\#\mu_Y})), \tag{2.48}$$

where the infimum is taken over all metric spaces (Z, d) with measurable isometric embeddings $f : X \to Z$ and $g : Y \to Z$.

To the best of our knowledge, despite the well-established theory, the L_p-transportation distance has not yet found any applications in practice, due to the computational intractability of the optimization problem that it encodes. In the OT and machine learning literature, there exist other works which also rely on this idea of projecting onto a common space. For example, Alaya et al. (2022) study a variant of Formulation (2.47) called *Sub-Embedding Robust Wasserstein*, or Paty and Cuturi (2019) propose the subspace robust Wasserstein projection, which projects onto a Grassmannian manifold, or Cai and Lim (2022) use orthogonal transformations to project one space onto the other.

Gromov-Wasserstein distance For notational convenience, given two metric spaces (X, d_X) and (Y, d_Y), we define the function $|d_X - d_Y| : (X \times Y) \times (X \times Y) \to \mathbb{R}_{\geq 0}$ by

$$|d_X - d_Y|((x_1, y_1), (x_2, y_2)) := \big|d_X(x_1, x_2) - d_Y(y_1, y_2)\big|. \tag{2.49}$$

Now, due to Theorem 7.3.25 in (Burago, Burago, and Ivanov, 2001),

$$\mathrm{GH}((X, d_X), (Y, d_Y)) = \frac{1}{2} \inf_R \sup_{\substack{(x_1,y_1)\in R \\ (x_2,y_2)\in R}} \big|d_X(x_1, x_2) - d_Y(y_1, y_2)\big| \tag{2.50}$$

$$= \frac{1}{2} \inf_R ||d_X - d_Y||_{L^\infty(R\times R)}, \tag{2.51}$$

where the infimum is taken over all $R \in \mathcal{R}(X, Y)$. By replacing the correspondance with the admissible coupling, we obtain the GW distance (Mémoli, 2007, 2011b): for $1 \leq p < \infty$,

$$\mathrm{GW}_p(\mathcal{X}, \mathcal{Y}) := \inf_{\pi \in U(\mu_X, \mu_Y)} ||d_X - d_Y||_{L^p(\pi \otimes \pi)} \tag{2.52}$$

$$= \inf_{\pi \in U(\mu_X, \mu_Y)} \left(\iint |d_X(x_1, x_2) - d_Y(y_1, y_2)|^p \, d\pi(x_1, y_1) \, d\pi(x_2, y_2) \right)^{1/p}, \tag{2.53}$$

and for $p = \infty$,

$$\mathrm{GW}_\infty(\mathcal{X}, \mathcal{Y}) := \inf_{\pi \in U(\mu_X, \mu_Y)} ||d_X - d_Y||_{L^\infty(R(\pi) \times R(\pi))} \tag{2.54}$$

$$= \inf_{\pi \in U(\mu_X, \mu_Y)} \sup_{\substack{(x_1, y_1) \in R(\pi) \\ (x_2, y_2) \in R(\pi)}} \big| d_X(x_1, x_2) - d_Y(y_1, y_2) \big|. \tag{2.55}$$

Thanks to Theorem 5.1 in (Mémoli, 2011a), we have the following relation between the GW and L_p-transportation distance: $\mathrm{GW}_\infty = D_\infty$ and $\mathrm{GW}_p \leq D_p$, for any $p \geq 1$. While bearing some similarity with Formulation (2.43), the optimization problem in GW distance is much more feasibility to solve in practice. In particular, the maximization over the set of admissible distances in Formulation (2.43) incurs an additional variable, which is difficult to optimize.

Bi-directional Gromov-Monge Thanks to Theorem 2.1 in (Kalton and Ostrovskii, 1999), for bounded metric spaces, we have

$$\mathrm{GH}((X, d_X), (Y, d_Y)) := \frac{1}{2} \inf_{\substack{f: X \to Y \\ g: Y \to X}} \sup_{\substack{(x_1, x_2) \in X^2 \\ (y_1, y_2) \in Y^2 \\ (x_i, y_i) \in G(f, g)}} \big| d_X(x_1, x_2) - d_Y(y_1, y_2) \big| \tag{2.56}$$

$$= \frac{1}{2} \inf_{\substack{f: X \to Y \\ g: Y \to X}} \max \Big\{ \Delta_\infty(f, X), \Delta_\infty(g, Y), \Delta_\infty(f, g, X, Y) \Big\}. \tag{2.57}$$

Here, the infimum is taken over all **arbitrary** maps f and g, and

- The union of graphs is defined by: $G(f, g) := \{(x, f(x)), x \in X\} \cup \{(g(y), y), y \in Y\}$.
- The first term is defined by

$$\Delta_\infty(f, X) := \sup_{(x_1, x_2) \in X^2} |d_X(x_1, x_2) - d_Y(f(x_1), f(x_2))| \tag{2.58}$$

$$= ||d_X - f^* d_Y||_{L^\infty(X^2)}. \tag{2.59}$$

- The second term is defined by

$$\Delta_\infty(g, Y) := \sup_{(y_1,y_2)\in Y^2} |d_X(g(y_1), g(y_2)) - d_Y(y_1, y_2)| \tag{2.60}$$

$$= ||g^* d_X - d_Y||_{L^\infty(Y^2)}. \tag{2.61}$$

- The last term is defined by

$$\Delta_\infty(f, g, X, Y) := \sup_{(x,y)\in X\times Y} |d_X(x, g(y)) - d_Y(f(x), y)| \tag{2.62}$$

$$= ||(\mathrm{Id}_X, g)^* d_X - (f, \mathrm{Id}_Y)^* d_Y||_{L^\infty(X\times Y)}. \tag{2.63}$$

The second equality holds, due to Remark 2 in (Mémoli and Sapiro, 2005). Interested readers can find more discussion on the practical use of Formulation (2.56) in Remark 4.5 in (Mémoli, 2011a). Recently, Zhang et al. (2022b) adapt this fourth formulation to the mm-space and propose the bi-directional Gromov-Monge (BGM) distance defined as

$$\mathrm{BGM}_p(\mathcal{X}, \mathcal{Y}) = \inf_{\substack{f\in\mathcal{T}(\mu_X,\mu_Y)\\ g\in\mathcal{T}(\mu_Y,\mu_X)}} \Delta_p(f, X) + \Delta_p(g, Y) + \Delta_p(f, g, X, Y), \tag{2.64}$$

where

- $\Delta_p(f, X) = ||d_X - f^* d_Y||_{L^p(X^2,\mu_X\otimes\mu_X)}$.
- $\Delta_p(g, Y) = ||g^* d_X - d_Y||_{L^p(Y^2,\mu_Y\otimes\mu_Y)}$.
- $\Delta_p(f, g, X, Y) = ||(\mathrm{Id}_X, g)^* d_X - (f, \mathrm{Id}_Y)^* d_Y||_{L^p(X\times Y,\mu_X\otimes\mu_Y)}$.

Since this formulation requires learning two transport maps which push from one distribution to the other, it shares some similarity with the cycle generative adversarial networks (Kim et al., 2017; Zhu et al., 2017), where one trains two transport maps which are roughly inverse to each other. Another concurrent variant of BGM is proposed by Hur, Guo, and Liang (2021), where they consider the reversible Gromov-Monge (RGM) distance

$$\mathrm{RGM}_2(\mathcal{X}, \mathcal{Y}) = \inf_{(f,g)\in I(\mu_X,\mu_Y)} \Delta_2(f, g, X, Y), \tag{2.65}$$

where $I(\mu_X, \mu_Y) := \{(f, g) : (g, \mathrm{Id}_Y)_\# \mu_Y = (\mathrm{Id}_X, f)_\# \mu_X\}$. We note that the constraints in RGM and BGM are related but not equivalent. In particular, as remarked in (Hur, Guo, and Liang, 2021), the equality $(g, \mathrm{Id}_Y)_\# \mu_Y = (\mathrm{Id}_X, f)_\# \mu_X$ implies that $f_\# \mu_X = \mu_Y$ and $g_\# \mu_Y = \mu_X$, but the reverse does not hold. In case of measure networks (Chowdhury and Mémoli, 2019), Proposition 1 in (Hur, Guo, and Liang, 2021) asserts that $\mathrm{RGM}_p \geq \mathrm{GW}_p$, for every $p \geq 1$.

To conclude, apart from the above metrics induced by the GH distance, there exist other alternatives for comparing mm-spaces. For example, by relying on the idea of Formulation (2.46),

one can use the Gromov-Hausdorff-Prokhorov distance (Villani, 2009) and its variants (Abraham, Delmas, and Hoscheit, 2013; Miermont, 2009)

$$\mathrm{GHP}(\mathcal{X}, \mathcal{Y}) = \inf_{Z,f,g} d_H^Z(f(X), g(Y)) + d_P(f_{\#}\mu_X, g_{\#}\mu_Y), \tag{2.66}$$

where d_P is the Prokhorov distance. The Prokhorov distance can also be replaced by, for example by the Wasserstein distance, which results in the Gromov-Hausdorff-Wasserstein distance (Villani, 2009). However, from the practical perspective, amongst all competitors, the GW distance still remains the most popular and attracting, given its well-established theory, computational feasibility, strong performance and handy OT plan.

2.2.3 Properties of GW distance

Isomorphism between mm-spaces Isomorphism is a central concept in the study of GW distance between the mm-spaces. Two mm-spaces are *isomorphic* if and only if there exists a measure-preserving isometry from one mm-space to the other. Notable examples of isomorphism include orthogonal transformations, for example rotation, translation and relection. It can be shown that the GW distance defines a metric on the space of mm-spaces, up to isomorphism (Mémoli, 2007; Sturm, 2012). Note that, this metric property still holds for other GH-based metrics, namely the RGM and L_p-transport distances (Hur, Guo, and Liang, 2021; Sturm, 2006). Unlike the GW distance, the Wasserstein distance is not invariant to orthogonal transformations. However, in Euclidean space, by additionally searching over the space of linear operators with bounded Schatten l_2-norm, one can recover the GW from the Wasserstein distance (Alvarez-Melis, Jegelka, and Jaakkola, 2019).

Isomorphism between measure networks By definition, the mm-space is a separable metric space. In particular, the distance is a measurable function. In practice, the distance matrix may represent the pairwise dissimilarity rather than similarity, or not even necessarily be symmetric, typically in a directed graph. Thus, the notion of distance in the mm-space can be relaxed and replaced with an appropriate measurable function, which results in a more generalized space, for example *measure network* (Chowdhury and Mémoli, 2019), or *gauged measure space* (Sturm, 2012). In the sequel, we focus on the *measure network* (Chowdhury and Mémoli, 2019), or simply *network*, if there is no risk of ambiguity.

Definition 2.2.5 (Measure network)**.** *A network is a triplet $\mathcal{X} = (X, c_X, \mu_X)$, where X is a Polish space equipped with the Borel probability measure μ_X and c_X is a bounded measurable function on X^2.*

We also assume that μ_X has full support, *i.e.*, $\mathrm{supp}(\mu_X) = X$. Unlike the mm-space, to characterize the equivalence class of networks, we need to introduce some forms of isomorphism.

First, let us start with a useful concept in (Mémoli and Needham, 2022b).

Definition 2.2.6 (Mass splitting)**.** *A network $\mathcal{Z} = (Z, c_Z, \mu_Z)$ is a **mass splitting** of a network $\mathcal{X} = (X, c_X, \mu_X)$ if there exists a measure-preserving map $\varphi : Z \to X$ such that the pullback equality $c_X = \varphi^* c_Z$ holds $\mu_Z \otimes \mu_Z$-almost everywhere. We denote MS($\mathcal{X}$) the set of all mass splittings of $\mathcal{X}$.*

Clearly, $\mathcal{X} \in \text{MS}(\mathcal{X})$, so MS($\mathcal{X}$) is always non-empty. Now, we define three forms of isomorphism.

Definition 2.2.7 (Isomorphism)**.** *Two measure networks are*

1. *strongly isomorphic if there exists a **bijective** measure-preserving map from one network to the other such that the pullback equality holds **everywhere**.*
2. *semi-strongly isomorphic if one is the mass splitting of the other and vice versa.*
3. *weakly isomorphic if they have a common mass splitting.*

There are several immediate observations from this definition.

1. Definition 2.2.7 can be easily adapted to the mm-space setting. In this case, by Lemma 1.10 in (Sturm, 2012), the three forms are equivalent.
2. The strong isomorphism is a very strict condition. First, it is not difficult to see that strong isomorphism implies semi-strong isomorphism and semi-strong isomorphism implies weak isomorphism. By Remark 2 in (Chowdhury and Mémoli, 2019), the reverse may not hold in general. Second, it requires the pullback equality to hold **everywhere**, rather than only almost-everywhere.
3. To the best of our knowledge, the semi-strong isomorphism has never been discussed in the literature. However, it can be useful to characterize the Gromov-Monge distance.

Similar to the Wasserstein distance, a Monge version of the GW distance, known as *Gromov-Monge* (GM) distance, is defined as

$$\text{GM}(\mathcal{X}, \mathcal{Y}) = \inf_{\varphi \in \mathcal{T}(\mu_X, \mu_Y)} ||c_X - \varphi^* c_Y||_{L^p(X^2, \mu_X \otimes \mu_X)}, \tag{2.67}$$

where, recall that $\mathcal{T}(\mu, \nu)$ is the set of transport maps from μ to ν. The GM distance suffers the same limitations as the Monge's problem, notably the existence of the transport map and the asymmetry of the metric. It is easy to see that $\mathcal{Z}$ is a mass splitting of $\mathcal{X}$ if and only if $\text{GM}(\mathcal{Z}, \mathcal{X}) = 0$. The equality between GM and GW distances has also attracted much interest in the mathematics community, see (Mémoli and Needham, 2022a) for a comprehensive and up-to-date review.

The following result is very handy as it allows us to flexibly switch between the GW and GM distances.

Corollary 2.2.1 (Theorem 14 in (Mémoli and Needham, 2022b)). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two measure networks, then*

$$GW(\mathcal{X},\mathcal{Y}) = \inf_{\mathcal{Z}\in MS(\mathcal{X})} GM(\mathcal{Z},\mathcal{Y}) = \inf_{\mathcal{Z}\in MS(\mathcal{Y})} GM(\mathcal{Z},\mathcal{X}). \tag{2.68}$$

Moreover, the infima are always attained: there exist two measure networks $\mathcal{Z}_x \in MS(\mathcal{X})$ and $\mathcal{Z}_y \in MS(\mathcal{Y})$ such that $GW(\mathcal{X},\mathcal{Y}) = GM(\mathcal{Z}_x,\mathcal{Y}) = GM(\mathcal{Z}_y,\mathcal{X})$.

Proposition 2.2.1 (Relations of GW and GM with isomorphism). *Let $\mathcal{X}$ and $\mathcal{Y}$ be two measure networks.*

1. *$\mathcal{X}$ and $\mathcal{Y}$ are weakly isomorphic if and only if $GW(\mathcal{X},\mathcal{Y}) = 0$.*
2. *$\mathcal{X}$ and $\mathcal{Y}$ are semi-strongly isomorphic if and only if $GM(\mathcal{X},\mathcal{Y}) = GM(\mathcal{Y},\mathcal{X}) = 0$.*

Proof. If $\mathcal{X}$ and $\mathcal{Y}$ are weakly isormorphic, then there exists a common mass splitting $\mathcal{Z}$ of $\mathcal{X}$ and $\mathcal{Y}$, which means $\mathrm{GM}(\mathcal{Z},\mathcal{Y}) = 0$. By Corollary 2.2.1, $\mathrm{GW}(\mathcal{X},\mathcal{Y}) = \mathrm{GM}(\mathcal{Z},\mathcal{Y}) = 0$. Conversely, if $\mathrm{GW}(\mathcal{X},\mathcal{Y}) = 0$, then by Corollary 2.2.1, there exists $\mathcal{Z} \in \mathrm{MS}(\mathcal{X})$ such that $\mathrm{GM}(\mathcal{Z},\mathcal{Y}) = 0$. But this means $\mathcal{Z} \in \mathrm{MS}(\mathcal{Y})$. We conclude that $\mathcal{X}$ and $\mathcal{Y}$ are weakly isormorphic since $\mathcal{Z}$ is the common mass splitting. The equivalence between semi-strong isormorphism and GM follows immediately from the definition. ■

Proposition 2.2.1 allows us to characterize the equivalence classes of GW distance in the setting of networks. More precisely,

Proposition 2.2.2 (Theorems 2.3 and 2.4 in (Chowdhury and Mémoli, 2019)). *GW distance defines a metric on the space of networks, up to weak isomorphism.*

Corollary 2.2.2 (Relation between weak isomorphism and mass splitting).

1. *If $\mathcal{Z}$ is a mass splitting of $\mathcal{X}$, then $\mathcal{X}$ and $\mathcal{Z}$ are weakly isomorphic (since $\mathcal{Z}$ is the common mass splitting).*
2. *If two networks are weakly isomorphic, then so are their mass splittings.*

Proof. If $\mathcal{Z}$ is a mass splitting of $\mathcal{X}$, then $\mathcal{Z}$ is the common mass splitting. So, $\mathcal{X}$ and $\mathcal{Z}$ are weakly isomorphic.

If $\mathcal{X}$ and $\mathcal{Y}$ are weakly isomorphic, then they have a common mass splitting $\mathcal{Z}$. By Proposition 2.2.1, we have $\mathrm{GW}(\mathcal{X},\mathcal{Z}) = \mathrm{GW}(\mathcal{Y},\mathcal{Z}) = 0$. Now, suppose $\mathcal{X}' \in \mathrm{MS}(\mathcal{X})$ and $\mathcal{Y}' \in \mathrm{MS}(\mathcal{Y})$, then $\mathrm{GW}(\mathcal{X},\mathcal{X}') = \mathrm{GW}(\mathcal{Y},\mathcal{Y}') = 0$. By triangle inequality, we deduce that $\mathrm{GW}(\mathcal{X}',\mathcal{Z}) = \mathrm{GW}(\mathcal{Y}',\mathcal{Z}) = 0$. So, $\mathcal{Z}$ is the common mass splitting of $\mathcal{X}'$ and $\mathcal{Y}'$, meaning that they are weakly isomorphic. ■

2.2.4 Optimization and algorithm

Given two matrices (not necessarily symmetric) $C^x \in \mathbb{R}^{m\times m}, C^y \in \mathbb{R}^{n\times n}$ and two histograms $\mu_X \in \Delta_m, \mu_Y \in \Delta_n$, the discrete GW problem reads

$$\min_{P\in U(\mu,\nu)} \sum_{i,j,k,l} |C^x_{ik} - C^y_{jl}|^p \; P_{ij}P_{kl}. \tag{2.69}$$

For convenience, we write $\sum_{i,j,k,l} |C^x_{ik} - C^y_{jl}|^p \; P_{ij}P_{kl} = \langle C \otimes P, P\rangle$, where the 4D-tensor cost $C = |C^x - C^y|^p$ is defined by $C_{ijkl} = (C^x_{ik} - C^y_{jl})^p$. Intuitively, the Wasserstein distance to linear assignment problem is the same as the GW distance to quadratic assignment problem (QAP) (Koopmans and Beckmann, 1957), where the permutation matrix is replaced by a doubly stochastic one. While the QAP is known to be NP-hard, introducing more flexibility to the alignment matrix does not necessarily ease the optimization. Furthermore, the computational complexity of the 4D-tensor cost is prohibitive in most applications. For these reasons, the current approach to solve the discrete GW usually consists of an efficient approximation technique and additional structures on the cost tensor. In this section, we discuss both components in more details.

Structure of the cost tensor

To handle the cost tensor, an efficient, yet easy-to-implement strategy is to make it as decomposable as possible. Given its form $C = |C^x - C^y|^p$, the decomposability can be achieved by choosing $p = 2$. To see this, let us recall the following results.

Lemma 2.2.1 (Generalization of Proposition 1 in (Peyré, Cuturi, and Solomon, 2016))**.** *Denote $\oplus$ the Kronecker sum. For any matrix M, we write $M^{\odot 2} := M \odot M$, where $\odot$ is the element-wise multiplication. Given $A \in \mathbb{R}^{n_1\times d_1}, B \in \mathbb{R}^{n_2\times d_2}$ and $P \in \mathbb{R}^{d_1\times d_2}$, we define $|A-B|^2 \otimes P \in \mathbb{R}^{n_1\times n_2}$ by $(|A-B|^2 \otimes P)_{ij} = \sum_{k,l} |A_{ik} - B_{jl}|^2 P_{kl}$. Then, $|A-B|^2 \otimes P = A^{\odot 2}P_{\#1} \oplus B^{\odot 2}P_{\#2} - 2APB^T$.*

In particular, in the case of GW distance, for any $P \in U(\mu_X, \mu_Y)$, we have

$$\langle |C^x - C^y|^2 \otimes P, P\rangle = \langle (C^x)^{\odot 2}\mu_X, \mu_X\rangle + \langle (C^y)^{\odot 2}\mu_Y, \mu_Y\rangle - 2\langle C^x P (C^y)^T, P\rangle, \tag{2.70}$$

where $\langle (C^x)^{\odot 2}\mu_X, \mu_X\rangle + \langle (C^y)^{\odot 2}\mu_Y, \mu_Y\rangle$ is a constant independent of P and has computational cost of $O(m^2 + n^2)$. The cost of computing $C^x P (C^y)^T$ is $O(m^2 n + n^2 m)$. So, the overall computational complexity is reduced from $O(m^2 n^2)$ to $O(m^2 n + n^2 m)$. Moreover, if the input matrices C^x, C^y are factorizable, then this cost can be even further reduced, for both constant term and $C^x P (C^y)^T$. For example, the squared distance matrix is factorizable, thanks to two lemmas below in (Scetbon, Peyré, and Cuturi, 2021).

Lemma 2.2.2 (Exact low-rank factorization of squared distance matrix)**.** *If* $A = (a_1, ..., a_m) \in \mathbb{R}^{m\times d}$ *and* $B = (b_1, ..., b_n) \in \mathbb{R}^{n\times d}$*, then the matrix* $D \in \mathbb{R}^{m\times n}$ *defined by* $D_{ij} = ||a_i - b_j||_2^2$ *can be decomposed as* $D = D_a D_b^T$*, where* $D_a = (A^{\odot 2}1_d, 1_m, -\sqrt{2}A) \in \mathbb{R}^{m\times(d+2)}$ *and* $D_b = (1_n, B^{\odot 2}1_d, \sqrt{2}B) \in \mathbb{R}^{n\times(d+2)}$.

Lemma 2.2.3 (Element-wise square trick for factorizable matrix)**.** *For* $x \in \mathbb{R}^d$*, define* $\varphi(x) = vec(xx^T) \in \mathbb{R}^{d^2}$*. Suppose* $A = BC^T$*, where* $B \in \mathbb{R}^{m\times d}$ *and* $C \in \mathbb{R}^{n\times d}$*. By writing* B *as* $[b_1, ..., b_m]^T$ *and* C *as* $[c_1, ..., c_n]^T$*, define* $\widetilde{B} = [\varphi(b_1), ..., \varphi(b_m)]^T \in \mathbb{R}^{m\times d^2}$ *and* $\widetilde{C} = [\varphi(c_1), ..., \varphi(c_n)]^T \in \mathbb{R}^{n\times d^2}$*. Then* $A^{\odot 2} = \widetilde{B}\widetilde{C}^T$*. In particular,* $A^{\odot 2}$ *is also factorizable and has rank* d^2*. In PyTorch,* $\widetilde{B}$ *can be calculated by:* `torch.einsum("ij,ik->ijk", B, B).reshape(B.shape[0], B.shape[1]**2)`.

We deduce that, if $C^x = D_x D_x^T, (C^x)^{\odot 2} = \widetilde{D}_x \widetilde{D}_x^T$ and $C^y = D_y D_y^T, (C^y)^{\odot 2} = \widetilde{D}_y \widetilde{D}_y^T$, with $D_x \in \mathbb{R}^{m\times d_x}, \widetilde{D}_x \in \mathbb{R}^{m\times d_x^2}, D_y \in \mathbb{R}^{n\times d_y}$ and $\widetilde{D}_y \in \mathbb{R}^{n\times d_y^2}$ are low-rank matrices with $d_x \ll m$ and $d_y \ll n$, we have

$$\langle |C^x - C^y|^2 \otimes P, P\rangle = (C^x)^{\odot 2} P_{\#1} \oplus (C^y)^{\odot 2} P_{\#2} - 2C^x P (C^y)^T \tag{2.71}$$

$$= \mu_X^T \widetilde{D}_x \widetilde{D}_x^T \mu_X + \mu_Y^T \widetilde{D}_y \widetilde{D}_y^T \mu_Y - 2\langle D_x D_x^T P D_y D_y^T, P\rangle. \tag{2.72}$$

Comparing to Equation (2.70), the computational costs are now much cheaper, where for the constant term, it is $O(md_x^2 + nd_y^2)$, and for $C^x P (C^y)^T$, it is $O(mn(d_x + d_y) + (m + n)d_x d_y)$. If the distance matrix is not factorizable, one can consider the low-rank approximation, whose implementation details can be found in (Scetbon, Cuturi, and Peyré, 2021).

We also stress that, exploiting the decomposability is not the only way to reduce the computational complexity. For an overview of more complicated strategies, interested readers may consult Table 1 in (Mengyu Li and Meng, 2023) and the reference therein.

Approximation technique

The common principle behind many current methods for solving the GW problem is the linearization of the objective function, so that each iteration boils down to solving a (possibly regularized) OT problem. However, given the nonconvex nature of the GW problem, they also share a common limitation, where only local, but not global convergence is guaranteed. In what follows, we briefly discuss several solvers using this principle. We note that, however, in some particular situations, it is still possible to achieve the global convergence. For example, when $p = 2$ and C^x, C^y are squared Euclidean distance, Ryner, Kronqvist, and Karlsson (2023) propose to apply the cutting-plane method to a tractable relaxation of the GW problem and show that this algorithm converges to the global optimum of the GW problem.

1. Condition gradient descent Vayer et al. (2019a) directly estimate the GW problem by using the Frank-Wolfe algorithm, also known as *condition gradient descent* (Frank and Wolfe,

1956; Jaggi, 2013). More precisely, each iteration requires solving an unregularized OT problem, whose cost $C \otimes P^{(t)}$ is the gradient involving the previous OT plan, then interpolating betweeen the newly acquired OT plan and the previous one using a certain line search strategy. Since the objective function is nonconvex, the Frank-Wolfe algorithm only guarantees the convergence to a local stationary point (Lacoste-Julien, 2016).

2. Projected gradient descent Similar to the Wasserstein distance, one can also approximate the GW distance with entropic regularization.

$$\min_{P \in U(\mu_X, \mu_Y)} \langle C \otimes P, P \rangle + \varepsilon \mathrm{KL}(P|\gamma), \tag{2.73}$$

for some relevant reference histogram $\gamma \in \mathbb{R}_{>0}^{m \times n}$, for example, $\gamma = 1_{m \times n}$ corresponds to using the negative entropy as regularizer. To solve this regularized problem, Peyré, Cuturi, and Solomon (2016) and Solomon et al. (2016) propose to use projected gradient descent, where each iteration boils down to solving an entropic OT problem. More precisely,

Lemma 2.2.4 (Generalization of Proposition 2 in (Peyré, Cuturi, and Solomon, 2016))**.** *Given a matrix $M \in \mathbb{R}^{m \times n}$, consider the following entropic (fused) GW problem*

$$\min_{P \in U(\mu_X, \mu_Y)} \langle C \otimes P, P \rangle + \langle M, P \rangle + \varepsilon KL(P|\gamma). \tag{2.74}$$

Then, the PGD iteration reads: for any learning rate $\eta > 0$,

$$P^{(t+1)} = \operatorname*{argmin}_{P \in U(\mu_X, \mu_Y)} \eta \, \langle C \otimes P^{(t)} + M, P \rangle + \varepsilon KL\left(P|\gamma^{\eta} \odot (P^{(t)})^{1-\eta}\right). \tag{2.75}$$

As observed in (Peyré and Cuturi, 2019), $\eta = 1$ usually works well in practice, where the sequence $(P^{(t)})_t$ converges empirically. However, the theoretical convergence analysis remains unexplored. As a side note, adding regularization to the GW distance also incurs bias, similar to the entropic OT. Inspired by the Sinkhorn divergence, Bunne et al. (2019) apply the same strategy to debiase the entropic GW, but there is no theoritical guarantee with their approach.

3. (Inexact) Bregman proximal point Inspired by the work of Xie et al. (2020) on Inexact Proximal Optimal Transport, Xu, Luo, and Carin (2019) and Xu et al. (2019) propose to apply the Bregman proximal point (BPP) method to the GW problem, where at each iteration, we solve

$$\min_{P \in U(\mu_X, \mu_Y)} \langle C \otimes P, P \rangle + \eta \, \mathrm{KL}(P|P^{(t)}). \tag{2.76}$$

This is nothing but the entropic GW problem presented in the second approach, thus can be solved with the PGD algorithm. The resulting sequence only guarantees that every limit point is a stationary point (Xu et al., 2019).

4. Block coordinate descent Apart from using entropic regularization, one can also consider the following problem

$$\min_{P,Q\in U(\mu_X,\mu_Y)} \langle C \otimes P, Q\rangle. \tag{2.77}$$

Clearly, this is a lower bound of the GW distance since we no longer require two couplings to be equal. The bound can become the equality, for example, when $p = 2$ and C^x, C^y are the Euclidean distances. This means that dropping the equality constraint does not change the minimum nor the minimizer. More discussion on such situations can be found in Section 3.1. In these cases, we can apply the block coordinate descent (BCD) algorithm, where in each iteration, we alternatively fix one coupling and solve the OT problem with respect to the other. Since the objective function is smooth, this algorithm can only guarantee the convergence to a stationary point (P.Tseng, 2001). This lower-bound strategy can also be easily extended to the general problem (2.74), where it is enough to rewrite the objective function as

$$\min_{\substack{P,Q\in U(\mu_X,\mu_Y)\\ P=Q}} \langle C \otimes P, Q\rangle + \frac{1}{2}\langle M, P\rangle + \frac{1}{2}\langle M, Q\rangle + \frac{\varepsilon}{2}\mathrm{KL}(P|\gamma) + \frac{\varepsilon}{2}\mathrm{KL}(Q|\gamma). \tag{2.78}$$

Moreover, one can ignore the equality constraint without impacting the minimum, under exactly the same conditions as in Problem (2.77).

5. Bregman alternating projected gradient Another way to use the lower bound is in (Li et al., 2022).

$$\min_{\substack{(P,Q)\in\mathcal{C}_1\times\mathcal{C}_2\\ P=Q}} \langle C \otimes P, Q\rangle, \tag{2.79}$$

where $\mathcal{C}_1 = \{P \geq 0 : P_{\#1} = \mu_X\}$ and $\mathcal{C}_2 = \{Q \geq 0 : Q_{\#2} = \mu_Y\}$, They propose to use the Bregman alternating projected gradient (BAPG), whose iterates read: for a given learning rate $\eta > 0$,

1. $P^{(t+1)} = \mathrm{argmin}_{P\in\mathcal{C}_1} \langle C \otimes Q^{(t)}, P\rangle + \eta\mathrm{KL}(P|Q^{(t)})$.
2. $Q^{(t+1)} = \mathrm{argmin}_{Q\in\mathcal{C}_2} \langle C \otimes P^{(t+1)}, Q\rangle + \eta\mathrm{KL}(Q|P^{(t+1)})$.

However, BAPG has two main drawbacks: it only converges asymptotically to a critical point and the iterates do not necessarily satisfy the marginal constraints. To overcome these issues, the authors propose the hybrid Bregman Projected Gradient (hBPG), which consists of initializing

with the solution of the entropic GW problem, then applying the BPP method discuss previously. Local linear convergence result is also established for hBPG.

6. Saddle point approximation Similar to (Koehl, Delarue, and Orland, 2019), Koehl, Delarue, and Orland (2023) use the framework from statistical physics and approximate the GW distance by the limit of a decreasing sequence of free energies. Rougly speaking, at each iteration, this boils down to iteratively estimating the free energy by using the saddle point approximation. Formally, we fix a increasing sequence of inverse temperatures $(\beta_t)_t$ converging to $+\infty$. At the iteration t, set $Q^{(0)} = P^{(t-1)}$. Then, for $k \geq 1$, run the following iterative scheme until convergence.

1. Calculate the new OT cost $C^{(k)} = C \otimes Q^{(k)}$.
2. Solve the non-linear system of equations for $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$.

$$\begin{cases} \sum_{j=1}^{n} \varphi\big(\beta_t(u + v_j + C^{(k)}_{\cdot,j})\big) = \mu_X. \\ \sum_{i=1}^{m} \varphi\big(\beta_t(u_i + v + C^{(k)}_{i,\cdot})\big) = \mu_Y. \end{cases} \tag{2.80}$$

3. Compute $Q^{(k+1)} = \varphi\big(\beta_t(u \oplus v + C^{(k)})\big)$.

Finally, set $P^{(t)} = Q^{(k)}$ and repeat the procedure. Here $\varphi(x) = \frac{e^{-x}}{e^{-x}-1} + \frac{1}{x}$, for $x \neq 0$ and $\varphi(0) = 1/2$. In theory, while the GW distance is known to be the limit of the free energy when the temperature tends to 0^+, the convergence analysis of this algorithm remains unexplored.

2.2.5 Beyond GW distance

Fused GW distance Structured data is an object made of two components: structure and feature information, is ubiquitous in practical situations. Typical examples include attributed graph, whose node is assigned with a label, or cortical surface whose vertice is associated to a vectorial representation of the functional activation map, or in supervised learning, the dataset is a set of example-label pairs. By construction, neither GW nor Wasserstein distance is designed to handle this kind of data. Vayer et al. (2019a) propose a simple, yet efficient method called *Fused GW* (FGW) distance, which is able to take into account both types of information. It is convenient to cast structured data as an attributed graph $\mathcal{X} = (C^x, F^x, \mu_X)$, where C^x is the similarity matrix, F^x is the set of features associated to the nodes and μ_X is the histogram assigned to the nodes. Given two attributed graphs $\mathcal{X} = (C^x, F^x, \mu_X)$ and $\mathcal{Y} = (C^y, F^y, \mu_Y)$, for $\alpha \in [0, 1]$, we define

$$\mathrm{FGW}_\alpha(\mathcal{X}, \mathcal{Y}) = \inf_{P \in U(\mu_X, \mu_Y)} (1-\alpha)\langle C \otimes P, P\rangle + \alpha\langle M, P\rangle, \tag{2.81}$$

where the 4D-tensor cost C is usually of the form $C = |C^x - C^y|^p$ as in the GW distance, The matrix M usually represents the pairwise similarity amongst features, typically $M_{ij} = ||a_i - b_j||^p$, where $a_i \in F^x$ and $b_j \in F^y$. Note that, the presence of feature information indicates that the measure-preserving isometries for structure information needs to be also feature-preserving. The structure of the objective function allows FGW to interpolate between the OT and GW distances, depending on the asymptotic behavior of α (Vayer et al., 2019a).

Any solver for GW can be easily applied to the FGW. Instead of using predefined feature, it is also possible to learn it simultaneously with the alignment matrix, for example in (Xu et al., 2019). It can also be easily integrated into many OT-based divergences, for example low-rank GW (Scetbon, Peyré, and Cuturi, 2021), Keypoint-guided OT (Gu et al., 2022), Information-maximizing OT (Chuang, Jegelka, and Alvarez-Melis, 2023), fused unbalanced GW (Thual et al., 2022). In supervised learning, the main idea of FGW is also shared in other OT-based divergences, for example the OT Dataset distance between datasets (Alvarez-Melis and Fusi, 2020), Joint-distribution OT (Courty et al., 2017), or Transportation L^p distance (Thorpe et al., 2017).

Marginal relaxation in GW distance Inspired by the recent interest and development on UOT (Liero, Mielke, and Savaré, 2018), the GH-based distances between mm-spaces can be extended to the unbalanced setting. More precisely, given two mm-spaces $\mathcal{X} = (X, d_X, \mu_X)$ and $\mathcal{Y} = (Y, d_Y, \mu_Y)$, Ponti and Mondino (2020) study the unbalanced extension of the L_p-transportation distance called *Sturm-Entropic-Transport distance*, defined as

$$\mathrm{D}_{\mathrm{ET}}(\mathcal{X}, \mathcal{Y}) = \inf_{Z,f,g} UW_{Z,p}((f(X), f_{\#}\mu_X), (g(Y), g_{\#}\mu_Y)), \tag{2.82}$$

where the infimum is taken over all complete and separable metric spaces Z and isometric embeddings $f : X \to Z$ and $g : Y \to Z$.

To compare compact mm-spaces, Séjourné, Vialard, and Peyré (2021b) propose the unbalanced GW (UGW) divergence: given two relaxation parameters $\lambda_1, \lambda_2 > 0$ and two Csizar divergences $D_{\varphi_1}, D_{\varphi_2}$, they define

$$\mathrm{UGW}_\lambda(\mathcal{X}, \mathcal{Y}) = \inf_{\pi \in \mathcal{M}^+(X \times Y)} \int |d_X - d_Y|^p d\pi d\pi + \lambda_1 D^{\otimes 2}_{\varphi_1}(\pi_{\#1}|\mu_X) + \lambda_2 D^{\otimes 2}_{\varphi_2}(\pi_{\#2}|\mu_Y). \tag{2.83}$$

Here, $D^{\otimes 2}_{\varphi}$ denotes the *quadratic divergence* $D^{\otimes 2}_{\varphi}(\mu|\nu) := D_{\varphi}(\mu \otimes \mu|\nu \otimes \nu)$. This structure of product measures is particularly useful in the study of theoretical and practical properties of UGW, namely the existence of solution, the relation with conic formulation (Séjourné, Vialard, and Peyré, 2021b) and the robustness to outliers (Tran et al., 2023). Moreover, thanks to

Corollary 2.1.1, when $D_{\varphi_1}, D_{\varphi_2}$ are KL divergences, we have

$$\mathrm{UGW}_\lambda(\mathcal{X}, \mathcal{Y}) = \lambda_1\, m(\mu_X)^2 + \lambda_2\, m(\mu_Y)^2 - (\lambda_1 + \lambda_2)\, m(\pi^*)^2, \tag{2.84}$$

meaning that the minimum is a **quadratic** function of the minimizer.

Zhang et al. (2022b) introduce the unbalanced bi-directional Gromov-Monge divergence defined as: for $\lambda_x, \lambda_y > 0$,

$$\mathrm{UBGM}_\lambda(\mathcal{X}, \mathcal{Y}) = \inf_{\substack{f:X\to Y \\ g:Y\to X}} \Delta_p(f, g, X, Y) + \lambda_x \mathrm{D}_{\varphi_x}(g_\# \mu_Y | \mu_X) + \lambda_y \mathrm{D}_{\varphi_y}(f_\# \mu_X | \mu_Y), \tag{2.85}$$

where D_{φ_x} and D_{φ_y} are divergences on $\mathcal{M}_1^+(X)$ and $\mathcal{M}_1^+(Y)$, respectively. Here, the authors define a divergence D on $\mathcal{M}_1^+(Z)$ as a function $D : \mathcal{M}_1^+(Z) \times \mathcal{M}_1^+(Z) \to \mathbb{R}_{\geq 0}$ such that $D(P|Q) = 0$ if and only if $P = Q$. In particular, when μ_X and μ_Y are probability measures and $D_{\varphi_x}, D_{\varphi_y}$ are maximum mean discrepancies, the convergence rate of the empirical measures can be computed.

Other unbalanced extensions of the GW distance include the semi-relaxed GW divergence (Vincent-Cuaz et al., 2022), where only one marginal is relaxed, or the partial GW (Chapel, Alaya, and Gasso, 2020) motivated by the partial OT (Caffarelli and McCann, 2010; Figalli, 2010), where the mass of the transport plan is bounded above. We note that both divergences can be obtained from UGW, by choosing appropriate relaxation parameters and divergences. Another alternative is the outlier-robust GW (RGW) (Kong et al., 2023). It combines GW distance with the marginal relaxation of UOT from (Liero, Mielke, and Savaré, 2018) and the KL relaxations as constraints from (Balaji, Chellappa, and Feizi, 2020), though the authors do not give proper credits to these prior works. Similar to UGW, RGW also enjoys provable robustness property and shows strong performance in various graph learning tasks, where outliers present.

Chapter 3

Contributions to CO-Optimal Transport

This chapter presents two contributions to the CO-Optimal Transport (COOT). The first one is summarized in (Tran et al., 2021), which studies a relaxation of COOT via multi-marginal OT (MMOT). It unifies several popular OT methods under its umbrella by promoting structural information on the coupling. We show that incorporating such information into MMOT results in an instance of a difference of convex (DC) programming problem allowing us to solve it numerically. Despite high computational cost, the solutions provided by DC optimization are usually as qualitative as those obtained using available optimization schemes.

The second contribution is on the continuous COOT and its entropic approximation. We consider a generalization of measure network called *measure hypernetwork* and show that continuous COOT can be used to compare such objects. We then study the convergence behavior of the entropic approximation of COOT under the framework of finite-dimensional measure network. In particular, we can quantify the approximation error of entropic COOT and easily

extend this analysis to the GW distance.

3.1 Background on discrete CO-Optimal Transport

In many practical applications, the tabular data are usually expressed as a matrix whose rows represent samples and columns represent features. In general, the usual OT-based divergences, notably the Wasserstein and GW distances, mostly make use of the pairwise distances, either within or across domains, to construct the cost matrix or tensor. This approach has two consequences. First, only sample correspondences are of interest, By contrast, one completely discards the feature alignments, which can be also interpretable, for example, in single-cell multi-omics tasks (Demetci et al., 2022b). [1]. Second, the distance averages out the features, thus incurs information loss. This can be problematic in the high-dimensional setting, where the Euclidean distance is usually not a good metric (see for example, (Aggarwal, Hinneburg, and Keim, 2001), or Theorem 3.1.1 and Remark 3.1.2 in (Vershynin, 2018)) [2].

One way to overcome these limitations is to use the *Co-Optimal Transport* (COOT) (Redko et al., 2020), which learns simultaneously the sample and feature alignments. In what follows, we denote by $\Delta_n = \{p \in \mathbb{R}^n_{>0} : \sum_{i=1}^n p_i = 1\}$ the simplex histogram with n bins. We call $\mathcal{X} = (X, \mu_1^X, \mu_2^X)$ a *weighted matrix* defined by a triplet comprised of a matrix $X \in \mathbb{R}^{n_x \times d_x}$ equipped with the histograms $\mu_1^X \in \Delta_{n_x}$ and $\mu_2^X \in \Delta_{d_x}$ on its rows and columns, respectively. For $p \geq 1$, we define the COOT between two weighted matrices $\mathcal{X} = (X, \mu_1^X, \mu_2^X)$ and $\mathcal{Y} = (Y, \mu_1^Y, \mu_2^Y)$ as

$$\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) := \inf_{\substack{P \in U(\mu_1^X, \mu_1^Y) \\ Q \in U(\mu_2^X, \mu_2^Y)}} \sum_{i,j,k,l} (X_{ij} - Y_{kl})^p P_{ik} Q_{jl}. \tag{3.1}$$

By Proposition 1 in (Redko et al., 2020), if the weights are uniforms, then COOT defines a distance on the space of weighted matrices, up to permutation of the matrix coordinates. We note that the formulation (3.1) can be easily extended to the multi-coupling setting, for example, in (Kerdoncuff et al., 2022).

From the perspective of matrix-comparison, COOT provides a principled way to compare any two arbitrary-size matrices. By contrast, this is not the case for many other existing divergences, whose applicability is summarized in Table 3.1.

1. In case of Wasserstein distance, there is no interest of feature correspondences because intuitively, they are equivalent to the identity matrix. However, it is no longer clear how to identify the such matching in the GW setting.

2. For more informative discussion, see https://stats.stackexchange.com/questions/99171/why-is-euclidean-distance-not-a-good-metric-in-high-dimensions.

Divergences	Input matrices	Requirement on histograms
Matrix norms	Same-size matrices	Not applicable
OT, UOT, SW	Matrices with the same number of columns	Only requires histogram on rows
GW, FGW, UGW, SGW	Square matrices	Histograms on rows and columns must be the same
COOT	Arbitrary-size matrices	Any histograms on rows and columns

Table 3.1 – Applicablility of some popular divergences. COOT is much more flexible than other OT-based divergences. UOT and UGW are the unbalaced OT and unbalaced GW divergence, respectively. FGW is the Fused GW divergence. SW and SGW denote the sliced Wasserstein (Bonneel et al., 2015; Rabin et al., 2012) and sliced GW (Vayer et al., 2019b) distances.

Co-Optimal Transport as lower bound of GW distance Under the framework of discrete GW, where the inputs are similarity matrices, for simplicity, we write $\mathcal{X} = (C^x, \mu_X)$, where C^x is the similarity matrix and μ_X is the sample histogram. Now, the GW distance can be reformulated as

$$\mathrm{GW}(\mathcal{X}, \mathcal{Y}) = \inf_{\substack{P,Q \in U(\mu_X, \mu_Y) \\ P=Q}} \sum_{i,j,k,l} (C^x_{ij} - C^y_{kl})^p P_{ik} Q_{jl}, \tag{3.2}$$

meaning that we optimize with respect to two independent couplings under the additional constraint that they must be equal. If it is relaxed, then one recovers the COOT distance between $\mathcal{X}$ and $\mathcal{Y}$. We also stress that COOT should not be confused with the third lower bound of the GW distance (Mémoli, 2007, 2011b) defined as

$$\mathrm{TLB}(\mathcal{X}, \mathcal{Y}) := \inf_{Q \in U(\mu_X, \mu_Y)} \Big(\inf_{P \in U(\mu_X, \mu_Y)} \sum_{i,j,k,l} (C^x_{ij} - C^y_{kl})^p P_{ik} \Big) Q_{jl}. \tag{3.3}$$

In particular, we have $\mathrm{GW}(\mathcal{X}, \mathcal{Y}) \geq \mathrm{COOT}(\mathcal{X}, \mathcal{Y}) \geq \mathrm{TLB}(\mathcal{X}, \mathcal{Y})$.

If the inputs are the Euclidean, or squared Euclidean distances, then equality holds between COOT and GW distances (Redko et al., 2020; Séjourné, Vialard, and Peyré, 2021b). These results are based on the prior works of Konno (1976) and Maron and Lipman (2018) and summarized in the folllowing proposition.

Definition 3.1.1. *A square matrix $A \in \mathbb{R}^{n \times n}$ is conditionally negative semi-definite (CND) if it is symmetric and for any $c \in \mathbb{R}^n$ such that $\sum_i c_i = 0$, we have $c^T A c \leq 0$.*

Proposition 3.1.1. *For $p = 2$, suppose that C^x and C^y are of the forms: $C^x_{ij} = f_i + f_j + A_{ij}$ and $C^y_{kl} = g_k + g_l + B_{kl}$, where f, g are vectors in $\mathbb{R}^m, \mathbb{R}^n$, respectively, and the matrices A, B are CND. Then $GW(\mathcal{X}, \mathcal{Y}) = COOT(\mathcal{X}, \mathcal{Y})$. Furthermore, if (P_1^*, P_2^*) is a solution of the COOT problem, then P_1^* and P_2^* are two solutions of the GW problem. In particular, if the semi-definiteness is*

replaced by the definiteness, then $P_1^* = P_2^*$.

In particular, if the similarity matrices is CND, then one can safely remove the equality constraint without changing the minimum. Thus, Proposition 3.1.1 justifies the rationale behind the alternative minimization procedure for GW distance presented in Section 2.2.4.

3.2 Continuous Co-Optimal Transport

In this section, we present our unpublished work on the continuous COOT and its entropic approximation. We will mostly follow the terminology and concepts proposed by Chowdhury et al. (2023).

3.2.1 Formulation and preliminary results

As already seen in discrete GW problem (3.2), by rewriting a measure network $\mathcal{X} = (X, c_X, \mu_X)$ as $\widetilde{\mathcal{X}} = ((X_1, \mu_1^X), (X_2, \mu_2^X), c_X)$, with $X_1 = X_2 = X$ and $\mu_1^X = \mu_2^X = \mu_X$, one can reformulate the GW problem as

$$\begin{aligned} &\inf_{\pi_1,\pi_2} \int_{X_1\times Y_1}\int_{X_2\times Y_2} |c_X(x_1,x_2) - c_Y(y_1,y_2)|^p \, d\pi_1(x_1,y_1)\, d\pi_2(x_2,y_2). \\ &\text{subject to: } \pi_k \in U(\mu_k^X, \mu_k^Y), \forall k = 1,2, \\ &\qquad\qquad \pi_1 = \pi_2. \end{aligned} \tag{3.4}$$

When the equality constraint on the two couplings is relaxed, we can allow that either $X_1 \neq X_2$ or $Y_1 \neq Y_2$. The interest of such situation can be found, for example, in heterogenous domain adaptation, where X_1 and Y_1 represent the "sample" spaces in the source and target domains, respectively, and X_2 and Y_2 represent the "feature" spaces in the source and target domains, respectively. As a result, the corresponding "sample" and "feature" couplings are also different in their natures.

Definition 3.2.1 (Measure hypernetwork (Chowdhury et al., 2023))**.** *Suppose* (X_1, μ_1^X) *and* (X_2, μ_2^X) *are two Polish measure spaces, and* c_X *is a bounded measurable function on* $X_1 \times X_2$. *We call the triplet* $\mathcal{X} = ((X_1, \mu_1^X), (X_2, \mu_2^X), c_X)$ *a* ***measure hypernetwork****. We also say* c_X *is the* ***interaction*** *between* X_1 *and* X_2.

Without risk of confusion, when $X_1 = X_2 = X$ and $\mu_1^X = \mu_2^X = \mu_X$, we use interchangeably "measure hypernetwork" and "measure network" in the context of GW (Chowdhury and Mémoli, 2019). When X_1 and X_2 are finite spaces (so μ_1^X and μ_2^X are histograms), we call $\mathcal{X}$ a finite measure hypernetwork. For convenience, we also refer the index 1 as "sample" and 2 as "feature", for example, X_1 is the sample space, π_2 is the feature coupling.

Definition 3.2.2 (COOT distance between measure hypernetworks (Chowdhury et al., 2023)). *For $p \geq 1$, the COOT distance between two measure hypernetworks $\mathcal{X}$ and $\mathcal{Y}$ is defined as*

$$COOT(\mathcal{X}, \mathcal{Y}) = \inf_{\substack{\pi_1 \in U(\mu_1^X, \mu_1^Y) \\ \pi_2 \in U(\mu_2^X, \mu_2^Y)}} \iint |c_X(x_1, x_2) - c_Y(y_1, y_2)|^p \, d\pi_1(x_1, y_1) \, d\pi_2(x_2, y_2). \tag{3.5}$$

It is not difficult to see that Definition 3.2.2 generalizes the discrete COOT (Redko et al., 2020). In practice, the input data is usually expressed as matrix, whose rows represent samples and columns represent features. In this case, the interaction value is precisely the coordinate of the data matrix. Meanwhile, the sample and feature spaces are unknown and have little interest and importance.

First, we can show that the COOT problem (3.5) is well defined.

Proposition 3.2.1 (Lemma 35 in (Chowdhury et al., 2023)). *The COOT problem always admits a minimizer.*

We provide the proof in Appendix 8.2.2. While the proofs of our result and of Lemma 35 in (Chowdhury et al., 2023) use the same proof technique of Theorem 2.2 in (Chowdhury and Mémoli, 2019), ours is slightly different. More precisely, we exploit a different reformulation of COOT, where it can be rewritten as a multi-marginal OT problem with additional factorization constraint on the coupling. Later, we will see that, this observation also allows to establish the convergence result of entropic COOT.

3.2.2 Metric properties

The framework on the GW isomorphism presented in Section 2.2.3 can be extended immediately to the COOT setting. In particular, while our presentation is different to that of Chowdhury et al. (2023), we still come up with the same metric properties.

Definition 3.2.3 (Relaxed mass splitting). *A measure hypernetwork $\mathcal{Z}$ is a **relaxed mass splitting** (RMS) of a measure hypernetwork $\mathcal{X}$ if there exist two measure-preserving maps $\varphi_k : Z_k \to X_k$, for $k = 1, 2$, such that the pullback equality $c_Z = (\varphi_1, \varphi_2)^* c_X$ holds $\mu_1^Z \otimes \mu_2^Z$-almost everywhere in $Z_1 \times Z_2$. We denote MS($\mathcal{X}$) the set of all mass splittings of $\mathcal{X}$. This pair of maps (φ_1, φ_2) is also called **basic weak isomorphism** in (Chowdhury et al., 2023).*

We denote RMS($\mathcal{X}$) the set of all relaxed mass splittings of $\mathcal{X}$. Clearly, $\mathcal{X} \in$ RMS($\mathcal{X}$), so RMS($\mathcal{X}$) is not empty. In particular, for measure networks, if $\mathcal{Z} \in$ MS($\mathcal{X}$), then $\mathcal{Z} \in$ RMS($\mathcal{X}$), meaning that MS($\mathcal{X}$) $\subset$ RMS($\mathcal{X}$). Now, we can define the isomorphism between the measure hypernetworks as follows.

Definition 3.2.4 (COOT-isomorphism). *Two measure hypernetworks are*

1. *strongly isomorphic if there exist two* ***bijective*** *measure-preserving map from one hypernetwork to the other such that the pullback equality holds* ***everywhere****.*
2. *semi-strongly isomorphic if one is the RMS of the other and vice versa.*
3. *weakly isomorphic if they have a common RMS.*

This is an immediate relaxation of the GW isomorphism. In particular, in case of measure networks, isomorphism in GW sense implies COOT isormorphism. The following result summarizes the relations amongst the three types of COOT isomorphism.

Corollary 3.2.1. *Given two measure hypernetworks $\mathcal{X}$ and $\mathcal{Y}$. Consider three statements*

(1) $\mathcal{X}$ and $\mathcal{Y}$ are strongly isomorphism.

(2) $\mathcal{X}$ and $\mathcal{Y}$ are semi-strongly isomorphism.

(3) $\mathcal{X}$ and $\mathcal{Y}$ are weakly isomorphism.

Then, the following relations hold

1. $(1) \implies (2) \implies (3)$.
2. *If $\mathcal{X}$ and $\mathcal{Y}$ are finite, then* $(2) \implies (1)$.
3. *If $\mathcal{X}$ and $\mathcal{Y}$ are finite such that $|X_k| = |Y_k|$ and μ_k^X, μ_k^Y are uniform distributions, for $k = 1, 2$, then* $(3) \implies (2)$. *This means all three forms are equivalent.*

Now, we can characterize the weak isomorphism by

Proposition 3.2.2. *Two measure hypernetworks $\mathcal{X}$ and $\mathcal{Y}$ are COOT-weakly isomorphic if and only if $COOT(\mathcal{X}, \mathcal{Y}) = 0$.*

Proposition 3.2.3 (Theorem 1 in (Chowdhury et al., 2023))**.** *$COOT^{1/p}$ defines a metric on the space of measure hypernetworks, up to COOT-weak isomorphism.*

3.2.3 Entropic regularization and approximation error

Similar to the Wasserstein and GW distances, one can approximate the COOT with entropic regularization. In this thesis, we are interested in the following formulation of entropic COOT: for $\varepsilon > 0$,

$$\mathrm{COOT}_\varepsilon(\mathcal{X}, \mathcal{Y}) = \inf_{\substack{\pi_1 \in U(\mu_1^X, \mu_1^Y) \\ \pi_2 \in U(\mu_2^X, \mu_2^Y)}} \iint |c_X(x_1, x_2) - c_Y(y_1, y_2)|^p \, d\pi_1(x_1, y_1) \, d\pi_2(x_2, y_2) \tag{3.6}$$

$$+ \varepsilon \, \mathrm{KL}(\pi_1 \otimes \pi_2 | (\mu_1^X \otimes \mu_1^Y) \otimes (\mu_2^X \otimes \mu_2^Y)). \tag{3.7}$$

Note that, this structure of the KL divergence term is particularly handy to prove all results related to the entropic COOT. It also bears similarity with the *quadratic divergence* (Séjourné,

Vialard, and Peyré, 2021b) defined by $\mathrm{KL}^{\otimes 2}(\mu, \nu) := \mathrm{KL}(\mu \otimes \mu | \nu \otimes \nu)$, which is used to define the unbalanced GW divergence. Note that, in the balanced setting, the joint penalization in terms of KL divergence is in fact equivalent to the independent KL-penalization. More precisely, given any $\pi_k \in U(\mu_k^X, \mu_k^Y)$, for $k = 1, 2$, we have

$$\mathrm{KL}(\pi_1 \otimes \pi_2 | (\mu_1^X \otimes \mu_1^Y) \otimes (\mu_2^X \otimes \mu_2^Y)) = \mathrm{KL}(\pi_1 | \mu_1^X \otimes \mu_1^Y) + \mathrm{KL}(\pi_2 | \mu_2^X \otimes \mu_2^Y). \tag{3.8}$$

In practice, since the couplings may not have the same nature (for example, when working directly with the input data, rather than via the similarity matrix), they can be penalized by different values of regularization.

Proposition 3.2.4. *The entropic COOT problem always admits a minimizer.*

One major practical interest of entropic COOT is that, for sufficiently small regularization, it provides a good proxy for unregularized COOT. Now, we establish the bound for the approximation error in the setting of finite-dimensional measure network.

Proposition 3.2.5. *Given two measure networks $\mathcal{X} = (X, \mu_X, c_X)$ and $\mathcal{Y} = (Y, \mu_Y, c_Y)$, where X is bounded subset of $\mathbb{R}^{d_x}$ and Y is a bounded subset of $\mathbb{R}^{d_y}$. Denote $d = \max\{d_x, d_y\}$ and $D = \max\{diam(X), diam(Y)\}$. Suppose there exists two constants $L, q > 0$ such that $|c_X(x_1, x_2)| \leq L||x_1 - x_2||^q$, for every $(x_1, x_2) \in X^2$, and $|c_Y(y_1, y_2)| \leq L||y_1 - y_2||^q$, for every $(y_1, y_2) \in Y^2$. Then, the approximation error between the p-COOT distance and its entropic approximation can be quantified as follows.*

$$COOT_\varepsilon(\mathcal{X}, \mathcal{Y}) - COOT(\mathcal{X}, \mathcal{Y}) \leq \frac{4d\varepsilon}{pq} + \frac{4d\varepsilon}{pq} \log\Big(\frac{(2D)^{pq} L^p d^q pq}{4d\varepsilon}\Big). \tag{3.9}$$

This bound is similar to the one in Wasserstein setting (Genevay et al., 2019). This is due to the fact that the entropic COOT can be reformulated as a variant of multi-marginal OT problem, thus the block approximation technique (Carlier et al., 2017) can be applied.

Let us consider two special cases of Proposition 3.2.5.

- When $p = 2$ and c_X, c_Y are Euclidean distances (*i.e.*, $q = 1$), the upper bound becomes $2d\varepsilon + 2d\varepsilon \log\left(\frac{2D^2L^2}{\varepsilon}\right)$.
- When $p = 2$ and c_X, c_Y are squared Euclidean distances (*i.e.*, $q = 2$), the upper bound becomes $d\varepsilon + d\varepsilon \log\left(\frac{16D^4L^2d}{\varepsilon}\right)$.

In both situations, the dependence of the bound on the maximal distance D between points within each space (even only at logarithmic scale) and on the dimension d indicates that either high dimensional space, or large intra-space distance (for example, due to outliers) may negatively impact the approximation error. On the other hand, by comparing these two bounds, we deduce that if $\log\left(\frac{4d\varepsilon}{L^2}\right) \leq 1$, then the squared Euclidean distances generates a provably better (smaller) upper bound.

With little modification of the proof, exactly the same upper bound in Proposition 3.2.5 holds for GW distance. We note that Zhang et al. (2022a) also establish a similar $O(\varepsilon \log \varepsilon)$-approximation error between the unregularized and entropic GW, but they rely on different assumptions to ours. In particular, their result only holds for 2-GW distance, when the distance function is the squared-Euclidean norm. By contrast, Proposition 3.2.5 holds for any p-GW distance and any L^q-norm as distance function.

3.3 Factored couplings in Multi-marginal Optimal Transport via Difference of Convex programming

3.3.1 Introduction

Broadly speaking, the classic OT problem provides a principled approach for transporting one probability distribution onto another following the principle of the least effort. Such a problem, and the distance on the space of probability distributions derived from it, arise in many areas of machine learning (ML) including generative modeling, transfer learning and information retrieval, where OT has been successfully applied. A natural extension of classic OT, in which the admissible transport plan can have more than two prescribed marginal distributions, is called the multi-marginal optimal transport (MMOT) (Gangbo and Swiech, 1998). The latter has several attractive properties: it enjoys a duality theory (Kellerer, 1984) and finds connections with the probabilistic graphical models (Haasler et al., 2020) and the Wasserstein barycenter problem (Agueh and Carlier, 2011) used for data averaging. While being less popular than the classic OT with two marginals, MMOT is a very useful framework on its own with some notable recent applications in generative adversarial networks (Cao et al., 2019), clustering (Mi and Bento, 2020) and domain adaptation (He et al., 2019; Hui et al., 2018), to name a few.

The recent success of OT in ML is often attributed to the entropic regularization (Cuturi, 2013) where the authors imposed a constraint on the coupling matrix forcing it to be closer to the independent coupling given by the rank-one product of the marginals. Such a constraint leads to the appearance of the strongly convex entropy term in the objective function and allows the entropic OT problem to be solved efficiently using simple Sinkhorn-Knopp matrix balancing algorithm. In addition to this, it was also noticed that structural constraints on the coupling and cost matrices allow to reduce the high computational cost and sample complexity of the classic OT problem (Forrow et al., 2019; Genevay et al., 2019; Lin, Azabou, and Dyer, 2021; Scetbon, Cuturi, and Peyré, 2021). However, none of these works considered a much more challenging case of doing so in a multi-marginal setting. On the other hand, while the work of Haasler et al. (2020) considers the MMOT problem in which the cost tensor induced by a graphical structure, it does not naturally promote the factorizability of transportation plans.

Contributions In this work, we define and study a general MMOT problem with structural penalization on the coupling matrix. We start by showing that a such formulation includes several popular OT methods as special cases and allows to gain deeper insights into them. We further consider a relaxed problem where the hard constraint is replaced by a regularization term and show that it leads to an instance of the difference of convex programming problem. A numerical study of the solutions obtained when solving the latter in cases of interest highlights their competitive performance when compared to solutions provided by the optimization strategies used previously.

3.3.2 Preliminary knowledge

Notations. In the discrete setting, the Kullback-Leibler divergence between two positive vectors $p, q \in \mathbb{R}^n_{>0}$ is defined as $\mathrm{KL}(p|q) = \sum_i p_i \log \frac{p_i}{q_i} - \sum_i p_i + \sum_i q_i$, with the convention that $0 \log 0 = 0$.

In what follows, given an integer $N \geq 1$, for any positive integers $a_1, ..., a_N$, we call $P \in \mathbb{R}^{a_1 \times ... \times a_N}$ a N-D tensor. In particular, a 1-D tensor is a vector and 2-D tensor is a matrix. A tensor is a probability tensor if its entries are nonnegative and the sum of all entries is 1. Given N probability vectors $\mu_1, ..., \mu_N$, we write $\mu = (\mu_n)_{n=1}^N$ and $\mu^{\otimes} := \mu_1 \otimes ... \otimes \mu_N$. We denote Σ the set of N-D probability tensors and $U(\mu) \subset \Sigma$ the set of nonnegative tensors whose N marginal distributions are $\mu_1, ..., \mu_N$. In this case, any coupling in $U(\mu)$ is said to be *admissible*.

Multi-marginal OT problem. Given a collection of N probability vectors $\mu = (\mu_n \in \mathbb{R}^{a_n})_{n=1}^N$ and a N-D cost tensor $C \in \mathbb{R}^{a_1 \times ... \times a_N}$, the MMOT problem reads

$$\mathrm{MMOT}(\mu) = \inf_{P \in U(\mu)} \langle C, P \rangle.$$

In practice, such a formulation is intractable to optimize in a discrete setting as it results in a linear program where the number of constraints grows exponentially in N. A more tractable strategy for solving MMOT is to consider the following entropic regularization problem

$$\inf_{P \in U(\mu)} \langle C, P \rangle + \varepsilon \mathrm{KL}(P|\mu^{\otimes}). \tag{3.10}$$

which can be solved using Sinkhorn's algorithm (Benamou et al., 2014). We refer the interested reader to Appendix 8.2.3 for algorithmic details.

3.3.3 Factored Multi-marginal Optimal Transport

In this section, we first define a factored MMOT (F-MMOT) problem where we seek to promote a structure on the optimal coupling given such as a factorization into a tensor product.

Interestingly, such a formulation can be shown to include several other OT problems as special cases. Then, we introduce a relaxed version called MMOT-DC where the factorization constraint is smoothly promoted through a Kullback-Leibler penalty.

Motivation

Before a formal statement of our problem, we first give a couple of motivating examples showing why and when structural constraints on the coupling matrix can be beneficial. To this end, first note that a trivial example of the usefulness of such constraints in OT is the famous entropic regularization. Indeed, while most of the works define the latter by adding negative entropy of the coupling to the classic OT objective function directly, the original idea was to constraint the sought coupling to remain close (to some extent) to a rank-one product of the two marginal distributions. The appearance of negative entropy in the final objective function is then only a byproduct of such constraint due to the decomposition of the KL divergence into a sum of three terms with two of them being constant. Below we give two more examples of real-world applications related to MMOT problem where a certain decomposition imposed on the coupling tensor can be desirable.

Multi-source multi-target translation. A popular task in computer vision is to match images across different domains in order to perform the so-called image translation. Such tasks are often tackled within the GAN framework where one source domain from which the translation is performed, is matched with multiple target domains modeled using generators. While MMOT was applied in this context by Cao et al. (2019) when only one source was considered, its application in a multi-source setting may benefit from structural constraints on the coupling tensor incorporating the human prior on what target domains each source domain should be matched to.

Multi-task reinforcement learning. In this application, the goal is to learn individual policies for a set of agents while taking into account the similarities between them and hoping that the latter will improve the individual policies. A common approach is to consider an objective function consisting of two terms where the first term is concerned with learning individual policies, while the second forces a consensus between them. Similar to the example considered above, MMOT problem was used to promote the consensus across different agents' policies in (Cohen, Kumar, and Deisenroth, 2021), even though such a consensus could have benefited from a prior regarding the semantic relationships between the learned tasks.

Factored MMOT and its relaxation

We start by giving several definitions used in the following parts of this section.

Definition 3.3.1 (Tuple partition). *Given two integers $N \geq M \geq 2$, a sequence of tuples $\mathcal{T} = (\mathcal{T}_m)_{m=1}^M$, is called a tuple partition of the N-tuple $(1, ..., N)$ if the tuples $\mathcal{T}_1, ..., \mathcal{T}_M$ are nonempty and disjoint, and their concatenation in this order gives $(1, ..., N)$.*

Here, we implicitly take into account the order of the tuple, which is not the case for the partition of the set $[N]$. If there exists a tuple in $\mathcal{T}$ which contains only one element, then we say $\mathcal{T}$ is *degenerate*.

Definition 3.3.2 (Marginal tensor). *Given a tensor $P \in \mathbb{R}^{a_1 \times ... \times a_N}$ and a tuple partition $\mathcal{T} = (\mathcal{T}_m)_{m=1}^M$, we call $P_{\#\mathcal{T}_m}$ its $\mathcal{T}_m$-marginal tensor, by summing P over all dimensions not in $\mathcal{T}_m$. We write $P_{\#\mathcal{T}} = P_{\#\mathcal{T}_1} \otimes ... \otimes P_{\#\mathcal{T}_M} \in \mathbb{R}^{a_1 \times ... \times a_N}$ the tensor product of its marginal tensors.*

For example, for $M = N = 2$, we have $\mathcal{T}_1 = (1)$ and $\mathcal{T}_2 = (2)$. So, given a matrix $P \in \mathbb{R}^{a_1 \times a_2}$, its marginal tensors $P_{\#\mathcal{T}_1}$ and $P_{\#\mathcal{T}_2}$ are simply vectors in $\mathbb{R}^{a_1}$ and $\mathbb{R}^{a_2}$, respectively, defined by $(P_{\#\mathcal{T}_1})_i = \sum_j P_{ij}$ and $(P_{\#\mathcal{T}_2})_j = \sum_i P_{ij}$ for $(i, j) \in [a_1] \times [a_2]$. The tensor product $P_{\#\mathcal{T}} \in \mathbb{R}^{a_1 \times a_2}$ is then defined by $(P_{\#\mathcal{T}})_{ij} = (P_{\#\mathcal{T}_1})_i (P_{\#\mathcal{T}_2})_j$. Clearly, if P is a probability tensor, then so are its marginal tensors and tensor product.

Suppose $\mathcal{T}_m = (p, ..., q)$ for some $m \in [M]$ and $1 \leq p \leq q \leq N$. We denote $\Sigma_{\mathcal{T}_m}$ the set of probability tensors in $\mathbb{R}^{a_p \times ... \times a_q}$ and $U_{\mathcal{T}_m} \subset \Sigma_{\mathcal{T}_m}$ the set of probability tensors in $\mathbb{R}^{a_p \times ... \times a_q}$ whose r^{th}-marginal vector is μ_r, for every $r = p, ..., q$. We also define $\mu_{\mathcal{T}_m}^{\otimes} := \mu_p \otimes ... \otimes \mu_q$.

Definition 3.3.3 (Factored MMOT). *Given a collection of histograms $\mu = (\mu_n)_{n=1}^N$ and a tuple partition $\mathcal{T} = (\mathcal{T}_m)_{m=1}^M$, we consider the following OT problem*

$$F\text{-}MMOT(\mathcal{T}, \mu) = \inf_{P \in U_{\mathcal{T}}} \langle C, P \rangle, \tag{3.11}$$

where $U_{\mathcal{T}} \subset U(\mu)$ is the set of admissible couplings which can be factorized as a tensor product of M component probability tensors in $\Sigma_{\mathcal{T}_1}, ..., \Sigma_{\mathcal{T}_M}$.

Several remarks are in order here. First, one should note that the partition considered above is in general not degenerate meaning that the decomposition can involve tensors of an arbitrary order $< N$. Second, the decomposition in this setting depicts the prior knowledge regarding the tuples of measures which should be independent: the couplings for the measures from different tuples will be degenerate and the optimal coupling tensor will be reconstructed from couplings of each tuple separately. Third, suppose the partition $(\mathcal{T}_m)_{m=1}^M$ is not degenerate and $M = 2$, i.e. the tensor is factorized as product of two tensors, Problem (3.11) is equivalent to a variation of low non-negative rank OT problem (see Appendix 8.2.3).

As for the existence of the solution to this problem, we have that $U_{\mathcal{T}}$ is compact because it is a close subset of the compact set $U(\mu)$, which implies that Problem (3.11) always admits a

solution. Furthermore, observe that

$$\begin{aligned} U_{\mathcal{T}} &= \{P \in U(\mu) : P = P_1 \otimes ... \otimes P_M, \text{where } P_m \in \Sigma_{\mathcal{T}_m}, \forall m = 1, ..., M\} \\ &= \{P \in \Sigma : P = P_1 \otimes ... \otimes P_M, \text{where } P_m \in U_{\mathcal{T}_m}, \forall m = 1, ..., M\}. \end{aligned} \tag{3.12}$$

Thus, the problem F-MMOT can be rewritten as

$$\text{F-MMOT}(\mathcal{T}, \mu) = \inf_{\substack{P_m \in U_{\mathcal{T}_m} \\ \forall m=1,...,M}} \langle C, P_1 \otimes ... \otimes P_M \rangle. \tag{3.13}$$

So, if $\mathcal{T}_1, ..., \mathcal{T}_M$ are 2-tuples and two marginal distributions corresponding to each $U_{\mathcal{T}_m}$ are identical and uniform, then by Birkhoff's theorem (Birkhoff, 1946), Problem (3.11) admits an optimal solution in which each component tensor P_m is a permutation matrix.

Two special cases. When $N = 4$ and $M = 2$ with $\mathcal{T}_1 = (1, 2)$ and $\mathcal{T}_2 = (3, 4)$, Problem (3.11) becomes the CO-Optimal transport (COOT), where the two component tensors are known as *sample* and *feature* couplings. If furthermore, $a_1 = a_3, a_2 = a_4$, and $\mu_1 = \mu_3, \mu_2 = \mu_4$, it becomes a lower bound of the discrete Gromov-Wasserstein (GW) distance. This means that our formulation can be seen as a generalization of several OT formulations.

Observe that if a probability tensor P can be factorized as a tensor product of probability tensors, i.e. $P = P_1 \otimes ... \otimes P_M$, then each P_m is also the $\mathcal{T}_m$-marginal tensor of P. In this case, we have $P = P_{\#\mathcal{T}}$. This prompts us to consider the following relaxation of factored MMOT, where the hard constraint $U_{\mathcal{T}}$ is replaced by a regularization term.

Definition 3.3.4 (Relaxed Factored MMOT)**.** *Given $\varepsilon \geq 0$, a collection of measures μ and a tuple partition $\mathcal{T}$, we define the following problem:*

$$\textit{MMOT-DC}_\varepsilon(\mathcal{T}, \mu) = \inf_{P \in U(\mu)} \langle C, P \rangle + \varepsilon KL(P | P_{\#\mathcal{T}}). \tag{3.14}$$

From the exposition above, one can guess that this relaxation is reminiscent of the entropic regularization in MMOT and coincides with it when $M = N$. As such, it also recovers the classical entropic OT. One should note that the choice of the KL divergence is not arbitrary and its advantage will become clear when it comes to the algorithm. A special case of Problem (3.14) is when $M = N$, we recover the entropic-regularized MMOT problem.

After having defined the two optimization problems, we now set on exploring their theoretical properties.

3.3.4 Theoretical properties

Intuitively, the relaxed problem is expected to allow for solutions with a lower value of the final objective function. We formally prove the validity of this intuition below.

Proposition 3.3.1 (Preliminary properties)**.** *Given a collection of histograms μ and a tuple partition $\mathcal{T}$,*

1. *For every $\varepsilon \geq 0$, we have MMOT$(\mu) \leq$ MMOT-DC$_\varepsilon(\mathcal{T}, \mu) \leq$ F-MMOT$(\mathcal{T}, \mu)$.*
2. *For every $\varepsilon > 0$, MMOT-DC$_\varepsilon(\mathcal{T}, \mu) = 0$ if and only if F-MMOT$(\mathcal{T}, \mu) = 0$.*

An interesting property of MMOT-DC is that it interpolates between MMOT and F-MMOT. Informally, for very large ε, the KL divergence term dominates, so the optimal transport plans tend to be factorizable. On the other hand, for very small ε, the KL divergence term becomes negligible and we approach MMOT. The result below formalizes this intuition.

Proposition 3.3.2 (Interpolation between MMOT and F-MMOT)**.** *For any tuple partition $\mathcal{T}$ and for $\varepsilon > 0$, let P_ε be a minimiser of the problem MMOT-DC$_\varepsilon(\mathcal{T}, \mu)$.*

1. *When $\varepsilon \to \infty$, one has MMOT-DC$_\varepsilon(\mathcal{T}, \mu) \to$ F-MMOT$(\mathcal{T}, \mu)$. In this case, any cluster point of the sequence of minimisers $(P_\varepsilon)_\varepsilon$ is a minimiser of F-MMOT$(\mathcal{T}, \mu)$.*
2. *When $\varepsilon \to 0$, then MMOT-DC$_\varepsilon(\mathcal{T}, \mu) \to$ MMOT(μ). In this case, any cluster point of the sequence of minimisers $(P_\varepsilon)_\varepsilon$ is a minimiser of MMOT(μ).*

3.3.5 Numerical solution

We now turn to the computational aspect of Problem (3.14). First, note that for any tuple partition $\mathcal{T} = (\mathcal{T}_m)_{m=1}^M$ and probability tensor P, the KL divergence term can be decomposed as

$$\mathrm{KL}(P|P_{\#\mathcal{T}}) = \mathrm{KL}(P|\mu^{\otimes}) - \sum_{m=1}^{m} \mathrm{KL}_m(P), \tag{3.15}$$

where the function KL_m defined by $\mathrm{KL}_m(P) := \mathrm{KL}(P_{\#\mathcal{T}_m}|\mu^{\otimes}_{\mathcal{T}_m})$ is continuous and convex with respect to P. Now, Problem (3.14) becomes

$$\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu) = \inf_{P \in U(\mu)} \langle C, P\rangle + \varepsilon \mathrm{KL}(P|\mu^{\otimes}) - \varepsilon \sum_{m=1}^{m} \mathrm{KL}_m(P). \tag{3.16}$$

This is nothing but a Difference of Convex (DC) programming problem (which explains the name MMOT-DC), thanks to the convexity of the set $U(\mu)$ and the KL divergence. Thus, it can be solved by the classic DC algorithm [3] (Pham and Bernoussi, 1986; Pham and Le, 1997) as follows: at the iteration t,

3. The DC algorithm is very closely related to Convex-concave procedure, majorization-minimization algorithm, Successive Linear Approximation. See (Le and Pham, 2018) for more details.

Algorithm 4 DC algorithm for Problem (3.14).

Input. Cost tensor C, tuple partition $(\mathcal{T}_m)_{m=1}^M$, collection of histograms $\mu = (\mu_n)_{n=1}^N$, hyperparameter $\varepsilon > 0$, initialization $P^{(0)}$, tuple of initial dual vectors for the Sinkhorn step $(f_1^{(0)}, ..., f_N^{(0)})$.
Output. Tensor $P \in U(\mu)$.
While not converge

1. Gradient step: compute the gradient of the convex term $G^{(t)} = \sum_{m=1}^{M} \nabla_P \mathrm{KL}_m(P^{(t)})$.
2. Sinkhorn step: solve

$$P^{(t+1)} = \underset{P \in U(\mu)}{\operatorname{argmin}} \langle C - \varepsilon G^{(t)}, P \rangle + \varepsilon \mathrm{KL}(P|\mu^{\otimes}), \tag{3.18}$$

using the Sinkhorn algorithm 8, with the tuple of initial dual vectors $(f_1^{(0)}, ..., f_N^{(0)})$.

1. Calculate $G^{(t)} \in \partial(\sum_{m=1}^M \mathrm{KL}_m)(P^{(t)})$.
2. Solve $P^{(t+1)} \in \operatorname{argmin}_{P \in U(\mu)} \langle C - \varepsilon G^{(t)}, P \rangle + \varepsilon \mathrm{KL}(P|\mu^{\otimes})$.

This algorithm is very easy to implement. Indeed, the second step is an entropic-regularized MMOT problem, which admits a unique solution, thanks to the strict convexity of the objective function. Such solution can be found by the Sinkhorn algorithm 8. In the first step, the gradient can be calculated explicitly. For the sake of simplicity, we illustrate the calculation in a simple case, where $M = 2$ and $N = 4$ with $\mathcal{T}_1$ and $\mathcal{T}_2$ are two 2-tuples. The function $\mathrm{KL}_1 + \mathrm{KL}_2$ is continuous, so $G^{(t)} = \nabla_P(\mathrm{KL}_1 + \mathrm{KL}_2)(P^{(t)})$. Given a 4-D probability tensor P, we have

$$\frac{\partial(\mathrm{KL}_1 + \mathrm{KL}_2)}{\partial P_{i,j,k,l}} = \log\left(\frac{\sum_{k,l} P_{i,j,k,l}}{(\mu_1)_i(\mu_2)_j}\right) + \log\left(\frac{\sum_{i,j} P_{i,j,k,l}}{(\mu_3)_k(\mu_4)_l}\right). \tag{3.17}$$

The complete DC algorithm for Problem (3.16) can be found in Algorithm 4.

3.3.6 Experimental evaluation

In this section, we illustrate the use of MMOT-DC on simulated data. Rather than performing experiments in full generality, we choose the setting where $N = 4$ and $M = 2$ with $\mathcal{T}_1 = (1, 2)$ and $\mathcal{T}_2 = (3, 4)$, so that we can compare MMOT-DC with other popular solvers of COOT and GW distance. Given two matrices X and Y, we always consider the 4-D cost tensor C, where $C_{i,j,k,l} = |X_{i,k} - Y_{j,l}|^2$. On the other hand, we are not interested in the 4-D minimiser of MMOT-DC, but only in its two $\mathcal{T}_1, \mathcal{T}_2$-marginal matrices.

Solving COOT on a toy example. We generate a random matrix $X \in \mathbb{R}^{30\times 25}$, whose entries are drawn independently from the uniform distribution on the interval $[0, 1)$. We equip the rows and columns of X with two discrete uniform distributions on $[30]$ and $[25]$. We fix two

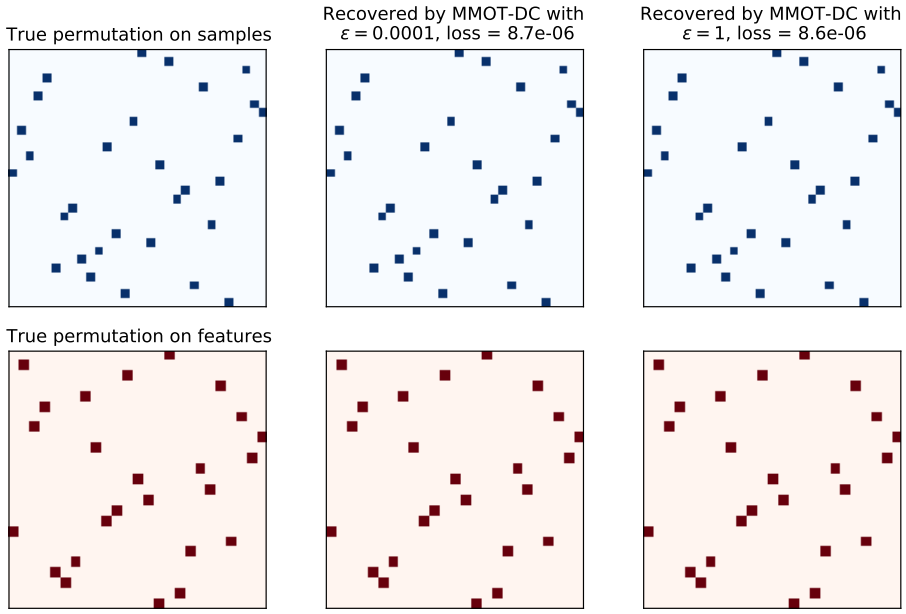

Figure 3.1 – Couplings generated by COOT and MMOT-DC on the matrix recovering task.

permutation matrices $Q_s \in \mathbb{R}^{30\times 30}$ (called sample permutation) and $Q_f \in \mathbb{R}^{25\times 25}$ (called feature permutation), then calculate $Y = Q_s X Q_f$. We also equip the rows and columns of Y with two discrete uniform distributions on [30] and [25].

It is not difficult to see that $\mathrm{COOT}(X, Y) = 0$ because (Q_s, Q_f) is a solution. As COOT is a special case of F-MMOT, we see that $\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu) = 0$, for every $\varepsilon > 0$, by Proposition 3.3.1. In this experiment, we will check if marginalizing the minimizer of MMOT-DC allows us to recover the permutation matrices Q_s and Q_f. As can be seen from Figure 3.1, MMOT-DC can recover the permutation positions, for various values of ε. On the other hand, it can not recover the true sparse permutation matrices because the Sinkhorn algorithm applied to the MMOT problem implicitly results in a dense tensor, thus having dense marginal matrices. For this reason, the loss only remains very close to zero, but never exactly. We also plot, with some abuse of notation, the histograms of the difference between the $(1,3), (1,4), (2,3), (2,4)$-marginal matrices of MMOT-DC and their corresponding counterparts from F-MMOT. In this example, in theory, as the optimal tensor P of F-MMOT can be factorized as $P = P_{\#\mathcal{T}_1} \otimes P_{\#\mathcal{T}_2} = Q_s \otimes Q_f$, it is immediate to see that $P_{\#(1,3)} = P_{\#(1,4)} = P_{\#(2,3)} = P_{\#(2,4)} \in \mathbb{R}^{30\times 25}$ are uniform matrices whose entries are $\frac{1}{750}$.

Quality of the MMOT-DC solutions. Now, we consider the situation where the true matching between two matrices is not known in advance and investigate the quality of the solutions returned by MMOT-DC to solve the COOT and GW problems. This means that we

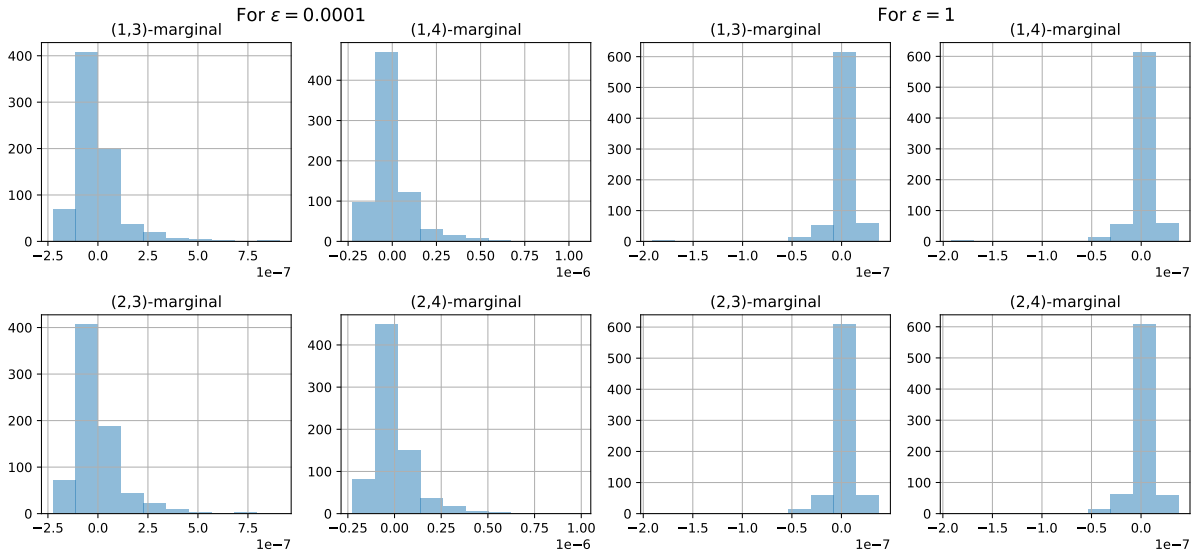

Figure 3.2 – Histograms of difference between true independent marginal matrices and their approximations. We see that the marginal matrices obtained by Algorithm 4 approximate well the theoretical uniform matrices.

will look at the COOT loss $\langle C, Q_s \otimes Q_f \rangle$, where the smaller the loss, the better when using both exact COOT and GW solvers and our relaxation.

We generate two random matrices $X \in \mathbb{R}^{20\times 3}$ and $Y \in \mathbb{R}^{30\times 2}$, whose entries are drawn independently from the uniform distribution on the interval $[0, 1)$. Then we calculate two corresponding squared Euclidean distance matrices of size 20 and 30. Their rows and columns are equipped with the discrete uniform distributions. In this case, Redko et al. (2020) show that the COOT loss coincides with the GW distance, and the Block Coordinate Descent (BCD) algorithm used to approximate COOT is equivalent to the Frank-Wolfe algorithm (Frank and Wolfe, 1956) used to solve the GW distance.

We compare four solvers:

1. The Frank-Wolfe algorithm to solve the GW distance (GW-FW).
2. The projected gradient algorithm to solve the entropic GW distance (Peyré, Cuturi, and Solomon, 2016) (EGW-PGD). We choose the regularization parameter from the set $\{0.0008, 0.0016, 0.0032, 0.0064, 0.0128, 0.0256\}$ and pick the one which corresponds to smallest COOT loss.
3. The Block Coordinate Descent algorithm to approximate the entropic COOT (Redko et al., 2020) (EGW-BCD), where two additional KL divergences corresponding to two couplings are introduced.

   The regularization parameters are tuned from the set $\{0, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$, where 0 means that there is no regularization term for the corresponding coupling and we

GW-FW	EGW-PGD	EGW-BCD	MMOT-DC
0.0829 (± 0.0354)	**0.0786 (± 0.0347)**	0.0804 (± 0.0353)	0.0822 (± 0.0364)

Table 3.2 – Average and standard deviation of COOT loss of the solvers. MMOT-DC is competitive to other solvers, except for EGW-PGD and EGW-BCD.

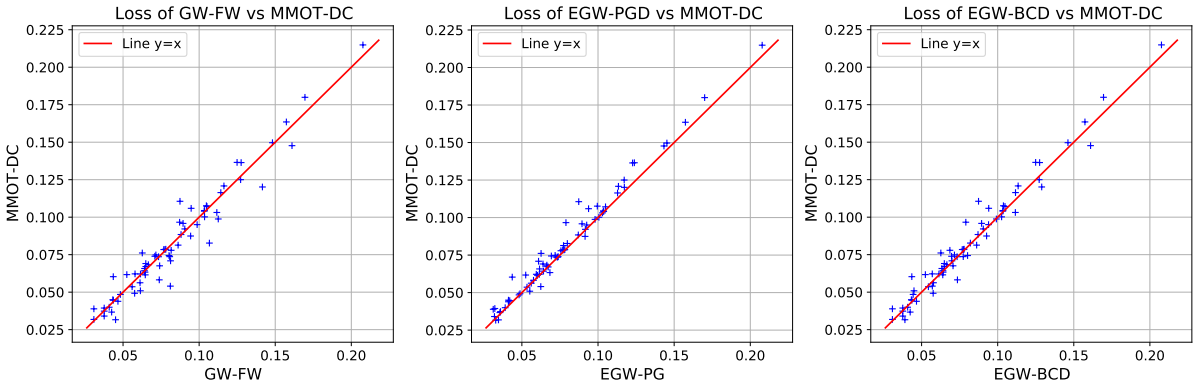

Figure 3.3 – Scatter plots of MMOT-DC versus other solvers. In all three plots, the points tend to concentrate around the line $y = x$, which indicates the comparable performance of MMOT-DC. On the other hand, the top-right plot shows the clear superiority of EGW-PGD.

pick the pair corresponding to the smallest COOT loss.

4. Algorithm 4 to solve the MMOT-DC. We tune $\varepsilon \in \{1, 1.4, 1.8, 2.2, 2.6\}$ and we pick the one which corresponds to smallest COOT loss.

For GW-FW and EGW-PGD, we use the implementation from the Python Optimal Transport package (Flamary et al., 2021).

Given two random matrices, we record the COOT loss corresponding to the solution generated by each method. We simulate this process 70 times and compare their overall performance. We can see in Table 3.2 the average value and standard deviation and the comparison for the values of the loss between the different algorithms in Figure 3.3. The performance is quite similar across methods with a slight advantage for EGW-PGD. This is in itself a very interesting result that has never been noted, to the best of our knowledge: the reason that the entropic version of GW can provide better solution than solving the exact problem, may be due to the "convexification" of the problem, thanks to the entropic regularization. Our approach is also interestingly better than the exact GW-FW, which illustrates that the relaxation might help in finding better solutions despite the non-convexity of the problem.

An empirical variation. Intuitively, for sufficiently large ε, the minimisation of the KL divergence is prioritised over the linear term in the objective function of the MMOT-DC problem,

which implies that the optimal tensor P^* is "close" to its corresponding tensor product $P^*_{\#\mathcal{T}}$. So, instead of calculating the gradient at P, one may calculate at $P_{\#\mathcal{T}}$. In this case, the gradient reads

$$\sum_{m=1}^{M} \nabla_P \mathrm{KL}_m(P_{\#\mathcal{T}}) = \log \frac{P_{\#\mathcal{T}_1}}{\mu^{\otimes}_{\mathcal{T}_1}} \oplus \ldots \oplus \log \frac{P_{\#\mathcal{T}_m}}{\mu^{\otimes}_{\mathcal{T}_m}}, \tag{3.19}$$

where $\oplus$ represents the tensor sum operator between two arbitrary-size tensors: $(A \oplus B)_{i,j} := A_i + B_j$, where with some abuse of notation, i or j can be understood as a tuple of indices. Thus, we avoid storing the N-D gradient tensor (as in the Algorithm 4) and only need to store M smaller-size tensors. Not only saving the memory, this variation also seems to be empirically competitive with the original Algorithm 4, if not sometimes better, in terms of COOT loss. The underlying reason might be related to the approximate DCA scheme (Vo, 2015), where one replaces both steps in each DC iteration by their approximation. We leave the formal theoretical justification of this variation to the future work. We call this variation *MMOT-DC-v1* and use the same setup as in Experiment 3.3.6.

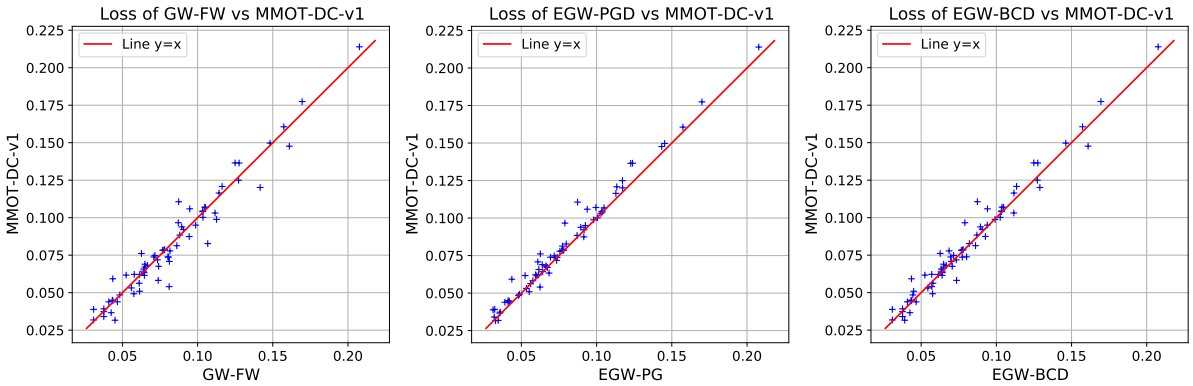

Figure 3.4 – Scatter plots of MMOT-DC-v1 versus other solvers. In all three plots, the points tend to concentrate around the line $y = x$, which indicates the comparable performance of MMOT-DC-v1. On the other hand, the top-right plot shows the clear superiority of EGW-PGD.

MMOT-DC	MMOT-DC-v1
0.0822 ($\pm$ 0.0364)	0.0820 ($\pm$ 0.0361)

Table 3.3 – Average and standard deviation of COOT loss of MMOT-DC and MMOT-DC-v1. The performance of the two algorithms is very similar.

3.3.7 Discussion

In this section, we present a novel relaxation of the factorized MMOT problem called *MMOT-DC*. More precisely, we replace the hard constraint on factorization constraint by a smooth regularization term. The resulting problem not only enjoys an interpolation property between MMOT and factorized MMOT, but also is a DC problem, which can be solved easily by the DC algorithm. We illustrate the use of MMOT-DC the via some simulated experiments and show that it is competitive with the existing popular solvers of COOT and GW distance. One limitation of the current DC algorithm is that, it is not scalable because it requires storing a full-size tensor in the gradient step computation. Thus, future work may focus on more efficiently designed algorithms, in terms of both time and memory footprint. Moreover, incorporating additional structure on the cost tensor may also be computationally and practically beneficial. From a theoretical viewpoint, it is also interesting to study the extension of MMOT-DC to the continuous setting, which can potentially allow us to further understand the connection between GW distance and COOT.

Chapter 4

Unbalanced Co-Optimal Transport

This chapter summarizes the results from the paper (Tran et al., 2023) and addresses the unbalaced extension of Co-Optimal transport. Optimal transport (OT) compares probability distributions by computing a meaningful alignment between their samples. Co-optimal transport (COOT) takes this comparison further by inferring an alignment between features as well. While this approach leads to better alignments and generalizes both OT and Gromov-Wasserstein distances, we provide a theoretical result showing that it is sensitive to outliers that are omnipresent in real-world data. This prompts us to propose unbalanced COOT for which we provably show its robustness to noise in the compared datasets. To the best of our knowledge, this is the first such result for OT methods in incomparable spaces. With this result in hand, we provide empirical evidence of this robustness for the challenging tasks of heterogeneous domain adaptation with and without varying proportions of classes and simultaneous alignment of samples and features across single-cell measurements.

4.1 Introduction

The last decade has witnessed many successful applications of optimal transport (OT) (Kantorovich, 1942; Monge, 1781) in machine learning, namely in domain adaptation (Courty et al., 2016), generative adversarial networks (Arjovsky, Chintala, and Bottou, 2017), classification (Frogner et al., 2015), dictionary learning (Rolet, Cuturi, and Peyré, 2016), semi-supervised

learning (Solomon et al., 2014). When the supports of the probability measures lie in the same ground metric space, it is natural to use the distance defined by the metric to induce the cost, which leads to the famous Wasserstein distance (Villani, 2003). When they do not, one can rely on the idea of Gromov-Hausdorff distance (Gromov, 1981) and its equivalent reformulations (Burago, Burago, and Ivanov, 2001; Gromov, 1999; Kalton and Ostrovskii, 1999), and adapt them to the setting of metric measure spaces (Gromov, 1999). This results in, for example, the Gromov-Wasserstein (GW) distance (Mémoli, 2007, 2011b; Sturm, 2012), which has been widely used in many applications, namely in shape matching (Mémoli, 2011b), comparing kernel matrices (Peyré, Cuturi, and Solomon, 2016), graphs (Vayer et al., 2019a; Xu, Luo, and Carin, 2019; Xu et al., 2019), computational biology (Demetci et al., 2022a), heterogeneous domain adaptation (Yan et al., 2018), correspondence alignment (Solomon et al., 2016), machine translation (Alvarez-Melis and Jaakkola, 2018).

By construction, the GW distance can only provide the sample alignment that best preserves the intrinsic geometry of the distributions and, as such, compares square pairwise relationship matrices. The CO-Optimal transport (COOT) (Chowdhury et al., 2023; Redko et al., 2020) goes beyond these limits by simultaneously learning two independent (feature and sample) correspondences, and thus provides greater flexibility over the GW distance in terms of usage and interpretability. First, it allows us to measure similarity between arbitrary-size matrices. An interesting use case is, for instance, on tabular data, which are usually expressed as a matrix whose rows represent samples and columns represent features. For the GW distance, the similarity or distance matrix (or any square matrix derived from the data) must be calculated in advance and the effect of the individual variables is lost during this computation. On the other hand, COOT can bypass this step as it can use either the tabular data directly or the similarity matrices as inputs. Second, COOT provides both sample and feature correspondences. These feature correspondences are also interpretable and allow to recover relations between the features of two different datasets even when they do not lie in the same space.

Similar to classical OT, COOT enforces hard constraints on the marginal distributions both between samples and features. These constraints lead to two main limitations: (1) unbalanced datasets where samples or features cannot be accurately matched; (2) mass transportation *must* be exhaustive: outliers, if any, must be matched regardless of the cost they induce. To circumvent these limitations, we propose to relax the mass preservation constraints in the COOT distance and study a broadly applicable and general OT framework that includes several well-studied cases presented in Table 4.1.

Related work. To relax the OT marginal constraints, a straightforward solution is to control the difference between the marginal distributions of the transportation plan and the data by some discrepancy measure, e.g., Kullback-Leibler divergence. In classical OT, this gives rise

	Across spaces	Sample alignment	Feature alignment	Robust to outliers
OT	✗	✓	✗	✗(Fatras et al., 2021)
UOT	✗	✓	✗	✓(Fatras et al., 2021)
GW	✓	✓	✗	✗(Prop. 4.3.1)
UGW	✓	✓	✗	✓(Thm. 4.3.1)
COOT	✓	✓	✓	✗(Prop. 4.3.1)
UCOOT	✓	✓	✓	✓(Thm. 4.3.1)

Table 4.1 – Properties of different OT formulations generalized by UCOOT. The proposed UCOOT is not only able to learn informative feature alignments, but also robust to outliers.

to the unbalanced OT (UOT), which was first proposed by Benamou (2003). The theoretical and numerical aspects of this extension have been studied extensively (Chizat et al., 2018a,b; Liero, Mielke, and Savaré, 2018; Pham et al., 2020) and are gaining increasing attention in the machine learning community, with wide-range applications, namely in domain adaptation (Fatras et al., 2021), generative adversarial networks (Balaji, Chellappa, and Feizi, 2020; Yang and Uhler, 2019), dynamic tracking (Lee, Bertrand, and Rozell, 2019), crowd counting (Ma et al., 2021), neuroscience (Bazeille et al., 2019a; Janati et al., 2019) or modeling cell developmental trajectories (Schiebinger et al., 2019).

Unbalanced OT and its variants are usually sought for their known robustness to outliers (Balaji, Chellappa, and Feizi, 2020; Fatras et al., 2021; Mukherjee et al., 2021). This appealing property goes beyond classical OT. For instance, to compare signed and non-negative measures in incomparable spaces, unbalanced OT (Liero, Mielke, and Savaré, 2018) can be blended with the L_p-transportation distance (Sturm, 2006), which leads to the Sturm-Entropic-Transport distance (Ponti and Mondino, 2020), or with the GW distance, which gives rise to the unbalanced GW (UGW) distance (Séjourné, Vialard, and Peyré, 2021b). Also motivated by the unbalanced OT, Zhang et al. (2022b) proposed a relaxation of the bidirectional Gromov-Monge distance called unbalanced bidirectional Gromov-Monge divergence.

Contributions. In this work, we introduce an unbalanced extension of COOT called "Unbalanced CO-Optimal transport" (UCOOT). UCOOT – defined for both discrete and continuous data – is a general framework that encompasses all the OT variants displayed in Table 4.1. Our main contribution is to show that UCOOT is provably robust to both samples and features outliers, while its balanced counterpart can be made arbitrarily large with strong enough perturbations. To the best of our knowledge, this is the first time such a general robustness result is established for OT across different spaces. Our theoretical findings are showcased in unsupervised heterogeneous domain adaptation and single-cell multi-omic data alignment, demonstrating a very competitive performance.

4.2 From COOT to Unbalanced Co-Optimal Transport

The ultimate goal behind the CO-Optimal Transport (COOT) framework is the simultaneous alignment of samples *and* features to allow for comparisons across spaces of different dimensions. In this section, we discuss OT formulations including OT, UOT, GW, UGW and COOT, then introduce the proposed UCOOT and show how the aforementioned distances fall into our framework.

From sample alignment to sample-feature alignment. Let (X_1^s, μ_1^s) and (X_2^s, μ_2^s) be a pair of compact measure spaces such that X_1^s and X_2^s belong to some common metric space $(\mathcal{E}, d)$. Classical optimal transport infers one alignment (or joint distribution) $\pi^s \in \mathcal{M}^+(X_1^s \times X_2^s)$ with marginals $(\pi^s_{\#1}, \pi^s_{\#2})$ close to (μ_1^s, μ_2^s) according to some appropriate divergence D such that the cost $\int c(x_1, x_2)\ \mathrm{d}\pi^s(x_1, x_2) + D(\pi^s_{\#1}|\mu_1^s) + D(\pi^s_{\#2}|\mu_2^s)$ is minimal. For instance, in balanced (resp. unbalanced) OT, D corresponds to the indicator divergence (resp. KL divergence or TV). To define a generalized OT beyond one single alignment, we must first introduce a new pair of measure spaces (X_1^f, μ_1^f) and (X_2^f, μ_2^f). Intuitively, the two transport plans that must be inferred: π^s across *samples* and π^f across *features*, must minimize a cost of the form $\iint c((x_1^s, x_1^f), (x_2^s, x_2^f))\ \mathrm{d}\pi^s(x_1^s, x_2^s)\ \mathrm{d}\pi^f(x_1^f, x_2^f)$ where $c((x_1^s, x_1^f), (x_2^s, x_2^f))$ is the *joint* cost of aligning the sample-feature pairs (x_1^s, x_1^f) and (x_2^s, x_2^f).

However, unlike OT, there is no underlying ambient metric space in which comparisons between these pairs are straightforward. Thus, we consider a simplified cost of the form: $c((x_1^s, x_1^f), (x_2^s, x_2^f)) = |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p$, for $p \geq 1$ and some scalar functions ξ_1, ξ_2 that define the sample-feature interactions. A similar definition was adopted by Chowdhury et al. (2023) to extend COOT to the continuous setting in the context of hypergraphs. Formally, our general formulation takes pairs of *sample-feature spaces* defined as follows.

Definition 4.2.1 (Sample-feature space)**.** *Let (X^s, μ^s) and (X^f, μ^f) be compact Polish measure spaces, where $\mu^f \in \mathcal{M}^+(X^f)$ and $\mu^s \in \mathcal{M}^+(X^s)$. Let ξ be a scalar integrable function in $L^p(X^s \times X^f, \mu^s \otimes \mu^f)$. We call the triplet $\mathcal{X} = ((X^s, \mu^s), (X^f, \mu^f), \xi)$ a sample-feature space and ξ is called an interaction.*

Definition 4.2.2 (Generalized COOT)**.** *Given two divergences D_1 and D_2, we define the generalized COOT of order p between $\mathcal{X}_1 = ((X_1^s, \mu_1^s), (X_1^f, \mu_1^f), \xi_1)$ and $\mathcal{X}_2 = ((X_2^s, \mu_2^s), (X_2^f, \mu_2^f), \xi_2)$ by:*

$$\inf_{\substack{\pi^s \in \mathcal{M}^+(X_1^s \times X_2^s) \\ \pi^f \in \mathcal{M}^+(X_1^f \times X_2^f) \\ m(\pi^s) = m(\pi^f)}} \underbrace{\iint |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p \ \mathrm{d}\pi^s\ \mathrm{d}\pi^f}_{\textit{transport cost of sample-feature pairs}} + \underbrace{\sum_{k=1}^{2} \rho_k D_k(\pi^s_{\#k} \otimes \pi^f_{\#k} | \mu_k^s \otimes \mu_k^f)}_{\textit{mass destruction / creation penalty}}, \tag{4.1}$$

for $\rho_1, \rho_2 > 0$ and $p \geq 1$.

As the multiplicative nature between π^s and π^f leads to an invariance by the scaling map $\alpha \mapsto (\alpha\pi^s, \frac{1}{\alpha}\pi^f)$, for $\alpha > 0$, we further impose the equal mass constraint $m(\pi^s) = m(\pi^f)$.

It is worth mentioning that Formulation (4.1) is not the only way to relax the marginal constraints. For example, instead of $D_k(\pi^s_{\#k} \otimes \pi^f_{\#k} | \mu^s_k \otimes \mu^f_k)$, one can consider $D_k(\pi^s_{\#k}|\mu^s_k) + D_k(\pi^f_{\#k}|\mu^f_k)$, or $D_s(\pi^s_{\#1} \otimes \pi^s_{\#2} | \mu^s_1 \otimes \mu^s_2)$, for some divergence D_s. However, amongst these choices, ours is the only one which can be recast as a variation of the unbalanced OT problem. This allows us to leverage the known techniques in unbalanced OT to justify the theoritical and practical properties, namely Proposition 4.2.1 and Theorem 4.3.1 below.

Note that the problem above is very general and can, with some additional constraints, recover exact OT, UOT, GW, UGW, COOT (see Table 4.2). In particular, if the measures (μ^s_1, μ^s_2) and (μ^f_1, μ^f_2) are probability measures, then setting $D_1 = D_2 = \iota_=$ leads to the COOT problem first introduced in the discrete case in (Redko et al., 2020) and recently generalized to the continuous setting in (Chowdhury et al., 2023) [1]. In this work, we relax the hard constraints and consider a more flexible formulation with the KL divergence:

Definition 4.2.3 (UCOOT)**.** *We define Unbalanced COOT (UCOOT) as in Equation (4.1) with $D_1 = D_2 = KL$. We write $UCOOT_\rho(\mathcal{X}_1, \mathcal{X}_2)$ to indicate the UCOOT between two sample-feature spaces $\mathcal{X}_1$ and $\mathcal{X}_2$, for a given pair of hyperparameters $\rho = (\rho_1, \rho_2)$.*

While various properties of the divergences D_k have been extensively studied in the context of unbalanced OT by several authors (Chizat, 2017; Frogner et al., 2015), the concept of sample-feature interaction requires more clarification. Let us consider some simple examples. In the discrete case, we consider n observations of d features represented by matrix $A \in \mathbb{R}^{n\times d}$. In this case, the space X^s (resp. X^f) is not explicitly known but can be characterized by the finite set $[n]$ (resp. $[d]$), up to an isomorphism. Assuming that all samples (resp. features) are equally important, the discrete empirical measures can be given by uniform weights $\mu^s = \frac{1_n}{n}$ (resp. $\mu^f = \frac{1_d}{d}$). The most natural sample-feature interaction ξ is simply the index function $\xi(i,j) = A_{ij}$. In the continuous case, we assume that data stream from a continuous random variable $a \sim \mu_s \in \mathcal{P}(\mathbb{R}^d)$ for which an interaction function can be $\xi(a,j) = a_j$.

Proposition 4.2.1. *For any D_1, $D_2 \in \{\iota_=, KL\}$, Problem (4.2.2) (in Equation (4.1)) admits a minimizer.*

Remark 4.2.1. The existence of minimizer shown in Proposition 4.2.1 can be extended to a larger family of Csiszár divergences (Csiszár, 1963). A general proof is given in the Appendix.

Relation between UCOOT and COOT. When μ^s_k and μ^f_k are probability measures, for $k = 1, 2$, then $\text{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) \leq \text{COOT}(\mathcal{X}_1, \mathcal{X}_2)$, for any $\rho_1, \rho_2 > 0$. Moreover, $\text{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) =$

1. Note that, Chowdhury et al. (2023) consider **bounded measurable** functions on the Polish measure space, whereas we work with **integrable** functions on the **compact** Polish measure space.

	Shape of inputs	Coupling constraint	Scalar function	Divergence
OT	$d_1 = d_2$	$\pi^f = I_{d_1} = I_{d_2}$	$\xi(i,j) = A_{ij}$	$\iota_=$
GW	$n_1 = d_1$, $n_2 = d_2$	$\pi^f = \pi^s$	$\xi(i,j) = \text{dist}(A_{i.}, A_{j.})$	$\iota_=$
COOT	–	–	$\xi(i,j) = A_{ij}$	$\iota_=$
semi-d. COOT	–	–	$\xi(a,j) = a_j$	$\iota_=$
UCOOT	–	–	$\xi(i,j) = A_{ij}$	KL

Table 4.2 – Conditions under which different OT formulations fall within the generalized framework of Definition 4.2.2. "semi-d" refers to "semi-discrete" setting, where μ_s is a continuous probability and $\mu_d = 1_d/d$. Here, I_d is the identity matrix in $\mathbb{R}^d$.

0 if and only if $\text{COOT}(\mathcal{X}_1, \mathcal{X}_2) = 0$. In particular, suppose that $\mathcal{X}_1$ and $\mathcal{X}_2$ are two finite sample-feature spaces such that (X_1^s, X_2^s) and (X_1^f, X_2^f) have the same cardinality and are equipped with the uniform measures $\mu_1^s = \mu_2^s$, $\mu_1^f = \mu_2^f$. Then $\text{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = 0$ if and only if there exist perfect alignments between rows (samples) and between columns (features) of the interaction matrices ξ_1 and ξ_2.

Relation between UCOOT and its solution. As a consequence of Corollary 2.1.1, UCOOT is a bilinear function of its minimizers, thanks to the property of the KL divergence. More precisely, if (π_*^s, π_*^f) is the equal-mass solution, then we have

$$\text{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = \sum_{k=1,2} \rho_k \, m(\mu_k^s) \, m(\mu_k^f) - (\rho_1 + \rho_2) \, m(\pi_*^s) m(\pi_*^f). \tag{4.2}$$

Interestingly, this equation depends only on the measures and hyperparameters. There is no dependency on the interactions because they have been fully captured in the masses of the optimal sample and feature couplings.

4.3 Robustness of Unbalanced Co-Optimal Transport

When discussing the concept of robustness, outliers are often considered as samples not following the underlying distribution of the data. In our general context of sample-feature alignments, we consider a *pair* $(x^s, x^f) \in X^s \times X^f$ to be an outlier if the magnitude of $|\xi(x^s, x^f)|$ is abnormally larger than other interactions between X^s and X^f. As a result, such outliers lead to abnormally large transportation costs $|\xi_1 - \xi_2|$. To study the robustness of COOT and UCOOT, we consider an outlier scenario where the marginal data distributions are contaminated by some additive noise distribution.

Assumption 4.3.1. *Consider two sample-feature spaces $\mathcal{X}_1$ and $\mathcal{X}_2$. Let ε^s (resp. ε^f) be a probability measure with compact support O^s (resp. O^f). For $a \in \{s, f\}$, define the noisy distribution $\widetilde{\mu}^a = \alpha_a \mu^a + (1 - \alpha_a)\varepsilon^a$, where $\alpha_a \in [0, 1]$. We assume that ξ_1 is defined on*

$(X_1^s \cup O^s) \times (X_1^f \cup O^f)$ *and that* ξ_1, ξ_2 *are continuous on their supports. We denote the contaminated sample-feature space by* $\widetilde{\mathcal{X}_1} = ((X_1^s \cup O^s, \widetilde{\mu}_1^s), (X_1^f \cup O^f, \widetilde{\mu}_1^f), \xi_1)$*. Finally, we define some useful minimal and maximal costs:*

$$\begin{cases} \Delta_0 := \min\limits_{\substack{x_1^s \in O^s, x_1^f \in O^f \\ x_2^s \in X_2^s, x_2^f \in X_2^f}} |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p \\ \Delta_\infty := \max\limits_{\substack{x_1^s \in X_1^s \cup O^s, x_1^f \in X_1^f \cup O^f \\ x_2^s \in X_2^s, x_2^f \in X_2^f}} |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p \ . \end{cases} \tag{4.3}$$

Here, Δ_0 *accounts for the minimal deviation of the cost between the outliers and target support, while* Δ_∞ *is the maximal deviation between the contaminated source and the target.*

The exact marginal constraints of COOT enforce conservation of mass. Thus, outliers *must* be transported no matter how large their transportation costs are. This intuition is captured by the following result.

Proposition 4.3.1 (COOT is sensitive to outliers)**.** *Consider* $\widetilde{\mathcal{X}_1}, \mathcal{X}_2$ *as defined in Assumption 4.3.1. Then*

$$COOT(\widetilde{\mathcal{X}_1}, \mathcal{X}_2) \geq (1 - \alpha_s)(1 - \alpha_f)\Delta_0. \tag{4.4}$$

Whenever the outlier proportion $(1 - \alpha_s)(1 - \alpha_f)$ is positive, COOT increases with the distance between the supports of the outliers and those of the clean data. Thus, the right hand side of Proposition 4.3.1 can be made arbitrarily large by taking outliers far from the supports of the clean data.

We can now state our main theoretical contribution. Relaxing the marginal constraints leads to a loss that saturates as outliers get further from the data:

Theorem 4.3.1 (UCOOT is robust to outliers)**.** *Consider two sample-feature spaces* $\widetilde{\mathcal{X}_1}, \mathcal{X}_2$ *as defined in Assumption 4.3.1. Let* $\delta := 2(\rho_1 + \rho_2)(1 - \alpha_s\alpha_f)$ *and* $K = M + \frac{1}{M} UCOOT(\mathcal{X}_1, \mathcal{X}_2) + \delta$*, where* $M = m(\pi^s) = m(\pi^f)$ *is the transported mass between clean data. Then:*

$$UCOOT(\widetilde{\mathcal{X}_1}, \mathcal{X}_2) \leq \alpha_s\alpha_f\, UCOOT(\mathcal{X}_1, \mathcal{X}_2) + \delta M \left[1 - \exp\left(-\frac{\Delta_\infty(1 + M) + K}{\delta M}\right)\right]. \tag{4.5}$$

The proof of Theorem 4.3.1 is provided in the Appendix and inspired from (Fatras et al., 2021), but in a much more general setting: (1) it covers both sample and feature outliers and (2) considers a noise distribution instead of a Dirac. Note that the bound in Proposition 4.3.1 indicates that outliers can make COOT arbitrary large, while UCOOT is upper bounded and discards the mass of outliers with high transportation cost.

This is well illustrated in Figure 4.1, where we simulate outliers by adding a perturbation to a row of the interaction matrix. More precisely, we first generate a matrix $A \in \mathbb{R}^{20\times 15}$ by $A_{ij} = \cos(\frac{i}{20}\pi) + \cos(\frac{j}{15}\pi)$. Then, we replace its last row by $\tau 1_{15}$, for $\tau \geq 0$. Figure 4.1 depicts COOT and UCOOT between A and its modified version as a function of τ. The higher the value of τ, the more likely that the last row contains the interaction of outliers. Consequently, as τ increases, so does COOT but at a much higher pace, whereas UCOOT remains stable.

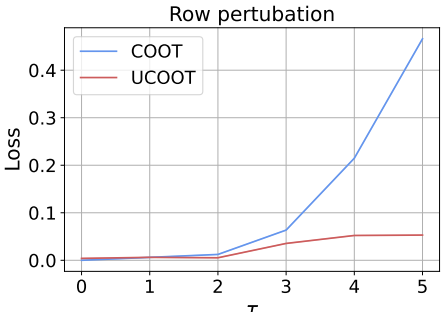

Figure 4.1 – Sensitivity of COOT and UCOOT under the presence of outliers.

It should be noted that, with minimal adaptation, Theorem 4.3.1 also holds for the unbalanced GW (UGW) distance. This provides a theoretical explanation of the empirical observation in (Séjourné, Vialard, and Peyré, 2021b) that unlike GW, the UGW distance is also robust to outliers.

4.4 Optimization algorithm and complexity

Solving COOT-type problems, in general, is not trivial. As highlighted in (Redko et al., 2020), the balanced case corresponds to a convex relaxation of the bilinear assignment problem, which seeks the pair of permutations minimizing the transport cost. Here we argue that relaxing the marginal constraints makes the problem easier in two different aspects: (1) the obtained problem is easier to solve through a sequence of GPU friendly iterations; (2) regularization leads to lower alignment costs and thus better local minima. In this section, we first describe how to compute UCOOT in practice.

Optimization strategy We consider two tabular datasets $A \in \mathbb{R}^{n_1\times d_1}$ and $B \in \mathbb{R}^{n_2\times d_2}$. Let u_k be the uniform histogram over sample-feature pairs: $u_k := \frac{1}{n_k d_k} 1_{n_k} \otimes 1_{d_k}$, for $k = 1, 2$. For the sake of simplicity, we assume uniform weights over both samples and features. Computing UCOOT can be done using block-coordinate descent (BCD) both with and without entropy regularization. More precisely, given a hyperparameter $\varepsilon \geq 0$, discrete UCOOT can be written as:

$$\begin{aligned}\min_{\substack{\pi^s,\pi^f \\ m(\pi^s)=m(\pi^f)}} & \sum_{i,j,k,l} (A_{ik} - B_{jl})^2 \pi^s_{ij}\pi^f_{kl} + \rho_1 \mathrm{KL}(\pi^s 1_{n_1} \otimes \pi^f 1_{d_1} | u_1) \\ & + \rho_2 \mathrm{KL}({\pi^s}^\top 1_{n_2} \otimes {\pi^f}^\top 1_{d_2} | u_2) + \varepsilon \mathrm{KL}(\pi^s \otimes \pi^f | \mu^s_1 \otimes \mu^s_2 \otimes \mu^f_1 \otimes \mu^f_2).\end{aligned} \tag{4.6}$$

The only difference between $\varepsilon = 0$ and $\varepsilon > 0$ lies in the inner-loop algorithm used to update one

Algorithm 5 BCD algorithm to solve UCOOT

Input: $A \in \mathbb{R}^{n_1,d_1}, B \in \mathbb{R}^{n_2,d_2}$, $\rho_1, \rho_2, \varepsilon$
Initialize π^s and π^f
repeat
  Update π^s using Sinkhorn or MM
  Rescale $\pi^s = \sqrt{\frac{m(\pi^f)}{m(\pi^s)}}\pi^s$
  Update π^s using Sinkhorn or MM
  Rescale $\pi^f = \sqrt{\frac{m(\pi^s)}{m(\pi^f)}}\pi^f$
until convergence

of transport plans (π^s, π^f) while the other one remains fixed. For $\varepsilon = 0$, we use the Majorization-Minimization (MM) algorithm (Chapel et al., 2021), which leads to a multiplicative update on the transport plan. For $\varepsilon > 0$, updating each transport plan boils down to an entropic UOT problem, which can be solved efficiently using the unbalanced variant of Sinkhorn's algorithm (Chizat et al., 2018a). The main benefit of entropy regularization is to reduce the number of variables from $(n_1 \times n_2) + (d_1 \times d_2)$ to $n_1 + n_2 + d_1 + d_2$. Moreover, by taking ε sufficiently small, we can recover solutions close to those in the non-entropic case. We formalize this claim in the following result.

Proposition 4.4.1. *Let $(\pi^s_\varepsilon, \pi^f_\varepsilon)$ be an equal-mass solution of the problem $UCOOT_{\rho,\varepsilon}(\mathcal{X}_1, \mathcal{X}_2)$. Denote $\mu^s = \mu^s_1 \otimes \mu^s_2$ and $\mu^f = \mu^f_1 \otimes \mu^f_2$.*

1. *When $\varepsilon \to \infty$, we have $\pi^s_\varepsilon \rightharpoonup \sqrt{\frac{m(\mu^f)}{m(\mu^s)}}\mu^s$ and $\pi^f_\varepsilon \rightharpoonup \sqrt{\frac{m(\mu^s)}{m(\mu^f)}}\mu^f$.*
2. *When $\varepsilon \to 0$, we have*
   (a) *$UCOOT_{\rho,\varepsilon}(\mathcal{X}_1, \mathcal{X}_2) \to UCOOT_{\rho}(\mathcal{X}_1, \mathcal{X}_2)$ and $m(\pi^s_\varepsilon) \to m(\pi^s_*)$, for any equal-mass solution (π^s_*, π^f_*) of the unregularized problem.*
   (b) *Any cluster point $(\widehat{\pi}^s, \widehat{\pi}^f)$ of the sequence $(\pi^s_\varepsilon, \pi^f_\varepsilon)_\varepsilon$ is an equal-mass solution of the unregularized problem. Furthermore,*

$$KL(\widehat{\pi}^s \otimes \widehat{\pi}^f | \mu^s \otimes \mu^f) = \min_{(\pi^s, \pi^f)} KL(\pi^s \otimes \pi^f | \mu^s \otimes \mu^f), \tag{4.7}$$

where the infimum is taken over all solutions of the unregularized problem.

4.5 Experiments

4.5.1 Illustration and interpretation on MNIST images

We illustrate the robustness of UCOOT and its ability to learn meaningful feature alignments under the presence of both sample and feature outliers in the MNIST dataset. We introduce

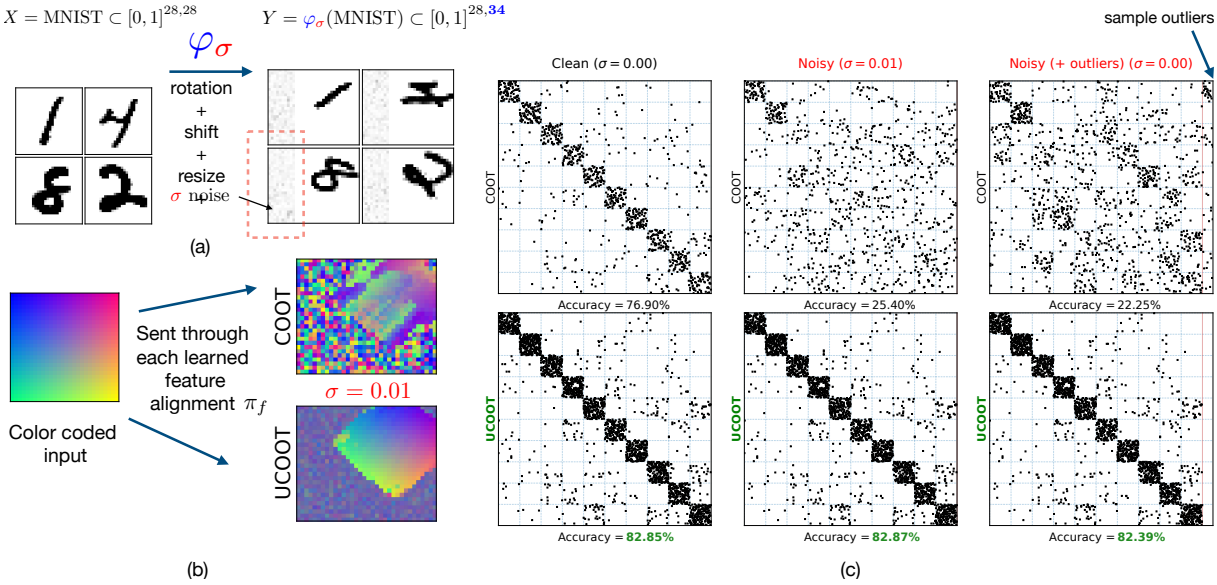

Figure 4.2 – Example illustrating the feature alignment π_f learned by UCOOT and its robustness to outliers. **(a)** Visualization of 4 random samples from both datasets. The added Gaussian noise only affects the first 10 columns of the images and is different across images. **(b)** The barycentric mapping (see Appendix for details) defined by UCOOT learns the transformation defined by φ_σ while disregarding non-informative features. **(c)** Alignments across samples from X and Y. We contaminated the target Y with 50 sample outliers (images with uniform entries in $[0, 1]$). A very small amount of noise is sufficient to derail COOT. Unlike COOT, UCOOT does not transport any outlier sample. Accuracy is computed as the percentage of mass within the block-diagonal structure.

the feature outliers by applying a handcrafted transformation φ_σ that performs a zero-padding (shift), a 45° rotation, a resize to (28, 34) and adds Gaussian noise $\mathcal{N}(0, \sigma^2)$ entries to the first 10 columns of the image.

Figure 4.2 (a) shows some examples of original and transformed images. We randomly sample 100 images per class (1000 total) from $X = \text{MNIST}$ and $Y = \varphi_\sigma(\text{MNIST})$. Regarding the sample outliers, we add 50 random images with uniform entries in [0, 1] to the target data Y. We then compute the optimal COOT and UCOOT alignments shown in Figure 4.2 (b) and (c). The flexibility of UCOOT with respect to mass transportation allows it to completely disregard: (1) noisy and uninformative pixels (features), which are all given the same weight as depicted by (b); (2) all the sample outliers of which none are transported as shown by the last blank column of the alignment (c). Moreover, notice how the color-coded input image is transformed according to the transformation φ_σ despite the fact that no spatial information is provided in the OT problem. On the other hand, a very small perturbation ($\sigma = 0.01$) is enough for the sample alignment given by COOT to lose its block-diagonal dominant structure (class information is lost), while the UCOOT alignment remains unscathed.

One may wonder whether the performance of UCOOT would still hold for different values of σ. Figure 4.3 answers this question positively. For $\sigma > 0$, we compute the average accuracy (defined by the percentage of mass within the block-diagonal structure) over 20 different runs. The performance of COOT not only degrades with noisier outliers but is also unstable. By contrast, the accuracy of UCOOT remains almost constant regardless of the level of noise.

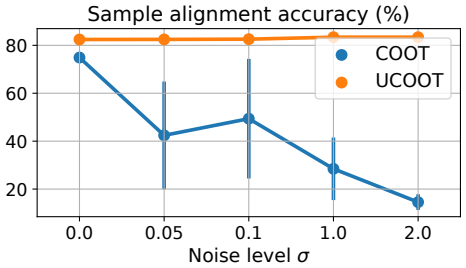

Figure 4.3 – Robustness of UCOOT vs. COOT on MNIST example, at different noise levels.

4.5.2 Heterogeneous Domain Adaptation

We now investigate the application of discrete UCOOT in semi-supervised and unsupervised Heterogeneous Domain Adaptation (HDA). It is a particularly difficult problem where one aims to predict classes on unlabeled data using labeled data lying in a different space. OT methods across spaces have recently shown good performance on such tasks, in particular using GW distance (Yan et al., 2018) and COOT (Redko et al., 2020).

Datasets and experimental setup. We consider the Caltech-Office dataset (Saenko et al., 2010) containing three domains: Amazon (A) (1123 images), Caltech-256 (C) (958 images) and Webcam (W) (295 images) with 10 overlapping classes amongst them. The image in each domain is representated by the output of the second last layer in the Google Net (Szegedy et al., 2015) and Caffe Net (Jia et al., 2014) neural network architectures, which results in 4096 and 1024-dimensional vectors, respectively (thus $d_s = 4096, d_t = 1024$). We compare 4 OT-based methods: GW, COOT, UGW, and UCOOT. For the semi-supervised HDA task, we additionally use k-NN, with $k = 3$ as baseline method, which corresponds to the situation where there is no adaptation. The hyper-parameters for each method are validated on a unique pair of datasets (W→W), then fixed for all other pairs in order to provide truly unsupervised HDA generalization.

We follow the same experimental setup as in (Redko et al., 2020). For each pair of domains, we randomly choose 20 samples per class (thus $n_s = n_t = 200$) and perform adaptation from CaffeNet to GoogleNet features, then calculate the accuracy of the generated predictions on the target domain using OT label propagation (Redko et al., 2019). This technique uses the OT plan to estimate the amount of mass transported from each class (since the sources are labeled) to a given target sample. The predicted class corresponds to the one which contains the most mass. We repeat this process 10 times and calculate the average and standard deviation of the performance. In both source and target domains, we assign uniform sample and feature distributions.

In the semi-supervised HDA task, we incorporate the prior knowledge on the target labels by

Domains	GW	UGW	COOT	UCOOT
C → C	16.25 (± 7.54)	10.85 (± 2.13)	36.40 (± 12.94)	**44.05 (± 19.33)**
C → A	12.95 (± 7.74)	11.60 (± 4.86)	28.30 (± 11.78)	**31.90 (± 7.43)**
C → W	18.95 (± 9.43)	14.15 (± 3.98)	19.55 (± 14.51)	**28.55 (± 6.60)**
A → C	16.40 (± 8.99)	10.25 (± 5.66)	**41.80 (± 14.81)**	39.15 (± 17.98)
A → A	14.75 (± 15.20)	20.20 (± 6.45)	**57.90 (± 16.84)**	42.45 (± 15.47)
A → W	14.55 (± 8.83)	20.65 (± 4.13)	42.10 (± 7.80)	**48.55 (± 13.06)**
W → C	20.65 (± 11.90)	14.20 (± 5.13)	8.60 (± 6.56)	**69.80 (± 14.91)**
W → A	17.00 (± 9.75)	7.10 (± 2.45)	16.65 (± 10.01)	**30.55 (± 10.09)**
W → W	19.30 (± 11.87)	24.40 (± 3.28)	**75.30 (± 3.26)**	51.50 (± 20.51)
Average	16.76 (± 10.14)	14.82 (± 4.23)	36.29 (± 10.95)	**42.94 (± 13.93)**

Table 4.3 – Unsupervised HDA from CaffeNet to GoogleNet.

adding an additional cost matrix to the training of sample coupling, so that a source sample will be penalized if it transfers mass to the target samples in the different classes. More precisely, we introduce the masked target label $\tilde{y}^{(t)} \in \mathbb{R}^{n_t}$ defined by randomly keeping $\tilde{n}_t \in \{1, 3, 5\}$ samples in each class in the target label $y^{(t)}$ and masking all other labels in $y^{(t)}$ by -1. Then the additional cost $M \in \mathbb{R}^{n_s \times n_t}$ between $y^{(s)}$ and $\tilde{y}^{(t)}$ is defined by

$$M_{ij} = \begin{cases} 0, \text{ if } y_i^{(s)} = \tilde{y}_j^{(t)}, \text{ or } \tilde{y}_j^{(t)} = -1 \\ v, \text{ otherwise.} \end{cases} \tag{4.8}$$

Here, $v > 0$ is a fixed value and we choose $v = 100$ in this experiment.

Once the sample coupling P is learned, the label propagation works as follows: suppose the labels contain K different classes, we apply the one-hot encoding to the source label $y^{(s)}$ to obtain $D^{(s)} \in \mathbb{R}^{K \times n_s}$ where $D^{(s)}_{ki} = 1_{\{y_i^{(s)}=k\}}$. The label proportions on the target data are estimated by: $L = D^{(s)}P \in \mathbb{R}^{K \times n_t}$. Then the prediction can be generated by choosing the label with the highest proportion, i.e. $\widehat{y}_j^{(t)} = \arg\max_k L_{kj}$. Note that, while the prediction is performed on the whole target samples, only those whose labels are masked as -1 during the training, are used in the calculation of accuracy. For the k-NN only, we train a classifier on the labelled target samples, then perform prediction on the unlabelled ones.

HDA results. The means and standard deviations of the accuracy on target data are reported in Table 4.3 for all the methods and all pairs of datasets. We observe that, thanks to its robustness, UCOOT outperforms COOT on 7 out of 9 dataset pairs, with higher average accuracy but also slightly larger variance. This is because of the difficulty of the unsupervised HDA problem and the instability present in all methods. In particular, GW-based approaches perform very poorly. This may be due to the fact that the pre-trained models contain meaningful but a very high-dimensional vectorial representation of the image. Thus, using the Euclidean distance

matrices as inputs not only causes information loss but also is less relevant (see for example, (Aggarwal, Hinneburg, and Keim, 2001), or Theorem 3.1.1 and Remark 3.1.2 in (Vershynin, 2018)).

However, under the presence of labelled target samples, Table 4.4 shows that the advantage of UCOOT diminishes significantly, as the level of certainty increases. In this case, both COOT and UCOOT are also much more stable but UCOOT is somewhat more volatile than COOT.

Robustness to target shift. We also illustrate the robustness of UCOOT to a change in class proportions, also known as target shift.

More precisely, we simulate a change in proportion only in the source domain by selecting $20p$ samples per class for 4 amongst 10 classes with p decreasing from $p = 1$ to $p = 0.2$. In this configuration, the classes in the source domain are imbalanced and the unlabeled HDA problem becomes more difficult. We report the performance of all the methods as a function of the Total Variation (TV) between the class marginal distributions on one pair of datasets in Figure 4.4. We can see that UCOOT is quite robust to change in class proportions, while COOT experiences a sharp decrease in accuracy when the class distributions become more imbalanced.

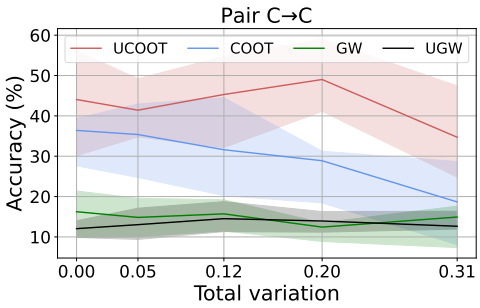

Figure 4.4 – Robustness to class proportion change for increasing TV on the class marginals.

4.5.3 Single-cell multi-omics alignment

Finally, we present a real-world application of UCOOT for the alignment of single-cell measurements. Recent advances in single-cell sequencing technologies allow biologists to measure a variety of cellular features at the single-cell resolution, such as expression levels of genes and epigenomic changes in the genome (Buenrostro et al., 2015; Chen, Lake, and Zhang, 2019a), or the abundance of surface proteins (Stoeckius et al., 2017). These multiple measurements produce single-cell multi-omics datasets. These datasets measuring different biological phenomena at the single-cell resolution allow scientists to study how the cellular processes are regulated, leading to finer cell variations during development and diseases. However, it is hard to obtain multiple types of measurements from the same individual cells due to experimental limitations. Therefore, many single-cell multi-omics datasets have disparate measurements from different sets of cells. As a result, computational methods are required to align the cells and the features of the different measurements to learn the relationships between them that help with data analysis and integration. Multiple tools (Cao, Hong, and Wan, 2021; Hao et al., 2021; Liu et al., 2019a), including GW (Cao, Hong, and Wan, 2021; Demetci et al., 2022a) and UGW (Demetci et al.,

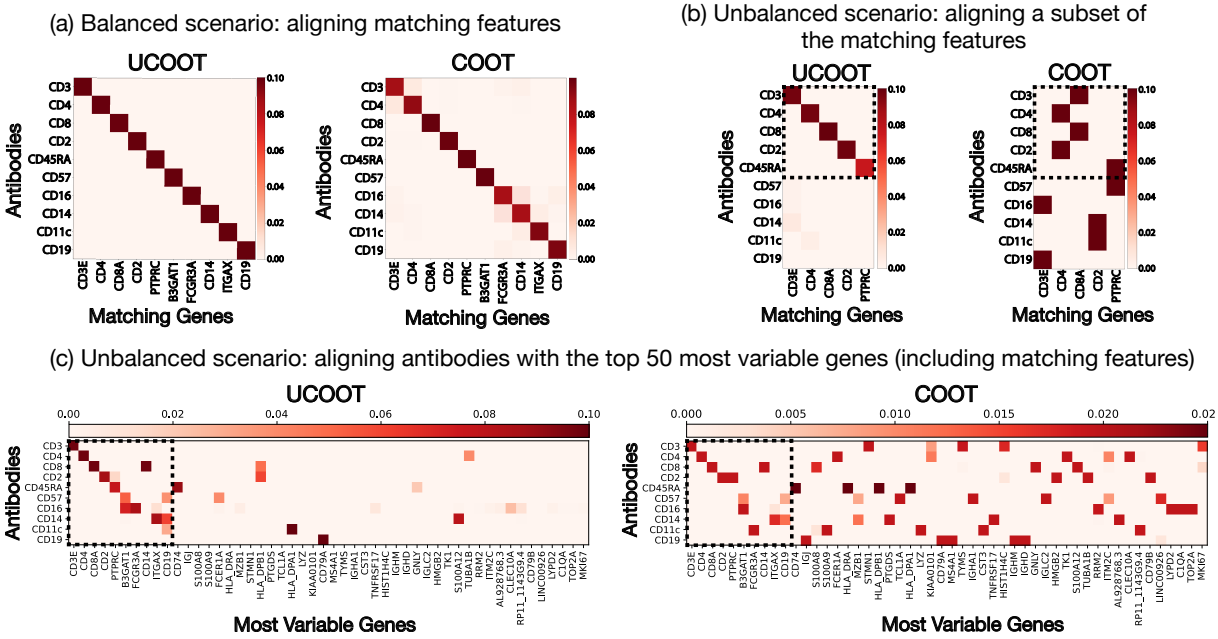

Figure 4.5 – Feature alignments on the single-cell multi-omics dataset of COOT and UCOOT between antibodies (surface proteins) and their matching genes (that encode them). **(a)** The features are sorted such that the correct alignment would yield a diagonal matrix. **(b)** Only five of the correct gene matches are kept (the last five genes from (a) are excluded). **(c)** Alignments between the ten antibodies and the top 50 most variable genes, including the matching genes. For **(b)** and **(c)**, the diagonal within the dashed square highlights the correct matches. Overall, UCOOT gives better feature alignments.

2021) based methods, have shown good performance for cell-to-cell alignments. However, aligning both samples and features is a more challenging and critical task that GW and UGW-based methods cannot address Here we provide an application of UCOOT to simultaneously align the samples and features in a single-cell multi-omics dataset.

Dataset For demonstration, we choose a dataset generated by the CITE-seq experiment (Stoeckius et al., 2017), which simultaneously measures gene expression and antibody (or surface protein) abundance in single cells. From this dataset, we use 1000 human peripheral blood cells, which have ten antibodies and 17,014 genes profiled. We selected this specific dataset as we know the ground-truth correspondences on both the samples (*i.e.*, cells) and the features (*i.e.*, genes and their encoded antibodies), thus allowing us to quantify and compare the alignment performance of UCOOT and COOT. As done previously (Cao, Hong, and Wan, 2021; Demetci et al., 2022a; Liu et al., 2019a), we quantify the cell alignment performance by calculating the fraction of cells closer than the true match (FOSCTTM) of each cell in the dataset and averaging it across all cells. This metric quantifies alignment error, so lower values are more desirable. The

feature alignments are measured by calculating the accuracy of correct matching.

Hyperparameter tuning Hyperparameters were tuned using grid search. For both COOT and UCOOT, we considered the following range for the entropic regularization coefficients $\varepsilon_f, \varepsilon_s \in \{1e-5, 5e-5, 1e-5, 5e-4, ..., 0.1, 0.5\}$. For the mass relaxation coefficients ρ_1, ρ_2 in UCOOT, the following range was considered $\rho_1, \rho_2 \in \{1e-3, 5e-3, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$. Each combination of hyperparameters were run on three randomly chosen subsets of the dataset that included 30% of the samples and the hyperparameter combinations that on average yielded the highest feature matches and lowest FOSCTTM were picked for the experiments on the full dataset.

Results in balanced scenarios First, we select and align the same number of samples and features across the two datasets. For this, we subset the gene expression domain with the ten genes that match to the ten antibodies they express. Original data contains the same number of cells across domains since both domains are simultaneously measured in the same single-cells. We observe that both UCOOT and COOT can correctly align features (Figure 4.5 (a)) and the cells across the two measurements. However, UCOOT gives better performance, as demonstrated by a lower FOCSTTM score (0.0062 vs 0.0127) for cells. Both COOT and UCOOT recover the diagonal for matching features (100% accuracy), but UCOOT recovers the exact matching, likely due to its robustness to noise, whereas COOT assigns weights to other features as well.

Results in unbalanced scenarios Next, we perform alignment with an unequal number of features. This setting is more likely to occur for real-world single-cell datasets as different features are measured. In the first simple scenario, we align the ten antibodies with only a subset (five) of their matching genes. As visualized in Figure 4.5 (b), COOT struggles to find the correct feature alignments (60% accuracy), which would lie in the diagonal of the highlighted square (dashed lines). However, the relaxation of the mass conservation constraint in UCOOT allows it to shrink the mass of antibodies that lack matches in the gene expression domain, leading do higher accuracy (100% accuracy).

Next, we align the ten antibodies with the 50 most variable genes in the dataset, including their matching genes. This alignment task is the most realistic scenario, as single-cell multi-omics data consists of high-dimensional datasets with a different number of features for different measurements. Therefore, biologists focus their analyses on the reduced set of most variable features (e.g. genes). It is also the most computationally challenging case among all our experiments on this dataset. Hence, we provide sample-level supervision to both methods by giving a cost penalization matrix based on the correct sample alignments to the sample alignment computation. We see in Figure 4.5(c) that in comparison to COOT (50% accuracy), UCOOT recovers more of the correct feature alignments (70% accuracy), and yields fewer redundant alignments. Note that

UCOOT avoids incorrect mappings by locally shrinking the mass of the features or samples that lack correspondences. This avoids subsequent incorrect downstream analysis of the integration results. This property can also help users to discover rare cell types by observing the extent of mass relaxation per cell or prune uninformative features in the single-cell datasets.

Lastly, we also consider the case of unequal number of samples across the two measurements. This case is common in real world single-cell multi-omics datasets that are not simultaneously measured. Demetci et al. (2021) have shown that single-cell alignment methods that do not account for this mismatch yield poor alignment results. Therefore, we downsample the number of cells in one of the domains by 25% and perform alignment with the full set of cells in the other domain. We compute the FOSCTTM score for all cells that have a true match in the dataset and report the average values. UCOOT continues to yield a low FOSCTTM score (0.0081 compared to 0.0062 in the balanced scenario), while COOT shows a larger drop in performance (0.1342 compared to 0.0127 in the balanced scenario).

4.6 Discussion

In this work, we present an extension of COOT called unbalanced COOT, where the hard constraint on the marginal distributions is replaced by a soft control via the KL divergence. The resulting problem not only benefits from the flexibility of COOT but also enjoys the provable robustness property under the presence of outliers, which is not the case for COOT. The experimental results confirm our findings, yielding a very competitive performance in the unsupervised HDA task, as well as meaningful feature couplings for the single-cell multi-omics alignment. Also, while UCOOT introduces additional hyper-parameters, domain knowledge can help narrow down the range of feasible values, thus reducing the time and computational cost of the tuning process. Further investigation should be carried out to fully understand and assess the observed efficiency of UCOOT in real-world applications, and also explore the possibilities of UCOOT in more diverse applicative settings, including its use as a loss in deep learning architectures. Lastly, from a theoretical perspective, statistical properties such as sample complexity or stability analysis are needed to better understand the intricate relations between the two sample and feature couplings.

Domains	Baseline	GW	UGW	COOT	UCOOT
			$\tilde{n}_t = 1$		
C → C	28.32 (± 6.28)	35.37 (± 8.85)	29.21 (± 6.54)	**87.37 (± 2.90)**	85.00 (± 2.03)
C → A	25.32 (± 9.61)	31.47 (± 8.76)	26.84 (± 7.93)	85.42 (± 6.21)	**85.79 (± 6.19)**
C → W	30.05 (± 5.90)	42.53 (± 6.31)	40.47 (± 8.32)	64.68 (± 7.88)	**67.00 (± 8.80)**
A → C	39.63 (± 7.21)	36.11 (± 8.04)	29.47 (± 5.50)	80.74 (± 9.87)	**84.11 (± 8.34)**
A → A	42.21 (± 7.60)	37.84 (± 16.34)	39.11 (± 11.56)	93.42 (± 1.32)	**93.58 (± 1.12)**
A → W	36.21 (± 9.45)	43.58 (± 3.75)	52.21 (± 8.26)	**91.63 (± 2.57)**	90.37 (± 6.51)
W → C	30.16 (± 7.22)	40.68 (± 8.11)	32.63 (± 7.70)	78.84 (± 4.24)	**79.05 (± 3.81)**
W → A	31.89 (± 6.55)	42.37 (± 7.35)	26.26 (± 4.67)	**95.84 (± 2.51)**	89.37 (± 11.34)
W → W	24.16 (± 6.79)	43.89 (± 4.64)	44.00 (± 5.10)	96.58 (± 5.54)	**98.00 (± 2.04)**
Average	32.57 (± 7.72)	39.32 (± 8.02)	35.58 (± 7.29)	**86.06 (± 4.78)**	85.81 (± 5.58)
			$\tilde{n}_t = 3$		
C → C	65.82 (± 4.28)	39.41 (± 9.83)	45.41 (± 5.56)	87.18 (± 2.05)	**87.76 (± 2.10)**
C → A	68.06 (± 5.89)	46.24 (± 10.45)	54.94 (± 7.29)	**86.94 (± 3.18)**	85.53 (± 3.13)
C → W	69.94 (± 4.92)	44.12 (± 4.99)	52.71 (± 6.25)	**83.76 (± 2.22)**	83.41 (± 5.60)
A → C	82.88 (± 4.44)	51.29 (± 3.71)	48.71 (± 6.10)	90.12 (± 1.76)	**90.18 (± 3.18)**
A → A	81.88 (± 4.13)	84.41 (± 4.69)	74.00 (± 6.73)	93.59 (± 1.91)	**94.65 (± 1.69)**
A → W	84.76 (± 2.91)	57.76 (± 9.23)	59.35 (± 4.20)	**93.82 (± 1.75)**	93.59 (± 1.40)
W → C	83.06 (± 4.75)	51.94 (± 8.68)	58.24 (± 2.44)	**95.76 (± 2.17)**	92.71 (± 6.19)
W → A	82.12 (± 3.69)	66.41 (± 10.75)	69.53 (± 5.62)	97.12 (± 0.61)	**97.71 (± 0.61)**
W → W	80.41 (± 3.56)	66.82 (± 3.26)	69.06 (± 5.39)	99.24 (± 0.75)	**99.41 (± 0.53)**
Average	77.18 (± 3.91)	56.49 (± 7.29)	59.11 (± 5.51)	**91.95 (± 1.82)**	91.66 (± 2.71)
			$\tilde{n}_t = 5$		
C → C	71.60 (± 4.12)	50.67 (± 3.78)	53.07 (± 4.68)	86.07 (± 2.36)	**87.53 (± 2.09)**
C → A	73.80 (± 3.14)	62.80 (± 5.16)	64.87 (± 3.78)	86.40 (± 2.62)	**86.60 (± 2.72)**
C → W	72.53 (± 4.13)	57.27 (± 2.74)	57.07 (± 4.08)	**88.20 (± 2.07)**	85.13 (± 4.28)
A → C	86.20 (± 3.07)	52.07 (± 3.16)	50.67 (± 4.32)	91.47 (± 2.12)	**93.87 (± 1.81)**
A → A	88.73 (± 2.66)	71.93 (± 5.50)	74.80 (± 7.84)	93.53 (± 1.71)	**94.13 (± 1.54)**
A → W	88.80 (± 2.60)	67.47 (± 5.26)	66.07 (± 5.66)	**93.13 (± 2.19)**	**93.13 (± 1.79)**
W → C	91.33 (± 2.49)	59.33 (± 4.15)	54.73 (± 3.68)	**95.87 (± 1.63)**	95.47 (± 1.90)
W → A	90.93 (± 3.40)	68.13 (± 4.01)	70.53 (± 3.71)	97.53 (± 1.27)	**98.73 (± 0.63)**
W → W	91.47 (± 3.40)	68.67 (± 3.71)	72.93 (± 3.17)	**99.40 (± 0.63)**	**99.40 (± 0.47)**
Average	83.84 (± 3.43)	62.04 (± 4.16)	62.75 (± 4.55)	92.40 (± 1.84)	**92.67 (± 1.91)**

Table 4.4 – Semi-supervised HDA from CaffeNet to GoogleNet, for different values of $\tilde{n}_t$.

Chapter 5

Fused Unbalanced Gromov-Wasserstein

This chapter presents the results from (Thual et al., 2022) and addresses the applications of unbalanced extension of fused Gromov-Wasserstein in human brains alignments. Individual brains vary in both anatomy and functional organization, even within a given species. Inter-individual variability is a major impediment when trying to draw generalizable conclusions from neuroimaging data collected on groups of subjects. Current co-registration procedures rely on limited data, and thus lead to very coarse inter-subject alignments. In this work, we present a novel method for inter-subject alignment based on Optimal Transport, denoted as Fused Unbalanced Gromov-Wasserstein (FUGW). The method aligns cortical surfaces based on the similarity of their functional signatures in response to a variety of stimulation settings, while penalizing large deformations of individual topographic organization. We demonstrate that FUGW is well-suited for whole-brain landmark-free alignment. The unbalanced feature allows to deal with the fact that functional areas vary in size across subjects. Our results show that FUGW alignment significantly increases between-subject correlation of activity for independent functional data, and leads to more precise mapping at the group level.

5.1 Introduction

The availability of millimeter or sub-millimeter anatomical or functional brain images has opened new horizons to neuroscience, namely that of mapping cognition in the human brain and detecting markers of diseases. Yet this endeavour has stumbled on the roadblock of inter-individual variability: while the overall organization of the human brain is largely invariant, two different brains (even from monozygotic twins (Pizzagalli et al., 2020)) may differ at the scale of centimeters in shape, folding pattern, and functional responses. The problem is further complicated by the fact that functional images are noisy, due to imaging limitations and behavioral differences across individuals that cannot be easily overcome. The status quo of the field is thus to rely on anatomy-based inter-individual alignment that approximately matches the outline of the brain (Avants et al., 2008) as well as its large-scale cortical folding patterns (Dale, Fischl, and Sereno, 1999; Fischl, 2012). Existing algorithms thus coarsely match anatomical features with diffeomorphic transformations, by warping individual data to a simplified template brain. Such methods lose much of the original individual detail and blur the functional information that can be measured in brain regions (see Figure 5.1).

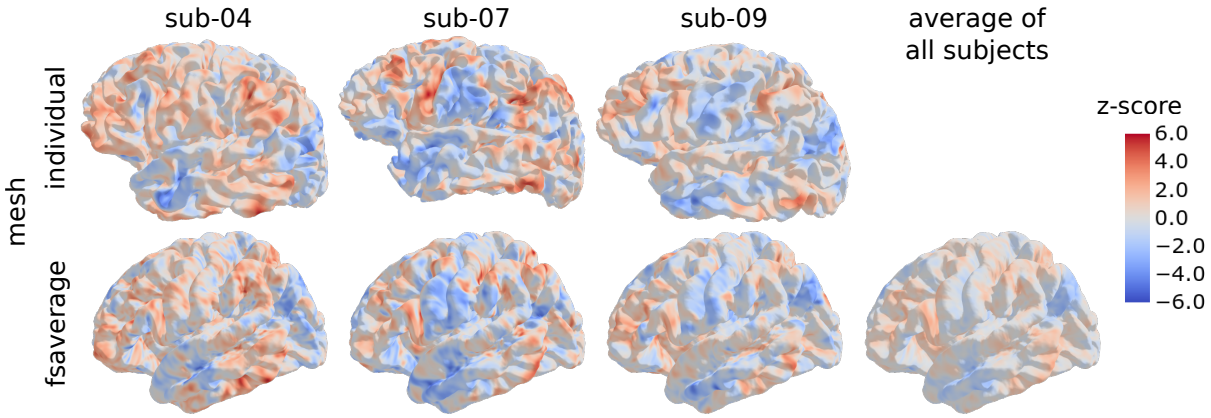

Figure 5.1 – **High variability in human anatomies and functional MRI responses across subjects** In this experiment contrasting areas of the brain which respond to mathematical tasks against other that don't, we observe great variability in locations and strength of brain activations across subjects (row 1). The classical approach consists in wrapping this data to a common surface template (row 2), where they can be averaged, often resulting in loss of individual details and detection power. These images were generated using Nilearn software (Abraham et al., 2014).

In order to improve upon the current situation, a number of challenges have to be addressed: *(i)* There exists no template brain with functional information, which by construction renders any cortical matching method blind to function. This is unfortunate, since functional information is arguably the most accessible marker to identify cortical regions and their boundaries (Glasser et al., 2016). *(ii)* When comparing two brains – coming from individuals or from a template – it is unclear what regularity should be imposed on the matching (Van Essen et al., 2012). While

it is traditional in medical imaging to impose diffeomorphicity (Avants et al., 2008), such a constrain does not match the frequent observation that brain regions vary across individuals in their fine-grained functional organization (Glasser et al., 2016; Schneider et al., 2019). *(iii)* Beyond the problem of aligning human brains, it is an even greater challenge to systematically compare functional brain organization in two different species, such as humans and macaques (Eichert et al., 2020; Mars et al., 2018; Neubert et al., 2014; Xu et al., 2020). Such inter-species comparisons introduce a more extreme form of variability in the correspondence model.

Related work Several attempts have been made to constrain the brain alignment process by using functional information. The first one consists in introducing functional maps into the diffeomorphic framework and search for a smooth transformation that matches functional information (Robinson et al., 2014; Sabuncu et al., 2010; Yeo et al., 2010), the most popular framework being arguably Multimodal Surface Matching (MSM) (Glasser et al., 2016; Robinson et al., 2014).

A second family of less constrained functional alignment approaches have been proposed, based on heuristics, by matching information in small, possibly overlapping, cortical patches (Bazeille et al., 2021; Haxby et al., 2011; Tavor et al., 2016). This popular framework has been called *hyperalignment* (Guntupalli et al., 2016; Haxby et al., 2011), or *shared response models* (Chen et al., 2015). Yet these approaches lack a principled framework and cannot be considered to solve the matching problem at scale. Neither do they allow to estimate a group-level template properly (Al-Wasity et al., 2020).

An alternative functional alignment framework has followed another path (Gramfort, Peyré, and Cuturi, 2015), considering functional signal as a three-dimensional distribution, and minimizing the transport cost. However, this framework imposes unnatural constraints of non-negativity of the signal and only works for one-dimensional contrasts, so that it cannot be used to learn multi-dimensional anatomo-functional structures. An important limitation of the latter two families of methods is that they operate on a fixed spatial context (mesh or voxel grid), and thus cannot be used on heterogeneous meshes such as between two individual human anatomies or, worse, between a monkey brain and a human brain.

Contributions Following (Bazeille et al., 2019b), we use the Wasserstein distance between source and target functional signals – consisting of contrast maps acquired with fMRI – to compute brain alignments. We contribute two notable extensions of this framework: *(i)* a Gromov-Wasserstein (GW) term to preserve global anatomical structure – this term introduces an anatomical penalization against improbably distant anatomical matches, yet without imposing diffeomorphic regularity – as well as *(ii)* an unbalanced correspondence that allows mappings from one brain to another to be incomplete, for instance because some functional areas are larger in some individuals than in others, or may simply be absent. We show that this approach successfully

addresses the challenging case of different cortical meshes, and that derived brain activity templates are sharper than those obtained with standard anatomical alignment approaches.

5.2 Methods

Optimal Transport yields a natural framework to address the alignment problem, as it seeks to derive a plan – a *coupling* – that can be seen as a soft assignment matrix between cortical areas of a source and target individual. As discussed previously, there is a need for a functional alignment method that respects the rich geometric structure of the anatomical features, hence the Wasserstein distance alone is not sufficient. By construction, the GW distance (Mémoli, 2007, 2011b) can help preserve the global geometry underlying the signal. The more recent fused GW distance (Vayer et al., 2019a) goes one step further by making it possible to integrate functional data simultaneously with anatomical information.

5.2.1 Fused Unbalanced Gromov-Wasserstein

We leverage (Séjourné, Vialard, and Peyré, 2021b; Vayer et al., 2019a) to present a new objective function which interpolates between a loss preserving the global geometry of the underlying mesh structure and a loss aligning source and target features, while simultaneously allowing not to transport some parts of the source and target distributions. We provide an open-source solver that minimizes this loss [1].

Formulation We denote $F^s \in \mathbb{R}^{n\times c}$ the matrix of features per vertex for the source subject. In the proposed application, they correspond to c functional activation maps, sampled on a mesh with n vertices representing the source subject's cortical surface. Let $D^s \in \mathbb{R}_+^{n\times n}$ be the matrix of pairwise geodesic distances [2] between vertices of the source mesh. Moreover, we assign the distribution $w^s \in \mathbb{R}_+^n$ on the source vertices. Comparably, we define $F^t \in \mathbb{R}^{p\times c}$, $D^t \in \mathbb{R}_+^{p\times p}$ and $w^t \in \mathbb{R}_+^p$ for the target subject, whose individual anatomy is represented by a mesh comprising p vertices. Eventually, w^s and w^t set the transportable mass per vertex, which, without prior knowledge, we choose to be uniform for the source and target vertices respectively: $w^s \triangleq (\frac{1}{n}, ..., \frac{1}{n})$, $w^t \triangleq (\frac{1}{p}, ..., \frac{1}{p})$.

Given a tuple of hyper-parameters $\theta \triangleq (\rho, \alpha, \varepsilon)$, where $\rho, \varepsilon \in \mathbb{R}_+$ and $\alpha \in [0, 1]$, for any

1. https://github.com/alexisthual/fugw provides a PyTorch (Paszke et al., 2019) solver with a scikit-learn (Pedregosa et al., 2011) compatible API
2. We compute geodesic distances using https://github.com/the-virtual-brain/tvb-gdist

coupling $P \in \mathbb{R}_{\geq 0}^{n \times p}$, we define the fused unbalanced Gromov-Wasserstein loss as

$$\begin{aligned} L_\theta(P) = (1-\alpha) \underbrace{\sum_{i,j} ||F_i^s - F_j^t||_2^2 P_{ij}}_{\text{Wasserstein loss } L_{\mathrm{W}}(P)} + \alpha \underbrace{\sum_{i,j,k,l} |D_{ik}^s - D_{jl}^t|^2 P_{ij} P_{kl}}_{\text{Gromov-Wasserstein loss } L_{\mathrm{GW}}(P)} \\ + \rho \underbrace{\left[\mathrm{KL}(P_{\#1} \otimes P_{\#1} | w^s \otimes w^s) + \mathrm{KL}(P_{\#2} \otimes P_{\#2} | w^t \otimes w^t) \right]}_{\text{Marginal constraints } L_{\mathrm{U}}(P)} + \varepsilon \underbrace{E(P)}_{\text{Entropy}} \end{aligned} \tag{5.1}$$

where $L_{\mathrm{W}}(P)$ matches vertices with similar features, $L_{\mathrm{GW}}(P)$ penalizes changes in geometry and $L_{\mathrm{U}}(P)$ fosters matching all parts of the source and target distributions. Throughout this paper, we refer to relaxing the hard marginal constraints of the underlying OT problem into soft ones as *unbalancing*. Here, $P_{\#1} \triangleq (\sum_j P_{i,j})_{0 \leq i < n}$ denotes the first marginal distribution of P, and $P_{\#2} \triangleq (\sum_i P_{i,j})_{0 \leq j < p}$ the second marginal distribution of P. The notation $\otimes$ represents the Kronecker product between two vectors or two matrices. $\mathrm{KL}(\cdot|\cdot)$ denotes the Kullback Leibler divergence, which is a typical choice to measure the discrepancy between two measures in the context of unbalanced optimal transport (Liero, Mielke, and Savaré, 2018). The last term $E(P) \triangleq \mathrm{KL}(P \otimes P | (w^s \otimes w^t) \otimes (w^s \otimes w^t))$ is mainly introduced for computational purposes, as it helps accelerate the approximation scheme of the optimisation problem. Typically, it is used in combination with a small value of ε, so that the impact of other terms is not diluted. On the other hand, the parameters α and ρ offer control over two other aspects of the problem: while α realizes a trade-off between the impact of different features and different geometries in the resulting alignment, ρ controls the amount of mass transported by penalizing configurations such that the marginal distributions of the transportation plan P are far from the prior weights w^s and w^t. This potentially helps adapting the size of areas where either the signal or the geometry differs too much between source and target.

Eventually, we define $\mathcal{X}^s \triangleq (F^s, D^s, w^s)$ and $\mathcal{X}^t \triangleq (F^t, D^t, w^t)$, and seek to derive an optimal coupling $P \in \mathbb{R}_{\geq 0}^{n \times p}$ minimizing

$$\mathrm{FUGW}(\mathcal{X}^s, \mathcal{X}^t) \triangleq \inf_{P \in \mathbb{R}_{\geq 0}^{n \times p}} L_\theta(P). \tag{5.2}$$

This can be seen as a natural combination of the fused GW (Vayer et al., 2019a) and the unbalanced GW (Séjourné, Vialard, and Peyré, 2021b) distances. To the best of our knowledge, it has never been considered in the literature.

Toy example illustrating the unbalancing property As exemplified in Figure 5.1, brain responses elicited by the same stimulus vary greatly between individuals. Figure 5.2 illustrates a similar yet simplified version of this problem, where the goal is to align two different signals supported on the same spherical meshes. In this example, for each of the $n = p = 3200$ vertices,

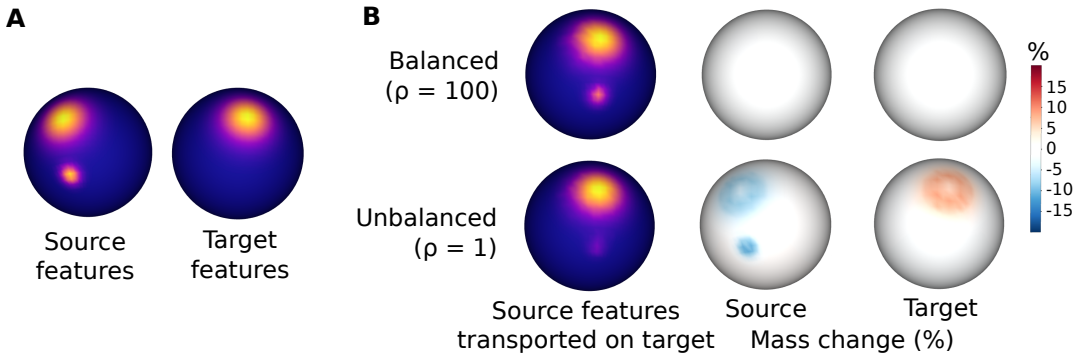

Figure 5.2 – **Unbalancing helps accounting for idiosyncrasies of the source and target signals** When trying to align the source and target signals (Panel A), the classical balanced setup (Panel B, top row) transports all parts of the source signal even if they have no counterpart in the target signal. In the unbalanced setup (Panel B, bottom row), less source-only signal is transported: in particular, less mass is transported from the source's small blob onto the target (Panel B, middle column).

the feature is simply a scalar. On the source mesh, the signal is constituted of two von Mises density functions that differ by their concentration (large and small), while on the target mesh, only the large one is present, but at a different location. We use the optimal coupling matrix P obtained from Equation (5.2) to transport the source signal on the target mesh. As shown in Figure 5.2.B, the parameter ρ allows to control the mass transferred from source to target. When $\rho = 100$, we approach the solution of the fused GW problem. Consequently, we observe the second mode on the target when transporting the source signal. When the mass control is weaker ($\rho = 1$), the smaller blob is partly removed because it has no counterpart in the target configuration, making the transport ill-posed.

5.2.2 Optimization

Estimating the unbalanced Gromov Wasserstein loss is numerically sensitive to initialization, due to the non-convexity of the problem. Therefore, FUGW is also *a priori* non-convex, and comparably difficult to estimate. Consequently, following (Séjourné, Vialard, and Peyré, 2021b), we instead compute a lower bound which is formulated as a bi-convex problem that relies on the joint estimation of two couplings.

$$\mathrm{FUGW}(\mathcal{X}^s, \mathcal{X}^t) = \inf_{\substack{P,Q \in \mathbb{R}_{\geq 0}^{n \times p} \\ P = Q}} L_\theta(P, Q) \geq \inf_{\substack{P,Q \in \mathbb{R}_{\geq 0}^{n \times p} \\ m(P) = m(Q)}} L_\theta(P, Q) \triangleq \mathrm{LB\text{-}FUGW}(\mathcal{X}^s, \mathcal{X}^t), \quad (5.3)$$

where $m(P) = \sum_{i,j} P_{i,j}$ denotes the mass of P and

$$L_\theta(P,Q) \triangleq (1-\alpha)\ L_{\mathrm{W}}(P,Q) + \alpha\ L_{\mathrm{GW}}(P,Q) + \rho\ L_{\mathrm{U}}(P,Q) + \varepsilon\ E(P,Q), \tag{5.4}$$

where

- $C \triangleq \left(||F_i^s - F_j^t||_2^2\right)_{i,j} \in \mathbb{R}_+^2$ (feature cost matrix)
- $G \triangleq \left(|D_{i,j}^s - D_{k,l}^t|\right)_{i,j,k,l} \in \mathbb{R}_+^4$ (geometry cost tensor)
- $L_{\mathrm{W}}(P,Q) \triangleq \langle C, \frac{P+Q}{2}\rangle = \frac{1}{2}(\sum_{i,j} C_{i,j}P_{i,j} + \sum_{i,j} C_{i,j}Q_{i,j})$ (Wasserstein)
- $L_{\mathrm{GW}}(P,Q) \triangleq \langle G, P\otimes Q\rangle = \sum_{i,j,k,l} G_{i,j,k,l}P_{i,j}Q_{k,l}$ (Gromov-Wasserstein)
- $L_{\mathrm{U}}(P,Q) \triangleq \ \mathrm{KL}\Big(P_{\#1}\otimes Q_{\#1}|w^s\otimes w^s\Big) + \ \mathrm{KL}\Big(P_{\#2}\otimes Q_{\#2}|w^t\otimes w^t\Big)$ (unbalancing)
- $E(P,Q) \triangleq \mathrm{KL}\Big(P\otimes Q|(w^s\otimes w^t)\otimes(w^s\otimes w^t)\Big)$ (entropy)

In particular, we have $L_\theta(P,P) = L_\theta(P)$, which is the objective function of FUGW introduced in Equation (5.1). It is difficult to study when equality holds between FUGW and its lower bound. Here, we attempt to understand the potential gap between them. First, let us introduce the following problem

$$\widetilde{\mathrm{FUGW}}(\mathcal{X}^s,\mathcal{X}^t) = \inf_{(P,Q)\in\mathcal{E}} L_\theta(P,Q), \tag{5.5}$$

where $\mathcal{E} = \{P,Q \in \mathbb{R}_{\geq 0}^{n\times p} : P_{\#1} = Q_{\#1}, P_{\#2} = Q_{\#2}\}$ is the set of pairs of transportation plans whose corresponding marginal distributions are equal. Clearly, we have

$$\text{LB-FUGW}(\mathcal{X}^s,\mathcal{X}^t) \leq \widetilde{\mathrm{FUGW}}(\mathcal{X}^s,\mathcal{X}^t) \leq \mathrm{FUGW}(\mathcal{X}^s,\mathcal{X}^t). \tag{5.6}$$

This inequality indicates that the difference between FUGW and LB-FUGW might be potentially large. However, this gap can be tightened under the conditions in Proposition 3.1.1.

Corollary 5.2.1. *If the distances D^s and D^t are of the forms: $D_{ij}^s = f_i + f_j + A_{ij}$ and $D_{kl}^t = g_k + g_l + B_{kl}$, where f, g are vectors in $\mathbb{R}^n, \mathbb{R}^p$, respectively, and the matrices A, B are both conditionally negative semi-definite, then we have $FUGW(\mathcal{X}^s,\mathcal{X}^t) = \widetilde{FUGW}(\mathcal{X}^s,\mathcal{X}^t)$.*

In our experiments, while the geodesic distances do not necessarily meet these conditions, we still observe that the two couplings of LB-FUGW are numerically equal. So it is enough to choose, for example, the first one, as alignment between source and target signals.

The lower bound of FUGW (5.3) involves solving a minimization problem with respect to two independent couplings. Using a Block-Coordinate Descent (BCD) scheme, we fix a coupling and minimize with respect to the other. This allows us to always be dealing with linear problems instead of a quadratic one. Eventually, each BCD iteration consists in alternatively solving two entropic unbalanced OT problems, whose solutions can be approximated using the Sinkhorn algorithm (Séjourné et al., 2019).

Algorithm 6 LB-FUGW barycenter for Problem (5.7)

Input: $(\mathcal{X}^s)_{s\in\mathcal{S}}, \rho, \alpha, \varepsilon$.
Output: Individual couplings $(P^{s,B})_{s\in\mathcal{S}}$, barycenter $\mathcal{X}^B$.
Initialize: $F^B = \mathbb{I}_k$; $D^B = 0_k$.
while $\mathcal{X}^B = (F^B, D^B, w^B)$ has not converged **do**
  Draw $\widetilde{S}$ subset of S.
  for $s \in \widetilde{S}$ **do**
    Align: $P^{s,B} \leftarrow$ LB-FUGW$(\mathcal{X}^s, \mathcal{X}^B, \rho, \alpha, \varepsilon)$. {Fixed $\mathcal{X}^B$}
  end for
  Update F^B, D^B: {Fixed $P^{s,B}$}

$$F^B = \frac{1}{|\widetilde{S}|}\sum_{s\in\widetilde{S}} \operatorname{diag}\left(\frac{1}{P^{s,B}_{\#2}}\right)(P^{s,B})^\top F^s \quad \text{and} \quad D_B = \frac{1}{|\widetilde{S}|}\sum_{s\in\widetilde{S}} \frac{(P^{s,B})^\top D^s P^{s,B}}{P^{s,B}_{\#2}(P^{s,B}_{\#2})^\top}.$$

end while

5.2.3 Barycenters

Barycenters represent common patterns across samples. Their role is instrumental in identifying a unique target for aligning a given group of individuals. As seen in Figure 5.1, the vertex-wise group average does not usually provide well-contrasted maps. Inspired by the success of the GW distance when estimating the barycenter of structured objects (Peyré, Cuturi, and Solomon, 2016; Vayer et al., 2019a), we use FUGW to find the barycenter $(F_B, D_B) \in \mathbb{R}^{k\times c} \times \mathbb{R}^{k\times k}$ of all subjects $s \in \mathcal{S}$, as well as the corresponding couplings $P^{s,B}$ from each subject to the barycenter. More precisely, we solve

$$\mathcal{X}^B = (F_B, D_B, w^B) \in \operatorname*{argmin}_{\mathcal{X}} \sum_{s\in\mathcal{S}} \text{FUGW}(\mathcal{X}^s, \mathcal{X}), \tag{5.7}$$

where we set the weights w_B to be the uniform distribution. By construction, the resulting barycenter benefits from the advantages of FUGW, i.e. equilibrium between geometry-preserving and feature-matching properties, while not forcing hard marginal constraints. The FUGW barycenter is estimated using a Block-Coordinate Descent (BCD) algorithm that consists in alternatively *(i)* minimizing the OT plans $P^{s,B}$ for each FUGW computation in (5.7) with fixed $\mathcal{X}^B$ and *(ii)* updating the barycenter $\mathcal{X}^B$ through a closed form with fixed $P^{s,B}$. See Algorithm 6 for more details. The first step simply uses the previously introduced solver. The second one takes advantage of the fact that the objective function introduced in (5.3) is differentiable in F^B and D^B, and the two couplings of LB-FUGW are numerically equal. This yields a closed form for F^B and D^B, as a function of $P^{s,B}$ and $\mathcal{X}^s$. We note that, during the barycenter estimation, the weight w^B is always fixed as uniform distribution.

5.3 Numerical experiments

We design three experiments to assess the performance of FUGW. In Experiments 1 and 2, we are interested in assessing if aligning pairs of individuals with FUGW increases correlation between subjects compared to a baseline correlation. We also compare the ensuing gains with those obtained when using the competing method MSM (Robinson et al., 2018, 2014) to align subjects. In Experiment 3, we derive a barycenter of individuals and assess its ability to capture fine-grained details compared to classical methods.

Dataset In all three experiments, we leverage data from the Individual Brain Charting dataset (Pinho et al., 2018). It is a longitudinal study on 12 human subjects, comprising 400 fMRI maps per subject collected on a wide variety of stimuli (motor, visual, auditory, theory of mind, language, mathematics, emotions, and more), movie-watching data, T1-weighted maps, as well as other features such as retinotopy which we don't use in this work. We leverage these 400 fMRI maps. The training, validation and test sets respectively comprise 326, 43 and 30 contrast maps acquired for each individual of the dataset. Tasks and MRI sessions differ between each of the sets.

Baseline alignment correlation For each pair of individuals (s, t) under study, and for each fMRI contrast c in the test set, we compute the Pearson correlation $\text{corr}(F^s_{\cdot,c}, F^t_{\cdot,c})$ after these maps have been projected onto a common surface anatomy (in this case, *fsaverage5* mesh). Throughout this work, such computations are made for each hemisphere separately.

Experiment 1 - Aligning pairs of humans with the same anatomy For each pair (s, t) under study, we derive an alignment $P^{s,t} \in \mathbb{R}^{n \times p}$ using FUGW on a set of training features. In this experiment, source and target data lie on the same anatomical mesh (*fsaverage5*), and $n = p = 10240$ for each hemisphere. Since each hemisphere's mesh is connected, we align one hemisphere at a time.

Computed couplings are used to align contrast maps of a the validation set from the source subject onto the target subject. Indeed, one can define $\phi_{s \to t} \colon X \in \mathbb{R}^{n \times q} \mapsto \left((P^{s,t})^T X\right) \oslash P^{s,t}_{\#2} \in \mathbb{R}^{p \times q}$ where $\oslash$ represents the element-wise division. $\phi_{s \to t}$ transports any matrix of features from the source mesh to the target mesh. We measure the Pearson correlation $\text{corr}(\phi_{s \to t}(F^s), F^t)$ between each aligned source and target maps.

We run a similar experiment for MSM and compute the correlation gain induced on a test set by FUGW and MSM respectively. For both models, we selected the hyper-parameters maximizing correlation gain on a validation set. In the case of FUGW, in addition to gains in correlation, hyper-parameter selection was influenced by three other metrics that help us assess the relevance of computed couplings:

Transported mass For each vertex i of the source subject, we compute $\sum_{0\leq j<p} P^{s,t}_{i,j}$.

Vertex displacement Taking advantage of the fact that the source and target anatomies are the same, we define $D = D^s = D^t$ and compute for each vertex i of the source subject the quantity $\sum_j P^{s,t}_{i,j} \cdot D_{i,j} / \sum_j P^{s,t}_{i,j}$, which measures the average geodesic distance on the cortical sheet between vertex i and the vertices of the target it has been matched with.

Vertex spread Large values of ε increase the entropy of derived couplings. To quantify this effect, and because we don't want the matching to be too blurry, we assess how much a vertex was *spread*. Considering $\tilde{P}_i = P^{s,t}_i / \sum_j P^{s,t}_{i,j} \in \mathbb{R}^p$ as a probability measure on target vertices, we estimate the anatomical variance of this measure by sampling q pairs (j_q, k_q) of $\tilde{P}_i$ and computing their average geodesic distance $\frac{1}{q} \sum_{j_q,k_q} D_{j_q,k_q}$.

Experiment 2 - Aligning pairs of humans with individual anatomies We perform a second alignment experiment, this time using individual meshes instead of an anatomical template. Importantly, in this case, there is no possibility to compare FUGW with baseline methods, since those cannot handle this case. However, individual meshes are significantly larger than the common anatomical template used in Experiment 1 ($n \approx m \approx$ 160k vs. 10k previously), resulting in couplings too large to fit on GPUs – for reference, a coupling of size 10k $\times$ 10k already weights 400Mo on disk. We thus reduce the size of the source and target data by clustering them into 10k small connected clusters using Ward's algorithm (Thirion et al., 2014).

Experiment 3 - Comparing FUGW barycenters with usual group analysis Since it is very difficult to estimate the barycentric mesh, we force it to be equal to the *fsaverage5* template. Empirically, this we force the distance matrix D^B to be equal to that of *fsaverage5*, and only estimate the functional barycenter F^B. We initialize it with the mean of $(F^s)_{s\in S}$ and derive F^B and $(P^{s,B})_{s\in S}$ from Problem (5.7). Then, for a given stimulus c, we compute its projection onto the barycenter for each subject. We use these projections to compute two maps of interest: *(i)* $M_{B,c}$ the mean of projected contrast maps across subjects and *(ii)* $T_{B,c}$ the t-statistic (for each vertex) of projected maps. We compare these two maps with their unaligned counterparts $M_{0,c}$ and $T_{0,c}$ respectively. The first map helps us to qualitatively evaluate the precision of FUGW

$$M_{B,c} \triangleq \frac{1}{|S|} \sum_{s\in S} \phi_{s\to t}(F^s_{\cdot,c}) \qquad T_{B,c} \triangleq \text{t-statistic}\Big((\phi_{s\to t}(F^s_{\cdot,c}))_{s\in S}\Big)$$

$$M_{0,c} \triangleq \frac{1}{|S|} \sum_{s\in S} F^s_{\cdot,c} \qquad T_{0,c} \triangleq \text{t-statistic}\Big((F^s_{\cdot,c})_{s\in S}\Big)$$

alignments and barycenter. The second one is classically used to infer the existence of areas of the brain that respond to specific stimuli. We assess whether FUGW helps find the same clusters of vertices. Eventually, we quantify the number of vertices significantly activated or deactivated with and without alignment respectively.

5.4 Results

5.4.1 Experiment 1 - Template anatomy

Aligning subjects on a fixed mesh We set $\alpha = 0.5$, $\rho = 1$ and $\varepsilon = 10^{-3}$. Pearson correlation between source and target contrast maps is systematically and significantly increased when aligned using FUGW, as illustrated in Figure 5.3 where correlation grows by almost 40% from 0.258 to 0.356.

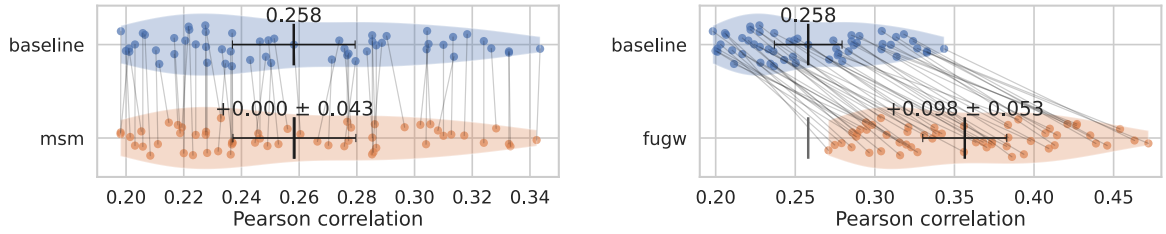

Figure 5.3 – **Comparison of gains in correlation after inter-subject alignment** For each pair of source and target subjects of the dataset, we compute the average Pearson correlation between 30 test contrasts, leading to the (baseline) correspondence score, and compare it with that of the same contrast maps aligned with either MSM (left) or FUGW (right). Correlation gains are much better for FUGW.

Hyper-parameters selection Hyper-parameters used to obtain these results were chosen after running a grid search on α, ε and ρ and evaluating it on the validation dataset. Computation took about 100 hours using 4 Tesla V100-DGXS-32GB GPUs. More precisely, it takes about 4 minutes to compute one coupling between a source and target 10k-vertex hemisphere on a single GPU, when the solver was set to run 10 BCD and 400 Sinkhorn iterations. In comparison, MSM takes about the same time on Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz CPUs. Results are reported in Figure 5.4 and provide multiple insights concerning FUGW.

Firstly, without anatomical constraint ($\alpha = 0$), source vertices can be matched with target vertices that are arbitrarily far on the cortical sheet. Even though this can significantly increase correlation, it also results in very high vertex displacement values (up to $100mm$). Such couplings are not anatomically plausible. Secondly, without functional information ($\alpha = 1$), couplings recover a nearly flawless matching between source and target meshes, so that, when $\varepsilon = 10^{-5}$ (ie when we force couplings to find single-vertex-to-single-vertex matches), vertex displacement and spread are close to 0 and correlation is unchanged. Fusing both constraints ($0 < \alpha < 1$) yields the largest gains in correlation while allowing to compute anatomically plausible reorganizations the cortical sheet between subjects.

The impact of ρ (controlling marginal penalizations) on correlation seems modest, with a slight tendency of increased correlation in unbalanced problems (low ρ).

Finally, it is worth noting that a relatively wide range of α and ρ yield comparable gains. The fact that FUGW performance is weakly sensitive to hyper-parameters makes it a good off-the-shelf tool for neuroscientists who wish to derive inter-individual alignments. However, ε is of dramatic importance in computed results and should be chosen carefully. Vertex spread is a useful metric to choose sensible values of ε; for human data one might consider that it should not exceed $20mm$.

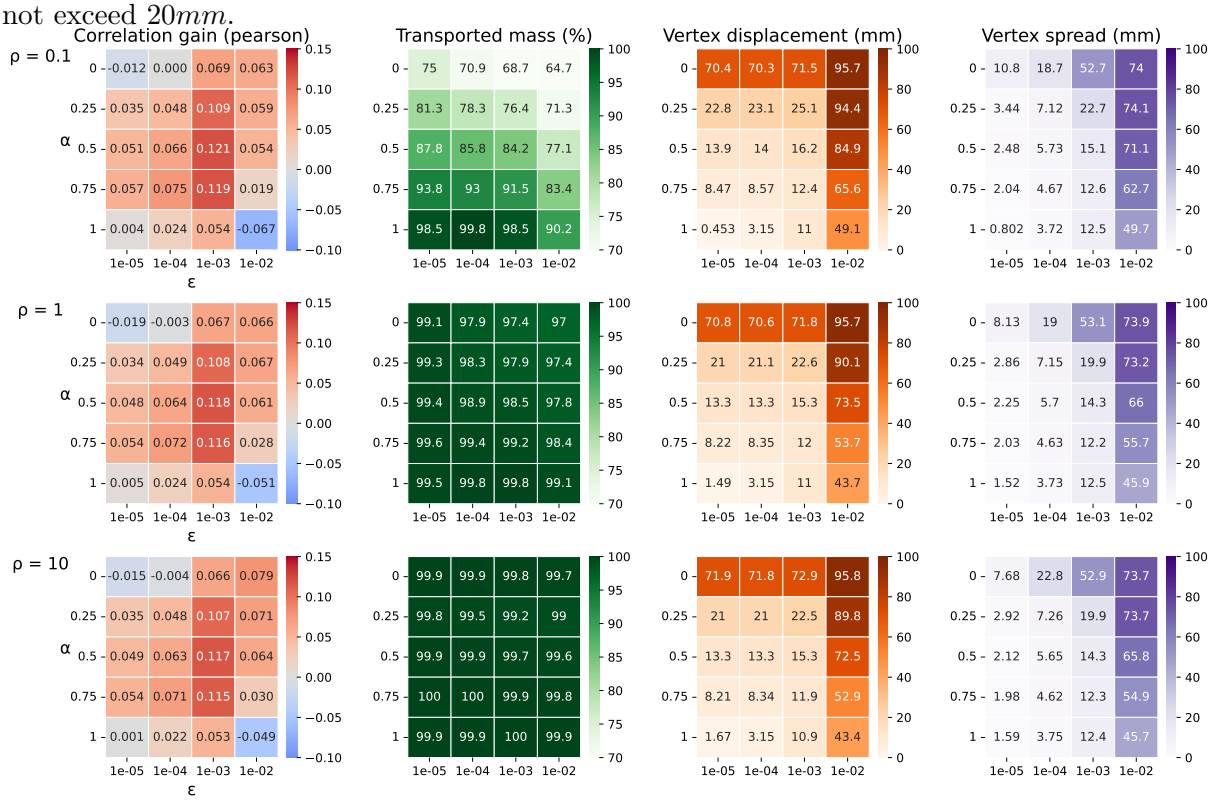

Figure 5.4 – **Exploring hyper-parameter space to find relevant couplings** Given a transport plan aligning a source and target subject, we evaluate how much this coupling (left) improves correlation between unseen contrast maps of the two subjects, (center left) actually transports data, (center right) moves vertices far from their original location on the cortical surface and (right) spreads vertices on the cortical sheet. We seek plans that maximize correlation gain, while keeping spread and displacement low enough.

Mass redistribution in unbalanced couplings Unbalanced couplings provide additional information about how functional areas might differ in size between pairs of individuals. This is illustrated in Figure 5.5, where we observe variation in size of the auditory area between a given pair of individuals. This feature is indeed captured by the difference of mass between subjects (although the displayed contrast was not part of the training set).

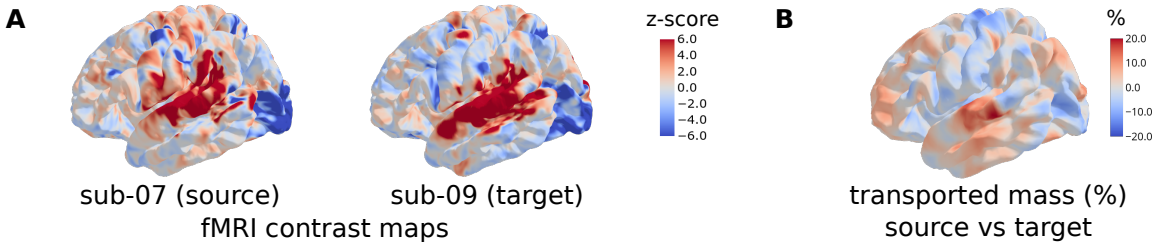

Figure 5.5 – **Transported mass indicates areas which have to be resized between subjects** (Panel A) We show a contrast map from the test set which displays areas showing stronger activation during auditory tasks versus equivalent visual tasks. It shows much more anterior activations on the target subject compared to the source subject. This is consistent with the observation that more mass is present in anterior auditory areas of the source subject than in the target subject (Panel B).

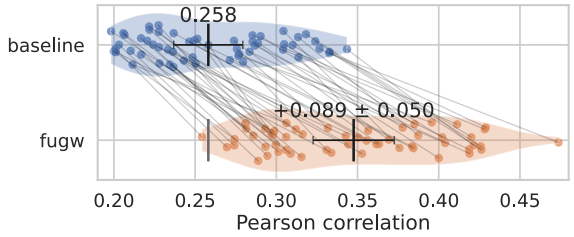

Figure 5.6 – **Correlation between pairs of subjects is significantly better after alignment on individual anatomies than after projecting subjects onto a common anatomical template**

5.4.2 Experiment 2 - Individual anatomies

As shown in Figure 5.6, we obtain correlation gains which are comparable to that of Experiment 1 (about 35% gain) while working on individual meshes. This tends to show that FUGW can compute meaningful alignments between pairs of individuals without the use of an anatomical template, which helps bridge most conceptual impediments listed in Section 5.1. Moreover, this opens the way for computation of simple statistics in cohorts of individuals in the absence of a template. Indeed, one can pick an individual of the cohort and use it as a reference subject on which to transport all other individuals. We give an example in Figure 5.7, showing that FUGW correctly preserved idiosyncrasies of each subject while transporting their functional signal in an anatomically sound way.

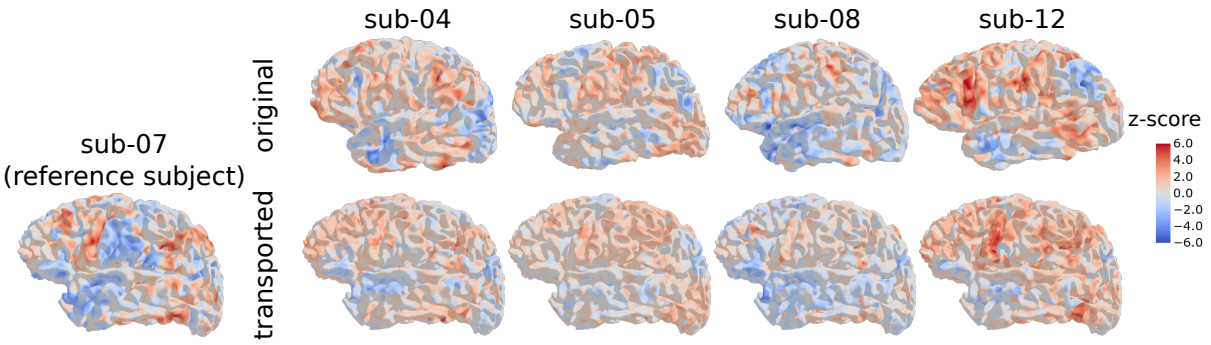

Figure 5.7 – **Transporting individual maps onto a reference subject** FUGW can help bridge the absence of template anatomies and derive pairs of alignments such that all individuals of the cohort are comparable. We display a map taken from the test set contrasting areas activated during mathematical reasoning against areas activated for other stimuli of the protocol.

5.4.3 Experiment 3 - Barycenter

In the absence of a proper metric to quantify the correctness of a barycenter, we first qualitatively compare the functional templates obtained with and without alignment. In Figure 5.8.A, we do so using brain maps taken from the test set. We can see that the barycenter obtained with FUGW yields sharper contrasts and more fine-grained details than the barycenter obtained by per-vertex averaging. We also display in Figure 5.8.B the result of a one-sample test for the same contrast, which can readily be used for inference. The one-sample test map obtained after alignment to the FUGW template exhibits the same supra-threshold clusters as the original approach, but also some additional spots which were likely lost due to inter-subject variability in the *fsaverage5* space. This approach is thus very useful to increase power in group inference. We quantify this result by counting the number of supra-threshold vertices with and without alignment for each contrast map of the test set. Our alignment method significantly finds more such vertices of interest, as shown in Figure 5.8.C.

5.5 Discussion

FUGW can derive meaningful couplings between pairs of subjects without the need of a pre-existing anatomical template. It is well-suited to computing barycenters of individuals, even for small cohorts.

In addition, we have shown clear evidence that FUGW yields gains that cannot be achieved by traditional diffeomorphic registration methods. These methods impose very strong constraints to the displacement field, that may prevent reaching optimal configurations. More deeply, this finding suggests that brain comparison ultimately requires lifting hard regularity constraints on the alignment models, and that two human brains differ by more than a simple continuous surface

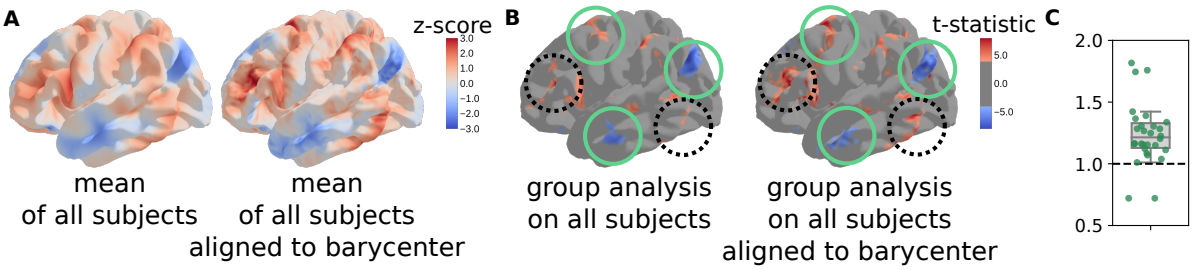

Figure 5.8 – **FUGW barycenter yields much finer-grained maps than group averages** We study the same statistical map as in Figure 5.1, which contrasts areas of the brain involved in mathematical reasoning. **A**. These complex maps projected onto the barycenter and averaged show more specific activation patterns than simple group averages, especially in cortical areas exhibiting more variability, such as the prefrontal cortex. **B**. Deriving a t-test on aligned maps captures the same clusters as the classical approach (plain green circles), but also new clusters in areas where inter-subject variability is high (dotted black circles). Peak t-statistics are also higher with FUGW. **C**. Ratio of number of activated vertices ($|\text{t-statistic}| \geq 4$) with versus without alignment for each map of the test set. Our method finds significantly more of such vertices (p-value $= 3 \cdot 10^{-4}$).

deformation. However, current results have not shown a strong correlation gain of unbalanced OT compared to balanced OT, likely because the cohort under study is too small. Leveraging datasets such as HCP (Van Essen et al., 2013) with a larger number of subjects will help lower the standard error on correlation gain estimates. In this work, we decided to rely on a predefined anatomical template (*fsaverage5*) to derive functional barycenters. It would be interesting to investigate whether more representative anatomical templates can be learned during the process. This would in particular help to customize templates to different populations or species.

Additionally, using an entropic solver introduces a new hyper-parameter ε that has a strong effect, but is hard to interpret. Future work may replace the Sinkhorn algorithm (Séjourné et al., 2019) used here by the majorization-minimization one (Chapel, Alaya, and Gasso, 2020), which does not require entropic smoothing. This solution can yield sparse couplings while being orders of magnitude faster, which will prove useful when computing barycenters on large cohorts.

Finally, we plan to make use of FUGW to derive alignments between human and non-human primates without anatomical priors. Indeed, the understanding of given brain mechanisms will benefit from more detailed invasive measurements made on other species *only if* brains can be matched across species; moreover, this raises the question of features that make the human brain unique, by identifying patterns that have no counterpart in other species. By maximizing the functional alignment between areas, but also allowing for some regions to be massively shrunk or downright absent in one species relative to the other, the present tool could shed an objective light on the important issue of whether and how the language-related areas of the human cortical sheet map onto the architecture of non-human primate brains.

Chapter 6

Breaking isometric ties and introducing priors in Gromov-Wasserstein distance

This chapter presents the results from (Demetci et al., 2024) and address the problem of incorporating priors into Gromov-Wasserstein (GW) distance. GW distance has many applications in machine learning due to its ability to compare measures across metric spaces and its invariance to isometric transformations. However, in certain applications, this invariant property can be too flexible, thus undesirable. Moreover, the GW distance solely considers pairwise sample similarities in input datasets, disregarding the raw feature representations. We propose a new optimal transport formulation, called Augmented Gromov-Wasserstein (AGW), that allows for some control over the level of rigidity to transformations. It also incorporates feature alignments, enabling us to better leverage prior knowledge on the input data for improved performance. We present theoretical insights into the proposed method. We then demonstrate its usefulness for single-cell multi-omic alignment tasks and heterogeneous domain adaptation in machine learning.

6.1 Introduction

Optimal transport (OT) theory provides a fundamental tool for comparing and aligning probability measures omnipresent in machine learning (ML) tasks. Following the least effort principle, OT and its associated metrics offer many attractive properties that other divergences, such as the popular Kullback-Leibler or Jensen-Shannon divergences, lack. For instance, OT borrows key geometric properties of the underlying "ground" space on which the distributions are defined (Villani, 2003) and enjoys non-vanishing gradients when measures have disjoint support (Arjovsky, Chintala, and Bottou, 2017). OT theory has also been extended to a much more challenging case of probability measures supported on different metric-measure spaces. In this scenario, the Gromov-Wasserstein (GW) distance seeks an optimal matching between points in the supports of the considered distributions that will minimize the distortion of intra-domain distances upon such matching.

Since its proposal by Mémoli (2011b) and further extensions by Peyré, Cuturi, and Solomon (2016), GW has been successfully used in a wide range of applications, including domain adaptation (Yan et al., 2018), computational biology (Cang and Nie, 2020; Cao et al., 2022; Cao, Hong, and Wan, 2021; Demetci et al., 2020, 2022a; Nitzan et al., 2019; Zeira et al., 2022), generative modeling (Bunne et al., 2019), and reinforcement learning (Nakagawa et al., 2022).

Limitations of prior work Successful applications of GW distance are often attributed to its invariance to distance-preserving transformations (also called "isometries") of the input domains. Since GW considers only intra-domain distances, it is naturally invariant to any transformation that does not alter them. While this is a blessing in many applications, for example, comparing graphs with the unknown ordering of nodes, it may become a curse when one has to choose the "right" isometry among many that yield the same GW distance. How could one break such ties while keeping the attractive properties of the GW distance? This question remains to be addressed in the field.

Additionally, GW distances are often used in tasks where one may have some *a priori* knowledge about the mapping between the two considered spaces. For example, in single-cell applications, mapping a group of cells in similar tissues across species helps understand the evolutionarily conserved and diverging cell types and functions (Kriebel and Welch, 2022). When performed using OT, this cross-species cell mapping may benefit from the knowledge about an overlapping set of orthologous genes [1]. GW formulation does not offer any straightforward way to incorporate this knowledge, which may lead to suboptimal performance.

1. Genes in two different species that originated from a common ancestor and largely maintained their function and sequence during speciation

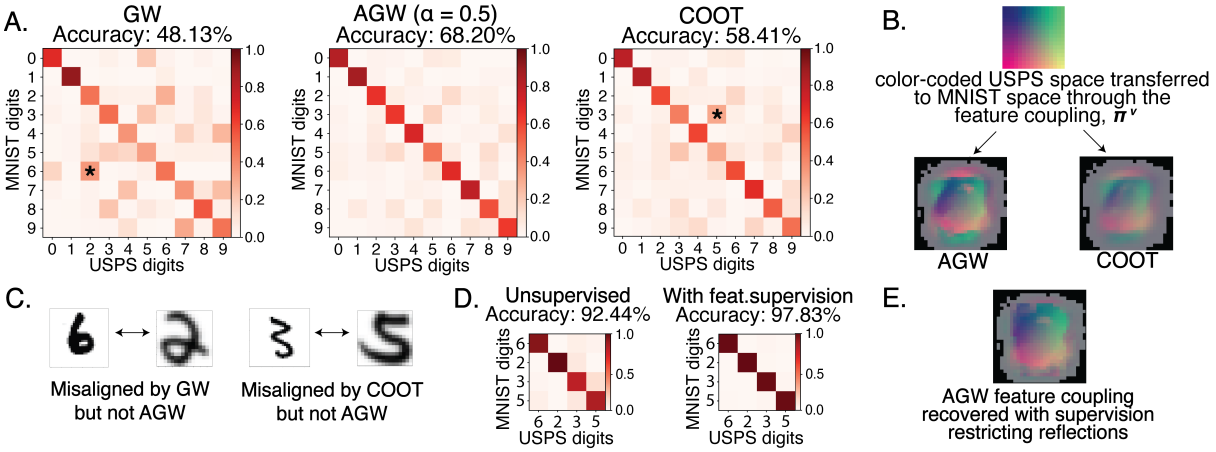

Figure 6.1 – Aligning digits from MNIST and USPS datasets. **(A)** Confusion matrices of GW, AGW with $\alpha = 0.5$ and COOT. (*) denote pair alignments ; **(B)** Feature coupling Q of AGW compared to COOT; **(C)** Illustration of a case from where GW's and COOT's invariants are detrimental for obtaining a meaningful comparison, while AGW remains informative. **(D)** Example showing improved digit alignment with feature-level supervision that restricts reflections **(E)** Feature coupling recovered by AGW ($\alpha = 0.5$) in the supervised setting of (D).

Our contributions In this chapter, we introduce a new OT formulation that addresses the drawbacks of the GW distance mentioned above. We summarize our contributions as follows:

1. We propose Augmented Gromov-Wasserstein (AGW), a new formulation that leverages both pairwise sample similarities in input datasets and their raw data representations;
2. We demonstrate that AGW allows for better control over the isometric transformations of the GW distance and helps break isometric ties;
3. We show that AGW can incorporate prior knowledge to guide how the two metric spaces should be compared, which improves object comparisons;
4. We provide a theoretical analysis of the properties of the proposed formulation and examples that concretely illustrate its unique features;
5. Our empirical results show that AGW outperforms previously proposed cross-domain OT methods in several downstream tasks and tends to converge in fewer iterations than GW. We first focus on real-world applications in computational biology, namely the single-cell data integration tasks. Then, we also illustrate its generalizability to the heterogeneous domain adaptation in ML.

6.2 Augmented Gromov-Wasserstein

Here, we start by outlining the motivation for our proposed formulation, highlighting the different properties of GW distance and COOT. Then, we detail our AGW method that interpolates

between the two, followed by a theoretical study of its properties.

6.2.1 Motivation

Invariants of GW GW distance remains unchanged under isometric transformations of the input data. This property has contributed much to the popularity of GW distance, as isometries naturally appear in many applications. However, not all isometries are equally desirable. For instance, a rotation of the handwritten digit 6 seen as a discrete measure can lead to its slight variation for small angles or to a digit 9 when the angle is close to 180 degrees. In both cases, however, the GW distance remains unchanged, making it insufficient to distinguish the two digits apart.

Invariants of COOT Unlike GW, COOT has fewer degrees of freedom in terms of invariance to global isometric transformations as it is limited to permutations of rows and columns of the two matrices, and not all isometric transformations can be achieved via such permutations.

Additionally, COOT is strictly positive for any two datasets of different sizes either in terms of features or samples, making it much more restrictive than GW. It thus provides a fine-grained control when comparing complex objects, yet it lacks the robustness of GW to frequently encountered transformations between the two datasets. Further, unlike GW, it is invariant to local isometries that can be achieved via permutations of a subset of features.

6.2.2 AGW formulation

Given the above discussion on the invariants of COOT and GW distance, interpolating between them will restrict each other's invariants. Additionally, interpolating with COOT is a natural way to introduce raw feature alignments in GW, which allows for leveraging priors on them. We call this interpolation **Augmented GW** (AGW). Recall that a weighted matrix is the triplet $\mathcal{X} = (X, \mu_1^X, \mu_2^X)$, where $X \in \mathbb{R}^{n_x \times d_x}, \mu_1^X \in \Delta_{n_x}$ and $\mu_2^X \in \Delta_{d_x}$. Given $\alpha \in [0, 1]$, we define the AGW between two weighted matrices $\mathcal{X}$ and $\mathcal{Y}$ as

$$\mathrm{AGW}_\alpha(\mathcal{X}, \mathcal{Y}) := \min_{\substack{P \in U(\mu_1^X, \mu_1^Y) \\ Q \in U(\mu_2^X, \mu_2^Y)}} \alpha \langle L(C^x, C^y) \otimes P, P \rangle + (1 - \alpha) \langle L(X, Y) \otimes P, Q \rangle, \tag{6.1}$$

where

- $\langle L(C^x, C^y) \otimes P, P \rangle = \sum_{i,j,k,l} (C^x_{ik} - C^y_{jl})^2 P_{ij} P_{kl}$ is the objective function of the GW distance. Here, the matrices $C^x \in \mathbb{R}^{n_x \times n_x}$ and $C^y \in \mathbb{R}^{n_y \times n_y}$ contain the intra-domain pairwise distances between the rows of X and Y, respectively.
- $\langle L(X, Y) \otimes P, Q \rangle = \sum_{i,j,k,l} (X_{ik} - Y_{jl})^2 P_{ij} Q_{kl}$ is the objective function of the COOT.

The AGW problem always admits a solution. Indeed, as the objective function is continuous and the sets of admissible couplings are compact, the existence of minimum and minimizer is guaranteed.

When $\alpha = 1$, we recover the GW distance, whereas $\alpha = 0$ corresponds to the COOT. Our interpolation offers several important benefits. First, COOT term ensures that AGW will take different values for any two isometries whenever $d_x \neq d_y$. Intuitively, AGW's value will then depend on how "far" a given isometry is from a permutation of rows and columns of the inputs. Thus, we restrict a broad class of (infinitely many) transformations that GW cannot distinguish and we tell them apart by assessing whether they can be approximately obtained by simply swapping 1D elements in input matrices.

Second, combining the objective functions of COOT and GW distance allows us to effectively influence the optimization of P by introducing priors on feature matchings through Q and vice versa. This can be achieved by penalizing the costs of matching certain features in the COOT term to influence the optimization of Q. This prior knowledge guides how the two metric spaces should be compared and improves empirical performance. These key properties explain our choice of calling it "augmented": we equip GW distance with an ability to provide finer-grained object comparisons by breaking isometric ties and/or guiding the matching using available prior knowledge.

Illustrations We illustrate AGW's properties on a task of aligning handwritten digits from MNIST (LeCun, Cortes, and Burges, 2010) (28×28 pixels) and USPS datasets (16×16 pixels) (Hull, 1994) in Figure 6.1, where AGW with $\alpha = 0.5$ outperforms both GW and COOT in alignment accuracy (Panel A). The black asterisks show the digit pairs that most benefit from AGW interpolation, which are $6 - 2$ for GW and $3 - 5$ for COOT. Panel C visualizes examples from these digit pairs that are misaligned by GW and COOT but not by AGW. [2]. Here, we observe that 6-2 misalignment by GW is likely because one is a close reflection of the other across the y-axis. Similarly, COOT mismatches 3 and 5 as one can obtain 3 from 5 by a local permutation of the upper half of the pixels. Panel B visualizes the feature couplings obtained by AGW (on the left) and COOT (on the right). The feature coupling by COOT confirms that COOT allows for a reflection across the y-axis on the upper half of the image but not on the lower half. With AGW, both of these misalignments improve, likely because (1) the correct feature alignments in the lower half of the images prevent 6 and 2 from being matched and (2) GW distance is non-zero for 5-3 matches since the transformation is not applied to the whole image. In Panels D and E, we also show that providing supervision on feature alignments to restrict local reflections further improves AGW's performance.

Similar improvement can be seen for aligning cells (samples) for two different single-cell

2. Here, we define "aligned pairs" as pairs of digits with the highest coupling probabilities.

Algorithm 7 BCD algorithm to solve AGW

1: Initialize P^* and Q^*.
2: **repeat**
3: Calculate $L_Q = L(X, Y) \otimes P^*$.
4: For fixed P, solve the OT problem: $Q^* \in \text{argmin}_{Q \in U(\mu_2^X, \mu_2^Y)} \langle L_Q, Q \rangle$.
5: Calculate $L_P = L(X, Y) \otimes Q^*$.
6: For fixed Q, solve the fused GW problem:

$$P^* \in \underset{P \in U(\mu_1^X, \mu_1^Y)}{\text{argmin}} \; \alpha \, \langle L(D_X, D_Y) \otimes P, P \rangle + (1 - \alpha) \langle L_P, P \rangle. \tag{6.2}$$

7: **until** convergence

measurements (features) (Chen, Lake, and Zhang, 2019b) in Figure S2. Panel A shows that AGW consistently maps the 4 cell types in the data better than GW (a popular method for this task (Cao et al., 2022; Cao, Hong, and Wan, 2021; Demetci et al., 2020, 2022a)) over 50 random subsampling of cells. The 2D projection of alignments in Panel B shows that GW sometimes completely swaps the cell type clusters when they have a similar number of cells, whereas AGW is more robust to this phenomenon.

Optimization For simplicity, suppose $n_x = n_y = n$ and $d_x = d_y = d$. With the squared loss in both GW and COOT terms, the computational trick by Peyré, Cuturi, and Solomon (2016) can be applied, which reduces the complexity of AGW from $O(n^4 + n^2 d^2)$ to $O(n^3 + dn^2 + nd^2)$. For optimization, we use the block coordinate descent (BCD) algorithm, where we alternatively fix one coupling and minimize AGW with respect to the other (Algorithm 7). Each iteration then consists of solving two OT problems. To further accelerate the optimization, entropic regularization (Cuturi, 2013) can be used on either P, Q, or both. In practice, we rely on the built-in functions of the Python Optimal Transport package (Flamary et al., 2021).

6.2.3 Theoretical analysis

Preliminary results Intuitively, we expect that AGW interpolates between GW and COOT and shares similar properties with Fused Gromov-Wasserstein (FGW) distance (Vayer et al., 2019a), namely the relaxed triangle inequality (since COOT and GW distances are both metrics). The following result summarizes these observations, whose proofs are presented in Appendix 8.5.1.

Proposition 6.2.1. *For $\alpha \in [0, 1]$,*

1. *Given two weighted matrices $\mathcal{X}$ and $\mathcal{Y}$, when $\alpha \to 0$ (or 1), one has $AGW_\alpha(\mathcal{X}, \mathcal{Y}) \to COOT(\mathcal{X}, \mathcal{Y})$ (or $GW(\mathcal{X}, \mathcal{Y})$).*

2. *AGW satisfies the relaxed triangle inequality: for any weighted matrices $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, one has $AGW_\alpha(\mathcal{X}, \mathcal{Y}) \leq 2(AGW_\alpha(\mathcal{X}, \mathcal{Z}) + AGW_\alpha(\mathcal{Z}, \mathcal{Y}))$.*

Invariants of AGW A more intriguing question is about the invariants that AGW exhibits. We denote $\mathcal{O}_d$ and $\mathcal{P}_d$ the sets of orthogonal and permutation matrices of size d, respectively. Given a matrix $X \in \mathbb{R}^{n\times d}$, we assume that

Assumption 6.2.1. *X is full-rank and has exactly $\min(n, d)$ distinct singular values.*

The full-rank assumption is not uncommon in the machine learning literature (Kawaguchi, 2016) and can be easily met in practice. Additionally, not only the Hermitian matrices with repeated eigenvalues are rare (see page 56 in (Tao, 2012)), but we can even show that

Corollary 6.2.1. *The set of Hermitian matrices with repeated eigenvalues has zero Lebesgue measure.*

Since the singular values of X are determined by the symmetric matrix XX^T, Corollary 6.2.1 assures that it is reasonable to exclude all symmetric matrices with repeated eigenvalues. With these, we present:

Theorem 6.2.1. *Given two weighted matrices $\mathcal{X}$ and $\mathcal{Y}$.*

1. *If $\mu_1^X = \mu_1^Y$ and Y is obtained by permuting columns of X via the permutation σ_c (so $\mu_2^Y = (\sigma_c)_\# \mu_2^X$), then $AGW_\alpha(\mathcal{X}, \mathcal{Y}) = 0$.*
2. *Suppose $X \in \mathbb{R}^{n\times d}$, where $n \geq d$, satisfies Assumption 6.2.1. For any $0 < \alpha < 1$, if $AGW_\alpha(\mathcal{X}, \mathcal{Y}) = 0$, then there exist a symmetric orthogonal matrix $O \in \mathcal{O}_d$ and a permutation matrix $P \in \mathcal{P}_d$ such that $Y = XOP$.*

Despite the simplicity of the interpolation structure, the invariants induced by AGW present novel and non-trivial challenges for theoretical analysis. While sharing basic invariants, such as feature swaps, AGW covers much fewer isometries than GW distance. Similar to COOT, AGW only has at most finitely many, whereas GW has infinitely many isometries. Under mild conditions, when AGW vanishes, only transformations with a particular structure (compositions of a permutation and a symmetric orthogonal transformation) are eligible. Given the superior empirical performance of AGW over GW and COOT, such isometries appear meaningful and relevant in real-world tasks.

Weak invariant to translation While enjoying the interpolation and metric properties, AGW does not inherit the invariant to the translation of the GW distance. However, we find that it satisfies a relaxed version of this invariant defined as follows.

Definition 6.2.1. *We call* $D = \inf_{\pi \in \mathcal{U}} F(\pi, X, Y)$*, where* X, Y *are input data and* $\mathcal{U}$ *is a set of feasible couplings and* F *is a real-valued functional, an OT-based divergence. Then* D *is weakly invariant to translation if for every* $a, b \in \mathbb{R}$*, we have* $\inf_{\pi \in U} F(\pi, X + a, Y + b) = \inf_{\pi \in \mathcal{U}} F(\pi, X, Y) + c$*, for some constant* C *depending on* a, b, X, Y *and* $\mathcal{U}$*.*

Here, we denote the translation of X as $X + a$, whose elements are of the form $X_{ij} + a$. Intuitively, an OT-based divergence is weakly invariant to translation if only the optimal transport plan is preserved under translation, but not necessarily the divergence itself. In practice, we would argue that the ability to preserve the optimal plan under translation is much more important than preserving the distance itself. In other words, the translation only shifts the minimum but has no impact on the optimization procedure, meaning that the minimizer remains unchanged. This is indeed the case of AGW.

Proposition 6.2.2. *For any* $\alpha \in [0, 1]$*, AGW is weakly invariant to translation.*

6.2.4 Related work

Most related to our work is the Fused Gromov-Wasserstein (FGW) distance (Vayer et al., 2019a) that compares structured objects. Its objective function is a convex combination of the GW term defined based on the pairwise intra-domain distances and the Wasserstein term defined over additional features that live in the same space for both input matrices. Despite the resemblance to FGW, AGW fundamentally differs from it in several ways. Firstly, AGW uses explicit control over the invariants of GW to provide more meaningful cross-domain matchings. No other OT-based metric in the literature (including FGW) leverages a similar idea. As such, Theorem 6.2.1 is the first result of its kind aiming at characterizing the invariances resulting from such interpolation.

Secondly, FGW is mostly used for structured objects endowed with additional information living in the same space, for example, two graphs where each node may be colored by a specific color ("additional" feature). On the other hand, AGW can be used on empirical measures defined for any set of objects across domains, including ones from different dimensional spaces, and requires no additional information.

Finally, the notion of feature space in FGW does not have the same meaning as in AGW. The feature space in FGW is associated with the sample space, whereas in AGW (and also in COOT), the two spaces are independent. Each element of the former is associated with a point in the sample space. By contrast, the features in AGW are precisely the coordinates of a point, in addition to its representation in the original and dissimilarity-induced spaces.

	Simulation 1 (300x1000, 300x2000)	Simulation 2 (300x1000, 300x2000)	Simulation 3 (300x1000, 300x2000)	Simulated RNA-seq (5000x50, 5000x500)	scGEM (177x28, 177x34)	SNARE-seq (1047x1000, 1047x3000)	CITE-seq (1000x25, 1000x24)
AGW	**0.0730**	**0.0041**	**0.0082**	**0.0**	0.183	**0.132**	**0.091**
GW	0.0866	0.0216	0.0084	7.1e-5	0.198	0.150	0.121
COOT	0.0752	**0.0041**	0.0088	**0.0**	0.206	0.153	0.132
UGW	0.0838	0.0522	0.0105	0.096	**0.161**	0.140	0.116
UCOOT	0.0850	0.0081	0.0122	0.115	0.181	0.188	0.127
bindSC	N/A	N/A	N/A	3.8e-4	N/A	0.242	0.144

Table 6.1 – **Single-cell alignment error**, as quantified by the average 'fraction of samples closer than true match' (FOSCTTM) metric (lower values are better). For each dataset, the size of the two domains they contain are expression in the format (number of samples x number of features) in the second row.

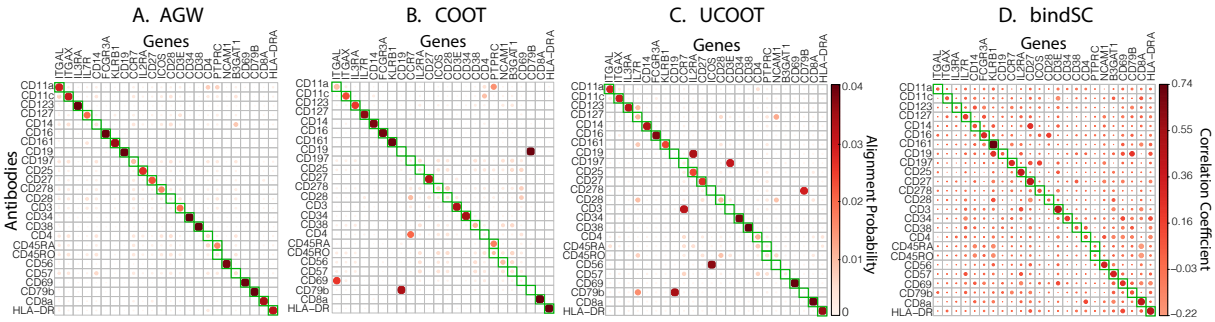

Figure 6.2 – Feature alignments for the CITE-seq dataset. Green boxes indicate where we expect matches (a notion of "ground-truth") based on domain knowledge.

6.3 Experimental evaluations

We apply AGW to the single-cell multi-omics alignment and heterogeneous domain adaptation tasks. Overall, we aim to empirically answer: **(1)** Does tightening invariances improve upon GW's performance in tasks where it has been previously used? **(2)** Does prior knowledge introduced in AGW help in obtaining better cross-domain matchings?

Baselines We pick other cross-domain OT methods as baselines, namely COOT, GW, and their unbalanced counterparts, UCOOT (Tran et al., 2023) and UGW (Séjourné, Vialard, and Peyré, 2021b). Note that we leave extending AGW to unbalanced scenarios for future work. We consider entropic regularization for all methods on both sample and (when applicable) feature couplings. We keep the hyperparameter values considered for all regularization coefficients consistent across all methods. We report the results of the best-performing hyperparameter combination after tuning on a validation set for each method in each experiment.

6.3.1 Integrating single-cell multi-omics datasets

Integrating data from different single-cell sequencing experiments is an important biological task for which OT has proven useful (Cao et al., 2022; Cao, Hong, and Wan, 2021; Demetci et al., 2020). Single-cell experiments measure various genomic features at the individual cell resolution. Jointly studying these can give scientists insight into the mechanisms regulating cells. However, experimentally combining multiple types of measurements for the same cell is challenging for most combinations. Scientists rely on the computational integration of multi-modal data taken on different but related cells (*e.g.*, by cell type or tissue) to study the relationships and interactions between different aspects of the genome.

We particularly focus on this task for two reasons. First, GW is used as a state-of-the-art method for this task (Cao et al., 2022; Cao, Hong, and Wan, 2021; Demetci et al., 2022a), so it is important to see if AGW improves upon it. Second, several single-cell benchmark datasets provide ground-truth matchings on the feature- and the sample-level alignments. This information allows us to assess the effect of guiding cross-domain matching with partial or full prior knowledge of these relationships.

Single-cell alignment We follow the first GW application in this domain (Demetci et al., 2020) and align samples (*i.e.*, cells) of simulated and real-world datasets from different measurement types. We have ground-truth information on cell-cell alignments for all datasets, which we only use for benchmarking. We demonstrate in Table 6.1 that AGW consistently yields higher quality cell alignments (with lower alignment error) compared to the state-of-the-art baselines, including GW, COOT, and their unbalanced counterparts.

We include bindSC as an additional baseline, which performs bi-order canonical correlation analysis to align single-cell datasets. Unlike other single-cell alignment methods, it internally computes a feature correlation matrix that users can extract. So, we include it as a baseline to compare its feature alignment performance against AGW in the next section. However, bindSC usage is limited to a few measurement types as it requires an input matrix that relates features across domains to bring the datasets into the same space at initialization. We do not have this information for most datasets, thus the "N/A" entries in Table 6.1.

Aligning genomic features AGW augments GW formulation with a feature coupling matrix. Therefore, we jointly align features and see whether AGW reveals relevant biological relationships. All current single-cell alignment methods only align samples (*i.e.*, cells), except for bindSC as discussed above.

Among the real-world datasets in Table 6.1, CITE-seq (Stoeckius et al., 2017) is the only one with ground-truth information on feature correspondences. This dataset has paired single-cell measurements on the abundance levels of 25 antibodies and activity (i.e., "expression") levels of

	A → A	A → C	A → W	C → A	C → C	C → W	W → A	W → C	W → W
AGW	**93.1±1.6**	**68.3±14.1**	**79.8±3.5**	55.4±7.1	76.4±5.6	**57.7±14.3**	60.1±9.1	**60.9±13.3**	**97.3±0.9**
GW	86.2±2.3	64.1±6.2	77.6±11.1	53.0±13.2	**81.9±10.5**	53.5±15.9	50.4±22.1	54.3±14.7	92.5±2.6
COOT	50.3±15.9	35.0±6.4	39.8±14.5	40.8±15.8	33.5±10.7	37.5±10.4	44.3±14.0	27.4±10.2	57.9±13.4
UGW	90.6±6.5	67.2±12.7	75.4±3.1	**56.3±14.6**	69.2±8.7	51.2±13.1	**66.7±9.9**	58.4±4.7	94.7±1.5
UCOOT	65.4±2.1	44.6±3.8	36.4±1.2	55.1±8.6	52.1±3.8	41.8±14.9	63.2±4.0	59.7±6.3	80.3±2.1

Table 6.2 – **Heterogeneous domain adaptation results (unsupervised)**. Best results are bolded, and second-bests are underlined. For AGW, the α values used are respectively $0.6, 0.9, 0.7, 0.9, 0.3, 0.8, 0.7, 0.2, 0.6$.

genes, including the genes that encode these 25 antibodies. So, we first present unsupervised feature alignment results on the CITE-seq dataset. We compare our feature alignments with bindSC, COOT, and UCOOT in Figure 6.2. The entries in the feature alignment matrices are arranged such that the "ground-truth" correspondences lie in the diagonal, marked by green squares. While AGW correctly assigns 19 out of 25 antibodies to their encoding genes with the highest alignment probability, this number is 16 for UCOOT, 15 for COOT and 13 for bindSC (which yields correlation coefficients instead of alignment probabilities). Additionally, the OT methods yield more sparse alignments thanks to the "least effort" requirement in their formulation.

Leveraging prior knowledge Finally, we show the advantage of providing priors by aligning a multi-species gene expression dataset containing measurements from the adult mouse prefrontal cortex (Bhattacherjee et al., 2019) and pallium of bearded lizard (Tosches et al., 2018). Since measurements come from two different species, the feature space (i.e., genes) differs, and there is no 1-1 correspondence between the samples (i.e., cells). However, there is a shared subset within the features, i.e., orthologous genes that descend from a common ancestor and maintain similar biological functions in both species. We also have domain knowledge on cells that belong to similar cell types across the two species. Thus, we expect AGW to recover these relationships.

Figure 6.3A visualizes the cell-type alignment probabilities yielded by AGW when full supervision is provided on the $10,816$ orthologous genes. The green boxes indicate alignment between similar types of cells. This matrix is obtained by averaging the sample alignment matrix (*i.e.*, cell-cell alignments) into cell-type groups. We observe that AGW yields biologically plausible alignments, as all the six cell types that have a natural match across the two species are correctly matched. We also show in Figure 6.3B that providing supervision on one alignment level (*e.g.*, features) improves the quality on the other alignment level (*e.g.*, samples).

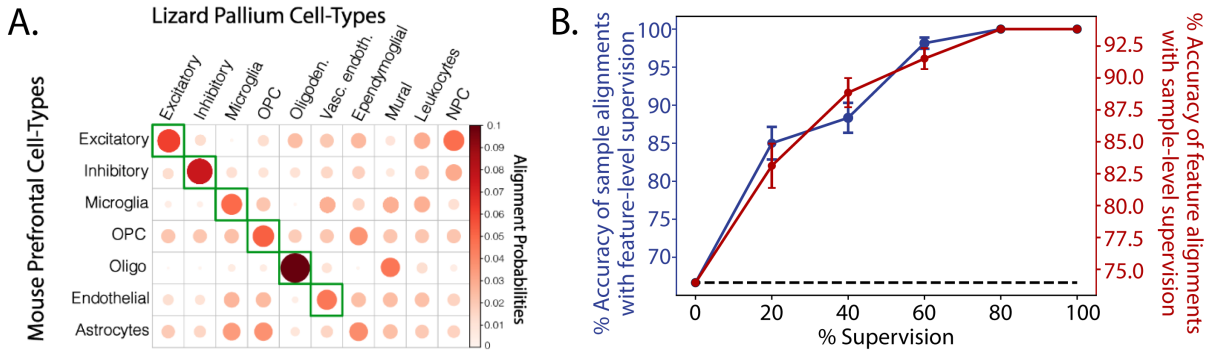

Figure 6.3 – **Aligning cross-species dataset.** A. AGW's cell-type alignments. B. Providing supervision on one level of alignment (*e.g.*, features) boosts alignments on the other. Standard errors computed over 10 random runs. Dashed line indicates the sample alignment performance of GW and bindSC (orthologous gene used in input).

6.3.2 Heterogeneous domain adaptation

Finally, we demonstrate the generalizability of our approach on a popular ML task, heterogeneous domain adaptation, where COOT and GW were previously successfully used. Domain adaptation (DA) refers to the problem in which a classifier learned on one domain (called *source*) can generalize to the other (called *target*). Here, we apply AGW to unsupervised and semi-supervised heterogeneous DA (HDA) tasks, where the source and target samples live in different spaces, and we have as few as zero labeled target samples.

Datasets and experimental setup We follow the experimental setup by Redko et al. (2020) and use source-target pairs from the Caltech-Office dataset (Saenko et al., 2010). We consider all pairs between three domains: Amazon (A), Caltech-256 (C), and Webcam (W), whose images are embeddings from the second last layer in the GoogleNet (Szegedy et al., 2015) (vectors in $\mathbb{R}^{4096}$) and CaffeNet (Jia et al., 2014) (vectors in $\mathbb{R}^{1024}$) neural network architectures.

In semi-supervised settings, we incorporate prior knowledge on a few target labels by adding an extra cost matrix to the training of sample coupling, so that a source sample will be penalized if it transfers mass to the target samples from different classes. Once the sample coupling P is learned, we obtain the final prediction using label propagation: $\widehat{y}_t = \text{argmax}_k \, L_{k\cdot}$, where $L = D_s P$ and D_s denotes one-hot encodings of the source labels y_s.

All hyperparameters are tuned on a validation set based on accuracy. We evaluate AGW against GW and COOT on source-target pairs from the Caltech-Office dataset (Saenko et al., 2010) by considering all pairs between the three domains: Amazon (A), Caltech-256 (C), and Webcam (W), similarly to (Redko et al., 2020). We randomly choose 20 samples per class and perform adaptation from CaffeNet to GoogleNet and repeat it 10 times. We report the average performance of each method along with the standard deviation. Differently than (Redko et al.,

2020), we (1) unit normalize the dataset prior to alignment as we empirically found it to boost all methods' average performance compared to using unnormalized datasets, (2) use cosine distances when defining intra-domain distance matrices for GW and AGW, as we found them to perform better than Euclidean distances, and (3) report results after hyperparameter tuning methods for each pair of datasets. Specifically, for each pair of (A)-(C), (A)-(W), etc, we sweep a hyperparameter grid over 5 runs of random sampling, choose the best-performing combination, and run 10 runs of random sampling to report results.

For all methods that allow for entropic regularization, we consider their version with no entropic regularization (either on the sample-level alignments, feature-level alignments, or both), along with various levels of regularization. For entropic regularization over sample alignments, we consider $\varepsilon_1 \in [5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 0.1]$. For entropic regularization over feature alignments in COOT and AGW, we consider $\varepsilon_2 \in [5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 0.1]$. As the interpolation coefficient of AGW, we consider $\alpha \in [0.1, 0.2, ..., 0.9]$.

Results Table 6.2 presents the performance of each method averaged across ten runs in the unsupervised setting, where AGW yields favorable results in 6 out of 9 cases. In two cases, UGW, and in one case, UCOOT, outperform AGW despite the lower performance of their balanced counterparts. In these cases, unbalanced formulations prove beneficial, and support extending AGW to unbalanced scenarios as future work.

6.4 Discussion

We present Augmented Gromov-Wasserstein (AGW), a new OT-based divergence for incomparable spaces. It interpolates between GW and CO-Optimal transport and allows to narrow down the choices of isometries induced by GW, while efficiently exploiting the prior knowledge on the input data. We study its basic properties and empirically show that such restrictions result in better performance for single-cell multi-omic alignment tasks and transfer learning. Future work will focus on refining the theoretical analysis of the AGW invariants to better understand their performance in practice. We will also extend AGW to the unbalanced and/or continuous setting, and other tasks where feature supervision by domain experts may be incorporated in OT framework.

Chapter 7

Conclusion

7.1 Contributions

The central motivation of this thesis is on the comparison between incomparable spaces. In particular, our work lies at the intersection of various extensions of OT, namely the marginal relaxation, the comparison between weighted objects, and the integration of prior knowledge. Our contributions can be organized in two main axes.

On the methodology side, we aimed to understand a few popular practices in the usage of divergences between incomparable spaces. First, in Section 3.2, we justify how entropic regularization can be used to approximate the GW distance and COOT. Then in Chapter 4, we show that the marginal constraints are tightly related to the impact of outliers. In particular, relaxing them with penalization via the Kullback-Leibler divergence allows for being very robust, whereas respecting them as in COOT and GW distance also paves the way for outliers to distort the minimum and mislead the alignments.

On the applications side, our proposed methods address the questions arised from real-world applications. In Chapter 5, we present the effectiveness of fused unbalanced GW in neuroscience, where we showcase how it can be used to align human cortical surfaces and learn better brain templates than the standard anatomical alignment approaches. In computational biology, we tackle different problematics in the integration of single-cell multi-omics data, including how to account for outliers (Chapter 4) and how to better exploit the input data (Chapter 6). In particular, we illustrate how the proposed variations of COOT are still able to correctly recover the relationships amongst genomic features, while providing meaningful cell correspondances between the multi-omics datasets and even outperforming many other OT-based competitors. These variations also show strong performance in the heterogeneous domain adaptation tasks, especially in the unsupervised setting.

7.2 Perspectives

From the methodology perspective, we discuss some remaining follow-ups left to explore from our work.

An alternative solver for unbalanced OT problem The core engine of many unbalanced OT-based methods, notably (fused) unbalanced GW and unbalanced COOT, relies on a solver for the unbalanced OT problem. As discussed in Section 2.1.2, there are few options. The *de facto* Sinkhorn-based algorithms (Séjourné et al., 2019; Séjourné, Vialard, and Peyré, 2021a) usually converges slowly for small regularization, while the Majorization-Minimization (MM) method (Chapel et al., 2021) suffers the same limitation for large marginal relaxation.

One possible alternative is the inexact Bregman Proximal Point (BPP) scheme, which is first applied to the balanced OT problem by Xie et al. (2020), known as *Inexact Proximal OT* (IPOT). Interestingly, this scheme can be easily extended to the unbalanced setting as follows. Denote F the objective function of the regularized unbalanced OT problem (2.22). For fixed learning rate $\eta > 0$, at iteration t, we solve

$$P^{(t+1)} \approx \operatorname*{argmin}_{P \in \mathbb{R}_{\geq 0}^{m \times n}} F(P) + \eta \, \mathrm{KL}(P|P^{(t)}), \tag{7.1}$$

or equivalently,

$$P^{(t+1)} \approx \operatorname*{argmin}_{P \in \mathbb{R}_{\geq 0}^{m \times n}} \left\langle C - \eta \log \frac{P^{(t)}}{\gamma}, P \right\rangle + \rho_1 \mathrm{KL}(P_{\#1}|\mu) + \rho_2 \mathrm{KL}(P_{\#2}|\nu) + (\varepsilon + \eta)\mathrm{KL}(P|\gamma). \tag{7.2}$$

The right-hand side of Equation (7.2) is nothing but an entropic UOT problem with modified cost and regularization. Thus, any solvers discussed in Section 2.1.2 can be used. Similar to IPOT, even as few as one Sinkhorn iteration may work in practice, meaning that the corresponding inexact solution may still empirically converge to the true minimizer of the UOT problem.

This inexact BPP scheme has two very appealing features. First, it is flexible and versatile since it can handle both balanced, semi-relaxed (which MM cannot), and unregularized (which Sinkhorn-based family cannot) settings.

Second, the presence of learning rate η increases the level of regularization in the inner entropic UOT subproblem, thus brings two important benefits. The first one is on the reduction of number of iterations: the larger the regularization, the faster the Sinkhorn algorithm converges. As a consequence, running only a few iterations is usually enough to obtain a decent approximation of the true solution. The second advantage is on the acceleration per BPP iteration. In practice, when the regularization is not too small, one can ignore the log-domain implementation and

employ the one with direct vector-matrix multiplication, without any concern about the numerical overflow issue. As a result, this allows to speed up the calculation of the iterates.

However, despite the simplicity, it appears to be difficult to study the convergence of this inexact scheme. In particular, while it is an immediate extension of the work of Xie et al. (2020) on the balanced OT, their proof techniques of the convergence can not be adapted to the unbalanced setting. This is because they rely on the property of the set of admissible couplings, which is not available in the UOT. Moreover, their assumptions and conditions are also not trivial to verify in practice, thus the convergence results are mostly of theoretical interest.

Perspectives on GW distance One potential application of MMOT-DC introduced in Section 3.3 is on the study of the sample complexity of GW distance. More precisely, given two measure networks $\mathcal{X} = (X, c_X, \mu_X)$ and $\mathcal{Y} = (Y, c_Y, \mu_Y)$, we want to quantify the convergence rate of $|\mathrm{GW}(\mathcal{X}, \mathcal{Y}) - \mathrm{GW}(\mathcal{X}_n, \mathcal{Y}_n)|$. Note that, Zhang et al., 2022a have also established the convergence rate for the case of 2-GW distance, in which the similarity is measured by squared Euclidean distance. The advantage of the approach via MMOT-DC, would be able to handle any conditionally negative semi-definite kernel.

The idea is as follows: by Proposition 3.1.1, COOT and GW distance are equivalent [1]. So, we can replace GW distance by COOT. Next, we extend MMOT-DC to the continuous setting. In the same spirit as Proposition 3.3.2, we expect that the interpolation property still holds, notably MMOT-DC converges to COOT as regularization tends to the infinity. Thanks to the Difference-of-Convex algorithm discussed in Section 3.3.5, we can linearize the MMOT-DC and obtain an entropic MMOT problem, which has two advantages. First, the technique used to study the sample complexity of entropic OT in (Genevay et al., 2019) can be extended to the multi-marginal setting. Second, this entropic problem can approximate COOT. Overall, the sample complexity of GW distance roughly boils down to that of an entropic MMOT problem.

Perspectives on Augmented Gromov-Wasserstein (AGW) While AGW has shown favorable performance over many other OT-based divergences, its isometries are still far from being fully understood. In particular, while we are able to explore the structure of **some** isometries via the singular-value decomposition, there are still some open questions.

1. What is the intuition behinds the isometries induced by AGW? To what extent are they "better" than those of GW distance?
2. Our analysis is only restricted to the case where $n \geq d$, meaning that the high-dimensional setting remains staying in the dark.
3. We can only establish the necessary conditions. But are they also sufficient?

1. Note that one needs to properly extend Proposition 3.1.1 to the continuous setting (of measure networks).

4. Given the discrete nature of AGW, it is natural the consider the continuous extension. In this case, what are the characteristics of its isometries?

In conclusion, we hope that this thesis contributes to the theory and practice of optimal transport for incomparable spaces, and that it will invite more applications of optimal transport in the interdisciplinary domains, including but not limited to computational biology and neuroscience.

Chapter 8

Annexes

8.1 Appendix of Chapter 2

8.1.1 Proofs related to Unbalanced Optimal Transport

Proof of Corollary 2.1.2. Following (Chapel et al., 2021), we can write Problem (2.37) as

$$\min_{t \in \mathbb{R}^{mn}_{\geq 0}} \langle c, t \rangle + \frac{||Mt - y||_2^2}{2}, \tag{8.1}$$

where $t = \text{vec}(P), c = \text{vec}(C)$ are the vectorizations of P, C, respectively. Here,

- $M = (\rho_1^{1/2} M_r^T, \rho_2^{1/2} M_c^T, \varepsilon^{1/2} I_{mn})^T \in \mathbb{R}^{(m+n+mn)\times(mn)}$.
- $y = (\rho_1^{1/2} \mu^T, \rho_2^{1/2} \nu^T, \varepsilon^{1/2} \text{vec}(\gamma)^T)^T \in \mathbb{R}^{m+n+mn}$.

- $M_r = \texttt{numpy.repeat(numpy.eye(n),m)} \in \mathbb{R}^{n\times mn}$.
- $M_c = [I_m, ..., I_m] = \texttt{numpy.tile(numpy.eye(m),n)} \in \mathbb{R}^{m\times mn}$.

See Appendix A in (Chapel et al., 2021) for more details of M_r, M_c. Remark that if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{m\times p}$, then

$$(A, B)\begin{pmatrix} A^T \\ B^T \end{pmatrix} = AA^T + BB^T. \tag{8.2}$$

Following Equation 23 in (Chapel et al., 2021), for $i \in [mn]$,

$$t_i^{(k+1)} = t_i^{(k)} \frac{\max\{0, (M^T y)_i - c_i\}}{(M^T M t^{(k)})_i}. \tag{8.3}$$

Now, we convert the vector t back to the transport plan. More precisely, $\text{mat}(M^T y) = (\rho_1\mu) \oplus (\rho_2\nu) + \varepsilon\gamma$, where matrization is the inverse operation of vectorization. Since, $M^T M = \rho_1 M_r^T M_r + \rho_2 M_c^T M_c + \varepsilon I_{mn}$, we obtain $\text{mat}(M^T M t^{(k)}) = (\rho_1 P_{\#1}^{(k)}) \oplus (\rho_2 P_{\#2}^{(k)}) + \varepsilon P^{(k)}$. The result then follows. ■

Proof of Corollary 2.1.1. We follow the same proof technique of Lemma 4 in (Pham et al., 2020). Given a Bregman divergence D_ψ, for any $t \in \mathbb{R}$ such that $tp \in \text{dom}(\psi)$, we have

$$\begin{aligned} D_\psi(tp|q) &= \psi(tp) - \psi(q) - \langle \nabla\psi(q), tp - q\rangle && (8.4)\\ &= \psi(tp) - \psi(q) - \Big[t\langle \nabla\psi(q), p - q\rangle + (t-1)\langle \nabla\psi(q), q\rangle\Big] && (8.5)\\ &= \psi(tp) - \psi(q) + t\Big[D_\psi(p|q) + \psi(q) - \psi(p)\Big] - (t-1)\langle \nabla\psi(q), q\rangle && (8.6)\\ &= tD_\psi(p|q) + [\psi(tp) - t\psi(p)] + (t-1)[\psi(q) - \langle \nabla\psi(q), q\rangle]. && (8.7) \end{aligned}$$

Denote g the objective function of Problem (2.25). We have,

$$g(tP) = tg(P) + \underbrace{\Big(\sum_{k=1}^{2} \rho_k[\varphi_k(tP_{\#k}) - t\varphi_k(P_{\#k})] + \varepsilon[\varphi(tP) - t\varphi(P)]\Big)}_{A(t)} \tag{8.8}$$

$$+ (t-1)\underbrace{\Big(\sum_{k=1}^{2} \rho_k[\varphi_k(\mu_k) - \langle \nabla\varphi_k(\mu_k), \mu_k\rangle] + \varepsilon[\varphi(\gamma) - \langle \nabla\varphi(\gamma), \gamma\rangle]\Big)}_{B} \tag{8.9}$$

$$= tg(P) + A(t) + (t-1)B. \tag{8.10}$$

For any minimizer $P^* \in E$, denote $E^* = \{t \in \mathbb{R} : tP^* \in E\}$. Clearly, E^* is not empty since

$1 \in E^*$. Since P^* is a (global) minimizer, we have $\frac{\partial g(tP^*)}{\partial t}\Big|_{t=1} = 0$, or equivalently,

$$g(P^*) = -B - \frac{\partial A(t)}{\partial t}\Big|_{t=1} \tag{8.11}$$

$$= \sum_{k=1}^{2} \rho_k \Big[\langle \nabla\varphi_k(\mu_k), \mu_k\rangle - \varphi_k(\mu_k)\Big] + \varepsilon\Big[\langle \nabla\varphi(\gamma), \gamma\rangle - \varphi(\gamma)\Big] \tag{8.12}$$

$$+ \sum_{k=1}^{2} \rho_k \left(\varphi_k(P^*_{\#k}) - \frac{\partial \varphi_k(tP^*_{\#k})}{\partial t}\Big|_{t=1}\right) + \varepsilon\left(\varphi(P^*) - \frac{\partial \varphi(tP^*)}{\partial t}\Big|_{t=1}\right). \tag{8.13}$$

The result then follows. ■

Corollary 8.1.1. *Equations* (2.37), (2.84) *and* (4.2) *are consequences of Corollary 2.1.1.*

Proof of Corollary 8.1.1. Equation (2.37) on the squared l_2-regularized UOT follows immediately from Corollary 2.1.1, where D_φ, D_{φ_1} and D_{φ_2} are the half squared Euclidean norm, and $E = \mathbb{R}^{m\times n}_{\geq 0}$. When D_φ, D_{φ_1} and D_{φ_2} are the KL divergence, Equation (2.84) on unbalanced Gromov-Wasserstein and Equation (4.2) on unbalanced Co-Optimal Transport can be justified in exactly the same manner. For this reason, we only show how to derive the relation in Equation (4.2). Recall that, from the proof of Proposition 8.3.1, Problem (8.101) can be rewritten as

$$\mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = \inf_{\pi\in E_{uco}} \int_S |\xi_1 - \xi_2|^p \mathrm{d}\pi + \sum_{k=1,2} \rho_k D_{\phi_k}(\pi_{\#k}|\mu_k), \tag{8.14}$$

where $\mu_k = \mu_k^s \otimes \mu_k^f$, for $k = 1, 2$, and

$$E_{uco} = \{\pi \in \mathcal{M}^+(S) | \pi = \pi^s \otimes \pi^f, \pi^s \in \mathcal{M}^+(X_1^s \times X_2^s), \pi^f \in \mathcal{M}^+(X_1^f \times X_2^f)\} \tag{8.15}$$

is the set of factorizable plans in $\mathcal{M}^+(S)$. Now, clearly, for any $t > 0$, if $\pi \in E_{uco}$, then $t\pi \in E_{uco}$. So, Corollary 2.1.1 can be applied and we deduce that, if $\pi_* = \pi_*^s \otimes \pi_*^f$ is the solution of Problem (8.14), then

$$\mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = \sum_{k=1}^{2} \rho_k m(\mu_k) - (\rho_1 + \rho_2) m(\pi_*) \tag{8.16}$$

$$= \sum_{k=1}^{2} \rho_k m(\mu_k^s) m(\mu_k^f) - (\rho_1 + \rho_2) m(\pi_*^s) m(\pi_*^f). \tag{8.17}$$

The result then follows. ■

8.1.2 Proofs related to Gromov-Wasserstein distance

Corollary 8.1.2. *The formulations* (2.42) *and* (2.46) *of the Gromov-Hausdorff distance are equivalent.*

Proof of Corollary 8.1.2. In the formulation (2.46), by choosing identity mappings (which are clearly isometric embeddings) $f = \mathrm{Id}_X$ and $g = \mathrm{Id}_Y$, and $Z = X \cup Y$ equipped with an admissible distance d, we have

$$\mathrm{GH}((X, d_X), (Y, d_Y)) \leq d_H^{(Z,d)}(X, Y) \tag{8.18}$$

As this is true for any $d \in \mathcal{D}(d_X, d_Y)$, we have $\mathrm{GH}((X, d_X), (Y, d_Y)) \leq \inf_d d_H^{(X\cup Y,d)}(X, Y)$. For the reverse direction, given any isometries $f : X \to X'$ and $g : Y \to Y'$ with $X', Y' \subset (Z, d_Z)$ (we call (X', Y') a *metric coupling* (Villani, 2009)), one can define a distance d on $X \cup Y$ (as shown in Proposition 27.1 in (Villani, 2009)). Then,

$$\begin{aligned} d_H^{(Z,d_Z)}(f(X), g(Y)) &= \max(\sup_{y\in Y} d_Z(g(y), f(X)), \sup_{x\in X} d_Z(f(x), g(Y))) \\ &= \max(\sup_{y\in Y} d(y, X), \sup_{x\in X} d(x, Y)) \\ &= d_H^{(X\cup Y,d)}(X, Y) \geq \inf_{d'} d_H^{(X\cup Y,d')}(X, Y) \end{aligned} \tag{8.19}$$

We deduce that $\mathrm{GH}((X, d_X), (Y, d_Y)) \geq \inf_d d_H^{(X\cup Y,d)}(X, Y)$, thus equality holds. ∎

Proof of Lemma 2.2.4. Denote $F(P) = \langle C \otimes P, P\rangle + \langle M, P\rangle + \varepsilon \mathrm{KL}(P|\gamma)$. Then $\nabla F(P) = C \otimes P + M + \varepsilon \log \frac{P}{\gamma}$. The PGD iterate reads: for $\tau > 0$,

$$P^{(t+1)} = \mathrm{proj}^{\mathrm{KL}}_{U(\mu_x,\mu_y)}\left(P^{(t)} \odot e^{-\tau\nabla F(P^{(t)})}\right), \tag{8.20}$$

where $\mathrm{proj}^{\mathrm{KL}}_{U(\mu_x,\mu_y)}(K) = \mathrm{argmin}_{P\in U(\mu_x,\mu_y)} -\varepsilon\langle \log K, P\rangle + \varepsilon H(P)$. Denote $\eta = \tau\varepsilon$. We have

$$\log K = \log\left(P^{(t)} \odot e^{-\tau\nabla F(P^{(t)})}\right) \tag{8.21}$$

$$= \log P^{(t)} - \tau(C \otimes P^{(t)} + M) - \tau\varepsilon \log \frac{P^{(t)}}{\gamma} \tag{8.22}$$

$$= -\tau(C \otimes P^{(t)} + M) + \log\left(\gamma^\eta \odot (P^{(t)})^{1-\eta}\right). \tag{8.23}$$

We deduce that

$$-\varepsilon\langle \log K, P\rangle + \varepsilon H(P) \tag{8.24}$$

$$= \eta\langle C \otimes P^{(t)} + M, P\rangle - \varepsilon\langle P, \log\left(\gamma^\eta \odot (P^{(t)})^{1-\eta}\right)\rangle + \varepsilon H(P) \tag{8.25}$$

$$= \eta\langle C \otimes P^{(t)} + M, P\rangle + \varepsilon \mathrm{KL}\left(P|\gamma^\eta \odot (P^{(t)})^{1-\eta}\right) - m\left(\gamma^\eta \odot (P^{(t)})^{1-\eta}\right), \tag{8.26}$$

where $m(Q) = \sum_{i,j} Q_{ij}$ is the mass of measure Q. The result then follows. ■

8.2 Appendix of Chapter 3

8.2.1 Proofs related to Discrete Co-Optimal Transport

Proof of Proposition 3.1.1. For any $P, Q \in U(\mu_X, \mu_Y)$, we have

$$\sum_{i,j,k,l} (C^x_{ik} - C^y_{jl})^2 P_{ij} Q_{kl} = \mu_X^T (C^x)^{\odot 2} m u_x + \mu_Y^T (C^y)^{\odot 2} \mu_y - 2 \sum_{i,j,k,l} C^x_{ik} C^y_{jl} P_{ij} Q_{kl} \tag{8.27}$$

$$= \mu_X^T (C^x)^{\odot 2} m u_x + \mu_Y^T (C^y)^{\odot 2} \mu_y - 2\mathrm{tr}(C^x Q C^y P^T). \tag{8.28}$$

Let (P, Q) and $\widehat{P}$ be the solutions of the COOT and GW problems, respectively. As $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) \leq \mathrm{GW}(\mathcal{X}, \mathcal{Y})$, we have $\mathrm{tr}(C^x Q C^y P^T) \geq \mathrm{tr}(C^x \widehat{P} C^y \widehat{P}^T) \geq \mathrm{tr}(C^x Q C^y Q^T)$, where the second inequality is due to the suboptimality of Q with respect to the GW problem. Given the form of C^x and C^y, we can further simplify this inequality as $\mathrm{tr}(AQBP^T) \geq \mathrm{tr}(AQBQ^T)$. Similarly, $\mathrm{tr}(AQBP^T) \geq \mathrm{tr}(APBP^T)$. So,

$$0 \leq 2\mathrm{tr}(AQBP^T) - \mathrm{tr}(AQBQ^T) - \mathrm{tr}(APBP^T) \tag{8.29}$$

$$= \mathrm{tr}(AQB(P-Q)^T) - \mathrm{tr}(A(P-Q)BP^T) \tag{8.30}$$

$$= \mathrm{vec}(Q^T)(B \otimes_K A)\mathrm{vec}(P-Q) - \mathrm{vec}(P)^T(B \otimes_K A)\mathrm{vec}(P-Q) \tag{8.31}$$

$$= -\mathrm{vec}(P-Q)^T(B \otimes_K A)\mathrm{vec}(P-Q), \tag{8.32}$$

where $\otimes_K$ denotes the Kronecker product. Now, we recall Theorem 1 in (Maron and Lipman, 2018).

Lemma 8.2.1. *If the matrices $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$ are CND, then $vec(X)^T(B \otimes_K A)vec(X) \geq 0$, for every $X \in lin(DS)$, where $lin(DS) = \{X \in \mathbb{R}^{m \times n} : X1_n = 0, X^T 1_m = 0\}$.*

As $P, Q \in U(\mu_X, \mu_Y)$, we have $P - Q \in \mathrm{lin}(DS)$. So, by Lemma 8.2.1, Inequality (8.29) is in fact an equality, which then implies that, $\mathrm{tr}(AQBP^T) = \mathrm{tr}(AQBQ^T) = \mathrm{tr}(APBP^T)$. We conclude that P and Q are two solutions of the GW problem and the equality between COOT and GW holds. When semi-definiteness is replaced by definiteness, Inequality (8.29) becomes an equality if and only if $P = Q$. ■

8.2.2 Proofs related to Continuous Co-Optimal Transport

Proof of Proposition 3.2.1. Denote $S = X_1 \times Y_1 \times X_2 \times Y_2$. For convenience, we also write $|c_X - c_Y|^p(x_1, y_1, x_2, y_2) := |c_X(x_1, x_2) - c_Y(y_1, y_2)|^p$. Now, the COOT problem can be rewritten

as

$$\mathrm{COOT}(\mathcal{X},\mathcal{Y}) = \inf_{\pi\in E_{co}} \int_S |c_X - c_Y|^p \, d\pi, \tag{8.33}$$

where the set

$$E_{co} = \{\pi \in \mathcal{P}(S) : \pi = \pi_1 \otimes \pi_2, \text{ where } \pi_k \in U(\mu_k^X, \mu_k^Y)\}, \tag{8.34}$$

contains the factorizable multi-marginal transport plans. Now, Lemmas 8.2.3 and 8.2.4 below imply that Problem (8.33) always admits a minimizer. ■

Lemma 8.2.2. *Countable product of Polish spaces is also a Polish space.*

Proof of Lemma 8.2.2. First, the countable product of completely metrizable spaces is also completely metrizable (see for example, Proposition 1.4 in (Dominique, 2020))

Second, we show that countable product of seperable spaces is also seperable. Given a sequence of seperable spaces $(X_k)_k$, let D_k be a countable dense subset of X_k. For each k, we fix a point $x_k \in X_k$. For each integer $j \geq 1$, define

$$\widetilde{D}_j = \prod_{k=1}^{j} D_k \times \prod_{k>j} \{x_k\} \quad \text{and} \quad \widetilde{D} = \cup_j \widetilde{D}_j. \tag{8.35}$$

Then clearly $\widetilde{D}_j$ is countable, for every $j \geq 1$, which implies $\widetilde{D}$ is also countable. To show that $\widetilde{D}$ is dense, every neighborhood E of $(x_1, ..., x_k, ...)$ (after reindexing) is of the form $E = \prod_{k=1}^{j} O_k \times \prod_{k>j}\{x_k\}$, for some integer $j \geq 1$ and $O_k \subset X_k$ is a neighborhood of x_k. As D_k is dense in X_k, we have $O_k \cap D_k \neq \emptyset$, thus $E \cap \widetilde{D} \neq \emptyset$. It follows that $\widetilde{D}$ is dense in $\prod_k X_k$. This concludes that the countable product of Polish spaces is also a Polish space. ■

Lemma 8.2.3. *If X_k and Y_k are Polish spaces, for every $k = 1, 2$, then E_{co} is non-empty and weakly compact in $\mathcal{P}(S)$.*

Proof of Lemma 8.2.3. Clearly, $(\mu_1^X \otimes \mu_1^Y) \otimes (\mu_2^X \otimes \mu_2^Y) \in E_{co}$, so E_{co} is not empty. We recall that $U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$ is the set of admissible couplings whose four marginals are $\mu_1^X, \mu_1^Y, \mu_2^X$ and μ_2^Y. First, observe that $E_{co} \subset U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$. Indeed, given $\pi_k \in U(\mu_k^X, \mu_k^Y)$, for $k = 1, 2$ and denote $\pi = \pi_1 \otimes \pi_2 \in E_{co}$. Then, for every $i, k \in \{1, 2\}$ and $i \neq k$, we have

$$\left(\int_{X_i \times Y_i} d\pi_i\right) \int_{Y_k} d\pi_k = d\mu_k^X. \tag{8.36}$$

Other marginal distributions can be calculated in a similar manner and we conclude that $\pi \in U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$.

As a direct generalization of Lemma 4.4 in (Villani, 2009), $U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$ is weakly compact in $\mathcal{P}(S)$. Thus, to show the compactness of E_{co}, it is enough to show that E_{co} is a weakly closed subset of $U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$. Take a sequence $(\pi^{(n)})_n \subset E_{co}$ such that $\pi^{(n)} \rightharpoonup \pi \in$

$U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$ (due to its compactness), we need to show that $\pi \in E_{co}$. As $(\pi^{(n)})_n \subset E_{co}$, there exist two sequences $(\pi_1^{(n)})_n \subset U(\mu_1^X, \mu_1^Y)$ and $(\pi_2^{(n)})_n \subset U(\mu_2^X, \mu_2^Y)$ such that $\pi^{(n)} = \pi_1^{(n)} \otimes \pi_2^{(n)}$.

For each $k = 1, 2$, due to the compactness of $U(\mu_k^X, \mu_k^Y)$ in $\mathcal{P}(X_k \times Y_k)$ (Lemma 4.4 in (Villani, 2009)), we can extract a converging subsequence $\pi_k^{(n_i^{(k)})} \rightharpoonup \pi_k \in U(\mu_k^X, \mu_k^Y)$, when $i \to \infty$. By applying Theorem 2.8 in (Billingsley, 1999) on the Polish space S (thanks to Lemma 8.2.2), we have $\pi_1^{(n_i^{(1)})} \otimes \pi_2^{(n_i^{(2)})} \rightharpoonup \pi_1 \otimes \pi_2 \in E_{co}$, when $i \to \infty$. This implies $\pi = \pi_1 \otimes \pi_2$, thus $\pi \in E_{co}$. ■

Lemma 8.2.4. *If c_X and c_Y are bounded measurable functions, then the functional $F : \pi \to \left(\int_S |c_X - c_Y|^p \, d\pi \right)^{1/p}$ is continuous on E_{co}.*

Proof of Lemma 8.2.4. It is enough to show that F is continuous on $U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$. To do this, we adapt the proof of Lemma 11 in (Chowdhury and Mémoli, 2019) by showing that there exists a sequence of continuous functions converging uniformly to F.

As $\mathcal{C}_b$ is dense in L^p (by applying Proposition 7.9 in (Folland, 1999) on Polish spaces endowed with finite measures), there exist two sequences of bounded continuous functions $(c_X^{(n)})_n \subset L^p(X, \mu^X)$ and $(c_Y^{(n)})_n \subset L^p(Y, \mu^Y)$ such that $||c_X - c_X^{(n)}||_{L^p(X,\mu^X)} \leq 1/n$ and $||c_Y - c_Y^{(n)}||_{L^p(Y,\mu^Y)} \leq 1/n$.

For each $n \in \mathbb{N}$, define $F_n : U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y) \to \mathbb{R}_{\geq 0}$ by $F_n(\pi) = ||c_X^{(n)} - c_Y^{(n)}||_{L^p(S,\pi)}$.

The compactness of $U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$ implies that, for every $\pi \in U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$, there exists a sequence $(\pi^{(m)})_m \subset U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$ such that $\pi^{(m)} \rightharpoonup \pi$. In particular, as $|c_X^{(n)} - c_Y^{(n)}|^p \in \mathcal{C}_b(S)$, we have

$$\lim_{m\to\infty} F_n(\pi^{(m)}) = \lim_{m\to\infty} \left(\int_S |c_X^{(n)} - c_Y^{(n)}|^p \, d\pi^{(m)} \right)^{1/p} = \left(\int_S |c_X^{(n)} - c_Y^{(n)}|^p \, d\pi \right)^{1/p} = F_n(\pi). \quad (8.37)$$

We deduce that F_n is sequentially continuous, thus continuous (by Remark 5.1.1 in (Ambrosio, Gigli, and Savaré, 2005)). Now, for any $\pi \in U(\mu_1^X, \mu_1^Y, \mu_2^X, \mu_2^Y)$, we have

$$\begin{aligned} |F_n(\pi) - F(\pi)| &= \left| ||c_X^{(n)} - c_Y^{(n)}||_{L^p(S,\pi)} - ||c_X - c_Y||_{L^p(S,\pi)} \right| \\ &\leq ||c_X^{(n)} - c_Y^{(n)} - (c_X - c_Y)||_{L^p(S,\pi)} \\ &\leq ||c_X^{(n)} - c_X||_{L^p(S,\pi)} + ||c_Y^{(n)} - c_Y||_{L^p(S,\pi)} \\ &= ||c_X^{(n)} - c_X||_{L^p(X,\mu^X)} + ||c_Y^{(n)} - c_Y||_{L^p(Y,\mu^Y)} \\ &\leq 2/n. \end{aligned} \quad (8.38)$$

The first inequality follows from a consequence of Minkowski's inequality: $\left| ||f|| - ||g|| \right| \leq ||f - g||$. The second inequality is the Minkowski's inequality. This implies that F_n converges uniformly to F, thus F is continuous. ■

Before proving the isomorphism and metric properties of COOT, let us first introduce the Monge's formulation of COOT.

$$\text{M-COOT}(\mathcal{X},\mathcal{Y}) = \inf_{\substack{T_1\in\mathcal{T}(\mu_1^X,\mu_1^Y)\\ T_2\in\mathcal{T}(\mu_2^X,\mu_2^Y)}} \iint \left|c_X(x_1,x_2) - c_Y\big(T_1(x_1),T_2(x_2)\big)\right|^p \, d\mu_1^X(x_1)\, d\mu_2^X(x_2), \quad (8.39)$$

where recall that $\mathcal{T}(\mu_k^X,\mu_k^Y) = \{T : X_k \to Y_k \text{ such that } T_\#\mu_k^X = \mu_k^Y\}$, for $k = 1,2$. It is not difficult to see that, M-COOT$(\mathcal{X},\mathcal{Y}) = 0$ if and only if $\mathcal{X} \in \text{RMS}(\mathcal{Y})$. Similar to the GW and Gromov-Monge distances, we have

Corollary 8.2.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two measure hypernetworks, then*

$$COOT(\mathcal{X},\mathcal{Y}) = \inf_{\mathcal{Z}\in RMS(\mathcal{X})} M\text{-}COOT(\mathcal{Z},\mathcal{Y}) = \inf_{\mathcal{Z}\in RMS(\mathcal{Y})} M\text{-}COOT(\mathcal{Z},\mathcal{X}). \quad (8.40)$$

Moreover, the infima are always attained: there exist two measure hypernetworks $\mathcal{Z}_x \in RMS(\mathcal{X})$ and $\mathcal{Z}_y \in RMS(\mathcal{Y})$ such that $COOT(\mathcal{X},\mathcal{Y}) = M\text{-}COOT(\mathcal{Z}_x,\mathcal{Y}) = M\text{-}COOT(\mathcal{Z}_y,\mathcal{X})$.

The proof is adapted directly from that of Theorem 14 in (Mémoli and Needham, 2022b). For self-contained purpose, we provide the complete proof here.

Proof of Corollary 8.2.1. Let (π_1^*,π_2^*) be a solution of the problem COOT$(\mathcal{X},\mathcal{Y})$. For $k = 1,2$, define the space $Z_k = X_k \times Y_k$ equipped with the probability measure $\mu_k^Z = \pi_k^* \in U(\mu_k^X,\mu_k^Y)$. Define the projection map $P_{X_k} : Z_k \to X_k$ by $P_{X_k}(x,y) = x$ and denote $c_Z = (P_{X_1},P_{X_2})^* c_X$. Clearly, the measure hypernetwork $\mathcal{Z} := ((Z_1,\mu_1^Z),(Z_2,\mu_2^Z),c_Z)$ is a RMS of $\mathcal{X}$. For $k = 1,2$, consider the canonical projection map $P_{Y_k} : Z_k \to Y_k$ defined by $P_{Y_k}(x,y) = y$, then P_{Y_k} is a transport map from μ_k^Z to μ_k^Y. Now, as $(P_{X_k},P_{Y_k})_\#\mu_k^Z = (P_{X_k},P_{Y_k})_\#\pi_k^* = \pi_k^*$, for $k = 1,2$, we have

$$\begin{aligned}
\text{M-COOT}(\mathcal{Z},\mathcal{Y}) &\le \int_{Z_1\times Z_2} |c_Z - (P_{Y_1},P_{Y_2})^* c_Y|^p \, d\mu_1^Z \, d\mu_2^Z && (8.41)\\
&= \int_{Z_1\times Z_2} |(P_{X_1},P_{X_2})^* c_X - (P_{Y_1},P_{Y_2})^* c_Y|^p \, d\mu_1^Z \, d\mu_2^Z && (8.42)\\
&= \int_S |c_X - c_Y|^p \, d(P_{X_1},P_{Y_1})_\#\mu_1^Z \, d(P_{X_2},P_{Y_2})_\#\mu_2^Z && (8.43)\\
&= \int_S |c_X - c_Y|^p \, d\pi_1^* \, d\pi_2^* && (8.44)\\
&= \text{COOT}(\mathcal{X},\mathcal{Y}), && (8.45)
\end{aligned}$$

and consequently,

$$\text{COOT}(\mathcal{X},\mathcal{Y}) \ge \inf_{\mathcal{Z}\in\text{RMS}(\mathcal{X})} \text{M-COOT}(\mathcal{Z},\mathcal{Y}). \quad (8.46)$$

For the reverse direction, let $\mathcal{Z}$ be a RMS of $\mathcal{X}$. Then, there exist two transport maps $f_k : Z_k \to$

X_k, for $k = 1, 2$ such that $c_Z = (f_1, f_2)^* c_X$, for $\mu_1^Z \otimes \mu_2^Z$-almost everywhere. The inequality (8.46) implies that we can safely exclude every $\mathcal{Z} \in \mathrm{RMS}(\mathcal{X})$ with $\text{M-COOT}(\mathcal{Z}, \mathcal{Y}) = \infty$ and consider only those with $\text{M-COOT}(\mathcal{Z}, \mathcal{Y}) < \infty$. In this case, there always exists a transport map g_k from μ_k^Z to μ_k^Y, for each $k = 1, 2$. If we define the map $(f_k, g_k) : Z_k \to X_k \times Y_k$ by $(f_k, g_k)(z_k) = (f_k(z_k), g_k(z_k))$, then $(f_k, g_k)_{\#\mu_k^Z} \in U(\mu_k^X, \mu_k^Y)$, for any $k = 1, 2$. Now,

$$\begin{aligned}
\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) &\leq \int_S |c_X - c_Y|^p \, d(f_1, g_1)_{\#}\mu_1^Z \, d(f_2, g_2)_{\#}\mu_2^Z && (8.47)\\
&= \int_{Z_1 \times Z_2} |(f_1, f_2)^* c_X - (g_1, g_2)^* c_Y|^p \, d\mu_1^Z \, d\mu_2^Z && (8.48)\\
&= \int_{Z_1 \times Z_2} |c_Z - (g_1, g_2)^* c_Y|^p \, d\mu_1^Z \, d\mu_2^Z. && (8.49)
\end{aligned}$$

As this is true for any $\mathcal{Z} \in \mathrm{RMS}(\mathcal{X})$ and any corresponding pair of transport maps (g_1, g_2), we have

$$\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) \leq \inf_{\mathcal{Z} \in \mathrm{RMS}(\mathcal{X})} \text{M-COOT}(\mathcal{Z}, \mathcal{Y}). \tag{8.50}$$

The equality then follows. Moreover, the first part of the proof also shows us how to construct a minimizer $\mathcal{Z}_x \in \mathrm{RMS}(\mathcal{X})$ such that $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = \text{M-COOT}(\mathcal{Z}_x, \mathcal{Y})$. Similarly, we have

$$\mathrm{COOT}(\mathcal{Y}, \mathcal{X}) = \inf_{\mathcal{Z} \in \mathrm{RMS}(\mathcal{Y})} \text{M-COOT}(\mathcal{Z}, \mathcal{X}). \tag{8.51}$$

By the symmetry of COOT, we deduce that

$$\inf_{\mathcal{Z} \in \mathrm{RMS}(\mathcal{X})} \text{M-COOT}(\mathcal{Z}, \mathcal{Y}) = \inf_{\mathcal{Z} \in \mathrm{RMS}(\mathcal{Y})} \text{M-COOT}(\mathcal{Z}, \mathcal{X}), \tag{8.52}$$

and there exists $\mathcal{Z}_y \in \mathrm{RMS}(\mathcal{Y})$ such that $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = \text{M-COOT}(\mathcal{Z}_y, \mathcal{X})$. ■

Proof of Corollary 3.2.1. Let us first prove the following simple lemma.

Lemma 8.2.5. *Let $f : X \to Y$ be a surjective map. If X and Y are finite and have the same cardinal, then f is bijective.*

Proof. Suppose $|X| = |Y| = n$. As f is surjective, for each $y \in Y$, the set $X(y) := \{x \in X : f(x) = y\}$ is not empty, *i.e.*, $|X(y)| \geq 1$. Clearly, $\cup_y X(y) \subset X$ and $X(y) \cap X(y') = \emptyset$, for any $y \neq y'$. Thus $|X| \geq |\cup_y X(y)| = \sum_y |X(y)| \geq n$. We deduce that $|X(y)| = 1$, for every $y \in Y$, thus f is bijective. ■

Now,

1. Clearly, strong isomorphism implies semi-strong isomorphism. Suppose $\mathcal{X}$ and $\mathcal{Y}$ are semi-strongly isomorphic, then $\mathcal{X}$ is a common RMS of $\mathcal{X}$ and $\mathcal{Y}$, meaning that $\mathcal{X}$ and $\mathcal{Y}$ are weakly isomorphic.

2. Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite measure hypernetworks. Suppose semi-strong isomorphism holds, then there exist four transport maps $f_k : X_k \to Y_k$ and $g_k : Y_k \to X_k$ such that $(f_k)_{\#}\mu_k^X = \mu_k^Y$ and $(g_k)_{\#}\mu_k^Y = \mu_k^X$, for $k = 1, 2$, and two pullback equalities hold everywhere. As a transport map is necessarily surjective, we must have $|Y_k| \geq |X_k|$ and $|X_k| \geq |Y_k|$, thus $|Y_k| = |X_k|$. By Lemma 8.2.5, we deduce that f_k (and g_k) are bijective. The strong isomorphism then follows.

3. By Proposition 3.2.2, the weak isomorphism implies $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = 0$. By Proposition 1 in (Redko et al., 2020), there exist two permutations (thus Borel measurable bijections) $\sigma_k : X_k \to Y_k$, for $k = 1, 2$, such that $c_X(i_1, i_2) = c_Y(\sigma_1(i_1), \sigma_2(i_2))$, for every $(i_1, i_2) \in X_1 \times X_2$. Furthermore, for every $j \in Y_k$, there exists a unique $i \in X_k$ such that $j = \sigma_k(i)$. Thus, by hypothesis, we have

$$\mu_k^Y(j) = \frac{1}{|Y_k|} = \frac{1}{|X_k|} = \mu_k^X(i) = \sum_{i' : \sigma_k(i') = j} \mu_k^X(i'), \tag{8.53}$$

which means $(\sigma_k)_{\#}\mu_k^X = \mu_k^Y$. So $\mathcal{X}$ is a RMS of $\mathcal{Y}$. Similarly, $\mathcal{Y}$ is also a RMS of $\mathcal{X}$. The semi-strong isomorphism then follows.

■

Proof of Proposition 3.2.2. Suppose two measure hypernetworks $\mathcal{X}$ and $\mathcal{Y}$ are weakly isomorphic, then there exists a measure hypernetwork $\mathcal{Z}^*$ which is a common RMS of $\mathcal{X}$ and $\mathcal{Y}$. In particular, $\text{M-COOT}(\mathcal{Z}^*, \mathcal{Y}) = 0$. By Corollary 8.2.1, we deduce that

$$\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = \inf_{\mathcal{Z} \in \mathrm{RMS}(\mathcal{X})} \text{M-COOT}(\mathcal{Z}, \mathcal{Y}) = 0. \tag{8.54}$$

Now, suppose $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = 0$, then by Corollary 8.2.1, there exists $\mathcal{Z}^* \in \mathrm{RMS}(\mathcal{X})$ such that $\text{M-COOT}(\mathcal{Z}^*, \mathcal{Y}) = 0$. But this also means $\mathcal{Z}^*$ is a RMS of $\mathcal{Y}$. We conclude that $\mathcal{X}$ and $\mathcal{Y}$ are weakly isomorphic. ■

Proof of Proposition 3.2.3. The proof of this result can be found in (Chowdhury et al., 2023). However, we still provide our proof here (slightly different but based on the same techniques).

For clarity, given three measure hypernetworks $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, we denote $S_{xy} = \prod_{k=1}^2 X_k \times Y_k, S_{yz} = \prod_{k=1}^2 Y_k \times Z_k, S_{xz} = \prod_{k=1}^2 X_k \times Z_k$ and $S_{xyz} = \prod_{k=1}^2 X_k \times Y_k \times Z_k$.

1. The positiveness is trivial. By Proposition 3.2.2, $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = 0$ if and only if $\mathcal{X}$ and $\mathcal{Y}$ are weakly isomorphic.

2. To show the symmetry, for each $k = 1, 2$, we define the bijection $f_k : X_k \times Y_k \to Y_k \times X_k$ by $f_k(x_k, y_k) = (y_k, x_k)$. Then, for any $\pi_k \in U(\mu_k^X, \mu_k^Y)$, where $k = 1, 2$, we have $(f_k)_{\#}\pi_k \in$

$U(\mu_k^Y, \mu_k^X)$ and

$$\begin{aligned}
\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) &= \inf_{\substack{\pi_k \in U(\mu_k^X, \mu_k^Y) \\ \forall k=1,2}} \int_{S_{xy}} |c_X - c_Y|^p \, d\pi_1 \, d\pi_2 \\
&= \inf_{\substack{\pi_k \in U(\mu_k^X, \mu_k^Y) \\ \forall k=1,2}} \int_{S_{yx}} |c_Y - c_X|^p \, d(f_1)_\# \pi_1 \, d(f_2)_\# \pi_2 \\
&= \inf_{\substack{\gamma_k \in U(\mu_k^Y, \mu_k^X) \\ \forall k=1,2}} \int_{S_{yx}} |c_Y - c_X|^p \, d\gamma_1 \, d\gamma_2 \\
&= \mathrm{COOT}(\mathcal{Y}, \mathcal{X}).
\end{aligned} \tag{8.55}$$

3. Now, we show the triangle inequality. Let $(\pi_1^{(YZ)}, \pi_2^{(YZ)})$ and $(\pi_1^{(XZ)}, \pi_2^{(XZ)})$ be the optimal couplings which minimize $\mathrm{COOT}(\mathcal{Y}, \mathcal{Z})$ and $\mathrm{COOT}(\mathcal{X}, \mathcal{Z})$, respectively, where $\pi_k^{(YZ)}$ and $\pi_k^{(XZ)}$ are in $U(\mu_k^Y, \mu_k^Z)$ and $U(\mu_k^X, \mu_k^Z)$, respectively, for every $k = 1, 2$.

   For each $k = 1, 2$, by the glueing lemma (Lemma 7.6 in (Villani, 2003)), there exists a probability measure $\sigma_k \in \mathcal{P}(X_k \times Y_k \times Z_k)$ such that $(P_{X_k Y_k})_\# \sigma_k = \pi_k^{(XY)}$ and $(P_{X_k Y_k})_\# \sigma_k = \pi_k^{(YZ)}$. Here, we define the projection maps

   - $P_{X_k} : X_k \times Y_k \times Z_k \to X_k$, where $P_{X_k}(x, y, z) = x$.
   - $P_{X_k Y_k} : X_k \times Y_k \times Z_k \to X_k \times Y_k$, where $P_{X_k Y_k}(x, y, z) = (x, y)$.
   - $P_{X_1 Z_1 X_2 Z_2} : S_{xyz} \to S_{xz}$, where $P_{X_1 Z_1 X_2 Z_2}(x_1, y_1, z_1, x_2, y_2, z_2) = (x_1, z_1, x_2, z_2)$.

   and all other projection maps are defined similarly. It follows that $(P_{X_k})_\# \sigma_k = \mu_k^X$ and $(P_{Z_k})_\# \sigma_k = \mu_k^Z$, thus $(P_{X_k Z_k})_\# \sigma_k \in U(\mu_k^X, \mu_k^Z)$. Furthermore, one also has $(P_{X_1 Z_1 X_2 Z_2})_\# \sigma = (P_{X_1 Z_1})_\# \sigma_1 \otimes (P_{X_2 Z_2})_\# \sigma_2$, where $\sigma = \sigma_1 \otimes \sigma_2$. Indeed, for any function $\phi \in \mathcal{C}_b(S_{xz})$, we have

$$\begin{aligned}
\int_{S_{xz}} \phi \, d(P_{X_1 Z_1 X_2 Z_2})_\# \sigma &= \int_{S_{xyz}} (\phi \circ P_{X_1 Z_1 X_2 Z_2}) \, d\sigma \\
&= \int_{S_{xyz}} (P_{X_1 Z_1}, P_{X_2 Z_2})^* \phi \, d\sigma_1 \, d\sigma_2 \\
&= \int_{S_{xz}} \phi \, d(P_{X_1 Z_1})_\# \sigma_1 \, d(P_{X_2 Z_2})_\# \sigma_2.
\end{aligned} \tag{8.56}$$

Here, with slight abuse of notation, we write $\phi(x_1, y_1, x_2, y_2) = \phi((x_1, y_1), (x_2, y_2))$. Now,

$$
\begin{aligned}
&\mathrm{COOT}(\mathcal{X}, \mathcal{Z})^{1/p} \\
&\leq \Big(\int_{S_{xz}} |c_X - c_Z|^p \, d(P_{X_1 Z_1})_{\#}\sigma_1 \, d(P_{X_2 Z_2})_{\#}\sigma_2 \Big)^{1/p} \\
&= \Big(\int_{S_{xz}} |c_X - c_Z|^p \, d(P_{X_1 Z_1 X_2 Z_2})_{\#}\sigma \Big)^{1/p} \\
&= \Big(\int_{S_{xyz}} (|c_X - c_Z|^p \circ P_{X_1 Z_1 X_2 Z_2}) \, d\sigma \Big)^{1/p} \\
&\leq \Big(\int_{S_{xyz}} (|c_X - c_Y|^p \circ P_{X_1 Y_1 X_2 Y_2}) \, d\sigma \Big)^{1/p} + \Big(\int_{S_{xyz}} (|c_Y - c_Z|^p \circ P_{Y_1 Z_1 Y_2 Z_2}) \, d\sigma \Big)^{1/p} \\
&= \Big(\int_{S_{xy}} |c_X - c_Y|^p \, d(P_{X_1 Y_1 X_2 Y_2})_{\#}\sigma \Big)^{1/p} + \Big(\int_{S_{yz}} |c_Y - c_Z|^p \, d(P_{Y_1 Z_1 Y_2 Z_2})_{\#}\sigma \Big)^{1/p} \\
&= \Big(\int_{S_{xy}} |c_X - c_Y|^p \, d(P_{X_1 Y_1})_{\#}\sigma_1 \, d(P_{X_2 Y_2})_{\#}\sigma_2 \Big)^{1/p} \\
&+ \Big(\int_{S_{yz}} |c_Y - c_Z|^p \, d(P_{Y_1 Z_1})_{\#}\sigma_1 \, d(P_{Y_2 Z_2})_{\#}\sigma_2 \Big)^{1/p} \\
&= \Big(\int_{S_{xy}} |c_X - c_Y|^p \, d\pi_1^{(XY)} \, d\pi_2^{(XY)} \Big)^{1/p} + \Big(\int_{S_{yz}} |c_Y - c_Z|^p \, d\pi_1^{(YZ)} \, d\pi_2^{(YZ)} \Big)^{1/p} \\
&= \mathrm{COOT}(\mathcal{X}, \mathcal{Y})^{1/p} + \mathrm{COOT}(\mathcal{Y}, \mathcal{Z})^{1/p}.
\end{aligned}
\tag{8.57}
$$

The first inequality is due to the sub-optimality of $((P_{X_1 Z_1})_{\#}\sigma_1, (P_{X_2 Z_2})_{\#}\sigma_2)$. The second one is the Minkowski inequality and the fact that: $|c_X(x_1, x_2) - c_Y(y_1, y_2)| \leq |c_X(x_1, x_2) - c_Z(z_1, z_2)| + |c_Z(z_1, z_2) - c_Y(y_1, y_2)|$, or more compactly

$$
|c_X - c_Z| \circ P_{X_1 Z_1 X_2 Z_2} \leq |c_X - c_Y| \circ P_{X_1 Y_1 X_2 Y_2} + |c_Y - c_Z| \circ P_{Y_1 Z_1 Y_2 Z_2}. \tag{8.58}
$$

■

Recall that

$$
E_{co} = \{\pi \in \mathcal{P}(S) : \pi = \pi_1 \otimes \pi_2, \text{ where } \pi_k \in U(\mu_k^X, \mu_k^Y)\}. \tag{8.59}
$$

First, observe that if $\pi \in E_{co}$ is a solution of COOT, then for any $\gamma \in E_{co}$, one has

$$
\begin{aligned}
0 &\leq \mathrm{COOT}_\varepsilon(\mathcal{X}, \mathcal{Y}) - \mathrm{COOT}(\mathcal{X}, \mathcal{Y}) \\
&\leq \left(\int_S |c_X - c_Y|^p \, d\gamma - \int_S |c_X - c_Y|^p \, d\pi \right) + \varepsilon \mathrm{KL}(\gamma | \mu_1 \otimes \mu_2).
\end{aligned}
\tag{8.60}
$$

The idea is to choose $\gamma \in E_{co}$ such that, for small ε, the quantity inside the bracket can be

arbitrarily small (but still positive, due to the optimality of π), and the KL divergence is always controlled, so that it does not blow up too fast. To do so, we extend the block approximation technique (Carlier et al., 2017) to the multi-marginal case.

Definition 8.2.1. *(Block approximation) Given an integer $K \geq 2$ and $p \geq 1$. For each $k = 1, ..., K$, let $\mu_k \in \mathcal{P}(\mathbb{R}^{n_k})$, for some integer $n_k \geq 1$, be a probability measure. For each tuple of integers $a_k = (a_k^{(1)}, ..., a_k^{(n_k)}) \in \mathbb{Z}^{n_k}$, we define the unit hypercube $Q_{a_k} = [a_k^{(1)}, a_k^{(1)} + 1[\times ... \times [a_k^{(n_k)}, a_k^{(n_k)} + 1[\subset \mathbb{R}^{n_k}$ and for $\Delta > 0$, we denote $Q_{a_k}^{\Delta} = [\Delta a_k^{(1)}, \Delta(a_k^{(1)} + 1)[\times ... \times [\Delta a_k^{(n_k)}, \Delta(a_k^{(n_k)} + 1)[\subset \mathbb{R}^{n_k}$ the rescaled hypercube by Δ of Q_{a_k}. For each $\pi \in \mathcal{P}(\prod_k \mathbb{R}^{n_k})$, we define its block approximation at scale Δ by*

$$\pi_{\Delta} := \sum_{\substack{a_k \in \mathbb{Z}^{n_k} \\ k=1,...,K}} \pi\Big(\prod_{k=1}^{K} Q_{a_k}^{\Delta}\Big)\big(\otimes_{k=1}^{K} \mu_k^{\Delta}\big), \tag{8.61}$$

where, for every Borel set $E_k \subset \mathbb{R}^{n_k}$, μ_k^{Δ} is the restriction of μ_k on $Q_{a_k}^{\Delta}$ defined by

$$\mu_k^{\Delta}(E_k) := \begin{cases} \frac{\mu_k(E_k \cap Q_{a_k}^{\Delta})}{\mu_k(Q_{a_k}^{\Delta})} & , \text{ if } \mu_k(Q_{a_k}^{\Delta}) > 0 \\ 0 & , \text{ otherwise.} \end{cases} \tag{8.62}$$

It is not difficult to see that block approximation of a product measure is also a product measure. Indeed, it is enough to consider the case $K = 2$. Suppose that $\pi = \pi_1 \otimes \pi_2$, then for any $\Delta > 0$,

$$\pi_{\Delta} = \sum_{a_1, a_2} \pi_1(Q_{a_1}^{\Delta})\, \pi_2(Q_{a_2}^{\Delta})\, \mu_1^{\Delta} \otimes \mu_2^{\Delta} = \left(\sum_{a_1} \pi_1(Q_{a_1}^{\Delta})\, \mu_1^{\Delta}\right) \otimes \left(\sum_{a_2} \pi_2(Q_{a_2}^{\Delta})\, \mu_2^{\Delta}\right). \tag{8.63}$$

Also, if a coupling is admissible, then so is its block approximation. More precisely, by Proposition 2.10 in (Carlier et al., 2017), for any $\pi \in U(\mu_1, ..., \mu_K)$ and $\Delta > 0$, we have $\pi_{\Delta} \in U(\mu_1, ..., \mu_K)$.

Proof of Proposition 3.2.5. This is an adaptation from the proof of quantitative bound between OT and regularized OT in (Genevay et al., 2019). Let $\pi^* \in E_{co}$ be a solution of COOT and denote π_{Δ} its block approximation at scale $\Delta > 0$. Clearly, $\pi_{\Delta} \in E_{co}$. Denote $\mu = \mu_X \otimes \mu_Y$. The sub-optimality of π_{Δ} implies

$$0 \leq \mathrm{COOT}_{\varepsilon}(\mathcal{X}, \mathcal{Y}) - \mathrm{COOT}(\mathcal{X}, \mathcal{Y}) \tag{8.64}$$

$$\leq \langle \pi_{\Delta}, |c_X - c_Y|^p \rangle - \langle \pi^*, |c_X - c_Y|^p \rangle + \varepsilon \mathrm{KL}(\pi_{\Delta} | \mu \otimes \mu). \tag{8.65}$$

Now, using the fact that $|x-y| \leq \max(x,y)$, for every $x, y \geq 0$, and for any i, j,

$$\sup_{(x_1,x_2)\in Q_{ij}^{\Delta}} |c_X(x_1,x_2)| \leq L \sup_{(x_1,x_2)\in Q_{ij}^{\Delta}} ||x_1 - x_2||^q \leq L(\Delta d_x^{1/p})^q, \tag{8.66}$$

we deduce that

$$\begin{aligned}
\langle \pi_\Delta, |c_X - c_Y|^p \rangle - \langle \pi^*, |c_X - c_Y|^p \rangle &\leq \sup_{i,j,k,l} \sup_{\substack{(x_1,x_2)\in Q_{ij}^{\Delta} \\ (y_1,y_2)\in Q_{kl}^{\Delta}}} |c_X(x_1,x_2) - c_Y(y_1,y_2)|^p && (8.67)\\
&\leq \max\Big\{ \sup_{(x_1,x_2)\in Q_{ij}^{\Delta}} |c_X(x_1,x_2)|^p, \sup_{(y_1,y_2)\in Q_{kl}^{\Delta}} |c_Y(y_1,y_2)|^p \Big\} && (8.68)\\
&\leq \max\big\{ L^p(\Delta d_x^{1/p})^{pq}, L^p(\Delta d_y^{1/p})^{pq} \big\} && (8.69)\\
&= L^p \Delta^{pq} d^q. && (8.70)
\end{aligned}$$

Following the proof of Theorem 1 in (Genevay et al., 2019), we obtain the bound for the KL term

$$\mathrm{KL}(\pi^{\Delta} | \mu \otimes \mu) \leq 2(d_x + d_y) \log(\frac{2D}{\Delta}) \leq 4d \log(\frac{2D}{\Delta}). \tag{8.71}$$

So, we have

$$\langle \pi_\Delta, |c_X - c_Y|^p \rangle - \langle \pi^*, |c_X - c_Y|^p \rangle + \varepsilon \mathrm{KL}(\pi^{\Delta} | \mu \otimes \mu) \leq L^p \Delta^{pq} d^q + 4d\varepsilon \log(\frac{2D}{\Delta}). \tag{8.72}$$

The RHS is a convex function of Δ, thus admits a minimizer $\Delta^{pq} = \frac{4d\varepsilon}{L^p d^q pq}$, thus

$$\langle \pi_\Delta, |c_X - c_Y|^p \rangle - \langle \pi^*, |c_X - c_Y|^p \rangle + \varepsilon \mathrm{KL}(\pi^{\Delta} | \mu \otimes \mu) \leq \frac{4d\varepsilon}{pq} + \frac{4d\varepsilon}{pq} \log\Big(\frac{(2D)^{pq} L^p d^q pq}{4d\varepsilon} \Big). \tag{8.73}$$

The result then follows. ■

8.2.3 Proofs related to MMOT-DC

Derivation of the Sinkhorn algorithm in entropic MMOT. The corresponding entropic dual problem of the primal problem (3.10) reads

$$\sup_{f_n \in \mathbb{R}^{a_n}} \sum_{n=1}^{N} \langle f_n, \mu_n \rangle - \varepsilon \sum_{i_1,\ldots,i_N} \exp\Big(\frac{\sum_n (f_n)_{i_n} - C_{i_1,\ldots,i_N}}{\varepsilon} \Big) + \varepsilon. \tag{8.74}$$

Algorithm 8 Sinkhorn algorithm for the entropic MMOT problem (3.10) from (Benamou et al., 2014).

Input. Histograms $\mu_1, ..., \mu_N$, hyperparameter $\varepsilon > 0$, cost tensor C and tuple of initial dual vectors $(f_1^{(0)}, ...f_N^{(0)})$.
Output. Optimal transport plan P and tuple of dual vectors $(f_1, ...f_N)$ (optional).

1. While not converge: for $n = 1, ..., N$,

$$f_n^{(t+1)} = \varepsilon \log \mu_n - \varepsilon \log \sum_{i_{-n}} \Big[\exp\Big(\frac{\sum_{j<n} (f_j^{(t+1)})_{i_j} + \sum_{j>n} (f_j^{(t)})_{i_j} - C_{\cdot, i_{-n}}}{\varepsilon} \Big) \Big]. \tag{8.79}$$

2. Return tensor P, where for $i_n \in [a_n]$, with $n \in [N]$,

$$P_{i_1,...,i_N} = \exp\Big(\frac{\sum_n (f_n)_{i_n} - C_{i_1,...,i_N}}{\varepsilon} \Big). \tag{8.80}$$

For each $n \in [N]$ and $i_n \in [a_n]$, the first order optimality condition reads

$$0 = (\mu_n)_{i_n} - \exp\big(\frac{(f_n)_{i_n}}{\varepsilon}\big) \sum_{i_{-n}} \exp\Big(\frac{\sum_{j \neq n} (f_j)_{i_j} - C_{i_1,...,i_N}}{\varepsilon} \Big), \tag{8.75}$$

where, with some abuse of notation, we write $i_{-n} = (i_1, ..., i_{n-1}, i_{n+1}, ..., i_N)$. Or, equivalently

$$(f_n)_{i_n} = \varepsilon \log(\mu_n)_{i_n} - \varepsilon \log \sum_{i_{-n}} \exp\Big(\frac{\sum_{j \neq n} (f_j)_{i_j} - C_{i_1,...,i_N}}{\varepsilon} \Big), \tag{8.76}$$

or a more compact form

$$f_n = \varepsilon \log \mu_n - \varepsilon \log \sum_{i_{-n}} \exp\Big(\frac{\sum_{j \neq n} (f_j)_{i_j} - C_{\cdot, i_{-n}}}{\varepsilon} \Big). \tag{8.77}$$

Using the primal-dual relation, we obtain the minimizer of the primal problem (3.10) by

$$P_{i_1,...,i_N} = \exp\Big(\frac{\sum_n (f_n)_{i_n} - C_{i_1,...,i_N}}{\varepsilon} \Big), \tag{8.78}$$

for $i_n \in [a_n]$, with $n \in [N]$. Similar to the entropic OT, the Sinkhorn algorithm 8 is also usually implemented in log-domain to avoid numerical instability.

F-MMOT of two components (*i.e.*, $M = 2$) is a variation of low nonnegative rank OT. For the sake of notational ease, we only consider the simplest case, where $N = 4$ and $M = 2$ with $\mathcal{T}_1 = (1, 2)$ and $\mathcal{T}_2 = (3, 4)$. However, the same argument still holds in the general case. First, we define three reshaping operations.

- Vectorization: concatenates rows of a matrix into a vector.

$$\text{vec} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}, \tag{8.81}$$

where each element $A_{i,j}$ of the matrix $A \in \mathbb{R}^{m\times n}$ is mapped to a unique element $b_{(i-1)n+j}$ of the vector $b \in \mathbb{R}^{mn}$, with $A_{i,j} = b_{(i-1)n+j}$, for $i = 1, ..., m$ and $j = 1, ..., n$. Conversely, each element b_k is mapped to a unique element $A_{k//n,n-k\%n}$, for every $k = 1, ..., mn$. Here, $k//n$ and $k\%n$ are the quotient and the remainder of the division of k by n, respectively, *i.e.*, if $k = qn + r$, with $0 \leq r < n$, then $k//n = q$ and $k\%n = r$.

- Matrization: transforms a 4D tensor to a 2D tensor (matrix) by vectorizing the first two and the last two dimensions of the tensor.

$$\text{mat} : \mathbb{R}^{n_1\times n_2\times n_3\times n_4} \to \mathbb{R}^{(n_1n_2)\times(n_3n_4)}, \tag{8.82}$$

where, similar to the vectorization, each element $P_{i,j,k,l}$ of the tensor $P \in \mathbb{R}^{n_1\times n_2\times n_3\times n_4}$ is mapped to a unique element $A_{(i-1)n_2+j,(k-1)n_4+l}$ of the matrix $A \in \mathbb{R}^{(n_1n_2)\times(n_3n_4)}$, with $P_{i,j,k,l} = A_{(i-1)n_2+j,(k-1)n_4+l}$.

- Concatenation: stacks vertically two equal-column matrices.

$$\begin{aligned} \text{con}_v :&\mathbb{R}^{m\times d} \times \mathbb{R}^{n\times d} \to \mathbb{R}^{(m+n)\times d} \\ &\big((u_1, ..., u_m), (v_1, ..., v_n)\big) \to (u_1, ..., u_m, v_1, ..., v_n)^T. \end{aligned} \tag{8.83}$$

Or, stacks horizontally two equal-row matrices

$$\begin{aligned} \text{con}_h :&\mathbb{R}^{n\times p} \times \mathbb{R}^{n\times q} \to \mathbb{R}^{n\times(p+q)} \\ &\big((u_1, ..., u_p), (v_1, ..., v_q)\big) \to (u_1, ..., u_p, v_1, ..., v_q). \end{aligned} \tag{8.84}$$

Lemma 8.2.6. *For any 4-D tensor $P \in \mathbb{R}^{n_1\times n_2\times n_3\times n_4}$, denote π its matrization. We have,*

$$vec\Big(\sum_{k,l} P_{\cdot,\cdot,k,l}\Big) = \sum_{n=1}^{n_3n_4} \pi_{\cdot,n} = \pi 1_{n_3n_4}, \tag{8.85}$$

where 1_n is the vector of ones in $\mathbb{R}^n$.

Proof. For $(i, j) \in [n_1] \times [n_2]$, we have

$$\text{vec}\Big(\sum_{k,l} P_{\cdot,\cdot,k,l}\Big)_{(i-1)n_2+j} = \sum_{k,l} P_{i,j,k,l} = \sum_{k,l} \pi_{(i-1)n_2+j,(k-1)n_4+l} = \sum_{n=1}^{n_3n_4} \pi_{(i-1)n_2+j,n}. \tag{8.86}$$

The result then follows. ■

Now, let $(e_i)_{i=1}^{n_1 n_2}$ be the standard basis vectors of $\mathbb{R}^{(n_1 n_2)}$, *i.e.*, $(e_i)_k = 1_{\{i=k\}}$. For each $P \in U(\mu)$, denote π its matrisation, then by Lemma 8.2.6, we have, for $i \in [n_1]$,

$$(\mu_1)_i = \sum_j \sum_{k,l} P_{i,j,k,l} = \sum_{j=1}^{n_2} \sum_{n=1}^{n_3 n_4} \pi_{(i-1)n_2+j,n}, \tag{8.87}$$

which can be recast in matrix form as $A_1^T \pi 1_{n_3 n_4} = \mu_1$, where the matrix $A_1 = \mathrm{con}_h(v_1, ..., v_{n_1}) \in \mathbb{R}^{(n_1 n_2) \times n_1}$, with $v_i \in \mathbb{R}^{(n_1 n_2)}$, where $v_i = \sum_{j=(i-1)n_2+1}^{in_2} e_j$, with $i \in [n_1]$. Similarly, $A_2 \pi 1_{n_3 n_4} = \mu_2$, where the matrix $A_2 = \mathrm{con}_h(I_{n_2}, ..., I_{n_2}) \in \mathbb{R}^{n_2 \times (n_1 n_2)}$, where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. Both conditions can be compactly written as $A_{12}^T \pi 1_{n_3 n_4} = \mu_{12}$, where the matrix $A_{12} = \mathrm{con}_h(A_1, A_2^T) \in \mathbb{R}^{(n_1 n_2) \times (n_1 + n_2)}$ and $\mu_{12} = \mathrm{con}_v(\mu_1, \mu_2) \in \mathbb{R}^{(n_1+n_2)}$. Note that μ_{12} is not a probability because its mass is 2. The matrix A_{12} has exactly $2n_1 n_2$ ones and the rest are zeros. An example of A_{12} is shown in Figure 8.1.

Similarly, for A_{34} and μ_{34} defined in the same way as A_{12} and μ_{12}, respectively, we establish the equality $A_{34}^T \pi^T 1_{n_1 n_2} = \mu_{34}$. As a side remark, both matrices A_{12}^T and A_{34}^T are *totally unimodular*, meaning that every square submatrix has determinant $-1, 0$, or 1. To handle the factorization

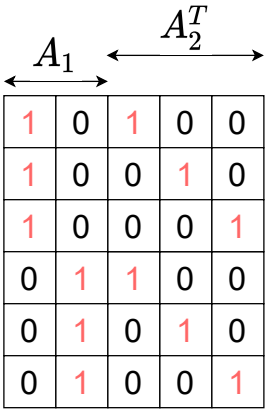

Figure 8.1 – An example of the matrix A_{12} when $n_1 = 2$ and $n_2 = 3$.

constraint, first we recall the following concept.

Definition 8.2.2. *Given a nonnegative matrix A, we define its nonnegative rank by*

$$\mathit{rank}_+(A) := \min\left\{r \geq 1 : A = \sum_{i=1}^{r} M_i, \ \mathit{where}\ \mathit{rank}(M_i) = 1, M_i \geq 0, \forall i\right\}. \tag{8.88}$$

By convention, zero matrix has zero (thus nonnegative) rank.

So, the constraint $P = P_1 \otimes P_2$ is equivalent to $\mathrm{mat}(P) = \mathrm{vec}(P_1)\mathrm{vec}(P_2)^T$. By Lemma 2.1 in (Cohen and Rothblum, 1993), $\mathrm{rank}_+(A) = 1$ if and only if there exist two nonnegative vectors u, v such that $A = uv^T$. Thus, the factorization constraint is equivalent to $\mathrm{rank}_+(\mathrm{mat}(P)) = 1$.

Denote $L = \mathrm{mat}(C)$ and $M = n_1 n_2, N = n_3 n_4$. Now, Problem (3.11) can be rewritten as

$$\min_{Q \in \mathbb{R}_{\geq 0}^{M \times N}} \langle L, Q \rangle \tag{8.89}$$

$$\text{such that } A_{12}^T Q 1_N = \mu_{12} \tag{8.90}$$

$$A_{34}^T Q^T 1_M = \mu_{34} \tag{8.91}$$

$$\mathrm{rank}_+(Q) = 1, \tag{8.92}$$

which is a variation of the low nonnegative rank OT problem studied in (Scetbon, Cuturi, and Peyré, 2021).

Proof of Proposition 3.3.1. The inequality $\mathrm{MMOT}(\mu) \leq \text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu)$ follows from the positivity of the KL divergence. On the other hand,

$$\text{F-MMOT}(\mathcal{T}, \mu) = \inf_{P \in U_\mathcal{T}} \langle C, P \rangle + \varepsilon \mathrm{KL}(P | P_{\#\mathcal{T}}), \tag{8.93}$$

because $\mathrm{KL}(P|P_{\#\mathcal{T}}) = 0$, for every $P \in U_\mathcal{T}$. As $U_\mathcal{T} \subset U(\mu)$, we have $\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu) \leq \text{F-MMOT}(\mathcal{T}, \mu)$.

Now, if $\text{F-MMOT}(\mathcal{T}, \mu) = 0$, then $\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu) = 0$. Conversely, if $\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu) = 0$, for $\varepsilon > 0$, then there exists $P^* \in U(\mu)$ such that $\langle C, P^* \rangle = 0$ and $P^* = P^*_{\#\mathcal{T}}$. Thus $\langle C, P^*_{\#\mathcal{T}} \rangle = 0$, which means $\text{F-MMOT}(\mathcal{T}, \mu) = 0$. ■

Proof of Proposition 3.3.2. The function $\varepsilon \to \text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu)$ is increasing on $\mathbb{R}_{\geq 0}$ and bounded, thus admits a finite limit $L \leq \text{F-MMOT}(\mathcal{T}, \mu)$, when $\varepsilon \to \infty$, and a finite limit $l \geq \mathrm{MMOT}(\mu)$, when $\varepsilon \to 0$.

Let P_ε be a solution of the problem $\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu)$. As $U(\mu)$ is compact, when either $\varepsilon \to 0$ or $\varepsilon \to \infty$, one can extract a converging subsequence (after reindexing) $(P_{\varepsilon_k})_k \to \widetilde{P} \in U(\mu)$, when either $\varepsilon_k \to 0$ or $\varepsilon_k \to \infty$. Thus, the convergence of the marginal distributions is also guaranteed, i.e $(P_{\varepsilon_k})_{\#\mathcal{T}_m} \to \widetilde{P}_{\#\mathcal{T}_m} \in U_{\mathcal{T}_m}$, for every $m \in [M]$, which implies that $P_{\varepsilon_k} - (P_{\varepsilon_k})_{\#\mathcal{T}} \to \widetilde{P} - \widetilde{P}_{\#\mathcal{T}}$.

When $\varepsilon \to 0$, let P^* be a solution of the problem $\mathrm{MMOT}(\mu)$. Then,

$$\langle C, P^* \rangle \leq \langle C, P_\varepsilon \rangle + \varepsilon \mathrm{KL}(P_\varepsilon | (P_\varepsilon)_{\#\mathcal{T}}) \leq \langle C, P^* \rangle + \varepsilon \mathrm{KL}(P^* | P^*_{\#\mathcal{T}}). \tag{8.94}$$

By the sandwich theorem, when $\varepsilon \to 0$, we have $\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu) \to \langle C, P^* \rangle = \mathrm{MMOT}(\mu)$. Furthermore, as

$$0 \leq \langle C, P_{\varepsilon_k} \rangle - \langle C, P^* \rangle \leq \varepsilon_k \mathrm{KL}(P^* | P^*_{\#\mathcal{T}}), \tag{8.95}$$

when $\varepsilon_k \to 0$, it follows that $\langle C, \widetilde{P} \rangle = \langle C, P^* \rangle$. So $\widetilde{P}$ is a solution of the problem $\mathrm{MMOT}(\mu)$. We conclude that any cluster point of the sequence of minimizers of $\text{MMOT-DC}_\varepsilon(\mathcal{T}, \mu)$ when $\varepsilon \to 0$

is a minimizer of MMOT(μ). As a byproduct, since

$$\mathrm{KL}(P^*|P^*_{\#\mathcal{T}}) - \mathrm{KL}(P_{\varepsilon_k}|(P_{\varepsilon_k})_{\#\mathcal{T}}) \geq \frac{\langle C, P_{\varepsilon_k}\rangle - \langle C, P^*\rangle}{\varepsilon_k} \geq 0, \tag{8.96}$$

we also deduce that $\mathrm{KL}(\widetilde{P}|\widetilde{P}_{\#\mathcal{T}}) \leq \mathrm{KL}(P^*|P^*_{\#\mathcal{T}})$ (so the cluster point $\widetilde{P}$ has minimal "mutual information").

On the other hand, when $\varepsilon \to \infty$, for $\mu^{\otimes N} = \mu_1 \otimes ... \otimes \mu_N$, one has

$$\langle C, \mu^{\otimes N}\rangle + \varepsilon \times 0 \geq \langle C, P_\varepsilon\rangle + \varepsilon \mathrm{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}) \geq \varepsilon \mathrm{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}). \tag{8.97}$$

Thus,

$$0 \leq \mathrm{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}) \leq \frac{1}{\varepsilon}\langle C, \mu^{\otimes N}\rangle \to 0, \text{ when } \varepsilon \to \infty, \tag{8.98}$$

which means $\mathrm{KL}(P_\varepsilon|(P_\varepsilon)_{\#\mathcal{T}}) \to 0$, when $\varepsilon \to \infty$. In particular, when $\varepsilon_k \to \infty$, we have $\mathrm{KL}(P_{\varepsilon_k}|(P_{\varepsilon_k})_{\#\mathcal{T}}) \to 0$. We deduce that $\mathrm{KL}(\widetilde{P}|\widetilde{P}_{\#\mathcal{T}}) = 0$, which implies $\widetilde{P} = \widetilde{P}_{\#\mathcal{T}}$.

Now, as MMOT-DC$_\varepsilon(\mathcal{T}, \mu) \geq \langle C, P_\varepsilon\rangle$, when $\varepsilon \to \infty$, we have $L \geq \langle C, \widetilde{P}\rangle = \langle C, \widetilde{P}_{\#\mathcal{T}}\rangle \geq$ F-MMOT$(\mathcal{T}, \mu)$. Thus $L = \langle C, \widetilde{P}\rangle =$ F-MMOT$(\mathcal{T}, \mu)$, i.e. MMOT-DC$_\varepsilon(\mathcal{T}, \mu) \to$ F-MMOT$(\mathcal{T}, \mu)$ when $\varepsilon \to \infty$. In this case, we also have that any cluster point of the sequence of minimizers of MMOT-DC$_\varepsilon(\mathcal{T}, \mu)$ is a minimizer of F-MMOT$(\mathcal{T}, \mu)$. ■

8.3 Appendix of Chapter 4

For later convenience, we define the function $|\xi_1 - \xi_2|^p : (X_1^s \times X_2^s) \times (X_1^f \times X_2^f) \to \mathbb{R}_{\geq 0}$ by

$$|\xi_1 - \xi_2|^p((x_1^s, x_2^s), (x_1^f, x_2^f)) := |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p, \tag{8.99}$$

and write the objective function of generalized COOT as

$$F_\rho(\pi^s, \pi^f) = \iint |\xi_1 - \xi_2|^p \mathrm{d}\pi^s \mathrm{d}\pi^f + \sum_{k=1}^{2} \rho_k D_k(\pi^s_{\#k} \otimes \pi^f_{\#k} | \mu^s_k \otimes \mu^f_k). \tag{8.100}$$

The generalized COOT now reads compactly as

$$\inf_{\substack{\pi^s \in \mathcal{M}^+(X_1^s \times X_2^s) \\ \pi^f \in \mathcal{M}^+(X_1^f \times X_2^f) \\ m(\pi^s) = m(\pi^f)}} F_\rho(\pi^s, \pi^f) \tag{8.101}$$

8.3.1 Proofs related to the properties of UCOOT

Claim 8.3.1. *When $D_k = \iota_=$ and μ_k^s, μ_k^f are probability measures, for $k = 1, 2$, then we recover COOT from generalized COOT.*

Proof of Claim 8.3.1. Under the above assumptions, the generalized COOT problem becomes

$$\begin{aligned}\inf_{\substack{\pi^s\in\mathcal{M}^+(X_1^s\times X_2^s)\\ \pi^f\in\mathcal{M}^+(X_1^f\times X_2^f)}} &\iint |\xi_1-\xi_2|^p \mathrm{d}\pi^s \mathrm{d}\pi^f \\ \text{subject to } &\pi^s_{\#1}\otimes\pi^f_{\#1}=\mu_1^s\otimes\mu_1^f \ \text{(C1)} \\ &\pi^s_{\#2}\otimes\pi^f_{\#2}=\mu_2^s\otimes\mu_2^f \ \text{(C2)} \\ &m(\pi^s)=m(\pi^f) \qquad \text{(C3)}\ .\end{aligned} \tag{8.102}$$

As $m(\pi)=m(\pi_{\#1})=m(\pi_{\#2})$, for any measure π, and μ_k^s,μ_k^f are probability measures, for $k=1,2$, one has $m(\pi^s)m(\pi^f)=1$, thus $m(\pi^s)=m(\pi^f)=1$. Now, the constraint C1 implies that $\int_{X_1^s}\mathrm{d}\pi^s_{\#1}\mathrm{d}\pi^f_{\#1}=\int_{X_1^s}\mathrm{d}\mu_1^s\mathrm{d}\mu_1^f$. Thus, $\pi^f_{\#1}=\mu_1^f$. Similarly, we have $\pi^s_{\#k}=\mu_k^s$ and $\pi^f_{\#k}=\mu_k^f$, for any $k=1,2$. We conclude that $\pi^f\in U(\mu_1^f,\mu_2^f)$ and $\pi^s\in U(\mu_1^s,\mu_2^s)$, and we obtain the COOT problem. ■

We will prove a more general version of Proposition 4.2.1.

Proposition 8.3.1 (Existence of minimizer)**.** *Denote* $S:=(X_1^s\times X_2^s)\times(X_1^f\times X_2^f)$. *Problem* (8.101) *admits a minimizer if at least one of the following conditions hold:*

1. *The entropy functions* ϕ_1 *and* ϕ_2 *are superlinear, i.e.,* $(\phi_1)'_\infty=(\phi_2)'_\infty=\infty$.
2. *The function* $|\xi_1-\xi_2|^p$ *has compact sublevels in* S *and* $\inf_S|\xi_1-\xi_2|^p+\rho_1(\phi_1)'_\infty+\rho_2(\phi_2)'_\infty>0$.

Proof of Proposition 8.3.1. We adapt the proof of Theorem 3.3 in (Liero, Mielke, and Savaré, 2018) and of Proposition 3 in (Séjourné, Vialard, and Peyré, 2021b). For convenience, we write $\mu_1=\mu_1^s\otimes\mu_1^f$ and $\mu_2=\mu_2^s\otimes\mu_2^f$. For each pair (π^s,π^f), denote $\pi=\pi^s\otimes\pi^f$. It can be shown that $\pi_{\#k}:=(P_{X_k^s\times X_k^f})_\#\pi=(P_{X_k^s})_\#\pi^s\otimes(P_{X_k^f})_\#\pi^f=\pi^s_{\#k}\otimes\pi^f_{\#k}$, for $k=1,2$. Indeed, for any function $\phi\in\mathcal{C}_b(X_k^s\times X_k^f)$, we have

$$\begin{aligned}\int_{X_k^s\times X_k^f}\phi\ \mathrm{d}(P_{X_k^sX_k^f})_\#\pi &= \int_S(\phi\circ P_{X_k^sX_k^f})\ \mathrm{d}\pi \\ &= \int_S\phi(x_k^s,x_k^f)\ \mathrm{d}\pi^s(x_1^s,x_2^s)\ \mathrm{d}\pi^f(x_1^f,x_2^f) \\ &= \int_{X_k^s\times X_k^f}\phi\ \mathrm{d}\pi^s_{\#k}\ \mathrm{d}\pi^f_{\#k}.\end{aligned} \tag{8.103}$$

Thus, Problem (8.101) can be rewritten as

$$\mathrm{UCOOT}_\rho(\mathcal{X}_1,\mathcal{X}_2)=\inf_{\pi\in E_{uco}}\int_S|\xi_1-\xi_2|^p\mathrm{d}\pi+\sum_{k=1,2}\rho_k D_{\phi_k}(\pi_{\#k}|\mu_k), \tag{8.104}$$

where

$$E_{uco} = \{\pi \in \mathcal{M}^+(S) | \pi = \pi^s \otimes \pi^f, \pi^s \in \mathcal{M}^+(X_1^s \times X_2^s), \pi^f \in \mathcal{M}^+(X_1^f \times X_2^f)\}. \tag{8.105}$$

Define

$$L(\pi) := \int_S |\xi_1 - \xi_2|^p \mathrm{d}\pi + \sum_{k=1,2} \rho_k D_{\phi_k}(\pi_{\#k} | \mu_k). \tag{8.106}$$

By Jensen's inequality, we have

$$\begin{aligned} L(\pi) &\geq m(\pi) \inf_S |\xi_1 - \xi_2|^p + \sum_{k=1,2} \rho_k m(\mu_k) \phi_k\Big(\frac{m(\pi_{\#k})}{m(\mu_k)}\Big) \\ &= m(\pi)\Big[\inf_S |\xi_1 - \xi_2|^p + \sum_{k=1,2} \rho_k \frac{m(\mu_k)}{m(\pi)} \phi_k\Big(\frac{m(\pi)}{m(\mu_k)}\Big)\Big], \end{aligned} \tag{8.107}$$

where, in the last equality, we use the relation $m(\pi) = m(\pi_{\#k})$, for $k = 1, 2$. It follows from the assumption that L is coercive. So, $L(\pi) \to \infty$ when $m(\pi) \to \infty$.

Clearly $\inf_{E_{uco}} L < \infty$ because $L((\mu_1^s \otimes \mu_2^s) \otimes (\mu_1^f \otimes \mu_2^f)) < \infty$. Let $(\pi_n)_n \subset E_{uco}$ be a minimizing sequence, meaning that $L(\pi_n) \to \inf_{E_{uco}} L$. Such sequence is necessarily bounded (otherwise, there exists a subsequence $(\pi_{n_k})_{n_k}$ with $m(\pi_{n_k}) \to \infty$ and the coercivity of L implies $L(\pi_{n_k}) \to \infty$, which is absurd). Suppose $m(\pi_n) \leq M$, for some $M > 0$. By Tychonoff's theorem, as X_k^s and X_k^f are compact spaces, so is the product space S. Thus, by Banach-Alaoglu theorem, the ball $B_M = \{\pi \in \mathcal{M}^+(S) : m(\pi) \leq M\}$ is weakly compact in $\mathcal{M}^+(S)$.

Consider the set $\overline{E}_{uco} = E_{uco} \cap B_M$, then clearly $(\pi_n)_n \subset \overline{E}_{uco}$. We will show that there exists a converging subsequence of $(\pi_n)_n$, whose limit is in $\overline{E}_{uco}$, thus $\overline{E}_{uco}$ is weakly compact. Indeed, by definition of E_{uco}, there exist two sequences $(\pi_n^s)_n$ and $(\pi_n^f)_n$ such that $\pi_n = \pi_n^s \otimes \pi_n^f$. We can assume furthermore that $m(\pi_n^s) = m(\pi_n^f) = \sqrt{m(\pi_n)} \leq \sqrt{M}$. As $m(\pi_n^s)$ and $m(\pi_n^f)$ are bounded, by reapplying Banach-Alaoglu theorem, one can extract two converging subsequences (after reindexing) $\pi_n^s \rightharpoonup \pi^s \in \mathcal{M}^+(X_1^s \times X_2^s)$ and $\pi_n^f \rightharpoonup \pi^f \in \mathcal{M}^+(X_1^f \times X_2^f)$, with $m(\pi^s) = m(\pi^f) \leq \sqrt{M}$. An immediate extension of Theorem 2.8 in (Billingsley, 1999) to the convergence of the products of bounded positive measures implies $\pi_n^s \otimes \pi_n^f \rightharpoonup \pi^s \otimes \pi^f \in \overline{E}_{uco}$.

Now, the lower semicontinuity of L implies that $\inf_{E_{uco}} L \geq L(\pi^s \otimes \pi^f)$, thus $L(\pi^s \otimes \pi^f) = \inf_{E_{uco}} L$ and (π^s, π^f) is a solution of Problem (8.101). ■

Claim 8.3.2. *Suppose that $\mathcal{X}_1$ and $\mathcal{X}_2$ are two finite sample-feature spaces such that (X_1^s, X_2^s) and (X_1^f, X_2^f) have the same cardinality and are equipped with the uniform measures $\mu_1^s = \mu_2^s$, $\mu_1^f = \mu_2^f$. Then $UCOOT_\rho(\mathcal{X}_1, \mathcal{X}_2) = 0$ if and only if there exist perfect alignments between rows (samples) and between columns (features) of the interaction matrices ξ_1 and ξ_2.*

Proof. Without loss of generality, we can assume that μ_k^s and μ_k^f are discrete uniform probability distributions, for $k = 1, 2$. By Proposition 1 in (Redko et al., 2020), under the assumptions on $\mathcal{X}_1$

and $\mathcal{X}_2$, we have $\mathrm{COOT}(\mathcal{X}_1, \mathcal{X}_2) = 0$ if and only if there exist perfect alignments between rows (samples) and between columns (features) of the interaction matrices ξ_1 and ξ_2. So, it is enough to prove that $\mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = 0$ if and only if $\mathrm{COOT}(\mathcal{X}_1, \mathcal{X}_2) = 0$.

Let (π^s, π^f) be a pair of equal-mass couplings such that $\mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = 0$. It follows that $\pi^s_{\#k} \otimes \pi^f_{\#k} = \mu^s_k \otimes \mu^f_k$, for $k = 1, 2$. Consequently, $m(\pi^s)m(\pi^f) = m(\mu^s_1)m(\mu^f_1) = 1$, so $m(\pi^s) = m(\pi^f) = 1$. Now, we have $\int_{X^s_k} \mathrm{d}\pi^s_{\#k}\, \mathrm{d}\pi^f_{\#k} = \int_{X^s_k} \mathrm{d}\mu^s_k\, \mathrm{d}\mu^f_k$, or equivalently, $\pi^f_{\#k} = \mu^f_k$. Similarly, $\pi^s_{\#k} = \mu^s_k$, meaning that $\pi^s \in U(\mu^s_1, \mu^s_2)$ and $\pi^f \in U(\mu^f_1, \mu^f_2)$. Thus, $\mathrm{COOT}(\mathcal{X}_1, \mathcal{X}_2) = \mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = 0$.

For the other direction, suppose that $\mathrm{COOT}(\mathcal{X}_1, \mathcal{X}_2) = 0$. Let (π^s, π^f) be a pair of couplings such that $\mathrm{COOT}(\mathcal{X}_1, \mathcal{X}_2) = 0$. As $\pi^s \in U(\mu^s_1, \mu^s_2)$ and $\pi^f \in U(\mu^f_1, \mu^f_2)$, one has $\mathrm{COOT}(\mathcal{X}_1, \mathcal{X}_2) = F_\rho(\pi^s, \pi^f) \geq \mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) \geq 0$, for every $\rho_1, \rho_2 > 0$. So, $\mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) = 0$. ■

8.3.2 Robustness of UCOOT and sensitivity of COOT

First, we recall our assumptions.

Assumption 8.3.1. *Consider two sample-feature spaces $\mathcal{X}_1$ and $\mathcal{X}_2$. Let ε^s (resp. ε^f) be a probability measure with compact support O^s (resp. O^f). For $a \in \{s, f\}$, define the noisy distribution $\widetilde{\mu}^a = \alpha_a \mu^a + (1 - \alpha_a)\varepsilon^a$, where $\alpha_a \in [0, 1]$. We assume that ξ_1 is defined on $(X^s_1 \cup O^s) \times (X^f_1 \cup O^f)$ and that ξ_1, ξ_2 are continuous on their supports. We denote the contaminated sample-feature space by $\widetilde{\mathcal{X}_1} = ((X^s_1 \cup O^s, \widetilde{\mu}^s_1), (\mathcal{X}^f_1 \cup O^f, \widetilde{\mu}^f_1), \xi_1)$. Finally, we define some useful minimal and maximal costs:*

$$\begin{cases} \Delta_0 = & \min\limits_{\substack{x^s_1 \in O^s, x^f_1 \in O^f \\ x^s_2 \in \mathcal{X}^s_2, x^f_2 \in \mathcal{X}^f_2}} \quad |\xi_1(x^s_1, x^f_1) - \xi_2(x^s_2, x^f_2)|^p \\ \Delta_\infty = & \max\limits_{\substack{x^s_1 \in \mathcal{X}^s_1 \cup O^s, x^f_1 \in \mathcal{X}^f_1 \cup O^f \\ x^s_2 \in \mathcal{X}^s_2, x^f_2 \in \mathcal{X}^f_2}} \quad |\xi_1(x^s_1, x^f_1) - \xi_2(x^s_2, x^f_2)|^p \end{cases}.$$

For convenience, we write $C = |\xi_1 - \xi_2|^p$ and $\widetilde{S} := (X^s_1 \cup O^s) \times X^s_2 \times (X^f_1 \cup O^f) \times X^f_2$.

Proof of Proposition 4.3.1. Consider a pair of feasible alignments (π^s, π^f). Since C is non-negative, taking the COOT integral over a smaller set leads to the lower bound:

$$\begin{aligned} \int_{\widetilde{S}} C \mathrm{d}\pi^s \mathrm{d}\pi^f &\geq \int_{O^s \times \mathcal{X}^s_2 \times O^f \times \mathcal{X}^f_2} C \mathrm{d}\pi^s \mathrm{d}\pi^f \\ &\geq \Delta_0 \int_{O^s \times \mathcal{X}^s_2 \times O^f \times \mathcal{X}^f_2} \mathrm{d}\pi^s \mathrm{d}\pi^f \\ &= \Delta_0 \int_{O^s \times O^f} \mathrm{d}\pi^s_{\#1} \mathrm{d}\pi^f_{\#1} \\ &\geq (1 - \alpha_s)(1 - \alpha_f)\Delta_0, \end{aligned} \tag{8.108}$$

where the last inequality follows from the marginal constraints. ■

Let us first recall the following lemma.

Lemma 8.3.1. *Let* $\varphi : t \in (0,1] \mapsto t\log(t) - t + 1$ *and* $f_{a,b} : t \in (0,1] \mapsto t \to at + b\varphi(t)$ *for some* $a, b > 0$. *Then:*

$$\min_{t\in(0,1]} f_{a,b}(t) = b(1 - e^{-a/b}) = f_{a,b}(e^{-\frac{a}{b}}). \tag{8.109}$$

Proof of Lemma 8.3.1. Since $f_{a,b}$ is convex, cancelling the gradient is sufficient for optimality. The solution follows immediately. ■

Proof of Theorem 4.3.1. The proof uses the same core idea of (Fatras et al., 2021) but is slightly more technical for two reasons: (1) we consider arbitrary outlier distributions instead of simple Diracs; (2) we consider sample-feature outliers which requires more technical derivations.

The idea of proof is as follows. First, we construct sample and feature couplings from the solution of "clean" UCOOT and the reference measures. Then, they are used to upper bound the "noisy" UCOOT. By manipulating this bound, the "clean" UCOOT term will appear. A variable $t \in (0,1)$ is also introduced in the fabricated couplings. The upper bound becomes a function of t and can be optimized to obtain the final bound.

Fabricating sample and feature couplings. Given the equal-mass solution (π^s, π^f) of the UCOOT problem, with $m(\pi^s) = m(\pi^f) = M$, consider, for $t \in (0,1)$, a pair of sub-optimal transport plans:

$$\widetilde{\pi}^s = \alpha_s \pi^s + t(1-\alpha_s)\varepsilon_s \otimes \mu_2^s \tag{8.110}$$

$$\widetilde{\pi}^f = \alpha_f \pi^f + t(1-\alpha_f)\varepsilon_f \otimes \mu_2^f. \tag{8.111}$$

Then, for $a \in \{s, f\}$, it holds:

- $\widetilde{\pi}^a_{\#1} = \alpha_k \pi^a_{\#1} + t(1-\alpha_a)\varepsilon_a$,
- $\widetilde{\pi}^a_{\#2} = \alpha_k \pi^a_{\#2} + t(1-\alpha_a)\mu_2^a$,
- $m(\widetilde{\mu}_1^a) = 1$ and $m(\widetilde{\pi}^a) = \alpha_a M + (1-\alpha_a)t$.

Establishing and manipulating the upper bound. Denote $q = (1-\alpha_s)(1-\alpha_f), s = \alpha_s(1-\alpha_f)+\alpha_f(1-\alpha_s)$ and recall that on $\widetilde{S}$, the cost C is upper bounded by $\Delta_\infty = \max_{\widetilde{S}} |\xi_1 - \xi_2|^p$.

First we upper bound the transportation cost:

$$\begin{aligned}
&\int_{\widetilde{S}} C \, \mathrm{d}\widetilde{\pi}^s \, \mathrm{d}\widetilde{\pi}^f \\
&= \alpha_s \alpha_f \int_{\widetilde{S}} C \, \mathrm{d}\pi^s \, d\pi^f + t \sum_{k \neq i} (1-\alpha_i)\alpha_k \int_{\widetilde{S}} C \, \mathrm{d}\varepsilon_i \, \mathrm{d}\mu_2^i \, \mathrm{d}\pi^k + qt^2 \int_{\widetilde{S}} C \, \mathrm{d}\varepsilon_s \, \mathrm{d}\mu_2^s \, \mathrm{d}\varepsilon_f \, \mathrm{d}\mu_2^f \\
&\leq \alpha_s \alpha_f \int_S C \, \mathrm{d}\pi^s \, \mathrm{d}\pi^f + \Delta_\infty (Ms + q)t,
\end{aligned} \tag{8.112}$$

since $t^2 \leq t$. Second, we turn to the KL marginal discrepancies. We would like to extract the KL terms involving only the clean transport plans from the contaminated ones. We first detail both joint KL divergences for the source measure indexed by 1. The same holds for the target measure:

$$\begin{aligned}
\mathrm{KL}(\widetilde{\pi}^s_{\#1} \otimes \widetilde{\pi}^f_{\#1} | \widetilde{\mu}^s_1 \otimes \widetilde{\mu}^f_1) &= \sum_{k \neq i} m(\widetilde{\pi}^i) \mathrm{KL}(\widetilde{\pi}^k_{\#1} | \widetilde{\mu}^k_1) + \prod_k \left(m(\widetilde{\pi}^k) - 1 \right) \\
\mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | \mu^s_1 \otimes \mu^f_1) &= M \sum_k \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + (M-1)^2.
\end{aligned} \tag{8.113}$$

Now we upper bound each smaller KL term using the joint convexity of the KL divergence:

$$\begin{aligned}
\mathrm{KL}(\widetilde{\pi}^k_{\#1} | \widetilde{\mu}^k_1) &\leq \alpha_k \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + (1-\alpha_k) \mathrm{KL}(t\varepsilon_k | \varepsilon_k) \\
&= \alpha_k \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + (1-\alpha_k)\varphi(t),
\end{aligned} \tag{8.114}$$

where $\varphi(t) = t \log t - t + 1$, for $t > 0$. Thus, for $k \neq i$:

$$\begin{aligned}
&m(\widetilde{\pi}^i) \mathrm{KL}(\widetilde{\pi}^k_{\#1} | \widetilde{\mu}^k_1) \leq m(\widetilde{\pi}^i) \alpha_k \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + m(\widetilde{\pi}^i)(1-\alpha_k)\varphi(t) \\
&= \alpha_i \alpha_k M \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + t(1-\alpha_i)\alpha_k \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + \alpha_i(1-\alpha_k)M\varphi(t) + tq\varphi(t).
\end{aligned} \tag{8.115}$$

Summing over f and s, we obtain:

$$\begin{aligned}
&\sum_{k \neq i} m(\widetilde{\pi}^i) \mathrm{KL}(\widetilde{\pi}^k_{\#1} | \widetilde{\mu}^k_1) \\
&\leq \alpha_s \alpha_f M \sum_k \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + t \sum_{k \neq i} (1-\alpha_i)\alpha_k \mathrm{KL}(\pi^k_{\#1} | \mu^k_1) + Ms\varphi(t) + 2qt\varphi(t) \\
&\leq (\alpha_s \alpha_f + \frac{ts}{M}) \left(\mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | \mu^s_1 \otimes \mu^f_1) - (1-M)^2 \right) + Ms\varphi(t) + 2qt\varphi(t).
\end{aligned} \tag{8.116}$$

where, in the last bound, we used the second equation of (8.113) and the fact that $\alpha_s(1-\alpha_f) \leq s$ and $\alpha_f(1-\alpha_s) \leq s$. The product of masses of (8.113) can be written:

$$\begin{aligned}\prod_k (m(\widetilde{\pi}^k) - 1) &= \prod_k (\alpha_k(M-1) + (1-\alpha_k)(t-1)) \\ &= \alpha_s\alpha_f(1-M)^2 + s(1-M)(1-t) + q(1-t)^2.\end{aligned} \tag{8.117}$$

Thus, combining these upper bounds for the source measure:

$$\begin{aligned}\mathrm{KL}(\widetilde{\pi}^s_{\#1} \otimes \widetilde{\pi}^f_{\#1} | \widetilde{\mu}^s_1 \otimes \widetilde{\mu}^f_1) \leq\ & \alpha_s\alpha_f \mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | \mu^s_1 \otimes \mu^f_1) \\ &+ \frac{ts}{M}\left(\mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | \mu^s_1 \otimes \mu^f_1) - (1-M)^2\right) \\ &+ \left[sM\varphi(t) + 2qt\varphi(t) + s(1-M)(1-t) + q(1-t)^2\right],\end{aligned} \tag{8.118}$$

and similarly, for the target measure:

$$\begin{aligned}\mathrm{KL}(\widetilde{\pi}^s_{\#2} \otimes \widetilde{\pi}^f_{\#2} | \mu^s_2 \otimes \mu^f_2) \leq\ & \alpha_s\alpha_f \mathrm{KL}(\pi^s_{\#2} \otimes \pi^f_{\#2} | \mu^s_2 \otimes \mu^f_2) \\ &+ \frac{ts}{M}\left(\mathrm{KL}(\pi^s_{\#2} \otimes \pi^f_{\#2} | \mu^s_2 \otimes \mu^f_2) - (1-M)^2\right) \\ &+ \left[sM\varphi(t) + 2qt\varphi(t) + s(1-M)(1-t) + q(1-t)^2\right].\end{aligned} \tag{8.119}$$

Then, for every $0 < t \leq 1$, by summing all bounds:

$$\begin{aligned}\mathrm{UCOOT}(\widetilde{\mathcal{X}_1}, \mathcal{X}_2) \leq\ & \alpha_s\alpha_f \mathrm{UCOOT}(\mathcal{X}_1, \mathcal{X}_2) + \Delta_\infty(Ms+q)t \\ &+ \frac{ts}{M}(\mathrm{UCOOT}(\mathcal{X}_1, \mathcal{X}_2) - (\rho_1+\rho_2)(1-M)^2) \\ &+ (\rho_1+\rho_2)\left[sM\varphi(t) + 2qt\varphi(t) + s(1-M)(1-t) + q(1-t)^2\right].\end{aligned} \tag{8.120}$$

Minimizing the upper bound with respect to t. To obtain the exponential bound, we would like have an upper bound of the form $at + b\varphi(t)$, so that Lemma 8.3.1 applies. Knowing that $1 \leq 2(t+\varphi(t))$ for any $t \in [0,1]$: Let's first isolate the quantity that is not of this form: We have:

$$\begin{aligned}2qt\varphi(t) + s(1-M) + q(t-1)^2 &= 2qt^2\log(t) - 2qt^2 + 2qt + s(1-M) + qt^2 - 2qt + q \\ &= 2qt^2\log(t) - qt^2 + s(1-M) + q \\ &= q\varphi(t^2) + s(1-M) \leq q + s(1-M) \\ &\leq 2(q + s(1-M))(t+\varphi(t)) \\ &= 2(1-\alpha_s\alpha_f - sM)(t+\varphi(t)).\end{aligned} \tag{8.121}$$

The new full bound is given by:

$$\mathrm{UCOOT}(\widetilde{\mathcal{X}_1}, \mathcal{X}_2) \leq \alpha_s \alpha_t \mathrm{UCOOT}(\mathcal{X}_1, \mathcal{X}_2) + A't + B'\varphi(t), \tag{8.122}$$

where

$$\begin{aligned} A' &= \Delta_\infty(Ms+q) + s(M-1) + \frac{s}{M}\mathrm{UCOOT}(\mathcal{X}_1, \mathcal{X}_2) - \frac{s}{M}(\rho_1+\rho_2)(1-M)^2 \\ &\quad + 2(\rho_1+\rho_2)(1-\alpha_s\alpha_f - sM) \\ &\leq \Delta_\infty(M+1) + M + \frac{1}{M}\mathrm{UCOOT}(\mathcal{X}_1, \mathcal{X}_2) + 2(\rho_1+\rho_2)(1-\alpha_s\alpha_f) = A \\ B' &= 2sM(\rho_1+\rho_2)(1-\alpha_s\alpha_f) \leq 2M(\rho_1+\rho_2)(1-\alpha_s\alpha_f) = B. \end{aligned} \tag{8.123}$$

In both inequalities, we use the fact that $s \leq 1 - \alpha_s\alpha_f \leq 1$. Using Lemma 8.3.1, we obtain

$$\mathrm{UCOOT}(\widetilde{\mathcal{X}_1}, \mathcal{X}_2) \leq \alpha_s\alpha_f \mathrm{UCOOT}(\mathcal{X}_1, \mathcal{X}_2) + B\left[1 - \exp\left(-\frac{A}{B}\right)\right]. \tag{8.124}$$

The upper bound of Theorem 4.3.1 then follows. ■

8.3.3 Numerical aspects

Proof of Proposition 4.4.1. Denote $\pi_\varepsilon = \pi_\varepsilon^s \otimes \pi_\varepsilon^f$.

1. When $\varepsilon \to \infty$: the sub-optimality of $\left(\sqrt{\frac{m(\mu^f)}{m(\mu^s)}}\mu^s, \sqrt{\frac{m(\mu^s)}{m(\mu^f)}}\mu^f\right)$ implies

$$\begin{aligned} \varepsilon\mathrm{KL}(\pi_\varepsilon|\mu^s\otimes\mu^f) &\leq F_\rho(\pi_\varepsilon^s, \pi_\varepsilon^f) + \varepsilon\mathrm{KL}(\pi_\varepsilon|\mu^s\otimes\mu^f) \\ &\leq F_\rho\left(\sqrt{\frac{m(\mu^f)}{m(\mu^s)}}\mu^s, \sqrt{\frac{m(\mu^s)}{m(\mu^f)}}\mu^f\right) + \varepsilon\mathrm{KL}(\mu^s\otimes\mu^f|\mu^s\otimes\mu^f) \\ &= \iint |\xi_1 - \xi_2|^p \mathrm{d}\mu^s\mathrm{d}\mu^f. \end{aligned} \tag{8.125}$$

   Thus,

$$0 \leq \mathrm{KL}(\pi_\varepsilon|\mu^s\otimes\mu^f) \leq \frac{1}{\varepsilon}\iint |\xi_1 - \xi_2|^p \mathrm{d}\mu^s\mathrm{d}\mu^f \to 0, \tag{8.126}$$

   whenever $\varepsilon \to \infty$. We deduce that $\mathrm{KL}(\pi_\varepsilon|\mu^s\otimes\mu^f)$, thus $\pi_\varepsilon \rightharpoonup \mu^s\otimes\mu^f$. The conclusion then follows.

2. Let (π_*^s, π_*^f) be a solution of $\mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2)$. The optimality of $(\pi_\varepsilon^s, \pi_\varepsilon^f)$ implies

$$\mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) \leq \mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2) + \varepsilon\mathrm{KL}(\pi_*^s\otimes\pi_*^f|\mu^s\otimes\mu^f). \tag{8.127}$$

   Thus, when $\varepsilon \to 0$, one has $\mathrm{UCOOT}_{\rho,\varepsilon}(\mathcal{X}_1, \mathcal{X}_2) \to \mathrm{UCOOT}_\rho(\mathcal{X}_1, \mathcal{X}_2)$. Moreover, following

the proof technique of Lemma 4 in (Pham et al., 2020), we can show that

$$\mathrm{UCOOT}_{\rho,\varepsilon}(\mathcal{X}_1,\mathcal{X}_2) = \sum_{k=1,2} \rho_k\, m(\mu_k^s)\, m(\mu_k^f) \tag{8.128}$$

$$+ \varepsilon \prod_{k=1,2} m(\mu_k^s)\, m(\mu_k^f) - (\rho_1+\rho_2+\varepsilon)\, m(\pi_\varepsilon^s)^2. \tag{8.129}$$

We deduce that $m(\pi_\varepsilon^s) \to m(\pi_*^s)$, when $\varepsilon \to 0$,
Now, for every $\varepsilon > 0$,

$$\begin{aligned}\langle C, \mu^s \otimes \mu^f\rangle &= F_\rho\left(\sqrt{\frac{m(\mu^f)}{m(\mu^s)}}\mu^s, \sqrt{\frac{m(\mu^s)}{m(\mu^f)}}\mu^f\right) + \varepsilon \mathrm{KL}(\mu^s\otimes\mu^f|\mu^s\otimes\mu^f)\\ &\geq F_\rho(\pi_\varepsilon^s,\pi_\varepsilon^f) + \varepsilon\mathrm{KL}(\pi_\varepsilon^s\otimes\pi_\varepsilon^f|\mu^s\otimes\mu^f)\\ &\geq F_\rho(\pi_\varepsilon^s,\pi_\varepsilon^f).\end{aligned} \tag{8.130}$$

On the other hand, following the same proof in Proposition 8.3.1, we can show that if $m(\pi_\varepsilon) \to \infty$, then $F_\rho(\pi_\varepsilon^s,\pi_\varepsilon^f) \to \infty$, which contradicts the above inequality. So, there exists $M > 0$ such that $m(\pi_\varepsilon) \leq M$, for every $\varepsilon > 0$.
The set $\widetilde{E}_{uco} = \{\pi \in \mathcal{M}^+(S) : m(\pi) \leq M\} \cap E_{uco}$ is clearly compact, thus from the sequence of minimisers $(\pi_\varepsilon)_\varepsilon \subset \widetilde{E}_{uco}$ (*i.e.*, $\pi_\varepsilon = \pi_\varepsilon^s \otimes \pi_\varepsilon^f$), we can extract a converging subsequence $(\pi_{\varepsilon_n})_{\varepsilon_n}$ such that $\pi_{\varepsilon_n} \to \widehat{\pi} = \widehat{\pi}^s \otimes \widehat{\pi}^f \in \widetilde{E}_{uco}$, with $m(\widehat{\pi}^s) = m(\widehat{\pi}^f)$. The continuity of the divergences implies that, $F_{\rho,\varepsilon}(\pi_{\varepsilon_n}^s,\pi_{\varepsilon_n}^f) \to F_\rho(\widehat{\pi}^s,\widehat{\pi}^f)$, when $\varepsilon \to 0$. We deduce that $\mathrm{UCOOT}_\rho(\mathcal{X}_1,\mathcal{X}_2) = F_\rho(\widehat{\pi}^s,\widehat{\pi}^f)$, or equivalently $(\widehat{\pi}^s,\widehat{\pi}^f)$ is a solution of $\mathrm{UCOOT}_\rho(\mathcal{X}_1,\mathcal{X}_2)$. Moreover, we have

$$\begin{aligned}0 &\leq F_\rho(\pi_{\varepsilon_n}^s,\pi_{\varepsilon_n}^f) - F_\rho(\pi_*^s,\pi_*^f)\\ &\leq \varepsilon_n\Big(\mathrm{KL}(\pi_*^s\otimes\pi_*^f|\mu^s\otimes\mu^f) - \mathrm{KL}(\pi_{\varepsilon_n}^s\otimes\pi_{\varepsilon_n}^f|\mu^s\otimes\mu^f)\Big).\end{aligned} \tag{8.131}$$

Dividing by ε_n in Equation (8.131) and let $\varepsilon_n \to 0$, we have

$$\mathrm{KL}(\widehat{\pi}^s\otimes\widehat{\pi}^f|\mu^s\otimes\mu^f) \leq \mathrm{KL}(\pi_*^s\otimes\pi_*^f|\mu^s\otimes\mu^f). \tag{8.132}$$

and we deduce that

$$\mathrm{KL}(\widehat{\pi}^s\otimes\widehat{\pi}^f|\mu^s\otimes\mu^f) = \min_{(\pi^s,\pi^f)} \mathrm{KL}(\pi^s\otimes\pi^f|\mu^s\otimes\mu^f), \tag{8.133}$$

where the infimum is taken over all solutions of $\mathrm{UCOOT}_\rho(\mathcal{X}_1,\mathcal{X}_2)$.

■

8.3.4 Experimental details

8.3.5 Heterogenous Domain Adaptation (HDA)

More details on label propagation Once the sample coupling P is learned, the label propagation works as follows: suppose the labels contain K different classes, we apply the one-hot encoding to the source label $y^{(s)}$ to obtain $D^{(s)} \in \mathbb{R}^{K \times n_s}$ where $D^{(s)}_{ki} = 1_{\{y^{(s)}_i = k\}}$. The label proportions on the target data are estimated by: $L = D^{(s)}P \in \mathbb{R}^{K \times n_t}$. Then the prediction can be generated by choosing the label with the highest proportion, *i.e.*, $\widehat{y}^{(t)}_j = \text{argmax}_k \, L_{kj}$.

Paramater validation We tune the hyperparameters of each method via grid search.

- For COOT, we choose the regularisation on the feature and sample couplings $\varepsilon_f, \varepsilon_s \in \{0, 0.01, 0.1, 0.5\}$.
- For GW, we choose the regularisation parameter $\varepsilon \in \{0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$.
- For UGW and UCOOT, we choose $\rho_1, \rho_2 \in \{1, 5, 20, 50\}$ and $\varepsilon \in \{0.01, 0.05, 0.1, 0.5\}$. Furthermore, for UGW and GW, before calculating the Euclidean distance matrix for each domain, the matrix of domain data is normalised by max scaling, so that its coordinates are bounded in $[-1, 1]$. This pre-processing step improves the performance of the for UGW and GW.

For each method, for each combination of tuple of hyperparameters, first, we choose a pair amongst 9 pairs, then repeat 10 times the training procedure, in which the optimal plan is estimated, then used to calculate the accuracy. We choose the tuple of hyperparameters corresponding to the highest average accuracy. This optimal tuple is then applied to all other 8 tasks, where in each task, the training procedure is repeated 10 times and we report the average accuracy.

When there is no regularization In the above hyperparameter tuning process, we only considered $\varepsilon > 0$ for UCOOT and UGW, so that the scaling algorithm (Chizat et al., 2018b) is applicable. We note that, the MM solver can allow us to handle the case $\varepsilon = 0$ (*i.e.*, we can estimate directly UCOOT, rather than via its entropic approximation). In this case, we also tune $\rho_1, \rho_2 \in \{1, 50, 20, 50\}$ and follow exactly the same tuning and testing procedure as in the case $\varepsilon > 0$. We report our finding in Table 8.1. We observe that, in many tasks, the performance remains competitive while enjoying lower variance.

Sensitivity analysis We report the sensitivity of UCOOT's performance to the hyperparameters ε, ρ_1 and ρ_2 for two tasks C→W and A→A in Tables 8.2 to 8.4. In general, the performance depends significantly on the choice of hyperparameters. In Table 8.2, given fixed values of ρ_1 and ρ_2, UCOOT performs badly for either too small or large values of ε, indicating

CaffeNet → GoogleNet			
Domains	COOT	UCOOT ($\varepsilon > 0$)	UCOOT ($\varepsilon = 0$)
C → C	36.40 (± 12.94)	**44.05 (± 19.33)**	38.60 (± 9.16)
C → A	28.30 (± 11.78)	**31.90 (± 7.43)**	29.45 (± 9.94)
C → W	19.55 (± 14.51)	28.55 (± 6.60)	**40.85 (± 12.53)**
A → C	**41.80 (± 14.81)**	39.15 (± 17.98)	18.00 (± 9.22)
A → A	**57.90 (± 16.84)**	42.45 (± 15.47)	40.40 (± 8.40)
A → W	42.10 (± 7.80)	48.55 (± 13.06)	**49.15 (± 6.64)**
W → C	8.60 (± 6.56)	**69.80 (± 14.91)**	19.70 (± 5.79)
W → A	16.65 (± 10.01)	**30.55 (± 10.09)**	25.90 (± 5.48)
W → W	**75.30 (± 3.26)**	51.50 (± 20.51)	49.55 (± 6.02)
Average	36.29 (± 10.95)	**42.94 (± 13.93)**	34.62 (± 11.17)

Table 8.1 – Unsupervised HDA from CaffeNet to GoogleNet for $\varepsilon > 0$ and $\varepsilon = 0$. UCOOT ($\varepsilon > 0$) corresponds to the model where ε, ρ_1 and ρ_2 are tuned, with $\varepsilon > 0$, and UCOOT ($\varepsilon = 0$) means that $\varepsilon = 0$ and only ρ_1, ρ_2 are tuned.

that regularization is necessary but should not be too strong. From Table 8.3, we see that large value of ρ_1 degrades the performance, meaning that the marginal constraints on the source distributions should not be too tight. Meanwhile, it seems that large ρ_2 is preferable, so the marginal distributions on the target spaces should not be too relaxed.

	CaffeNet → GoogleNet				
Domains	$\varepsilon = 0.03$	0.05	0.1	0.2	0.4
C → W	27.65 (± 11.34)	37.20 (± 9.35)	34.75 (± 13.04)	17.00 (± 5.92)	11.25 (± 1.66)
A → A	21.95 (± 9.46)	35.30 (± 15.11)	41.15 (± 19.16)	58.45 (± 15.54)	8.90 (± 1.34)

Table 8.2 – Sensitivity of UCOOT to ε in tasks C→W and A→A. We fix $\rho_2 = 50$ and $\rho_1 = 1$ and show the accuracy for various value of ε.

	CaffeNet → GoogleNet				
Domains	$\rho_1 = 20$	40	50	60	80
C → W	35.80 (± 9.33)	37.35 (± 13.82)	27.45 (± 8.33)	32.45 (± 11.62)	30.15 (± 12.89)
A → A	55.20 (± 18.44)	44.15 (± 21.54)	24.30 (± 15.58)	36.10 (± 23.97)	24.80 (± 15.08)

Table 8.3 – Sensitivity of UCOOT to ρ_1 in tasks C→W and A→A. We fix $\rho_2 = 1$ and $\varepsilon = 0.1$ and show the accuracy for various value of ρ_1.

	CaffeNet → GoogleNet				
Domains	$\rho_2 = 0.3$	0.5	1	2	4
C → W	34.20 (± 9.83)	34.45 (± 10.80)	34.75 (± 13.04)	29.70 (± 10.55)	32.30 (± 18.81)
A → A	20.75 (± 10.11)	29.00 (± 15.79)	41.15 (± 19.16)	32.65 (± 8.80)	49.95 (± 15.75)

Table 8.4 – Sensitivity of UCOOT to ρ_2 in tasks C→W and A→A. We fix $\rho_1 = 50$ and $\varepsilon = 0.1$ and show the accuracy for various value of ρ_2.

8.4 Appendix of Chapter 5

8.4.1 Proofs related to Fused Unbalanced Gromov-Wasserstein

Proof of Corollary 5.2.1. Let (P, Q) be a solution of the problem $\widetilde{\text{FUGW}}(\mathcal{X}^s, \mathcal{X}^t)$. This means $P_{\#1} = Q_{\#1}, P_{\#2} = Q_{\#2}$. Using the relation

$$\text{KL}(p \otimes q | a \otimes b) = m(q)\text{KL}(p|a) + m(p)\text{KL}(q|b) + \big(m(P) - m(a)\big)\big(m(Q) - m(b)\big), \tag{8.134}$$

we have,

1. $\text{KL}(P_{\#1} \otimes Q_{\#1} | w^s \otimes w^s) = \text{KL}(P_{\#1} \otimes P_{\#1} | w^s \otimes w^s) = \text{KL}(Q_{\#1} \otimes Q_{\#1} | w^s \otimes w^s)$.
2. $\text{KL}(P_{\#2} \otimes Q_{\#2} | w^t \otimes w^t) = \text{KL}(P_{\#2} \otimes P_{\#2} | w^t \otimes w^t) = \text{KL}(Q_{\#2} \otimes Q_{\#2} | w^t \otimes w^t)$.
3. $\text{KL}\Big(P \otimes Q | (w^s \otimes w^t) \otimes (w^s \otimes w^t)\Big) = \text{KL}\Big(P \otimes P | (w^s \otimes w^t) \otimes (w^s \otimes w^t)\Big) = \text{KL}\Big(P \otimes Q | (w^s \otimes w^t) \otimes (w^s \otimes w^t)\Big)$.

So, the inequality $\widetilde{\text{FUGW}}(\mathcal{X}^s, \mathcal{X}^t) \leq \text{FUGW}(\mathcal{X}^s, \mathcal{X}^t)$ and the suboptimality of P, Q with respect to the problem $\text{FUGW}(\mathcal{X}^s, \mathcal{X}^t)$ imply that

$$2\langle G, P \otimes Q \rangle \leq \langle G, P \otimes P \rangle + \langle G, Q \otimes Q \rangle. \tag{8.135}$$

Now, following exactly the same proof of Proposition 3.1.1, the above inequality becomes an equality. ■

8.5 Appendix of Chapter 5

8.5.1 Proofs related to Augmented Gromov-Wasserstein

Proof of Proposition 6.2.1. The proof of this proposition can be adapted directly from (Vayer et al., 2019a). For self-contained purpose, we reproduce the proof here. Denote

- (P_α, Q_α) the optimal sample and feature couplings for $\text{AGW}_\alpha(\mathcal{X}, \mathcal{Y})$.

- (P_0, Q_0) the optimal sample and feature couplings for $\text{COOT}(\mathcal{X}, \mathcal{Y})$ (corresponding to $\alpha = 0$).
- P_1 the optimal sample coupling for $\text{GW}(\mathcal{X}, \mathcal{Y})$ (corresponding to $\alpha = 1$).

Due to the suboptimality of P_α for GW and (P_1, Q_0) for AGW, we have

$$\begin{aligned}\alpha\langle L(C^x, C^y) \otimes P_1, P_1\rangle &\leq \alpha\langle L(C^x, C^y) \otimes P_\alpha, P_\alpha\rangle + (1-\alpha)\langle L(X, Y) \otimes Q_\alpha, P_\alpha\rangle &(8.136)\\ &\leq \alpha\langle L(C^x, C^y) \otimes P_1, P_1\rangle + (1-\alpha)\langle L(X, Y) \otimes Q_0, P_1\rangle, &(8.137)\end{aligned}$$

or equivalently

$$\alpha\text{GW}(\mathcal{X}, \mathcal{Y}) \leq \text{AGW}_\alpha(\mathcal{X}, \mathcal{Y}) \leq \alpha\text{GW}(\mathcal{X}, \mathcal{Y}) + (1-\alpha)\langle L(X, Y) \otimes Q_0, P_1\rangle. \tag{8.138}$$

Similarly, we have

$$\begin{aligned}(1-\alpha)\text{COOT}(\mathcal{X}, \mathcal{Y}) &\leq \text{AGW}_\alpha(\mathcal{X}, \mathcal{Y}) &(8.139)\\ &\leq (1-\alpha)\text{COOT}(\mathcal{X}, \mathcal{Y}) + \alpha\langle L(C^x, C^y) \otimes P_0, P_0\rangle. &(8.140)\end{aligned}$$

The interpolation property then follows by the sandwich theorem.

Regarding the relaxed triangle inequality, given three weighted matrices $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, denote $(P^{XY}, Q^{XY}), (P^{YZ}, Q^{YZ})$ and (P^{XZ}, Q^{XZ}) the solutions of $\text{AGW}_\alpha(\mathcal{X}, \mathcal{Y}), \text{AGW}_\alpha(\mathcal{Y}, \mathcal{Z})$ and $\text{AGW}_\alpha(\mathcal{X}, \mathcal{Z})$, respectively. We define $P = P^{XY}\text{diag}\left(\frac{1}{\mu_1^Y}\right)P^{YZ}$ and $Q = Q^{XY}\text{diag}\left(\frac{1}{\mu_2^Y}\right)Q^{YZ}$.

Then, $P \in U(\mu_1^X, \mu_1^Z)$ and $Q \in U(\mu_2^X, \mu_2^Z)$. The suboptimality of (P, Q) implies that

$$\frac{\mathrm{AGW}_\alpha(\mathcal{X}, \mathcal{Z})}{2} \tag{8.141}$$

$$\leq \alpha \sum_{i,j,k,l} \frac{|C^x_{i,j} - C^z_{k,l}|^2}{2} P_{i,k} P_{j,l} + (1-\alpha) \sum_{i,j,k,l} \frac{|X_{i,j} - Z_{k,l}|^2}{2} P_{i,k} Q_{j,l} \tag{8.142}$$

$$= \alpha \sum_{i,j,k,l} \frac{|C^x_{i,j} - C^z_{k,l}|^2}{2} \left(\sum_e \frac{P^{XY}_{i,e} P^{YZ}_{e,k}}{(\mu_1^Y)_e} \right) \left(\sum_o \frac{P^{XY}_{j,o} P^{YZ}_{o,l}}{(\mu_1^Y)_o} \right) \tag{8.143}$$

$$+ (1-\alpha) \sum_{i,j,k,l} \frac{|X_{i,j} - Z_{k,l}|^2}{2} \left(\sum_e \frac{P^{XY}_{i,e} P^{YZ}_{e,k}}{(\mu_1^Y)_e} \right) \left(\sum_o \frac{Q^{XY}_{j,o} Q^{YZ}_{o,l}}{(\mu_2^Y)_o} \right) \tag{8.144}$$

$$\leq \alpha \sum_{i,j,k,l,e,o} |C^x_{i,j} - C^y_{e,o}|^2 \frac{P^{XY}_{i,e} P^{YZ}_{e,k}}{(\mu_1^Y)_e} \frac{P^{XY}_{j,o} P^{YZ}_{o,l}}{(\mu_1^Y)_o} \tag{8.145}$$

$$+ (1-\alpha) \sum_{i,j,k,l,e,o} |X_{i,j} - Y_{e,o}|^2 \frac{P^{XY}_{i,e} P^{YZ}_{e,k}}{(\mu_1^Y)_e} \frac{Q^{XY}_{j,o} Q^{YZ}_{o,l}}{(\mu_2^Y)_o} \tag{8.146}$$

$$+ \alpha \sum_{i,j,k,l,e,o} |C^y_{e,o} - C^z_{k,l}|^2 \frac{P^{XY}_{i,e} P^{YZ}_{e,k}}{(\mu_1^Y)_e} \frac{P^{XY}_{j,o} P^{YZ}_{o,l}}{(\mu_1^Y)_o} \tag{8.147}$$

$$+ (1-\alpha) \sum_{i,j,k,l,e,o} |Y_{e,o} - Z_{k,l}|^2 \frac{P^{XY}_{i,e} P^{YZ}_{e,k}}{(\mu_1^Y)_e} \frac{Q^{XY}_{j,o} Q^{YZ}_{o,l}}{(\mu_2^Y)_o} \tag{8.148}$$

$$= \alpha \sum_{i,j,e,o} |C^x_{i,j} - C^y_{e,o}|^2 P^{XY}_{i,e} P^{XY}_{j,o} + (1-\alpha) \sum_{i,j,e,o} |X_{i,j} - Y_{e,o}|^2 P^{XY}_{i,e} Q^{XY}_{j,o} \tag{8.149}$$

$$+ \alpha \sum_{k,l,e,o} |C^y_{e,o} - C^z_{k,l}|^2 P^{YZ}_{e,k} P^{YZ}_{o,l} + (1-\alpha) \sum_{k,l,e,o} |Y_{e,o} - Z_{k,l}|^2 P^{YZ}_{e,k} Q^{YZ}_{o,l} \tag{8.150}$$

$$= \mathrm{AGW}_\alpha(\mathcal{X}, \mathcal{Y}) + \mathrm{AGW}_\alpha(\mathcal{Y}, \mathcal{Z}). \tag{8.151}$$

where the second inequality follows from the inequality: $(x+y)^2 \leq 2(x^2+y^2)$. ■

Proof of Corollary 6.2.1. This proof is based on the personal communication with professor Will Sawin on his discussion on https://mathoverflow.net/questions/420319/why-is-the-set-of-hermitian- We thank him for his invaluable support during the submission of our paper.

First, let us recall the Schwartz-Zippel lemma. Denote $F(x_1, ..., x_n)$ a multivariate polynomial. Its total degree is the maximum of the sums of the powers of the variables in any monomial. The Schwartz-Zippel lemma states that: let $F(x_1, ..., x_n)$ be a nonzero multivariate polynomial of total degree d and S be a finite subset of $\mathbb{R}$. Denote $Z_S := \{(x_1, ..., x_n) \in S^n : F(x_1, ..., x_n) = 0\}$ the set of zeros of F on S^n. Then $|Z_S| \leq d|S|^{n-1}$.

Note that, the set of Hermitian matrices of size n forms a finite-dimensional real vector space. In particular, it is isomorphic to the Euclidean space $\mathbb{R}^{n^2}$. Denote I set of Hermitian matrices of size n with repeated eigenvalues. It is enough to show that I has measure zero. We have $I \simeq E$,

for some $E \subset \mathbb{R}^{n^2}$. By Proposition 4 in (Stein and Shakarchi, 2005), since I is closed (see page 56 in (Tao, 2012)), it is measurable. If I does not have zero measure, then the intersection $E \cap [0,1]^{n^2}$ has positive measure $p > 0$. If, for each $i \in [n^2]$, we sample m i.i.d coordinates uniformly in $[0,1]$, then we have m^{n^2} points uniformly distributed in $[0,1]^{n^2}$. So, the expected number of points lying in E is pm^{n^2}.

On the other hand, recall that a (Hermitian) matrix has repeated eigenvalues if and only if the discriminant of its characteristic polynomial is zero. Moreover, the discriminant of the characteristic polynomial is a polynomial in n^2 entries of the matrix. Thus, the measure of I (or, equivalently E) is the measure of the set of values of these n^2 variables which make a certain polynomial of total degree d vanish. By Schwartz-Zippel lemma, on average, there are at most dm^{n^2-1} points in E. By choosing $m > d/p$, we obtain a contradiction. Thus E (or equivalently I) must have zero measure. ■

Proof of Theorem 6.2.1. Regarding the first claim, note that $Y = XQ$, where Q is a permutation matrix corresponding to the permutation σ_c. Since Y is obtained by swapping columns of X, it is easy to see that $\mathrm{GW}(\mathcal{X}, \mathcal{Y}) = 0$ and the optimal plan between X and Y is $P^* = \frac{1}{n^2}\mathrm{Id}_n$. Similarly, $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = 0$, where P^* and $Q^* = \frac{1}{n}Q$ are the optimal sample and feature couplings, respectively. In other words, $\langle L(C^x, C^y) \otimes P^*, P^* \rangle = 0$ and $\langle L(X, Y) \otimes Q^*, P^* \rangle = 0$. We deduce that $\mathrm{AGW}_\alpha(\mathcal{X}, \mathcal{Y}) = 0$.

Now, for $0 < \alpha < 1$, if $\mathrm{AGW}_\alpha(\mathcal{X}, \mathcal{Y}) = 0$, then $\mathrm{GW}(\mathcal{X}, \mathcal{Y}) = \mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = 0$. In particular, X and Y must have the same shape, so $X, Y \in \mathbb{R}^{n \times d}$. As $\mathrm{GW}(\mathcal{X}, \mathcal{Y}) = 0$, there exists an isometry from X to Y. Note that every isometry from $\mathbb{R}^d$ to $\mathbb{R}^d$ is a composition of at most $d+1$ reflections (see, for example, Corollary A.7 in (*Isometries of* $\mathbb{R}^n$)). So, $Y = XO$, for some $O \in \mathcal{O}_d$. As $\mathrm{COOT}(\mathcal{X}, \mathcal{Y}) = 0$, there exist two permutations σ_r and σ_c such that $X_{i,j} = Y_{\sigma_r(i), \sigma_c(j)}$, or equivalently two permutation matrices $P \in \mathcal{P}_n, Q_1 \in \mathcal{P}_d$ such that $Y = PXQ_1$. We deduce that $XO = PXQ_1$, or equivalently $X = PXQ$, for $Q = Q_1 O^T \in \mathcal{O}_d$. We will show that Q is symmetric.

Indeed, consider the singular value decomposition of X, *i.e.*, $X = U\Sigma V^T$, where $U \in \mathbb{R}^{n \times d}$ such that $U^T U = I_d$, $V \in \mathcal{O}_d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose diagonal contains d strictly decreasing singular values (since $n \geq d$). As $X = PXQ$, we have $U\Sigma V^T = (PU)\Sigma(V^T Q)$. For $i \in [d]$, let $u_i \in \mathbb{R}^n$ and $v_i \in \mathbb{R}^d$ be columns of U and V, respectively. As the singular values are positive and distinct, the columns are unique up to the sign change of both columns in U and V. This means $u_i = \pm P u_i$ and $v_i = \pm Q^T v_i$. In other words, ± 1 are eigenvalues of P and Q^T, and u_i, v_i are their corresponding eigenvectors, respectively. Denote $D \in \mathbb{R}^{d \times d}$ any diagonal matrix whose diagonal values are in $\{\pm 1\}$, then $Q^T = VDV^{-1} = VDV^T = Q$. So, Q is symmetric. Theorem 6.2.1 then follows by observing that $O = Q^T Q_1$. ■

Lemma 8.5.1. *COOT is weakly invariant to translation.*

Proof of Lemma 8.5.1. For any $P \in U(\mu_1^X, \mu_1^Y), Q \in U(\mu_2^X, \mu_2^Y)$ and $c \in \mathbb{R}$, we have

$$\sum_{i,j,k,l} (X_{ik} - Y_{jl} - c)^2 P_{ij} Q_{kl} = \sum_{i,j,k,l} (X_{ik} - Y_{jl})^2 P_{ij} Q_{kl} - 2c \sum_{i,j,k,l} (X_{ik} - Y_{jl}) P_{ij} Q_{kl} + c^2. \quad (8.152)$$

Now,

$$\begin{aligned}
\sum_{i,j,k,l} (X_{ik} - Y_{jl}) P_{ij} Q_{kl} &= \sum_{i,j,k,l} X_{ik} P_{ij} Q_{kl} - \sum_{ijkl} Y_{jl} P_{ij} Q_{kl} && (8.153)\\
&= \sum_{i,k} X_{ik} \left(\sum_j P_{ij} \right) \left(\sum_l Q_{kl} \right) - \sum_{j,l} Y_{jl} \left(\sum_i P_{ij} \right) \left(\sum_k Q_{kl} \right) && (8.154)\\
&= \sum_{i,k} X_{ik} (\mu_1^X)_i (\mu_2^X)_j \mu'_k - \sum_{j,l} Y_{jl} (\mu_1^Y)_j (\mu_2^Y)_l && (8.155)\\
&= (\mu_1^X)^T X \mu_2^X - (\mu_1^Y)^T Y \mu_2^Y . && (8.156)
\end{aligned}$$

So, $\mathrm{COOT}(\mathcal{X}, \mathcal{Y} + c) = \mathrm{COOT}(\mathcal{X}, \mathcal{Y}) - 2c \left((\mu_1^X)^T X \mu_2^X - (\mu_1^Y)^T Y \mu_2^Y \right) + c^2$. This implies that COOT is weakly invariant to translation. ■

Proof of Proposition 6.2.2. Note that the GW term in AGW remains unchanged by translation. By adapting the proof of Lemma 8.5.1, we obtain

$$\mathrm{AGW}_\alpha(\mathcal{X}, \mathcal{Y} + c) = \mathrm{AGW}_\alpha(\mathcal{X}, \mathcal{Y}) + (1 - \alpha) \left[c^2 - 2c \left((\mu_1^X)^T X \mu_2^X - (\mu_1^Y)^T Y \mu_2^Y \right) \right]. \quad (8.157)$$

The result then follows. ■

8.5.2 Experimental Set-up Details

MNIST Illustrations

We align 1000 images of hand-written digits from the MNIST dataset with 1000 images from the USPS dataset. Each dataset is subsampled to contain 100 instances of each of the 10 possible digits (0 through 9), using the random seed of 1976. We set all marginal distributions to uniform, and use cosine distances for GW and AGW. We consider both the entropically regularized and non-regularized versions for all methods. For entropic regularization, we sweep a grid of $\varepsilon_1, \varepsilon_2$(=if applicable) $\in [5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1, 5e-1]$. For AGW, we consider $[0.1, 0.2, 0.3, ..., 0.9]$, and present results with the best-performing hyperparameter combination of each method, as measured by the percent accuracy of matching images from the same digit across the two datasets.

Single-cell multi-omic alignment experiments

Datasets We largely follow the first paper that applied OT to single-cell multi-omic alignment task (Demetci et al., 2020) in our experimental set-up and use four simulated datasets and three real-world single-cell multi-omic datasets to benchmark our cell alignment performance.

Three of the simulated datasets have been generated by (Liu et al., 2019b) by non-linearly projecting 600 samples from a common 2-dimensional space onto different 1000- and 2000-dimensional spaces with 300 samples in each.

We include a fourth simulated dataset generated by (Demetci et al., 2020) using a single-cell RNA-seq data simulation package in R, called Splatter (Zappia, Phipson, and Oshlack, 2017). We refer to this dataset as "Synthetic RNA-seq". This dataset includes a simulated gene expression domain with 50 genes and 5000 cells divided across three cell types and another domain created by non-linearly projecting these cells onto a 500-dimensional space. As a result of their generation schemes, all simulated datasets have ground-truth 1-1 cell correspondence information. We use this information solely for benchmarking. We do not have access to ground-truth feature relationships in these datasets, so we exclude them from feature alignment experiments.

Additionally, we include three real-world single-cell sequencing datasets in our experiments. To have ground-truth information on cell correspondences for evaluation, we choose three co-assay datasets which have paired measurements on the same individual cells: an scGEM dataset (Cheow et al., 2016), a SNARE-seq dataset (Chen, Lake, and Zhang, 2019b), and a CITE-seq dataset (Stoeckius et al., 2017) (these are exceptions to the experimental challenge described above). These first two datasets have been used by existing OT-based single-cell alignment methods (Cao et al., 2020; Cao, Hong, and Wan, 2021; Demetci et al., 2020, 2022a; Singh et al., 2020), while the last one was included in the evaluations of a non-OT-based alignment method, bindSC (Dou et al., 2022).

In addition to these three datasets, we include a fourth single-cell dataset, which contains data from the same measurement modality (*i.e.*, gene expression) but from two different species: mouse (Bhattacherjee et al., 2019) and bearded lizard (Tosches et al., 2018). Our motivation behind including this dataset is to demonstrate the effects of both sample-level (*i.e.*, cell-level) and feature-level (*i.e.*, gene-level) supervision on alignment qualities. We refer to this dataset as the "cross-species dataset", which contains 4,187 cells from lizard pallium (a brain region) and 6,296 cells from the mouse prefrontal cortex. The two species share a subset of their features: 10,816 paralogous genes. Each also has species-specific genes: 10,184 in the mouse dataset and 1,563 in the lizard dataset.

Baselines and hyperparameter tuning We benchmark AGW's performance on single-cell alignment tasks against three algorithms: (1) COOT (Redko et al., 2020), (2) SCOT (Demetci et al., 2020), which is a Gromov-Wasserstein OT-based algorithm that uses k-nearest neighbor

(kNN) graph distances on dimensionality reduced datasets (top 30 principal components for gene expression domains and simulated domains, 15-25 topics with latent dirichlet allocation for other measurement domains) as intra-domain distance matrices. This choice of distances has been shown to perform better than Euclidean distances, cosine distances by (Demetci et al., 2020), and bindSC (Dou et al., 2022). For consistency, we keep the intra-domain distance computations the same for AGW and UGW, too. Among all baselines, bindSC is not an OT-based algorithm: It employs bi-order canonical correlation analysis to perform alignment. We include it as a benchmark as it is the only existing single-cell alignment algorithm that can perform feature alignments (in addition to cell alignments) for a few limited types of measurement modalities.

When methods share similar hyperparameters in their formulation (*e.g.*, entropic regularization constant, ε for methods that employ OT), we use the same hyperparameter grid to perform their tuning. Otherwise, we refer to the publication and the code repository for each method to choose a hyperparameter range. For SCOT, we tune four hyperparameters: $k \in \{20, 30, \ldots, 150\}$, the number of neighbors in the cell neighborhood graphs, $\varepsilon \in \{5e{-}4, 3e{-}4, 1e{-}4, 7e{-}3, 5e{-}3, \ldots, 1e{-}2\}$, the entropic regularization coefficient for the optimal transport formulation. Similarly, for both COOT and AGW, we sweep $\varepsilon_1, \varepsilon_2 \in \{5e-4, 3e-4, 1e-4, 7e-3, 5e-3, \ldots, 1e-2\}$ for the coefficients of entropic regularization over the sample and feature alignments. We use the same intra-domain distance matrices in AGW as in SCOT (based on kNN graphs). For all OT-based methods, we perform barycentric projection to complete the alignment.

For bindSC, we choose the coupling coefficient that assigns weight to the initial gene activity matrix $\alpha \in \{0, 0.1, 0.2, \ldots 0.9\}$ and the coupling coefficient that assigns a weight factor to multi-objective function $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$. Additionally, we choose the number of canonical vectors for the embedding space $K \in \{3, 4, 5, 10, 30, 32\}$. For all methods, we report results with the best-performing hyperparameter combinations.

Evaluation Metrics When evaluating cell alignments, we use a metric previously used by other single-cell multi-omic integration tools (Cao et al., 2020; Cao, Hong, and Wan, 2021; Demetci et al., 2020, 2022a; Dou et al., 2022; Liu et al., 2019b; Singh et al., 2020) called "fraction of samples closer than the true match" (FOSCTTM). For this metric, we compute the Euclidean distances between a fixed sample point and all the data points in the other domain. Then, we use these distances to compute the fraction of samples that are closer to the fixed sample than its true match and then average these values for all the samples in both domains. This metric measures alignment error, so the lower values correspond to higher-quality alignments.

We investigate the accuracy of feature correspondences recovered to assess feature alignment performance. We mainly use two real-world datasets for this task - CITE-seq, and the cross-species scRNA-seq datasets (results on SNARE-seq and scGEM datasets are qualitatively evaluated due to the lack of ground-truth information). For the CITE-seq dataset, we expect the feature

correspondences to recover the relationship between the 25 antibodies and the genes that encode them. To investigate this, we simultaneously align the cells and features of the two modalities using the 25 antibodies and 25 genes in an unsupervised manner. We compute the percentage of 25 antibodies whose strongest correspondence is their encoding gene.

For the cross-species RNA-seq dataset, we expect alignments between (1) the cell-type annotations common to the mouse and lizard datasets, namely excitatory neurons, inhibitory neurons, microglia, OPC (Oligodendrocyte precursor cells), oligodendrocytes, and endothelial cells and (2) between the paralogous genes. For this dataset, we generate cell-label matches by averaging the rows and columns of the cell-cell alignment matrix yielded by AGW based on these cell annotation labels. We compute the percentage of these six cell-type groups that match as their strongest correspondence. For feature alignments, we compute the percentage of the 10,816 shared genes that are assigned to their corresponding paralogous gene with their highest alignment probability. For this dataset, we consider providing supervision at increasing levels on both sample and feature alignments. For feature-level supervision, 20% supervision means setting the alignment cost of $\sim 20\%$ of the genes with their paralogous pairs to 0. For sample-level supervision, 20% supervision corresponds to downscaling the alignment cost of $\sim 20\%$ of the mouse cells from the aforementioned seven cell types with the $\sim 20\%$ of lizard cells from their corresponding cell-type by $\frac{1}{\#\text{ lizard cells in the same cell-type}}$.

Bibliography

Abid, Brahim Khalil and Robert Gower (2018), « Stochastic algorithms for entropy-regularized optimal transport problems », *in*: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, vol. 84, Proceedings of Machine Learning Research, PMLR, page(s): 1505–1512 (Cited on page 22).

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux (2014), « Machine learning for neuroimaging with scikit-learn », *in*: *Frontiers in Neuroinformatics* 8, ISSN: 1662-5196, URL: https://www.frontiersin.org/article/10.3389/fninf.2014.00014 (visited on 05/19/2022) (Cited on page 85).

Abraham, Romain, Jean-François Delmas, and Patrick Hoscheit (2013), « A note on the Gromov-Hausdorff-Prokhorov distance between (locally) compact metric measure spaces », *in*: *Electronic Journal of Probability* 18, page(s): 1–21 (Cited on page 38).

Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim (2001), « On the Surprising Behavior of Distance Metrics in High Dimensional Space », *in*: *Database Theory — ICDT 2001*, Springer Berlin Heidelberg, page(s): 420–434 (Cited on pages 49, 79).

Agueh, Martial and Guillaume Carlier (2011), « Barycenters in the Wasserstein Space », *in*: *SIAM Journal on Mathematical Analysis* 43 (2), page(s): 904–924 (Cited on page 55).

Al-Wasity, S., S. Vogt, A. Vuckovic, and F. E. Pollick (Mar. 2020), « Hyperalignment of motor cortical areas based on motor imagery during action observation », *in*: *Sci Rep* 10.*1*, page(s): 5362 (Cited on page 86).

Alaya, Mokhtar Z., Maxime Berar, Gilles Gasso, and Alain Rakotomamonjy (2019), « Screening Sinkhorn Algorithm for Regularized Optimal Transport », *in*: *Advances in Neural Information Processing Systems 32*, page(s): 12169–12179 (Cited on page 22).

Alaya, Mokhtar Z., Maxime Bérar, Gilles Gasso, and Alain Rakotomamonjy (2022), « Theoretical guarantees for bridging metric measure embedding and optimal transport », *in*: *Neurocomputing* 468, page(s): 416–430 (Cited on page 35).

Altschuler, Jason, Jonathan Weed, and Philippe Rigollet (2017), « Near-Linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration », *in*: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page(s): 1961–1971 (Cited on pages 21, 22).

Alvarez-Melis, David and Nicolo Fusi (2020), « Geometric Dataset Distances via Optimal Transport », *in*: *Advances in Neural Information Processing Systems*, vol. 33, page(s): 21428–21439 (Cited on page 46).

Alvarez-Melis, David and Tommi S. Jaakkola (2018), « Gromov-Wasserstein Alignment of Word Embedding Spaces », *in*: *Empirical Methods in Natural Language Processing (EMNLP)*, page(s): 1881–1890 (Cited on pages 32, 68).

Alvarez-Melis, David, Stefanie Jegelka, and Tommi S. Jaakkola (2019), « Towards Optimal Transport with Global Invariances », *in*: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, vol. 89, Proceedings of Machine Learning Research, PMLR, page(s): 1870–1879 (Cited on page 38).

Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savaré (2005), *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics. ETH Zürich, Birkhäuser Basel (Cited on pages 17, 122).

Amos, Brandon, Lei Xu, and J. Zico Kolter (2017), « Input Convex Neural Networks », *in*: *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, Proceedings of Machine Learning Research, page(s): 146–155 (Cited on page 18).

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017), « Wasserstein Generative Adversarial Networks », *in*: *International Conference on Machine Learning* 70, page(s): 214–223 (Cited on pages 7, 8, 18, 67, 100).

Avants, B.B., C.L. Epstein, M. Grossman, and J.C. Gee (2008), « Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain », *in*: *Medical Image Analysis* 12.*1*, page(s): 26–41, ISSN: 1361-8415, DOI: 10.1016/j.media.2007.06.004, URL: http://www.sciencedirect.com/science/article/pii/S1361841507000606 (Cited on pages 85, 86).

Balaji, Yogesh, Rama Chellappa, and Soheil Feizi (2020), « Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation », *in*: *Advances in Neural Information Processing Systems* (Cited on pages 23, 47, 69).

Bao, Han and Shinsaku Sakaue (2022), « Sparse Regularized Optimal Transport with Deformed q-Entropy », *in*: *Entropy* 24.*11* (Cited on page 22).

Bazeille, Thomas, Elizabeth DuPre, Hugo Richard, Jean-Baptiste Poline, and Bertrand Thirion (Dec. 15, 2021), « An empirical evaluation of functional alignment using inter-subject decoding », *in*: *NeuroImage* 245, page(s): 118683, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2021.118683 (Cited on page 86).

Bazeille, Thomas, Hugo Richard, Hicham Janati, and Bertrand Thirion (2019a), « Local optimal transport for functional brain template estimation », *in*: *International Conference on Information Processing in Medical Imaging*, Springer, page(s): 237–248 (Cited on pages 23, 69).

— (June 2019b), « Local Optimal Transport for Functional Brain Template Estimation », *in*: *IPMI 2019 - 26th International Conference on Information Processing in Medical Imaging*, Hong Kong, China, DOI: 10.1007/978-3-030-20351-1_18 (Cited on page 86).

Benamou, Jean-David (2003), « Numerical resolution of an "unbalanced" mass transport problem », *in*: *ESAIM: Mathematical Modelling and Numerical Analysis* 37 (5), page(s): 851–868 (Cited on pages 8, 23, 69).

Benamou, Jean-David, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré (2014), « Iterative Bregman Projections for Regularized Transportation Problems », *in*: *SIAM Journal on Scientific Computing* 37 (2), page(s): 1111–1138 (Cited on pages 56, 130).

Bhattacherjee, Aritra, Mohamed Nadhir Djekidel, Renchao Chen, Wenqiang Chen, Luis M. Tuesta, and Yi Zhang (2019), « Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction », *in*: *Nature Communications* 10.*1*, page(s): 4169, ISSN: 2041-1723, DOI: 10.1038/s41467-019-12054-3, URL: https://doi.org/10.1038/s41467-019-12054-3 (Cited on pages 109, 150).

Billingsley, Patrick (1999), *Convergence of probability measures*, 2nd ed., Wiley Series in Probability, Statistics: Probability, and Statistics (Cited on pages 122, 136).

Birkhoff, George David (1946), « Tres observaciones sobre el algebra lineal », *in*: *Universidad Nacional de Tucuman, Revista* 5, page(s): 147–150 (Cited on pages 15, 59).

Blondel, Mathieu, Vivien Seguy, and Antoine Rolet (2018), « Smooth and Sparse Optimal Transport », *in*: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, vol. 84, Proceedings of Machine Learning Research, page(s): 880–889 (Cited on pages 22, 31).

Bonneel, Nicolas and Julie Digne (2023), « A survey of Optimal Transport for Computer Graphics and Computer Vision. », *in*: *Computer Graphics Forum* 42, page(s): 439–460, URL: https://doi.org/10.1111/cgf.14778 (Cited on pages 8, 19).

Bonneel, Nicolas, Gabriel Peyré, and Marco Cuturi (2016), « Wasserstein barycentric coordinates: histogram regression using optimal transport », *in*: *ACM Transactions on Graphics* 35.*4*, ISSN: 0730-0301, DOI: 10.1145/2897824.2925918, URL: https://doi.org/10.1145/2897824.2925918 (Cited on page 8).

Bonneel, Nicolas, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister (2015), « Sliced and Radon Wasserstein Barycenters of Measures », *in*: *Journal of Mathematical Imaging and Vision volume* 51 (1), page(s): 22–45 (Cited on pages 19, 50).

Bourgain, Jean (1985), « On Lipschitz embedding of finite metric spaces in Hilbert space », *in*: *Israel Journal of Mathematics* (52), page(s): 46–52 (Cited on page 34).

Brenier, Yann (1987), « Décomposition polaire et réarrangement monotone des champs de vecteurs », *in*: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 305, page(s): 805–808 (Cited on pages 15, 18).

Brualdi, Richard A. (2006), *Combinatorial Matrix Classes*, Encyclopedia of Mathematics and its Applications, Cambridge University Press (Cited on page 16).

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf (2015), « ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide », *in*: *Current Protocols in Molecular Biology* 109.*1*, page(s): 21.29.1–21.29.9, DOI: https://doi.org/10.1002/0471142727.mb2129s109, eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb2129s109, URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb2129s109 (Cited on page 79).

Bunne, Charlotte, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka (2019), « Learning Generative Models across Incomparable Spaces », *in*: 97 (Cited on pages 43, 100).

Burago, Dmitri, Yuri Burago, and Sergei Ivanov (2001), *A Course in Metric Geometry*, vol. 33, Graduate Studies in Mathematics, American Mathematical Society (Cited on pages 15, 33, 35, 68).

Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyou Zhu (1995), « A Limited Memory Algorithm for Bound Constrained Optimization », *in*: *SIAM Journal on Scientific Computing* 16.*5*, page(s): 1190–1208 (Cited on page 25).

Caffarelli, Luis and Robert J. McCann (2010), « Free boundaries in optimal transport and Monge-Ampère obstacle problems », *in*: *Annals of Mathematics* 171 (2), page(s): 673–730 (Cited on page 47).

Cai, Yuhang and Lek-Heng Lim (2022), « Distances Between Probability Distributions of Different Dimensions », *in*: *IEEE Transactions on Information Theory* 68.*6*, page(s): 4020–4031 (Cited on page 35).

Cang, Zixuan and Qing Nie (2020), « Inferring spatial and signaling relationships between cells from single cell transcriptomic data », *in*: *Nature communications* 11.*1*, page(s): 1–13 (Cited on page 100).

Cao, Jiezhang, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan (2019), « Multi-marginal Wasserstein GAN », *in*: *Advances in Neural Information Processing Systems*, page(s): 1774–1784 (Cited on pages 55, 57).

Cao, Kai, Xiangqi Bai, Yiguang Hong, and Lin Wan (2020), « Unsupervised topological alignment for single-cell multi-omics integration », *in*: *Bioinformatics* 36.*Supplement_1*, page(s): i48–i56 (Cited on pages 150, 151).

Cao, Kai, Qiyu Gong, Yiguang Hong, and Lin Wan (2022), « A unified computational framework for single-cell data integration with optimal transport », *in*: *Nature Communications* 13.*1*, page(s): 7419 (Cited on pages 100, 104, 108).

Cao, Kai, Yiguang Hong, and Lin Wan (Aug. 2021), « Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona », *in*: *Bioinformatics*, btab594, ISSN:

1367-4803, DOI: 10.1093/bioinformatics/btab594, eprint: https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btab594/40081868/btab594.pdf, URL: https://doi.org/10.1093/bioinformatics/btab594 (Cited on pages 79, 80, 100, 104, 108, 150, 151).

Carlier, Guillaume, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer (2017), « Convergence of Entropic Schemes for Optimal Transport and Gradient Flows », *in*: *arXiv preprint arXiv:1512.02783* (Cited on pages 10, 20, 54, 128).

Chapel, Laetitia, Mokhtar Z. Alaya, and Gilles Gasso (2020), « Partial Optimal Tranport with applications on Positive-Unlabeled Learning », *in*: *Advances in Neural Information Processing Systems* (Cited on pages 47, 98).

Chapel, Laetitia, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso (2021), « Unbalanced Optimal Transport through Non-negative Penalized Linear Regression », *in*: *Neural Information Processing Systems (NeurIPS)* (Cited on pages 29–31, 75, 113, 116, 117).

Chartrand, Rick and Brendt Wohlberg (2009), « A Gradient Descent Solution to the Monge-Kantorovich Problem », *in*: *Applied Mathematical Sciences* 3.*22*, page(s): 1071–1080 (Cited on page 18).

Chen, Po-Hsuan (Cameron), Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge (2015), « A Reduced-Dimension fMRI Shared Response Model », *in*: *Advances in Neural Information Processing Systems*, ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, vol. 28, Curran Associates, Inc. (Cited on page 86).

Chen, Song, Blue B. Lake, and Kun Zhang (2019a), « High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell », *in*: *Nature Biotechnology* 37.*12*, page(s): 1452–1457, DOI: 10.1038/s41587-019-0290-0, URL: https://doi.org/10.1038/s41587-019-0290-0 (Cited on page 79).

— (2019b), « High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell », *in*: *Nature Biotechnology* 37.*12*, page(s): 1452–1457, DOI: 10.1038/s41587-019-0290-0, URL: https://doi.org/10.1038/s41587-019-0290-0 (Cited on pages 104, 150).

Cheow, Lih Feng, Elise T Courtois, Yuliana Tan, Ramya Viswanathan, Qiaorui Xing, Rui Zhen Tan, Daniel S Q Tan, Paul Robson, Loh Yuin-Han, Stephen R Quake, and William F Burkholder (2016), « Single-cell multimodal profiling reveals cellular epigenetic heterogeneity », *in*: *Nature Methods* 13.*10*, page(s): 833–836 (Cited on page 150).

Chizat, Lénaïc, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré (2020), « Faster Wasserstein Distance Estimation with the Sinkhorn Divergence », *in*: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Curran Associates Inc., ISBN: 9781713829546 (Cited on page 20).

Chizat, Lénaïc (2017), « Unbalanced Optimal Transport: Models, Numerical Methods, Applications », PhD thesis, PSL Research University (Cited on page 71).

Chizat, Lénaïc, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard (2018a), « Scaling algorithms for unbalanced optimal transport problems », *in*: *Mathematics of Computation* 87, page(s): 2563–2609 (Cited on pages 21, 23, 25, 27, 31, 69, 75).

— (2018b), « Unbalanced optimal transport: Dynamic and Kantorovich formulations », *in*: *Journal of Functional Analysis* 274.*11*, page(s): 3090–3123 (Cited on pages 23, 24, 31, 69, 143).

Chowdhury, Samir and Facundo Mémoli (2019), « The Gromov-Wasserstein distance between networks and stable network invariants », *in*: *Information and Inference: A Journal of the IMA* 8 (4), page(s): 757–787 (Cited on pages 10, 37–40, 51, 52, 122).

Chowdhury, Samir and Tom Needham (2021), « Generalized Spectral Clustering via Gromov-Wasserstein Learning », *in*: Proceedings of Machine Learning Research 130 (Cited on page 32).

Chowdhury, Samir, Tom Needham, Ethan Semrad, Bei Wang, and Youjia Zhou (2023), « Hypergraph Co-Optimal Transport: Metric and Categorical Properties », *in*: *Journal of Applied and Computational Topology* 40 (Cited on pages 10, 51–53, 68, 70, 71, 125).

Chuang, Ching-Yao, Stefanie Jegelka, and David Alvarez-Melis (2023), « InfoOT: Information Maximizing Optimal Transport », *in*: *International Conference on Machine Learning*, PMLR (Cited on pages 9, 46).

Cohen, Joel E. and Uriel G. Rothblum (1993), « Nonnegative ranks, decompositions, and factorizations of nonnegative matrices », *in*: *Linear Algebra and its Applications* 190, page(s): 149–168 (Cited on page 132).

Cohen, Samuel, K. S. Sesh Kumar, and Marc Peter Deisenroth (2021), « Sliced Multi-Marginal Optimal Transport », *in*: *ICML* (Cited on page 57).

Cominetti, R. and J. San Martín (1994), « Asymptotic analysis of the exponential penalty trajectory in linear programming », *in*: *Mathematical Programming* 67, page(s): 169–187 (Cited on page 20).

Courty, Nicolas, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy (2017), « Joint distribution optimal transportation for domain adaptation », *in*: *Neural Information Processing Systems* (Cited on page 46).

Courty, Nicolas, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy (2016), « Optimal transport for domain adaptation », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, page(s): 1853–1865 (Cited on pages 8, 9, 22, 67).

Csiszár, Imre (1963), « Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten », *in*: *Magyar Tud. Akad. Mat. Kutató Int. Közl* 8, page(s): 85–108 (Cited on pages 23, 71).

Cuturi, Marco (2013), « Sinkhorn Distances: Lightspeed Computation of Optimal Transport », *in*: *Advances in Neural Information Processing Systems*, page(s): 2292–2300 (Cited on pages 19, 55, 104).

Cuturi, Marco, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul (2022), « Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein », *in*: *arXiv preprint arXiv:2201.12324* (Cited on page 20).

Dale, Anders M., Bruce Fischl, and Martin I. Sereno (1999), « Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction », *in*: 9.*2*, page(s): 179–194, ISSN: 1053-8119, DOI: 10.1006/nimg.1998.0395, URL: http://www.sciencedirect.com/science/article/pii/S1053811998903950 (Cited on page 85).

Demetci, Pinar, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh (2020), « Gromov-Wasserstein optimal transport to align single-cell multi-omics data », *in*: *bioRxiv* (Cited on pages 32, 100, 104, 108, 150, 151).

— (2022a), « SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport », *in*: *Journal of computational biology* 29.*1*, page(s): 3–18 (Cited on pages 68, 79, 80, 100, 104, 108, 150, 151).

Demetci, Pinar, Rebecca Santorella, Björn Sandstede, and Ritambhara Singh (2021), « Unsupervised integration of single-cell multi-omics datasets with disparities in cell-type representation », *in*: *bioRxiv*, DOI: 10.1101/2021.11.09.467903, eprint: https://www.biorxiv.org/content/early/2021/11/11/2021.11.09.467903.full.pdf, URL: https://www.biorxiv.org/content/early/2021/11/11/2021.11.09.467903 (Cited on pages 79, 82).

Demetci, Pinar, Quang Huy Tran, Ievgen Redko, and Ritambhara Singh (2022b), « Jointly aligning cells and genomic features of single-cell multi-omics data with co-optimal transport », *in*: *bioRxiv* (Cited on page 49).

— (2022c), « Jointly aligning cells and genomic features of single-cell multi-omics data with co-optimal transport », *in*: *bioRxiv*, DOI: 10.1101/2022.11.09.515883, eprint: https://www.biorxiv.org/content/early/2022/12/11/2022.11.09.515883.full.pdf, URL: https://www.biorxiv.org/content/early/2022/12/11/2022.11.09.515883.

— (2024), « Breaking isometric ties and introducing priors in Gromov-Wasserstein distances », *in*: *The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)* (Cited on pages 12, 99).

Dessein, Arnaud, Nicolas Papadakis, and Jean-Luc Rouas (2016), « Regularized Optimal Transport and the Rot Mover's Distance », *in*: *Journal of Machine Learning Research* 19, page(s): 1–53 (Cited on page 22).

Dominique, Lecomte (2020), *Cours de Théorie Descriptive des Ensembles*, URL: https://webusers.imj-prg.fr/~dominique.lecomte/Chapitres/2-Polish%20spaces.pdf (Cited on page 121).

Dou, Jinzhuang, Shaoheng Liang, Vakul Mohanty, Qi Miao, Yuefan Huang, Qingnan Liang, Xuesen Cheng, Sangbae Kim, Jongsu Choi, Yumei Li, Li Li, May Daher, Rafet Basar, Katayoun Rezvani, Rui Chen, and Ken Chen (2022), « Bi-order multimodal integration of single-cell data », *in*: *Genome Biology* 23.*1*, page(s): 112, DOI: 10.1186/s13059-022-02679-x, URL: https://doi.org/10.1186/s13059-022-02679-x (Cited on pages 150, 151).

Dvurechensky, Pavel, Alexander Gasnikov, and Alexey Kroshnin (2018), « Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm », *in*: *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, Proceedings of Machine Learning Research, PMLR, page(s): 1367–1376 (Cited on page 22).

Edelsbrunner, Herbert, David Letscher, and Afra Zomorodian (2002), « Topological Persistence and Simplification », *in*: *Discrete Computational Geometry* 28, page(s): 511–533 (Cited on page 7).

Eichert, Nicole, Emma C Robinson, Katherine L Bryant, Saad Jbabdi, Mark Jenkinson, Longchuan Li, Kristine Krug, Kate E Watkins, and Rogier B Mars (Mar. 23, 2020), « Cross-species cortical alignment identifies different types of anatomical reorganization in the primate temporal lobe », *in*: *eLife* 9, ed. by Timothy Verstynen, Joshua I Gold, Timothy Verstynen, and Katja Heuer, Publisher: eLife Sciences Publications, Ltd, page(s): e53232, ISSN: 2050-084X, DOI: 10.7554/eLife.53232, URL: https://doi.org/10.7554/eLife.53232 (Cited on page 86).

Essid, Montacer and Justin Solomon (2018), « Quadratically Regularized Optimal Transport on Graphs », *in*: *SIAM Journal on Scientific Computing* 40.*4*, page(s): A1961–A1986.

Fatras, Kilian, Thibault Séjourné, Nicolas Courty, and Rémi Flamary (2021), « Unbalanced minibatch Optimal Transport; applications to Domain Adaptation », *in*: *International Conference on Machine Learning* (Cited on pages 23, 24, 69, 73, 138).

Fatras, Kilian, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty (2020), « Learning with minibatch Wasserstein: asymptotic and gradient properties », *in*: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, vol. 108, page(s): 2131–2141 (Cited on page 19).

Ferradans, Sira, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol (2014), « Regularized Discrete Optimal Transport », *in*: *SIAM Journal on Imaging Sciences* 7.*3*, page(s): 1853–1882 (Cited on page 8).

Feydy, Jean (2020), « Geometric data analysis, beyond convolutions », PhD thesis, Université Paris Saclay (Cited on page 21).

Feydy, Jean, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré (2019), « Interpolating between Optimal Transport and MMD using Sinkhorn Divergences », *in*: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)* (Cited on pages 19–21).

Figalli, Alessio (2010), « The optimal partial transport problem », *in*: *Archive for rational mechanics and analysis* 109 (2), page(s): 533–560 (Cited on page 47).

Fischl, Bruce (Aug. 15, 2012), « FreeSurfer », *in*: *NeuroImage*, 20 YEARS OF fMRI 62.*2*, page(s): 774–781, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2012.01.021, URL: https://www.sciencedirect.com/science/article/pii/S1053811912000389 (Cited on page 85).

Flamary, Rémi, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer (2021), « POT: Python Optimal Transport », *in*: *Journal of Machine Learning Research* 22.*78*, page(s): 1–8, URL: http://jmlr.org/papers/v22/20-451.html (Cited on pages 11, 20, 22, 64, 104).

Folland, Gerald B (1999), *Real analysis: modern techniques and their applications*, John Wiley & Sons (Cited on page 122).

Forrow, Aden, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed (2019), « Statistical Optimal Transport via Factored Couplings », *in*: *The 22nd International Conference on Artificial Intelligence and Statistics*, page(s): 2454–2465 (Cited on page 55).

Frank, Marguerite and Philip Wolfe (1956), « An algorithm for quadratic programming », *in*: *Naval Research Logistics Quarterly* 3.*1-2*, page(s): 95–110 (Cited on pages 42, 63).

Franklin, Joel and Jens Lorenz (1989), « On the scaling of multidimensional matrices », *in*: *Linear Algebra and its Applications* 114, page(s): 717–735 (Cited on page 20).

Frogner, Charlie, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio (2015), « Learning with a Wasserstein loss », *in*: *Advances in Neural Information Processing Systems*, page(s): 2053–2061 (Cited on pages 8, 25, 31, 67, 71).

Gangbo, Wilfrid and Andrzej Swiech (1998), « Optimal maps for the multidimensional Monge-Kantorovich problem », *in*: *Communications on Pure and Applied Mathematics* 51 (1), page(s): 23–45 (Cited on page 55).

Genevay, Aude, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré (2019), « Sample Complexity of Sinkhorn Divergences », *in*: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* 89, page(s): 1574–1583 (Cited on pages 20, 54, 55, 114, 128, 129).

Genevay, Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach (2016), « Stochastic optimization for large-scale optimal transport », *in*: *Advances in Neural Information Processing Systems* 25, page(s): 3440–3448 (Cited on page 22).

Genevay, Aude, Gabriel Peyre, and Marco Cuturi (2018), « Learning generative models with Sinkhorn divergences », *in*: *International Conference on Artificial Intelligence and Statistics* 84, page(s): 1608–1617 (Cited on page 18).

Gibbs, Alison L. and Francis Edward Su (2002), « On Choosing and Bounding Probability Metrics », *in*: *International Statistical Review / Revue Internationale de Statistique* 70.*3*, page(s): 419–435 (Cited on page 17).

Glasser, Matthew F, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, Stephen M Smith, and David C Van Essen (2016), « A multi-modal parcellation of human cerebral cortex », en, *in*: *Nature* 536.*7615*, page(s): 171–178 (Cited on pages 85, 86).

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014), « Generative adversarial nets », *in*: *Advances in Neural Information Processing Systems*, page(s): 2672–2680 (Cited on page 18).

Gramfort, Alexandre, Gabriel Peyré, and Marco Cuturi (2015), *Fast Optimal Transport Averaging of Neuroimaging Data*, DOI: 10.48550/ARXIV.1503.08596 (Cited on page 86).

Gromov, Mikhael (1981), « Groups of polynomial growth and expanding maps (with an appendix by Jacques Tits) », *in*: *Publications Mathématiques de l'IHES* 53, page(s): 53–78 (Cited on pages 33, 68).

Gromov, Misha (1999), *Metric Structures for Riemannian and Non-Riemannian Spaces*, vol. 152, Progress in Mathematics, Birkhäuser, Boston, US (Cited on pages 33, 34, 68).

Gu, Xiang, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu (2022), « Keypoint-Guided Optimal Transport with Applications in Heterogeneous Domain Adaptation », *in*: *Advances in Neural Information Processing Systems* (Cited on pages 9, 46).

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville (2017), « Improved training of Wasserstein GANs », *in*: *Advances in Neural Information Processing Systems*, page(s): 5769–5779 (Cited on page 18).

Guntupalli, J. Swaroop, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby (June 1, 2016), « A Model of Representational Spaces in Human Cortex », *in*: *Cerebral Cortex* 26.*6*, Publisher: Oxford Academic, page(s): 2919–2934, ISSN: 1047-3211, DOI: 10.1093/cercor/bhw068 (Cited on page 86).

Guo, Xin, Johnny Hong, Tianyi Lin, and Nan Yang (2021), « Relaxed Wasserstein with Applications to GANs », *in*: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page(s): 3325–3329 (Cited on page 17).

Haasler, Isabel, Rahul Singh, Qinsheng Zhang, Johan Karlsson, and Yongxin Chen (2020), « Multi-marginal optimal transport and probabilistic graphical models », *in*: *arXiv preprint arXiv:2006.14113* (Cited on page 55).

Hao, Yuhan, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija (2021), « Integrated analysis of multimodal single-cell data », *in*: *Cell*, DOI: 10.1016/j.cell.2021.04.048, URL: https://doi.org/10.1016/j.cell.2021.04.048 (Cited on page 79).

Haxby, James V., J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge (Oct. 20, 2011), « A common, high-dimensional model of the representational space in human ventral temporal cortex », *in*: *Neuron* 72.*2*, page(s): 404–416, ISSN: 1097-4199, DOI: 10.1016/j.neuron.2011.08.026 (Cited on page 86).

He, Zhenliang, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen (2019), « AttGAN: Facial Attribute Editing by Only Changing What You Want », *in*: *IEEE Trans. Image Process.* 28.*11*, page(s): 5464–5478 (Cited on page 55).

Hui, Le, Xiang Li, Jiaxin Chen, Hongliang He, and Jian Yang (2018), « Unsupervised Multi-Domain Image Translation with Domain-Specific Encoders/Decoders », *in*: *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*, IEEE Computer Society, page(s): 2044–2049 (Cited on page 55).

Huizing, Geert-Jan, Laura Cantini, and Gabriel Peyré (2022), « Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors », *in*: *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, Proceedings of Machine Learning Research, PMLR, page(s): 9429–9443 (Cited on page 17).

Hull, J.J. (1994), « A database for handwritten text recognition research », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.*5*, page(s): 550–554, DOI: 10.1109/34.291440 (Cited on page 103).

Hunter, David R and Kenneth Lange (2004), « A Tutorial on MM Algorithms », *in*: *The American Statistician* 58.*1*, page(s): 30–37 (Cited on page 29).

Hur, YoonHaeng, Wenxuan Guo, and Tengyuan Liang (2021), « Reversible Gromov-Monge Sampler for Simulation-Based Inference », *in*: *arXiv preprint arXiv:2109.14090* (Cited on pages 37, 38).

Jaggi, Martin (2013), « Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization », *in*: *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, Proceedings of Machine Learning Research 1, PMLR, page(s): 427–435 (Cited on page 43).

Janati, Hicham, Thomas Bazeille, Bertrand Thirion, Marco Cuturi, and Alexandre Gramfort (2019), « Group level MEG/EEG source imaging via optimal transport: minimum Wasserstein estimates », *in*: *International Conference on Information Processing in Medical Imaging*, Springer, page(s): 743–754 (Cited on pages 23, 69).

Jia, Yangqing, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell (2014), « Caffe: Convolutional Architecture for Fast Feature Embedding », *in*: *Proceedings of the 22nd ACM International Conference on Multimedia*, page(s): 675–678 (Cited on pages 77, 110).

Jiao, Jiantao, Thomas A. Courtade, Albert No, Kartik Venkat, and Tsachy Weissman (2014), « Information Measures: The Curious Case of the Binary Alphabet », *in*: *IEEE Transactions on Information Theory* 60.*12*, page(s): 7616–7626, DOI: 10.1109/TIT.2014.2360184 (Cited on page 26).

Johnson, William B. and Joram Lindenstrauss (1984), « Extensions of Lipschitz mappings into Hilbert space », *in*: *Contemporary mathematics* (26), page(s): 186–206 (Cited on page 34).

Jolicoeur-Martineau, Alexia and Ioannis Mitliagkas (2019), « Connections between Support Vector Machines, Wasserstein distance and gradient-penalty GANs », *in*: *arXiv preprint arXiv:1910.06922* (Cited on page 18).

Kalton, Nigel J and Mikhail I Ostrovskii (1999), « Distances between Banach spaces », *in*: *Forum Mathematicum* 11.*1*, page(s): 17–48 (Cited on pages 36, 68).

Kantorovich, Leonid (1942), « On the transfer of masses (in Russian) », *in*: *Doklady Akademii Nauk* 37, page(s): 227–229 (Cited on pages 7, 15, 17, 67).

Kawaguchi, Kenji (2016), « Deep Learning without Poor Local Minima », *in*: *Advances in Neural Information Processing Systems*, vol. 29 (Cited on page 105).

Kellerer, Hans G (1984), « Duality theorems for marginal problems », *in*: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 67 (4), page(s): 399–432 (Cited on page 55).

Kerdoncuff, Tanguy, Rémi Emonet, Michaël Perrot, and Marc Sebban (2022), « Optimal Tensor Transport », *in*: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, page(s): 7124–7132 (Cited on page 49).

Kim, Taeksoo, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim (2017), « Learning to discover cross-domain relations with generative adversarial networks », *in*: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, Sydney, NSW, Australia: JMLR.org, page(s): 1857–1865 (Cited on page 37).

Koehl, Patrice, Marc Delarue, and Henri Orland (2019), « Optimal transport at finite temperature », *in*: *Physical Review E* 100 (1), page(s): 013310 (Cited on pages 19, 45).

— (2023), « Computing the Gromov-Wasserstein Distance between Two Surface Meshes Using Optimal Transport », *in*: *Algorithms* 16.*3* (Cited on page 45).

Kong, Lemin, Jiajin Li, Jianheng Tang, and Anthony Man-Cho So (2023), « Outlier-Robust Gromov-Wasserstein for Graph Data », *in*: *Advances in Neural Information Processing Systems* 37 (Cited on page 47).

Konno, Hiroshi (1976), « Maximization of a convex quadratic function under linear constraints », *in*: *Mathematical Programming* 11 (1), page(s): 117–127 (Cited on page 50).

Konrad, Keith, *Isometries of* $\mathbb{R}^n$, https://kconrad.math.uconn.edu/blurbs/grouptheory/isometryRn.pdf, Accessed: 2023-05-01 (Cited on page 148).

Koopmans, Tjalling C. and Martin Beckmann (1957), « Assignment Problems and the Location of Economic Activities », *in*: *Econometrica* 25 (1), page(s): 53–76 (Cited on page 41).

Korotin, Alexander, Vage Egiazarian, Arip Asadulaev, and Evgeny Burnaev (2019), « Wasserstein-2 Generative Networks », *in*: *arXiv preprint arXiv:1909.13082* (Cited on page 18).

Kostic, Vladimir R, Saverio Salzo, and Massimiliano Pontil (2021), « Linear Convergence of Batch Greenkhorn for Regularized Multimarginal Optimal Transport », *in*: *NeurIPS Workshop on Optimal Transport and Machine Learning* (Cited on page 22).

Kriebel, April R and Joshua D Welch (2022), « UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization », *in*: *Nature communications* 13.*1*, page(s): 780 (Cited on page 100).

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger (2015), « From Word Embeddings To Document Distances », *in*: *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, Proceedings of Machine Learning Research, PMLR, page(s): 957–966 (Cited on pages 7, 8).

Lacoste-Julien, Simon (2016), « Convergence Rate of Frank-Wolfe for Non-Convex Objectives », *in*: *arXiv preprint arXiv:1607.00345* (Cited on page 43).

Le, An Hoai Thi and Tao Dinh Pham (2018), « DC programming and DCA: thirty years of developments », *in*: *Mathematical Programming* 169, page(s): 5–68 (Cited on page 60).

Le, Khang, Huy Nguyen, Quang Nguyen, Tung Pham, Hung Bui, and Nhat Ho (2021), « On Robust Optimal Transport: Computational Complexity and Barycenter Computation », *in*: *Neural Information Processing Systems (NeurIPS)* (Cited on page 23).

LeCun, Yann, Corinna Cortes, and CJ Burges (2010), « MNIST handwritten digit database », *in*: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (Cited on page 103).

Lee, John, Nicholas P. Bertrand, and Christopher J. Rozell (2019), « Parallel Unbalanced Optimal Transport Regularization for Large Scale Imaging Problems », *in*: *arXiv preprint arXiv:1909.00149* (Cited on pages 23, 31, 69).

Lehmann, Tobias, Max-K. von Renesse, Alexander Sambale, and André Uschmajew (2020), « A note on overrelaxation in the Sinkhorn algorithm », *in*: *arXiv preprint arXiv:2012.12562* (Cited on page 22).

Lei, Na, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng David Gu (2019), « A geometric view of optimal transportation and generative model », *in*: *Computer Aided Geometric Design* 68, page(s): 1–21 (Cited on page 18).

Li, Jiajin, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet (2022), « Fast and Provably Convergent Algorithms for Gromov-Wasserstein in Graph Data », *in*: *arXiv preprint arXiv:2205.08115* (Cited on page 44).

Liero, Matthias, Alexander Mielke, and Giuseppe Savaré (2018), « Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures », *in*: *Inventiones Mathematicae* 211 (3), page(s): 969–1117 (Cited on pages 8, 23, 24, 31, 46, 47, 69, 88, 135).

Lin, Chi-Heng, Mehdi Azabou, and Eva Dyer (2021), « Making transport more robust and interpretable by moving data through a small number of anchor points », *in*: *Proceedings of the 38th International Conference on Machine Learning* 139, page(s): 6631–6641 (Cited on page 55).

Lin, Tianyi, Nhat Ho, Marco Cuturi, and Michael I. Jordan (2020), « On the Complexity of Approximating Multimarginal Optimal Transport », *in*: *arXiv preprint arXiv:1910.00152* (Cited on page 22).

Lin, Tianyi, Nhat Ho, and Michael I. Jordan (2022), « On the Efficiency of Entropic Regularized Algorithms for Optimal Transport », *in*: *Journal of Machine Learning Research* 23, page(s): 1–42 (Cited on page 22).

Lindbäck, Jacob, Zesen Wang, and Mikael Johansson (2023), « Bringing regularized optimal transport to lightspeed: a splitting method adapted for GPUs », *in*: *arXiv preprint arXiv:2305.18483* (Cited on page 22).

Liu, Jie, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble (2019a), « Jointly embedding multiple single-cell omics measurements », *in*: *BioRxiv*, page(s): 644310 (Cited on pages 79, 80).

— (2019b), « Jointly Embedding Multiple Single-Cell Omics Measurements », *in*: *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, vol. 143, page(s): 10:1–10:13 (Cited on pages 150, 151).

Liu, Tianlin, Joan Puigcerver, and Mathieu Blondel (2022), « Sparsity-Constrained Optimal Transport », *in*: *arXiv preprint arXiv:2209.15466* (Cited on page 22).

Lorenz, Dirk A., Paul Manns, and Christian Meyer (2021), « Quadratically Regularized Optimal Transport », *in*: *Applied Mathematics and Optimization* 83, page(s): 1919–1949 (Cited on page 22).

Luise, Giulia, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto (2018), « Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance », *in*: *Advances in Neural Information Processing Systems*, page(s): 5859–5870 (Cited on page 20).

Léonard, Christian (2012), « From the Schrödinger problem to the Monge-Kantorovich problem », *in*: *Journal of Functional Analysis* 262.*4*, page(s): 1879–1920 (Cited on page 20).

Ma, Zhiheng, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong (2021), « Learning to Count via Unbalanced Optimal Transport », *in*: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.*3*, page(s): 2319–2327 (Cited on pages 23, 69).

Makkuva, Ashok, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee (2020), « Optimal transport mapping via input convex neural networks », *in*: *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, page(s): 6672–6681 (Cited on page 18).

Manupriya, Piyushi, J. Saketha Nath, and Pratik Jawanpuria (2023), « MMD-Regularized Unbalanced Optimal Transport », *in*: *arXiv preprint arXiv:2011.05001* (Cited on page 31).

Marino, Simone Di and Augusto Gerolin (2020), « Optimal Transport losses and Sinkhorn algorithm with general convex regularization », *in*: *arXiv preprint arXiv:2007.00976* (Cited on page 22).

Maron, Haggai and Yaron Lipman (2018), « (Probably) Concave Graph Matching », *in*: *Advances in Neural Information Processing Systems* 31, page(s): 406–416 (Cited on pages 50, 120).

Mars, Rogier B, Stamatios N Sotiropoulos, Richard E Passingham, Jerome Sallet, Lennart Verhagen, Alexandre A Khrapitchev, Nicola Sibson, and Saad Jbabdi (May 11, 2018), « Whole brain comparative anatomy using connectivity blueprints », *in*: *eLife* 7, ed. by Klaas Enno Stephan, Publisher: eLife Sciences Publications, Ltd, page(s): e35237, ISSN: 2050-084X, DOI: 10.7554/eLife.35237, URL: https://doi.org/10.7554/eLife.35237 (Cited on page 86).

Matousek, Jiri (2013), *Lecture Notes on Metric Embeddings* (Cited on page 34).

Mena, Gonzalo and Jonathan Niles-Weed (2019), « Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem », *in*: *Advances in Neural Information Processing Systems*, page(s): 4543–4553 (Cited on page 20).

Mengyu Li Jun Yu, Hongteng Xu and Cheng Meng (2023), « Efficient Approximation of Gromov-Wasserstein Distance Using Importance Sparsification », *in*: *Journal of Computational and Graphical Statistics* 0.*0*, page(s): 1–12 (Cited on page 42).

Mi, Liang and José Bento (2020), « Multi-Marginal Optimal Transport Defines a Generalized Metric », *in*: *CoRR* abs/2001.11114 (Cited on page 55).

Miermont, Grégory (2009), « Tessellations of random maps of arbitrary genus », *in*: *Annales scientifiques de l'Ecole Normale Supérieure* Series 4, 42.*5*, page(s): 725–781 (Cited on page 38).

Monge, Gaspard (1781), « Mémoire sur la théorie des déblais et des remblais », *in*: *Histoire de l'Académie Royale des Sciences*, page(s): 666–704 (Cited on pages 7, 14, 67).

Mroueh, Youssef, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng (2017), « Sobolev GAN », *in*: *arXiv preprint arXiv:1711.04894* (Cited on page 18).

Mukherjee, Debarghya, Aritra Guha, Justin M Solomon, Yuekai Sun, and Mikhail Yurochkin (2021), « Outlier-Robust Optimal Transport », *in*: *Proceedings of the 38th International*

Conference on Machine Learning, vol. 139, Proceedings of Machine Learning Research, PMLR, page(s): 7850–7860 (Cited on pages 23, 69).

Muzellec, Boris, Richard Nock, Giorgio Patrini, and Frank Nielsen (2017), « Tsallis Regularized Optimal Transport and Ecological Inference », *in*: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, AAAI Press, page(s): 2387–2393 (Cited on page 22).

Mémoli, Facundo (2007), « On the use of Gromov-Hausdorff Distances for Shape Comparison », *in*: *Eurographics Symposium on Point-Based Graphics*, The Eurographics Association (Cited on pages 8, 32, 36, 38, 50, 68, 87).

— (2011a), « A spectral notion of Gromov-Wasserstein distance and related methods », *in*: *Applied and Computational Harmonic Analysis* 30.*3*, page(s): 363–401 (Cited on pages 32, 36, 37).

— (2011b), « Gromov-Wasserstein distances and the metric approach to object matching », *in*: *Foundations of Computational Mathematics*, page(s): 1–71 (Cited on pages 8, 15, 16, 32, 33, 36, 50, 68, 87, 100).

Mémoli, Facundo and Tom Needham (2022a), « Comparison Results for Gromov-Wasserstein and Gromov-Monge Distances », *in*: *arXiv preprint arXiv:2212.14123* (Cited on page 39).

— (2022b), « Distance distributions and inverse problems for metric measure spaces », *in*: *Studies in Applied Mathematics* 149.*4*, page(s): 943–1001 (Cited on pages 39, 40, 123).

Mémoli, Facundo and Guillermo Sapiro (2005), « A Theoretical and Computational Framework for Isometry Invariant Recognition of Point Cloud Data », *in*: *Foundations of Computational Mathematics* 5.*3*, page(s): 313–347 (Cited on page 37).

Müller, Alfred (1997), « Integral Probability Metrics and Their Generating Classes of Functions », *in*: *Advances in Applied Probability* 29.*3*, page(s): 429–443 (Cited on page 17).

Nakagawa, Nao, Ren Togo, Takahiro Ogawa, and Miki Haseyama (2022), « Gromov-Wasserstein Autoencoders », *in*: *arXiv preprint arXiv:2209.07007* (Cited on page 100).

Neubert, Franz-Xaver, Rogier B. Mars, Adam G. Thomas, Jerome Sallet, and Matthew F. S. Rushworth (Feb. 5, 2014), « Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex », *in*: *Neuron* 81.*3*, page(s): 700–713, ISSN: 1097-4199, DOI: 10.1016/j.neuron.2013.11.012 (Cited on page 86).

Nguyen, Quang Minh, Hoang H. Nguyen, Yi Zhou, and Lam M. Nguyen (2022), « On Unbalanced Optimal Transport: Gradient Methods, Sparsity and Approximation Error », *in*: *arXiv preprint arXiv:2202.03618* (Cited on page 31).

Nietert, Sloan, Ziv Goldfeld, and Rachel Cummings (2022), « Outlier-Robust Optimal Transport: Duality, Structure, and Statistical Analysis », *in*: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, vol. 151, Proceedings of Machine Learning Research, PMLR, page(s): 11691–11719 (Cited on page 23).

Nikolentzos, Giannis, Polykarpos Meladianos, and Michalis Vazirgiannis (2017), « Matching Node Embeddings for Graph Similarity », *in*: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.*1* (Cited on page 7).

Nitzan, Mor, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky (2019), « Gene expression cartography », *in*: *Nature* 576.*7785*, page(s): 132–137, ISSN: 1476-4687, DOI: 10.1038/s41586-019-1773-3, URL: https://doi.org/10.1038/s41586-019-1773-3 (Cited on page 100).

Nutz, Marcel (2022), *Introduction to Entropic Optimal Transport* (Cited on page 19).

Orlin, James (1988), « A Faster Strongly Polynomial Minimum Cost Flow Algorithm », *in*: *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, STOC '88, page(s): 377–387 (Cited on page 19).

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019), « PyTorch: An Imperative Style, High-Performance Deep Learning Library », *in*: *Advances in Neural Information Processing Systems 32*, ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Curran Associates, Inc., page(s): 8024–8035 (Cited on page 87).

Paty, François-Pierre and Marco Cuturi (2019), « Subspace Robust Wasserstein Distances », *in*: *Proceedings of the 36th International Conference on Machine Learning* 97, page(s): 5072–5081 (Cited on page 35).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), « Scikit-learn: Machine Learning in Python », *in*: *Journal of Machine Learning Research* 12, page(s): 2825–2830 (Cited on page 87).

Pele, Ofir and Michael Werman (2009), « Fast and robust Earth Mover's Distances », *in*: *2009 IEEE 12th International Conference on Computer Vision*, page(s): 460–467 (Cited on page 19).

Petzka, Henning, Asja Fischer, and Denis Lukovnicov (2018), « On the regularization of Wasserstein GANs », *in*: *arXiv preprint arXiv:1709.08894* (Cited on page 18).

Peyré, Gabriel and Marco Cuturi (2019), « Computational Optimal Transport », *in*: *Foundations and Trends in Machine Learning* 11 (Cited on pages 10, 16, 18–21, 43).

Peyré, Gabriel, Marco Cuturi, and Justin Solomon (2016), « Gromov-Wasserstein Averaging of Kernel and Distance Matrices », *in*: *International Conference on Machine Learning* 48 (Cited on pages 32, 41, 43, 63, 68, 91, 100, 104).

Pham, Khiem, Khang Le, Nhat Ho, Tung Pham, and Hung Bui (2020), « On Unbalanced Optimal Transport: An Analysis of Sinkhorn Algorithm », *in*: *Proceedings of the 37 th International Conference on Machine Learning* 119 (Cited on pages 23, 26, 69, 117, 142).

Pham, Tao Dinh and Souad El Bernoussi (1986), « Algorithms for Solving a Class of Nonconvex Optimization Problems. Methods of Subgradients », *in*: *North-Holland Mathematics Studies* 129, page(s): 249–271 (Cited on page 60).

Pham, Tao Dinh and An Hoai Thi Le (1997), « Convex analysis approach to D.C. programming: Theory, Algorithm and Applications », *in*: *Acta Mathematica Vietnamica* 22 (1), page(s): 289–355 (Cited on page 60).

Pinho, Ana Luísa, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, Véronique Joly-Testault, Gaëlle Médiouni-Cloarec, Christine Doublé, Bernadette Martins, Philippe Pinel, Evelyn Eger, Gael Varoquaux, Christophe Pallier, Stanislas Dehaene, Lucie Hertz-Pannier, and Bertrand Thirion (2018), « Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping », *in*: *Scientific Data* 5, page(s): 180105, DOI: 10.1038/sdata.2018.105, URL: https://hal.archives-ouvertes.fr/hal-01817528 (Cited on page 92).

Pizzagalli, F., G. Auzias, Q. Yang, S. R. Mathias, J. Faskowitz, J. D. Boyd, A. Amini, D. Rivière, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, J. F. Mangin, D. C. Glahn, J. Blangero, M. J. Wright, P. M. Thompson, P. Kochunov, and N. Jahanshad (Sept. 2020), « The reliability and heritability of cortical folds and their genetic correlations across hemispheres », *in*: *Commun Biol* 3.*1*, page(s): 510 (Cited on page 85).

Ponti, Nicoló De and Andrea Mondino (2020), « Entropy-Transport distances between unbalanced metric measure spaces », *in*: *arXiv preprint arXiv:2009.10636* (Cited on pages 46, 69).

P.Tseng (2001), « Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization », *in*: *Journal of Optimization Theory and Applications* 109, page(s): 475–494 (Cited on pages 29, 44).

Rabin, Julien, Gabriel Peyré, Julie Delon, and Marc Bernot (2012), « Wasserstein Barycenter and Its Application to Texture Mixing », *in*: *Scale Space and Variational Methods in Computer Vision*, page(s): 435–446 (Cited on pages 19, 50).

Ramdas, Aaditya, Nicolás García Trillos, and Marco Cuturi (2017), « On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests », *in*: *Entropy* 19 (2) (Cited on page 19).

Redko, Ievgen, Nicolas Courty, Rémi Flamary, and Devis Tuia (2019), « Optimal Transport for Multi-source Domain Adaptation under Target Shift », *in*: *Proceedings of Machine Learning Research* 89 (Cited on pages 8, 77).

Redko, Ievgen, Titouan Vayer, Rémi Flamary, and Nicolas Courty (2020), « CO-Optimal Transport », *in*: *Advances in Neural Information Processing Systems* 33 (Cited on pages 7, 8, 49, 50, 52, 63, 68, 71, 74, 77, 110, 125, 136, 150).

Roberts, Lucas, Leo Razoumov, Lin Su, and Yuyang Wang (2017), « Gini-regularized Optimal Transport with an Application to Spatio-Temporal Forecasting », *in*: *arXiv preprint arXiv:1712.02512* (Cited on page 22).

Robinson, Emma C., Kara Garcia, Matthew F. Glasser, Zhengdao Chen, Timothy S. Coalson, Antonios Makropoulos, Jelena Bozek, Robert Wright, Andreas Schuh, Matthew Webster, Jana Hutter, Anthony Price, Lucilio Cordero Grande, Emer Hughes, Nora Tusor, Philip V. Bayly, David C. Van Essen, Stephen M. Smith, A. David Edwards, Joseph Hajnal, Mark Jenkinson, Ben Glocker, and Daniel Rueckert (2018), « Multimodal surface matching with higher-order smoothness constraints », *in*: *NeuroImage* 167, page(s): 453–465, ISSN: 1095-9572, DOI: 10.1016/j.neuroimage.2017.10.037 (Cited on page 92).

Robinson, Emma C., Saad Jbabdi, Matthew F. Glasser, Jesper Andersson, Gregory C. Burgess, Michael P. Harms, Stephen M. Smith, David C. Van Essen, and Mark Jenkinson (Oct. 15, 2014), « MSM: a new flexible framework for Multimodal Surface Matching », *in*: *NeuroImage* 100, page(s): 414–426, ISSN: 1095-9572, DOI: 10.1016/j.neuroimage.2014.05.069 (Cited on pages 86, 92).

Rolet, Antoine, Marco Cuturi, and Gabriel Peyré (2016), « Fast Dictionary Learning with a Smoothed Wasserstein Loss », *in*: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* 51, page(s): 630–638 (Cited on pages 8, 67).

Ryner, Martin and Johan Karlsson (2022), « Orthogonalization of data via Gromov-Wasserstein type feedback for clustering and visualization », *in*: *arXiv preprint arXiv:2207.12279* (Cited on page 32).

Ryner, Martin, Jan Kronqvist, and Johan Karlsson (2023), « Globally solving the Gromov-Wasserstein problem for point clouds in low dimensional Euclidean spaces », *in*: *Advances in Neural Information Processing Systems* 36 (Cited on page 42).

Sabuncu, Mert R., Benjamin D. Singer, Bryan Conroy, Ronald E. Bryan, Peter J. Ramadge, and James V. Haxby (Jan. 2010), « Function-based intersubject alignment of human cortical anatomy », *in*: *Cerebral Cortex (New York, N.Y.: 1991)* 20.*1*, page(s): 130–140, ISSN: 1460-2199, DOI: 10.1093/cercor/bhp085 (Cited on page 86).

Saenko, Kate, Brian Kulis, Mario Fritz, and Trevor Darrell (2010), « Adapting visual category models to new domains », *in*: *Proceedings of the 11th European Conference on Computer Vision*, page(s): 213–226 (Cited on pages 77, 110).

Salimans, Tim, Han Zhang, Alec Radford, and Dimitris Metaxas (2018), « Improving GANs Using Optimal Transport », *in*: *arXiv preprint arXiv:1803.05573* (Cited on page 18).

Sanjabi, Maziar, Jimmy Ba, Meisam Razaviyayn, and Jason D. Lee (2018), « On the convergence and robustness of training GANs with regularized optimal transport », *in*: *Advances in Neural Information Processing Systems*, page(s): 7091–7101 (Cited on pages 18, 20).

Santambrogio, Filippo (2015), *Optimal transport for applied mathematicians*, Springer (Cited on pages 14, 17, 18).

Scetbon, Meyer, Marco Cuturi, and Gabriel Peyré (2021), « Low-Rank Sinkhorn Factorization », *in*: *Proceedings of the 38th International Conference on Machine Learning* 139, page(s): 9344–9354 (Cited on pages 42, 55, 133).

Scetbon, Meyer, Gabriel Peyré, and Marco Cuturi (2021), « Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs », *in*: *arXiv preprint arXiv:2106.01128* (Cited on pages 41, 46).

Schiebinger, Geoffrey, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander (2019), « Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming », *in*: *Cell* 176.*4*, page(s): 928–943 (Cited on pages 8, 23, 69).

Schmitzer, Bernhard (2019), « Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems », *in*: *SIAM Journal on Scientific Computing* 41 (3), page(s): 1443–1481 (Cited on page 21).

Schneider, Marian, Valentin G. Kemper, Thomas C. Emmerling, Federico De Martino, and Rainer Goebel (2019), « Columnar clusters in the human motion complex reflect consciously perceived motion axis », *in*: *Proceedings of the National Academy of Sciences* 116.*11*, page(s): 5096–5101, DOI: 10.1073/pnas.1814504116, eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1814504116, URL: https://www.pnas.org/doi/abs/10.1073/pnas.1814504116 (Cited on page 86).

Schrödinger, Erwin (1932), « Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique », *in*: *Annales de l'institut Henri Poincaré* 2.*4*, page(s): 269–310 (Cited on page 19).

Seguy, Vivien, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel (2018), « Large-Scale Optimal Transport and Mapping Estimation », *in*: *Proceedings of the International Conference in Learning Representations* (Cited on page 22).

Singh, Ritambhara, Pinar Demetci, Giancarlo Bonora, Vijay Ramani, Choli Lee, He Fang, Zhijun Duan, Xinxian Deng, Jay Shendure, Christine Disteche, and William Stafford Noble (2020), « Unsupervised manifold alignment for single-cell multi-omics data », *in*: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '20, Virtual Event, USA: Association for Computing Machinery, ISBN: 9781450379649, DOI: 10.1145/3388440.3412410, URL: https://doi.org/10.1145/3388440.3412410 (Cited on pages 150, 151).

Sinkhorn, Richard and Paul Knopp (1967), « Concerning nonnegative matrices and doubly stochastic matrices », *in*: *Pacific Journal of Mathematics* 21 (2), page(s): 343–348 (Cited on page 20).

Solomon, Justin, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas (2015), « Convolutional wasserstein distances: efficient optimal transportation on geometric domains », *in*: *ACM Transactions on Graphics* 34.*4*, ISSN: 0730-0301, DOI: 10.1145/2766963, URL: https://doi.org/10.1145/2766963 (Cited on page 8).

Solomon, Justin, Gabriel Peyré, Vladimir G. Kim, and Suvrit Sra (2016), « Entropic Metric Alignment for Correspondence Problems », *in*: *ACM Transactions on Graphics* 35.*4* (Cited on pages 32, 43, 68).

Solomon, Justin, Raif Rustamov, Leonidas Guibas, and Adrian Butscher (2014), « Wasserstein Propagation for Semi-Supervised Learning », *in*: *Proceedings of the 31st International Conference on Machine Learning* 32 (1), page(s): 306–314 (Cited on page 68).

Sommerfeld, Max, Jörn Schrieber, Yoav Zemel, and Axel Munk (2019), « Optimal Transport: Fast Probabilistic Approximation with Exact Solvers », *in*: *Journal of Machine Learning Research* 20.*105*, page(s): 1–23 (Cited on page 19).

Sonthalia, Rishi and Anna C. Gilbert (2020), « Dual Regularized Optimal Transport », *in*: *arXiv preprint arXiv:2012.03126* (Cited on page 31).

Stein, Elias M. and Rami Shakarchi (2005), *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Princeton Lectures in Analysis, Princeton University Press (Cited on page 148).

Stoeckius, Marlon, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert (2017), « Simultaneous epitope and transcriptome measurement in single cells », *in*: *Nature Methods* 14.*9*, page(s): 865–868, DOI: 10.1038/nmeth.4380, URL: https://doi.org/10.1038/nmeth.4380 (Cited on pages 79, 80, 108, 150).

Sturm, Karl-Theodor (2006), « On the geometry of metric measure spaces », *in*: *Acta Mathematica* 196, page(s): 65–131 (Cited on pages 34, 35, 38, 69).

— (2012), « The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces », *in*: *arXiv preprint arXiv:1208.0434* (Cited on pages 32, 38, 39, 68).

Sun, Ying, Prabhu Babu, and Daniel P. Palomar (2017), « Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning », *in*: *IEEE Transactions on Signal Processing* 65.*3*, page(s): 794–816 (Cited on page 29).

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich (2015), « Going Deeper with Convolutions », *in*: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page(s): 1–9 (Cited on pages 77, 110).

Séjourné, Thibault, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré (2019), « Sinkhorn Divergences for Unbalanced Optimal Transport », *in*: *arXiv preprint arXiv:1910.12958* (Cited on pages 23, 25, 27, 31, 90, 98, 113).

Séjourné, Thibault, Gabriel Peyré, and François-Xavier Vialard (2022), « Unbalanced Optimal Transport, from Theory to Numerics », *in*: *arXiv preprint arXiv:2211.08775* (Cited on page 10).

Séjourné, Thibault, François-Xavier Vialard, and Gabriel Peyré (2021a), « Faster Unbalanced Optimal Transport: Translation invariant Sinkhorn and 1-D Frank-Wolfe », *in*: *NeurIPS Optimal Transport and Machine Learning Workshop* (Cited on pages 28, 29, 113).

— (2021b), « The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation », *in*: *Advances in Neural Information Processing Systems* 34 (Cited on pages 9, 46, 50, 53, 69, 74, 87–89, 107, 135).

Taghvaei, Amirhossein and Amin Jalali (2019), « 2-Wasserstein Approximation via Restricted Convex Potentials with Application to Improved Training for GANs », *in*: *arXiv preprint arXiv:1902.07197* (Cited on page 18).

Tao, Terence (2012), *Topics in random matrix theory*, Graduate Studies in Mathematics, American Mathematical Society (Cited on pages 105, 148).

Tavor, Ido, O Parker Jones, Rogier B Mars, SM Smith, TE Behrens, and Saad Jbabdi (2016), « Task-free MRI predicts individual differences in brain activity during task performance », *in*: *Science* 352.*6282*, page(s): 216–220 (Cited on page 86).

Terjék, Dávid and Diego González-Sánchez (2022), « Optimal transport with f-divergence regularization and generalized Sinkhorn algorithm », *in*: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, vol. 151, Proceedings of Machine Learning Research, PMLR, page(s): 5135–5165 (Cited on page 22).

Thibault, Alexis, Lénaïc Chizat, Charles Dossal, and Nicolas Papadakis (2021), « Overrelaxed Sinkhorn-Knopp Algorithm for Regularized Optimal Transport », *in*: *Algorithms* 14.*5* (Cited on page 22).

Thirion, Bertrand, Gaël Varoquaux, Elvis Dohmatob, and Jean-Baptiste Poline (2014), « Which fMRI clustering gives good brain parcellations? », *in*: *Frontiers in Neuroscience* 8.*167*, page(s): 13, DOI: 10.3389/fnins.2014.00167, URL: https://hal.inria.fr/hal-01015172 (Cited on page 93).

Thornton, James and Marco Cuturi (2023), « Rethinking Initialization of the Sinkhorn Algorithm », *in*: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, vol. 206, Proceedings of Machine Learning Research, PMLR, page(s): 8682–8698 (Cited on page 21).

Thorpe, Matthew, Serim Park, Soheil Kolouri, Gustavo K. Rohde, and Dejan SlepăźEv (2017), « A Transportation L^p Distance for Signal Analysis », *in*: *Journal of Mathematical Imaging and Vision* 59.*2*, page(s): 187–210 (Cited on page 46).

Thual, Alexis, Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion (2022), « Aligning individual brains with Fused Unbalanced Gromov-Wasserstein », *in*: *Advances in Neural Information Processing Systems* (Cited on pages 11, 46, 84).

Tosches, Maria Antonietta, Tracy M. Yamawaki, Robert K. Naumann, Ariel A. Jacobi, Georgi Tushev, and Gilles Laurent (2018), « Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles », *in*: *Science* 360.*6391*, page(s): 881–888, DOI: 10.1126/science.aar4237, eprint: https://www.science.org/doi/pdf/10.1126/science.aar4237, URL: https://www.science.org/doi/abs/10.1126/science.aar4237 (Cited on pages 109, 150).

Tran, Quang Huy, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, and Ritambhara Singh (2023), « Unbalanced CO-optimal Transport », *in*: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.*8*, page(s): 10006–10016 (Cited on pages 11, 46, 67, 107).

Tran, Quang Huy, Hicham Janati, Ievgen Redko, Rémi Flamary, and Nicolas Courty (2021), « Factored couplings in multi-marginal optimal transport via difference of convex programming », *in*: *NeurIPS Optimal Transport and Machine Learning Workshop* (Cited on pages 10, 48).

Van Essen, D. C., M. F. Glasser, D. L. Dierker, J. Harwell, and T. Coalson (Oct. 2012), « Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases », *in*: *Cereb Cortex* 22.*10*, page(s): 2241–2262 (Cited on page 85).

Van Essen, D. C. et al. (Oct. 2013), « The WU-Minn Human Connectome Project: an overview », *in*: *Neuroimage* 80, page(s): 62–79 (Cited on page 98).

Vayer, Titouan, Laetita Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty (2019a), « Optimal Transport for structured data with application on graphs », *in*: *International Conference on Machine Learning* 97 (Cited on pages 9, 32, 42, 45, 46, 68, 87, 88, 91, 104, 106, 145).

Vayer, Titouan, Flamary Rémi, Courty Nicolas, Tavenard Romain, and Chapel Laetitia (2019b), « Sliced Gromov-Wasserstein », *in*: *Advances in Neural Information Processing Systems* 32, page(s): 14726–14736 (Cited on page 50).

Vershynin, Roman (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, DOI: 10.1017/9781108231596 (Cited on pages 49, 79).

Villani, Cédric (2003), *Topics in optimal transportation*, vol. 58, Graduate Studies in Mathematics, American Mathematical Society (Cited on pages 15, 17, 18, 68, 100, 126).

— (2009), *Optimal transport: old and new*, vol. 338, Grundlehren der mathematischen Wissenschaften, Springer-Verlag (Cited on pages 17, 18, 32, 38, 119, 121, 122).

Vincent-Cuaz, Cédric, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty (2021), « Online Graph Dictionary Learning », *in*: Proceedings of Machine Learning Research 139, page(s): 10564–10574 (Cited on page 32).

Vincent-Cuaz, Cédric, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty (2022), « Semi-relaxed Gromov-Wasserstein divergence with applications on graphs », *in*: (Cited on page 47).

Vo, Thanh Xuan (2015), « Learning with sparsity and uncertainty by Difference of Convex functions optimization », PhD thesis, Université de Lorraine (Cited on page 65).

Weed, Jonathan (2018), « An explicit analysis of the entropic penalty in linear programming », *in*: *Proceedings of the 31st Conference On Learning Theory*, vol. 75, Proceedings of Machine Learning Research, PMLR, page(s): 1841–1855 (Cited on page 20).

Xie, Yujia, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha (2020), « A Fast Proximal Point Method for Computing Exact Wasserstein Distance », *in*: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, Proceedings of Machine Learning Research 115, page(s): 433–453 (Cited on pages 43, 113, 114).

Xu, Hongteng, Dixin Luo, and Lawrence Carin (2019), « Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching », *in*: *Advances in Neural Information Processing Systems* 32 (Cited on pages 32, 43, 68).

Xu, Hongteng, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke (2019), « Gromov-Wasserstein Learning for Graph Matching and Node Embedding », *in*: *Proceedings of the 36th International Conference on Machine Learning* (97), page(s): 6932–6941 (Cited on pages 32, 43, 44, 46, 68).

Xu, Ting, Karl-Heinz Nenning, Ernst Schwartz, Seok-Jun Hong, Joshua T. Vogelstein, Alexandros Goulas, Damien A. Fair, Charles E. Schroeder, Daniel S. Margulies, Jonny Smallwood, Michael P. Milham, and Georg Langs (Dec. 1, 2020), « Cross-species functional alignment reveals evolutionary hierarchy within the connectome », *in*: *NeuroImage* 223, page(s): 117346, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2020.117346, URL: http://www.sciencedirect.com/science/article/pii/S1053811920308326 (Cited on page 86).

Yan, Yuguang, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu (2018), « Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation », *in*: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, page(s): 2969–2975 (Cited on pages 32, 68, 77, 100).

Yang, Karren D. and Caroline Uhler (2019), « Scalable Unbalanced Optimal Transport using Generative Adversarial Networks », *in*: *7th International Conference on Learning Representations* (Cited on pages 23, 69).

Yeo, B.T. Thomas, Mert R. Sabuncu, Tom Vercauteren, Nicholas Ayache, Bruce Fischl, and Polina Golland (Mar. 2010), « Spherical Demons: Fast Diffeomorphic Landmark-Free Surface Registration », *in*: *IEEE transactions on medical imaging* 29.*3*, page(s): 650–668, ISSN: 0278-0062, DOI: 10.1109/TMI.2009.2030797, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2862393/ (Cited on page 86).

Zappia, Luke, Belinda Phipson, and Alicia Oshlack (2017), « Splatter: simulation of single-cell RNA sequencing data », *in*: *Genome biology* 18.*1*, page(s): 1–15 (Cited on page 150).

Zeira, Ron, Max Land, Alexander Strzalkowski, and Benjamin J Raphael (2022), « Alignment and integration of spatial transcriptomics data », *in*: *Nature Methods* 19.*5*, page(s): 567–575 (Cited on page 100).

Zhang, Zhengxin, Ziv Goldfeld, Youssef Mroueh, and Bharath K. Sriperumbudur (2022a), « Gromov-Wasserstein Distances: Entropic Regularization, Duality, and Sample Complexity », *in*: *arXiv preprint arXiv:2212.12848* (Cited on pages 10, 55, 114).

Zhang, Zhengxin, Youssef Mroueh, Ziv Goldfeld, and Bharath Sriperumbudur (2022b), « Cycle Consistent Probability Divergences Across Different Spaces », *in*: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, vol. 151, Proceedings of Machine Learning Research, PMLR, page(s): 7257–7285 (Cited on pages 37, 47, 69).

Zhu, Ciyou, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal (1997), « Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization », *in*: *ACM Transactions on Mathematical Software* 23.*4*, page(s): 550–560, ISSN: 0098-3500, DOI: 10.1145/279232.279236, URL: https://doi.org/10.1145/279232.279236 (Cited on page 25).

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros (2017), « Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks », *in*: *2017 IEEE International Conference on Computer Vision (ICCV)*, page(s): 2242–2251 (Cited on page 37).

Titre : Transport optimal pour l'apprentissage par transfert entre les espaces

Mot clés : Transport Optimal Non-équilibré, Gromov-Wasserstein, Co-Optimal Transport, Adaptation de Domaine Hétérogène

Résumé : Au cours des dernières années, la remarquable puissance de la théorie du transport optimal a largement dépassé la comparaison classique des mesures de probabilité évoluant dans le même espace sous-jacent. Dans cette thèse, nous nous intéressons aux problèmes de transport optimal entre des espaces incomparables. Plus précisément, nous nous concentrons sur la relaxation marginale du transport optimal (équilibré) de Gromov-Wasserstein et du Co-Optimal Transport, ainsi que sur l'intégration de connaissances préalables dans la distance de Gromov-Wasserstein et sa formulation non-équilibrée. Nous commençons par le Co-Optimal Transport en cadre continu, qui sert de première étape vers l'étude de l'approximation entropique et de l'extension non-équilibrée. Ensuite, nous introduisons la formulation non-équilibrée du Co-Optimal Transport et montrons sa robustesse aux valeurs aberrantes, contrairement à son homologue équilibré. Ensuite, nous proposons d'utiliser la divergence de Fused Gromov-Wasserstein non-équilibrée pour aligner les surfaces corticales, en exploitant simultanément les signaux fonctionnels et la structure anatomique du cerveau humain. Enfin, nous renforçons davantage la distance de Gromov-Wasserstein avec la capacité de manipuler plus efficacement les données brutes et d'effectuer un appariement des features génomiques.

Title: Optimal transport for transfer learning across domains

Keywords: Unbalanced Optimal Transport, Gromov-Wasserstein, Co-Optimal Transport, Heterogeneous Domain Adaptation

Abstract: In the recent years, the remarkable versatility of optimal transport theory has gone far beyond the classic comparison of the probability measures living in the same underlying space. In this thesis, we are interested in the optimal transport problems between incomparable spaces. More precisely, we focus on the marginal relaxation of the (balanced) Gromov-Wasserstein and Co-Optimal Transport, as well as the integration of prior knowledge into Gromov-Wasserstein distance and its unbalanced formulation. We start with the Co-Optimal Transport in continuous setting, which serves as the first step towards the study of the entropic approximation and unbalanced extension. Then, we introduce the unbalanced formulation of the Co-Optimal Transport and show its robustness to outliers, by contrast to the balanced counterpart. Next, we propose to use the fused unbalanced Gromov-Wasserstein divergence to align the cortical surfaces, by simultaneously exploiting the functional signals and anatomical structure of human brain. Finally, we further empower the Gromov-Wasserstein distance with the ability to manipulate more efficiently the input data and to perform meaningful genomic feature matching.