THÈSE DE DOCTORAT DE

L'UNIVERSITÉ BRETAGNE SUD
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies
de l'Information et de la Communication*
Spécialité : *Informatique*

Par

Titouan VAYER

A contribution to Optimal Transport on incomparable spaces

**Thèse présentée et soutenue à Vannes, le Le 05 Novembre 2020
Unité de recherche : IRISA
Thèse N° : 573**

Rapporteurs avant soutenance :

Gabriel Peyré   Directeur de Recherche CNRS, DMA, École Normale Supérieure.
Marc Sebban   Professeur des universités. Université Jean Monnet, Saint-Etienne.

Composition du Jury :

Président :   Filippo Santambrogio   Professeur des universités, Université Claude Bernard, Lyon 1.
Examinateurs :   Julie Delon   Professeure des universités, Université Paris Descartes.
                    Pierre Vandergheynst   Full Professor, Ecole Polytechnique fédérale de Lausanne.
Dir. de thèse :   Nicolas Courty   Professeur des universités, Université Bretagne Sud.
Co-dir. de thèse : Laetitia Chapel   Maître de Conférence, Université Bretagne Sud.
                    Romain Tavenard   Maître de Conférence, Université de Rennes 2.

Invité(s) :

Rémi Flamary   Maître de Conférence, HDR, CMAP, École Polytechnique.

Acknowledgements

Résumé (Français)

Introduction

"Ainsi, l'on voit dans les Sciences, tantôt des théories brillantes, mais longtemps inutiles, devenir tout à coup le fondement des applications les plus importantes, et tantôt des applications très simples en apparence, faire naître l'idée de théories abstraites dont on n'avait pas encore le besoin, diriger vers les théories des travaux des Géomètres, et leur ouvrir une carrière nouvelle." C'est ainsi que Nicolas De Condorcet [De Condorcet 1781] au 18ème siècle a introduit les travaux de Gaspard Monge [Monge 1781] qui sont au coeur de la théorie du transport optimal. Comment déplacer des masses d'un endroit à un autre de sorte à minimiser l'effort global de déplacement ? Condorcet avait raison : cette "idée simple" a maturé au fil des ans en une élégante théorie au croisement des mathématiques et de l'optimisation et est aujourd'hui au centre de nombreuses applications en machine learning.

D'une manière générale, l'intérêt du transport optimal réside à la fois dans sa capacité à fournir des relations, des correspondances, entre des ensembles de points et dans le fait qu'il induise une notion géométrique de distance entre des distributions de probabilité (voir Figure 1). Ces deux propriétés se sont révélées très utiles pour un large éventail de tâches qui sont, pour n'en citer que quelques-unes, le recalage d'images [Haker 2001], la recherche d'image par contenu [Rubner 1998], l'adaptation de domaine [Courty 2017], le traitement du signal [Kolouri 2017], l'apprentissage non supervisé [Arjovsky 2017, Genevay 2018], l'apprentissage supervisé et semi-supervisé [Frogner 2015, Solomon 2014], le traitement automatique du langage [Kusner 2015], l'équité en machine learning [Gordaliza 2019], ou encore en biologie [Schiebinger 2019] ou en astrophysique [Frisch 2002].

Malgré ses nombreuses propriétés, le problème de transport optimal reste difficile à résoudre en pratique et il est connu pour souffrir de problèmes de scalabilité qui empêchent son utilisation sur des données volumineuses, omniprésentes en machine learning. L'émergence du transport optimal a été grandement favorisée par de récentes avancées dans le domaine de l'optimisation [Cuturi 2013, Altschuler 2017, Genevay 2016].

De plus, dans sa formulation originelle, le transport optimal reste assez limité aux applications où il existe un "moyen direct" pour comparer les points, appelés *samples*, provenant des distributions. Il est donc souvent limité aux cas où les samples font partie du même espace métrique, qui est la plupart du temps un espace euclidien. Cette limitation empêche notamment l'utilisation du transport optimal pour une variété de tâches dans lesquelles il existe une *information de structure* supplémentaire sur les données, qui ne peut généralement pas être décrite par des espaces euclidiens. On peut citer par exemple le cas où les samples sont décrits par des graphes, des arbres ou des séries temporelles. Cette limitation empêche également son utilisation lorsque les samples se trouvent dans des espaces métriques différents, potentiellement non liés, ou lorsqu'une notion de distance entre les samples ne peut pas être facilement définie. Tous ces cas font partie de ce qu'on appellera par la suite le cas de figure *incomparable*. Une solution intéressante se trouve dans l'élégante théorie de la distance de Gromov-Wasserstein [Memoli 2011]
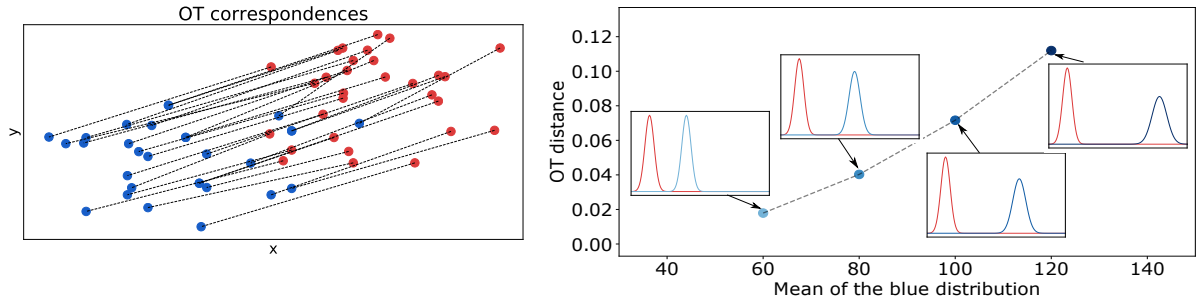
Figure 1: Le Transport Optimal fournit des outils pour trouver des correspondances, relations, entre des ensembles de points et donne une notion géométrique de distance entre des distributions de probabilité. **(à gauche)** Deux ensembles de points en 2D. Les correspondances trouvées par l'OT sont illustrées en pointillées. **(à droite)** Les distances définies par l'OT entre deux distributions de probabilité bleue et rouge. Plus la moyenne de la distribution bleue est éloignée de la rouge, plus la distance est grande.

qui ne nécessite pas la comparaison des échantillons *entre* les distributions. Cependant, cette distance est connue pour être encore plus difficile à résoudre que le transport optimal classique. L'objectif de cette thèse est d'aider à surmonter ces différents obstacles en:

(i) Définissant de nouveaux problèmes de transport optimal sur des espaces incomparables et en particulier pour des données structurées.

(ii) Réduisant l'écart entre la compréhension théorique du transport optimal classique et celle du problème de Gromov-Wasserstein.

Données structurées et espaces incomparables en machine learning   Avant d'entrer dans le détail des contributions de cette thèse il est important de rendre explicite la notion de *structure* et *d'espaces incomparables*. Une première approche est de considérer une *information structurelle* comme étant l'élément d'information qui encode les relations spécifiques qui existent entre les composants d'un objet. Cette définition peut être mise en relation avec le concept de *relationnal reasoning* [Battaglia 2018] dans lequel des *entités* (ou des éléments ayant des attributs tels que l'intensité d'un signal) coexistent avec certaines *relations* ou propriétés entre eux.

Des tels cas de données structurées apparaissent naturellement lorsque la structure est *explicite*. Par exemple, dans le contexte des graphes, les arêtes sont représentatives de la structure, de sorte que chaque attribut du graphe (généralement un vecteur de $\mathbb{R}^d$) peut être lié à d'autres par les arêtes entre les noeuds. Ces objets apparaissent notamment pour modéliser des composés chimiques ou des molécules [Kriege 2016], la connectivité du cerveau [Ktena 2017] ou encore les réseaux sociaux [Yanardag 2015]. Cette famille générique de données structurées comprend également les arbres [Day 1985] ou encore les séries temporelles dont les valeurs sont corrélées dans le temps, de sorte leur comparaison nécessite de prendre en compte dans la modélisation la structure temporelle, la direction du temps.

Les informations de structure des données en machine learning peuvent apparaître de manière plus subtile voire même *implicitement*. Elles peuvent se matérialiser lorsqu'on construit une *prior* ou un *biais* structurel sur la représentation des données. Par exemple, dans le contexte de l'apprentissage profond, de nombreuses architectures exploitent *l'équivariance* à une transformation pour améliorer la généralisation des modèles. Les réseaux de neurones convolutifs (CNN) sont un excellent exemple de cette prior structurelle: un CNN est équivariant aux translation, *i.e.* si nous translatons l'entrée du réseau de

neurone, la sortie des convolutions sera également translatée. Ce biais structurel est connu pour révéler une certaine hiérarchie spatiale sur les pixels utile dans de nombreuses applications [Wang 2018, Chen 2018]. D'autres travaux ont étudié la conception de couches avec des équivariances à d'autres transformations telles que les permutations, rotations ou réflexions [Kondor 2018, Cohen 2016, Gens 2014].

Contrairement aux méthodes "end-to-end" telles que les réseaux de neurones, d'autres approches plus "hand-engineering", basées sur la segmentation des images, permettent de dévoiler une certaine structure utile sur les images [Bach 2007, Jianbo Shi 2000]. Les structures implicites sont également au cœur de nombreux outils de traitement du langage naturel (NLP) utilisés pour trouver de bonnes représentations vectorielles des mots [Mikolov 2013b, Mikolov 2013a, Pennington 2014], pour la reconnaissance vocale [Hinton 2012] ou plus généralement pour l'apprentissage de séquences [Sutskever 2014]. Dans ces cas, les biais structurels passent soit par l'utilisation de variables latentes ou celle de probabilités conditionnelles. Lorsqu'ils sont disponibles, les labels ou les classes induisent également une structure implicite sur les features des données. Par exemple, en adaptation de domaines, on peut souhaiter que les samples du domaine source ayant le même label soient appariés de manière cohérente dans la même région de l'espace cible, et ainsi éviter qu'ils ne soient divisés en des emplacements trop éloignés [Courty 2017, Alvarez-Melis 2018b]. La tendance récente dans la communauté de machine learning des réseaux de neurones pour les graphes (GNN) [Wu 2020] est l'un des nombreux exemples soulignant l'importance des données structurées de nos jours.

Alors que les exemples précédents considèrent les données structurées comme des *entrées* du processus d'apprentissage, la prédominance des données structurées en machine learning se manifeste également dans de nombreux travaux où les données structurées sont des *sorties.* On peut citer par exemple le domaine de la prédiction structurée dans lequel on veut apprendre à produire des objets structurés tels que des séquences, des arbres ou des assignements [Taskar 2005, Lafferty 2001, Collins 2002, Blondel 2020, Mensch 2018, Korba 2018].

En bref, la notion de *structure* en machine learning est omniprésente et apparaît aussi souvent qu'il y a une information supplémentaire sur les objets qui va au-delà de leurs représentations caractéristiques, de leurs features. Comme le montrent de nombreux contextes de machine learning tels que les modèles graphiques [Pearl 1986, Pearl 2009], l'apprentissage par renforcement [D**v**zeroski 2001] ou les modèles bayésiens non-paramétriques [Hjort 2010], considérer les objets comme une composition complexe d'entités avec certaines interactions est particulièrement utile, afin d'apprendre à partir de petites quantités de données.

La notion précédente de données structurées peut se voir comme étant un cas particulier de données définies sur des espaces incomparables. Dans cette situation, chaque sample possède une "caractéristique" qui lui est propre et qui peut ne pas être partagée avec les autres samples. Par exemple, lorsque l'on considère un ensemble de données composé de plusieurs graphes, la structure d'un graphe n'est généralement pas partagée avec les autres graphes. Cette notion, volontairement large, englobe également le cas où les données proviennent de sources hétérogènes. Un exemple particulier de ce problème est l'adaptation de domaines hétérogènes [Yeh 2014, Zhou 2014, Wang 2011] qui vise à exploiter les connaissances provenant de domaines sources hétérogènes pour améliorer les performances d'apprentissage dans un domaine cible avec, potentiellement, différents espaces pour les features entre les domaines source et cible. Le cas des datasets MNIST/USPS [LeCun 2010, Hull 1994] illustre assez bien cette situation : sur la base de la connaissance d'images de chiffre de taille 28×28 (*i.e.* des vecteurs de $\mathbb{R}^{784}$) de MNIST, comment construire *e.g.* un classifieur qui fonctionne sur des images de chiffres 16×16 (*i.e.* des vecteurs de $\mathbb{R}^{256}$) de USPS ? Il va sans dire que ce problème se pose souvent dans tous les domaines du machine learning : il est courant que les données soient recueillies à partir de sources diverses et hétérogènes et les méthodes qui s'appuient sur
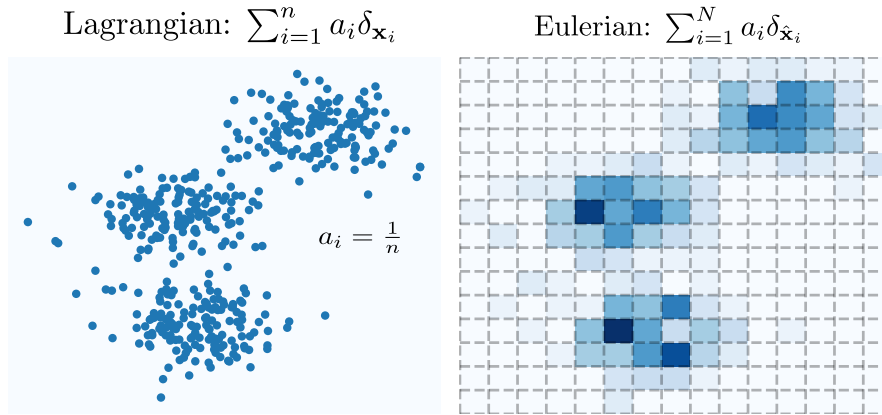
Figure 2: On associe à un dataset $(\mathbf{x}_i)_{i \in [\![n]\!]}$ une distribution de probabilité le décrivant entièrement. Dans cet exemple $\mathbf{x}_i \in \mathbb{R}^2$. **(à gauche)** Formulation Lagragienne (ou nuage de points): $\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$ est une mesure de probabilité discrète où $(a_i)_{i \in [\![n]\!]}$ est un vecteur de probabilité, *i.e.* $a_i \geq 0$ et $\sum_{i=1}^{n} a_i = 1$ et δ est la mesure de dirac *i.e.* $\delta_{\mathbf{x}_i}(\mathbf{x}) = 1$ si $\mathbf{x} = \mathbf{x}_i$ et 0 sinon. **(à droite)** Formulation Euleriennne (histogrammes): on associe une mesure de probabilité $\sum_{i=1}^{N} a_i \delta_{\hat{\mathbf{x}}_i}$ où $(\hat{\mathbf{x}}_i)_{i \in [\![N]\!]}$ est une grille régulière qui discrétise l'espace $\mathbb{R}^2$ et $(a_i)_{i \in [\![N]\!]}$ est un vecteur de probabilité.

cette diversité présentent souvent un grand intérêt.

Le transport optimal pour le machine learning La question centrale qui se pose souvent en machine learning est la suivante : comment représenter les données et comment les comparer ? Le cadre des distributions de probabilités apporte certaines réponses à cette question en associant une mesure de probabilité à une collection de samples qui forment un dataset $(\mathbf{x}_i)_{i \in [\![n]\!]}$. La représentation *Lagragienne* du dataset résulte en une mesure de probabilité discrète $\sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$ dans laquelle on associe à chaque point $\mathbf{x}_i$ un dirac $\delta_{\mathbf{x}_i}(\mathbf{x}) = 1$ si $\mathbf{x} = \mathbf{x}_i$ sinon 0 ainsi qu'un poids $a_i \geq 0$ tel que $(a_i)_{i \in [\![n]\!]}$ est un vecteur de probabilité qui satisfait $\sum_{i=1}^{n} a_i = 1$. Lorsqu'aucune information concernant l'importance relative des samples dans le dataset n'est disponible les poids peuvent être choisis uniformes de sorte que $a_i = \frac{1}{n}$. De même une représentation *Eulerienne* peut être construite *via* la distribution de probabilité $\sum_{i=1}^{N} a_i \delta_{\hat{\mathbf{x}}_i}$ dans laquelle $(\hat{\mathbf{x}}_i)_{i \in [\![N]\!]}$ est une grille régulière qui discrétise l'espace. Cette formulation est équivalente à construire un histogramme sur nos données (voir Figure 2). Ces points de vue sur les données préconisent de trouver un moyen approprié de comparer leur représentation sous forme de distributions de probabilité et, à ce titre, la question de trouver des outils pour les comparer est au cœur de nombreux algorithmes de machine learning.

Bien qu'il existe diverses divergences telles que ϕ-divergences [Csiszar 1975] ou Maximum Mean Discrepancies (MMD) [Gretton 2007], la richesse du transport optimal réside dans sa capacité à intégrer la géométrie de l'espace sous-jacent dans sa formulation et à prêter attention aux relations, aux correspondances des échantillons au sein de leurs représentations respectives. Pour souligner en bref l'avantage de représenter les données par des histogrammes/distributions de probabilités couplé à l'utilisation du

transport optimal, nous pouvons citer par exemple [Rubner 2000] pour la recherche d'images par contenu ou [Kusner 2015] pour le traitement du langage naturel. À ce stade, des questions naturelles se posent : cette représentation théorique des données est-elle utilisable lorsque leur nature est intrinsèquement structurée, ou lorsque les samples se trouvent dans des espaces incomparables ? Dans ce cas, comment pouvons-nous représenter les données sous forme de distributions de probabilités ? Dans quelle mesure cette représentation est-elle valable ? Le transport optimal est-il toujours applicable et, si ce n'est pas le cas, comment comparer ces distributions de probabilités ? L'objectif de cette thèse, entre autres, est de fournir des réponses à ces questions.

Contributions

Cette thèse couvre la majeure partie des travaux de l'auteur et se concentre sur un seul axe de recherche qui est *Le transport optimal sur des espaces incomparables.* Des travaux supplémentaires [Vayer 2020a] sur les séries temporelles sur des espaces incomparables, qui ne sont pas basés sur le transport optimal, ne sont pas inclus dans cette thèse. Le lecteur intéressé peut cependant trouver les détails dans la bibliographie.

Chapitre 2

Ce chapitre présente les résultats fondamentaux de la théorie classique du transport optimal et résume/illustre ses différentes formulations ainsi que quelques solveurs numériques connus. La philosophie de ce chapitre est de fournir un aperçu de haut niveau du transport optimal, tant en théorie qu'en pratique. Ce chapitre se conclut sur la théorie de Gromov-Wasserstein qui est au cœur de la thèse. Un lecteur familier avec les concepts de base du transport optimal peut passer cette partie bien qu'elle contienne des concepts et des notations essentiels qui seront abordés tout au long de la thèse.

Chapitre 3

Ce chapitre est consacré au transport optimal pour les données structurées, et notamment dans le contexte des graphes. Il est basé sur les travaux des articles [Vayer 2019a] et [Vayer 2020b] et fournit des réponses à la question de la définition d'un cadre mathématique pour le transport optimal dans le cas de données structurées. Nous fournissons un cadre général, basé sur la notion de Fused Gromov-Wasserstein qui définit une distance de transport optimal entre des objets structurés tels que des graphes labelés non dirigés.

En résumé, nous considérons les graphes labelés non orientés comme des tuples de la forme $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \ell_f, \ell_s)$ où $(\mathcal{V}, \mathcal{E})$ est l'ensemble des noeuds et des arêtes du graphe. $\ell_f : \mathcal{V} \rightarrow \Omega_f$ est une fonction qui associe à chaque noeud $v_i \in \mathcal{V}$ un feature $a_i \stackrel{\text{def}}{=} \ell_f(v_i)$ dans un espace métrique (Ω_f, d). Nous appelons *information de feature* l'ensemble de tous les features $(a_i)_i$ du graphe. De même, $\ell_s : \mathcal{V} \rightarrow \Omega_s$ associe à un nœud v_i du graphe un point $x_i \stackrel{\text{def}}{=} \ell_s(v_i)$ appartenant à un espace métrique (Ω_s, C) spécifique à chaque graphe. $C : \Omega_s \times \Omega_s \rightarrow \mathbb{R}$ est une application symétrique qui vise à mesurer la similarité entre les nœuds du graphe. Cependant, contrairement à l'espace des features, Ω_s est implicite et, en pratique, il suffit de connaître la mesure de similarité C. Avec un léger abus de notation, C est utilisé pour désigner à la fois la mesure de similarité de la structure et la matrice qui encode cette similarité entre les paires de nœuds du graphe $(\mathbf{C}(i, k) = C(x_i, x_k))_{i,k}$. Selon le contexte, $\mathbf{C}$ peut soit encoder les informations de voisinage des nœuds, les informations des arrêtes du graphe, ou plus généralement, il peut modéliser une distance entre
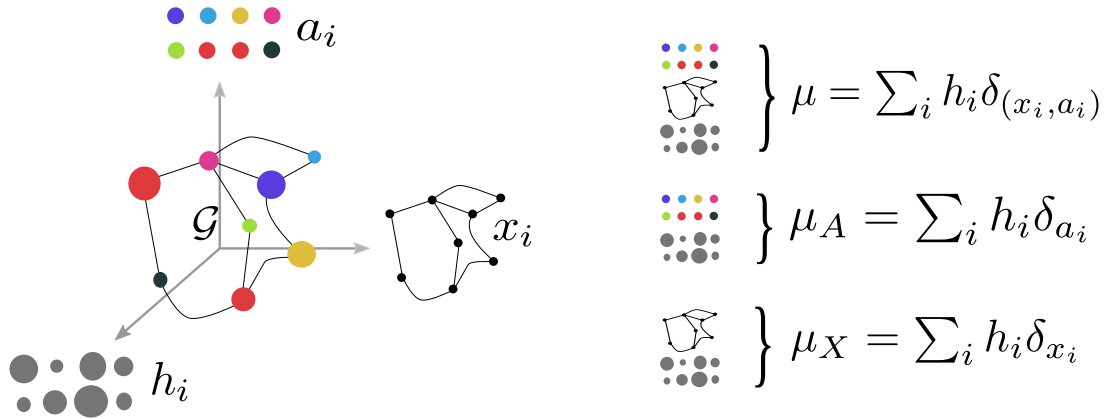
Figure 3: **(à gauche)** Graphe labelé avec $(a_i)_i$ l'information de feature, $(x_i)_i$ l'information de structure et $(h_i)_i$ un vecteur de poids qui mesure l'importance relative des noeuds. **(à droite)** Les données structurées sont entièrement décrites par une mesure de probabilité μ sur l'espace produit des features et de la structure, avec respectivement des marginales μ_X et μ_A sur la structure et les features.

les nœuds telle que la distance shortest-path. Nous désignons par *information de structure* l'ensemble de tous les points de structure $(x_i)_i$ du graphe.

Nous proposons d'enrichir le graphe décrit précédemment avec un vecteur de poids qui a pour but d'encoder l'importance relative des noeuds du graphe. Pour ce faire, si nous supposons que le graphe a n noeuds, nous associons aux noeuds des poids $(h_i)_i \in \Sigma_n$. Par cette procédure, nous obtenons la notion de *données structurées* comme un tuple $\mathcal{S} = (\mathcal{G}, h_{\mathcal{G}})$ où $\mathcal{G}$ est un graphe labelé et $h_{\mathcal{G}}$ est une fonction qui associe un poids à chaque noeud. Cette définition permet au graphe d'être représenté par une mesure de probabilité $\mu = \sum_{i=1}^{n} h_i \delta_{(x_i, a_i)}$ sur l'espace produit feature/structure qui décrit l'ensemble de la donnée structurée (voir Figure 3).

Considérons à présent deux données structurées $\mu = \sum_{i=1}^{n} h_i \delta_{(x_i, a_i)}$ et $\nu = \sum_{i=1}^{m} g_j \delta_{(y_j, b_j)}$, où $\mathbf{h} \in \Sigma_n$ et $\mathbf{g} \in \Sigma_m$ sont des histogrammes. Nous notons $\mathbf{M_{AB}} = (d(a_i, b_j))_{i,j}$ la matrice $n \times m$ de distance entre les features et $\mathbf{C}_1, \mathbf{C}_2$, les matrices de structure des graphes.

Nous définissons une nouvelle distance de transport optimal appelée la distance de Fused Gromov-Wasserstein. Elle est donnée pour un paramètre $\alpha \in [0, 1]$ par :

$$FGW(\mathbf{C}_1, \mathbf{C}_2, \mathbf{h}, \mathbf{g}) = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{h}, \mathbf{g})} \sum_{i,j,k,l} (1-\alpha) d(a_i, b_j)^q + \alpha |C_1(i,k) - C_2(j,l)|^q \pi_{i,j} \pi_{k,l} \tag{1}$$

Nous prouvons que cette fonction définit bien une distance entre les graphes labelés (Theorem 3.3.1) et qu'elle permet aussi de donner naissance à une notion de barycentre, de moyenne, de graphes basée sur la moyenne de Fréchet (Section 3.3.2). Nous établissons un algorithme pour calculer ces différents objets (Section 3.3.3) qui nous permet en particulier de résoudre aussi le problème de Gromov-Wasserstein classique, et nous prouvons qu'elle s'avère utile dans de nombreux scénarii de machine learning sur les graphes comme la classification, la simplification de graphes ou encore le clustering de graphes (Section 3.4). Nous concluons cette partie en élargissant la définition précédente au cas de données structurées continues (Section 3.5) où nous montrons que FGW possède des propriétés de distance similaires et définit de plus une géodésique sur l'espace des données structurées.

Chapitre 4

Ce Chapitre vise à combler l'écart entre la théorie classique du transport optimal et la théorie de Gromov-Wasserstein. Il s'ouvre sur le cas particulier des distributions 1D dont l'étude est basée sur les travaux de l'article [Vayer 2019b]. Nous établissons la première closed-form de Gromov-Wasserstein dans le cas des mesures discrètes sur la droite réelle et du coût euclidien au carré. Nous prouvons que celle-ci peut être calculée rapidement avec une complexité en $O(n \log(n))$. En particulier nous montrons qu'un couplage optimal pour Gromov-Wasserstein est trouvé en considérant soit le couplage diagonal ou le couplage anti-diagonal lorsque les points sont triés. Nous proposons, en utilisant cette closed-form, une nouvelle divergence appelée Sliced Gromov-Wasserstein, à la manière de Sliced Wasserstein. Nous établissons ses propriétés (Theorem 4.1.3) et l'utilisons dans des scénarii tels que la comparaison de meshes 3D et les réseaux de neurones génératifs.

Une deuxième partie plus prospective se concentre sur la théorie de Gromov-Wasserstein pour les espaces euclidiens en la liant à la théorie classique du transport optimal. Nous abordons notamment la question de la régularité des plans de transport de Gromov-Wasserstein. Nous donnons des conditions nécessaires sous lesquelles nous pouvons prouver que le couplage optimal de Gromov-Wasserstein est supporté par une fonction déterministe, de la même manière que le couplage optimal pour le cas de Wasserstein avec un coût quadratique entre des mesures régulières (théorème de Brenier [Brenier 1991]). Pour cela nous considérons les cas où les mesures de distance ou similarité dans chaque espace $c_{\mathcal{X}}, c_{\mathcal{Y}}$ sont définies par les produits scalaires ou par des coûts quadratiques. Nous montrons que résoudre GW équivaut à résoudre conjointement un problème de transport linéaire et un problème d'alignement. Ainsi, la régularité des plans optimaux de GW peut être étudiée au travers de formulations équivalentes, plus simples à analyser. Cela nous permet également de construire des solutions algorithmiques pour GW sur des espaces euclidiens. En résumé :

(i) Dans la Section 4.2.2 nous considérons le cas où $c_{\mathcal{X}}, c_{\mathcal{Y}}$ sont définis par des produits scalaires dans chaque espace. A condition que la mesure de probabilité source soit régulière par rapport à la mesure de Lebesgue, nous donnons une condition suffisante pour l'existence d'un plan de transport optimal *déterministe*, *i.e.* supporté par une fonction déterministe T. Nous montrons que cette fonction est de la forme $\nabla u \circ \mathbf{P}$ où u est une fonction convexe et $\mathbf{P}$ est une application linéaire qui peut être considérée comme une transformation globale "recalant" les mesures de probabilité dans le même espace (Theorem 4.2.1). Nous utilisons cette formulation pour montrer que la distance GW entre les mesures de probabilité 1D admet une solution closed-form. Plus précisément, nous montrons que le couplage optimal est déterminé par les fonctions de distribution cumulative et anticumulative de la distribution source (Theorem 4.2.4).

(ii) Dans la Section 4.2.3 nous considérons $c_{\mathcal{X}}, c_{\mathcal{Y}}$ définis comme étant le carré des distances euclidiennes dans chaque espace. Nous montrons que le problème équivaut à une maximisation d'une fonction convexe sur $\Pi(\mu, \nu)$. Nous utilisons la dualité Fenchel-Legendre dans l'espace des mesures pour en déduire un problème équivalent à celui de Gromov-Wasserstein (Theorem 4.2.5). Nous l'analysons plus en profondeur et montrons que la régularité des plans de transport optimaux est plus compliquée à établir que dans le cas précédent.

(iii) Dans la Section 4.2.4 nous utilisons les formulations précédentes pour obtenir des solutions numériques efficaces pour le problème GW basées sur des BCD. Nous montrons que ces algorithmes se comparent favorablement par rapport aux solveurs standards tels que le gradient conditionnel ou avec la régularisation entropique.
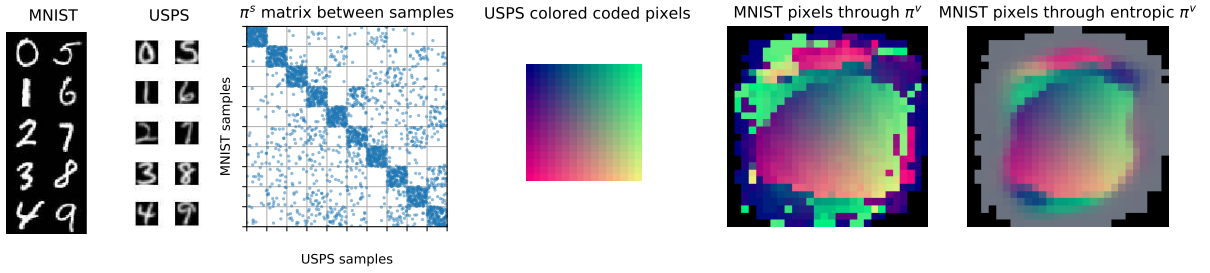
Figure 4: Illustration de COOT entre les datasets MNIST et USPS. **(gauche)** samples des datasets MNIST et USPS ; **(centre gauche)** Matrice de transport π^s entre les samples, triés par classe ; **(centre)** Image USPS avec les pixels colorés en respectant leur position 2D ; **(centre droite)** couleurs transportées sur image MNIST en utilisant π^v, les pixels noirs correspondent aux pixels MNIST non informatifs (toujours à 0); **(droite)** couleurs transportées sur image MNIST en utilisant π^v avec régularisation entropique.

(iv) Nous concluons par la Section 4.2.5 en considérant le problème *Gromov-Monge* dans les espaces euclidiens, qui est l'équivalent du problème de Monge du transport linéaire dans le contexte de Gromov-Wasserstein. Nous discutons du cas particulier de Gromov-Monge entre les mesures gaussiennes et nous montrons que ce problème admet une closed-form quand on se limite aux push-forward linéaires (Theorem 4.2.6). Nous donnons des interprétations géométriques de ce résultat et nous comparons le push-forward optimal avec la celui de la théorie du transport optimal classique dans le cas des mesures gaussiennes.

Chapitre 5

Ce chapitre présente un nouveau cadre théorique pour comparer des mesures de probabilité sur des espaces incomparables, à savoir le problème de co-transport optimal. Contrairement à l'approche de Gromov-Wasserstein, cette approche permet d'optimiser simultanément deux couplages entre les samples et les features des données. Ce chapitre fournit une analyse théorique approfondie de ce cadre et, du point de vue des applications, il aborde le problème de l'adaptation des domaines hétérogènes et du co-clustering. Ce chapitre est basé sur l'article [Redko 2020].

Plus précisément nous considérons deux datasets quelconques $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ et $\mathbf{X}' = [\mathbf{x}'_1, \ldots, \mathbf{x}'_{n'}]^T \in \mathbb{R}^{n' \times d'}$, avec en général $n \neq n'$ et $d \neq d'$. Les lignes sont appelées *samples* et les colonnes *features*. Nous associons aux samples $(\mathbf{x}_i)_{i \in [\![n]\!]}$ et $(\mathbf{x}'_i)_{i \in [\![n']\!]}$ des poids $\mathbf{w} = [w_1, \ldots, w_n]^\top \in \Sigma_n$ et $\mathbf{w}' = [w'_1, \ldots, w'_{n'}]^\top \in \Sigma_{n'}$. De la même manière on associe aux features des poids $\mathbf{v} \in \Sigma_d$ et $\mathbf{v}' \in \Sigma_{d'}$. Le problème de Co-transport optimal est défini par:

$$COOT(\mathbf{X}, \mathbf{X}', \mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}') = \min_{\substack{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}') \\ \boldsymbol{\pi}^v \in \Pi(\mathbf{v}, \mathbf{v}')}} \quad \sum_{i,j,k,l} L(X_{i,k}, X'_{j,l}) \boldsymbol{\pi}^s_{i,j} \boldsymbol{\pi}^v_{k,l} \tag{2}$$

Pour illustrer cette définition, nous résolvons le problème d'optimisation entre deux datasets classiques : MNIST et USPS. Ils contiennent des images de différentes résolution (les images de USPS sont de taille 16×16 et MNIST 28×28) qui appartiennent aux mêmes classes (chiffres entre 0 et 9). En outre, les chiffres sont également différemment centrés comme l'illustrent les exemples de la partie gauche de la Figure 4. Cela signifie que sans prétraitement, les images ne se trouvent pas dans le même espace topologique et donc ne peuvent pas être comparées directement à l'aide des distances conventionnelles. Les images représentent les *samples* tandis que chaque pixel agit comme un *feature* conduisant à 256 et 784 features

pour USPS et MNIST

Le résultat de la résolution du problème est reporté sur la Figure 4. Dans la partie centrale gauche, nous fournissons le couplage optimal $\boldsymbol{\pi}^s$ entre les samples, *i.e.* les différentes images, triés par classe. Le couplage $\boldsymbol{\pi}^v$, à son tour, décrit les relations entre les features, *i.e.* les pixels, dans les deux domaines. Pour le visualiser, nous codons en couleur les pixels de l'image USPS source et utilisons $\boldsymbol{\pi}^v$ pour transporter les couleurs sur une image MNIST cible de sorte que ses pixels soient définis comme des combinaisons convexes de couleurs de la première avec des coefficients donnés par $\boldsymbol{\pi}^v$. Les résultats correspondants sont présentés dans la partie droite de la Figure 4.

Nous montrons que cette nouvelle formulation inclut Gromov-Wasserstein comme cas particulier (Proposition 5.5.2), et qu'elle présente l'avantage de travailler directement sur les données brutes sans avoir à calculer, stocker et choisir les mesures de similarité. De plus, COOT fournit deux couplages interprétables entre les features et les samples. Nous montrons que COOT définit une notion de distance entre les datasets (Proposition 5.3.1) et nous établissons une procédure d'optimisation basée sur la résolution de transports linéaires (Section 5.4). Sur le plan pratique, nous apportons la preuve de l'utilité de COOT pour le machine learning notamment en adaptation de domaines hétérogènes (Section 5.6.1) et pour le co-clustering (Section 5.6.2).

Contents

Linear algebra

$\mathbf{a}, \mathbf{A}$	all vectors in $\mathbb{R}^d$ and matrices are written in bold. The coordinates will be written $A_{i,j}$ or $A(i,j)$ for matrices and a_i for vectors without bold.
$[\![n]\!]$	the subset $\{1, \cdots, n\}$ of $\mathbb{N}$
$\mathbf{I_n}, \mathbf{J_n}$	the $n \times n$ identity matrix and anti-identity matrix
$\|.\|, \langle, \rangle$	a norm (depends on the context) and an inner product
ℓ_p	Denotes the standard $\|.\|_p$ norm
$\otimes_K$	is the Kronecker product of matrices, *i.e.* for two matrices $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{B}^{p \times q}$, $\mathbf{A} \otimes_K \mathbf{B} \in \mathbb{R}^{np \times mq}$ defined by $(A_{i,j}\mathbf{B})_{i,j}$
$\otimes$	the the tensor-matrix multiplication, *i.e.* for a tensor $\mathbf{L} = (L_{i,j,k,l})$ and a matrix $\mathbf{A}$, $\mathbf{L} \otimes \mathbf{A}$ is the matrix $\left(\sum_{k,l} L_{i,j,k,l} A_{k,l} \right)_{i,j}$.
$\mathrm{tr}, \det$	is the trace operator for matrices *i.e.* $\mathrm{tr}(\mathbf{P}) = \sum_{i,j} P_{i,j}$ and the determinant operator
$\|.\|_{\mathcal{F}}$	is the Frobenius norm for matrices *i.e.* $\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{\mathrm{tr}(\mathbf{P}^T \mathbf{P})}$
$\langle ., . \rangle_{\mathcal{F}}$	is the inner product for matrices *i.e.* $\langle \mathbf{P}, \mathbf{Q} \rangle_{\mathcal{F}} = \mathrm{tr}(\mathbf{Q}^T \mathbf{P})$
$\mathrm{rank}(\mathbf{A}), \ker(\mathbf{A})$	The rank and the kernel of a matrix $\mathbf{A}$
DS	is the set of doubly-stochastic matrices, *i.e.* $DS = \{\mathbf{X} \in \mathbb{R}^{n \times n} \mid \mathbf{X}\mathbf{1}_n = \mathbf{1}_n, \mathbf{X}^T \mathbf{1}_n = \mathbf{1}_n, \mathbf{X} \geq 0\}$
Π_n	the set of permutation matrices, *i.e.* $\mathbf{X} = (x_{ij})_{(i,j) \in [\![n]\!]^2} \in \Pi_n$ if $x_{ij} \in \{0,1\}$ and $\sum_{i=1}^n x_{ij} = \sum_{j=1}^n x_{ij} = 1$.
$\mathcal{O}(p)$	the subset of $\mathbb{R}^{p \times p}$ of all orthogonal matrices.
$\mathbb{V}_p(\mathbb{R}^q)$	is the Stiefel manifold, *i.e.* the set of all orthonormal p-frames in $\mathbb{R}^q$ or equivalently $\mathbb{V}_p(\mathbb{R}^q) = \{\mathbf{B} \in \mathbb{R}^{q \times p} \mid \mathbf{B}^T \mathbf{B} = \mathbf{I_p}\}$.
S_n	the set of all permutations of $[\![n]\!]$.
$\oslash$	Element-wise division operator for two vectors, *i.e.* $\mathbf{u} \oslash \mathbf{v} = (\frac{u_i}{v_i})_i$
Σ_n	the set of probability vectors in $\mathbb{R}_+^n$ (or histograms with n bins), *i.e.* $\Sigma_n = \{\mathbf{a} \in \mathbb{R}_+^n \mid \sum_{i=1}^n a_i = 1\}$

Measure theory

$\mathcal{P}(\mathcal{X})$	the set of probability measures on a space $\mathcal{X}$
$\mathcal{M}(\mathcal{X})$	the set of Borel finite signed measures on a space $\mathcal{X}$
$\mathcal{L}$	the Lebesgue measure on $\mathbb{R}$ or $\mathbb{R}^n$ depending on the context
$\mathrm{supp}(\mu)$	the support of μ (see definition 2.1.4)
δ_x	the dirac measure on x, *i.e.* $\delta_x(y) = 0$ if $y \neq x$ else 1
$\#$	the push forward operator.
$\mu \otimes \nu$	the product measure of two probability measures μ, ν, *i.e.*, $\mu \otimes \nu(A \times B) = \mu(A)\nu(B)$
$(id \times T)\#\mu$	Is the measure $\mu \otimes (T\#\mu)$
$\Pi(\mu, \nu)$	the set of couplings of two probability measures μ, ν (see definition 2.1.1)
$M_c(\mu, \nu)$	The Monge cost between two probability measures μ, ν (see definition MP)
$\mathcal{T}_c(\mu, \nu)$	The Kantorovitch cost between two probability measures μ, ν (see definition KP)
$W_p(\mu, \nu)$	The p-Wasserstein distance between two probability measures μ, ν
$GW_p(\mu, \nu)$	The p-Gromov-Wasserstein distance between two probability measures μ, ν
$\mathcal{F}_\mu$	the Fourier transform of the probability measure μ *i.e.* $\mathcal{F}_\mu(s) = \int e^{-2i\pi\langle s,x \rangle} d\mu(x)$
$\mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$	multivariate Gaussian (normal) distribution with mean $\mathbf{m}$ and covariance $\mathbf{\Sigma}$

Functions

DT	the Jacobian of a function T
$C(\mathcal{X}), C(\mathcal{X}, \mathcal{Y})$	the set of continuous functions form $\mathcal{X}$ to $\mathbb{R}$ (*resp.* from $\mathcal{X}$ to $\mathcal{Y}$)
$C_b(\mathcal{X}), C_b(\mathcal{X}, \mathcal{Y})$	the set of continuous and bounded functions form $\mathcal{X}$ to $\mathbb{R}$ (*resp.* from $\mathcal{X}$ to $\mathcal{Y}$)
$C^p(\mathcal{X})$	the set of function of class C^p.
$Lip_k(\mathcal{X})$	the set of Lipschitz functions on $\mathcal{X}$ with Lipschitz constant k
$L^p(\mu)$	The set of p-integrable functions with respect to a measure μ, *i.e.* $f \in L^p(\mu)$ if $\int |f|^p \mathrm{d}\mu < +\infty$
∇	the gradient operator
$d_{\mathcal{X}} \oplus d_{\mathcal{Y}}$	The "sum" distance on the cartesian product $(\mathcal{X}, d_{\mathcal{X}}) \times (\mathcal{Y}, d_{\mathcal{Y}})$. More precisley it is defined by $d_{\mathcal{X}} \oplus d_{\mathcal{Y}}((x, x'), (y, y')) = d_{\mathcal{X}}(x, x') + d_{\mathcal{Y}}(y, y')$.
$\mathbb{S}^{d-1}$	The d-dimensional hypersphere, *i.e.* $\mathbb{S}^{d-1} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\|_2 = 1 \right\}$

Acronyms

OT	Is the acronym for Optimal Transport
W,SW,GW, SGW	Stands respectively for Wasserstein, Sliced Wasserstein, Gromov-Wasserstein and Sliced Gromov-Wasserstein
QAP,QP,BAP,BP	Stands respectively for Quadratic Assignment Problem, Quadratic Program, Bilinear Assignment Problem, Bilinear Program

CHAPTER 1

Introduction

Contents

1.1 Moving probability measures: a least effort problem

"Thus, we see in science, sometimes brilliant but for a long time useless theories, suddenly becoming the basis of the most important applications, and sometimes seemingly very simple applications, giving rise to the idea of abstract theories that are not yet needed, directing towards the theories of the Surveyors' work, and opening up a new career for them" [1]. That is how Nicolas De Condorcet [De Condorcet 1781] in the 18th century introduced the work of Gaspard Monge [Monge 1781] which is at the core of the Optimal Transport (OT) theory. How to move some masses from one location to another so as to minimize the overall effort? Condorcet was right: this "simple idea" has evolved over the years to become a theory at the crossroads of mathematics/optimization and is today at the center of many machine learning applications.

Broadly speaking the interest of Optimal Transport lies in both its ability to provide correspondences between sets of points and its ability to induce a geometric notion of distance between probability distributions (see Figure 1.1) Both have proved to be very useful for a wide range of tasks that are, to name a few, image registration [Haker 2001], image retrieval [Rubner 1998], domain adaptation [Courty 2017], signal processing [Kolouri 2017], unsupervised learning [Arjovsky 2017, Genevay 2018], supervised and semi-supervised learning [Frogner 2015, Solomon 2014], natural language processing [Kusner 2015], fairness [Gordaliza 2019], in biology [Schiebinger 2019] or in astrophysics [Frisch 2002].

Despite its many properties, the optimal transport problem remains difficult to solve in practice and is known to suffer from scalability problems that prevent its use on large, ubiquitous data in machine learning. The emergence of optimal transport in the machine learning community has been greatly favoured by recent achievements on the optimization side [Cuturi 2013, Altschuler 2017, Genevay 2016] which tend to circumvent the heavy computational complexity of solving OT problems.

---

[1] Original quote: "Ainsi, l'on voit dans les Sciences, tantôt des théories brillantes, mais longtemps inutiles, devenir tout à coup le fondement des applications les plus importantes, et tantôt des applications très simples en apparence, faire naître l'idée de théories abstraites dont on n'avait pas encore le besoin, diriger vers les théories des travaux des Géomètres, et leur ouvrir une carrière nouvelle."
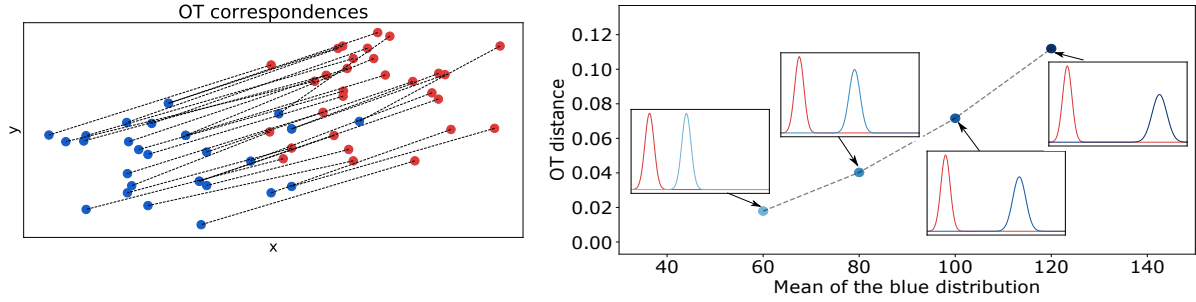
Figure 1.1: Optimal Transport provides tools for finding correspondences between set of point and gives a geometric notion of distance between probability distributions. **(left)** Two sets of points in 2D. The correspondences found by OT is depicted in dashed line. **(right)** OT distance between two probability distributions in red and blue. The further away the mean of the red probability distribution is from the blue one the higher is the OT distance.

Moreover optimal transport, in its early formulation, is quite restricted to applications where there exists a *direct way* of comparing the samples of the data. Its applicability is thus often limited to the case where the samples are part of a common ground metric space, that is most of the time Euclidean. This limitation prevents its use for a variety of machine learning tasks where there is an additional *structural information* on the data which can not usually be described in the Euclidean setting *e.g.* when the samples are described by graphs, trees or time series. It also prevents the use of optimal transport when the samples lie in different, seemingly not related, metric spaces or when a meaningful notion of distance between the samples can not be easily defined. All of these instances can be framed into the *incomparable* setting, that is when the samples lie on *incomparable spaces*. An interesting remedy in this situation can be found in the theory of the Gromov-Wasserstein distance [Memoli 2011] which does not require the comparison of the samples *across* the distributions. However, it is well-known to be arduous to solve and to suffer from tedious scalability issues. The purpose of this thesis is to help overcoming these obstacles by:

(i) Defining new Optimal Transport frameworks for incomparable spaces and especially for structured data.

(ii) Reducing the theoretical gap between classical Optimal Transport and the Gromov-Wasserstein theory in Euclidean spaces and in particular derive scalable and applicable formulations.

1.2   Structured Data and Incomparable Spaces in Machine Learning

Before going into the details of this thesis, it is important to make explicit what is behind the notion of *structure* through the manuscript. As a starter we can see the *structural information* as the piece of information which encodes the specific relationships that exist among the components of an object. This definition can be related with the concept of *relational reasoning* [Battaglia 2018] where some *entities* (or elements with attributes such as the intensity of a signal) coexist with some *relations* or properties between them.

Natural instances of such structured data arise when the structure is *explicit*. For example, in a graph context, edges are representative of this notion so that each attribute of the graph (typically $\mathbb{R}^d$ vectors) may be linked to some others through the edges between the nodes. Notable examples are found

in chemical compounds or molecules modeling [Kriege 2016], brain connectivity [Ktena 2017], or social networks [Yanardag 2015]. This generic family of structured data also encompasses trees [Day 1985] or time series where the signals' values are correlated through time so that comparing times series requires one to take the direction of time into account.

Structure information of data in machine learning can also be more subtle or even *implicit.* For example it can appear where one build structural priors, or *inductive bias* [Battaglia 2018], on the objects representation. For instance in the context of deep learning, many successful architectures exploit the equivariance to a symmetry transformation to improve generalization. The convolutional neural network (CNN) is a prime example of this "built-in" inductive bias which satisfies translation equivariance, *i.e.* if we translate the input, the output of the convolutions will also be translated. This inductive bias is known to reveal useful spatial hierarchy, or structure, on the pixels [Wang 2018, Chen 2018] and other works have studied designing layers with equivariances to other transformations such as permutation, rotation, reflection [Kondor 2018, Cohen 2016, Gens 2014]. Unlike end-to-end methods other more "hand-engineering" approaches based *e.g.* on images segmentation can leverage some structure in images that can further be usefully exploited [Bach 2007, Jianbo Shi 2000]. Implicit structure is also at the core of many natural language processing (NLP) tools used to find good word representations [Mikolov 2013b, Mikolov 2013a, Pennington 2014]. It is also the main ingredient of speech recognition [Hinton 2012] or sequence learning [Sutskever 2014]. In these cases structural assumptions on the sequences of words are made whether by the mean of latent variables or *via* conditional probabilities. When available, labels or classes also induce an implicit structure on the feature space of the data. For instance in Domain Adaptation one may desire the source samples with the same label to be matched consistently within the same region of the target space preventing them from being split into disjointed far locations [Courty 2017, Alvarez-Melis 2018b]. The recent trend of graph neural networks (GNN) [Wu 2020] in the machine learning community is one of the many examples emphasizing that structured data remains an important and challenging setting nowadays. While previous instances consider structured data as inputs of the learning process the prevalence of this notion in machine learning also arises in another line of works where it is an output. This setting is *e.g.* considered in the structured prediction approach where one wants to learn to produce structured object such as sequences, trees or assignments [Taskar 2005, Lafferty 2001, Collins 2002, Blondel 2020, Mensch 2018, Korba 2018].

In short, the notion of *structure* in machine learning is omnipresent and appears as often as there is an additional information about the objects that goes beyond their feature representations. As shown in many contexts in machine learning such as graphical models [Pearl 1986, Pearl 2009], relational reinforcement learning [D**v**zeroski 2001] or Bayesian nonparametrics [Hjort 2010], considering objects as a complex composition of entities together with some interactions is crucial in order to learn from small amounts of data.

The previous notion of structure data can be seen as a special case of data defined on incomparable spaces. Informally in this situation, each sample has its own "characteristic" that may not be shared with the other samples. For instance when considering a dataset of multiple graphs (each graph being a data point) the structure of one graph is usually not shared among the other graphs. This notion, purposely broad, encompasses also the case where data come from heterogeneous sources. As a particular instance of this problem Heterogeneous Domain Adaptation [Yeh 2014, Zhou 2014, Wang 2011] aims to exploit knowledge from heterogeneous source domains to improve the learning performance in a target domain with, potentially, different feature spaces between the source and target domains. The MNIST/USPS [LeCun 2010, Hull 1994] case is a prime example of this situation: based on the knowledge of 28×28 digit images (*i.e.* $\mathbb{R}^{784}$ vectors) from MNIST how to build *e.g.* a classifier that works well on
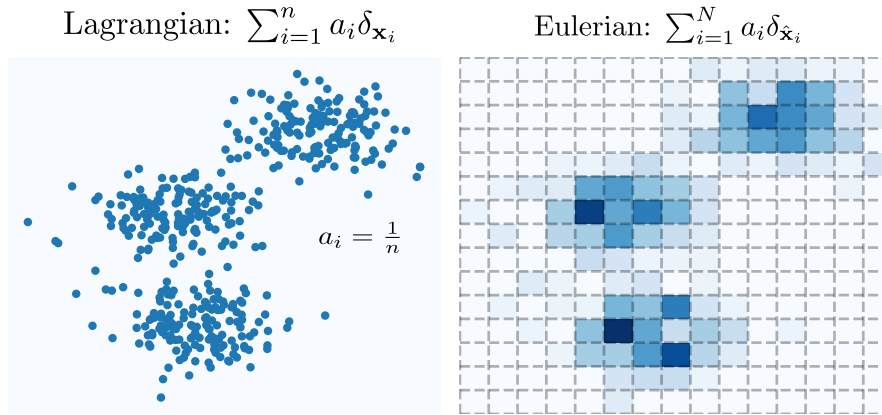
Figure 1.2:   We associate to a dataset $(\mathbf{x}_i)_{i \in [\![n]\!]}$ a probability measure describing the dataset. In this example $\mathbf{x}_i \in \mathbb{R}^2$. **(left)** Lagragian formulation (points clouds): $\sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$ is a discrete probability measure where $(a_i)_{i \in [\![n]\!]}$ is a probability vector, *i.e.* $a_i \geq 0$ and $\sum_{i=1}^n a_i = 1$ and δ is the dirac measure $\delta_{\mathbf{x}_i}(\mathbf{x}) = 1$ if $\mathbf{x} = \mathbf{x}_i$ else 0. **(right)** Eulerian formulation (histograms): $\sum_{i=1}^N a_i \delta_{\hat{\mathbf{x}}_i}$ where $(\hat{\mathbf{x}}_i)_{i \in [\![N]\!]}$ is a regular grid on $\mathbb{R}^2$ and $(a_i)_{i \in [\![N]\!]}$ is a probability vector.

16×16 digit images (*i.e.* $\mathbb{R}^{256}$ vectors) of USPS? Needless to say this problem often arises in all the fields of machine learning: it is common that the data are gathered from heterogeneous sources in practice and methods that build upon this diversity are often of high interest.

1.3   Motivating Optimal Transport

The central question that often arises in machine learning is: how to represent data and how to compare them? The framework of probability distributions provides an answer to this query by associating a probability measure to a collection of samples that forms a dataset $(\mathbf{x}_i)_{i \in [\![n]\!]}$. The *Lagragian* representation of the dataset results in a discrete probability measure $\sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$ in which one associates to each point $\mathbf{x}_i$ a dirac $\delta_{\mathbf{x}_i}(\mathbf{x}) = 1$ if $\mathbf{x} = \mathbf{x}_i$ otherwise 0 as well as a weight $a_i \geq 0$ such that $(a_i)_{i \in [\![n]\!]}$ is a probability vector which satisfies $\sum_{i=1}^n a_i = 1$. When no information about the relative importance of the samples in the dataset is available the weights can be chosen as uniform so that $a_i = \frac{1}{n}$. Similarly a *Eulerian* representation can be constructed *via* the probability distribution $\sum_{i=1}^N a_i \delta_{\hat{\mathbf{x}}_i}$ in which $(\hat{\mathbf{x}}_i)_{i \in [\![N]\!]}$ is a regular grid on the space. This formulation produces an histogram of our data (see Figure 2). This point of view on the data advocates finding an appropriate way of comparing their representation as probability distributions and, as such, the question of finding adequate measures of "how far" are two probability distributions is at the core of many machine learning algorithms. Although various divergences exist such as ϕ-divergences [Csiszar 1975] or Maximum Mean Discrepancies (MMD) [Gretton 2007], the richness of optimal transport lie in its ability to incorporate the geometry of the underlying space in its formulation and to pay attention to the relations, the correspondences of the samples within their respective representations. To highlight in short the benefit of representing data through probability

distributions coupled with OT we can cite [Rubner 2000] for image retrieval or [Kusner 2015] for natural language processing. At this point natural questions arise: is this framework applicable when the nature of the data is inherently structured or when the different data points lie in incomparable spaces? In this case how can we represent data as probability distributions? To what extent this representation is valuable? Is the Optimal transport framework still applicable, and, if not, how do we compare these probability distributions? The purpose of this thesis, *inter alia*, is to give some answers to these questions.

1.4 Outline of the Thesis

This thesis covers mostly all the author's work conducted and focus on a single line of research that is *Optimal Transport on Incomparable Spaces*. Additional line of works [Vayer 2020a] on time series on incomparable spaces, which is not based on optimal transport, is not included in this thesis but the interested reader can find the details in the bibliography. The rest of the thesis is hinged so that all chapters can be read separately in any order except for Chapter 2 that provides all the mathematical background and tools used in the other chapters.

Chapter 2 sets up the mathematical and numerical background of optimal transport. It presents the fundamental results of classical optimal transport theory and summarizes/illustrates its different formulations as well as some well-known solvers. The philosophy of this chapter is to provide a high-level overview of OT both in theory and practice. This chapter concludes with the Gromov-Wasserstein theory which is at the core of thesis. A reader familiar with the basic concepts of optimal transport may skip this part although it contains crucial concepts and notations that will be discussed throughout the thesis.

Chapter 3 is dedicated to optimal transport for structured data, and especially in the context of graphs. It is based on the works of the articles [Vayer 2019a] and [Vayer 2020b] and gives some answers to the question of defining a mathematical framework for optimal transport in the case of structured data. A general framework for this setting is given, based on the Fused Gromov-Wasserstein distance that defines a OT distance between structured data such as undirected graphs, and applied on real world graph data applications.

Chapter 4 aims at bridging the gap between the classical optimal transport theory and the Gromov-Wasserstein theory. The chapter opens with the special case of 1D distributions which results in the Sliced Gromov-Wasserstein formulation based on the works in [Vayer 2019b]. A second more prospective part focuses on the Gromov-Wasserstein theory for Euclidean spaces and connects with the classical optimal transport theory by questioning the regularity of Gromov-Wasserstein optimal transport plans.

Chapter 5 presents a new framework for comparing probability measures on incomparable spaces, namely the CO-Optimal transportation problem. Contrary to the Gromov-Wasserstein approach this approach simultaneously optimizes two transport maps between both samples and features of the data. This chapter provides a thorough theoretical analysis of this framework and, from an application side, this work tackles the problem of Heterogeneous Domain Adaptation and co-clustering/data summarization. This chapter is based on the article [Redko 2020].

Generality about optimal transport

Contents

Optimal transport is a long-standing mathematical problem whose theory has matured over the years. A good gateway for this theory can be found in [Santambrogio 2015]. A more mathematical oriented overview can be found in [Villani 2008] while the most complete document about numerical aspects of OT can be found in [Peyré 2019]. The objective of this chapter to present in short the main results of the "classical" OT theory both mathematically and numerically. We will discuss in the last part of this chapter the theory related to the Gromov-Wasserstein transportation problem for which we refer the reader to [Memoli 2011, Sturm 2012] for its foundation.

2.1   Linear Optimal Transport theory

The Monge problem   The OT problem has been historically introduced by Gaspard Monge [Monge 1781] and can be described as the following "least effort problem": given two probability distributions μ and ν how do we *transfer* all the probability mass of μ onto ν so that the overall *effort* of transferring this mass is minimized? Originally the idea was to move dirt (déblais) to one place to another (remblais) in the most efficient way.

For properly defining this problem we need to define the notions of *transfer* and *effort*: the former can be expressed through the notion of *push-forward* and the latter through the notion of *cost*. More precisely

and given two Polish spaces $\mathcal{X}, \mathcal{Y}$[1] and two probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ a cost is a function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{+\infty\}$ which values $c(x, y)$ aim at measuring how far is $x \in \mathcal{X}$ from $y \in \mathcal{Y}$ and quantifies somehow the "effort" of moving x forward to y. The *push-forward*[2] of a probability measure μ through a function $T : \mathcal{X} \to \mathcal{Y}$ is defined as the probability measure $T\#\mu \in \mathcal{P}(\mathcal{Y})$ which satisfies equivalently one of the two following conditions:

(i) $T\#\mu(A) = \mu(T^{-1}(A)) = \mu(\{x \in \mathcal{X} \mid T(x) \in A\})$ for every measurable set A

(ii) $\int_{\mathcal{Y}} \phi(y)\mathrm{d}(T\#\mu)(y) = \int_{\mathcal{X}} \phi \circ T(x)\mathrm{d}\mu(x)$ for every measurable function ϕ.

These conditions simply state that we transform, or *push*, the probability measure μ thanks to T so as to create another probability measure on $\mathcal{Y}$. When we consider a discrete probability distribution $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ the push-forward measure $T\#\mu$ is simply defined by $T\#\mu = \sum_{i=1}^{n} a_i \delta_{T(x_i)}$.

As described at the beginning, one wants to move the *source* distribution μ forward to the *target* distribution ν. This translates mathematically as finding a map T which satisfies $T\#\mu = \nu$. When we consider the Euclidean setting and when the probability measures have densities f, g with respect to the Lebesgue measure by the change of variable formula the push-forward condition writes:

$$g(T(\mathbf{x})) \det(DT(\mathbf{x})) = f(\mathbf{x}) \tag{2.1}$$

where DT stands for the Jacobian of T. Among all these possible push-forwards, OT aims at finding the map T which minimizes the total cost of having moved μ forward to ν that is $\int_{\mathcal{X}} c(x, T(x))\mathrm{d}\mu(x)$. Overall the problem of Monge (MP) can be formulated as the following non-convex optimization problem:

$$M_c(\mu, \nu) = \inf_{T\#\mu = \nu} \int_{\mathcal{X}} c(x, T(x))\mathrm{d}\mu(x). \tag{MP}$$

In general finding such optimal map T of the (MP) problem is quite difficult to solve since the solution may not be unique and may not even exists (see Figure 2.1). Even in the regular setting where μ, ν have densities, equation (2.1) is highly non-linear in T which is one of the major difficulty preventing from an easy analysis of the Monge Problem. As such the Monge problem remained an open question for many years and results about the existence and unicity of the optimal Monge map were limited to special cases until the works of Brenier [Brenier 1991] which implications will be detailed after.

Kantorovitch formulation   Major breakthroughs in the OT theory were made possible thanks to Kantorovitch [Kantorovich 1942] who proposes a relaxation of the (MP) problem. The key idea is to consider a *probabilistic mapping* instead of a deterministic map T to push the source measure forward to the target one. In the Kantorovitch formulation it is allowed to *split* the mass of the probability measures into pieces and transport them towards several targets points. This translates mathematically by replacing the push-forward of a measure by a probabilistic *coupling*:

Definition 2.1.1 (Couplings)**.** *Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. A coupling π of μ and ν is a probability distribution on $\mathcal{X} \times \mathcal{Y}$ such that both marginals of π are respectively μ and ν. More precisely π is part of the following set:*

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid A \subset \mathcal{X}, B \subset \mathcal{Y} \text{ measurable } \pi(A \times \mathcal{Y}) = \mu(A) ; \pi(\mathcal{X} \times B) = \nu(B)\} \tag{2.2}$$

---

[1]A Polish space $\mathcal{X}$ is a separable completely metrizable topological space.

[2]T is often called a map or Monge map in the OT literature
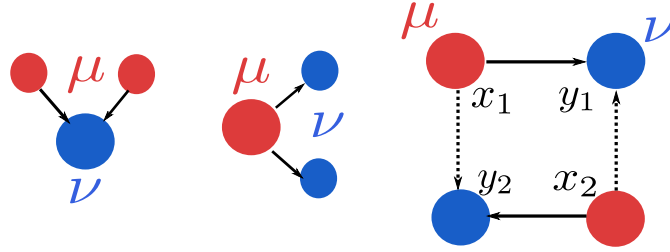
Figure 2.1: Push-forward between two discrete probability measures μ and ν in three *scenarii*. **(left)** μ is supported on x_1, x_2 with corresponding weights $\frac{1}{2}, \frac{1}{2}$ and ν on y_1 with weight 1. The only possible push-forward T is $T(x_1) = y_1$ and $T(x_2) = y_1$. **(center)** In this situation there is no push-forward of μ onto ν because no function can satisfies $T(x_1) = y_1$ and $T(x_1) = y_2$ when $y_1 \neq y_2$. The problem of Monge admits no solution in this case. **(right)** All points are equidistant form each other that is $c(x_i, y_j) = 1$ for $i, j \in [\![2]\!]^2$. In this case the solution of problem MP is not unique it may associate x_1 whether with y_1 or y_2 with the same overall cost (same with x_2).

Remark 2.1.1. *Unlike the set of push-forwards of μ forward to ν the set of couplings of two probability measures is always non-empty as the product measure $\mu \otimes \nu$ is in $\Pi(\mu, \nu)$.*

An important illustration of the former definition is when μ and ν are discrete probability measures. This situation will be omnipresent thorough the manuscript we propose to detail the notations in the following example:

Example 2.1.1 (The case of discrete probability measures). *Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \nu = \sum_{i=1}^{m} b_j \delta_{y_j}$ be discrete probability measures where $x_i \in \mathcal{X}$, $y_j \in \mathcal{Y}$ and $\mathbf{a} = (a_i)_{i \in [\![n]\!]} \in \Sigma_n, \mathbf{b} = (b_j)_{j \in [\![m]\!]} \in \Sigma_m$ are probability vectors which belong to the following probability simplex:*

$$\Sigma_n \overset{def}{=} \{\mathbf{a} \in \mathbb{R}_+^n \mid \sum_{i=1}^{n} a_i = 1\}. \tag{2.3}$$

We will use interchangeably the term histogram or probability vector for an element $\mathbf{a} \in \Sigma_n$. In this case a coupling $\boldsymbol{\pi}$ is a matrix of the following set:

$$\Pi(\mathbf{a}, \mathbf{b}) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{n \times m} \mid \boldsymbol{\pi} \mathbf{1}_m = \mathbf{a} ; \boldsymbol{\pi}^T \mathbf{1}_n = \mathbf{b}\}$$
$$= \{\boldsymbol{\pi} \in \mathbb{R}_+^{n \times m} \mid \forall (i, j) \in [\![n]\!] \times [\![m]\!], \sum_{j=1}^{m} \pi_{ij} = a_i ; \sum_{i=1}^{n} \pi_{ij} = b_j\} \tag{2.4}$$

Using the coupling instead of a deterministic map allows defining the OT problem for a very large class of probability measures under very mild assumptions. More precisely let $\mathcal{X}, \mathcal{Y}$ be Polish spaces and

---

Memo 2.1.1 (Lower semi-continuity). *On a metric space $(\mathcal{X}, d)$ a function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is said to be lower semi-continuous (l.s.c.) if for every sequence $x_n \to x$ we have $f(x) \leq \liminf f(x_n)$. Such functions have the following properties:*

- *If $(f_k)_k$ is a sequence of l.s.c. functions on $\mathcal{X}$ then $f = \sup_k f_k$ is l.s.c.*

- *If f is l.s.c. and bounded from below then there exists a sequence of continuous and bounded functions $(f_k)_{k \in \mathbb{N}}$ converging increasingly to f. We can also suppose that each f_k is k-Lipsichtz.*
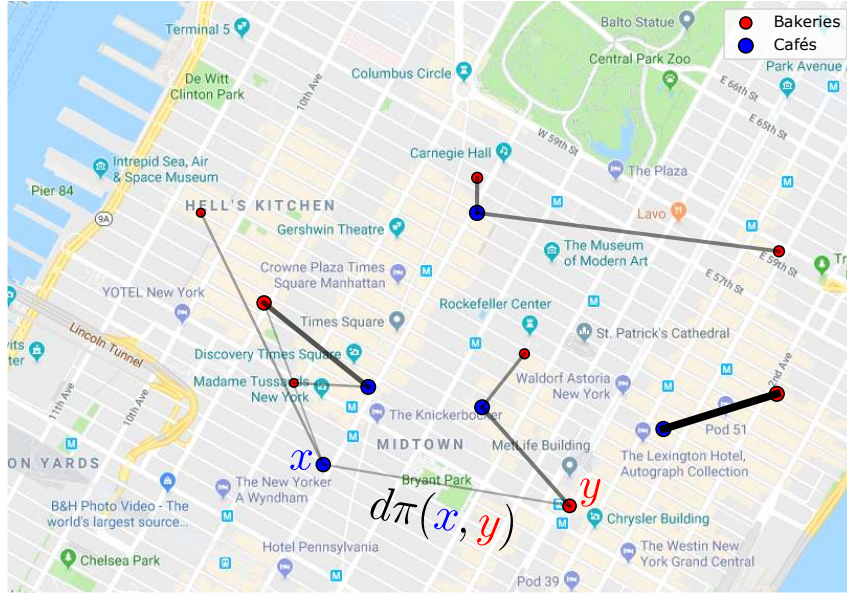
Figure 2.2: The bakery/cafés problem in Manhattan based on example 2.1.2. The bakeries (*resp* cafés) are in blue (*resp* red) and the dots' size denote the amount of bread disposable (*resp* needed). Lines represent the amount of bread transferred d$\pi(x,y)$: the wider the larger. The optimal transport plan is depicted in the Figure and represents the cheapest transport plan so that all breads are transferred from bakeries to cafés.

$\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. Given a cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \cup \{+\infty\}$ lower semi-continuous (see Memo 2.1.1), then Kantorovitch problem aims at finding:

$$\mathcal{T}_c(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x,y) \mathrm{d}\pi(x,y). \tag{KP}$$

The resulting cost, potentially infinite without further assumptions, corresponds to the minimal cost of moving μ forward to ν by splitting their masses and transporting forward the pieces according to the transport plan π. The good news about this formulation is that the infimum is always well defined providing that the cost is positive and lower semi-continuous (actually it suffices that c is bounded from below see [Santambrogio 2015, Theorem 1.7]). The problem KP is defined regardless of the nature of the probability distributions: they can be both discrete or continuous (see Figure 2.3). The problem appears to be *linear* in π so that we will denote (KP) as the *linear transportation problem*.

Relying on the Kantorovitch formulation (KP) appears to be very useful in order to find a solution of the Monge problem (MP). Indeed a push forward $T\#\mu = \nu$ induces a coupling $\pi = (id \times T)\#\mu$[3] then it is easy to verify that $\mathcal{T}_c(\mu,\nu) \le M_c(\mu,\nu)$. To find the converse inequality it suffices to find an optimal solution π^* of (KP) which is of the form $\pi^* = (id \times T^*)\#\mu$ where $T^*\#\mu = \nu$. In this case we would have proven that both problems are equal and that T^* is optimal for (MP). In other words if there is an optimal coupling supported on a *deterministic* function then both (KP) and (MP) are equivalent. We will see in next sections cases where we can ensure that necessarily the optimal coupling is of this form.

Example 2.1.2. *The famous bakery analogy of Villani's book [Villani 2003] provides a simple illustration of the linear OT problem. Suppose that someone is in charge of the distribution of bread from bakeries*

---

[3]$(id \times T)\#\mu$ is the measure $\mu \otimes T\#\mu$ or equivalently $\mathrm{d}(id \times T)\#\mu(x,y) = \mathrm{d}\mu(x)\mathrm{d}(T\#\mu(y))$
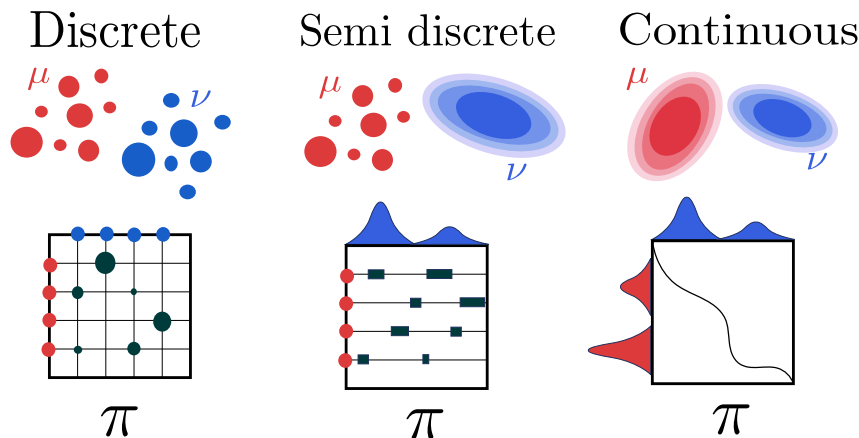
Figure 2.3: Different OT problems between **(left)** two discrete probability distributions, the coupling π is a matrix **(center)** one continuous and one discrete probability distribution, this is called the *semi-discrete* OT problem **(right)** two continuous distributions. Inspired from [Peyré 2019]

to cafés in Manhattan. The bakeries are located at some points y and the cafés at some points x distant from each other by $c(x, y)$. At 8 a.m sharp all the bread from the bakeries has to be transferred to the cafés in order for the citizens of Manhattan to have a good day. The company in charge of the distribution wants to route the breads from bakeries to cafés the cheapest way possible. This problem can be recast into a linear OT problem. Considering two distributions $\mu = $ "all available breads in bakeries" and $\nu = $ "all the demands in bread of cafés", the company seeks for a transport plan π such that $\mathrm{d}\pi(x, y)$ is the amount of breads transferred from bakery y to café x. The best transport plan minimizes the overall cost of moving all the breads from bakeries to cafés which is $\int c(x, y)\mathrm{d}\pi(x, y)$ (see Figure 2.2)

Wasserstein distance   The most notable scenario in many OT applications is when $\mathcal{X} = \mathcal{Y} = \Omega$ where (Ω, d) is a Polish space, *e.g.* an Euclidean space. In this case there is a natural way of defining the cost c since the space Ω is already endowed with a notion of distance between the points. In this situation we can define the so-called *p-Wasserstein distance* for $p \in [1, +\infty[$ as $W_p(\mu, \nu) = (\mathcal{T}_{d^p}(\mu, \nu))^{\frac{1}{p}}$ or precisely:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} d^p(x, y) \mathrm{d}\pi(x, y) \right)^{\frac{1}{p}} \tag{2.5}$$

The name is not misleading: this function satisfies all the axioms of distance on the space of probability distributions with bounded p-moments as stated in the next theorem (see [Villani 2008, Definition 6.4]):

Theorem 2.1.1 (The Wasserstein distance is a distance). *Let (Ω, d) be a Polish space, $p \in [1, +\infty[$ and:*

$$\mathcal{P}_p(\Omega) \stackrel{def}{=} \{ \mu \in \mathcal{P}(\Omega) \mid \int_\Omega d(x_0, x)^p \mathrm{d}\mu(x) < +\infty \} \tag{2.6}$$

with $x_0 \in \Omega$ arbitrary. Let $\mu, \nu \in \mathcal{P}_p(\Omega)$. Then $W_p(\mu, \nu) < +\infty$. Moreover:

(i) $W_p(\mu, \nu) = W_p(\nu, \mu)$ (symmetry)

> *(ii)* $W_p(\mu, \nu) = 0 \iff \mu = \nu$ *(identity of indiscernibles)*
>
> *(iii)* *Let* $\zeta \in \mathcal{P}_p(\Omega)^2$ *then* $W_p(\mu, \nu) \leq W_p(\mu, \zeta) + W_p(\zeta, \nu)$ *(triangle inequality)*

Example 2.1.3. *One direct example of Wasserstein distance is between two diracs supported on x,y. In this case* $W_p(\delta_x, \delta_y) = d(x, y)$ *which is quite the intuitive behavior, that is the more the dirac are located far from each others the larger their Wasserstein distance is.*

The distance property renders W_p a powerful tool for comparing probability measures. Another valuable feature of the former distance is that it gives a characterization of the weak convergence of probability measure. Informally a sequence of probability measures gets as close as possible to a probability measure μ if the Wasserstein distance tends to zero. The convergence is based on the following definition:

Definition 2.1.2 (Weak-convergence). *Let* $(\mu_n)_{n \in \mathbb{N}}$ *be a sequence of probability measures on* $\mathcal{X}$ *a Polish space. We say that* $(\mu_n)_{n \in \mathbb{N}}$ *converges weakly to* μ *in* $\mathcal{X}$ *if for all continuous and bounded functions* $f : \mathcal{X} \to \mathbb{R}$:

$$\int_{\mathcal{X}} f \, \mathrm{d}\mu_n \to \int_{\mathcal{X}} f \, \mathrm{d}\mu \tag{2.7}$$

The Wasserstein distance *metrizes the weak convergence* of probability measures, in other words $(\mu_n)_{n \in \mathbb{N}}$ converges weakly to μ if and only if $W_p(\mu_n, \mu) \to 0$ [Villani 2003, Theorem 6.9]. Note that other distances can be proposed to metrize the space of probability measures *e.g.* the Lévy–Prokhorov distance, but the richness of W lies in its ability to incorporate a lot of the geometry of the underlying space through the distance d. Consequently Wasserstein spaces $(\mathcal{P}_p(\Omega), W_p)$ are very large and many metric spaces can be embed into Wasserstein spaces with low distortion [Bourgain 1986, Andoni 2015, Frogner 2019].

2.1.1 The main theorem of the linear Optimal Transport theory

A fundamental result of the linear OT theory is the cyclical monotonicity property of its optimal transport plans. Basically it illustrates that an optimal transport plan can not be improved *locally* and more importantly that is also sufficient for being a global optimal transport plan. Consequently it characterizes the set of optimal couplings using the notion of *c*-concave functions which appears to be very useful for defining the notions of duality and for solving the Monge problem (MP) by relying on the Kantorovitch relaxation (KP). As such the cyclical monotonicity is maybe the main ingredient of linear OT. This section aims at presenting in short both this result and its consequences.

Definition 2.1.3 (Cyclically Monotone Set). *Let* $c : \mathcal{X} \times \mathcal{Y} \to] - \infty, +\infty]$ *be a real valued function on arbitrary sets* $\mathcal{X}, \mathcal{Y}$. *A set* $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ *is said to be cyclically monotone if for* $(x_i, y_i)_{i=1}^N \in \Gamma^N$ *and* σ *a permutation of* $[1, .., N]$

$$\sum_{i=1}^N c(x_i, y_i) \leq \sum_{i=1}^N c(x_i, y_{\sigma(i)}) \tag{2.8}$$

We will call c-CM such sets.

c-CM sets is an important notion in OT theory since it characterizes optimal transport plans for well behaved costs. We consider the following definition:

Definition 2.1.4 (Support). *Let* $(\mathcal{X}, d)$ *be a Polish space and* $\mu \in \mathcal{P}(\mathcal{X})$. *The support of* μ *is defined as the smallest closed set* F *such that* $\mu(F) = 1$ *or equivalently:*

$$supp(\mu) = \{x \in \mathcal{X} \mid \forall \varepsilon > 0, \mu(B(x, \varepsilon)) > 0\} \tag{2.9}$$

Informally the support of a distribution is where the distribution "lives", *i.e.* where it is not zero. In the discrete case $\boldsymbol{\pi}$ is a matrix and the support is found in the indices (i, j) such that $\pi_{ij} > 0$. The following theorem states that the support of an optimal coupling is actually a *c*-CM set:

> **Theorem 2.1.2** (Theorem 1.38 in [Santambrogio 2015]). *Let $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty[$ continuous and $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ with $\mathcal{T}_c(\mu, \nu) < +\infty$. If a coupling $\pi \in \Pi(\mu, \nu)$ is optimal for (KP) then $supp(\pi)$ is a c-CM set.*

Interestingly enough the *c*-CM sets are characterized by specific functions based on the notion of *c*-transforms. This property gives another way of computing optimal transport plans based on deterministic functions.

Definition 2.1.5 (*c*-transforms). *Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces and $\psi : \mathcal{X} \to \overline{\mathbb{R}}$ be a function. We define its c-transform as the function $\psi^c : \mathcal{Y} \to \overline{\mathbb{R}}$:*

$$\psi^c(y) = \inf_{x \in \mathcal{X}} c(x, y) - \psi(x) \tag{2.10}$$

and the $\bar{c}$-transform of a function $\phi : \mathcal{Y} \to \overline{\mathbb{R}}$ as the function $\phi^{\bar{c}} : \mathcal{X} \to \overline{\mathbb{R}}$:

$$\phi^{\bar{c}}(x) = \inf_{y \in \mathcal{Y}} c(x, y) - \phi(y) \tag{2.11}$$

Fonctions that can be written as ψ^c or $\phi^{\bar{c}}$ are called respectively c-concave or $\bar{c}$-concave functions.

Remark 2.1.2. *The c-transform is a a generalization of the Legendre transform that is well-known in convex analysis [Rockafellar 1970]. More precisely for function $u : \mathbb{R}^d \to \mathbb{R}$ its Legendre transform is defined as $u^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \mathbf{y}, \mathbf{x} \rangle - u(\mathbf{x})$ (see Memo 2.1.2). The c-transform corresponds to this notion by considering $c(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$ (up to the change of sign). Another special case deserves attention that is when $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$. Consider $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, then the function ψ is c-concave if and only if the function $u : \mathbf{x} \to \frac{1}{2}\|\mathbf{x}\|^2 - \psi(\mathbf{x})$ is convex and lower semi-continuous and the Legendre transform of u is the function $\mathbf{x} \to \frac{1}{2}\|\mathbf{x}\|^2 - \psi^c(\mathbf{x})$ (see [Santambrogio 2015, Proposition 1.21]).*

When $\mathcal{X} = \mathcal{Y}$ and c is symetric both notions are equivalent and in this case we will drop this distinction. One important property about *c*-transform is that it satisfies:

$$\forall x, y \in \mathcal{X} \times \mathcal{Y}, \psi(x) + \psi^c(y) \leq c(x, y) \tag{2.12}$$

The case of equality of (2.12) is attained on special subsets of $\mathcal{X} \times \mathcal{Y}$ that are precisely the *c*-CM sets as stated in the next theorem:

> **Memo 2.1.2** (Convex analysis). *For any function $u : \mathbb{R}^d \to \mathbb{R}$ its convex conjugate or Legendre transform is defined by $u^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \mathbf{y}, \mathbf{x} \rangle - u(\mathbf{x})$. The subdiffenrential of u, denoted as ∂u, is defined for $\mathbf{x} \in \mathbb{R}^d$ as $\partial u(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^d \mid \forall \mathbf{z} \in \mathbb{R}^d\ u(\mathbf{x}) - u(\mathbf{z}) \geq \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle\}$ which reduces to $\partial u(\mathbf{x}) = \{\nabla u(\mathbf{x})\}$ when u is differentiable at $\mathbf{x}$. When u is convex differentiable the convex conjugate has the following important properties (see [Rockafellar 1970, Theorem 23.5]):*
>
>   *i $u(\mathbf{x}) + u^*(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle$ (Fenchel-Young inequality)*
>
>   *ii $u(\mathbf{x}) + u^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle \iff \mathbf{y} \in \partial u(\mathbf{x})$*
>
>   *iii In particular $u(\mathbf{x}) + u^*(\nabla u(\mathbf{x})) = \langle \mathbf{x}, \nabla u(\mathbf{x}) \rangle$*

Theorem 2.1.3 (Theorem 1.37 in [Santambrogio 2015])**.** *If $\Gamma \neq \emptyset$ is a c-CM set in $\mathcal{X} \times \mathcal{Y}$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, then there exists a c-concave function $\psi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ such that:*

$$\Gamma \subset \{(x, y) \in X \times Y : \psi(x) + \psi^c(y) = c(x, y)\} \tag{2.13}$$

To summarize, the support of optimal couplings are necessarily *c*-CM sets and these *c*-CM sets are also characterized by functions using the notion of *c*-transform. The fundamental theorem of optimal transport states that all these results are in fact equivalent:

Theorem 2.1.4 (Fundamental theorem of linear OT)**.** *Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces, $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ and $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty[$ lower-semi continuous such that $\mathcal{T}_c(\mu, \nu) < \infty$. Let $\pi \in \Pi(\mu, \nu)$ then the following conditions are equivalent:*

(i) π is optimal for (KP)

(ii) The support of π is c-cyclically monotone

(iii) There exists a measurable c-concave function ψ such that $\psi(x) + \psi^c(y) = c(x, y)$ π a.e.

Proof. We will give a sketch of proof, for completeness the reader can refer to Theorem 5.10 in [Villani 2008]. Theorems 2.1.2 and 2.1.3 already proved that $(i) \implies (ii) \implies (iii)$ in the case where c is continuous. To pass from continuity to lower semi-continuity we can consider a sequence $(c_k)_k$ of costs that converges increasingly to c and observe that [Santambrogio 2015, Lemma 1.41]:

$$\lim_{k \to +\infty} \inf_{\pi \in \Pi(\mu, \nu)} \int c_k(x, y) \mathrm{d}\pi(x, y) = \inf_{\pi \in \Pi(\mu, \nu)} \int c(x, y) \mathrm{d}\pi(x, y)$$

and, using some subtleties, this can prove $(i) \implies (ii) \implies (iii)$ when c is lower semi-continuous. For the converse we can easily prove that $(iii) \implies (i)$. By hypothesis $\int c(x, y) \mathrm{d}\pi(x, y) = \int \psi(x) \mathrm{d}\mu(x) + \int \psi^c(y) \mathrm{d}\nu(y)$. However for any other coupling π' we have $\int c(x, y) \mathrm{d}\pi'(x, y) \geq \int \psi(x) \mathrm{d}\mu(x) + \int \psi^c(y) \mathrm{d}\nu(y)$ by relation (2.12) and so $\int \int c(x, y) \mathrm{d}\pi'(x, y) \geq \int c(x, y) \mathrm{d}\pi(x, y)$ so that π is optimal. Technical details are hidden here for proving the measurability and integrability of ψ, ψ^c. $\qquad\square$

The main theorem of OT for duality A first implication of the fundamental theorem is related to a duality principle which is a widely used property in linear programming (see Section 2.1.4 for more details). This property can be extended in full generality in the context of OT as stated in the next theorem:

Theorem 2.1.5 (Duality theorem)**.** *Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces, $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ and $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ be lower semi-continuous (l.s.c.) such that $\mathcal{T}_c(\mu, \nu) < +\infty$ then strong duality holds. More precisely the* dual problem*:*

$$\sup_{\substack{\phi, \psi \in C_b(\mathcal{X}) \times C_b(\mathcal{Y}) \\ \forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \ \phi(x) + \psi(y) \leq c(x,y)}} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) \tag{DKP}$$

leads to the same optimum as the (KP) *problem. Equivalently:*

$$\mathcal{T}_c(\mu, \nu) = \sup_{\phi, \psi \in \Phi_c} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) \tag{2.14}$$

> *where* Φ_c *is the set of continuous bounded functions which verifies:*
>
> $$\forall x, y \in \mathcal{X} \times \mathcal{Y}, \phi(x) + \psi(y) \leq c(x, y) \tag{2.15}$$
>
> *This result holds when* Φ_c *is replaced by* $\Phi_c(\mu, \nu)$ *the set of integrable functions which satisfies* (2.15).

Proof. For completeness we will give here an idea of the proof, the interested reader can refer to Theorem 5.10 in [Villani 2008] for more details. If $\phi, \psi \in \Phi_c(\mu, \nu)$ and $\pi \in \Pi(\mu, \nu)$ then by hypothesis:

$$\int \phi(x) \mathrm{d}\mu(x) + \int \psi(y) \mathrm{d}\nu(y) = \int \phi(x) + \psi(y) \mathrm{d}\pi(x, y) \leq \int c(x, y) \mathrm{d}\pi(x, y) \tag{2.16}$$

Which implies that $\sup_{\phi, \psi \in \Phi_c(\mu, \nu)} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y) \leq \mathcal{T}_c(\mu, \nu)$. To show the converse inequality we will use the cyclical monotonicity properties of optimal transport plans. Let π^* be an optimal coupling for the (KP) problem. Using Theorem 2.1.4 we know that there exists a c-concave function ψ such that $\psi(x) + \psi^c(y) = c(x, y)$ for all $x, y \in \mathcal{X} \times \mathcal{Y}$. In this way:

$$
\begin{aligned}
\int c(x, y) \mathrm{d}\pi^*(x, y) &= \int \psi(x) \mathrm{d}\mu(x) + \int \psi^c(y) \mathrm{d}\nu(y) \\
&\leq \sup_{\phi, \psi \in \Phi_c} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \psi(y) \mathrm{d}\nu(y)
\end{aligned}
\tag{2.17}
$$

Last inequality stems from the property (2.12) of c-transform since $(\psi, \psi^c) \in \Phi_c$. If c is continuous and bounded then so are ψ, ψ^c so last inequality is valid. If c is only *l.s.c.* then we can show that there is a sequence $(c_k)_k$ bounded and n-Lipschitz such that $c = \sup_k c_k$. A limit argument suffices to conclude for this case. $\qquad \square$

The functions ϕ, ψ are usually called *Kantorovitch potentials* and play an important role in OT problems. Given two admissible potentials ϕ, ψ *i.e.* that satisfy $\phi(x) + \psi(y) \leq c(x, y)$ we can always cook up a pair of "better" potentials using the c-transform. Indeed, due to (2.12), one can check that the pairs (ϕ, ϕ^c), $(\psi, \psi^{\bar{c}})$ are also admissible potentials and improve the objective function. It turns out that after one iteration of this procedure we can not improve the potentials anymore. Based on this remark we can also write the duality as the maximization over one *single* potential which is the *semi-dual* formulation:

$$\sup_{\phi \ c-\text{concave}} \int_{\mathcal{X}} \phi(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} \phi^c(y) \mathrm{d}\nu(y) \tag{2.18}$$

Example 2.1.4. *When* $c = d$ *is a distance on some space* Ω *then there is a tight connection between* c-*transform and 1-Lipschitz functions. Indeed suppose that* ϕ *is a 1-Lipschitz function, then for* $x, y \in \Omega^2$, $\phi(y) \leq \phi(x) + d(x, y)$ *so that* $\phi(y) = \inf_{x \in \Omega} \phi(x) + d(x, y) = (-\phi)^d(y)$ *which proves that the d-transform of* $-\phi$ *is* ϕ. *The converse is also true so that the semi-dual formulation can be written:*

$$\sup_{\phi \in Lip_1(\Omega)} \int_{\Omega} \phi(x) \mathrm{d}\mu(x) - \int_{\Omega} \phi(y) \mathrm{d}\nu(y) \tag{2.19}$$

This formulation is very useful in practice in the context of generative modeling [Arjovsky 2017] (see Section 2.1.5).

The main theorem of OT for the Monge problem From a theoretical perspective one fundamental question that arises is the *regularity* of these potentials and, with some assumptions, we can use them to solve the Monge problem (MP) based on the Kantorovitch relaxation (KP). We have the following result:

Proposition 2.1.1 (Proposition 1.15 in [Santambrogio 2015]). *Let $\mathcal{X} = \mathcal{Y} = \Omega \subset \mathbb{R}^d$ and $c \in C^1$. In the following ϕ denotes a Kantorovitch potential. If $(\mathbf{x}_0, \mathbf{y}_0) \in supp(\pi^*)$ then:*

$$\nabla \phi(\mathbf{x}_0) = \nabla_{\mathbf{x}} c(\mathbf{x}_0, \mathbf{y}_0) \tag{2.20}$$

provided that ϕ is differentiable at $\mathbf{x}_0$.

This proposition suggests the following strategy in order to find an optimal coupling: (1) Ensure that ϕ is differentiable μ *a.e.* This can be guaranteed when μ is absolutely continuous with respect to the Lebesgue measure and when ϕ is regular enough, such as Lipschitz. (2) Deduce from previous proposition that π is characterized by a deterministic function that is the map associating y_0 to each x_0. The idea here is to "inverse" ∇c in (2.20) and deduce from (x_0, y_0) that y_0 is uniquely determined from x_0. This can be done using some regularity assumptions on c and the spaces $\mathcal{X}, \mathcal{Y}$. When conditions (1) and (2) are satisfied we can deduce that the optimal coupling is unique since it was constructed using ϕ and c only.

The step (2) can be verified *e.g.* when we have the following condition:

Definition 2.1.6 (Twist condition). *For $\Omega \subset \mathbb{R}^d$ we say that $c : \Omega \times \Omega \to \mathbb{R}$ satisfies the Twist condition whenever c is differentiable w.r.t. $\mathbf{x}$ at every point, and the map $\mathbf{y} \to \nabla_{\mathbf{x}} c(\mathbf{x}_0, \mathbf{y})$ is injective for every $\mathbf{x}_0$.*

When working on Euclidean domains and when the cost $c \in C^2$ this condition corresponds to $det\left(\frac{\partial^2 c}{\partial \mathbf{y}_i \partial \mathbf{x}_j}\right) \neq 0$. The squared Euclidean cost is an important example of costs which satisfies the Twist condition and leads to the celebrated Brenier theorem [Brenier 1991]:

Theorem 2.1.6 (Brenier). *Let $\Omega = \mathbb{R}^p$, $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2$ and $\mu \in \mathcal{P}(\mathbb{R}^p)$ absolutely continuous with respect to the Lebesgue measure and $\nu \in \mathcal{P}(\mathbb{R}^p)$ with $\int \|\mathbf{x}\|_2^2 d\mu(\mathbf{x}) < +\infty, \int \|\mathbf{y}\|_2^2 d\nu(\mathbf{y}) < +\infty$.*

The optimal transport plan π^* of (KP) *is unique and supported on the gradient of a convex function. More precisely it can be written as $\pi^* = (id \times T) \# \mu$ where $T = \nabla \phi$ and $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex and finite almost everywhere.*

Moreover T is the unique solution of (MP). *If T' is another optimal solution then $T = T'$ μ a.e.*

This result can be generalized to costs $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{y} - \mathbf{x})$ with h strictly convex and in this case T can be written as $T(\mathbf{x}) = \mathbf{x} - (\nabla h)^{-1} \nabla \phi(\mathbf{x})$ where ϕ is a c-concave function (see *e.g.* [Gangbo 1996]).

The regularity of potential functions ϕ, ψ and its consequences for Optimal Transport problems is a long-standing line of research. Other more general hypothesis on the cost function c than the Twist condition can be built thanks to Ma, Trudinger and Wang who found a key assumption on the cost c that requires fourth-order condition on the cost functions [Ma 2005]. The resulting MTW conditions turned out to be sufficient to prove the regularity of the Kantorovitch potentials. We refer the reader to [Figalli 2010] for a survey on this topic.

2.1.2   Special cases: 1D transportation and transport between Gaussians

Two important special cases will be considered in this manuscript, namely the cases where μ and ν are probability distributions on $\mathbb{R}$ or when they are Gaussians distributions. Theses cases are well-known in linear OT for having *closed-form solutions* which are given in the next results (respectively [Santambrogio 2015, Theorem 2.9] and [Peyré 2019, Remark 2.30]).
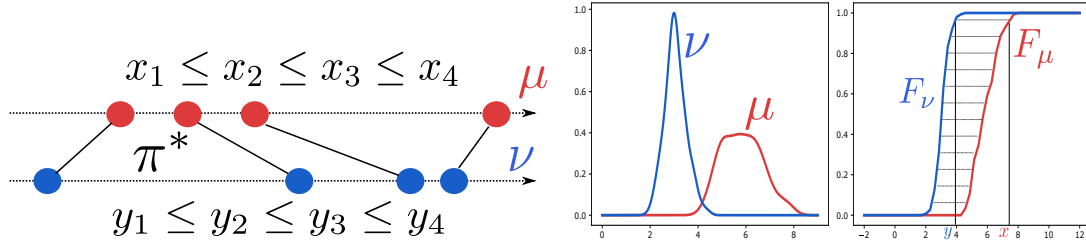
Figure 2.4: OT for 1D probability measures can be computed using simple sorts. **(left)** Optimal coupling between two discrete probability measures with uniform weights. When the points are sorted it associates first point of the source with the first point of the target and so on. **(right)** Generic case: the optimal coupling associates horizontally the points *w.r.t.* the cumulative distributions of the probability measures. In this example x is associated with y.

Theorem 2.1.7 (Closed-form expression on the real-line). *Asssume that $\Omega = \mathbb{R}$, $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Let F_μ be the cumulative distribution function:*

$$\forall t \in \mathbb{R}, \ F_\mu(t) = \mu(]-\infty, t]) \tag{2.21}$$

and F_μ^{-1} its pseudo inverse, namely:

$$\forall x \in [0, 1], \ F_\mu^{-1}(x) = \inf\{t \in \mathbb{R} \ | F_\mu(t) \geq x\} \tag{2.22}$$

If $c(x, y) = h(y - x)$ where h is stricly convex then (KP) has a unique solution given by $\pi^* = (F_\mu^{-1} \times F_\nu^{-1})\#\mathcal{L}_{[0,1]}$ where $\mathcal{L}_{[0,1]}$ is the Lebesgue measure restricted to $[0, 1]$.

Moreover if μ is atomless π^* is supported on $T_{mon}(x) = F_\nu^{-1}(F_\mu(x))$, i.e. $\pi^* = (id \times T_{mon})\#\mu$. If h is only convex then π^* is still optimal but uniqueness can not be guaranteed.

This theorem states that it suffices to sort the support of the distributions in order to recover the optimal coupling (see Figure 2.4). In the special case where $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ this corresponds to sort $x_1 \leq \cdots \leq x_n$, $y_1 \leq \cdots \leq y_n$ and to associate x_1 with y_1, x_2 with y_2 and so on. In the case $\mu = \sum_{i=1}^n a_i\delta_{x_i}, \nu = \sum_{i=1}^m b_j\delta_{y_j}$ the previous theorem states that, after sorting the points, the optimal mapping is obtained by putting as much mass as possible from x_1 to y_1 and to add the remaining mass to y_2. This procedure is repeated until there is no more mass left. This corresponds to a *monotone rearrangement* $\pi_{ij}, \pi_{i'j'} > 0$ then $x_i \leq x_{i'}$ implies that $y_j \leq y_{j'}$. Overall the Wasserstein distance in 1D can be solved using simple sorts. This result is the main ingredient of the sliced-Wasserstein distance (see Section 2.1.5).

Another special case arises when the probability measures are Gaussian. This is a well known result in the literature of OT geometry [Givens 1984, McCann 1997, Takatsu 2011] which is recalled in the following theorem:

Theorem 2.1.8 (Closed form expression for Gaussians). *Let $\mu = \mathcal{N}(\mathbf{m}_\mu, \boldsymbol{\Sigma}_\mu), \nu = \mathcal{N}(\mathbf{m}_\nu, \boldsymbol{\Sigma}_\nu)$ and suppose that $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{y} - \mathbf{x})$ with h striclty convex.*
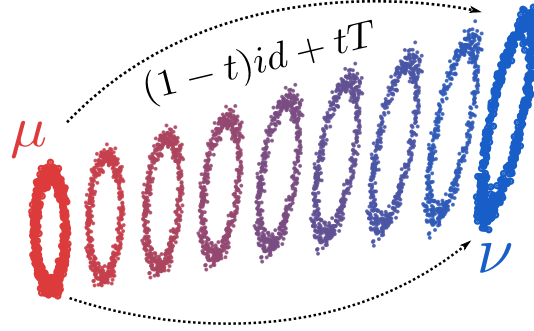
Figure 2.5:   Linear displacement interpolation between two discrete probability measure μ, ν on $\mathbb{R}^2$ using the linear map T defined in 2.1.8. The figure shows the displacement interpolant which is a probability measure is defined by $((1-t)id + T)\#\mu$ for $t \in [0,1]$ [McCann 1997] (see Section 2.1.6 for more details)

Let $T : \mathbf{x} \to \mathbf{m}_\nu + \mathbf{A}(\mathbf{x} - \mathbf{m}_\nu)$ where:

$$\mathbf{A} = \mathbf{\Sigma}_\mu^{-1/2}(\mathbf{\Sigma}_\mu^{1/2}\mathbf{\Sigma}_\nu\mathbf{\Sigma}_\mu^{1/2})^{\frac{1}{2}}\mathbf{\Sigma}_\mu^{-1/2} \tag{2.23}$$

then T is the unique optimal solution of (MP) and $\pi^* = (id \times T)\#\mu$ is the unique optimal solution of (KP).

In particular when $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ is the Euclidean distance on $\mathbb{R}^d$ the 2-Wasserstein distance is given by:

$$W_2^2(\mu, \nu) = \|\mathbf{m}_\mu - \mathbf{m}_\nu\|_2^2 + \mathcal{B}(\mathbf{\Sigma}_\mu, \mathbf{\Sigma}_\nu)^2 \tag{2.24}$$

where $\mathcal{B}(\mathbf{\Sigma}_\mu, \mathbf{\Sigma}_\nu) = tr\left(\mathbf{\Sigma}_\mu + \mathbf{\Sigma}_\nu - 2(\mathbf{\Sigma}_\mu^{1/2}\mathbf{\Sigma}_\nu\mathbf{\Sigma}_\mu^{1/2})^{\frac{1}{2}}\right)$ is the Bures metric [Bures 1969].

Interestingly enough the problem of computing OT between Gaussian measures draws connections with the general case. Indeed for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$ the optimal map T defined in Theorem 2.1.8 is actually the optimal Monge map of (MP) when restricted to the class of linear Monge map [Flamary 2019, Proposition 1]. Figure 2.5 illustrates the behavior of this map for two discrete probability measures on $\mathbb{R}^2$.

Note that a generalization of the previous result exists for elliptical distributions which are somehow generalizations of Gaussian densities. In this case the W_2 admits also a closed-form (see [Muzellec 2018]).

2.1.3   Some statistical aspects of OT

In most of machine learning applications we do not have access to the true distributions μ, ν but only to samples from these distributions. As such a natural question arises: can we infer from this samples good estimates of OT objects such as couplings or OT distances? One particular question is how well can we estimate the Wasserstein distance by relying only on samples of the distribution? If we consider a probability distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ and an empirical distribution $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ where $x_i \sim \mu$ are *iid* samples does μ_n is a good proxy for μ? Unfortunately the sample complexity of the estimation of the Wasserstein distance is exponential in the dimension of the ambient space. More precisely $\mathbb{E}[W_p(\mu_n, \mu)] = O(n^{-\frac{1}{d}})$ so that the Wasserstein distance suffers from the curse of dimensionality [Dudley 1969, Weed 2017]. It was shown in [Weed 2017] that this result can be refined to $O(n^{-\frac{1}{p}})$ where p is the intrinsic dimension of the

$$\sigma = \begin{pmatrix} 1234 \\ 3241 \end{pmatrix} \quad \pi_\sigma = \tfrac{1}{4} \begin{pmatrix} 0010 \\ 0100 \\ 0001 \\ 1000 \end{pmatrix}$$
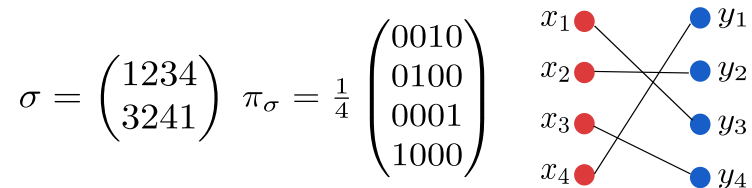
Figure 2.6: An assignment of x_1, x_2, x_3, x_4 to y_1, y_2, y_3, y_4 can be described whether by a permutation σ or using assignment matrix $\boldsymbol{\pi}_\sigma$. In this example both correspond to the permutation σ: $1 \to 3, 2 \to 2, 3 \to 4,\ 4 \to 1$

data but, generally, this is a major bottleneck for the use of OT in high-dimensional machine learning problems. Previous analysis can be extended to the infinite dimensional setting as analysed in [Lei 2020]. The problem of estimating the optimal coupling by relying on small batches of μ when it is discrete was further analyzed in [Fatras 2020].

To circumvent this limitation some robust projection formulations have been proposed [Niles-Weed 2019, Lin 2020] as well as strategies such as gaussian-smoothing [Goldfeld 2020] or based on wavelet estimator [Weed 2019a]. The entropic regularization presented in the next section in also one of the tool that facilitates the estimation of W_p for high-dimensional settings. For more details about statistical aspects of OT we refer the reader to [Weed 2019b].

2.1.4 A quick numerical tour: solving Optimal Transport

In this section we consider the problem of computing OT between discrete probability measures $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \nu = \sum_{i=1}^{m} b_j \delta_{y_j}$. The problem can be solved in many ways and we aim here at giving a brief summary these possibilities. We denote by $\mathbf{C} = (c_{ij})_{i \in [\![n]\!], j \in [\![m]\!]}$ the matrix of all pair-to-pair costs between the samples x_i, y_j, *i.e.* $c_{ij} = c(x_i, y_j)$ for all $i \in [\![n]\!], j \in [\![m]\!]$. In the discrete case the underlying problem reads:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \boldsymbol{\pi} \rangle_{\mathcal{F}} = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \sum_{ij} c_{ij} \pi_{ij}. \tag{2.25}$$

As described previously the problem is *linear* in $\boldsymbol{\pi}$, in this way the discrete case corresponds to a linear program (LP) [Dantzig 1997]. Before discussing potential algorithms for solving equation (2.25) we detail one important special case.

Assignment problems Suppose that $m = n$. In this case we can look for an *assignment* of the points, that is a one-to-one correspondence between the points x_i, y_i. This translates mathematically by looking at the permutation $\sigma \in S_n$ of the points or at the permutation matrix $\boldsymbol{\pi}_\sigma = \begin{cases} \frac{1}{n}, & \text{if } j = \sigma(i) \\ 0, & \text{otherwise} \end{cases}$ such that the overall cost is minimized (see Figure 2.6). In this situation one aims at solving:

$$\min_{\sigma \in S_n} \frac{1}{n} \sum_{i=1}^{n} c_{i,\sigma(i)} = \min_{\sigma \in S_n} \langle \mathbf{C}, \boldsymbol{\pi}_\sigma \rangle_{\mathcal{F}} \tag{2.26}$$

This problem is well known in the literature as the *linear assignment problem* (see *e.g.* [Burkard 1999]). It is worth pointing that, in this case, it exactly corresponds to the Monge problem (MP) in the discrete case. Interestingly enough when the weights of the OT problem (2.25) are set as uniform *i.e.* $\mathbf{a} = \mathbf{b} = \frac{1}{n} \mathbf{1}_n$

Algorithm 1 North-West corner rule

1: $\mathbf{a}, \mathbf{b}, i, j = 1$
2: **while** $i <= n$, $j <= m$ **do**
3: $\quad \pi_{ij} = \min\{a_i, b_j\}$ // Send as many units as possible form i to j
4: $\quad a_i = a_i - \pi_{ij}$ // Adjust the supply
5: $\quad b_j = b_j - \pi_{ij}$ // Adjust the demand
6: $\quad$ If $a_i = 0$, $i = i + 1$, if $b_j = 0$, $j = j + 1$
7: **end while**

---

both problems (2.25) and (2.26) are equivalent. More precisely by combining the fundamental theorem of linear programming [Bertsimas 1997], which states that the minimum of a linear program is reached at an extremal point of the polyhedron, and Birkhoff's theorem [Birkhoff 1946], which states that the extremal points of $\Pi(\frac{1}{n}\mathbf{1}_n, \frac{1}{n}\mathbf{1}_n)$ is the set of permutation matrices, we can conclude that the optimal map of (2.25) is reached at $\boldsymbol{\pi}_\sigma$ which is optimal for (2.26).

Algorithmic solutions To solve the OT problem (2.25) in general one can rely on classical algorithms for solving (LP) [Dantzig 1997]. We make here a brief overview of possible numerical solutions and we refer the reader to Section 3 in [Peyré 2019] for more details.

As seen in Theorem 2.14 the OT problem can be solved using duality which reads in the discrete case:

$$\max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\beta}\in\mathbb{R}^m \\ \forall(i,j)\in[\![n]\!]\times[\![m]\!], \alpha_i+\beta_j \leq c_{ij}}} \boldsymbol{\alpha}^T\mathbf{a} + \boldsymbol{\beta}^T\mathbf{b} \tag{2.27}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$ denotes the Kantorovitch potentials. Thanks to the fundamental Theorem 2.1.4 an optimal solution $\boldsymbol{\pi}^*$ of the primal problem is found when $\alpha_i^* + \beta_j^* = c_{ij}$ for $\pi_{ij}^* > 0$ where $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ are solutions of the dual problem. Using this remark we can solve (2.25) by relying on the Network Simplex algorithm which philosophy is to find feasible solutions $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ such that $\alpha_i + \beta_j = c_{ij}$ whenever $\pi_{ij} > 0$ (in this case we say that $\boldsymbol{\pi}$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are complementary *w.r.t.* $\mathbf{C}$). The complexity of this algorithm is $O(n^3 \log(n))$ when $m = n$. The special case of uniform weights for assignment problems can be solved using the Auction algorithm which has a cubic complexity $O(n^3)$.

Special cases: Monge property The case where $\mathbf{C}$ has special structure deserves attention. In particular when $\mathbf{C}$ satisfies the following *Monge property* [Burkard 1996]:

$$\forall(i,j)\in[\![n]\!]\times[\![m]\!], \ c_{i,j} + c_{i+1,j+1} \leq c_{i+1,j} + c_{i,j+1} \tag{2.28}$$

which can be tested in $O(mn)$ operations. This property has some interesting historical background. It is actually based on the original observation of Monge who states that if quantity must be transported from locations x_1, y_1 to locations x_2, y_2 then the route from x_1 and the route from y_1 must not intersect: better not to cross the paths![4] In this case the simple North-West corner rule (see Algorithm 1) produces an optimal solution in $O(n + m)$.

---

[4]The original quote by Monge is [Monge 1781]: "Lorsque le transport du deblai se fait de manière que la somme des produits des molécules par l'espace parcouru est un minimum, les routes de deux points quelconques A & B, ne doivent plus se couper entre leurs extrémités, car la somme Ab + Ba, des routes qui se coupent, est toujours plus grande que la somme Aa + Bb, de celles qui ne se coupent pas."
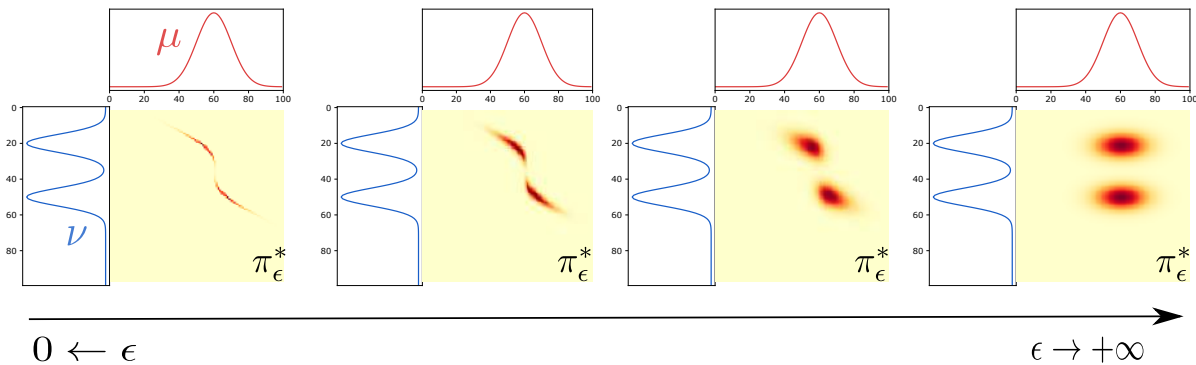
Figure 2.7: Effect of the entropic regularization parameter ε on the optimal coupling $\boldsymbol{\pi}_\varepsilon^*$ between two 1D probability distributions. As ε increases the coupling tends to blur and converges to the marginals' product coupling.

Special cases: 1D probability distributions As seen in Section 2.1.2 the case of 1D probability distributions can be solved efficiently using simple sorts when $\mathbf{C}$ is *e.g.* a squared Euclidean distance matrix. The complexity of computing the Wasserstein distance is $O(n\log(n))$ when $n = m$ and weights are uniform and in general it suffices to compute the two cumulative distribution functions which is $O(n\log(n) + m\log(m))$.

Special cases: Gaussian distributions When μ and ν are Gaussian distributions (and with a Euclidean cost) the OT problem is also quite easy to solve. In the discrete case, when relying on samples from μ, ν, and using the empirical version of the means and covariances, finding the optimal solution has a $O((n + m)d^2 + d^3)$ complexity.

Although previous special cases exist solving the OT problem in general remains costly. The next section presents a regularization scheme that tends to lower this computational complexity and was one of the major breakthrough in OT past years.

Entropic regularization

The idea of penalizing the entropy of the joint coupling can be traced back to Schrödinger [Schrodinger 1931] and its use for linear OT to Wilson [Wilson 1969], yet it was made popular quite recently in the OT community [Cuturi 2013]. The entropic regularization has multiple virtues in practice: 1) it turns the optimal transport problem into a strongly-convex minimization problem which solution is unique 2) solving an entropic regularized OT only involves simple iterations of matrix-vectors products which can be plugged easily into modern differentiable frameworks 3) it can be accelerated on GPU and can solve in parallel several OT problems 4) it has many desirable properties for high-dimensional problems statistically speaking.

The entropy term for a coupling $\boldsymbol{\pi}$ reads as:

$$H(\boldsymbol{\pi}) = -\sum_{ij}(\log(\pi_{ij}) - 1)\pi_{ij} \tag{2.29}$$

which corresponding entropic regularized OT problem:

Algorithm 2 Sinkhorn-Knopp Algorithm for entropic transport

1: $\mathbf{a}, \mathbf{b}, \mathbf{C}, \varepsilon > 0, \mathbf{u}^{(0)}, \mathbf{v}^{(0)} = \mathbf{1}, \mathbf{K} = \exp(-\frac{\mathbf{C}}{\varepsilon})$

2: **for** $i = 1, \ldots, n_{it}$ **do**

3:     $\mathbf{u}^{(i)} = \mathbf{a} \oslash \mathbf{K}^{\top} \mathbf{v}^{(i-1)}$ // Update left scaling

4:     $\mathbf{v}^{(i)} = \mathbf{b} \oslash \mathbf{K}\mathbf{u}^{(i-1)}$ // Update right scaling

5: **end for**

6: **return** $\boldsymbol{\pi}^* = \mathrm{Diag}(\mathbf{u})\mathbf{K}\,\mathrm{Diag}(\mathbf{v})$

$$\mathcal{T}_c^{\varepsilon}(\mu, \nu) = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \boldsymbol{\pi} \rangle_{\mathcal{F}} - \varepsilon H(\boldsymbol{\pi}) \tag{ε-KP}$$

Interestingly enough the optimal cost of (KP) can be obtained as $\varepsilon \to 0$, *i.e.* $\lim_{\varepsilon \to 0} \mathcal{T}_c^{\varepsilon}(\mu, \nu) = \mathcal{T}_c(\mu, \nu)$ [Peyré 2019, Propositon 4.1]. As a side effect, the entropic term tends to blur the optimal coupling so that more points are associated compared to the sparse optimal solution of the original problem. In other words entropy forces the solution to have a spread support. In the limit setting where $\varepsilon \to +\infty$ all points are coupled together such that $\lim_{\varepsilon \to +\infty} \boldsymbol{\pi}_{\varepsilon}^* = \mathbf{a}\mathbf{b}^T$ where $\boldsymbol{\pi}_{\varepsilon}^*$ denotes the optimal coupling of (ε-KP) (see Figure 2.7). Note that the entropy regularization can also be defined when the probability measures are not discrete and in this case reads $H(\pi) = \int_{\mathcal{X} \times \mathcal{Y}} (\log(\frac{d\pi(x,y)}{d\mu(x)d\nu(y)}) - 1) d\pi(x, y)$.

Sinkhorn-Knopp and Bregman projections   A simple analytic solution of (ε-KP) can be found using the lagragian duality as expressed in the following proposition [Peyré 2019, Proposition 4.3]:

Proposition 2.1.2. *Problem* (ε-KP) *has a unique solution of the form* $\boldsymbol{\pi}^* = \mathrm{Diag}(\mathbf{u})\mathbf{K}\,\mathrm{Diag}(\mathbf{v})$ *with* $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$ *and* $\mathbf{u}, \mathbf{v} \in \mathbb{R}_+^n \times \mathbb{R}_+^n$.

As written in [Sinkhorn 1967] there is a unique solution of the form $\boldsymbol{\pi}^* = \mathrm{Diag}(\mathbf{u})\mathbf{K}\,\mathrm{Diag}(\mathbf{v})$ with marginals $\mathbf{a}, \mathbf{b}$ providing that $\mathbf{K}$ is positive definite. Moreover it can be recovered based on the Sinkhorn-Knopp Matrix scaling algorithm that relies on matrices multiplications by alternatively updating $\mathbf{u}$ and $\mathbf{v}$ in order for $\boldsymbol{\pi}^*$ to have the prescribed marginals (see in Algorithm 2). When $n = m$ and by setting $\tau = \frac{4 \log(n)}{\varepsilon}$ the Sinkhorn algorithm produces an optimal solution $\boldsymbol{\pi}^*$ such that $\langle \mathbf{C}, \boldsymbol{\pi}^* \rangle_{\mathcal{F}} \leq \mathcal{T}_c(\mu, \nu) + \tau$ after $O(\|C\|_{\infty}^3 \log(n)\tau^{-3})$ iterations [Altschuler 2017]. In particular this implies that a τ-approximate solution of the original unregularized problem can be computed in $O(n^2 \log(n)\tau^{-3})$ time.

From a practical point of view the Sinkhorn's algorithm suffers from stability issues when $\varepsilon \to 0$ as the kernel $\mathbf{K}$ vanishes rapidly which results in divisions by 0 during the algorithms' iterations. To avoid such underflows for small value of ε [Schmitzer 2016] suggest a log-sum-exp stabilization trick whose iterations turn to be mathematically equivalent to the original iterations.

This problem is also a special case of a Kullback-Leiber minimization problem where one wants to find a coupling matrix $\boldsymbol{\pi}^*$ the closest possible to a kernel $\mathbf{K}$ in the sense of the Kullback-Leiber geometry. More precisely (ε-KP) is equal to:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \mathbf{KL}(\boldsymbol{\pi}|\mathbf{K}) \tag{2.30}$$

where $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$ and $\mathbf{KL}(\boldsymbol{\pi}, \mathbf{K}) = \sum_{ij} \pi_{ij} \log(\frac{\pi_{ij}}{K_{ij}}) - \pi_{ij} + K_{ij}$ is the Kullback-Leiber divergence between $\boldsymbol{\pi}$ and $\mathbf{K}$. Reformulating OT problems as a minimization of a Kullback-Leiber divergence allows the use of the machinery of Bregman projections in order to find a solution and to analyse the convergence [Benamou 2015]. This formulation is particularly interesting for solving multi-marginals

OT problems [Nenna 2016], regularized OT barycenter [Benamou 2015, Bigot 2019b] and the Gromov-Wasserstein problem [Peyré 2016] in short.

Sinkhorn divergences   One drawback of entropic regularized OT is that it induces a bias $\mathcal{T}_c^\varepsilon(\mu, \mu) \neq 0$ which can be problematic for learning using $\mathcal{T}_c^\varepsilon$. In [Genevay 2018] authors propose to correct this bias by considering the so-called Sinkhorn divergence:

$$SD_{c,\varepsilon}(\mu, \nu) = \mathcal{T}_c^\varepsilon(\mu, \nu) - \frac{1}{2}\mathcal{T}_c^\varepsilon(\mu, \mu) - \frac{1}{2}\mathcal{T}_c^\varepsilon(\nu, \nu). \tag{2.31}$$

This divergence enjoys many valuable properties. First it defines a symmetric positive definite smooth function on the space of probability measures that is convex in both μ, ν and that metrizes the weak convergence of probability measures [Feydy 2019]. Second it interpolates, through ε, between Wasserstein distance and Kernel norms (MMD) allowing finding a trade-off between both. Finally it is more suited for high-dimensional problems where the estimation of the Wasserstein distance is known to suffer from the curse of dimensionality (the sample complexity if $O(n^{-\frac{1}{d}})$ as explained in Section 2.1.3) whereas the sample complexity of $SD_{c,\varepsilon}$ is $O(\varepsilon^{-\frac{d}{2}} n^{-\frac{1}{2}})$ [Genevay 2019].

Stochastic Optimal Transport: going large scale.   The regularization of linear OT allows deriving stochastic formulations that are useful in practice to handle large scale datasets. This setting was considered in [Genevay 2016, Seguy 2018] where authors rely on the dual formulation (2.14) or the semi-dual formulation (2.18) in the regularized case. More precisely for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $\varepsilon > 0$ the regularized dual (*resp.* semi-dual) boils down to solve the following unconstrained maximization problems:

$$\sup_{\phi,\psi \in C(\mathbb{R}^d) \times C(\mathbb{R}^d)} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu, \mathbf{y} \sim \nu} [F_\varepsilon(\phi(\mathbf{x}), \psi(\mathbf{y}))] \tag{s-D}$$

$$\sup_{\psi \in C(\mathbb{R}^d)} \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} [H_\varepsilon(\mathbf{x}, \psi)] \tag{s-SD}$$

where $F_\varepsilon(\phi(\mathbf{x}), \psi(\mathbf{y})) = \phi(\mathbf{x}) + \psi(\mathbf{y}) - \varepsilon e^{\frac{\phi(\mathbf{x}) + \psi(\mathbf{y}) - c(\mathbf{x},\mathbf{y})}{\varepsilon}}$ and $H_\varepsilon(\mathbf{x}, \psi) = \int_{\mathcal{Y}} \psi(\mathbf{y}) \mathrm{d}\nu(\mathbf{y}) - \varepsilon \log(\int e^{\frac{1}{\varepsilon}(\psi(\mathbf{y}) - c(\mathbf{x},\mathbf{y}))} \mathrm{d}\nu(\mathbf{y})) - \varepsilon$ when entropic regularization is used. Since the problem is recast in the form of an unconstrained maximization of an expectation, the idea is to use stochastic gradients tools such as Stochastic Gradient Descent (SGD), or Stochastic Averaged Gradient (SAG) to compute a solution of (s-D),(s-SD). When both $\mu = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}, \nu = \sum_{j=1}^m b_j \delta_{\mathbf{y}_j}$ are discrete (SGD) or (SAG) are directly applicable to maximize the following finite sums:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{R}^n \times \mathbb{R}^m} \sum_{i=1,j=1}^{n,m} F_\varepsilon(\alpha_i, \beta_j) a_i b_j \tag{s-Ddis}$$

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^m} \sum_{i=1}^n H_\varepsilon(\mathbf{x}_i, \boldsymbol{\beta}) a_i \tag{s-SDdis}$$

In [Genevay 2016] authors propose to use (SAG) to compute (s-SDdis) which operates at each iteration by sampling a point $\mathbf{x}_k$ from μ then to compute the gradient of $H_\varepsilon(\mathbf{x}_k, \boldsymbol{\beta})$ corresponding to that sample while keeping in memory a copy of past gradients. This approach costs $O(n)$ per iteration due to the computation of the gradient and converges to a solution within $O(k^{-1})$ iterations. In contrast in [Seguy 2018] propose to solve (s-Ddis) by applying an (SGD) on mini-batches of both μ, ν which comes with a $O(p^2)$ cost per iteration where p is the mini-batch size and also converges in $O(k^{-1})$. In the continuous setting the problem is infinite dimensional so that it can not be solved using (SGD) anymore. In [Genevay 2019]

authors propose to represent the dual variables as kernel expansions while in [Seguy 2018] the dual variables are parametrized by a neural network. Another line of works rely on the unregularized problem and on the special case of W_1. In this case the duality reads $\sup_\phi \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu, \mathbf{y} \sim \nu}[\phi(\mathbf{x}) - \phi(\mathbf{y})]$ where the maximization is done over all 1-Lipschitz function ϕ. In [Arjovsky 2017] authors tackled this problem in the context of generative modelling. They parametrized μ, ν using a neural network and used the same (SGD)+mini-batch procedure resulting on a $O(p)$ cost per iteration. Their approach however relies on a weight clipping of the (NN) weights in-between gradient updates to enforce the Lipschitz constraint which lead to optimization difficulties [Gulrajani 2017]. Note all approaches comes at the price of biasing the optimal coupling due to the mini-batch sampling. This effect was further analyzed in [Fatras 2020].

2.1.5   Other formulations

Apart from entropic-regularized OT there are a lot of other methods for approximating OT. One of them relies on the closed-form expression of OT for probability distributions over the real line resulting on the so-called *Sliced Wasserstein distance* (SW) [Rabin 2011]. Considering $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ the key idea is to randomly select lines in $\mathbb{R}^d$, to project the measures into these lines and to compute the resulting 1D-Wasserstein distance which can be done using simple sorts as seen previously. The sliced-Wasserstein distance is the average of all these 1D-Wasserstein distances over all drawn lines. More precisely:

Definition 2.1.7 (Sliced Wasserstein distance)**.** *Let λ_{d-1} be the uniform measure on $\mathbb{S}^{d-1}$.*

For $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ we note $P_{\boldsymbol{\theta}}$ the projection on $\boldsymbol{\theta}$, i.e. $P_{\boldsymbol{\theta}}(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\theta} \rangle$. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)^2$. The Sliced Wasserstein distance between μ and ν is defined as:

$$SW_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} W_p^p(P_{\boldsymbol{\theta}} \# \mu, P_{\boldsymbol{\theta}} \# \nu) d\lambda_{d-1}(\boldsymbol{\theta}) \tag{2.32}$$

where the Wasserstein distance is defined with the standard Euclidean distance on $\mathbb{R}^d$.

SW enjoys several interesting properties. First SW_2 induces a similar topology than W_2: it defines a distance on $\mathcal{P}_p(\mathbb{R}^d)$ [Bonnotte 2013] that metrizes the weak convergence [Nadjahi 2019] and which is equivalent to the Wasserstein distance for measures with compact supports [Nadjahi 2020, Bonnotte 2013]. Second it defines a positive definite kernel $e^{-\gamma SW_2^2(\mu,\nu)}$ for $\gamma > 0$ over the space of probability distributions that can be easily plugged into an SVM [Kolouri 2016]. This contrasts with the Wasserstein distance W_2 which is not *Hilbertian* and consequently does not define a positive definite kernel (see Section 8.3 in [Peyré 2019]). In terms of sample complexity SW is known to be dimension independant [Nadjahi 2020] such as $O(n^{-1/2})$ when $p = 2$ [Lin 2020, Nadjahi 2020] and better samples complexities can be found by projecting on subspaces of dimension $k > 1$ instead of random lines [Lin 2020, Paty 2019], yet raising tractability issues as the sorting trick is no longer valid. Moreover, as a side effect of its definition, SW is unable to find the correspondences between the samples of the distributions as it does not provide an optimal transport map π which is valuable for certain application such as domain adaptation [Courty 2017].

From a practical side estimating SW requires the calculation of an integral over the hypersphere which can be done using a simple Monte-Carlo scheme. Hence for discrete probability measures with n atoms the overall complexity of computing SW is $O(Ln \log(n))$ where L is the number of projection directions on $\mathbb{S}^{d-1}$. The quality of the Monte Carlo estimates is impacted by the number of projections as well as the variance of the evaluations of the Wasserstein distance has pointed out empirically in [Kolouri 2019a, Deshpande 2019] and more formally in [Nadjahi 2020]. The low computational complexity of SW makes it very attractive for a number of *scenarii* such as in deep learning for generative modeling [Deshpande 2018, Deshpande 2019],

for barycenter computation [Bonneel 2015] or topological data analysis [Carrière 2017] to name a few. This "projection" idea was further developed and improved in several works which proposed to project on k-dimensional subspaces [Paty 2019], to use non-linear projection [Kolouri 2019a] or to generalize *SW* for the unbalanced setting [Bonneel 2019].

Many other interesting formulations can be derived from the original OT formulation. Since they are not considered in this manuscript we just give a brief overview here. In [Ferradans 2014] author propose to regularize the linear OT with a quadratic term, resulting on a quadratic regularized OT. For regular grids [Solomon 2015] define a Wasserstein distance that can be computed efficiently in $O(n^2 \log(n))$ using convolutions. Another line of works consider an unbalanced setting where the source probability measure is partially transferred to the target probability measure resulting on the unbalanced formulation [Chizat 2017]. A case of particular interest is when the target probability measure is discrete and the source continuous, namely the semi-discrete OT. It founds many applications in practice and can be tackled using Laguerre cells [Lévy 2018]. Finally the multi-marginal OT aims at solving an linear OT problem where there are many target/source probability measures and one optimal coupling for transporting them all [Nenna 2016].

2.1.6 Wasserstein barycenter

The Wasserstein distance is also an interesting tool in order to compute a notion of *barycenter* of probability distributions. In an Euclidean setting the traditional barycenter of points $(\mathbf{x}_i)_{i \in [\![m]\!]}$ can be computed by solving $\inf_{\mathbf{x} \in \mathbb{R}^d} \sum_{i=1}^n \lambda_i \|\mathbf{x} - \mathbf{x}_i\|_2^2$ where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. The barycenter vector is then given by $\mathbf{x} = \sum_{i=1}^n \lambda_i \mathbf{x}_i$. This can be generalized to arbitrary metric spaces $(\mathcal{X}, d)$ using the so-called Fréchet (or Karcher) mean [Karcher 2014]:

$$\inf_{x \in \mathcal{X}} \sum_{i=1}^n \lambda_i d(x, x_i)^p \tag{2.33}$$

for $p \in \mathbb{N}$. The problem (2.33) motivates the use of Wasserstein barycenter by considering the metric space $(\mathcal{P}_p(\Omega), W_p)$. Generally (2.33) is non-convex and difficult to solve for arbitrary metric space, however in the case of the Wasserstein distance the situation is somehow easier since it can be formulated as a convex problem for which existence can be proved and efficient numerical solvers exist. For a set of input probability measures $(\nu_i)_{i \in [\![n]\!]} \in \mathcal{P}(\Omega)$ the Wasserstein barycenter reads as the following variational problem:

$$\inf_{\mu \in \mathcal{P}(\Omega)} \sum_{i=1}^n \lambda_i \mathcal{T}_c(\mu, \nu_i). \tag{2.34}$$

The barycentric formulation finds many applications in machine learning such in Bayesian inference [Srivastava 2015], fairness [Gordaliza 2019], in image processing for texture synthesis and mixing [Rabin 2012] or in neuroimaging [Gramfort 2015] to name a few. As proven in [Agueh 2011] in the context of W_2^2 for $\Omega = \mathbb{R}^d$ this problem is convex and when one of the input measure has a density the barycenter is well-defined and unique. Even though there exist special cases (see Section 9.2 in [Peyré 2019]) in practice finding a solution in the general setting is difficult. In the following we detail one solution for the scenario where the input measures are discrete. More formally let $(\nu_i)_{i=1}^n$ be discrete probability measures with weights $\mathbf{b}_i \in \Sigma_{n_i}$ and that are supported on $\mathbf{Y}_i = (\mathbf{y}_q^i)_{q \in [\![n_i]\!]} \in \mathbb{R}^{n_i \times d}$ for each $i \in [\![n]\!]$. Instead of looking at all possible discrete probability measures we can search a k atoms probability measure *i.e.* of the form $\hat{\mu} = \sum_{p=1}^k a_p \delta_{\mathbf{x}_p}$ where $\mathbf{X} = (\mathbf{x}_p)_{p \in [\![k]\!]} \in \mathbb{R}^{k \times d}$ and $\mathbf{a} \in \Sigma_k$. Overall the resulting problem is:

$$\min_{\mathbf{a}\in\Sigma_k,\mathbf{X}\in\mathbb{R}^{k\times d}} \sum_{i=1}^{n} \lambda_i \min_{\boldsymbol{\pi}\in\Pi(\mathbf{a},\mathbf{b}_i)} \langle\boldsymbol{\pi},\mathbf{C}_{\mathbf{XY}_i}\rangle_{\mathcal{F}} = \min_{\substack{\mathbf{a}\in\Sigma_k,\mathbf{X}\in\mathbb{R}^{k\times d}\\ \forall i\in[\![n]\!],\boldsymbol{\pi}_i\in\Pi(\mathbf{a},\mathbf{b}_i)}} \sum_{i=1}^{n} \lambda_i\langle\boldsymbol{\pi}_i,\mathbf{C}_{\mathbf{XY}_i}\rangle_{\mathcal{F}} \tag{2.35}$$

where $\mathbf{C}_{\mathbf{XY}_i} \in \mathbb{R}^{k\times n_i}$ is the matrix defined by all pair to pair costs between the points of the barycenter and ν_i, *i.e.* $\mathbf{C}_{XY_i} = (c(\mathbf{x}_p,\mathbf{y}_q^i))_{p,q\in[\![k]\!]\times[\![n_i]\!]}$. In [Cuturi 2014] author propose to solve (2.35) using Block Coordinate Descent (BCD) that alternates between minimizing *w.r.t.* $\mathbf{a},\mathbf{X}$ and $\boldsymbol{\pi}_i$ while keeping others fixed:

(i) The minimization *w.r.t.* all $\boldsymbol{\pi}_i$ with $\mathbf{a},\mathbf{X}$ fixed involves solving n OT problems which can be done using algorithms described in Section 2.1.4.

(ii) The minimization *w.r.t.* $\mathbf{X}$ with $\mathbf{a},\boldsymbol{\pi}_i$ fixed can be performed in closed-form in the case $\Omega = \mathbb{R}^d$ and $c(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$ [Cuturi 2014, Equation 8]:

$$\mathbf{X} = \text{Diag}\left(\frac{1}{\mathbf{a}}\right)\left(\sum_{i=1}^{n}\lambda_i\boldsymbol{\pi}_i\mathbf{Y}_i\right) \tag{2.36}$$

(iii) The minimization *w.r.t.* the weight $\mathbf{a}$ with $\mathbf{X},\boldsymbol{\pi}_i$ fixed relies on the optimal dual variables of all OT sub-problems of step (i) and applies a projected subgradient minimization *w.r.t.* $\mathbf{a}$ as described in Algortihm 1 in [Cuturi 2014].

These three steps are repeated until convergence of $\mathbf{X}$ and $\mathbf{a}$. The major bottleneck of this approach is its computational complexity which is driven by the calculation of many OT problems. When the support $\mathbf{X}$ is fixed and by denoting $\mathbf{C}_{\mathbf{XY}_i} = \mathbf{C}_i$ the problem reduces to:

$$\min_{\substack{\mathbf{a}\in\Sigma_k\\ \forall i\in[\![n]\!],\boldsymbol{\pi}_i\in\Pi(\mathbf{a},\mathbf{b}_i)}} \sum_{i=1}^{n} \lambda_i\langle\boldsymbol{\pi}_i,\mathbf{C}_i\rangle \tag{2.37}$$

which is an (LP) with $kn^2 + n$ variables and $2Nn$ constraints. Note that first order methods such as subgradient descent on the dual have been proposed in [Carlier 2015] to solve (2.37) but in general its scale forbids the use generic solvers even for medium scale problems. These remarks advocate for the use of entropic regularized OT to obtain fast and smooth approximations of the original barycenter problem as given by:

$$\min_{\substack{\mathbf{a}\in\Sigma_k\\ \forall i\in[\![n]\!],\boldsymbol{\pi}_i\in\Pi(\mathbf{a},\mathbf{b}_i)}} \sum_{i=1}^{n} \lambda_i\langle\boldsymbol{\pi}_i,\mathbf{C}_i\rangle_{\mathcal{F}} - \varepsilon H(\boldsymbol{\pi}_i) \tag{2.38}$$

The resulting problem is a smooth convex minimization problem, which can be tackled using gradient descent [Cuturi 2014] or with descent method on the semi-dual [Cuturi 2018]. Another possibility is to rewrite (2.38) as a the following weighted KL minimization problem [Benamou 2015]:

$$\min_{\substack{(\boldsymbol{\pi}_i)_i\\ \forall i\in[\![n]\!],\boldsymbol{\pi}_i^T\mathbf{1}=\mathbf{b}_i\\ \boldsymbol{\pi}_1\mathbf{1}=\cdots=\boldsymbol{\pi}_n\mathbf{1}}} \sum_{i=1}^{n} \lambda_i\varepsilon KL(\boldsymbol{\pi}_i|\mathbf{K}_i) \tag{2.39}$$

where $\mathbf{K}_i = e^{-\frac{\mathbf{C}_i}{\varepsilon}}$. In this formulation the barycenter $\mathbf{a}$ is encoded in the row marginals of all the couplings $\boldsymbol{\pi}_i$ such that $\mathbf{a} = \boldsymbol{\pi}_1\mathbf{1} = \cdots = \boldsymbol{\pi}_n\mathbf{1}$. It is shown in [Benamou 2015] that this problem can also be solved using a generalized Sinkhorn algorithm which involves iterative projections. As such

the entropic regularization is quite suited for the barycenter problem and was further analyzed for the general case of continuous probability measures in [Bigot 2019a, Bigot 2019b]. Note that other methods have been proposed which rely *e.g.* on the sliced Wasserstein formulation [Bonneel 2015], unbalanced formulation [Chizat 2017] or on convolutions for geometric domains [Solomon 2015].

The case $n = 2$**: McCann interpolant** One special case deserves attention that is when $n = 2$ and in the case $\Omega = \mathbb{R}^d$ equipped with $\|.\|_2$. This setting corresponds to the so-called *McCann interpolant* [McCann 1997] where one wants to find:

$$\inf_{\mu \in \mathcal{P}(\mathbb{R}^d)} (1 - t)W_2^2(\mu, \nu_1) + tW_2^2(\mu, \nu_2) \tag{2.40}$$

with $t \in [0, 1]$ and ν_1 is regular with respect to the Lebesgue measure. Using Brenier theorem we know that there exists a unique push-forward such that $T \# \nu_1 = \nu_2$. In this case the barycenter is unique and obtained with $\mu_t = ((1-t)id + tT) \# \nu_1$. In practice when the probability measures ν_1, ν_2 are discrete with respectively n and m atoms this interpolant can be computed by $\mu_t = \sum_{i=1, j=1}^{m,n} \pi_{ij}^* \delta_{(1-t)\mathbf{x}_i + t\mathbf{y}_j}$ where π^* is an optimal coupling between ν_1, ν_2.

2.2 The Gromov-Wasserstein problem

2.2.1 Problem statement

Despite its valuable properties the linear OT problem faces the challenging problem of probability measures whose supports lie in incomparable spaces, that is to say when $\mathcal{X}, \mathcal{Y}$ are not part of a common ground metric space. For example when $\mu \in \mathcal{P}(\mathbb{R}^3), \nu \in \mathcal{P}(\mathbb{R}^2)$ the definition of a meaningful cost $c : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}_+$ is not straightforward. In particular in this setting we can not define a distance between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3 \times \mathbb{R}^2$ so that the Wasserstein distance can no longer be defined. Moreover the Wasserstein distance is not invariant to important families of invariants, such translations or rotations or more generally *isometries* which is an important flaw of linear OT for certain applications such as shape matching.

The Gromov-Wasserstein (GW) framework is an elegant remedy for this situation. It is built upon a quadratic Optimal Transport problem, as opposed to a linear one for the linear OT problem, and, informally its optimal value quantifies the metric distortion when transporting points from one space to another. This section aims at presenting the GW problem, its fundamental metric properties as well as numerical solvers. We refer the reader to [Sturm 2012, Memoli 2011, Chowdhury 2019a] for further readings.

We consider two polish spaces $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$. Let $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be continuous measurable functions and $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ be probability measures on $\mathcal{X}, \mathcal{Y}$. The Gromov-Wasserstein (GW) problem aims at finding:

$$GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{\mathcal{X} \times \mathcal{Y}} \int_{\mathcal{X} \times \mathcal{Y}} \left| c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(y, y') \right|^p \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y') \right)^{\frac{1}{p}} \tag{2.41}$$

for $p \in \mathbb{N}^*$ (see Figure 2.8). GW depends on the choice of similarities $c_{\mathcal{X}}, c_{\mathcal{Y}}$ between points in $\mathcal{X}$ and $\mathcal{Y}$. When it is clear form the context we will simply note $GW_p(\mu, \nu)$ instead of $GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu)$. Since $\mathcal{X}, \mathcal{Y}$ are already endowed with a natural metric one choice would be to consider $c_{\mathcal{X}} = d_{\mathcal{X}}$ and $c_{\mathcal{Y}} = d_{\mathcal{Y}}$. This setting brings in light the notion of *metric measure spaces*, as triplets of the form $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ where $(\mathcal{X}, d_{\mathcal{X}})$ is a complete separable metric space and μ is a Borel probability measure on $\mathcal{X}$. It was studied
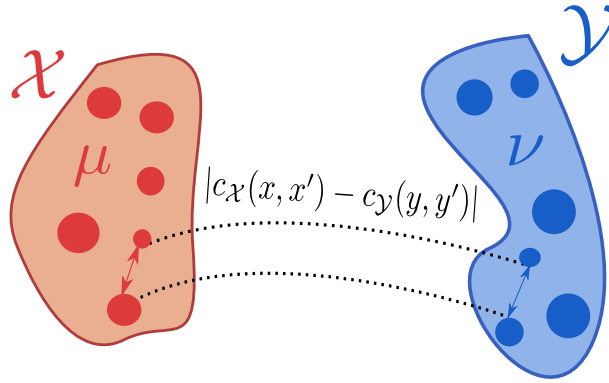
Figure 2.8:   The GW problem considers two probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ over two spaces that do not necessarily share a common metric. It is built upon the similarities $c_\mathcal{X}, c_\mathcal{Y}$ *within* each space and on a measure of the distortion between each pair of points $\left| c_\mathcal{X}(x, x') - c_\mathcal{Y}(y, y') \right|$.

in depth in [Sturm 2012]. Another possibility is to consider triplets $(\mathcal{X}, c_\mathcal{X}, \mu)$ where $c_\mathcal{X}$ is a integrable function, this notion refers to *measure networks* and was studied in [Chowdhury 2019a].

The GW objective is constructed so that if an optimal coupling π maps x to y and x' to y' then the couple (x, x') should be "as similar" in $\mathcal{X}$ as (y, y') in $\mathcal{Y}$. When $c_\mathcal{X}, c_\mathcal{Y}$ are distances it implies that x, x' are as close in $\mathcal{X}$ as y, y' in $\mathcal{Y}$. In this work we consider a general setting where $c_\mathcal{X}, c_\mathcal{Y}$ are continuous and $\mathcal{X}, \mathcal{Y}$ are Polish spaces and we will detail the two previous settings.

As for the linear OT problem the equation (2.41) always admits a solution. To show that we define $L(x, x', y, y') = \left| c_\mathcal{X}(x, x') - c_\mathcal{Y}(y, y') \right|$. If $\Pi(\mu, \nu)$ is compact and the functionnal $\pi \to \int \int L \mathrm{d}\pi \mathrm{d}\pi$ is l.s.c. for the weak-convergence, Weierstrass theorem (see Memo 2.2.1) proves that the infimum will be attained at some optimal coupling. The first condition is a well-known result in OT theory provided that $\mathcal{X}, \mathcal{Y}$ are Polish spaces [Santambrogio 2015, Theorem 1.7]. For the lower semi-continuity *w.r.t.* the weak-convergence we can show that it suffices that L be itself l.s.c. using the following lemma:

Lemma 2.2.1. *Let Ω be a Polish space. If $f : \Omega \times \Omega \to \mathbb{R}_+ \cup \{+\infty\}$ is lower semi-continuous, then the functional $J : \mathcal{P}(\Omega) \to \mathbb{R} \cup \{+\infty\}$ with $J(\mu) = \int \int f(w, w') \mathrm{d}\mu(w) \mathrm{d}\mu(w')$ is l.s.c. for the weak convergence of measures.*

Proof. Since f is l.s.c. and bounded from below by 0 we can consider $(f_k)_k$ a sequence of continuous and bounded functions converging increasingly to f (see e.g [Santambrogio 2015]). By the monotone convergence theorem $J_k(\mu) \to J(\mu) \stackrel{def}{=} \sup_k J_k(\mu) = \sup_k \int \int f_k d\mu d\mu$. Moreover every J_k is continuous for the weak convergence. Using theorem 2.8 [Billingsley 1999] on the Polish space $\Omega \times \Omega$ we know that if μ_n converges weakly to μ then the product measure $\mu_n \otimes \mu_n$ converges weakly to $\mu \otimes \mu$. In this way $\lim_{n \to \infty} J_k(\mu_n) = J_k(\mu)$ since f_k are continuous and bounded. In particular every J_k is l.s.c. We can conclude that J is l.s.c. as the supremum of l.s.c. functionals on the metric space of $(\mathcal{P}(\Omega), \delta)$ (see *e.g.* [Santambrogio 2015]). Here we equipped $\mathcal{P}(\Omega)$ with a metric δ as *e.g.* $\delta(\mu, \nu) = \sum_{k=1}^\infty 2^{-k} |\int_\Omega f_k d\mu - \int_\Omega f_k d\nu|$ (see remark 5.11 in [Ambrosio 2005]). $\qquad\square$

Memo 2.2.1 (Weierstrass theorem)**.** *The Weierstrass theorem states that if $f : \mathcal{X} \to \mathbb{R} \cup +\infty$ is l.s.c. and $\mathcal{X}$ is compact then there exists $x^* = \inf_{x \in \mathcal{X}} f(x)$ (see box 1.1 in [Santambrogio 2015]).*

Overall since L is l.s.c. due to the continuity of $c_{\mathcal{X}}, c_{\mathcal{Y}}$ we can apply Lemma 2.2.1 with $\Omega = \mathcal{X} \times \mathcal{Y}$ and conclude that $\pi \to \int \int L \mathrm{d}\pi \mathrm{d}\pi$ is l.s.c. for the weak-convergence and by the means of Weierstrass theorem equation (2.41) always admits a minimizer.

Remark 2.2.1. *As a consequence of our formulation the resulting GW cost main be infinite. A simple condition to remedy this possibility would be $\int \int \left| c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y') \right|^p \mathrm{d}(\mu \otimes \mu)(x,x') \mathrm{d}(\nu \otimes \nu)(y,y') < \infty$. Since:*

$$\left| c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y') \right|^p \le (|c_{\mathcal{X}}(x,x')| + |c_{\mathcal{Y}}(y,y')|)^p \overset{*}{\le} 2^{p-1}(|c_{\mathcal{X}}(x,x')|^p + |c_{\mathcal{Y}}(y,y')|^p) \tag{2.42}$$

if $c_{\mathcal{X}}, c_{\mathcal{Y}}$ are p-integrable functions, i.e. $c_{\mathcal{X}} \in L^p(\mu \otimes \mu)$ and $c_{\mathcal{Y}} \in L^p(\nu \otimes \nu)$ then the cost is finite (we used Hölder's inequality in (*) see Memo 2.2.2)

GW for measures on Euclidean spaces One special case will be consider in Chapter 4, that is when $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}^q$ and $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ with p not necessarily equal to q. We can consider the GW problem with the standard Euclidean distances for $c_{\mathcal{X}}, c_{\mathcal{Y}}$ on respectively $\mathbb{R}^p, \mathbb{R}^q$. This setting illustrates the invariance property of the GW problem *w.r.t.* rotations and translations. More precisely let $\mathbf{O} \in \mathcal{O}(p)$ and $\mathbf{x}_0 \in \mathbb{R}^p$ associated with $T(\mathbf{x}) = \mathbf{O}\mathbf{x} + \mathbf{x}_0$. Then the GW problem is invariant by T that is $GW_p^p(T\#\mu, \nu) = GW_p^p(\mu, \nu)$ (same applies for ν). To see that we simply used for all $\mathbf{x}, \mathbf{x}'$ $c_{\mathcal{X}}(T(\mathbf{x}), T(\mathbf{x}')) = \|\mathbf{O}\mathbf{x} + \mathbf{x}_0 - \mathbf{O}\mathbf{x}' - \mathbf{x}_0\|_2 = \|\mathbf{O}(\mathbf{x} - \mathbf{x}')\|_2 = \|\mathbf{x} - \mathbf{x}'\|_2$ since $\mathbf{O} \in \mathcal{O}(p)$. This property will be generalized for any metric space by considering the notion of *isometry* and contrasts with the Wasserstein distance which is not invariant neither to translation or rotation of the support of one probability measure.

Example 2.2.1. *As a first illustration of the GW problem we consider two discrete probability measures $\mu \in \mathcal{P}(\mathbb{R}^2), \nu \in \mathcal{P}(\mathbb{R}^3)$ following respectively a spiral in $\mathbb{R}^2$, $\mathbb{R}^3$ and composed of a mixture of Gaussian distributions. $c_{\mathcal{X}}, c_{\mathcal{Y}}$ are defined by the Euclidean distances between the points. We compute an optimal coupling of the GW problem using the FW solver presented in Chapter 3 (see Section 2.2.3 for more details). The result is depicted in Figure 2.9.*

Memo 2.2.2 (Hölder's inequality). *Let $(\mathcal{X}, \mu)$ be a measurable space and $(f, g) \in L^p(\mu) \times L^q(\mu)$ with $p, q > 0$ verifying $\frac{1}{p} + \frac{1}{q} = 1$. The Hölder's inequality states:*

$$\int |fg| \mathrm{d}\mu \le \left(\int |f|^p \mathrm{d}\mu \right)^{\frac{1}{p}} \left(\int |g|^q \mathrm{d}\mu \right)^{\frac{1}{q}} \tag{2.43}$$

As a corollary for $q \ge 1$ we have:

$$\forall x, y \in \mathbb{R}_+, \ (x + y)^q \le 2^{q-1}(x^q + y^q). \tag{2.44}$$

Indeed, if $q > 1$:
$(x+y)^q = \left(\left(\frac{1}{2^{q-1}}\right)^{\frac{1}{q}} \frac{x}{\left(\frac{1}{2^{q-1}}\right)^{\frac{1}{q}}} + \left(\frac{1}{2^{q-1}}\right)^{\frac{1}{q}} \frac{y}{\left(\frac{1}{2^{q-1}}\right)^{\frac{1}{q}}} \right)^q \le \left[\left(\frac{1}{2^{q-1}}\right)^{\frac{1}{q-1}} + \left(\frac{1}{2^{q-1}}\right)^{\frac{1}{q-1}} \right]^{q-1} \left(\frac{x^q}{\frac{1}{2^{q-1}}} + \frac{y^q}{\frac{1}{2^{q-1}}} \right)$
$= \frac{x^q}{\frac{1}{2^{q-1}}} + \frac{y^q}{\frac{1}{2^{q-1}}}.$
Last inequality is a consequence of Hölder's inequality. The result remains valid for $q = 1$.
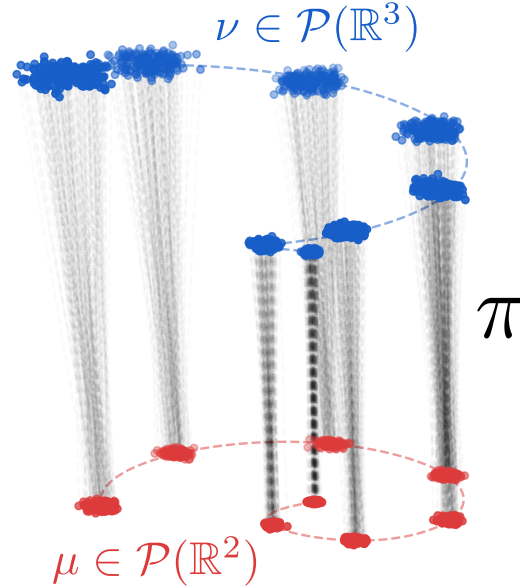
Figure 2.9: GW problem between two discrete probability measures $\mu \in \mathcal{P}(\mathbb{R}^2), \nu \in \mathcal{P}(\mathbb{R}^3)$. The optimal coupling π is depicted in dashed lines. It associates points so as to minimize the distortion between all pair-to-pair distances *within* the support of each measures.

2.2.2 Properties of GW

One of the main property of the *GW* problem is that it allows for comparing probability measures whose supports dwell in different, potentially non-related, spaces by defining a notion of equivalence of two probability distributions in this case. This is made possible thanks to the concepts of *isometry* and *isomorphism*.

Definition 2.2.1 (Isometry). *Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two metric spaces. An isometry is a sujective map $\phi : \mathcal{X} \to \mathcal{Y}$ that preserves the distances:*

$$\forall x, x' \in X, d_{\mathcal{Y}}(\phi(x), \phi(x')) = d_{\mathcal{X}}(x, x'). \tag{2.45}$$

An isometry is necessarily bijective, since for $\phi(x) = \phi(x')$ we have $d_{\mathcal{Y}}(\phi(x), \phi(x')) = 0 = d_{\mathcal{X}}(x, x')$ and hence $x = x'$ (in the same way ϕ^{-1} is also a isometry). When it exists, $\mathcal{X}$ and $\mathcal{Y}$ share the same "size" and any statement about $\mathcal{X}$ which can be expressed through its distance is transported to $\mathcal{Y}$ by the isometry ϕ.

Example 2.2.2. *Let us consider the two following graphs whose discrete metric spaces are obtained as shortest path between the vertices (see corresponding graphs in Figure 2.10).*

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 1 & 2 & 2 & 0 \end{pmatrix}}_{d_{\mathcal{X}}(x_i, x_j)} \quad and \quad \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}}_{d_{\mathcal{Y}}(y_i, y_j)}.$$
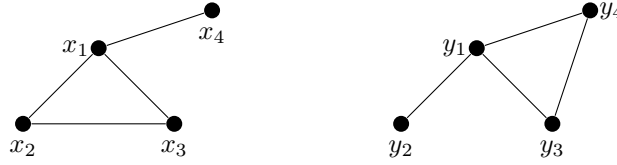
Figure 2.10: Two isometric metric spaces. Distances between the nodes are given by the shortest path, and the weight of each edge is equal to 1.

Figure 2.11: Two isometric but not isomorphic spaces.

These spaces are isometric since the surjective map ϕ such that $\phi(x_1) = y_1$, $\phi(x_2) = y_3$, $\phi(x_3) = y_4$, $\phi(x_4) = y_2$ verifies equation (2.45).

Another natural and straightforward example is two point clouds rotated from each other. More precisely if we consider $(\mathbf{x}_i)_{i\in[\![n]\!]}$, $(\mathbf{y}_i)_{i\in[\![n]\!]}$ where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^p \times \mathbb{R}^p$ equipped with the Euclidean norm $\|.\|_2$. Suppose that there exists a orthogonal matrix $\mathbf{O} \in \mathcal{O}(p)$ such that $\mathbf{y}_i = \mathbf{O}\mathbf{x}_i$ for all $i \in [\![n]\!]$ (with a slight abuse of notations we identify the matrix with its linear application). Then for all $(i,j) \in [\![n]\!]^2$ we have:

$$\|\mathbf{y}_i - \mathbf{y}_j\|_2 = \|\mathbf{O}\mathbf{x}_i - \mathbf{O}\mathbf{x}_j\|_2 = \|\mathbf{O}(\mathbf{x}_i - \mathbf{x}_j)\|_2 = \|\mathbf{x}_i - \mathbf{x}_j\|_2 \tag{2.46}$$

since $\mathbf{O} \in \mathcal{O}(p)$ so that $\mathcal{X} = (\mathbf{x}_i)_{i\in[\![n]\!]}$ and $\mathcal{Y} = (\mathbf{y}_i)_{i\in[\![n]\!]}$ are isometric.

This notion can be enriched in order to take into account the measures, which results in the notion of *strong isomorphism*:

Definition 2.2.2 (Strong isomorphism). *Let $(\mathcal{X}, d_\mathcal{X}), (\mathcal{Y}, d_\mathcal{Y})$ be Polish spaces and $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. We say that $(\mathcal{X}, d_\mathcal{X}, \mu)$ is strongly isomorphic to $(\mathcal{Y}, d_\mathcal{Y}, \nu)$ if there exists a bijection $\phi : supp(\mu) \to supp(\nu)$ such that:*

i ϕ is an isometry, i.e. $d_\mathcal{Y}(\phi(x), \phi(x')) = d_\mathcal{X}(x, x')$ for $x, x' \in supp(\mu)^2$

ii ϕ pushes μ forward to ν, i.e. $\phi\#\mu = \nu$

When it is clear from the context we will simply say that μ is strongly isomorphic to ν when previous conditions are satisfied.

Example 2.2.3. *Let us consider two mm-spaces $(\mathcal{X} = \{x_1, x_2\}, d_\mathcal{X} = \{1\}, \mu = \{\frac{1}{2}, \frac{1}{2}\})$ and $(\mathcal{Y} = \{y_1, y_2\}, d_\mathcal{Y} = \{1\}, \nu = \{\frac{1}{4}, \frac{3}{4}\})$ as depicted in Figure 2.11. These spaces are isometric but not isomorphic as there exists no measure preserving map which pushes μ forward to ν*

Another notion of isomorphism deserves attention especially when $c_\mathcal{X}, c_\mathcal{Y}$ are not distances. In this case we will consider the following *weak isomorphism* property:

Definition 2.2.3 (Weak isomorphism). *Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces and $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. We say that $(\mathcal{X}, c_\mathcal{X}, \mu)$ is weakly isomorphic to $(\mathcal{Y}, c_\mathcal{Y}, \nu)$ if there exists $(\mathcal{Z}, c_\mathcal{Z}, m)$, with $supp(m) = \mathcal{Z}$ and maps $\phi_0 : \mathcal{Z} \to \mathcal{X}, \phi_1 : \mathcal{Z} \to \mathcal{Y}$ such that:*

$\quad i\ c_{\mathcal{Z}}(z, z') = c_{\mathcal{X}}(\phi_0(x), \phi_0(x')) = c_{\mathcal{Y}}(\phi_1(x), \phi_1(x'))$ *for* $z, z' \in \mathcal{Z}^2$

$\quad ii\ \phi_0 \# m = \mu$ *and* $\phi_1 \# m = \nu$

When it is clear from the context we will simply say that μ is weakly isomorphic to ν when previous conditions are satisfied.

The weak isomorphism brings to light a kind of "tripod structure" in which the isomorphism is defined trough a third space $\mathcal{Z}$. In fact both notions are equivalent when $c_{\mathcal{X}}, c_{\mathcal{Y}}$ are distances as stated in the next proposition [Sturm 2012, lemma 1.10]:

Proposition 2.2.1. *The spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ are strongly isomorphic if and only if $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ are weakly isomorphic.*

However the weak-isomorphism property has its own interest when working with arbitrary similarity measures $c_{\mathcal{X}}, c_{\mathcal{Y}}$. The following theorem is fundamental for GW and aims to unify the metric properties of *GW* given in [Sturm 2012, Chowdhury 2019a]. It proves that GW defines a metric *w.r.t.* the isomorphism notions:

Theorem 2.2.1 (Metric properties of *GW*). *In the following $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$ are Polish spaces and $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$*

$\quad i\ GW_p$ *is symmetric, positive and satisfies the triangle inequality. More precisely for $(\mathcal{X}, c_{\mathcal{X}}, \mu)$, $(\mathcal{Y}, c_{\mathcal{Y}}, \nu)$, $(\mathcal{Z}, c_{\mathcal{Z}}, m)$ we have:*

$$GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) \leq GW_p(c_{\mathcal{X}}, c_{\mathcal{Z}}, \mu, m) + GW_p(c_{\mathcal{Z}}, c_{\mathcal{Y}}, m, \nu) \qquad (2.47)$$

$\quad ii\ GW_p(d_{\mathcal{X}}, d_{\mathcal{Y}}, \mu, \nu) = 0$ *if and only if $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ are strongly isomorphic.*

$\quad iii\ GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = 0$ *if and only if $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, c_{\mathcal{Y}}, \nu)$ are weakly isomorphic.*

$\quad iv\ $ *More generally, for any $q \geq 1$, $GW_p(d_{\mathcal{X}}^q, d_{\mathcal{Y}}^q, \mu, \nu) = 0$ if and only if $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ are strongly isomorphic.*

Proof. For the first point (i) positiveness is straightforward. For the triangle inequality and symmetry see [Chowdhury 2019a, Theorem 16]. For (ii) see [Sturm 2012, Lemma 1.10] and for (iii) see [Chowdhury 2019a, Theorem 18]. Note in the proof that for (iii) the result is still valid even if $c_{\mathcal{X}}, c_{\mathcal{Y}}$ are not p-integrable. For (iv) see [Sturm 2012, Lemma 9.2]. $\qquad\square$

This theorem can endow the space of all spaces of the form $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ with a distance defined by GW which, however, requires the finiteness of *GW*. More precisely:

Definition 2.2.4. *Let $\mathcal{X}$ be a Polish space, $\mu \in \mathcal{P}(\mathcal{X})$ and $c_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ be measurable. We define the size $size_p$ of $\mathcal{X}$, given $c_{\mathcal{X}}$ and μ by $size_p(\mathcal{X}, c_{\mathcal{X}}, \mu) = (\int c_{\mathcal{X}}(x, x')^p d\mu(x) d\mu(x'))^{\frac{1}{p}}$*

We define $\mathbb{X}_p$ be the space of all metric measure spaces with finite L^p-size, i.e $\mathbb{X}_p = \{(\mathcal{X}, d_{\mathcal{X}}, \mu) | size_p(\mathcal{X}, d_{\mathcal{X}}, \mu) < +\infty\}$ where $(\mathcal{X}, d_{\mathcal{X}})$ is a Polish space and $\mu \in \mathcal{P}(\mathcal{X})$.

We define also $\mathbb{N}_p$ be the space of all network measure spaces [Chowdhury 2019a] with finite L^p-size, i.e. $\mathbb{N}_p = \{(\mathcal{X}, c_{\mathcal{X}}, \mu) | size_p(\mathcal{X}, c_{\mathcal{X}}, \mu) < +\infty\}$ where $\mathcal{X}$ is a Polish space, $\mu \in \mathcal{P}(\mathcal{X})$ and $c_{\mathcal{X}}$ a continuous measurable function.

The function $size_p$ quantifies somehow an average diameter of $\mathcal{X}$ given a probability measure and a function $c_{\mathcal{X}}$. Using this notion and Theorem 2.2.1 we now state the main theorem about the metric properties of GW:

Theorem 2.2.2 (GW is a distance). *GW_p is a distance on $\mathbb{X}_p$ quotiented by the strong isomorphisms. GW_p is a distance on $\mathbb{N}_p$ quotiented by the weak isomorphisms.*

This theorem has a lot of implications. It endows the space of all metric (network) measure spaces with a topology, a geometric structure, induced by Gromov-Wasserstein and, as such, allows the use of a wide family of geometric tools and a notion of convergence of metric measure spaces. Moreover it indicates that GW is well suited for comparing objects with respect to a large class of invariants that are for instance rotations, translations or permutations. This property is important *e.g.* for shape comparison where the orientation of a shape does not define its nature or for graphs where any permutation of the nodes result in the same graph. It is sometimes valuable to have a notion of distance which is insensitive to these transformations so as to focus properly on what matters rather than to encode the invariance (see Remark 2.2.2). Finally if GW vanishes it implies necessarily that the objects are isomorphic which is interesting for detecting such cases.

Note also that GW is deeply connected to the Gromov-Hausdorff distance [Gromov 1999] that aims at measuring how far are $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ from being isometric and can be used for studying the convergence of metric spaces [Burago 2001]. However computing this distance results in a highly non-convex optimization problem whose global solution is untractable. As shown in [Memoli 2011] the introduction of measures turns out to "smooth" the definition of the Gromov-Hausdorff distance and results in the Gromov-Wasserstein distance.

Remark 2.2.2 (Implicit or explicit encoding of the invariances). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ and $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a cost. In [Alvarez-Melis 2019] authors propose the following OT problem:*

$$InvOT(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \min_{f \in F} \int c(\mathbf{x}, f(\mathbf{y})) \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \tag{2.48}$$

where F is a class of functions from $\mathbb{R}^d$ to $\mathbb{R}^d$ aiming at encoding a global transformation of the features. For example F can be defined as $\mathcal{O}(d)$ the set of orthogonal transformations or any linear transformation with bounded Schatten norm (see [Alvarez-Melis 2019] or Chapter 4 for more details). When considering $F = \mathcal{O}(d)$ the resulting InvOT becomes invariant by rotation of the support of the target measure. It could be interesting in a setting where one wants to match two distributions modulo a rotation as e.g. in unsupervised word translation where word embedding algorithms are known to produce vectors intrinsically invariant to angle [Alvarez-Melis 2019, Grave 2019]. This approach can be put into perspective with the GW distance as, in the GW case, one makes the implicit assumption, or prior, that the invariances are the isometric transformations of the data, whereas in the InvOT approach one makes the prior that we know somehow which class of invariances is of interest for the problem and we encode it into the loss. Both approach are relevant and related (see Chapter 4): if we have another prior than isometric transformations it is maybe more suited to encode it directly via F and if we know that isometries are relevant we can directly build upon GW.

Geodesics and GW interpolation The space of all mm-spaces $\mathbb{X}_p$ endow with GW has also a nice geodesic structure which is important in order to derive dynamic formulations and gradient flows [Ambrosio 2005]. Informally in $\mathbb{X}_p$ we can connect any two points (that are mm-spaces) with a curve

that somehow represents the shortest path connecting these points. More precisely given two mm-spaces $(\mathcal{X}, d_\mathcal{X}, \mu), (\mathcal{Y}, d_\mathcal{Y}, \nu)$ the curve $t \in [0,1] \to (\mathcal{X} \times \mathcal{Y}, d_t, \pi^*)$ where π^* is an optimal coupling of the GW problem between μ and ν and:

$$\forall (x, y), (x', y') \in (\mathcal{X} \times \mathcal{Y})^2, \ d_t((x, y), (x', y')) = (1 - t) d_\mathcal{X}(x, x') + t d_\mathcal{Y}(y, y') \tag{2.49}$$

is a geodesic [Sturm 2012]. However computing this geodesic is often intractable in practice since it implies the calculation of the cartesian product $\mathcal{X} \times \mathcal{Y}$. One can rely instead on the barycenter formulation defined in Section 2.2.4.

2.2.3 Solving GW

In this section we describe some numerical solutions to the GW problem. In the following $\mu = \sum_{i=1}^n a_i \delta_{x_i} \in \mathcal{P}(\mathcal{X})$, $\nu = \sum_{j=1}^m b_j \delta_{y_j} \in \mathcal{P}(\mathcal{Y})$ are discrete probability measures over respectively $(\mathcal{X}, d_\mathcal{X}), (\mathcal{Y}, d_\mathcal{Y})$. We note also $\mathbf{C}_1, \mathbf{C}_2$ the matrices of pair-to-pair distances inside each space, *i.e.* $\forall (i, k) \in [\![n]\!]^2, \ C_1(i, k) = d_\mathcal{X}(x_i, x_k)$ and $\forall (j, l) \in [\![m]\!]^2, \ C_2(j, l) = d_\mathcal{Y}(y_j, y_l)$. The GW problem aims at solving:

$$GW_p^p(\mathbf{C}_1, \mathbf{C}_2, \mathbf{a}, \mathbf{b}) = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i,j,k,l} |C_1(i, k) - C_2(j, l)|^p \pi_{i,j} \pi_{k,l} = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^p \otimes \boldsymbol{\pi}, \boldsymbol{\pi} \rangle_\mathcal{F} \tag{2.50}$$

where we define $L(\mathbf{C}_1, \mathbf{C}_2)$ as the tensor $\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2) = (|C_1(i, k) - C_2(j, l)|)_{i,j,k,l}$ and $\otimes$ is the the tensor-matrix multiplication, *i.e.* for a tensor $\mathbf{L} = (L_{i,j,k,l})$, $\mathbf{L} \otimes \boldsymbol{\pi}$ is the matrix $\left(\sum_{k,l} L_{i,j,k,l} \pi_{k,l} \right)_{i,j}$.

The optimization problem (2.50) is a non-convex Quadratic Program (QP) which is NP-hard in general [Loiola 2007] and notoriously hard to approximate. When $p = 2$, *i.e.* when $\mathbf{L} = |.|^2$ equation (2.50) can be recast as:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \text{tr}(\mathbf{c}_{\mathbf{C}_1, \mathbf{C}_2} \boldsymbol{\pi}^T) - 2\text{tr}(\mathbf{C}_1 \boldsymbol{\pi} \mathbf{C}_2 \boldsymbol{\pi}^T) \tag{2.51}$$

where $\mathbf{c}_{\mathbf{C}_1, \mathbf{C}_2} = (\mathbf{C}_1)^2 \mathbf{a} \mathbf{1}_m^T + \mathbf{1}_n \mathbf{b}^T (\mathbf{C}_2)^2$ (see Proposition 1 in [Peyré 2016]). In standard QP form this problem reads also:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \mathbf{c}^T \mathbf{x}(\boldsymbol{\pi}) + \frac{1}{2} \mathbf{x}(\boldsymbol{\pi})^T \mathbf{Q} \mathbf{x}(\boldsymbol{\pi}) \tag{2.52}$$

where $\mathbf{x}(\boldsymbol{\pi}) = \text{vec}(\boldsymbol{\pi}), \mathbf{c} = \text{vec}(\mathbf{c}_{\mathbf{C}_1, \mathbf{C}_2})$ and $\mathbf{Q} = -4\mathbf{C}_1 \otimes_K \mathbf{C}_2$ with $\otimes_K$ the Kronecker product of matrices defined for two arbitrary matrices $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{B}^{p \times q}$ as $\mathbf{A} \otimes_K \mathbf{B} \in \mathbb{R}^{np \times mq}$ with $\mathbf{A} \otimes_K \mathbf{B} = (A_{i,j} \mathbf{B})_{i,j}$ (see Memo 2.2.3). Equation (2.52) is a non-convex QP as the Hessian $\mathbf{Q}$ is not positive semi-definite in general (its eigenvalues are the products of the eigeinvalues of $\mathbf{C}_1, \mathbf{C}_2$).

Relations with Quadratic Assignment Problem and Graph Matching The *GW* problem is very related to the so called *Quadratic Assignment Problem* (QAP). This problem was first introduced by Koopmans and Beckmann [Koopmans 1957] to model a plant location problem and plays today many roles in optimization. Given two matrices $\mathbf{A} = (a_{i,j})_{(i,j) \in [\![n]\!]^2}$ and $\mathbf{B} = (b_{i,j})_{(i,j) \in [\![n]\!]^2}$ the standard form for the QAP reads:

$$\min_{\sigma \in S_n} \sum_{i=1, j=1}^n a_{\sigma(i)\sigma(j)} b_{ij} \tag{2.53}$$

The QAP can be understood as a *facility location* problem: given n facilities and n locations, one wants to assign each facility to a location with a_{ij} the flow of material moving from facility i to facility j

and b_{ij} the distance from facility i to facility j. In this context the cost of *simultaneously* locating facility $\sigma(i)$ to location i and facility $\sigma(j)$ to location j is $a_{\sigma(i)\sigma(j)}b_{ij}$. In this model one wants to find the assignment that minimizes the overall cost of locating each facility. This problem was considered for example in [Elshafei 1977] for locating hospital departments so as to minimize the total distance traveled by patients but it also covers a large variety of applications such as scheduling [Geoffrion 1976], parallel and distributed computing [Bokhari 1981] or balancing of turbine runners [Laporte 1988]. For a comprehensive survey on this topic we refer to [Çela 2013, Loiola 2007]. Unfortunately the QAP is NP-Hard in general and only few special cases are known to be solvable in polynomial time. That is the case for example when matrices $\mathbf{A}$ and $\mathbf{B}$ have simple known structures, such as a diagonal structure and Toeplitz or separability properties such as $a_{i,j} = \alpha_i\alpha_j$ [Çela 2018, Çela 2011, Çela 2015]). The question of finding a polynomial time algorithm that solves the QAP when $\mathbf{A}$ and $\mathbf{B}$ satisfy the Monge and the anti-Monge properties are, to the best of our knowledge, still open [Çela 2013].

The QAP is intrinsically linked with the *graph matching* problem whose literature is also extensive (see *e.g.* [Berg 2005, Lyzinski 2016, Zaslavskiy 2009, Maron 2018, Caetano 2009]). The graph matching problem refers to optimization problems where the goal is to match edge affinities of two graphs that are represented by symmetric matrices $\mathbf{A} = (a_{i,j})_{(i,j)\in[\![n]\!]^2}$ and $\mathbf{B} = (b_{i,j})_{(i,j)\in[\![n]\!]^2}$. A common approach for these types of problem is to attempt to solve:

$$\min_{\mathbf{X}\in\Pi_n} \|\mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B}\|_{\mathcal{F}}^2 \tag{2.54}$$

where Π_n is the set of permutation matrices, *i.e.* $\mathbf{X} = (x_{ij})_{(i,j)\in[\![n]\!]^2} \in \Pi_n$ if $x_{ij} \in \{0,1\}$ and $\sum_{i=1}^n x_{ij} = \sum_{j=1}^n x_{ij} = 1$. By noticing that $\|\mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{B}\|_{\mathcal{F}}^2 = \|\mathbf{A}\mathbf{X}\|_{\mathcal{F}}^2 + \|\mathbf{X}\mathbf{B}\|_{\mathcal{F}}^2 - 2\mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}^T)$ and that $\|\mathbf{A}\mathbf{X}\|_{\mathcal{F}}^2 = \mathrm{tr}(\mathbf{X}^T\mathbf{A}^T\mathbf{A}\mathbf{X}) = \mathrm{tr}(\mathbf{X}\mathbf{X}^T\mathbf{A}^T\mathbf{A}) = \mathrm{tr}(\mathbf{A}^T\mathbf{A}) = \|\mathbf{A}\|_{\mathcal{F}}^2$ the problem (2.54) is equivalent to:

$$\min_{\mathbf{X}\in\Pi_n} -\mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}^T) \tag{2.55}$$

As such the graph matching problem is a QAP and consequently is NP-Hard in general. A way of finding a approximate solution is to consider a relaxation of the constraints by replacing Π_n by its convex-hull, namely the set of doubly stochastic matrices $DS = \{\mathbf{X} \in \mathbb{R}^{n\times n} | \mathbf{X}\mathbf{1}_n = \mathbf{1}_n, \mathbf{X}^T\mathbf{1}_n = \mathbf{1}_n, \mathbf{X} \geq 0\}$ (see [Dym 2017, Bernard 2018, Schellewald 2001]).

The previous discussion can relate the *GW* problem with the graph matching one. Indeed when we consider two discrete probability measures with the same number of atoms and with uniform weights (*i.e.* $\mathbf{a} = \mathbf{b} = \mathbf{1}_n/n$) then *GW* is equivalent to the relaxation of the graph matching problem with affinity matrices $\mathbf{C}_1, \mathbf{C}_2$. To see this it suffices to notice that, in this case, $\Pi(\mathbf{a}, \mathbf{b})$ is the set of doubly stochastic matrices (modulo a factor n which has no impact). We will see in Chapter 4 that the QAP and graph matching point of views can be quite enlightening for deriving properties of the GW distance.

Entropic regularization   In [Peyré 2016, Solomon 2016] authors propose to solve (2.50) using entropic regularization which results in the following optimization problem:

$$\min_{\boldsymbol{\pi}\in\Pi(\mathbf{a},\mathbf{b})} \langle \mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^p \otimes \boldsymbol{\pi}, \boldsymbol{\pi}\rangle_{\mathcal{F}} - \varepsilon H(\boldsymbol{\pi}). \tag{2.56}$$

This is a non-convex optimization which was tackled using projected gradient descent using the geometry of the KL divergence for both the gradient step and the projection step [Peyré 2016]. More precisely by denoting $E_\varepsilon(\boldsymbol{\pi})$, the loss in (2.56) the iterations of this algorithm read:

$$\boldsymbol{\pi} \leftarrow \mathrm{Proj}_{\Pi(\mathbf{a},\mathbf{b})}^{KL}(\boldsymbol{\pi} * e^{-\tau\nabla E_\varepsilon(\boldsymbol{\pi})}) \tag{2.57}$$

Algorithm 3 Solving Entropic-regularized GW

1: $\mathbf{a}, \mathbf{b}, \mathbf{C}_1, \mathbf{C}_2, \varepsilon > 0$

2: Initalize $\boldsymbol{\pi} = \mathbf{a}\mathbf{b}^T, \mathbf{u}^{(0)}, \mathbf{v}^{(0)} = \mathbf{1}$

3: **for** $i = 1, \ldots, n_{it}^1$ **do**

4:     Compute the gradient of the GW loss $\mathbf{C} = 2\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^p \otimes \boldsymbol{\pi}$

5:     Set $\mathbf{K} = \exp(-\frac{\mathbf{C}}{\varepsilon})$

6:      // Do Sinkhorn-Knopp Algorithm 2

7:     **for** $i = 1, \ldots, n_{it}^2$ **do**

8:         $\mathbf{u}^{(i)} = \mathbf{a} \oslash \mathbf{K}^\top \mathbf{v}^{(i-1)}$ // Update left scaling

9:         $\mathbf{v}^{(i)} = \mathbf{b} \oslash \mathbf{K}\mathbf{u}^{(i-1)}$ // Update right scaling

10:    **end for**

11: **end for**

12: **return** $\boldsymbol{\pi}^* = \text{Diag}(\mathbf{u})\mathbf{K}\,\text{Diag}(\mathbf{v})$

---

where $\tau > 0$ is the step size of the gradient descent and $*$ denotes elementwise (Hadamard) matrix multiplication. The projection operator is defined as the result of the minimization problem:

$$\text{Proj}_{\Pi(\mathbf{a},\mathbf{b})}^{KL}(\mathbf{K}) = \underset{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})}{\arg\min}\, KL(\boldsymbol{\pi}|\mathbf{K}) \tag{2.58}$$

As shown in Section 2.1.4 the projection can be solved using the efficient Sinkhorn-Knopp Algorithm (see Algorithm 2). The gradient $\nabla E_\varepsilon(\boldsymbol{\pi})$ can be calculated as $2\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^p \otimes \boldsymbol{\pi} + \varepsilon \log(\boldsymbol{\pi})$, and, as noted in [Peyré 2016], the special case where the step size τ is defined as $\tau = \frac{1}{\varepsilon}$ the iterations (2.57) boil down to solving an entropic regularized linear OT problem with ground cost $2\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^p \otimes \boldsymbol{\pi}$. Overall the procedure to solve GW with entropic regularization is a projected gradient procedure where the projection step can be solved using an entropic linear OT as described in Algorithm 3. Note that, as for the linear OT case, the resulting optimal coupling is not sparse since the entropy term tends to blur the solution. Moreover one usually wants to have a gradient step τ which is not "too big" in order to ensure the convergence of the algorithm. Since the gradient step τ is inversely proportional to the regularization parameter ε this comes at the price of blurring the resulting optimal solution. In practice, there is a trade-off between regularization and convergence of the algorithm (see [Peyré 2016] and experiments of Chapter 4).

Memo 2.2.3 (Kronecker product and vec operator). *Let* $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{B}^{p \times q}$ *be two arbitrary matrices. The vec operator converts the matrix into a column vector. It is defined as* $vec(\mathbf{A}) \in \mathbb{R}^{nm \times 1} = (A_{1,1}, \ldots, A_{n,1}, A_{1,2}, \ldots, A_{n,2}, \ldots, A_{n,m})^T$. *The Kornecker product of two matrices result in a block matrix* $\mathbf{A} \otimes_K \mathbf{B} \in \mathbb{R}^{np \times mq}$ *defined by* $\mathbf{A} \otimes_K \mathbf{B} = (A_{i,j}\mathbf{B})_{i,j}$. *These two operators satisfy the following properties (see [Petersen 2012]):*

- $vec(\mathbf{A}\mathbf{B}) = (\mathbf{I} \otimes_K \mathbf{A})vec(\mathbf{B}) = (\mathbf{B}^T \otimes_K \mathbf{I})vec(\mathbf{A})$

- $(\mathbf{A} \otimes_K \mathbf{B})^T = (\mathbf{A}^T \otimes_K \mathbf{B}^T)$

- $(\mathbf{A} \otimes_K \mathbf{B})(\mathbf{B} \otimes_K \mathbf{A}) = (\mathbf{A}\mathbf{B} \otimes_K \mathbf{B}\mathbf{A})$

- $tr(\mathbf{A}^T\mathbf{B}) = vec(\mathbf{A})^T vec(\mathbf{B})$

When $p = 2$ the previous problem reduces to the *softassign quadratic assignement* problem. In the special case where the problem is convex the convergence of the previous scheme was analyzed in [Rangarajan 1997b, Rangarajan 1999] but, with arbitrary matrices $\mathbf{C}_1, \mathbf{C}_2$ there is no known results to the best of our knowledge.

Computing a lower-bound Originally (2.50) was tackled by computing a lower bound (called the **TLB**) in [Memoli 2011]. More precisely, authors propose to solve the following problem:

$$\min_{\boldsymbol{\pi}^1 \in \Pi(\mathbf{a},\mathbf{b})} \sum_{k,l} \left(\min_{\boldsymbol{\pi}^2 \in \Pi(\mathbf{a},\mathbf{b})} \sum_{i,j} |C_1(i,k) - C_2(j,l)|^p \pi^2_{i,j} \right) \pi^1_{k,l} \qquad (2.59)$$

This problem is actually a "Wasserstein of Wasserstein distances": it is equivalent to solve an OT problem which ground cost results itself on OT problems between the 1D empirical distributions of the lines of $\mathbf{C}_1, \mathbf{C}_2$. More precisely by considering $\mu_k = \sum_i a_i \delta_{C_1(i,k)} \in \mathcal{P}(\mathbb{R})$ and $\nu_l = \sum_j b_j \delta_{C_2(j,l)} \in \mathcal{P}(\mathbb{R})$ (2.59) is equivalent to:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \sum_{k,l} W_p^p(\mu_k, \nu_l) \pi_{k,l} \qquad (2.60)$$

where the Wasserstein distance is computed using $|.|$ as ground cost. The advantage of this formulation is that (2.60) only involves linear OT problems that can be solved using tools presented in Section 2.1.4. This idea is based on the *local distribution of distances* and was also successfully applied in computer graphics for 3D shape comparison [Gelfand 2005, Memoli 2011].

Computational complexities One major bottleneck for computing GW is first the calculation of the big tensor $(|C_1(i,k) - C_2(j,l)|^p)_{i,j,k,l}$ which is $O(n^2 m^2)$ is general. By noticing that it suffices instead to compute $\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^p \otimes \boldsymbol{\pi}$ authors in [Peyré 2016] show that, in the case $p = 2$, one can rely on the separability of $\mathbf{L}$ which results in a $O(n^2 m + m^2 n)$ complexity. A second bottleneck is the complexity of finding an optimal solution which is driven by the algorithmic method. The convergence of Algorithm 3 for the entropic-regularized GW problem is still not well-understood and slow in practice as shown for example in Chapter 4. We will present in Chapter 3 an algorithm based on Frank-Wolfe (FW) to find a sparse local optimal solution with a cubic complexity. Using the FW properties we will show that this algorithm also converges to a local stationary point with a $O(\frac{1}{\sqrt{t}})$ rate. Regarding the lower bound computation one can rely on the sorting strategies for 1D distributions to compute an inner Wasserstein distance $W_p^p(\mu_i, \nu_j)$ in $O(n \log(n) + m \log(m))$ hence a $O(n^2 m \log(n) + m^2 n \log(m))$ complexity for all pairs. Then finding a solution $\boldsymbol{\pi}$ has the same complexity as computing a linear OT problem and one can rely on entropic regularization to reach a quadratic time complexity. More recently, in [Sato 2020] authors propose to fix the outer optimal transport plan of the lower bound to $\mathbf{a}\mathbf{b}^T$ which results in a divergence that can be computed in $O((n^2 + m^2) \log(nm))$ using a sweep line strategy.

Illustration of previous solvers In order to give a simple illustration of the different solvers presented above we consider two unlabeled graphs with the same number of nodes ($n = m = 20$) and with 4 communities. Each node of the graph lies in the implicit metric space defined by the shortest-path distance inside the graph so that $\mathbf{C}_1$ is the shortest path distance matrix between each node (same for $\mathbf{C}_2$). We consider uniform weights, *i.e.* $\mathbf{a} = \mathbf{b} = \frac{1}{n}\mathbf{1}_n$ and $p = 2$ for GW. We solve the GW problem by relying on (1) The FW algorithm defined in Chapter 3, (2) The entropic regularized GW problem with $\varepsilon = 8e - 4$, (3) The lower bound approach **TLB**. The behavior of the three approaches is depicted in Figure 2.12
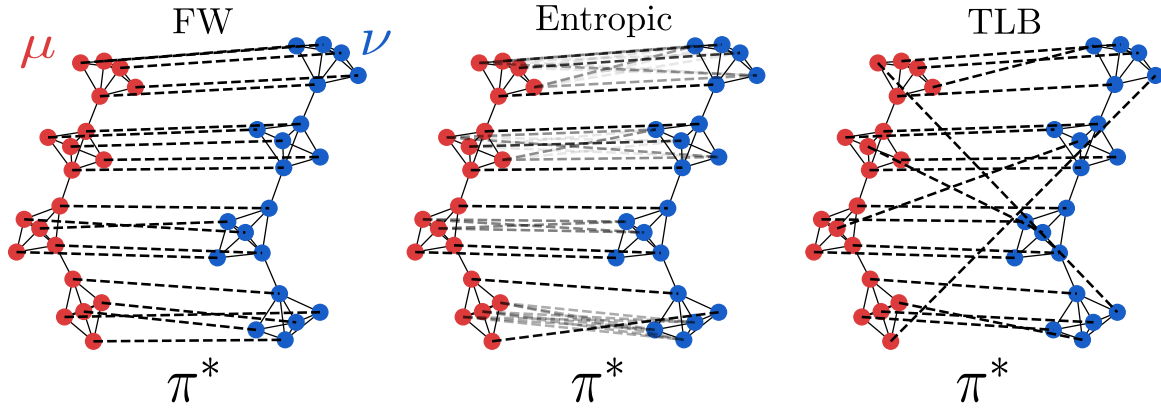
Figure 2.12:   Illustration of the optimal coupling for the GW problem between two graphs described by their shortest-path. **(left)** FW solver of Chapter 3 **(middle)** Entropic regularized GW (Algorithm 3) **(right)** Lower bound approach **TLB**. The optimal coupling $\boldsymbol{\pi}^*$ is depicted in dashed lines. The darker the stronger.

2.2.4   Gromov-Wasserstein Barycenter

In the same vein as the Wasserstein barycenter we can build upon the Gromov-Wasserstein distance a notion of barycenter. This setting was first tackled in [Peyré 2016] where authors consider the problem of computing the barycenter of a family of discrete probability measures over different metric spaces *w.r.t.* the Gromov-Wasserstein geometry. More formally let $(\mathbf{C}_i, \mathbf{b}_i)_{i \in [\![n]\!]}$ be this family where $\mathbf{C}_i$ is a arbitrary matrix and $\mathbf{b}_i$ is a probability vector. $\mathbf{C}_i$ can be chosen to be a distance matrix or more generally any similarity matrix, such as a kernel matrix encoding a notion of similarity between the points inside each distribution as in [Chowdhury 2019a]. When the weights of the barycenter are given and fixed to $\mathbf{a} \in \Sigma_k$ with $k \in \mathbb{N}^*$ the GW barycenter problem aims at finding:

$$\min_{\mathbf{C} \in \mathbb{R}^{k \times k}} \sum_{i=1}^{n} \lambda_i GW_p^p(\mathbf{C}, \mathbf{C}_i, \mathbf{a}, \mathbf{b}_i) = \min_{\substack{\mathbf{C} \in \mathbb{R}^{k \times k} \\ \forall i \in [\![n]\!], \boldsymbol{\pi}_i \in \Pi(\mathbf{a}, \mathbf{b}_i)}} \sum_{i=1}^{n} \lambda_i \langle \mathbf{L}(\mathbf{C}, \mathbf{C}_i)^p \otimes \boldsymbol{\pi}, \boldsymbol{\pi} \rangle_{\mathcal{F}} \qquad (2.61)$$

with $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. In [Peyré 2016] authors consider this problem based on the entropic regularized version of GW. They propose to solve (2.61) by relying on a BCD procedure which alternates between solving n GW problems with $\mathbf{C}$ fixed and by finding $\mathbf{C}$ with all $\boldsymbol{\pi}_i$ fixed. The latter is given in closed-form when $p = 2$ by [Peyré 2016, Equation 14]:

$$\mathbf{C} = \frac{1}{\mathbf{a}\mathbf{a}^T} \sum_{i=1}^{n} \lambda_i \boldsymbol{\pi}_i^T \mathbf{C}_i \boldsymbol{\pi}_i \qquad (2.62)$$

where the division is made element-wise. Solving the n GW problems can be done using Algorithm 3 as in [Peyré 2016] or with the GW or the lower bound approach as presented in the previous section. This approach was further generalized in [Chowdhury 2019b] where authors leverage the Riemannian geometry of Gromov-Wasserstein space to treat *e.g.* the case where $\mathbf{C}_i$ are not necessarily symmetric.

Illustration   To illustrate the GW barycenter we consider a simple dataset consisting in 4 2D-shapes from the apple class of the MPEG-7 computer vision database. Each shape is associated with a discrete probability measure with uniform weights. The matrices $\mathbf{C}_i$ are simply the Euclidean distances between
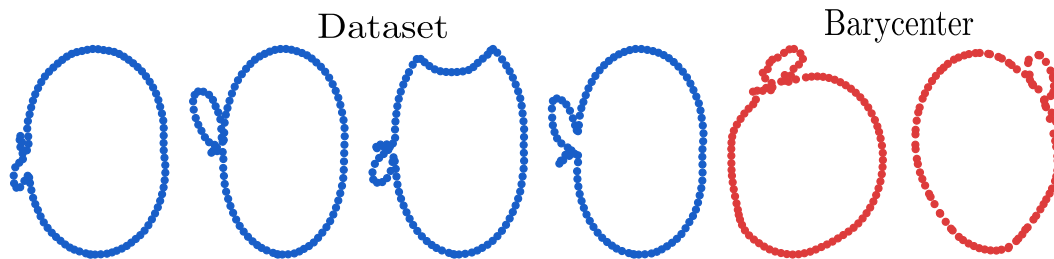
Figure 2.13: GW barycenter of 4 apple shapes in blue. The barycenters are depicted in the right side of the Figure in red. The first barycenter uses the FW algorithm for solving the GW problems and the second is computed using entropic regularization solved with Algorithm 3. Note that the barycenter are arbitrarily rotated due to the MDS procedure.

the points in each shape. We compute the barycenter using the previous discussion with both the FW solver for solving the GW problems and using the entropic regularized GW ($\varepsilon = 1e-3$) with Algorithm 3. We compute a MDS on the **C** matrix obtained by the barycenter procedure to recover 2D points. Results are depicted in Figure 2.13.

Applications of GW The GW problem is well suited for comparing heterogeneous data while being invariant to the isometries of the data. As such it has first received attention for shapes comparison [Memoli 2011, Solomon 2016] or in computer vision [Schmitzer 2013] where it is often valuable to compare objects without any assumption on their orientation. GW was further exploited to handle unstructured geometric data such as point clouds or meshes in [Ezuz 2017] where authors use GW to learn regular 2D grids that faithfully represent the 3D meshes while being applicable for standards CNN architectures. More recently GW has been the subject of much attention in the graph community as a graph matching tool [Xu 2019b, Xu 2019a, Fey 2020] or for graphs representation [Kwon 2020] (see Chapter 4 for more details). It also proves its usefulness in cellular biology thanks to its ability for aligning heterogeneous types of single-cell measurements [Demetci 2020]. Closer to the machine learning community GW has been applied in Domain Adaptation (DA) in the complex settings of Unsupervised DA [Xia 2020] and Heterogeneous (DA) [Yan 2018] (see Chapter 5 for more details), in generative modeling on incomparable spaces *i.e.* when the data generated do not share the same Euclidean space as the source data [Bunne 2019], or for cross-lingual correspondences of word embeddings [Alvarez-Melis 2018a].

2.3 Conclusion

The optimal transport framework is a powerful tool for comparing probability distributions by relying on both Wasserstein and Gromov-Wasserstein distances. The linear OT theory is a well-studied problem both from the theoretical and numerical side. In contrast a lot of questions remain unanswered for the Gromov-Wasserstein theory and interesting connections with the Wasserstein theory can be made. From a theoretical perspective there is no known result yet about the regularity of the optimal transport plans. As for the practical side the GW problem remains very costly to solve and difficult to approximate. The purpose of the next chapters is, *inter alia*, to work towards these directions and to address the following questions:

    i Is there any favorable cases where Optimal Transport plans of the GW are supported on a Monge

map as in the Brenier theorem? (see Chapter 4)

ii Can we find some special cases where GW admits a closed-form solution such as the 1D or the Gaussian cases? If so, can we derive scalable and useful formulations from these cases? (see Chapter 4)

iii Is the GW framework suited for the structured data setting? How does it behave for concrete structured data problems such as graph applications? (see Chapter 3)

iv More generally can we derive other formulations than GW that are maybe more useful for data on incomparable spaces? (see Chapter 5)

Optimal Transport for structured data

Perhaps as you went along you did learn something.
– Ernest Hemingway, The Sun Also Rises

Contents

Summary of the contributions

This chapter is based on the papers [Vayer 2019a, Vayer 2020b] and considers the problem of computing distances between structured objects such as undirected graphs, seen as probability distributions in a specific metric space. We consider a new transportation distance (*i.e.* that minimizes a total cost of transporting probability masses) that unveils the geometric nature of the structured objects space. Unlike Wasserstein or Gromov-Wasserstein metrics that focus solely and respectively on features (by considering a metric in the feature space) or structure (by seeing structure as a metric space), our new distance exploits jointly both information, and is consequently called Fused Gromov-Wasserstein (FGW). After discussing its properties and computational aspects, we show results on a graph classification task, where our method outperforms both graph kernels and deep graph convolutional networks. Exploiting further on the metric properties of FGW, interesting geometric objects such as Fréchet means or barycenters of graphs are illustrated and discussed in a clustering context. We provide in a second part the mathematical framework for this distance in the continuous setting, prove its metric, geodesic and interpolation properties and provide a concentration result for the convergence of finite samples.

3.1   Introduction

There is a longstanding line of research on learning from structured data, *i.e.* objects that are a combination of a feature and structural information (see for example [Bakir 2007, Battaglia 2018]). As immediate instances, graph data are usually ensembles of nodes with attributes (typically $\mathbb{R}^d$ vectors) linked by some specific relation. Notable examples are found in chemical compounds or molecules modeling [Kriege 2016], brain connectivity [Ktena 2017], or social networks [Yanardag 2015]. This generic family of objects also encompasses time series [Cuturi 2017], trees [Day 1985] or even images [Bach 2007].

Being able to leverage on both feature and structural information in a learning task is a tedious task, that requires the association in some ways of those two pieces of information in order to capture the similarity between the structured data. Several kernels have been designed to perform this task [Shervashidze 2011, Vishwanathan 2010]. As a good representative of those methods, the Weisfeiler-Lehman kernel [Vishwanathan 2010] captures in each node a notion of vicinity by aggregating, in the sense of the topology of the graph, the surrounding features. Recent advances in graph convolutional networks [Bronstein 2017, Kipf 2016, Defferrard 2016, Wu 2020] allows learning end-to-end the best combination of features by relying on parametric convolutions on the graph, *i.e.* learnable linear combinations of features. In the end, and in order to compare two graphs that might have different number of nodes and connections, those two categories of methods build a new representation for every graph that shares the same space, and that is amenable to classification.

Contrasting with those previous methods, we suggest in this chapter to see graphs as probability distributions, embedded in a specific metric space. We propose to define a specific notion of distance between these probability distributions, that can be used in most of the classical machine learning approaches. Beyond its mathematical properties, disposing of a distance between structured data, provided it is meaningful, is desirable in many ways: *i)* it can then be plugged into distance-based machine learning algorithms such as k-nn or t-SNE *ii)* its quality is not dependent on the learning set size, and *iii)* it allows considering interesting quantities such as geodesic interpolation or barycenters.

Yet, defining this distance is not a trivial task. While features can always be compared using a standard metric, such as Euclidean distances, comparing structures requires a notion of similarity which can be found *via* the notion of *isometry*, since the graph nodes are not ordered (we define later on which cases two graphs are considered identical). We use the notion of transportation distance to compare two graphs represented as probability distributions. Optimal transport have inspired a number of recent breakthroughs

in machine learning (*e.g.* [Huang 2016, Courty 2017, Arjovsky 2017]) because of its capacity to compare empirical distributions, and also the recent advances in solving the underlying problem [Peyré 2019]. Yet, the natural formulation of OT cannot leverage the structural information of objects since it only relies on a cost function that compares their feature representations.

However, some modifications over OT formulation have been proposed in order to compare structural information of objects. Following the pioneering work by Mémoli [Memoli 2011], Peyré *et al.* [Peyré 2016] propose a way of comparing two distance matrices that can be seen as representations of some objects' structures. They use the Gromov-Wasserstein distance (see Chapter 2) capable of comparing two distributions even if they do not lie in the same ground space and apply it to compute barycenter of molecular shapes. Even though this approach has wide applications, it only encodes the intrinsic structural information in the transportation problem. To the best of our knowledge, the problem of including both structural and feature information in a unified OT formulation remains largely under-addressed.

OT distances that include both features and structures. Recent approaches tend to incorporate some structure information as a regularization of the OT problem. For example in [Alvarez-Melis 2018b] and [Courty 2017], authors constrain transport maps to favor some assignments in certain groups. These approaches require a known and simple structure such as class clusters to work but do not generalize well to more general structural information. In their work [Thorpe 2017], propose an OT distance that combines both a Lagrangian formulation of a signal and its temporal structural information. They define a metric, called Transportation L^p distance, that can be seen as a distance over the coupled space of time and feature. They apply it for signal analysis and show that combining both structure and feature tends to better capture the signal information. Yet, for their approach to work, the structure and feature information should lie in the same ambiant space, which is not a valid assumption for more general problems such as similarity between graphs. In [Nikolentzos 2017], authors propose a graph similarity measure for discrete labeled graph with OT. Using the eigenvector decomposition of the adjency matrix, which captures graph connectivities, nodes of a graph are first embedded in a new space, then a ground metric based on the distance in both this embedding and the labels is used to compute a Wasserstein distance serving as a graph similarity measure.

Contributions. After defining structured data as discrete probability measures (Section 3.2), we propose a new framework, namely FGW, capable of taking into account both structure and feature information into the optimal transport problem. The framework can compare any usual structured machine learning data even if the feature and structure information dwell in spaces of different dimensions, allowing the comparison of undirected labeled graphs. It is based on a distance that embeds a trade-off parameter which allows balancing the importance of the features and the structure. We provide a conditional-gradient algorithm for computing FGW (Section 3.3), and we evaluate it (Section 3.4) on both synthetic and real-world graph datasets on various tasks. We show that FGW is particularly useful for both supervised and unsupervised learning on graphs.

Among the contributions of this Chapter the numerical solution presented in Section 3.3 can also be used to compute the Gromov-Wasserstein distance. To the best of our knowledge this is the first optimization scheme for GW that does not require entropic regularization and which results in a sparse optimal solution.

We also define and illustrate a notion of labeled graph barycenters using FGW (Section 3.3.2), based on the Fréchet mean, and apply it for clustering and coarsening of graphs problems.

In a last part (Section 3.5) we generalize the definition of structured data to compact metric spaces. We
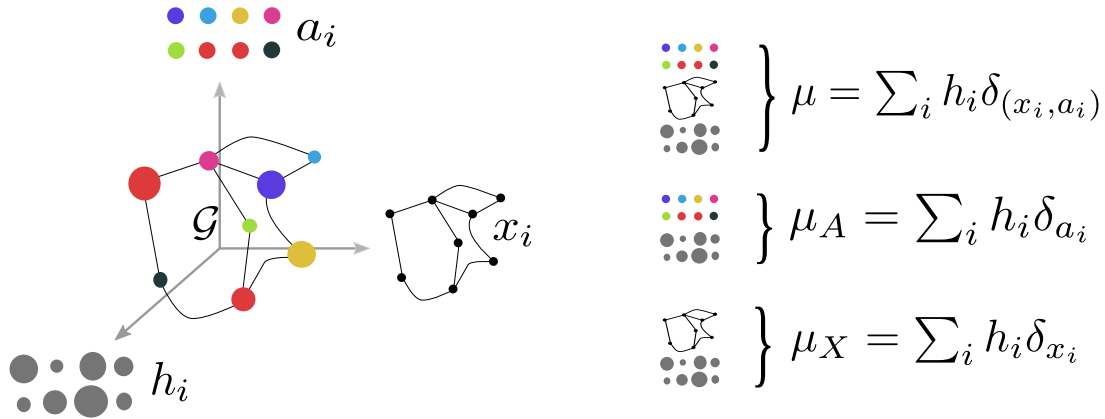
Figure 3.1: **(left)** Labeled graph with $(a_i)_i$ its feature information, $(x_i)_i$ its structure information and a probability vector $(h_i)_i$ that measures the relative importance of the vertices. **(right)** Associated structured data which is entirely described by a fully supported probability measure μ over the product space of feature and structure, with marginals μ_X and μ_A on the structure and the features respectively.

present the theoretical foundations of our framework in this general setting and states the mathematical properties of FGW. Notably, we show that it is a metric in the space of structured objects with respect to an intuitive equivalence relation between structured objects, we give a concentration result for the convergence of finite samples, and we study its interpolation and geodesic properties.

3.2   Structured data as probability measures

In this chapter, we focus on comparing structured data which combine a feature **and** a structure information. In order to give a good intuition about the method we first consider the discrete setting which corresponds to labeled graphs. More formally, we consider undirected labeled graphs as tuples of the form $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \ell_f, \ell_s)$ where $(\mathcal{V}, \mathcal{E})$ are the set of vertices and edges of the graph. $\ell_f : \mathcal{V} \to \Omega_f$ is a labelling function which associates each vertex $v_i \in \mathcal{V}$ with a feature $a_i \stackrel{\text{def}}{=} \ell_f(v_i)$ in some feature metric space (Ω_f, d). We will denote by *feature information* the set of all the features $(a_i)_i$ of the graph. Similarly, $\ell_s : \mathcal{V} \to \Omega_s$ maps a vertex v_i from the graph to its structure representation $x_i \stackrel{\text{def}}{=} \ell_s(v_i)$ in some structure space (Ω_s, C) specific to each graph. $C : \Omega_s \times \Omega_s \to \mathbb{R}$ is a symmetric application which aims at measuring the similarity between the nodes in the graph. Unlike the feature space however, Ω_s is implicit and in practice, knowing the similarity measure C will be sufficient. With a slight abuse of notation, C will be used in the following to denote both the structure similarity measure and the matrix that encodes this similarity between pairs of nodes in the graph $(\mathbf{C}(i, k) = C(x_i, x_k))_{i,k}$. Depending on the context, $\mathbf{C}$ can either encode the neighborhood information of the nodes, the edge information of the graph or more generally it can model a distance between the nodes such as the shortest path distance or the harmonic distance [Verma 2017]. When C is a metric, such as the shortest-path distance, we naturally endow the structure with the metric space (Ω_s, C). We will denote by *structure information* the set of all the structure embeddings $(x_i)_i$ of the graph.

We propose to enrich the previously described graph with a probability vector which serves the purpose of signaling the relative importance of the vertices in the graph. To do so, if we assume that the graph has n vertices, we equip those vertices with weights $(h_i)_i \in \Sigma_n$. Through this procedure, we derive the notion of *structured data* as a tuple $\mathcal{S} = (\mathcal{G}, h_{\mathcal{G}})$ where $\mathcal{G}$ is a graph as described previously and $h_{\mathcal{G}}$ is a
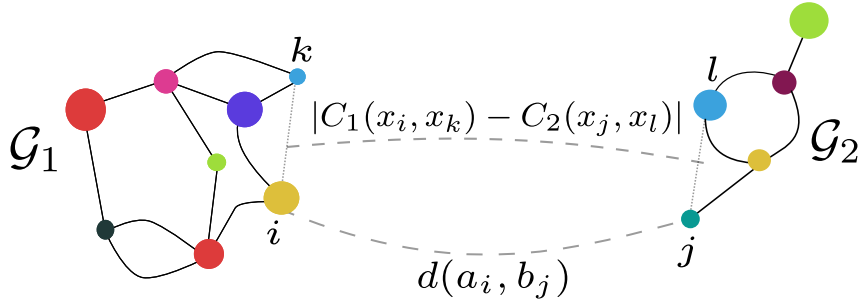
Figure 3.2: *FGW* loss E_q for a coupling π depends on both a similarity between each feature of each node of each graph $(d(a_i, b_j))_{i,j}$ and between all intra-graph structure similarities $(|C_1(x_i, x_k) - C_2(x_j, x_l)|)_{i,j,k,l}$.

function that associates a weight to each vertex. This definition allows the graph to be represented by a fully supported probability measure over the product space feature/structure $\mu = \sum_{i=1}^n h_i \delta_{(x_i, a_i)}$ which describes the entire structured data (see Figure 3.1). When all the weights are equal (*i.e.* $h_i = \frac{1}{n}$), so all vertices have the same relative importance, the structured data holds the exact same information as its graph. However, weights can be used to encode some *a priori* information. For instance on segmented images, one can construct a graph using the spatial neighborhood of the segmented zones, the features can be taken as the average color in the zone, and the weights as the ratio of image pixels in the zone.

3.3 Fused Gromov-Wasserstein approach for structured data

We aim at defining a distance between two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, described respectively by their probability measure $\mu = \sum_{i=1}^n h_i \delta_{(x_i, a_i)}$ and $\nu = \sum_{j=1}^m g_j \delta_{(y_j, b_j)}$, where $\mathbf{h} \in \Sigma_n$ and $\mathbf{g} \in \Sigma_m$ are probability vectors. Without loss of generality we suppose $(x_i, a_i) \neq (x_j, a_j)$ for $i \neq j$ (same for y_j and b_j). We recall that $\Pi(\mathbf{h}, \mathbf{g})$ the set of all admissible couplings between $\mathbf{h}$ and $\mathbf{g}$. To that extent, the matrix $\boldsymbol{\pi} \in \Pi(\mathbf{h}, \mathbf{g})$ describes a probabilistic matching of the nodes of the two graphs. We note $\mathbf{M_{AB}} = (d(a_i, b_j))_{i,j}$ the $n \times m$ matrix standing for the distance between the features. The structure matrices are denoted $\mathbf{C}_1$ and $\mathbf{C}_2$, and μ_X and μ_A (resp. ν_Y and ν_B) are representative of the marginals of μ (resp. ν) *w.r.t.* the structure and feature respectively (see Figure 3.1). We also define the similarity between the structures by measuring the similarity between all pairwise distances within each graph thanks to the 4-dimensional tensor $\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)$:

$$\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2) = (L_{i,j,k,l}(\mathbf{C}_1, \mathbf{C}_2))_{i,j,k,l} = (|C_1(i,k) - C_2(j,l)|)_{i,j,k,l}.$$

3.3.1 Fused Gromov-Wasserstein distance

We define a novel Optimal Transport discrepancy called the Fused Gromov-Wasserstein distance. It is defined for a trade-off parameter $\alpha \in [0,1]$ as

$$FGW(\mathbf{M_{AB}}, \mathbf{C}_1, \mathbf{C}_2, \mathbf{h}, \mathbf{g}) = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{h}, \mathbf{g})} E_q(\mathbf{M_{AB}}, \mathbf{C}_1, \mathbf{C}_2, \boldsymbol{\pi}) \qquad (3.1)$$

where

$$
\begin{aligned}
E_q(\mathbf{M_{AB}}, \mathbf{C}_1, \mathbf{C}_2, \boldsymbol{\pi}) &= \langle (1-\alpha)\mathbf{M_{AB}}^q + \alpha \mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^q \otimes \boldsymbol{\pi}, \boldsymbol{\pi} \rangle_{\mathcal{F}} \\
&= \sum_{i,j,k,l} (1-\alpha)d(a_i, b_j)^q + \alpha |C_1(i,k) - C_2(j,l)|^q \pi_{i,j} \pi_{k,l}
\end{aligned} \qquad (3.2)
$$

The FGW distance looks for the coupling $\boldsymbol{\pi}$ between the vertices of the graph that minimizes the cost E_q which is a linear combinaison of a cost $d(a_i, b_j)$ of transporting one feature a_i to a feature b_j and a cost $|C_1(i, k) - C_2(j, l)|$ of transporting pairs of nodes in each structure (see Figure 3.2). As such, the optimal coupling tends to associate pairs of feature and structure points with similar distances within each structure pair and with similar features. α acts as a trade-off parameter between the cost of the structures represented by $\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)$ and the cost on the features $\mathbf{M}_{\mathbf{AB}}$. In this way, the convex combination of both terms leads to the use of both information in one formalism resulting on a single map $\boldsymbol{\pi}$ which "moves" the mass from one joint probability measure forawrd to the other. As an important feature of FGW, by relying on a sum of (inter- and intra-)vertex-to-vertex distances, it can handle structured data with continuous attributed or discrete labeled nodes (thanks to the definition of d) and can also be computed even if the graphs have different number of nodes.

This new distance is called the FGW distance as it acts as a generalization of the Wasserstein and Gromov-Wasserstein distances. Indeed as α tends to zero, the FGW distance recovers the Wasserstein distance between the features $W_q(\mu_A, \nu_B)^q$ and as α tends to one, we recover the Gromov-Wasserstein distance $GW_q(\mu_X, \nu_Y)^q$ between the structures (see Proposition 3.5.2 of Section 3.5.5).

More importantly FGW enjoys metric properties on labeled graphs as stated in the following theorem:

Theorem 3.3.1 (FGW *defines a metric for* $q = 1$ *and a semi-metric for* $q > 1$). *If* $q = 1$, *and if* $\mathbf{C}_1$, $\mathbf{C}_2$ *are distance matrices such as shortest-path matrices then* FGW *defines a metric over the space of structured data quotiented by the measure preserving isometries that are also feature preserving. More precisely,* FGW *satisfies the triangle inequality and vanishes iff* $n = m$ *and there exists a permutation* $\sigma \in S_n$ *such that:*

$$\forall i \in [\![n]\!], \ h_i = g_{\sigma(i)} \tag{3.3}$$

$$\forall i \in [\![n]\!], \ a_i = b_{\sigma(i)} \tag{3.4}$$

$$\forall i, k \in [\![n]\!]^2, \ C_1(i, k) = C_2(\sigma(i), \sigma(k)) \tag{3.5}$$

If $q > 1$, *the triangle inequality is relaxed by a factor* 2^{q-1} *such that* FGW *defines a semi-metric.*

This results is a direct consequence of Theorem 3.5.1 in Section 3.5.2 where FGW is defined for general metric spaces. The resulting permutation σ preserves the weight of each node (equation (3.3)), the features (equation (3.4)) and the pairwise structure relation between the nodes (equation (3.5)). For example, comparing two graphs with uniform weights on the vertices and with shortest-path structure matrices, the FGW distance vanishes *iff* the graphs have the same number of vertices and there exists a one-to-one mapping between the vertices of the graphs which preserves both the shortest-paths and the features. More informally, in this case graphs have vertices with the same labels connected by the same edges, and thus FGW can be used to determine if two graphs are isomorphic [West 2000].

The metric FGW is fully unsupervised and can be used in a wide set of applications such as k-nearest-neighbors, distance-substitution kernels, pseudo-Euclidean embeddings, or representative-set methods. Arguably, such a distance also allows for a fine interpretation of the similarity (through the optimal mapping $\boldsymbol{\pi}$), contrary to end-to-end learning machines such as neural networks.

3.3.2 Fused Gromov-Wasserstein barycenter

OT barycenters have many desirable properties and applications (see Chapter 2 for more details), yet no formulation can leverage both structural and feature information in the barycenter computation. In this

section, we consider the FGW distance to define a barycenter of a set of structured data as a Fréchet mean. We look for the structured data μ that minimizes the sum of (weighted) FGW distances within a given set of structured data $(\mu_k)_{k \in [\![K]\!]}$ associated with structure matrices $(\mathbf{C}_k)_{k \in [\![K]\!]}$, features $(\mathbf{B}_k)_{k \in [\![K]\!]}$ and base histograms $(\mathbf{h}_k)_{k \in [\![K]\!]}$. For simplicity, we assume that the histogram $\mathbf{h}$ associated to the barycenter is known and fixed; in other words, we set the number of vertices N and the weight associated to each of them.

In this context, for a fixed $N \in \mathbb{N}$ and $(\lambda_k)_k$ such that $\sum_{k=1}^{K} \lambda_k = 1$, we aim to find the set of features $\mathbf{A} = (a_i)_i$ and the structure matrix $\mathbf{C}$ of the barycenter that minimize the following equation:

$$\min_{\mu} \sum_{k=1}^{K} \lambda_k FGW_{q,\alpha}(\mu, \mu_k) = \min_{\substack{\mathbf{C} \in \mathbb{R}^{N \times N}, \ \mathbf{A} \in \mathbb{R}^{N \times d} \\ \forall k \in [\![K]\!], \boldsymbol{\pi}_k \in \Pi(\mathbf{h}, \mathbf{h}_k)}} \sum_{k=1}^{K} \lambda_k E_q(\mathbf{M_{AB}}_k, \mathbf{C}, \mathbf{C}_k, \boldsymbol{\pi}_k) \tag{3.6}$$

Note that this problem is jointly convex *w.r.t.* $\mathbf{C}$ and $\mathbf{A}$ but not *w.r.t.* $\boldsymbol{\pi}_k$. We discuss the proposed algorithm to solve this problem in the next section. Interestingly enough, one can derive several variants of this problem, where the features or the structure matrices of the barycenter can be fixed. Solving the related simpler optimization problem extends straightforwardly. We give examples of such barycenters both in the experimental section where we solve a graph based k-means problem.

3.3.3 Optimization and algorithmic solution

In this section we discuss the numerical optimization problem for computing the FGW distance between discrete distributions.

Solving the Quadratic Optimization problem. Equation (3.1) is clearly a quadratic problem *w.r.t.* $\boldsymbol{\pi}$ which is NP-hard in general [Loiola 2007]. However finding a solution in practice can be done quite efficiently. We propose here a method based on the Frank-Wolfe algorithm [Jaggi 2013] (*aka* Conditional Gradient). When considering $q = 2$ the FGW computation problem can be re-written as finding $\boldsymbol{\pi}^*$ such that:

$$\boldsymbol{\pi}^* = \arg\min_{\boldsymbol{\pi} \in \Pi(\mathbf{h}, \mathbf{g})} \quad \text{vec}(\boldsymbol{\pi})^T \mathbf{Q}(\alpha) \text{vec}(\boldsymbol{\pi}) + \text{vec}(\mathbf{D}(\alpha))^T \text{vec}(\pi) \tag{3.7}$$

where $\mathbf{Q} = -2\alpha \mathbf{C}_2 \otimes_K \mathbf{C}_1$ and $\mathbf{D}(\alpha) = (1-\alpha)\mathbf{M_{AB}}$. $\otimes_K$ denotes the Kronecker product of two matrices, vec the column-stacking operator. With such form, the resulting optimal map can be seen as a quadratic regularized map from initial Wasserstein [Ferradans 2014, Flamary 2014]. However, unlike these approaches, we have a quadratic but provably non convex term. The gradient $\mathbf{G}$ that arises from equation (3.1) can be expressed with the following partial derivative *w.r.t.* $\boldsymbol{\pi}$:

$$\mathbf{G}(\boldsymbol{\pi}) = (1-\alpha)\mathbf{M_{AB}}^q + 2\alpha \mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^q \otimes \boldsymbol{\pi} \tag{3.8}$$

Note that despite the apparent $O(m^2 n^2)$ complexity of computing the tensor product $\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^q \otimes \boldsymbol{\pi}$ given $\boldsymbol{\pi}$, one can simplify the sum to complexity $O(mn^2 + m^2 n)$ [Peyré 2016] operations when $q = 2$.

Solving a large scale QP with a classical solver can be computationally expensive. In [Ferradans 2014], authors propose a solver for a graph regularized optimal transport problem whose resulting optimization problem is also a QP. We can then directly use their conditional gradient scheme to solve our optimization problem as presented in Algorithm 4. It only needs at each iteration to compute the gradient in equation (3.8) and to solve a linear OT problem with classical solvers (see Chapter 2 for more details). The line-search part is a constrained minimization of a second degree polynomial function and, as such, admits a closed form expression written in Algorithm 5. While the problem is non convex, conditional gradient is

---

Algorithm 4 Conditional Gradient (CG) for *FGW*

1: $\pi^{(0)} \leftarrow \mathbf{hg}^\top$
2: **for** $i = 1, \ldots,$ **do**
3:     $\mathbf{G} \leftarrow$ Gradient from equation (3.8) *w.r.t.* $\boldsymbol{\pi}^{(i-1)}$
4:     $\tilde{\boldsymbol{\pi}}^{(i)} \leftarrow$ Solve OT with ground loss $\mathbf{G}$
5:     $\tau^{(i)} \leftarrow$ Line-search for loss (3.1) with $\tau \in (0, 1)$ using Alg. 5
6:     $\boldsymbol{\pi}^{(i)} \leftarrow (1 - \tau^{(i)})\boldsymbol{\pi}^{(i-1)} + \tau^{(i)}\tilde{\boldsymbol{\pi}}^{(i)}$
7: **end for**

Algorithm 5 Line-search for CG $(q = 2)$

1: $\mathbf{c}_{\mathbf{C}_1, \mathbf{C}_2}$ from equation (6) in [Peyré 2016]
2: $a = -2\alpha \langle \mathbf{C}_1 \tilde{\boldsymbol{\pi}}^{(i)} \mathbf{C}_2, \tilde{\boldsymbol{\pi}}^{(i)} \rangle$
3: $b = \langle (1-\alpha)\mathbf{M_{AB}} + \alpha \mathbf{c}_{\mathbf{C}_1, \mathbf{C}_2}, \tilde{\boldsymbol{\pi}}^{(i)} \rangle_\mathcal{F} - 2\alpha \big(\langle \mathbf{C}_1 \tilde{\boldsymbol{\pi}}^{(i)} \mathbf{C}_2, \boldsymbol{\pi}^{(i-1)} \rangle_\mathcal{F} + \langle \mathbf{C}_1 \boldsymbol{\pi}^{(i-1)} \mathbf{C}_2, \tilde{\boldsymbol{\pi}}^{(i)} \rangle_\mathcal{F} \big)$
4: **if** $a > 0$ **then**
5:     $\tau^{(i)} \leftarrow \min(1, \max(0, \frac{-b}{2a}))$
6: **else**
7:     $\tau^{(i)} \leftarrow 1$ if $a + b < 0$ else $\tau^{(i)} \leftarrow 0$
end if

---

known to converge to a local stationary point with a $O(\frac{1}{\sqrt{t}})$ rate [Lacoste-Julien 2016]. More precisely we note $g^{(i)}$ the Frank-Wolfe gap at iteration i defined by:

$$g^{(i)} = \max_{\boldsymbol{\pi} \in \Pi(\mathbf{h}, \mathbf{g})} \langle \boldsymbol{\pi} - \boldsymbol{\pi}^{(i)}, -\mathbf{G}^i \rangle_\mathcal{F} \tag{3.9}$$

We note also $|||.|||$ the dual norm for tensors: $|||\mathbf{L}||| = \sup_{\|\mathbf{A}\|_\mathcal{F}=1} \|\mathbf{L} \otimes \mathbf{A}\|_\mathcal{F}$ where $\mathbf{L}$ is a 4-dimensional tensor and $\mathbf{A}$ a matrix. Then using this dual norm we have that the gradient $\mathbf{G}(.)$ is $|||2\alpha\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^q|||$-Lipschitz:

$$\forall (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) \in \Pi(\mathbf{h}, \mathbf{g})^2, \ \|\mathbf{G}(\boldsymbol{\pi}_1) - \mathbf{G}(\boldsymbol{\pi}_2)\|_\mathcal{F} = \|2\alpha\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^q \otimes (\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2)\|_\mathcal{F} \\ \leq 4\alpha^2 |||\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^q||| \ \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\|_\mathcal{F} \tag{3.10}$$

We note also $\text{diam}_{\|.\|_\mathcal{F}}(\Pi(\mathbf{h}, \mathbf{g}))^2$ the diameter of $\Pi(\mathbf{h}, \mathbf{g})$ (see [Jaggi 2013]):

$$\text{diam}_{\|.\|_\mathcal{F}}(\Pi(\mathbf{h}, \mathbf{g}))^2 = \max_{(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) \in \Pi(\mathbf{h}, \mathbf{g})^2} \|\boldsymbol{\pi}_1 - \boldsymbol{\pi}_2\|_\mathcal{F}^2 \tag{3.11}$$

Then by using Theorem 1 in [Lacoste-Julien 2016] then minimal gap encountered by the iterates during the algorithm after t iterations satisfies:

$$\min_{0 \leq i \leq t} g^{(i)} \leq \frac{\max\{2h_0, C\}}{\sqrt{t+1}} \tag{3.12}$$

where $C = 4\alpha^2 |||\mathbf{L}(\mathbf{C}_1, \mathbf{C}_2)^q||| \ \text{diam}_{\|.\|_\mathcal{F}}(\Pi(\mathbf{h}, \mathbf{g}))^2$ and h_0 is is the initial global suboptimality.

Solving the barycenter problem with Block Coordinate Descent (BCD). We propose to minimize equation (3.6) using a BCD algorithm, *i.e.* iteratively minimizing with respect to the couplings $\boldsymbol{\pi}_k$, to the metric $\mathbf{C}$ and the feature vector $\mathbf{A}$. The minimization of this problem *w.r.t.* $(\boldsymbol{\pi}_k)_{k \in \llbracket K \rrbracket}$ is equivalent to compute S independent Fused Gromov-Wasserstein distances as discussed above. We suppose
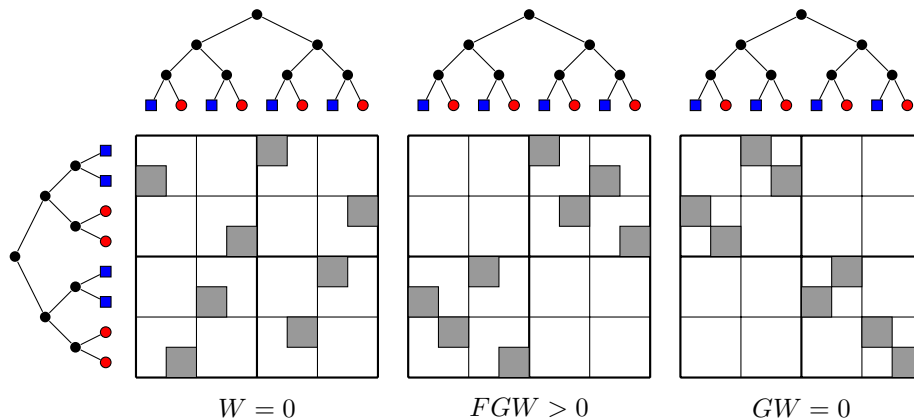
Figure 3.3: Example of FGW, GW and W on synthetic trees. Dark grey color represents a non null $\pi_{i,j}$ value between two nodes i and j. **(left)** the W distance between the features with $\alpha = 0$, **(middle)** FGW **(right)** the GW between the structures $\alpha = 1$.

that the feature space is $\Omega_f = (\mathbb{R}^d, \|.\|_2^2)$ and we consider $q = 2$. Minimization *w.r.t.* $\mathbf{C}$ in this case has a closed form (see Proposition 4 in [Peyré 2016] and Chapter 2) :

$$\mathbf{C} \leftarrow \frac{1}{\mathbf{h}\mathbf{h}^T} \sum_{k=1}^{K} \lambda_k \boldsymbol{\pi}_k^T \mathbf{C}_k \boldsymbol{\pi}_k \qquad (3.13)$$

where the division is computed elementwise and $\mathbf{h}$ is the histogram of the barycenter as discussed in section 3.3.2. Minimization *w.r.t.* $\mathbf{A}$ can be computed with [Cuturi 2014, Equation 8]:

$$\mathbf{A} \leftarrow \sum_{k=1}^{K} \lambda_k \mathbf{B}_k \boldsymbol{\pi}_k^T \mathrm{diag}(\frac{1}{\mathbf{h}}) \qquad (3.14)$$

3.4   Experimental results

We illustrate in this section the behavior of our method on synthetic and real datasets. The algorithm presented in the previous section have been implemented in the Python Optimal Transport toolbox [Flamary 2017].

3.4.1   Illustration of FGW on trees

We construct two trees as illustrated in Figure 3.3, where the 1D node features are shown with colors (in red, features belong to $[0, 1]$ and in blue in $[9, 10]$). The structure similarity matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ are the shortest-paths between the nodes. Both trees have the same individual structure and the same features up to a permutation. However when combining both informations the trees are not the same, as they do not have the same labels at the same place. Figure 3.3 illustrates the behavior of the FGW distance when the trade-off parameter α changes. The left part recovers the Wasserstein distance between the features ($\alpha = 0$): red nodes are coupled to red ones and the blue nodes to the blue ones. For an alpha close to 1 (right), we recover the Gromov-Wasserstein distance between the structures of the trees: all couples of points are coupled to another couple of points, without taking into account the features. Both approaches fail in discriminating the two trees. Finally, for an intermediate α in FGW (center), the bottom and first level structure are preserved as well as the feature matching (red on red and blue on blue), resulting on a positive distance. Note that FGW preserves also the substructures of the trees through its coupling.
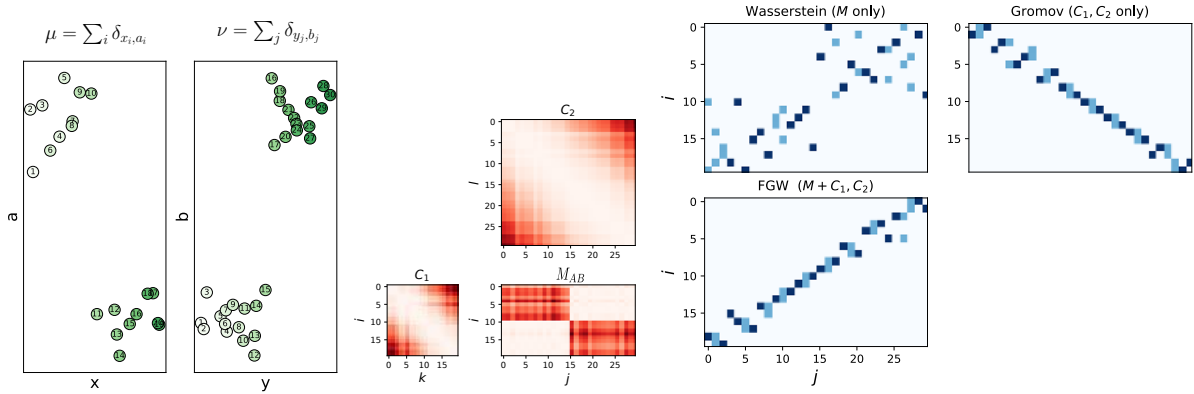
Figure 3.4: Illustration of the difference between W, GW and FGW couplings. **(left)** empirical distributions μ with 20 samples and ν with 30 samples which color is proportional to their index. **(middle)** Cost matrices in the feature ($\mathbf{M_{AB}}$) and structure domains ($\mathbf{C_1}, \mathbf{C_2}$) with similar samples in white. **(right)** Solution for all methods. Dark blue indicates a non zero coefficient of the transportation map. Feature distances are large between points laying on the diagonal of $\mathbf{M_{AB}}$ such that Wasserstein maps is anti-diagonal but unstructured. Fused Gromov-Wasserstein incorporates both feature and structure maps in a single transport map.

3.4.2   Illustration of FGW on 1D distributions

Figure 3.4 illustrates the differences between Wasserstein, Gromov-Wasserstein and Fused Gromov-Wasserstein couplings $\boldsymbol{\pi}^*$ on 1D distributions. In this example both the feature and structure are 1-dimensional, that is $(\mathbf{x}_i, \mathbf{y}_j) \in \mathbb{R}^2$ and $(\mathbf{a}_i, \mathbf{b}_j) \in \mathbb{R}^2$ (Figure 3.4 left). The feature space (vertical axis) denotes two clusters among the elements of both objects illustrated in the OT matrix $\mathbf{M_{AB}}$, the structure space (horizontal axis) denotes a noisy temporal sequence along the indexes illustrated in the matrices $\mathbf{C_1}$ and $\mathbf{C_2}$ (Figure 3.4 center). Wasserstein respects the clustering but forgets the temporal structure, Gromov-Wasserstein respects the structure but do not take the clustering into account. Only FGW retrieves a transport matrix respecting both feature and structure.

3.4.3   Illustration of FGW on simple images

We extract a 28×28 image from the MNIST dataset and generate a second one through translation or mirroring of the digit in the original image. We use pixel gray levels as the features, and the structure is defined as the city-block distance on the pixel coordinate grid. We use equal weights for all the pixels in the image. Figure 3.5 shows the different couplings obtained when considering either the features only, the structure only or both information. FGW aligns the pixels of the digits, recovering the correct order of the pixels, while both Wassertein and Gromov-Wasserstein distances fail at providing a meaningful transportation map. Note that in the Wasserstein and Gromov-Wasserstein case, the distances are equal to 0, whereas FGW manages to spot that the two images are different. Also note that, in the FGW sense, the original digit and its mirrored version are also equivalent as there exists an isometry between their structure spaces, making FGW invariant to rotations or flips in the structure space in this case.

3.4.4   Illustration of FGW on time series data

One of the main assets of FGW is that it can be used on a wide class of structured data such as graphs and also time series. We consider here 25 monodimensional time series composed of two humps in $[0, 1]$
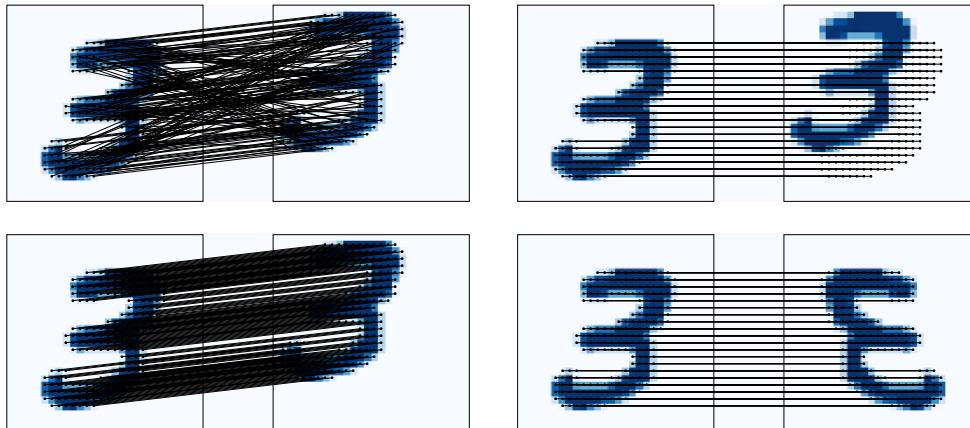
Figure 3.5: Couplings obtained when considering **(top left)** the features only, where we have $W = 0$ **(top right)** the structure only, with $GW = 0$ **(bottom left and right)** both the features and the structure, with $FGW_{0.1,2}$. For readibility issues, only the couplings starting from non white pixels on the left picture are depicted.

with random uniform height between 0 and 1. Signals are distributed according to two classes translated from each other with a fixed gap. The FGW distance is computed by considering $\mathbf{M_{AB}}$ as the Euclidean distance between the features of the signals (here the value of the signal in each point) and $\mathbf{C}_1$ and $\mathbf{C}_2$ as the euclidean distance between timestamps.

A 2D embedding is computed from a FGW distance matrix between a number of examples in this dataset with multidimensional scaling (MDS) in Figure 3.6 (top). One can clearly see that the representation with a reasonable α value in the center is the most discriminant one. This can be better understood by looking as the OT matrices between the classes. Figure 3.6 (bottom) illustrates the behavior of FGW on one pair of examples when going from Wasserstein to Gromov-Wasserstein. The black line depicts the matching provided by the transport matrix and one can clearly see that while Wasserstein on the left assigns samples completely independently to their temporal position, the Gromov-Wasserstein on the right tends to align perfectly the samples (note that it could have reversed exactly the alignment with the same loss) but discards the values in the signal. Only the true FGW in the center finds a transport matrix that both respects the time sequences and aligns similar values in the signals.

3.4.5 Graph-structured data classification

We now use FGW on real-world dataset where we study its behavior on a graph classification task. More precisely we address the question of training a classifier for graph data and evaluate the FGW distance used in a kernel with SVM.

Datasets We consider 12 widely used benchmark datasets divided into 3 groups. BZR, COX2 [Sutherland 2003], PROTEINS, ENZYMES [Borgwardt 2005], CUNEIFORM [Kriege 2018] and SYNTHETIC [Feragen 2013] are vector attributed graphs. MUTAG [Debnath 1991], PTC-MR [Kriege 2016] and NCI1 [Wale 2008] contain graphs with discrete attributes derived from small molecules. IMDB-B, IMDB-M [Yanardag 2015] contain unlabeled graphs derived from social networks. All datasets are available in [Kersting 2016].

2D MDS embeddings from distances
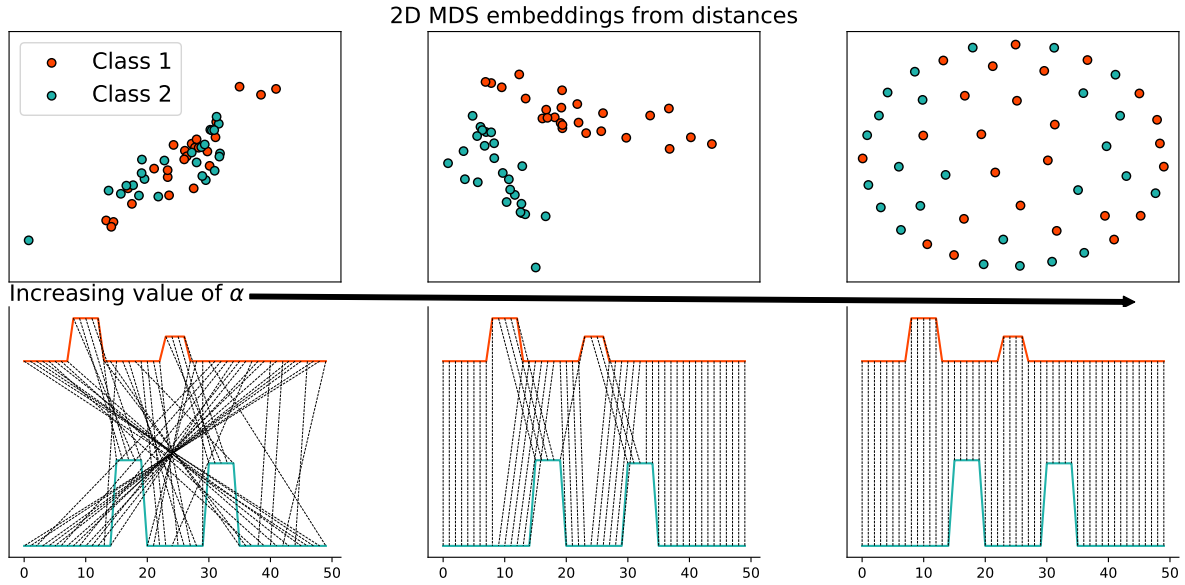
Figure 3.6: Behavior of trade-off parameter α on a toy time series classification problem. α is increasing from left ($\alpha = 0$: Wasserstein distance) to right ($\alpha = 1$: Gromov-Wasserstein distance). **(top row)** 2D-embedding is computed from the set of pairwise distances between samples with MDS **(bottom row)** illustration of couplings between two sample time series from opposite classes.

Experimental setup   Regarding the feature distance matrix $\mathbf{M_{AB}}$ between node features, when dealing with real valued vector attributed graphs, we consider the ℓ_2 distance between the labels of the vertices. In the case of graphs with discrete attributes, we consider two settings: in the first one, we keep the original labels (denoted as RAW); we also consider a Weisfeiler-Lehman labeling (denoted as WL) by concatenating the labels of the neighbors. A vector of size H is created by repeating this procedure H times [Vishwanathan 2010, Kriege 2016]. In both cases, we compute the feature distance matrix by using $d(a_i, b_j) = \sum_{k=0}^{H} \delta(\tau(a_i^k), \tau(b_j^k))$ where $\delta(x,y) = 1$ if $x \neq y$ else $\delta(x,y) = 0$ and $\tau(a_i^k)$ denotes the concatenated label at iteration k (for $k = 0$ original labels are used). Regarding the structure distances C, they are computed by considering a shortest path distance between the vertices.

For the classification task, we run a SVM using the indefinite kernel matrix $e^{-\gamma FGW}$ which is seen as a noisy observation of the true positive semidefinite kernel [Luss 2007]. We compare classification accuracies with the following state-of-the-art graph kernel methods: (SPK) denotes the shortest path kernel [Borgwardt 2005], (RWK) the random walk kernel [Gärtner 2003], (WLK) the Weisfeler Lehman kernel [Vishwanathan 2010], (GK) the graphlet count kernel [Shervashidze 2009]. For real valued vector attributes, we consider the HOPPER kernel (HOPPERK) [Feragen 2013] and the propagation kernel (PROPAK) [Neumann 2016]. We build upon the GraKel library [Siglidis 2018] to construct the kernels and C-SVM to perform the classification. We also compare *FGW* with the PATCHY-SAN framework for CNN on graphs (PSCN) [Niepert 2016] building on our own implementation of the method[1].

To compare the methods, most papers about graph classification usually perform a nested cross validation (using 9 folds for training, 1 for testing, and reporting the average accuracy of this experiment repeated 10 times) and report accuracies of the other methods taken from the original papers. However, these comparisons are not fair because of the high variance on most datasets *w.r.t.* the folds chosen for

---

[1] https://github.com/tvayer/PSCN

Table 3.1: Average classification accuracy on the graph datasets with vector attributes.

VECTOR ATTRIBUTES	BZR	COX2	CUNEIFORM	ENZYMES	PROTEIN	SYNTHETIC
FGW	**85.12±4.15***	77.23±4.86	**76.67±7.04**	71.00±6.76	**74.55±2.74**	**100.00±0.00**
HOPPERK	84.15±5.26	**79.57±3.46**	32.59±8.73	45.33±4.00	71.96±3.22	90.67±4.67
PROPAK	79.51±5.02	77.66±3.95	12.59±6.67	**71.67±5.63***	61.34±4.38	64.67±6.70
PSCN κ=10	80.00±4.47	71.70±3.57	25.19±7.73	26.67±4.77	67.95±11.28	**100.00±0.00**
PSCN κ=5	82.20±4.23	71.91±3.40	24.81±7.23	27.33±4.16	71.79±3.39	**100.00±0.00**

training and testing. This is why, in our experiments, the nested cross validation is performed on the same folds for training and testing for *all* methods. In the result tables 3.1, 3.2 and 3.3 we add a (*) when the best score does not yield to a significative improvement (based on a Wilcoxon signed rank test on the test scores) compared to the second best one. Note that, because of their small sizes, we repeat the experiments 50 times for MUTAG and PTC-MR datasets. For all methods using SVM, we cross validate the parameter $C \in \{10^{-7}, 10^{-6}, ..., 10^7\}$. The range of the WL parameter H is $\{0, 1..., 10\}$, and we also compute this kernel with H fixed at 2, 4. The decay factor λ for RWK $\{10^{-6}, 10^{-5}..., 10^{-2}\}$, for the GK kernel we set the graphlet size $\kappa = 3$ and cross validate the precision level ε and the confidence δ as in the original paper [Shervashidze 2009]. The $t_{\max}$ parameter for PROPAK is chosen within $\{1, 3, 5, 8, 10, 15, 20\}$. For PSCN, we choose the normalized betweenness centrality as labeling procedure and cross validate the batch size in $\{10, 15, ..., 35\}$ and number of epochs in $\{10, 20, ..., 100\}$. Finally for FGW, γ is cross validated within $\{2^{-10}, 2^{-9}, ..., 2^{10}\}$ and α is cross validated *via* a logspace search in $[0, 0.5]$ and symmetrically $[0.5, 1]$ (15 values are drawn).

Results and discussion

Vector attributed graphs. The average accuracies reported in Table 3.1 show that FGW is a clear state-of-the-art method and performs best on 4 out of 6 datasets with performances in the error bars of the best methods on the other two datasets. Results for CUNEIFORM are significantly below those from the original paper [Kriege 2018] which can be explained by the fact that the method in this paper uses a graph convolutional approach specially designed for this dataset and that the experimental setting is different. In comparison, the other competitive methods are less consistent as they exhibit some good performances on some datasets only.

Discrete labeled graphs. We first note in Table 3.2 that FGW using WL attributes outperforms all competitive methods, including FGW with raw features. Indeed, the WL attributes allow encoding more finely the neighborood of the vertices by stacking their attributes, whereas FGW with raw features only consider the shortest path distance between vertices, not their sequence of labels. This result calls for using meaningful feature and/or structure matrices in the FGW definition, that can be dataset-dependant, in order to enhance the performances. We also note that FGW with WL attributes outperforms the WL kernel method, highlighting the benefit of an optimal transport-based distance over a kernel-based similarity. Surprisingly results of PSCN are significantly lower than those from the original paper. We believe that it comes from the difference between the folds assignment for training and testing, which suggests that PSCN is difficult to tune.

Table 3.2: Average classification accuracy on the graph datasets with discrete attributes.

DISCRETE ATTR.	MUTAG	NCI1	PTC-MR
FGW RAW	83.26±10.30	72.82±1.46	55.71±6.74
FGW WL H=2	86.42±7.81	85.82±1.16	63.20±7.68
FGW WL H=4	**88.42±5.67**	**86.42±1.63**	**65.31±7.90**
GK K=3	82.42±8.40	60.78±2.48	56.46±8.03
RWK	79.47±8.17	58.63±2.44	55.09±7.34
SPK	82.95±8.19	74.26±1.53	60.05±7.39
WLK	86.21±8.48	85.77±1.07	62.86±7.23
WLK H=2	86.21±8.15	81.85±2.28	61.60±8.14
WLK H=4	83.68±9.13	85.13±1.61	62.17±7.80
PSCN K=10	83.47±10.26	70.65±2.58	58.34±7.71
PSCN K=5	83.05±10.80	69.85±1.79	55.37±8.28

Table 3.3: Average classification accuracy on the graph datasets with no attributes.

WITHOUT ATTRIBUTE	IMDB-B	IMDB-M
GW	**63.80±3.49**	**48.00±3.22**
GK K=3	56.00±3.61	41.13±4.68
SPK	55.80±2.93	38.93±5.12

Non-attributed graphs. The particular case of the GW distance for graph classification is also illustrated on social datasets, that contain no labels on the vertices. Accuracies reported in Table 3.3 show that it greatly outperforms SPK and GK graph kernel methods.

Comparison between FGW, W and GW During the validation step, the optimal value of α was consistently selected inside the $]0, 1[$ interval, excluding 0 and 1, suggesting that both structure and feature pieces of information are necessary (details are given in Section 6.1.1).

3.4.6   Graph barycenter and compression

In this section we use the barycentric formulation of FGW using the Fréchet mean formulation in equation (3.6) in two settings. The first one is the computation of barycenter of several toy labeled graphs and the second the compression of one graph into a smaller graph, also known as coarsening [Loukas 2019].

Graph barycenter In this part, we use FGW to compute barycenter of toy graphs. In a first example, we generate graphs following either a circle or 8 symbol with 1D features following a sine and linear variation respectively. For each example, the number of nodes is drawn randomly between 10 and 25, Gaussian noise is added to the features and a small noise is applied to the structure (some connections are randomly added). An example graph with no noise is provided for each class in the first column of Figure
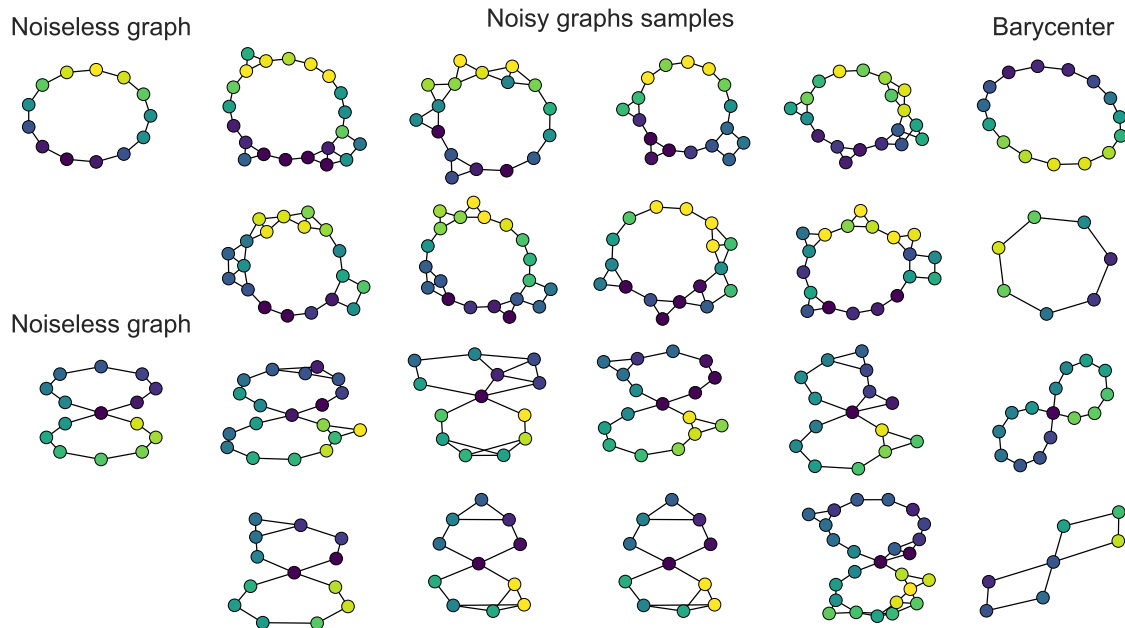
Figure 3.7: Illustration of FGW graph barycenter. The first column illustrates the original settings with the noiseless graphs, and columns 2 to 5 are the noisy samples that constitute the datasets. Column 6 show the barycenters for each setting, with different number of nodes. Blue nodes indicates a feature value close to -1, yellow nodes close to 1.

3.7. One can see from there that the circle class has a feature varying smoothly (sine) along the graph but the 8 has a sharp feature change at its center (so that low pass filtering would loose some information). Some examples of the generated graphs are provided in the 2nd-to-7th columns of Figure 3.7. We compute the FGW barycenter containing 10 samples using the shortest path distance between the nodes as the structural information and the distance induced by the Euclidean norm $\|.\|_2$ for the features.

Note that the iterations of the barycenter defined in equation (3.13) result in a dense $\mathbf{C}$ matrix and to visualize properly the graph barycenter we need a adjacency matrix. We propose a simple heuristic procedure to recover an adjacency matrix for the graphs' barycenter based on a thresholding of the matrix $\mathbf{C}$. Given a threshold t the matrix $thresh_t(\mathbf{C})$ is given by 1 if $C_{ij} <= t$ and 0 elsewhere. The threshold t is tuned so as to minimize the Frobenius norm between the original $\mathbf{C}$ matrix and the shortest path matrix constructed after thresholding $\mathbf{C}$. More precisely if SP denotes the algorithm which takes as input an adjacency matrix and outputs a shortest-path matrix then the threshold is given:

$$\underset{t \in \mathbb{R}_+}{\arg\min} \|\mathbf{C} - SP(thresh_t(\mathbf{C}))\|_F^2 \tag{3.15}$$

The idea behind equation (3.15) is that $\mathbf{C}$ represents somehow the shortest-path matrix of a graph so that we want the adjacency matrix $thresh_t(\mathbf{C})$ to give a shortest-path matrix as close as possible to $\mathbf{C}$. Unfortunately equation (3.15) is not differentiable with respect to t. We use a simple brute force strategy to find a suitable threshold t by looking at $\arg\min_{t \in \{t_1,\dots,t_L\}} \|\mathbf{C} - SP(thresh_t(\mathbf{C}))\|_F^2$ where $t_1, \dots, t_L$ are drawn from $\mathbb{R}_+$.
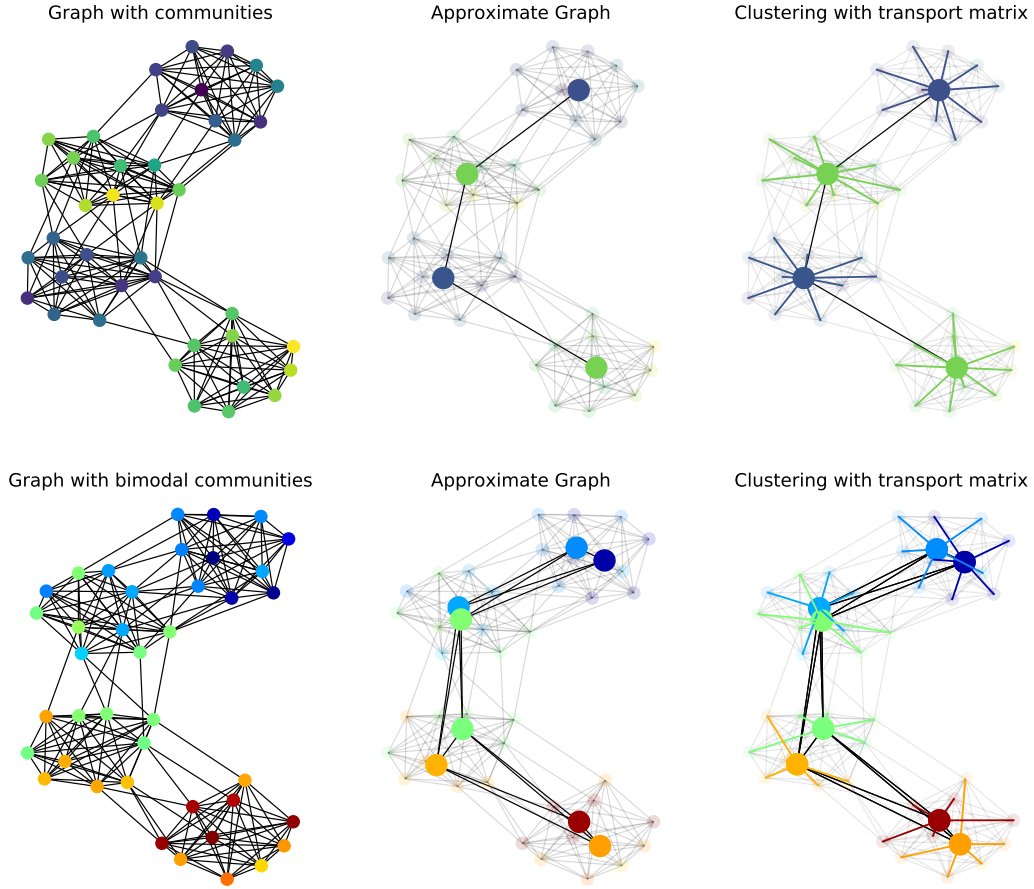
Figure 3.8:   Example of community clustering on graphs using FGW. **(top)** Community clustering with 4 communities and uniform features per cluster. **(bottom)** Community clustering with 4 communities and bimodal features per cluster (and two nodes per cluster in the approximate graph).

Resulting barycenters are showed in Figure 3.7 for $n = 15$ and $n = 7$ nodes. First, one can see that the barycenters are denoised both in the feature space and the structure space. Also note that the sharp change at the center of the 8 class is conserved in the barycenters which is a nice result compared to other divergences that tend to smooth-out their barycenters (ℓ_2 for instance). Finally, note that by selecting the number of nodes in the barycenter one can compress the graph or estimate a "high resolution" representation from all the samples. To the best of our knowledge, no other method can compute such graph barycenters. Finally, note that FGW is interpretable because the resulting OT matrix provides correspondence between the nodes from the samples and those from the barycenter.

Graph compression   In the second experiment, we evaluate the ability of FGW to perform graph approximation and compression on a Stochastic Block Model graph [Wang 1987, Nowicki 2001]. The question is to see if estimating an approximated graph can recover the relation between the blocks and perform simultaneously a community clustering on the original graph (using the coupling matrix $\boldsymbol{\pi}$). We generate two community graphs illustrated in the left column of Figure 3.8. The coarsened graph is obtained by solving the Fréchet mean formulation with $k = 1$. More precisely given an original graph μ_0

described by its features $\mathbf{B}_0$ and its structure $\mathbf{C}_0$ we look for a graph of N nodes, with N smaller than the number of nodes of the original graph, which solves:

$$\min_{\mu} FGW_{q,\alpha}(\mu, \mu_0) = \min_{\mathbf{C} \in \mathbb{R}^{N \times N}, \ \mathbf{A} \in \mathbb{R}^{N \times n}, \boldsymbol{\pi}} E_q(\mathbf{M}_{\mathbf{AB}_0}, \mathbf{C}, \mathbf{C}_0, \boldsymbol{\pi}) \tag{3.16}$$

The results are depicted in Figure 3.8. We can see that the relation between the blocks is sparse and has a "linear" structure, the example in the first line has features that follow the blocks (noisy but similar in each block) whereas the example in the second line has two modes per block. The first graph approximation (top line) is done with $N = 4$ nodes and we can recover both the blocks in the graph and the average feature on each blocks (colors on the nodes). The second problem is more complex due to the two modes per block but one can see that when approximating the graph with $N = 8$ nodes we recover both the structure between the blocks and the sub-clusters in each block, which illustrates the strength of FGW that encodes both features and structures.

3.4.7 Unsupervised learning: graphs clustering

In the last experiment, we evaluate the ability of FGW to perform a clustering of multiple graphs and to retrieve meaningful barycenters of such clusters. To do so, we generate a dataset of 4 groups of community graphs. Each graph follows a simple Stochastic Block Model [Wang 1987, Nowicki 2001] and the groups are defined *w.r.t.* the number of communities inside each graph and the distribution of their labels. The dataset is composed of 40 graphs (10 graphs per group) and the number of nodes of each graph is drawn randomly from $\{20, 30, ..., 50\}$ as illustrated in Figure 3.9. We perform a k-means clustering using the FGW barycenter defined in equation (3.6) as the centroid of the groups and the FGW distance for the cluster assignment. We fix the number of nodes of each centroid to 30. We perform a thresholding on the pairwise similarity matrix C of the centroid at the end in order to obtain an adjacency matrix for visualization purposes. The threshold value is empirically chosen with the procedure described in the previous section. The evolution of the barycenters along the iterations is reported in Figure 3.9. We can see that these centroids recover community structures and feature distributions that are representative of their cluster content. On this example, note that the clustering recovers perfectly the known groups in the dataset. To the best of our knowledge, there exists no other method able to perform a clustering of graphs and to retrieve the average graph in each cluster without having to solve a pre-image problem.

3.4.8 Other applications of Fused Gromov-Wasserstein

Since its introduction the Fused Gromov-Wasserstein distance was also used in various contexts where structured data are involved. It has found applications in molecular biology for the analysis of Single-cell RNA (scRNA-seq) sequencing data in [Cang 2020] where authors use FGW as a building block to recover spatial properties of scRNA-sequation In machine learning FGW was used for learning structured autoencoders [Xu 2020], to study the continuity of Graph Neural Network [Béthune 2020] or on domain adaptation tasks on functional Near-Infrared Spectroscopy (fNIRS) data [Lyu 2020]. The Fused Gromov-Wasserstein approach was further analysed in [Barbe 2020] where authors propose to improve the FGW distance using a smoothing strategy on the features of the graphs. They propose to incorporate a diffusion kernel on the features which result in a more robust similarity measures of labeled graphs. Authors tackle domain adaptation tasks between labeled graphs where the label information is only available in a target domain. FGW was also used on point clouds in [Puy 2020] for the estimation of scene flows *i.e.* of the 3D motion of points at the surface of objects in a scene.
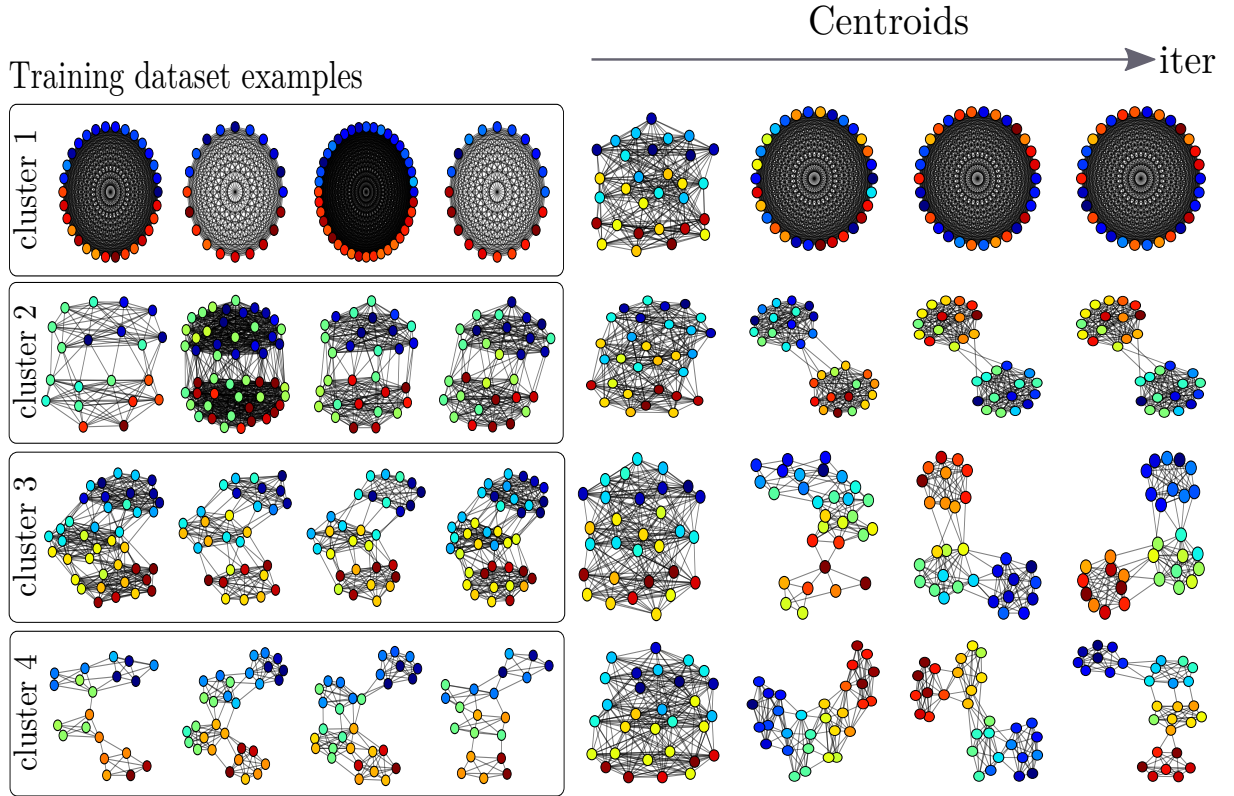
Figure 3.9: **(left)** Examples from the clustering dataset, color indicates the labels. **(right)** Evolution of the centroids of each cluster in the k-means clustering, from the random initialization until convergence to the final centroid.

3.5　FGW in the continuous setting

The previous graph representation for objects with a finite number of points/vertices extends naturally to the continuous setting. The purpose of this section is to generalize *FGW* to general probability distributions and to state some of its mathematical properties. We consider the following definition of structured objects:

Definition 3.5.1 (Structured objects). *A structured object over a metric space (Ω, d) is a triplet $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$, where: $(\mathcal{X}, d_{\mathcal{X}})$ is a metric space and μ is a probability measure over $\mathcal{X} \times \Omega$. (Ω, d) is denoted as the* feature space, *such that $d : \Omega \times \Omega \to \mathbb{R}_+$ is the distance in the feature space and $(\mathcal{X}, d_{\mathcal{X}})$ the* structure space, *such that $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the distance in the structure space. We will note μ_X and μ_A the structure and feature marginals of μ.*

Definition 3.5.2 (Space of structured objects). *We note $\mathbb{X}$ the set of all metric spaces. The space of all structured objects over (Ω, d) will be written as $\mathbb{S}(\Omega)$ and is defined by all the triplets $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ where $(\mathcal{X}, d_{\mathcal{X}}) \in \mathbb{X}$ and $\mu \in \mathcal{P}(\mathcal{X} \times \Omega)$. To avoid finiteness issues we define for $p \in \mathbb{N}^*$ the space $\mathbb{S}_p(\Omega) \subset \mathbb{S}(\Omega)$, $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu) \in \mathbb{S}_p(\Omega)$ if:*

$$\int_\Omega d(a, a_0)^p \mathrm{d}\mu_A(a) < +\infty \tag{3.17}$$

(the finiteness of this integral does not depend on the choice of a_0)

$$\int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x, x')^p \mathrm{d}\mu_X(x) d\mu_X(x') < +\infty. \tag{3.18}$$

For the sake of simplicity, and when it is clear from the context, we will sometimes denote only by μ the whole structured object. In the same way as the discrete case we will note μ_X and μ_A the structure and feature marginals of μ. We recall that those marginals encode very partial information since they focus only on independent feature distributions or only on the structure. This definition encompasses the discrete setting discussed in above. More precisely let us consider a labeled graph of n nodes with features $A = (a_i)_{i=1}^n$ with $a_i \in \Omega$ and $\mathcal{X} = (x_i)_{i=1}^n$ the structure representation of the nodes. Let $(h_i)_{i=1}^n$ be an histogram, then the probability measure $\mu = \sum_{i=1}^n h_i \delta_{(x_i, a_i)}$ defines structured object in the sense of Definition 3.5.1 since it lies in $\mathcal{P}(\mathcal{X} \times \Omega)$. In this case, an example of μ, μ_X and μ_A is provided in Figure 3.1.

Note that the set of structured objects is quite general and allows also considering discrete probability measures of the form $\mu = \sum_{i,j=1}^{p,q} h_{i,j} \delta_{(x_i, a_j)}$ with p, q possibly different than n. We propose to focus on a particular type of structured objects, namely the *generalized labeled graphs* as described in the following definition:

Definition 3.5.3 (Generalized labeled graph)**.** *We call* generalized labeled graph *a structured object* $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu) \in \mathbb{S}_p(\Omega)$ *such that μ can be expressed as $\mu = (id \times \ell_f)\#\mu_X$ where $\ell_f : \mathcal{X} \to \Omega$ is surjective and pushes μ_X forward to μ_A, i.e. $\ell_f\#\mu_X = \mu_A$.*

This definition implies that there exists a function ℓ_f which associates a feature $a = \ell_f(x)$ to a structure point $x \in \mathcal{X}$ and, since ℓ_f is surjective, one structure point can not have two different features. The labeled graph described by $\mu = \sum_{i=1}^n h_i \delta_{(x_i, a_i)}$ is a particular instance of a generalized labeled graph in which ℓ_f is defined by $\ell_f(x_i) = a_i$.

3.5.1 Comparing structured objects

We now aim to define a notion of equivalence between two structured objects $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \nu)$. We note in the following ν_Y, ν_B the marginals of ν. Intuitively, two structured objects are the same if they share the same feature information, if their structure information are lookalike and if the probability measures are corresponding in some sense. In this section, we present mathematical tools for individual comparison of the elements of structured objects. For completeness we recall here some useful mathematical tools defined in Chapter 2 for comparing the elements of structured objects and we refer the reader to this chapter for more details.

Definition 3.5.4 (Isometry)**.** *Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two metric spaces. An isometry is a sujective map $\phi : \mathcal{X} \to \mathcal{Y}$ that preserves the distances:*

$$\forall (x, x') \in \mathcal{X}^2, d_{\mathcal{Y}}(\phi(x), \phi(x')) = d_{\mathcal{X}}(x, x') \tag{3.19}$$

We refer the reader to section 2.2.2 for wider explanations about isometry. The previous map ϕ can be used in order to compare the structure information of two structured objects. When the metric spaces are enriched with a probability measure they define a *measurable metric spaces* also called *mm-spaces* (see section 2.2.2). In this case the notion of *strong isomorphism* can be used for comparing mm-spaces:

Definition 3.5.5 (Strong isomorphism)**.** *Two mm-spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu_X), (\mathcal{Y}, d_{\mathcal{Y}}, \mu_Y)$ are strongly isomorphic if there exists an isometry $\phi : supp(\mu_X) \to supp(\nu_Y)$ which pushes μ_X forward to μ_Y, i.e. $\phi\#\mu_X = \mu_Y$. In this case we say that ϕ is* measure preserving*.*

All this considered, we can now define a notion of equivalence between structured objects:

Definition 3.5.6 ((II)-Strong isomorphism of structured objects.)**.** *Two structured objects are said to be* (II)-strongly isomorphic *if there exists an isomorphism $I : supp(\mu_X) \rightarrow supp(\nu_Y)$ between the structures such that $\phi = (I, id)$ is bijective between $supp(\mu)$ and $supp(\nu)$ and measure preserving. More precisely ϕ satisfies the following properties:*

P.1 $\phi \# \mu = \nu$.

P.2 *The function ϕ statisfies:*
$$\forall (x, a) \in supp(\mu), \phi(x, a) = (I(x), a).$$

P.3 *The function $I : supp(\mu_X) \rightarrow supp(\nu_Y)$ is surjective, satisfies $I \# \mu_X = \nu_Y$ and:*
$$\forall x, x' \in supp(\mu_X)^2, d_{\mathcal{X}}(x, x') = d_{\mathcal{Y}}(I(x), I(x')).$$

Remark 3.5.1. *It is easy to check that the (II)-strong isomorphism defines an equivalence relation over $\mathbb{S}_p(\Omega)$. Moreover the function ϕ described in this definition can be seen as a feature, structure and measure preserving function. Indeed from **P.1** ϕ is measure preserving. Moreover $(\mathcal{X}, d_{\mathcal{X}}, \mu_X)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu_Y)$ are isomorphic through I. Finally using **P.1** and **P.2** we have that $\mu_A = \nu_B$ so that the feature information is also preserved.*

To illustrate this definition, we consider a simple example of two structured objects in the discrete case:

Example 3.5.1. *Let two structured objects defined by:*

$$\underbrace{\begin{pmatrix} (x_1, a_1) \\ (x_2, a_2) \\ (x_3, a_3) \\ (x_4, a_4) \end{pmatrix}}_{x_i, a_i}, \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 0 & 2 \\ 1 & 2 & 2 & 0 \end{pmatrix}}_{d_{\mathcal{X}}(x_i, x_j)}, \underbrace{\begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}}_{h_i} \quad and \quad \underbrace{\begin{pmatrix} (y_1, b_1) \\ (y_2, b_2) \\ (y_3, b_3) \\ (y_4, b_4) \end{pmatrix}}_{y_i, b_i}, \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}}_{d_{\mathcal{Y}}(y_i, y_j)}, \underbrace{\begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}}_{h_i'}$$

*with for i, $a_i = b_i$ and for $i \neq j$, $a_i \neq a_j$ (see Figure 3.10). The two structured objects have isometric structures and same features individually but they are not (II)-strongly isomorphic. One possible map $\phi = (\phi_1, \phi_2) : \mathcal{X} \times \Omega \rightarrow \mathcal{Y} \times \Omega$ such that ϕ_1 leads to an isometry is $\phi(x_1, a_1) = (y_1, b_1)$, $\phi(x_2, a_2) = (y_3, b_3)$, $\phi(x_3, a_3) = (y_4, b_4)$, $\phi(x_4, a_4) = (y_2, b_2)$. Yet this map does not satisfy $\phi_2(x, .) = id$ for any x since $\phi(x_2, a_2) = (y_3, b_3)$ and $a_2 \neq b_3$. The other possible functions such that ϕ_1 leads to an isometry are simply permutations of this example, yet it is easy to check that none of them verifies **P.2** (for example with $\phi(x_2, a_2) = (y_4, b_4)$).*

We generalize the definition (3.1) of Fused Gromov-Wasserstein (FGW) to the continuous setting as follows:

Definition 3.5.7. *The Fused-Gromov-Wasserstein distance is defined for $\alpha \in [0, 1]$ and $p, q \geq 1$ as:*

$$FGW_{\alpha, p, q}(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} E_{p, q, \alpha}(\pi) \right)^{\frac{1}{p}} \tag{3.20}$$

where:

$$E_{p, q, \alpha}(\pi) = \int \int \left((1 - \alpha) d(a, b)^q + \alpha |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|^q \right)^p d\pi((x, a), (y, b)) d\pi((x', a'), (y', b'))$$

We will write in the following $L(x, y, x', y') = |d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')|$.
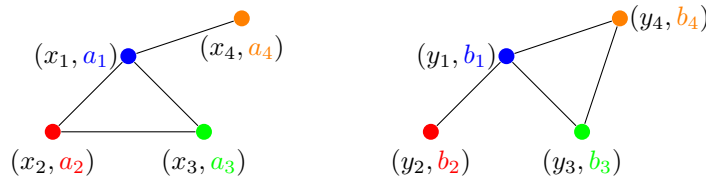
Figure 3.10: Two structured objects with isometric structures and identical features that are not (II)-strongly isomorphic. The color of the nodes represent the node feature and each edge represents a distance of 1 between the connected nodes.
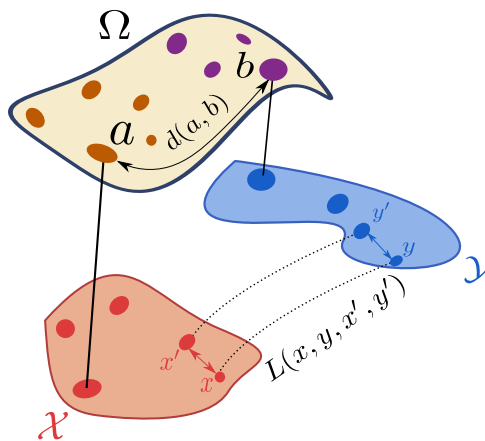
Figure 3.11: Illustration of definition 3.5.7. The figure shows two structured objects $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \mu)$. The feature space Ω is the common space for all features. The two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ represent the structures of the two structured objects, the similarity between all pair-to-pair distances of the structure points is measured by $L(x, y, x', y')$. μ and ν are the joint measures on the structure space and the feature space.

Note that this definition is coherent with the definition given in equation (3.1) when $p = 1$. For brevity we will simply note FGW instead for $FGW_{\alpha,p,q}$ when it is clear from the context. Many desirable properties arise from this definition. Among them, one can define a topology over the space of structured objects using the FGW distance to compare structured objects, in the same philosophy as for Wasserstein and Gromov-Wasserstein distances. The definition also implies that FGW acts as a generalization of both Wasserstein and Gromov-Wasserstein distances, with FGW achieving an interpolation between these two distances. More remarkably, FGW distance also realizes geodesic properties over the space of structured objects, allowing the definition of gradient flows. Before reviewing all these properties, we first compare FGW with GW and W (by assuming for now that FGW exists, which will be shown later in Theorem 3.5.1).

Proposition 3.5.1 (Comparison between FGW, GW and W.)**.** *With previous notations:*

- *The following inequalities hold:*

$$FGW_{\alpha,p,q}(\mu, \nu) \geq (1 - \alpha)W_{pq}(\mu_A, \nu_B)^q \tag{3.21}$$

$$FGW_{\alpha,p,q}(\mu, \nu) \geq \alpha GW_{pq}(\mu_X, \nu_Y)^q \tag{3.22}$$

- *Let us suppose that the structure spaces $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$ are part of a single ground space $(\mathcal{Z}, d_{\mathcal{Z}})$ (i.e. $\mathcal{X}, \mathcal{Y} \subset \mathcal{Z}$ and $d_{\mathcal{X}} = d_{\mathcal{Y}} = d_{\mathcal{Z}}$). We consider the Wasserstein distance between μ and ν for the*

distance on $\mathcal{Z} \times \Omega$: $\tilde{d}((x,a),(y,b)) = (1-\alpha)d(a,b) + \alpha d_{\mathcal{Z}}(x,y)$. Then:

$$FGW_{\alpha,p,1}(\mu,\nu)(\mu,\nu) \leq 2W_p(\mu,\nu). \tag{3.23}$$

Proof of this proposition can be found in Section 6.1.2. In particular, following this proposition, when the *FGW* distance vanishes then both *GW* and *W* distances vanish so that the structure and the feature of the structure object are individually "the same" (with respect to their corresponding equivalence relation). However the converse is not necessarily true as shown further in Section 3.5.3.

In the following we establish some mathematical properties of the *FGW* distance. The first result relates to the existence of the *FGW* distance and the topology of the space of structured objects. We prove that the *FGW* distance is indeed a distance regarding the equivalence relation between structured objects as defined in Defintion 3.5.6, allowing us to derive a topology on $\mathbb{S}(\Omega)$.

3.5.2 Topology of the structured object space

The *FGW* distance has the following properties:

Theorem 3.5.1 (Metric properties). *Let $p, q \geq 1$, $\alpha \in]0,1[$ and $(\mu,\nu) \in \mathbb{S}_{pq}(\Omega) \times \mathbb{S}_{pq}(\Omega)$. The functional $\pi \to E_{p,q,\alpha}(\pi)$ always achieves an infimum π^* in $\Pi(\mu,\nu)$ s.t. $FGW_{\alpha,p,q}(\mu,\nu) = E_{p,q,\alpha}(\pi^*) < +\infty$. Moreover:*

- *$FGW_{\alpha,p,q}$ is symmetric and, for $q = 1$, satisfies the triangle inequality. For $q \geq 2$, the triangular inequality is relaxed by a factor 2^{q-1}.*

- *For $\alpha \in]0,1[$, $FGW_{\alpha,p,q}(\mu,\nu) = 0$ if an only if there exists a bijective function $\phi = (\phi_1, \phi_2)$: $supp(\mu) \to supp(\nu)$ such that:*

$$\phi\#\mu = \nu \tag{3.24}$$

$$\forall (x,a) \in supp(\mu) \ , \phi_2(x,a) = a \tag{3.25}$$

$$\forall (x,a), (x',a') \in supp(\mu)^2, \ d_{\mathcal{X}}(x,x') = d_{\mathcal{Y}}(\phi_1(x,a), \phi_1(x',a')) \tag{3.26}$$

- *If (μ,ν) are generalized labeled graphs then $FGW_{\alpha,p,q}(\mu,\nu) = 0$ if and only if $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \nu)$ are (II)-strongly isomorphic.*

Proof of this theorem can be found in Section 6.1.3. The identity of indiscernibles is the most delicate part to prove and is based on using the Gromov-Wasserstein distance between the spaces $X \times \Omega$ and $\mathcal{Y} \times \Omega$. The previous theorem states that *FGW* is a distance over the space of generalized labeled graphs endowed with the strong isomorphism as equivalence relation defined in Definition 3.5.6. More generally for any structured objects the equivalence relation is given by equations (3.24), (3.25) and (3.26). Informally, invariants of the *FGW* are structured objects that have both the same structure and the same features in the same place. Despite the fact that $q = 1$ leads to a proper metric the case $q = 2$ can be computed more efficiently using a separability trick from [Peyré 2016] as seen in Section 3.3.3.

Remark 3.5.2. *Note that the previous theorem actually proves Theorem 3.3.1. Indeed if we consider $\mu = \sum_{i=1}^n h_i \delta_{(x_i,a_i)}$ and $\nu = \sum_{j=1}^m g_j \delta_{(y_j,b_j)}$ describing two labeled graphs as discussed in the previous section. Then μ and ν are generalized labeled graph. Using Theorem 3.5.1 μ and ν are (II)-strongly isomorphic if and only if there exists a bijection between the supports satisfies (**P.1**), (**P.2**) and (**P.3**). Since the supports are discrete this is equivalent to the condition $n = m$ and there exists a permutation*

$\sigma \in S_n$ which satisfies the conditions of Theorem 3.3.1. The triangle inequality property of Theorem 3.3.1 derives directly from the triangle inequality of Theorem 3.5.1.

There are some special cases where W and GW can be adapted to structured objects and can be used also to compare them. These cases result in different notions of equivalence as described in the following discussion.

3.5.3 Can we adapt W and GW distances for structured objects?

Despite the appealing properties of both Wasserstein and Gromov-Wasserstein distances, they fail at comparing structured objects by focusing only on the feature and structure marginals respectively. However, with some hypotheses, one could adapt these distances for structured objects.

Adapting Wasserstein: common structure space   If the structure spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ are part of a same ground space $(\mathcal{Z}, d_\mathcal{Z})$ one can build a distance $\hat{d} = d_\mathcal{Z} \oplus d$ between couples (x, a) and (y, b) and apply the Wasserstein distance. In this case, when the Wasserstein distance vanishes then the structured objects are equal in the sense $\mu = \nu$ which implies that μ and ν are *de facto* (II)-strongly isomorphic. This approach is very related with the one discussed in [Thorpe 2017] where authors define the Transportation L^p distance for signal analysis purposes. Their approach can be viewed as a transport between two joint measures:

$$\mu(\mathcal{X} \times \Omega) = \mathcal{L}(\{(\mathbf{x}, f(\mathbf{x})) \mid \mathbf{x} \in X \subset Z = \mathbb{R}^d \mid f(\mathbf{x}) \in \Omega \subset \mathbb{R}^m\}) \tag{3.27}$$

$$\nu(\mathcal{Y} \times \Omega) = \mathcal{L}(\{(\mathbf{y}, g(\mathbf{y})) \mid \mathbf{y} \in Y \subset Z = \mathbb{R}^d \mid g(\mathbf{y}) \in \Omega \subset \mathbb{R}^m\}) \tag{3.28}$$

for functions $f, g : Z \to \mathbb{R}^m$ representative of the signal values and $\mathcal{L}$ the Lebesgue measure. The distance for the transport is defined as $\hat{d}((\mathbf{x}, f(\mathbf{x})), (\mathbf{y}, g(\mathbf{y}))) = \frac{1}{\alpha}\|\mathbf{x} - \mathbf{y}\|_p^p + \|f(\mathbf{x}) - g(\mathbf{y})\|_p^p$ for $\alpha > 0$ and $\|\cdot\|_p$ the l_p norm. In this case $f(\mathbf{x})$ and $g(\mathbf{y})$ can be interpreted as encoding the feature information of the signal while $\mathbf{x}, \mathbf{y}$ encode its structure information. This approach is interesting but cannot be used on structured objects such as graphs that will not share a common structure embedding space.

Adapting Gromov-Wasserstein   The Gromov-Wasserstein distance can also be adapted to structured objects by considering the distances $(1 - \beta)d_\mathcal{X} \oplus \beta d$ and $(1 - \beta)d_\mathcal{Y} \oplus \beta d$ within each space $\mathcal{X} \times \Omega$ and $\mathcal{Y} \times \Omega$ respectively and $\beta \in]0, 1[$. When the resulting GW distance vanishes, structured objects are strongly isomorphic with respect to $(1 - \beta)d_\mathcal{X} \oplus \beta d$ and $(1 - \beta)d_\mathcal{Y} \oplus \beta d$. However the (II)-strong isomorphism is stronger than this notion since the strong isomorphism allows for "permuting the labels" but not the (II)-strong isomorphism. More precisely we have the following lemma:

Lemma 3.5.1. *Let $(\mathcal{X} \times \Omega, d_\mathcal{X}, \mu), (\mathcal{Y} \times \Omega, d_\mathcal{Y}, \nu)$ be two structured objects and $\beta \in]0, 1[$.*
If $(\mathcal{X} \times \Omega, d_\mathcal{X}, \mu)$ and $(\mathcal{Y} \times \Omega, d_\mathcal{Y}, \nu)$ are (II)-strongly isomorphic then $(\mathcal{X} \times \Omega, (1 - \beta)d_\mathcal{X} \oplus \beta d, \mu)$ and $(\mathcal{Y} \times \Omega, (1 - \beta)d_\mathcal{Y} \oplus \beta d, \nu)$ are strongly isomorphic. However the converse is not true in general.

Proof. To see this, if we consider ϕ as defined in Theorem 3.5.1, then for $(x, a), (x', b) \in \text{supp}(\mu)^2$ we have $d_\mathcal{X}(x, x') = d_\mathcal{Y}(I(x), I(x'))$. In this way:

$$(1 - \beta)d_\mathcal{X}(x, x') + \beta d(a, b) = (1 - \beta)d_\mathcal{Y}(I(x), I(x')) + \beta d(a, b) \tag{3.29}$$

which can be rewritten as:

$$(1 - \beta)d \oplus \beta d_\mathcal{X}((x, a), (x', b)) = (1 - \beta)d \oplus \beta d_\mathcal{Y}(\phi(x, a), \phi(x', b)) \tag{3.30}$$

and so ϕ is an isometry with respect to $(1 - \beta)d \oplus \beta d_{\mathcal{X}}$ and $(1 - \beta)d \oplus \beta d_{\mathcal{Y}}$. Since ϕ is also measure preserving and surjective $(\mathcal{X} \times \Omega, (1-\beta)d_{\mathcal{X}} \oplus \beta d, \mu)$ and $(\mathcal{Y} \times \Omega, (1-\beta)d_{\mathcal{Y}} \oplus \beta d, \nu)$ are strongly isomorphic.

However the converse is not necessarily true as it is easy to cook up an example with the same structure but with permuted labels so that objects are strongly isomorphic but not (II)-strongly isomorphic. For example in the tree example depicted in Figure 3.10, the structures are isometric and the distances between the features within each space are the same between each structured objects so that $(\mathcal{X} \times \Omega, (1-\beta)d_{\mathcal{X}} \oplus \beta d, \mu)$ and $(\mathcal{Y} \times \Omega, (1 - \beta)d_{\mathcal{Y}} \oplus \beta d, \nu)$ are strongly isomorphic, yet not (II)-strongly isomorphic as shown in the example since $FGW > 0$. $\qquad\qquad\square$

3.5.4 Convergence of finite samples

The metric properties of FGW naturally endow the structured object space with a notion of convergence as described in the next definition:

Definition 3.5.8 (Convergence of structured objects.)**.** *Let* $\left((\mathcal{X}_n \times \Omega, d_{\mathcal{X}_n}, \mu_n)\right)_{n \in \mathbb{N}}$ *be a sequence of structured objects. It converges to* $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ *in the Fused Gromov-Wasserstein sense if:*

$$\lim_{n \to \infty} FGW_{\alpha,p,1}(\mu_n, \mu) = 0 \tag{3.31}$$

We consider in this definition only the case $q = 1$ as it gives a proper metric (with $q > 1$ the triangle inequality is relaxed by a factor 2^{q-1}). Using Prop. 3.5.1, it is straightforward to see that if the sequence converges in the FGW sense, both the features and the structure converge respectively in the Wasserstein and Gromov-Wasserstein sense.

An interesting question arises from this definition. If we consider a structured object $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ and if we sample the joint distribution so as to consider $(\{(x_i, a_i)\}_{i \in \{1,..,n\}}, d_{\mathcal{X}}, \mu_n)_{n \in \mathbb{N}}$ with $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i, a_i}$ where $(x_i, a_i) \in \mathcal{X} \times \Omega$ are sampled from μ. Does this sequence converges to $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ in the FGW sense and how fast is the convergence?

This question can be answered thanks to a notion of "size" of a probability measure. For the sake of conciseness we will not present exhaustively the theory but the reader can refer to [Weed 2017] for more details. Given a measure μ on Ω we denote as $dim_p^*(\mu)$ its *upper Wasserstein dimension*. It coincides with the intuitive notion of "dimension" when the measure is sufficiently well behaved. For example, for any absolutely continuous measure μ with respect to the Lebesgue measure on $[0, 1]^d$, we have $dim_p^*(\mu) = d$ for any $p \in [1, \frac{d}{2}]$. Using this definition and the results in [Weed 2017], we can answer the question of convergence of finite sample in the following proposition (proof can be found in Section 6.1.4):

Theorem 3.5.2 (Convergence of finite samples and a concentration inequality)**.** *With previous notations. Let* $p \geq 1$. *We have:*

$$\lim_{n \to \infty} FGW_{\alpha,p,1}(\mu_n, \mu) = 0 \tag{3.32}$$

Moreover, suppose that $s > d_p^*(\mu)$. *Then there exists a constant* C *that does not depend on* n *such that:*

$$\mathbb{E}[FGW_{\alpha,p,1}(\mu_n, \mu)] \leq Cn^{-\frac{1}{s}}. \tag{3.33}$$

The expectation is taken over the i.i.d samples (x_i, a_i). *A particular case of this inequality is when* $\alpha = 1$ *so that we can use the result above to derive a concentration result for the Gromov-Wasserstein distance. More precisely, if* $\nu_n = \frac{1}{n} \sum_i \delta_{x_i}$ *denotes the empirical measure of* $\nu \in \mathcal{P}(\mathcal{X})$ *and if*

$s' > d_p^*(\nu)$ *we have:*

$$\mathbb{E}[GW_p(\nu_n, \nu)] \leq C' n^{-\frac{1}{s'}}. \tag{3.34}$$

This result is a simple application of the convergence of finite sample properties of the Wasserstein distance, since in this case μ_n and μ are part of the same ground space so that equation (3.33) derives naturally from equation (3.23) and the properties of Wasserstein. In contrast to the Wasserstein distance case this inequality is not necessarily sharp and future work will be dedicated to the study of its tightness.

3.5.5 Interpolation properties between W and GW distances

FGW distance is a generalization of both Wasserstein and Gromov-Wasserstein distances in the sense that it achieves an interpolation between them. More precisely, we have the following theorem:

Proposition 3.5.2 (Interpolation properties.)**.** *As α tends to zero, one recovers the Wasserstein distance between the features information and as α goes to one, one recovers the Gromov-Wasserstein distance between the structure information:*

$$\lim_{\alpha \to 0} FGW_{\alpha,p,q}(\mu, \nu) = (W_{pq}(\mu_A, \nu_B))^q \tag{3.35}$$

$$\lim_{\alpha \to 1} FGW_{\alpha,p,q}(\mu, \nu) = (GW_{pq}(\mu_X, \nu_Y))^q \tag{3.36}$$

Proof of this proposition can be found in Section 6.1.5. This result shows that FGW can revert to one of the other distances and thus acts as a generalization of Wasserstein and Gromov-Wasserstein distances, as claimed in the discrete case section.

3.5.6 Geodesic properties

One desirable property in OT is the underlying geodesics defined by the mass transfer between two probability distributions. These properties are useful in order to define dynamic formulation of OT problems. This dynamic point of view is inspired by fluid dynamics and finds its origin in the Wasserstein context with [Benamou 2000]. Various applications in machine learning can be derived from this formulation: interpolation along geodesic paths was used in computer graphics for color or illumination interpolations [Bonneel 2011]. More recently, [Chizat 2018] used Wasserstein gradient flows in an optimization context, deriving global minima results for non-convex particles gradient descent. In [Zhang 2018] authors used Wasserstein gradient flows in the context of reinforcement learning for policy optimization.

The main idea of this dynamic formulation is to describe the optimal transport problem between two measures as a curve in the space of measures minimizing its total length. We first describe some generality about geodesic spaces and recall classical results for dynamic formulation in both Wasserstein and Gromov-Wasserstein contexts. In a second part, we derive new geodesic properties in the FGW context.

Geodesic spaces Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and x, y two points in $\mathcal{X}$. We say that a curve $w : [0, 1] \to \mathcal{X}$ joining the *endpoints* x and y (*i.e.* with $w(0) = x$ and $w(1) = y$) is a *constant speed geodesic* if it satisfies $d_{\mathcal{X}}(w(t), w(s)) \leq |t - s| d(w(0), w(1)) = |t - s| d_{\mathcal{X}}(x, y)$ for $t, s \in [0, 1]$. Moreover, if $(\mathcal{X}, d_{\mathcal{X}})$ is a length space (*i.e.* if the distance between two points of $\mathcal{X}$ is equal to the infimum of the lengths of the curves connecting these two points) then the converse is also true and a constant speed geodesic satisfies $d_{\mathcal{X}}(w(t), w(s)) = |t - s| d_{\mathcal{X}}(x, y)$. It is easy to compute distances along such curve as they are directly embedded into $\mathbb{R}$.

In the Wasserstein context, if the ground space is a complete separable, locally compact length space and if the endpoints of the geodesic are given, then there exists a geodesic curve. Moreover, if the transport between the endpoints is unique then there is a unique displacement interpolation between the endpoints (see Corollary 7.22 and 7.23 in [Villani 2008]). For example, if the ground space is $\mathbb{R}^d$ and the distance between the points is measured via the $\|.\|_2$ norm, then the geodesics exist and are uniquely determined (note that this can be generalized to costs of the form $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{y} - \mathbf{x})$ where h is strictly convex). In the Gromov-Wasserstein context, there always exists constant speed geodesics as long as the endpoints are given. These geodesics are unique modulo strong isomorphisms (see [Sturm 2012]).

The _FGW_ case   In this paragraph, we suppose that $\Omega = \mathbb{R}^d$. We are interested in finding a geodesic curve in the space of structured objects _i.e._ a constant speed curve of structured objects joining two structured objects. As for Wasserstein and Gromov-Wasserstein, the structured object space endowed with the Fused Gromov-Wasserstein distance maintains some geodesic properties. The following result proves the existence of such a geodesic and characterizes it:

Theorem 3.5.3 (Constant speed geodesic.). _Let $p \geq 1$ and $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu_0)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \mu_1)$ in $\mathbb{S}_p(\mathbb{R}^d)$. Let π^* be an optimal coupling for the Fused Gromov-Wasserstein distance between μ_0, μ_1 and $t \in [0, 1]$. We equip $\mathbb{R}^d$ with ℓ_m norm for $m \geq 1$._

We define $\eta_t : \mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega \to \mathcal{X} \times \mathcal{Y} \times \Omega$ such that:

$$\forall (x, \mathbf{a}), (y, \mathbf{b}) \in \mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega, \ \eta_t(x, \mathbf{a}, y, \mathbf{b}) = (x, y, (1-t)\mathbf{a} + t\mathbf{b}) \tag{3.37}$$

Then:

$$(\mathcal{X} \times \mathcal{Y} \times \Omega, (1-t)d_{\mathcal{X}} \oplus td_{\mathcal{Y}}, \mu_t = \eta_t \# \pi^*)_{t \in [0,1]} \tag{3.38}$$

is a constant speed geodesic connecting $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu_0)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \mu_1)$ in the metric space $\big(\mathbb{S}_p(\mathbb{R}^d), FGW_{\alpha,p,1}\big)$.

Proof of the previous theorem can be found in Section 6.1.6. In a sense this result combines the geodesics in the Wasserstein space and in the space of all mm-spaces since it suffices to interpolate the distances in the structure space and the features to construct a geodesic. The main interest is that it defines the minimum path between two structured objects. For example, considering two discrete structured objects represented by the measures $\mu_0 = \sum_{i=1}^n h_i \delta_{(x_i, \mathbf{a}_i)}$ and $\mu_1 = \sum_{j=1}^m g_j \delta_{(y_j, \mathbf{b}_j)}$, the interpolation path is given for $t \in [0, 1]$ by the measure $\mu_t = \sum_{i=1}^n \sum_{j=1}^m \pi^*(i,j) \delta_{(x_i, y_j, (1-t)\mathbf{a}_i + t\mathbf{b}_j)}$ where π^* is an optimal coupling for the _FGW_ distance. However this geodesic is difficult to handle in practice since it requires the computation of the cartesian product $\mathcal{X}_0 \times \mathcal{X}_1$. The Fréchet mean defined in Section 3.3.2 seems to be more suited in practice. The proper definition and properties of velocity fields associated to this geodesic is postponed to further works.

3.6   Discussion and conclusion

Countless problems in machine learning involve structured data, usually stressed in light of the graph formalism. We consider here labeled graphs enriched by an histogram, which naturally leads to represent structured data as probability measures in the joint space of their features and structures. Widely known for their ability to meaningfully compare probability measures, transportation distances are generalized in this chapter so as to be suited in the context of structured data, motivating the so-called Fused

Gromov-Wasserstein distance. We theoretically prove that it defines indeed a distance on structured data, and consequently on graphs of arbitrary sizes. FGW provides a natural framework for analysis of labeled graphs as we demonstrate on classification, where it reaches and surpasses most of the time the state-of-the-art performances, and in graph-based k-means where we develop a novel approach to represent the clusters centroids using a barycentric formulation of FGW. We believe that this metric can have a significant impact on challenging graph signal analysis problems.

While we considered a unique measure of distance between nodes in the graph structure (shortest path), other choices could be made with respect to the problem at hand, or eventually learned in an end-to-end manner. The same applies to the distance between features. We also envision a potential use of this distance in deep learning applications where a distance between graph is needed (such as graph auto-encoders). Another line of work will also try to lower the computational complexity of the underlying optimization problem to ensure better scalability to very large graphs.

The Gromov-Wasserstein problem in Euclidean spaces

Contents

Summary of the contributions

This chapter is based on the paper [Vayer 2019b] and addresses the problem of GW in Euclidean spaces. Recently used in various machine learning contexts, the Gromov-Wasserstein distance allows for comparing distributions which supports do not necessarily lie in the same metric space. However, this optimal transport distance requires solving a complex non convex quadratic program which is most of the time very costly both in time and memory. Contrary to GW, the Wasserstein distance enjoys several properties (*e.g.* duality) that permit large scale optimization. Among those, the solution of W on the real line, that only requires sorting discrete samples in 1D, allows defining the Sliced Wasserstein (SW) distance. This first part of this chapter presents a new divergence based on GW akin to SW. More precisely the contributions of are the following:

- We derive the first closed form solution for GW when dealing with discrete 1D distributions, based on a new result for the related quadratic assignment problem (Theorem 4.1.1 and Theorem 4.1.2).

- Based on this result we define a novel OT discrepancy which can deal with large scale distributions via a slicing approach and we show how it relates to the GW distance while being $O(n \log(n))$ to compute.

- We illustrate the behavior of this so called Sliced Gromov-Wasserstein (SGW) discrepancy in experiments where we demonstrate its ability to tackle similar problems as GW while being several order of magnitudes faster to compute.

The second part of this chapter is more prospective and tackle the problem of probability distributions which supports lie on Euclidean spaces with, potentially, different dimensions. This part investigate the regularity of GW optimal transport plan in the cases of inner product similarities and Euclidean distances. The contributions of this part are, in summary:

- We show that the GW problem in Euclidean spaces is equivalent to jointly solve a linear transportation problem and a "alignment" problem (Theorem 4.2.1 and Theorem 4.2.5).

- We give necessary conditions under which a GW optimal transport plan is supported on a deterministic function (Theorem 4.2.3 and Proposition 4.2.4). This allows to derive a closed-form expression for GW with inner product similarities between 1D probability distributions (not necessarily discrete, see Theorem 4.2.4).

- We study the Gromov-Monge problem in Euclidean spaces, and in particular the *linear* Gromov-Monge problem for which we exhibit a closed-form expression between Gaussian distributions (Theorem 4.2.6).

4.1 Sliced Gromov-Wasserstein

4.1.1 Introduction

As described in Chapter 2 the linear optimal transport problem aims at defining ways to compare probability distributions, through *e.g.* the Wasserstein distance. It has proved to be very useful for a wide range of machine learning tasks including generative modelling (Wasserstein GANs [Arjovsky 2017]), domain adaptation [Courty 2017] or supervised embeddings for classification purposes [Huang 2016]. However one limitation of this approach is that it implicitly assumes *aligned* distributions, *i.e.* that lie in the same metric space or at least between spaces where a meaningful distance *across* domains can be computed. From another perspective, the Gromov-Wasserstein distance benefits from more flexibility when it comes to the more challenging scenario where heterogeneous distributions are involved, *i.e.* distributions which supports do not necessarily lie on the same metric space. It only requires modelling the topological or relational aspects of the distributions *within* each domain in order to compare them. As such, it has recently received a high interest in the machine learning community, solving learning tasks such as heterogenous domain adaptation [Yan 2018], deep metric alignment [Ezuz 2017], graph classification (see Chapter 3 for more details) or generative modelling [Bunne 2019].

OT is known to be a computationally difficult problem: the Wasserstein distance involves a linear program that most of the time prevents its use to settings with more than a few tens of thousands of points. For medium to large scale problems, some methods relying *e.g.* on entropic regularization or dual formulation (as seen in Chapter 2) have been investigated in the past years. Among them, one builds upon the mono-dimensional case where computing the Wasserstein distance can be trivially solved in $O(n \log n)$ by sorting points in order and pairing them from left to right. While this 1D case has a limited interest *per se*, it is one of the main ingredients of the *Sliced* Wasserstein distance [Rabin 2011]: high-dimensional data are linearly projected into sets of mono-dimensional distributions, the sliced Wasserstein distance being the average of the Wasserstein distances between all projected measures. This framework provides an efficient algorithm that can handle millions of points and has similar properties to the Wasserstein distance [Bonnotte 2013]. As such, it has attracted attention and has been successfully used in various tasks such as barycenter computation [Bonneel 2015], classification [Kolouri 2016] or generative modeling [Kolouri 2019b, Deshpande 2018, Liutkus 2019, Wu 2019].

Regarding GW, the optimization problem is a non-convex quadratic program, with a prohibitive computational cost for problems with more than a few thousands of points: the number of terms grows quadratically with the number of samples and one cannot rely on a dual formulation as for Wasserstein. However several approaches have been proposed to tackle its computation. Initially approximated by a linear lower bound, GW was thereafter estimated through an entropy regularized version that can be efficiently computed by iterating Sinkhorn projections (see Chapter 2) or using a conditional gradient scheme relying on linear program OT solvers (see Chapter 3). However, all these methods are still too costly for large scale *scenarii*. In this section, we propose a new formulation related to GW that lowers its computational cost. To that extent, we derive a novel OT discrepancy called Sliced Gromov-Wasserstein (SGW). It is similar in spirit to the Sliced Wasserstein distance as it relies on the exact computation of 1D GW distances of distributions projected onto random directions. We notably provide the first 1D closed-form solution of the GW problem by proving a new result about the Quadratic Assignment Problem for matrices that are squared euclidean distances of real numbers. Computation of SGW for discrete distributions of n points is $O(L \, n \log(n))$, where L is the number of sampled directions. This complexity is the same as the Sliced-Wasserstein distance and is even lower than computing the value of GW which is $O(n^3)$ for a known coupling (once the optimization problem solved) in the general case [Peyré 2016]. Experimental validation shows that SGW retains various properties of GW while being much cheaper to compute, allowing its use in difficult large scale settings such as large mesh matching or generative adversarial networks.

4.1.2 From 1D GW to Sliced Gromov-Wasserstein

We first provide and prove a solution for an 1D Quadratic Assignement Problem with a quasilinear time complexity of $O(n \log(n))$. This new special case of the QAP is shown to be equivalent to the *hard assignment* version of GW, called the Gromov-Monge (GM) problem, with squared Euclidean cost for distributions lying on the real line. We also show that, in this context, solving GM is equivalent to solving GW. We derive a new discrepancy named Sliced Gromov-Wasserstein (SGW) that relies on these findings for efficient computation.

Solving a Quadratic Assignment Problem in 1D In Koopmans-Beckmann form [Koopmans 1957] a QAP takes as input two $n \times n$ matrices $\mathbf{A} = (a_{ij})$, $\mathbf{B} = (b_{ij})$. The goal is to find a permutation $\sigma \in S_n$, the set of all permutations of $[\![n]\!]$, which minimizes the objective function $\sum_{i,j=1}^{n} a_{i,j} b_{\sigma(i),\sigma(j)}$. In

full generality this problem is NP-hard (see Section 2.2.3 for more details). The following theorem is a new result about QAP and states that it can be solved in polynomial time when $\mathbf{A}$ and $\mathbf{B}$ are squared Euclidean distance matrices of sorted real numbers:

Theorem 4.1.1 (A new special case for the Quadratic Assignment Problem). *For real numbers $x_1 < \cdots < x_n$ and $y_1 < \cdots < y_n$,*

$$\min_{\sigma \in S_n} \sum_{i,j} -(x_i - x_j)^2 (y_{\sigma(i)} - y_{\sigma(j)})^2 \tag{4.1}$$

is achieved either by the identity permutation $\sigma(i) = i$ (Id) or the anti-identity permutation $\sigma(i) = n + 1 - i$ (anti $-$ Id). In other words:

$$\exists \sigma \in \{Id, anti - Id\}, \; \sigma \in \arg\min_{\sigma \in S_n} \sum_{i,j} -(x_i - x_j)^2 (y_{\sigma(i)} - y_{\sigma(j)})^2 \tag{4.2}$$

To the best of our knowledge, this result is new. It states that if one wants to find the best one-to-one correspondence of real numbers such that their pairwise distances are best conserved, it suffices to sort the points and check whether the identity has a better cost than the anti-identity. Proof of this theorem can be found in Section 6.2.1. We postulate that this result also holds for $a_{ij} = |x_i - x_j|^k$ and $b_{ij} = -|y_i - y_j|^k$ with any $k \geq 1$ but leave this study for future works.

Gromov-Wasserstein distance on the real line   When $n = m$ and $a_i = b_j = \frac{1}{n}$, one can look for the *hard assignment* version of the *GW* distance resulting in the Gromov-Monge problem [Mémoli 2018] associated with the following *GM* distance:

$$GM_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = \min_{\sigma \in S_n} \frac{1}{n^2} \sum_{i,j} \left| c_{\mathcal{X}}(x_i, x_j) - c_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)}) \right|^2 \tag{4.3}$$

where $\sigma \in S_n$ is a one-to-one mapping $[\![n]\!] \to [\![n]\!]$. Interestingly when the permutation σ is known, the computation of the cost is $O(n^2)$ which is far better than $O(n^3)$ for the general *GW* case. It is easy to see that this problem is equivalent to minimizing $\sum_{i,j=1}^n a_{i,j} b_{\sigma(i),\sigma(j)}$ with $a_{ij} = c_{\mathcal{X}}(x_i, x_j)$ and $b_{ij} = -c_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)})$. Indeed we have:

$$\sum_{i,j} \left| c_{\mathcal{X}}(x_i, x_j) - c_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)}) \right|^2 = \sum_{i,j} c_{\mathcal{X}}(x_i, x_j)^2 + \sum_{i,j} c_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)})^2 - 2 \sum_{i,j} c_{\mathcal{X}}(x_i, x_j) c_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)})$$

$$= \sum_{i,j} c_{\mathcal{X}}(x_i, x_j)^2 + \sum_{i,j} c_{\mathcal{Y}}(y_i, y_j)^2 - 2 \sum_{i,j} c_{\mathcal{X}}(x_i, x_j) c_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)})$$

So that only the term $-2 \sum_{i,j} c_{\mathcal{X}}(x_i, x_j) c_{\mathcal{Y}}(y_{\sigma(i)}, y_{\sigma(j)})$ depends on σ. Thus, when squared Euclidean costs are used for distributions lying on the real line, Theorem 4.1.1 exactly recovers the solution of the *GM* problem defined in equation (4.3). As matter of consequence, Theorem 4.1.1 provides an efficient way of solving the Gromov-Monge problem.

Moreover, this theorem also allows finding a closed-form for the *GW* distance. Indeed, some recent advances in graph matching state that, under some conditions on $\mathbf{A}$ and $\mathbf{B}$, the assignment problem is equivalent to its *soft-assignment* counterpart [Maron 2018]. This way, using both Theorem 4.1.1 and [Maron 2018], one can find a solvable case for the *GW* distance as stated in the following theorem:

> **Theorem 4.1.2** (Equivalence between GW and GM for discrete measures)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$, $\nu \in \mathcal{P}(\mathbb{R}^q)$ be discrete probability measures with same number of atoms and uniform weights, i.e. $\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{x}_i}, \nu = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{y}_i}$ with $\mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^q$. For $\mathbf{x} \in \mathbb{R}^p$ we note $\|\mathbf{x}\|_{2,p} = \sqrt{\sum_{i=1}^{p} |x_i|^2}$ the ℓ_2 norm on $\mathbb{R}^p$ (same for $\mathbb{R}^q$). Let $c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_{2,p}^2$, $c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_{2,q}^2$. Then:*
>
> $$GW_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = GM_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) \tag{4.4}$$
>
> *Moreover when $p = q = 1$, i.e. $c_{\mathcal{X}}(x, x') = c_{\mathcal{Y}}(x, x') = |x - x'|^2$, and $x_1 < \cdots < x_n$ and $y_1 < \cdots < y_n$ the optimal values are achieved by considering either the identity or the anti-identity permutation.*

A detailed proof is provided in Section 6.2.2. Note also that, while both possible solutions for problem (4.3) can be computed in $O(n \log(n))$, finding the best one requires the computation of the cost which seems, at first sight, to have a $O(n^2)$ complexity. However, under the hypotheses of squared Euclidean distances, the cost can be computed in $O(n)$. Indeed, in this case, one can develop the sum in equation (4.3) to compute it in $O(n)$ operations using binomial expansion (see details in Section 6.2.3) so that the overall complexity of finding the best assignment and computing the cost is $O(n \log(n))$ which is the same complexity as the Wasserstein for 1D distributions.

Sliced Gromov-Wasserstein discrepancy Theorem 4.1.2 can be put in perspective with the Wasserstein distance for 1D distributions which is achieved by the identity permutation when points are sorted [Peyré 2019]. As explained in Chapter 2, this result was used to approximate the Wasserstein distance between measures of $\mathbb{R}^q$ using the so called Sliced Wasserstein (SW) distance [Bonneel 2015]. The main idea is to project the points of the measures on lines of $\mathbb{R}^q$ where computing a Wasserstein distance is easy since it only involves a simple sort and to average these distances. In the same philosophy we build upon Theorem 4.1.2 to define a "sliced" version of the GW distance. In the following, we consider $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ be probability distributions (not necessarily discrete).

Let $\mathbb{S}^{q-1} = \{\boldsymbol{\theta} \in \mathbb{R}^q | \|\boldsymbol{\theta}\|_2 = 1\}$ be the q-dimensional hypersphere and λ_{q-1} the uniform measure on $\mathbb{S}^{q-1}$. For $\boldsymbol{\theta}$ we note $P_{\boldsymbol{\theta}}$ the projection on $\boldsymbol{\theta}$, i.e. $P_{\boldsymbol{\theta}}(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\theta} \rangle$. For a linear map $\Delta \in \mathbb{R}^{q \times p}$ (identified with slight abuses of notation by its corresponding matrix), we define the Sliced Gromov-Wasserstein (SGW) discrepancy as follows:

$$SGW_\Delta(\mu, \nu) = \underset{\boldsymbol{\theta} \sim \lambda_{q-1}}{\mathbb{E}} [GW_2^2(P_{\boldsymbol{\theta}} \# \mu_\Delta, P_{\boldsymbol{\theta}} \# \nu)] = \int_{\mathbb{S}^{q-1}} GW_2^2(d^2, P_{\boldsymbol{\theta}} \# \mu_\Delta, P_{\boldsymbol{\theta}} \# \nu) d\lambda_{q-1}(\boldsymbol{\theta}) \tag{4.5}$$

where $\mu_\Delta = \Delta \# \mu \in \mathcal{P}(\mathbb{R}^q)$. The function Δ acts as a mapping for a point in $\mathbb{R}^p$ of the measure μ onto $\mathbb{R}^q$. When $p = q$ and when we consider Δ as the identity map we simply write $SGW(\mu, \nu)$ instead of $SGW_{\mathbf{I_p}}(\mu, \nu)$. When $p < q$, one straightforward choice is $\Delta = \Delta_{pad}$ the "uplifting" operator which pads each point of the measure with zeros: $\Delta_{pad}(\mathbf{x}) = (x_1, \ldots, x_p, \underbrace{0, \ldots, 0}_{q-p})$. The procedure is illustrated in Fig 4.1.

In general fixing Δ implies that some properties of GW, such as the rotational invariance, are lost. Consequently, we also propose a variant of SGW that does not depends on the choice of Δ called Rotation Invariant SGW ($RISGW$) and expressed for $p \geq q$ as the following:

$$RISGW(\mu, \nu) = \min_{\Delta \in \mathbb{V}_p(\mathbb{R}^q)} SGW_\Delta(\mu, \nu). \tag{4.6}$$
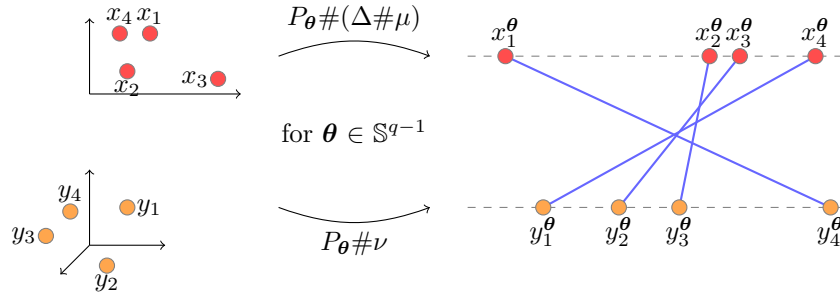
Figure 4.1: Example in dimension $p = 2$ and $q = 3$ **(left)** that are projected on the line **(right)**. The solution for this projection is the anti-diagonal coupling.

We propose to minimize SGW_Δ with respect to Δ in the Stiefel manifold $\mathbb{V}_p(\mathbb{R}^q)$ [Absil 2009] which is defined as $\mathbb{V}_p(\mathbb{R}^q) = \{\Delta \in \mathbb{R}^{q \times p} | \Delta^T \Delta = \mathbf{I_p}\}$. It can be seen as finding an optimal projector of the measure μ [Paty 2019, Deshpande 2019]. This formulation comes at the cost of an additional optimization step but allows recovering one key property of GW. When $p = q$ this encompasses for *e.g.* all rotations of the space, making $RISGW$ invariant by rotation.

Interestingly enough, SGW holds various properties of the GW distance as summarized in the following theorem:

Theorem 4.1.3 (Properties of SGW).

- *For all Δ, SGW_Δ and $RISGW$ are translation invariant. $RISGW$ is also rotational invariant when $p = q$, more precisely if $\mathbf{Q} \in \mathcal{O}(p)$ is an orthogonal matrix, $RISGW(\mathbf{Q}\#\mu, \nu) = RISGW(\mu, \nu)$ (same for any $\mathbf{Q}' \in \mathcal{O}(q)$ applied on ν).*

- *SGW and $RISGW$ are pseudo-distances on $\mathcal{P}(\mathbb{R}^p)$, i.e. they are symmetric, satisfy the triangle inequality and $SGW(\mu, \mu) = RISGW(\mu, \mu) = 0$.*

- *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ be probability distribution with compact supports. If $SGW(\mu, \nu) = 0$ then μ and ν are isomorphic for the distance induced by the ℓ_1 norm on $\mathbb{R}^p$, i.e. $d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^p |x_i - x_i'|$ for $(x, x') \in \mathbb{R}^p \times \mathbb{R}^p$. In particular this implies:*

$$SGW(\mu, \nu) = 0 \implies GW_2(d, d, \mu, \nu) = 0 \tag{4.7}$$

(with a slight abuse of notation we identify the matrix $\mathbf{Q}$ by its linear application). A proof of this theorem can be found in Section 6.2.4. This theorem states that if SGW vanishes then measures must be isomorphic, as it is the case for GW. It states also that $RISGW$ holds most of the properties of GW in term of invariants.

Remark 4.1.1. *The Δ map can also be used in the context of the Sliced Wasserstein distance so as to define $SW_\Delta(\mu, \nu)$, $RISW(\mu, \nu)$ for $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^q)$ with $p \neq q$. Please note that from a purely computational point of view, complexities of these discrepancies are the same as SGW and RISGW when μ and ν are discrete measures with the same number of atoms $n = m$, and uniform weights. Also, unlike SGW and RISGW, these discrepancies are not translation invariant. This approach was studied in [Lai 2014] for the case $p = q$ in the context of point cloud registration. More details are given in Section 6.2.5.*

Algorithm 6 Sliced Gromov-Wasserstein for discrete measures

1: $p < q$, $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i} \in \mathcal{P}(\mathbb{R}^p)$ and $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{y}_j} \in \mathcal{P}(\mathbb{R}^q)$

2: $\forall i, \mathbf{x}_i \leftarrow \Delta(\mathbf{x}_i)$, sample uniformly $(\boldsymbol{\theta}_l)_{l=1,\dots,L} \in \mathbb{S}^{q-1}$

3: **for** $l = 1,\dots,L$ **do**

4:      Sort $(\langle \mathbf{x}_i, \boldsymbol{\theta}_l\rangle)_i$ and $(\langle \mathbf{y}_j, \boldsymbol{\theta}_l\rangle)_j$ in increasing order

5:      Solve (4.3) for reals $(\langle \mathbf{x}_i, \boldsymbol{\theta}_l\rangle)_i$ and $(\langle \mathbf{y}_j, \boldsymbol{\theta}_l\rangle)_j$, $\sigma_{\boldsymbol{\theta}_l}$ (Anti-Id or Id is a solution)

6: **end for**

7: return $\frac{1}{n^2 L}\sum_{l=1}^{L}\sum_{i,k=1}^{n}\left(\langle\mathbf{x}_i-\mathbf{x}_k,\boldsymbol{\theta}_l\rangle^2-\langle\mathbf{y}_{\sigma_{\boldsymbol{\theta}_l}(i)}-y_{\sigma_{\boldsymbol{\theta}_l}(k)},\boldsymbol{\theta}_l\rangle^2\right)^2$

Computational aspects In the following μ,ν are *discrete measures* with the *same* number of atoms $n = m$, and *uniform weights*, *i.e.* $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i}, \nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{y}_i}$ with $\mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^q$ so that we can apply Theorem 4.1.2. Similarly to Sliced Wasserstein, SGW can be approximated by replacing the integral by a finite sum over randomly drawn directions. In practice we compute SGW as the average of GW_2^2 projected on L directions θ. While the sum in (4.5) can be implemented with libraries such as Pykeops [Charlier 2018], Theorem 4.1.2 shows that computing (4.5) is achieved by an $O(n\log(n))$ sorting of the projected samples and by finding the optimal permutation which is either the identity or the anti identity. Moreover computing the cost is $O(n)$ for each projection as explained previously. Thus the overall complexity of computing SGW with L projections is $O(Ln(p+q)+Ln\log(n)+Ln) = O(Ln(p+q+\log(n)))$ when taking into account the cost of projections. The pseudo-code for SGW is presented in Algorithm 6 Note that these computations can be efficiently implemented in parallel on GPUs with modern toolkits such as Pytorch [Paszke 2017].

The complexity of solving $RISGW$ is higher but one can rely on efficient algorithms for optimizing on the Stiefel manifold [Absil 2009] that have been implemented in several toolboxes [Townsend 2016, Meghwanshi 2018]. Note that each iteration in a manifold gradient decent requires the solution of SGW, that can be computed and differentiated efficiently with the frameworks described above. Moreover, the optimization over the Stiefel manifold does not depend on the number of points but only on the dimension d of the problem so that overall complexity is $n_{\text{iter}}(Ln(d+\log(n))+d^3)$, which is affordable for small d. In practice, we observed in the numerical experiments that RISGW converges in few iterations (the order of 10).

4.1.3 Experimental results

The goal of this section is to validate SGW and its rotational invariant on both quantitative (execution time) and qualitative sides. All the experiments were conducted on a standard computer equipped with a NVIDIA Titan X GPU.

SGW and RISGW on spiral dataset As a first example, we use the spiral dataset from sklearn toolbox and compute GW, SGW and $RISGW$ on $n = 100$ samples with $L = 20$ sampled lines for different rotations of the target distribution. The optimization of Δ on the Stiefel manifold is performed using Pymanopt [Townsend 2016] with automatic differentiation with autograd [Maclaurin 2015]. Some examples of empirical distributions are available in Figure 4.2 (left). The mean value of GW, SGW and $RISGW$ are reported on Figure 4.2 (right) where we can see that $RISGW$ is invariant to rotation as GW whereas SGW with $\Delta = \mathbf{I_p}$ is clearly not.
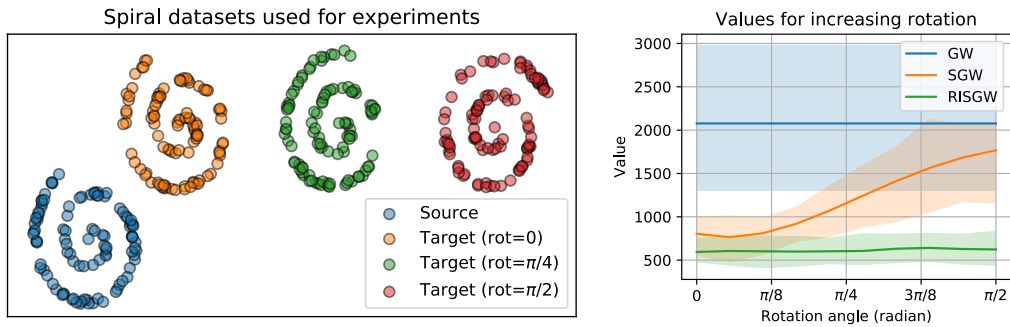
Figure 4.2: Illustration of SGW, $RISGW$ and GW on spiral dataset for varying rotations on discrete 2D spiral dataset. **(left)** Examples of spiral distributions for source and target with different rotations. **(right)** Average value of SGW, GW and $RISGW$ with $L = 20$ as a function of rotation angle of the target. Colored areas correspond to the 20% and 80% percentiles.

Runtimes comparison   We perform a comparison between runtimes of SGW, GW and its entropic counterpart [Solomon 2016]. We calculate these distances between two 2D random measures of $n \in \{1e2, ..., 1e6\}$ points. For SGW, the number of projections L is taken from $\{50, 200\}$. We use the Python Optimal Transport (POT) toolbox [Flamary 2017] to compute GW distance on CPU. For entropic-GW we use the Pytorch GPU implementation from [Bunne 2019] that uses the log-stabilized Sinkhorn algorithm [Schmitzer 2016] with a regularization parameter $\varepsilon = 100$. For SGW, we implemented both a Numpy implementation and a Pytorch implementation running on GPU. Figure 4.3 illustrates the results.

SGW is the only method which scales *w.r.t.* the number of samples and allows computation for $n > 10^4$. While entropic-GW uses GPU, it is still slow because the gradient step size in the algorithm is inversely proportional to the regularization parameter [Peyré 2016] which highly curtails the convergence of the method. On CPU, SGW is two orders of magnitude faster than GW. On GPU, SGW is five orders of magnitude faster than GW and four orders of magnitude faster than entropic GW. Still the slope of both GW implementations are surprisingly good, probably due to their maximum iteration stopping criteria. In this experiment we were able to compute SGW between 10^6 points in 1s. Finally note that we recover exactly a quasi-linear slope, corresponding to the $O(n \log(n))$ complexity for SGW.

Meshes comparison   In the context of computer graphics, GW can be used to quantify the correspondances between two meshes. A direct interest is found in shape retrieval, search, exploration or organization of databases. In order to recover experimentally some of the desired properties of the GW distance, we reproduce an experiment originally conducted in [Rustamov 2013] and presented in [Solomon 2016] with the use of entropic-GW.

From a given time series of 45 meshes representing a galloping horse, the goal is to conduct a multi-dimensional scaling (MDS) of the pairwise distances, computed with SGW between the meshes, that allows ploting each mesh as a 2D point. As one can observe in Figure 4.4, the cyclical nature of this motion is recovered in this 2D plot, as already illustrated in [Solomon 2016] with the GW distance. Each horse mesh is composed of approximately $9,000$ vertices. The average time for computing one distance is around 30 minutes using the POT implementation, which makes the computation of the full pairwise distance matrix impractical (as already mentioned in [Solomon 2016]). In contrast, our method only requires 25 minutes to compute the full distance matrix, with an average of 1.5s per mesh pair, using our CPU implementation. This clearly highlights the benefits of our method in this case.
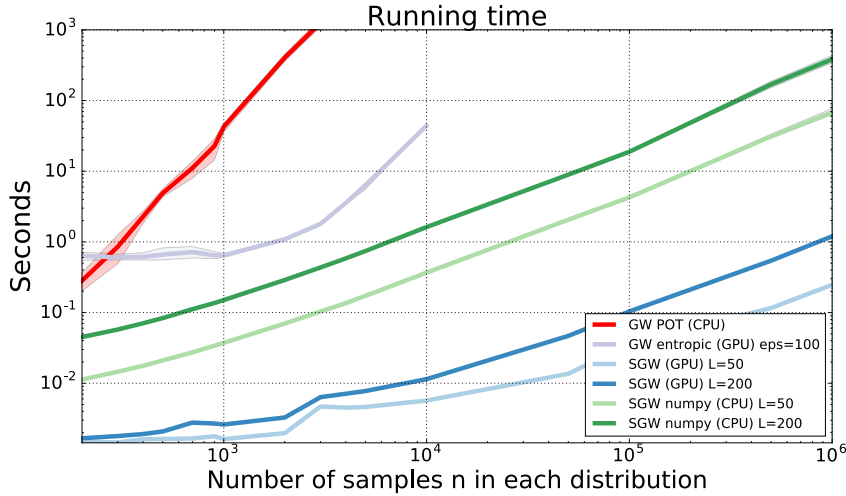
Figure 4.3: Runtimes comparison between SGW, GW, entropic-GW between two 2D random distributions with varying number of points from 0 to 10^6 in log-log scale. The time includes the calculation of the pair-to-pair distances.

SGW as a generative adversarial network (GAN) loss In a recent paper [Bunne 2019], Bunne and colleagues propose a new variant of GAN between incomparable spaces, *i.e.* of different dimensions. In contrast with classical divergences such as Wasserstein, they suggest to capture the intrinsic relations between the samples of the target probability distribution by using GW as a loss for learning. More formally, this translates into the following optimization problem over a desired generator G:

$$G^* = \arg\min GW_2^2(c_X, c_{G(Z)}, \mu, \nu_G), \tag{4.8}$$

where Z is a random noise following a prescribed low-dimensional distribution (typically Gaussian), $G(Z)$ performs the uplifting of Z in the desired dimensional space, and $c_{G(Z)}$ is the corresponding metric. μ and ν_G correspond respectively to the target and generated distributions, that we might want to align in the sense of GW. Following the same idea, and the fact that sliced variants of the Wasserstein distance have been successfully used in the context of GAN [Deshpande 2018], we propose to use SGW instead of GW as a loss for learning G. As a proof of concept, we reproduce the simple toy example of [Bunne 2019]. Those examples consist in generating 2D or 3D distributions from target distributions either in 2D or 3D spaces (Figure 4.5 and Figure 4.6). These distributions are formed by $3,000$ samples. We do not use their adversarial metric learning as it might confuse the objectives of this experiment and as it is not required for these low dimensional problems [Bunne 2019]. The generator G is designed as a simple multilayer perceptron with 2 hidden layers of respectively 256 and 128 units with ReLu activation functions, and one final layer with 2 or 3 output neurons (with linear activation) as output, depending on the experiment. The Adam optimizer is used, with a learning rate of 2.10^{-4} and $\beta_1 = 0.5, \beta_2 = 0.99$. The convergence to a visually acceptable solution takes a few hundred epochs. Contrary to [Bunne 2019], we directly back-propagate through our loss, without having to explicit a coupling matrix and resorting to the envelope Theorem. Compared to [Bunne 2019] and the use of entropic-GW , the time per epoch is more than one order of magnitude faster, as expected from previous experiment.
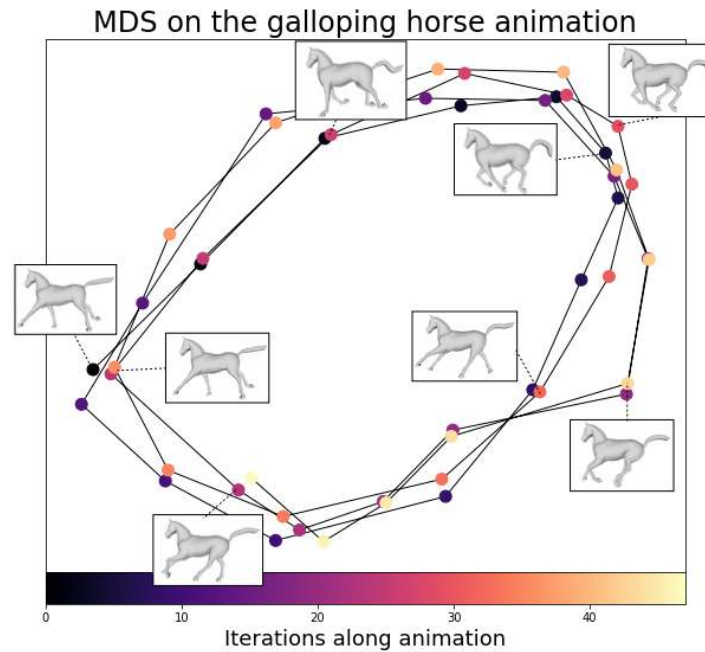
Figure 4.4: Each sample in this Figure corresponds to a mesh and is colored by the corresponding time iteration. One can see that the cyclical nature of the motion is recovered.
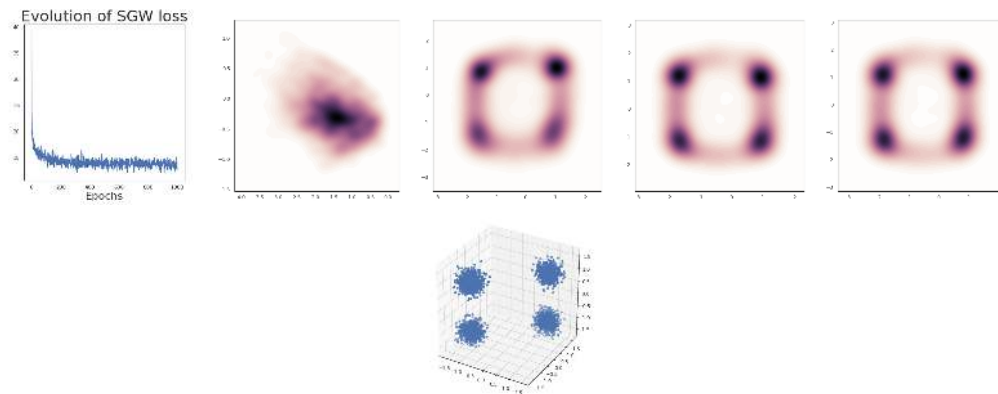
Figure 4.5: Using SGW in a GAN loss. First image shows the loss value along epochs. The next 4 images are produced by sampling the generated distribution ($3,000$ samples, plotted as a continuous density map). Last image shows the target 3D distribution.

4.1.4 Discussion and conclusion

In this section we establish a new result about Quadratic Assignment Problem when matrices are squared euclidean distances on the real line, and use it to state a closed-form expression for GW between monodimensional measures. Building upon this result we define a new similarity measure, called the Sliced Gromov-Wasserstein and a variant Rotation-invariant SGW and prove that both conserve various properties of the GW distance while being cheaper to compute and applicable in a large-scale setting. Notably SGW can be computed in 1 second for distributions with 1 million samples each. This paves the way for novel promising machine learning applications of optimal transport between metric spaces.
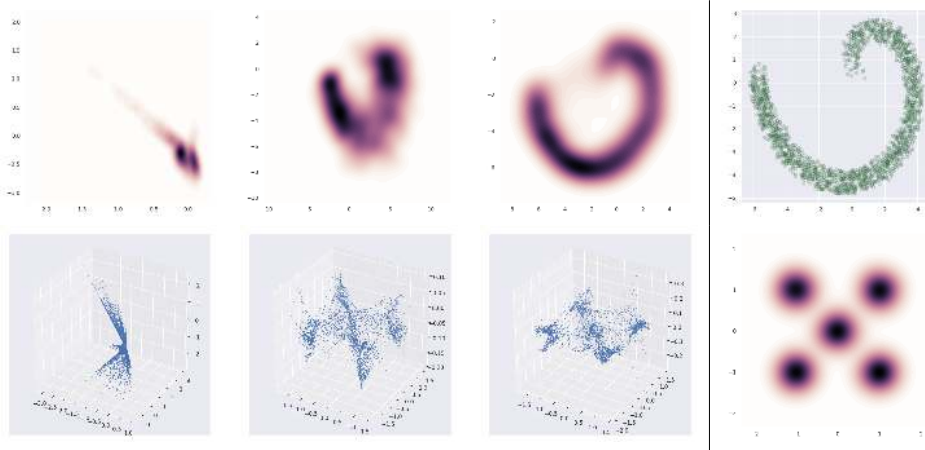
Figure 4.6: Using SGW in a GAN loss. The three rows depicts three different examples. First row is 2D (Generator) to 2D (Target), Second 3D to 2D. First column is initialization, second one is at 100 Epochs, third one at 1000. Last column depicts the target distribution.

Yet, several questions are raised in this work. Notably, our method perfectly fits the case when the two distributions are given empirically through samples embedded in an Hilbertian space, that allows for projection on the real line. This is the case in most of the machine learning applications that use the Gromov-Wasserstein distance. However, when only distances between samples are available, the projection operation can not be carried anymore, while the computation of GW is still possible. One can argue that it is possible to embed either isometrically those distances into a Hilbertian space, or at least with a low distortion, and then apply the presented technique. Our future line of work considers this option, as well as a possible direct reasoning on the distance matrix. For example, one should be able to consider geodesic paths (in a graph for instance) as the equivalent appropriate geometric object related to the line. This constitutes the direct follow-up of this work, as well as a better understanding of the accuracy of the estimated discrepancy with respect to the ambiant dimension and the projections number.

4.2 Regularity & formulations of GW problems in Euclidean spaces

4.2.1 Introduction

In the previous part we built upon the special case of 1D probability discrete measures. We consider in this section general probability measures $\mu \in \mathcal{P}(\mathbb{R}^p)$ and $\nu \in \mathcal{P}(\mathbb{R}^q)$ supported on Euclidean spaces $\mathbb{R}^p$ and $\mathbb{R}^q$ with (possibly) $p \neq q$. The corresponding inner products are denoted by $\langle \mathbf{x}, \mathbf{x}' \rangle_p$ (*resp.* $\langle \mathbf{y}, \mathbf{y}' \rangle_q$) for vectors in $\mathbb{R}^p$ (*resp.* $\mathbb{R}^q$) associated with the Euclidean norms which are denoted both by $\|.\|_2$ to avoid overloading notations.

We tackle in this section the problem of the *regularity* of the optimal transport plans of GW_2 in the Euclidean setting. More precisely we consider the following problem:

Problem 1. *Let $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$. Can we find an* deterministic *transport map to the GW_2 problem? More precisely does the following statement hold?*

$\exists T : \mathbb{R}^p \rightarrow \mathbb{R}^q$ such that $T \# \mu = \nu$ and $\gamma_T = (id \times T) \# \mu$ is optimal for the GW_2 problem:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int \int |c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') - c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}')|^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \tag{4.9}$$

The Euclidean setting is motivated by the linear transportation theory where the first regularity solution for Optimal Transport was proved by Brenier for probability measures in $\mathbb{R}^p$ (see Chapter 2). In this case we recall that the optimal transport plan γ_T between two probability measure $(\mu, \nu) \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$, with μ "well behaved" and with cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$, is unique and supported by a map T such that $\gamma_T = (id \times T)\#\mu$. The purpose of this section is to show that the Euclidean setting is also quite suited for the GW case. The problem of regularity of GW optimal transport plans was first addressed by Sturm in his seminal work about GW [Sturm 2012, Challenges 3.6]. More precisely Sturm asks the following question: are there some "nice" spaces in which we are able to prove Brenier's like results for GW? We give in this section some partial answers to this query by considering the two cases where $c_\mathcal{X}, c_\mathcal{Y}$ are defined by the inner products or by the squared Euclidean distances in each space.

The main results of this section is to derive equivalent formulations of the GW in these two cases. More precisely we show that solving GW is equivalent to jointly solve a linear transportation problem and a "alignment" problem. As such the regularity of GW optimal plans can be observed in the light of these "dual problems". As another consequence it allows also to derive algorithmic solutions for the GW problems in Euclidean spaces based on simple Block Coordinate Descent procedures. This section is organized as follow:

(i) In Section 4.2.2 we consider the case where $c_\mathcal{X}, c_\mathcal{Y}$ are defined by the inner products in each space. Providing that the source probability measure is regular with respect to the Lebesgue measure we give a sufficient condition for the existence of a *deterministic* optimal transport plan, *i.e.* supported on a deterministic function T. We show that this function is of the form $\nabla u \circ \mathbf{P}$ where u is a convex function and $\mathbf{P}$ is a linear application which can be seen as a global transformation "realigning" the probability measures in the same space. We use this formulation to show that the GW distance between 1D probability measures admits a closed-form solution. More precisely we show that the optimal coupling is determined by the cumulative and the anti-cumulative distribution functions of the source distribution. We further discuss the difference between the linear OT problem W_2 and the GW problem when the target measure is a perturbed version of the source measure.

(ii) In Section 4.2.3 we consider $c_\mathcal{X}, c_\mathcal{Y}$ as the squared Euclidean distances in each space. We show that this setting is equivalent to a maximization of a convex function on $\Pi(\mu, \nu)$. We use the Fenchel-Legendre duality in the space of measures to derive a problem equivalent to that of Gromov-Wasserstein. We further analyse it and show that the regularity of optimal transport plans is more complicated to state than in the previous case.

(iii) In Section 4.2.4 we use the previous formulations to derive efficient numerical solutions for the GW problem based on Block Coordinate Descent. We show that these procedures compare favourably with respect to standard solvers such as Conditional Gradient (see Chapter 3) or with entropic regularization (see Chapter 2).

(iv) We conclude in Section 4.2.5 by considering the *Gromov-Monge* problem in Euclidean spaces, which is the exact counterpart of the Monge problem of linear transportation but in the Gromov-Wasserstein context. We discuss the special case of the Gromov-Monge between Gaussian measures and we show that this problem admits a closed-form solution when restricting to *linear push-forward*. We give geometric interpretations of this result and compare the optimal push-forward with the standard optimal map of linear OT theory in the case of Gaussian measures (see Chapter 2).

This section is more prospective and somehow opens more doors than it closes. We hope that it will paves the path for further interesting works on this topic and believe that it could help for bridging the

gap between the understanding of the linear OT problem and the Gromov-Wasserstein theory. We have chosen to include in the main text some proofs that are reasonably long and that we consider interesting for the overall understanding of the section. The other proofs, which require more space, are postponed to Section 6.2.

4.2.2 The inner product case

To encompass the two cases described in the introduction we consider the following lemma (a proof can be found in Section 6.2.6):

Lemma 4.2.1. *For a coupling $\pi \in \Pi(\mu, \nu)$ we note:*

$$J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi) \stackrel{def}{=} \int_{\mathcal{X} \times \mathcal{X}} \int_{\mathcal{Y} \times \mathcal{Y}} |c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') - c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}')|^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \tag{4.10}$$

the GW_2 loss. Suppose that there exist scalars a, b, c such that $c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = a\|\mathbf{x}\|_2^2 + b\|\mathbf{x}'\|_2^2 + c\langle\mathbf{x}, \mathbf{x}'\rangle_p$ and $c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = a\|\mathbf{y}\|_2^2 + b\|\mathbf{y}'\|_2^2 + c\langle\mathbf{y}, \mathbf{y}'\rangle_q$. Then:

$$J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi) = C_{\mu,\nu} - 2Z(\pi) \tag{4.11}$$

where $C_{\mu,\nu} = \int c_{\mathcal{X}}^2 \mathrm{d}\mu\mathrm{d}\mu + \int c_{\mathcal{Y}}^2 \mathrm{d}\nu\mathrm{d}\nu - 4ab \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 d\mu(\mathbf{x})d\nu(\mathbf{y})$ and:

$$Z(\pi) = (a^2 + b^2) \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + c^2 \| \int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}^2$$
$$+ (a + b)c \int \left[\|\mathbf{x}\|_2^2 \langle \mathbb{E}_{Y \sim \nu}[Y], \mathbf{y}\rangle_q + \|\mathbf{y}\|_2^2 \langle \mathbb{E}_{X \sim \mu}[X], \mathbf{x}\rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \right] \tag{4.12}$$

In this section we study the *GW* problem with $c_{\mathcal{X}} = \langle\mathbf{x}, \mathbf{x}'\rangle_p$ and $c_{\mathcal{Y}} = \langle\mathbf{y}, \mathbf{y}'\rangle_q$. This corresponds to $a, b = 0$ and $c = 1$ case of Lemma 4.2.1. With a small abuse of notation we will denote by $\mathbf{P} \in \mathbb{R}^{q \times p}$ both the linear application $\mathbf{P} : \mathbb{R}^p \to \mathbb{R}^q$ and its associated matrix. Moreover we will often make no distinction between a vector $\mathbf{x} \in \mathbb{R}^p$ and the matrix associated to $\mathbf{x} \in \mathbb{R}^{p \times 1}$ such that $\mathbf{x}^T \in \mathbb{R}^{1 \times p}$. The next theorem gives a equivalent formulation of GW_2 in this context:

Theorem 4.2.1 (Equivalence of GW for the inner product case)**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ with $\int \|\mathbf{x}\|_2^4 \mathrm{d}\mu(x) < +\infty, \int \|\mathbf{y}\|_2^4 \mathrm{d}\nu(\mathbf{y}) < +\infty$. Suppose without loss of generality that $p \geq q$ and let:*

$$F_{p,q} \stackrel{def}{=} \{\mathbf{P} \in \mathbb{R}^{q \times p} |\ \|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}\} \tag{4.13}$$

Then problems:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int \int (\langle\mathbf{x}, \mathbf{x}'\rangle_p - \langle\mathbf{y}, \mathbf{y}'\rangle_q)^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \qquad \text{(innerGW)}$$

$$\sup_{\pi \in \Pi(\mu, \nu)} \sup_{\mathbf{P} \in F_{p,q}} \int \langle\mathbf{P}\mathbf{x}, \mathbf{y}\rangle_q \ \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \qquad \text{(MaxOT)}$$

are equivalent. In other words, $\pi^* \in \Pi(\mu, \nu)$ is an optimal solution of (innerGW) if and only if π^* is an optimal solution of (MaxOT).

Remark 4.2.1. *The condition $\int \|\mathbf{x}\|_2^4 \mathrm{d}\mu(\mathbf{x}) < +\infty, \int \|\mathbf{y}\|_2^4 \mathrm{d}\nu(\mathbf{y}) < +\infty$ suffices to prove that both (innerGW) and (MaxOT) are finite and that (MaxOT) admits an optimal solution $\pi^* \in \Pi(\mu, \nu)$ (we postponed this study to Lemma 6.2.7 in Section 6.2.7).*

This theorem gives another interesting formulation of the Gromov-Wasserstein problem. It proves that GW is equivalent to a linear OT problem combined with an "alignment" of the measures μ and ν on the same space using a linear application $\mathbf{P}$. The set $F_{p,q}$ can be regarded as the set of matrices with fixed Schatten ℓ_2 norms, that is $\mathbf{P} \in F_{p,q}$ if $\|\sigma(\mathbf{P})\|_2 = \sqrt{p}$ where $\sigma(\mathbf{P})$ is a vector containing the singular values of $\mathbf{P}$. When $p = q$ any orthogonal matrix $\mathbf{O} \in \mathcal{O}(p)$ is in $F_{p,q}$ since $\|\mathbf{O}\|_{\mathcal{F}} = \sqrt{\mathrm{tr}(\mathbf{O}^T\mathbf{O})} = \sqrt{\mathrm{tr}(\mathbf{I_p})} = \sqrt{p}$. More generally when $p > q$ any matrix in the Stiefel manifold $\Delta \in \mathbb{V}_p(\mathbb{R}^q)$ is an element of $F_{p,q}$ since $\Delta^T\Delta = \mathbf{I_p}$. Interestingly enough, the problem (MaxOT) can be related to the work of Alvarez and coauthors [Alvarez-Melis 2019] where they proposed a linear optimal transport problem which takes into account a latent global transformation of the measures. More precisely they consider two probability measures $(\mu, \nu) \in \mathcal{P}(\mathbb{R}^p) \times \mathbb{R}^p$ (*i.e.* $p = q$) and propose two minimize the following problem:

$$InvOT(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \min_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \int \|\mathbf{Px} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) \tag{4.14}$$

If we note $\boldsymbol{\Sigma}_\mu = \int \mathbf{xx}^T \mathrm{d}\mu(\mathbf{x})$ and we suppose that $\boldsymbol{\Sigma}_\mu = \mathbf{I_p}$ (which is called the μ-whitened property in [Alvarez-Melis 2019]) then problem (4.14) is equivalent to (MaxOT). To see this is suffices to develop $\int \|\mathbf{Px} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y})$ as:

$$\int \|\mathbf{Px} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y}) = \int \|\mathbf{Px}\|_2^2 \mathrm{d}\mu(\mathbf{x}) + \int \|\mathbf{y}\|_2^2 \mathrm{d}\nu(\mathbf{y}) - 2\int \langle \mathbf{Px}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x},\mathbf{y}) \tag{4.15}$$

Then we can check that $\int \|\mathbf{Px}\|_2^2 \mathrm{d}\mu(\mathbf{x})$ does not depend on $\mathbf{P}$ since:

$$\int \|\mathbf{Px}\|_2^2 \mathrm{d}\mu(\mathbf{x}) = \int \mathbf{x}^T\mathbf{P}^T\mathbf{Px} \mathrm{d}\mu(\mathbf{x}) \stackrel{*}{=} \int \mathrm{tr}(\mathbf{x}^T\mathbf{P}^T\mathbf{Px}) \mathrm{d}\mu(\mathbf{x}) \stackrel{**}{=} \int \mathrm{tr}(\mathbf{P}^T\mathbf{Pxx}^T) \mathrm{d}\mu(\mathbf{x})$$
$$\stackrel{***}{=} \mathrm{tr}(\mathbf{P}^T\mathbf{P} \int \mathbf{xx}^T \mathrm{d}\mu(\mathbf{x})) = \mathrm{tr}(\mathbf{P}^T\mathbf{P}) = \|\mathbf{P}\|_{\mathcal{F}}^2 = p \tag{4.16}$$

where in (*) we used $\mathbf{x}^T\mathbf{P}^T\mathbf{Px} \in \mathbb{R}$, in (**) that the trace is invariant by cyclical permutation and in (***) the linearity of the trace. Finally we used that $\mathrm{tr}(\mathbf{P}^T\mathbf{P}) = \|\mathbf{P}\|_{\mathcal{F}}^2 = p$ by hypothesis. However note that in general both problems may differ since the μ-whitened property does not hold in general.

Theorem 4.2.1 is based on the following generalization of the Frobenius norm duality to the continuous setting:

Lemma 4.2.2. *For any* $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ *and* $\pi \in \Pi(\mu,\nu)$. *Then:*

$$\sup_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \int \langle \mathbf{Px}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x},\mathbf{y}) = \sqrt{p} \| \int \mathbf{yx}^T \mathrm{d}\pi(\mathbf{x},\mathbf{y}) \|_{\mathcal{F}} \tag{4.17}$$

This supremum is achieved for $\mathbf{P}^* = \frac{\sqrt{p}}{\|\int \mathbf{yx}^T \mathrm{d}\pi(\mathbf{x},\mathbf{y})\|_{\mathcal{F}}} \int \mathbf{yx}^T \mathrm{d}\pi(\mathbf{x},\mathbf{y})$

Proof. We have:

$$\int \langle \mathbf{Px}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x},\mathbf{y}) = \int \mathbf{y}^T\mathbf{Px} \mathrm{d}\pi(\mathbf{x},\mathbf{y}) \stackrel{*}{=} \int \mathrm{tr}(\mathbf{y}^T\mathbf{Px}) \mathrm{d}\pi(\mathbf{x},\mathbf{y}) \stackrel{**}{=} \int \mathrm{tr}(\mathbf{Pxy}^T) \mathrm{d}\pi(\mathbf{x},\mathbf{y})$$
$$\stackrel{***}{=} \mathrm{tr}(\mathbf{P} \int \mathbf{xy}^T \mathrm{d}\pi(\mathbf{x},\mathbf{y})) = \langle \int \mathbf{yx}^T \mathrm{d}\pi(\mathbf{x},\mathbf{y}), \mathbf{P} \rangle_{\mathcal{F}}$$

where in (*) we used that $\mathbf{y}^T\mathbf{Px} \in \mathbb{R}$, in (**) we used the cyclical permutation invariance of the trace and in (***) its linearity. Hence $\sup_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \int \langle \mathbf{Px}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x},\mathbf{y}) = \sup_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \langle \mathbf{P}, \int \mathbf{yx}^T \mathrm{d}\pi(\mathbf{x},\mathbf{y}) \rangle_{\mathcal{F}}$. We note $\mathbf{V}_\pi = \int \mathbf{yx}^T \mathrm{d}\pi(\mathbf{x},\mathbf{y})$. We want to solve:

$$\sup_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \langle \mathbf{P}, \mathbf{V}_\pi \rangle_{\mathcal{F}} \tag{4.18}$$

Let $\mathbf{P}$ such that $\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}$. Then by Cauchy-Schwartz (see Memo 4.2.1) $\langle \mathbf{P}, \mathbf{V}_\pi \rangle_{\mathcal{F}} \leq \|\mathbf{P}\|_{\mathcal{F}} \|\mathbf{V}_\pi\|_{\mathcal{F}} = \sqrt{p}\|\mathbf{V}_\pi\|_{\mathcal{F}}$. Hence, $\sup_{\|\mathbf{P}\|_F = \sqrt{p}} \langle \mathbf{P}, \mathbf{V}_\pi \rangle_{\mathcal{F}} \leq \sqrt{p}\|\mathbf{V}_\pi\|_{\mathcal{F}}$.

Conversely, take $\mathbf{P}^* = \sqrt{p}\frac{\mathbf{V}_\pi}{\|\mathbf{V}_\pi\|_{\mathcal{F}}}$. Then $\|\mathbf{P}^*\|_{\mathcal{F}} = \sqrt{p}$ so $\sup_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \langle \mathbf{P}, \mathbf{V}_\pi \rangle_{\mathcal{F}} \geq \langle \mathbf{P}^*, \mathbf{V}_\pi \rangle_{\mathcal{F}} = \langle \sqrt{p}\frac{\mathbf{V}_\pi}{\|\mathbf{V}_\pi\|_F}, \mathbf{V}_\pi \rangle_{\mathcal{F}} = \sqrt{p}\|\mathbf{V}_\pi\|_{\mathcal{F}}$ which concludes the proof.

$\square$

Combining Lemma 4.2.1 and Lemma 4.2.2 actually proves Theorem 4.2.1. Indeed using Lemma 4.2.1 we see that (innerGW) is equivalent to maximizing $Z(\pi) = \|\int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}^2$ over the couplings $\pi \in \Pi(\mu, \nu)$ since the other terms are constant. In this way it is equivalent to maximize $\|\int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}$ which is equivalent by Lemma 4.2.2 to maximize $\sup_{\mathbf{P} \in F_{p,q}} \int \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$ *w.r.t.* π.

Regularity of (innerGW) **OT plans**   Theorem 4.2.1 proves that it is equivalent to study the problem (MaxOT) for studying the regularity of *GW* optimal transport plans. Interestingly enough the problem (MaxOT) echoes the linear transportation problem $\sup_{\pi \in \Pi(\mu, \nu)} \int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$ when $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ which is widely studied in the literature an can be tackled using tools from convexity analysis such as the Legendre transform. The following result due to McCann is particularly useful in this case:

Theorem 4.2.2 ([McCann 1995]). *Let* $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^p)$. *Suppose that* μ *is absolutely continuous with respect to the Lebesgue measure, then there exists a convex function* $u : \mathbb{R}^p \to \mathbb{R}$ *whose gradient* ∇u *pushes* μ *forward to* ν, *i.e.* $\nabla u \# \mu = \nu$. *Moreover* ∇u *is unique* μ *a.e.*

By noticing that, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p \times \mathbb{R}^p$, $\langle \mathbf{x}, \mathbf{y} \rangle_p \leq u^*(\mathbf{x}) + u(\mathbf{y})$ where u^* is the Legendre transform of the convex function u the result of McCann proves that the map ∇u defines an optimal coupling $\gamma = (id \times \nabla u)\#\mu$ for the problem $\sup_{\pi \in \Pi(\mu, \nu)} \int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$ between $(\mu, \nu) \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$. Indeed for any coupling $\pi \in \Pi(\mu, \nu)$:

$$
\begin{aligned}
\int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) &\leq \int u^*(\mathbf{x}) + u(\mathbf{y}) \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \\
&= \int u^*(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u(\mathbf{y}) \mathrm{d}\nu(\mathbf{y})
\end{aligned}
\tag{4.20}
$$

Using that ∇u pushes μ forward to ν implies:

$$
\begin{aligned}
\int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) &\leq \int u^*(\mathbf{x}) \mathrm{d}\mu(x) + \int u(\nabla u(\mathbf{x})) \mathrm{d}\mu(\mathbf{x}) \\
&= \int \langle \mathbf{x}, \nabla u(\mathbf{x}) \rangle_p \mathrm{d}\mu(\mathbf{x})
\end{aligned}
\tag{4.21}
$$

since for any convex function $u^*(\mathbf{x}) + u(\mathbf{x}) = \langle \mathbf{x}, \nabla u(\mathbf{x}) \rangle$. Overall $\int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \leq \int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\gamma(\mathbf{x}, \mathbf{y})$ for any coupling π which proves that γ is optimal. The idea is to use the same reasoning to find an optimal solution of (MaxOT). In order to invoke McCann's theorem we will need the regularity of the probability measure $\mathbf{P}\#\mu$ for $\mathbf{P} \in F_{p,q}$:

Memo 4.2.1 (Cauchy-Schwartz inequality). *Let* Ω *be a vector space associated with an inner product* $\langle, \rangle$ *which defines a norm* $\|.\|$ *through* $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ *for* $\mathbf{x} \in \Omega$. *The Cauchy-Schwartz inequality reads:*

$$
\forall \mathbf{x}, \mathbf{y} \in \Omega^2, \ |\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|
\tag{4.19}
$$

Proposition 4.2.1. *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$ regular with respect to the Lebesgue measure in $\mathbb{R}^p$ and a linear map $l : \mathbb{R}^p \to \mathbb{R}^q$ associated with a matrix $\mathbf{L} \in \mathbb{R}^{q \times p}$*

- *If $p < q$ (we go from lower to higher dimension), then $l\#\mu \in \mathbb{R}^q$ is not regular with respect to the Lebesgue measure on $\mathbb{R}^q$.*

- *If $p \geq q$ (we go from higher to lower dimension), then $l\#\mu \in \mathbb{R}^q$ is regular with respect to the Lebesgue measure on $\mathbb{R}^q$ if and only if the linear map l is surjective, that is $rank(\mathbf{L}) = q$.*

Proof. We have $l\#\mu(l(\mathbb{R}^p)) = \mu(l^{-1}(l(\mathbb{R}^p))) = \mu(\mathbb{R}^p) = 1$ so that $l\#\mu$ gives measure 1 to the image of l. However for the first point the image of l is a strict linear subspace of $\mathbb{R}^q$ and therefore has Lebesgue measure zero. Using Radon-Nykodym theorem this implies that $l\#\mu$ can not have a density with respect to the Lebesgue on $\mathbb{R}^q$. For the second point, suppose that l is surjective. Then rank($\mathbf{L}$) = q which implies that $\mathbf{L}\mathbf{L}^T$ is invertible so that $\det(\mathbf{L}\mathbf{L}^T) \neq 0$. We define $J = \sqrt{\det(\mathbf{L}\mathbf{L}^T)}$. Let g be the density of μ with respect to the Lebesgue measure in $\mathbb{R}^p$. Then by the coarea formula the density of $l\#\mu$ with respect to the Lebesgue measure on $\mathbb{R}^q$ is:

$$h(y) = \int_{l^{-1}(y)} \frac{g(\mathbf{x})}{J} dV_{l^{-1}(\mathbf{y})}(\mathbf{x}) \tag{4.22}$$

where denotes $dV_{l^{-1}(\mathbf{y})}(\mathbf{x})$ the volume element. Conversely if l is not surjective then rank($\mathbf{L}$) < q. Then the image of l is a strict linear subspace of $\mathbb{R}^q$ with Lebesgue measure zero and therefore $l\#\mu$ can not have a density with respect to the Lebesgue on $\mathbb{R}^q$. $\qquad\square$

In the light of previous results we can give the following sufficient condition so that (innerGW) problem admits an optimal transport plan supported on a deterministic function:

Theorem 4.2.3. *Let $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ with $\int \|\mathbf{x}\|_2^4 d\mu(\mathbf{x}) < +\infty, \int \|\mathbf{y}\|_2^4 d\nu(\mathbf{y}) < +\infty$. Suppose that $p \geq q$ and that μ is regular with respect to the Lebesgue measure in $\mathbb{R}^p$. Suppose that there exists $(\pi^*, \mathbf{P}^*)$ an optimal solution of (MaxOT) with $\mathbf{P}^*$ surjective.*

Then there exists $u : \mathbb{R}^q \to \mathbb{R}$ convex such that $\nabla u \circ \mathbf{P}^*$ pushes μ forward to ν. Moreover the coupling $\gamma = (id \times \nabla u \circ \mathbf{P}^*)\#\mu$ is optimal for (innerGW). In particular problem 1 holds.

Proof. Let $(\pi^*, \mathbf{P}^*)$ be maximizers of (MaxOT) with $\mathbf{P}^*$ surjective. Using Proposition 4.2.1 we know that $\mathbf{P}^*\#\mu$ is regular with respect to the Lebesgue measure on $\mathbb{R}^q$. Using Theorem 4.2.2 there exists $u : \mathbb{R}^q \to \mathbb{R}$ convex such that $\nabla u\#\mathbf{P}^*\#\mu = \nu$ or equivalently $\nabla u \circ \mathbf{P}^*$ pushes μ forward to ν. Moreover we have:

$$\int \langle \mathbf{P}^*\mathbf{x}, \mathbf{y}\rangle_d d\pi^*(\mathbf{x},\mathbf{y}) \overset{1}{\leq} \int (u(\mathbf{P}^*x) + u^*(\mathbf{y})) \, d\pi^*(\mathbf{x},\mathbf{y}) = \int u(\mathbf{P}^*\mathbf{x}) d\mu(\mathbf{x}) + \int u^*(\mathbf{y}) d\nu(\mathbf{y})$$

$$\overset{2}{=} \int u(\mathbf{P}^*\mathbf{x}) d\mu(\mathbf{x}) + \int u^*(\mathbf{y}) d(\nabla u \circ \mathbf{P}^*\#\mu)(\mathbf{y})$$

$$= \int u(\mathbf{P}^*\mathbf{x}) d\mu(x) + \int u^*(\nabla u(\mathbf{P}^*\mathbf{x})) d\mu(\mathbf{x})$$

$$\overset{3}{=} \int \langle \mathbf{P}^*\mathbf{x}, \nabla u(\mathbf{P}^*\mathbf{x})\rangle_q d\mu(\mathbf{x})$$

where in (1) we used that u is convex which implies $u(\mathbf{Px}) + u^*(\mathbf{y}) \geq \langle \mathbf{Px}, \mathbf{y}\rangle_q$ by Fenchel-Young inequality, in (2) we used $\nabla u \circ \mathbf{P}^*$ pushes μ forward to ν, in (3) we used that for any x and convex function $u^*(\mathbf{x}) + u(\mathbf{x}) = \langle \mathbf{x}, \nabla u(\mathbf{x})\rangle$.

If we define $T = \nabla u \circ \mathbf{P}^*$ and $\gamma = (id \times T)\#\mu \in \Pi(\mu, \nu)$ we can deduce from (3) that:

$$\sup_{\pi \in \Pi(\mu, \nu)} \sup_{\mathbf{P} \in F_{p,q}} \int \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \leq \int \langle \mathbf{P}^*\mathbf{x}, y \rangle_q \mathrm{d}\gamma(\mathbf{x}, \mathbf{y})$$

By suboptimality the converse inequality is also true (since $\gamma = (id \times T)\#\mu \in \Pi(\mu, \nu)$). In this way we have:

$$\sup_{\pi \in \Pi(\mu, \nu)} \sup_{\mathbf{P} \in F_{p,q}} \int \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) = \int \langle \mathbf{P}^*\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\gamma(\mathbf{x}, \mathbf{y})$$

Overall the couple $(\gamma, \mathbf{P}^*)$ is optimal for the problem (MaxOT) and γ is optimal for (innerGW) using Theorem 4.2.1 which concludes the proof.

$\square$

The last result indicates that Monge map of the form $\nabla u \circ \mathbf{P}$ where $\mathbf{P}$ is a linear application may be of interest to study optimal solutions of (innerGW). When considering $p = q$ and $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ looking at such maps can generate the optimal Monge map ∇u of linear optimal transport problem $\sup_{\pi \in \Pi(\mu, \nu)} \int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$ when $\mathbf{P} = \mathbf{I_p}$. Note that a couple $(u, \mathbf{P})$ which satisfies both $\nabla u \circ \mathbf{P}$ pushes μ onto ν and $\gamma = (id \times \nabla u \circ \mathbf{P})$ is optimal for (innerGW) is not guaranteed to be unique. This difference should be put in perspective with the theory of linear transportation with ground cost $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$. Indeed in this context if one finds a mapping $T = \nabla u$ with any convex function u then Brenier's theorem states that this mapping is optimal since there is a unique Monge map satisfying this property. This unicity result is particularly interesting *e.g.* in order to prove that the linear transport between Gaussian measure admits a closed-form expression [Takatsu 2011]. However in our context the map $\nabla u \circ \mathbf{P}$ may fail to be unique as shown in the following example:

Example 4.2.1. *Consider $p = q$, a source measure $\mu \in \mathcal{P}(\mathbb{R}^p)$ and a target measure $\nu \in \mathcal{P}(\mathbb{R}^p)$ whose support is invariant by rotation, i.e. such that $\mathbf{O}\#\nu = \nu$ for any $\mathbf{O} \in \mathcal{O}(p)$. We can consider e.g. any isotropic Gaussian measure $\mathcal{N}(\mathbf{m}_\nu, \sigma^2 \mathbf{I_p})$. Since $\mathbf{O}$ preserves the angles then the problem (innerGW) is invariant by $\mathbf{O}$ which implies that any optimal map defined in Theorem 4.2.3 will fail to be unique. More precisely let suppose that $\nabla u \circ \mathbf{P}$ pushes μ forward to ν and that $\gamma = (id \times \nabla u \circ \mathbf{P})$ is optimal for (innerGW). Then for any $\mathbf{O} \in \mathcal{O}(p)$, $\mathbf{O}^T \circ \nabla u \circ \mathbf{P}$ also pushes μ forward to ν since $\nabla u \circ \mathbf{P}\#\mu = \nu = \mathbf{O}\#\nu$ by rotational invariance of the support of ν. This implies that $\mathbf{O}^T \circ \nabla u \circ \mathbf{P}\#\mu = \nu$. Moreover the map $\gamma' = (id \times \mathbf{O}^T \circ \nabla u \circ \mathbf{P})$ is also optimal with the same cost as γ. Indeed:*

$$\int \int \left(\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle \mathbf{y}, \mathbf{y}' \rangle_p \right)^2 \mathrm{d}\gamma'(\mathbf{x}, \mathbf{y}) \mathrm{d}\gamma'(\mathbf{x}', \mathbf{y}') = \int \int \left(\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle \mathbf{O}^T \nabla u(\mathbf{P}\mathbf{x}), \mathbf{O}^T \nabla u(\mathbf{P}\mathbf{x}') \rangle_p \right)^2 \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}')$$

$$\overset{*}{=} \int \int \left(\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle \nabla u(\mathbf{P}\mathbf{x}), \nabla u(\mathbf{P}\mathbf{x}') \rangle_p \right)^2 \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}')$$

$$= \int \int \left(\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle \mathbf{y}, \mathbf{y}' \rangle_p \right)^2 \mathrm{d}\gamma(\mathbf{x}, \mathbf{y}) \mathrm{d}\gamma(\mathbf{x}', \mathbf{y}')$$

where in (*) we used that $\mathbf{O} \in \mathcal{O}(p)$.

The condition of Theorem 4.2.3 seems reasonable and not too restrictive: it suffices that one optimal solution of (MaxOT) where $\mathbf{P}^*$ is surjective exists in order to prove that an optimal coupling of innerGW is supported by a deterministic map. We believe that some simple assumption can be made on μ, ν in order to meet this condition. In particular this is satisfied when μ, ν are 1D distributions as detailed below.

Application of Theorem 4.2.3 for 1D probability distributions   The sufficient condition given in Theorem 4.2.3 can be used to derived a closed-form expression for the problem (innerGW) between 1D probability distributions. Let us consider $(\mu, \nu) \in \mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ and F_μ and F_ν be the cumulative distribution functions of μ and ν and F_μ^{-1}, F_ν^{-1} its pseudo inverses (see Chapter 2). We suppose that μ is regular with respect to the Lebesgue measure in $\mathbb{R}$.

In this case the linear application $\mathbf{P}$ in (MaxOT) reduces to a scalar $p \in \mathbb{R}$:

$$\sup_{\pi \in \Pi(\mu,\nu)} \sup_{|p|=1} \int (px).y \, \mathrm{d}\pi(x,y) \tag{4.23}$$

In this way an optimal any optimal solution (π^*, p^*) of (4.23) satisfies $p^* \in \{-1, 1\}$ so that p defines a surjective linear application. Then by applying Theorem 4.2.3 there exists $u : \mathbb{R} \to \mathbb{R}$, convex such that $u' \circ p^*$ pushes μ forward to ν and that $\gamma = (id \times u' \circ p^*)\#\mu$ is optimal for (innerGW). In other words there exists $f = u'$ non-decreasing such that $\gamma = (id \times f \circ p^*)\#\mu$ is optimal for (innerGW). However we known from linear transport theory that there is a unique non-decreasing map $T_{asc} : \mathbb{R} \to \mathbb{R}$ such that $T_{asc}\#\mu = \nu$ and it is given by $T_{asc}(x) = F_\nu^{-1}(F_\mu(x))$ (see theorem 2.5 in [Santambrogio 2015]). This proves that if $p^* = 1$ then $f \circ p^* = T_{asc}$ so that $\gamma = (id \times T_{asc})\#\mu$ is optimal for (innerGW). If $p^* = -1$ then we have to consider the "anti" cumulative distribution function. Indeed in this case $f \circ p^*$ is non-increasing and pushes μ forward to ν which is equivalent to say that f is non-decreasing and pushes $\tilde{\mu}$ forward to ν where "$\mathrm{d}\tilde{\mu}(x) = \mathrm{d}\mu(-x)$". This discussion leads to the following result:

Theorem 4.2.4 (Closed Form expression for (innerGW) between 1D distributions). *Let $(\mu, \nu) \in \mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ with μ regular with respect to the Lebesgue measure. Let $F_\mu^{\nearrow}(x) = \mu(] - \infty, x])$ be the cumulative distribution function and $F_\mu^{\searrow}(x) = \mu(] - x, +\infty])$ be the anti-cumulative distribution function. Let $T_{asc} : \mathbb{R} \to \mathbb{R}$ defined by $T_{asc}(x) = F_\nu^{-1}(F_\mu^{\nearrow}(x))$ and $T_{desc} : \mathbb{R} \to \mathbb{R}$ defined by $T_{desc}(x) = F_\nu^{-1}(F_\mu^{\searrow}(x))$.*

Then an optimal solution of (innerGW) is achieved by the map $\gamma = (id \times T_{asc})\#\mu$ or by the map $\gamma = (id \times T_{desc})\#\mu$.

Theorem 4.2.4 proves that it suffices to compute the CDF or the anti-CDF of the distribution to recover an optimal coupling. It can be put in light of the results of Section 4.1 where we proved that for discrete probability measures with uniform weights and same number of atoms an optimal coupling for the GW problem with squared Euclidean distances can be found in the diagonal or the anti-diagonal coupling when samples are sorted. As such the previous theorem is stronger since it can be applied for general 1D probability distribution when considering inner product similarities. Can we find other examples of optimal couplings for (innerGW), for example when the dimension is larger than 1? The next discussion gives another example which answers this question.

Construction of an optimal couple $(u, \mathbf{P})$ As seen in the previous discussion the couples $(u, \mathbf{P})$ where u is convex so that $\nabla u \circ \mathbf{P}$ pushes μ forward to ν may lead to an optimal map *w.r.t.* the Gromov-Wasserstein distance with inner product similarities. In this part we wish to give an example of such couple $(u, \mathbf{P})$ which leads to an optimal map. Moreover this discussion will also highlight another difference between the linear OT problem and the Gromov-Wasserstein problem. It is known that when $\mu \in \mathcal{P}(\mathbb{R}^p)$ and when $u : \mathbb{R}^p \to \mathbb{R}$ is convex and differentiable μ *a.e* then the optimal transport plan for the linear OT problem $\inf_{\pi \in \Pi(\mu,\nu)} \int \|\mathbf{x} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$ between μ and $\nu \overset{def}{=} \nabla u \#\mu$ is given by $\gamma = (id \times \nabla u)\#\mu$. In other words if one perturbs the source measure μ with a transformation which is the gradient of

a convex function then the cheapest way (in terms of W_2) of moving the source measure forward to the target is by the means of this transformation. To see this we can do the same reasoning as in (4.20) by noticing that $\inf_{\pi \in \Pi(\mu,\nu)} \int \|\mathbf{x} - \mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$ is equivalent to $\sup_{\pi \in \Pi(\mu,\nu)} \int \langle \mathbf{x}, \mathbf{y} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$ (see also [Santambrogio 2015, Theorem 1.48]). In our case however the situation is a little bit different as detailed in the next proposition:

Proposition 4.2.2. *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$ with $\int \|\mathbf{x}\|_2^4 \mathrm{d}\mu(\mathbf{x}) < +\infty$. Let $u : \mathbb{R}^p \to \mathbb{R}$ be a convex function. We consider:*

$$\sup_{\mathbf{Q} \in F_{p,q}} \int u(\mathbf{Q}\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \tag{E_u}$$

Let $\mathbf{P} \in F_{p,q}$ be a solution to (E_u) with $\int \|\nabla u(\mathbf{P}\mathbf{x})\|_2^4 \mathrm{d}\mu(\mathbf{x}) < +\infty$ then $\gamma = (id \times \nabla u \circ \mathbf{P}) \# \mu$ is an optimal solution of (innerGW) between μ and $\nu \overset{def}{=} \nabla u \circ \mathbf{P} \# \mu$.

Proof. Let $(\pi^*, \mathbf{P}^*)$ be maximizers of (MaxOT). We have:

$$\int \langle \mathbf{P}^*\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y}) \overset{1}{\leq} \int (u(\mathbf{P}^*\mathbf{x}) + u^*(\mathbf{y})) \, \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y}) = \int u(\mathbf{P}^*\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\mathbf{y}) \mathrm{d}\nu(\mathbf{y})$$

$$\overset{2}{=} \int u(\mathbf{P}^*\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\mathbf{y}) \mathrm{d}(\nabla u \circ \mathbf{P} \# \mu)(\mathbf{y})$$

$$= \int u(\mathbf{P}^*\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\nabla u(\mathbf{P}\mathbf{x})) \mathrm{d}\mu(\mathbf{x})$$

$$\leq \sup_{\mathbf{Q} \in F_{p,q}} \int u(\mathbf{Q}\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\nabla u(\mathbf{P}\mathbf{x})) \mathrm{d}\mu(\mathbf{x})$$

$$\overset{3}{=} \int u(\mathbf{P}\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\nabla u(\mathbf{P}\mathbf{x})) \mathrm{d}\mu(\mathbf{x})$$

$$\overset{4}{=} \int \langle \mathbf{P}\mathbf{x}, \nabla u(\mathbf{P}\mathbf{x}) \rangle_q \mathrm{d}\mu(\mathbf{x})$$

where in (1) we used that u is convex, in (2) we used $(\nabla u \circ \mathbf{P}) \# \mu = \nu$, in(3) we used that $\mathbf{P}$ maximizes $\sup_{\mathbf{Q} \in F_{p,q}} \int u(\mathbf{Q}\mathbf{x}) \mathrm{d}\mu(\mathbf{x})$ and in (4) we used that for any x and convex function $u^*(\mathbf{x}) + u(\mathbf{x}) = \langle \mathbf{x}, \nabla u(\mathbf{x}) \rangle$.
We can deduce from (4) that:

$$\sup_{\pi \in \Pi(\mu,\nu)} \sup_{\mathbf{P} \in F_{p,q}} \int \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \leq \int \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\gamma(\mathbf{x}, \mathbf{y})$$

By suboptimality the converse inequality is also true so that $(\gamma, \mathbf{P})$ is an optimal solution of (MaxOT) and consequently γ is an optimal solution of (innerGW) using Theorem 4.2.1.

$\square$

This results states that if one perturbs a source measure $\mu \in \mathcal{P}(\mathbb{R}^p)$ with a map $\nabla u \circ \mathbf{P}$ with the condition that $\mathbf{P} : \mathbb{R}^p \to \mathbb{R}^q$ achieves $\sup_{\mathbf{Q} \in F_{p,q}} \int u(\mathbf{Q}\mathbf{x}) \mathrm{d}\mu(\mathbf{x})$ then the cheapest way *w.r.t.* (innerGW) of moving μ forward to $\nabla u \circ \mathbf{P} \# \mu$ is by the means of the transformation $\nabla u \circ \mathbf{P}$. To illustrate this proposition we can look at simple convex transformations $u(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{U}\mathbf{x}$ where $\mathbf{U}$ is symmetric positive semi-definite (which means that the transformation ∇u is linear). In this case the following proposition exhibits a couple $(u, \mathbf{P})$ which leads to an optimal map for (innerGW):

Proposition 4.2.3 (An optimal couple $(u, \mathbf{P})$). *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$. We note $\boldsymbol{\Sigma}_\mu = \int \mathbf{x}\mathbf{x}^T \mathrm{d}\mu(\mathbf{x})$. Let $u : \mathbb{R}^p \to \mathbb{R}$ convex defined by $u(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{U}\mathbf{x}$ where $\mathbf{U}$ is symmetric positive semi-definite. Let $\mathbf{v}, \mathbf{w}$ be the eigenvectors corresponding to the largest eigenvalue of respectively $\boldsymbol{\Sigma}_\mu$ and $\mathbf{U}$ and $\mathbf{P} = \sqrt{p} \, \mathbf{v}\mathbf{w}^T \in F_{p,p}$*

The coupling $\gamma = (id \times \nabla u \circ \mathbf{P})$ is optimal for (innerGW) between μ and $\nu \overset{def}{=} \nabla u \circ \mathbf{P} \# \mu$

Proof. In this case the problem $\sup_{\mathbf{Q} \in F_{p,p}} \int u(\mathbf{Q}\mathbf{x}) \mathrm{d}\mu(\mathbf{x})$ reduces to $\sup_{\|\mathbf{Q}\|_{\mathcal{F}}^2 = p} = \int \mathbf{x}^T \mathbf{Q}^T \mathbf{U} \mathbf{Q} \mathbf{x} \mathrm{d}\mu(\mathbf{x}) = \sup_{\|\mathbf{Q}\|_{\mathcal{F}}^2 = p} \mathrm{tr}(\mathbf{Q}^T \mathbf{U} \mathbf{Q} \mathbf{\Sigma}_\mu)$ where $\mathbf{\Sigma}_\mu = \int \mathbf{x}\mathbf{x}^T \mathrm{d}\mu(\mathbf{x})$. By vectorizing the matrix $\mathbf{Q}$ it is equivalent to:

$$\max_{\substack{\mathbf{q} \in \mathbb{R}^{p^2} \\ \|\mathbf{q}\|_2^2 = p}} \mathbf{q}^T \mathbf{M}_{u,\mu} \mathbf{q} \tag{4.24}$$

where $\mathbf{M}_u = \mathbf{U} \otimes_K \mathbf{\Sigma}_\mu$ with $\otimes_K$ the Kronecker product of matrices. $\mathbf{M}_{u,\mu}$ is symmetric positive semi-definite. We can rewrite this problem with $\tilde{\mathbf{q}} = \sqrt{p}\, \mathbf{q}$:

$$\max_{\tilde{\mathbf{q}}^T \tilde{\mathbf{q}} = 1} \tilde{\mathbf{q}}^T \mathbf{M}_{u,\mu} \tilde{\mathbf{q}} \tag{4.25}$$

Which is a maximization of a Rayleigh quotient problem. It is well known that a solution of this problem is found at any eigenvector associated to the largest eigenvalue of $\mathbf{M}_{u,\mu}$ (see *e.g.* [Anstreicher 1998]). However the eigenvalues of $\mathbf{M}_{u,\mu}$ are given by all the products of the eigenvalues of $\mathbf{\Sigma}_\mu$ and $\mathbf{U}$ (see *e.g.* [Horn 1991]). Since they are all positive the largest eigenvalue of $\mathbf{M}_{u,\mu}$ is found at the largest eigenvalue of $\mathbf{\Sigma}_\mu$ and $\mathbf{U}$ with corresponding eigenvector $\mathbf{v}, \mathbf{w}$ and the optimal $\tilde{\mathbf{q}}$ is $\mathbf{v}, \mathbf{w}^T$. This implies that the optimal $\mathbf{Q}$ is $\mathbf{Q} = \sqrt{p}\, \mathbf{v}\mathbf{w}^T$. We can apply Proposition 4.2.2 to conclude. $\qquad\square$

4.2.3 The squared Euclidean case

The *GW* problem is usually considered with distances, as it provides a metric with respect to strong isomorphisms. The goal of this section is to study the case $c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2^2, c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_2^2$ with $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$. As for the inner product case we can prove that this problem is equivalent to another linear OT problem parametrized by a linear application. More precisely:

Theorem 4.2.5. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact subset of respectively $\mathbb{R}^p$ and $\mathbb{R}^q$. Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. Assume without loss of generality that $\mathbb{E}_{X \sim \mu}[X] = 0$ and $\mathbb{E}_{Y \sim \nu}[Y] = 0$. Then problems:*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int (\|\mathbf{x} - \mathbf{x}'\|_2^2 - \|\mathbf{y} - \mathbf{y}'\|_2^2)^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \tag{sqGW}$$

and

$$\sup_{\pi \in \Pi(\mu,\nu)} \sup_{\mathbf{P} \in \mathbb{R}^{q \times p}} \int (\langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2) \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) - \frac{1}{8}\|\mathbf{P}\|_{\mathcal{F}}^2 \tag{dual-sqGW}$$

are equivalent.

To prove the previous theorem we will rely on the calculus of Lemma 4.2.1 and the observation that the cost $J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi)$ is invariant by translation of the support of the measures so that they can be centered without loss of generaly. This implies that the term $\int \left[\|\mathbf{x}\|_2^2 \langle \mathbb{E}_{Y \sim \nu}[Y], \mathbf{y} \rangle_q + \|\mathbf{y}\|_2^2 \langle \mathbb{E}_{X \sim \mu}[X], \mathbf{x} \rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \right]$ in Lemma 4.2.1 vanishes. More presicely we have the following result:

Lemma 4.2.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact subset of respectively $\mathbb{R}^p$ and $\mathbb{R}^q$. Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. We can assume without loss of generality that $\mathbb{E}_{X \sim \mu}[X] = 0$ and $\mathbb{E}_{Y \sim \nu}[Y] = 0$. In this case (sqGW) is equivalent to:*

$$\sup_{\pi \in \Pi(\mu,\nu)} \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + 2\| \int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \|_{\mathcal{F}}^2 \tag{4.26}$$

A proof of this lemma can be found in Section 6.2.8. In the following we will note $F(\pi) = \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + 2\| \int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \|_{\mathcal{F}}^2$. To prove Theorem 4.2.5 the idea is to observe that this problem is a maximization of a convex function of π. We can use standard convex analysis tools such that the Fenchel-Moreau duality to derive the Fenchel dual of $F(\pi)$ as detailed below.

Duality in the space of measure We suppose in the following that $\mathcal{X}, \mathcal{Y}$ are general compact spaces. In this case the dual space of $C(\mathcal{X} \times \mathcal{Y})$ is $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ (see *e.g.* Memo 1.3 in [Santambrogio 2015]). We recall the main definitions of the Legendre-Fenchel tansform:

Definition 4.2.1 (Legendre–Fenchel Transformation). *Let $\mathcal{X}, \mathcal{Y}$ be compact Hausdorff spaces. For a function $F : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ we define its convex conjugate $F^* : C(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ and its bi-conjugate $F^{**} : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ as:*

$$F^*(h) = \sup_{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})} \int h(x, y) \mathrm{d}\pi(x, y) - F(\pi)$$

$$F^{**}(\pi) = \sup_{h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \int h(x, y) \mathrm{d}\pi(x, y) - F^*(h) \tag{4.27}$$

One remarkable property of the convex conjugate F^* is that it is always l.s.c and convex because $h \to \int h \mathrm{d}\pi - F(\pi)$ is an affine function which is always convex and continuous. In this way F^* is the pointwise supremum of continuous linear functions which is convex and l.s.c. We denote by $\mathrm{dom}(F) = \{\pi | F(\pi) < +\infty\}$ the domain of F. A function is called *proper* if $\mathrm{dom}(F) \neq \emptyset$. A fundamental result in convex analysis is the Fenchel-Moreau theorem states that the bi-conjugate of a proper and l.s.c convex function equals to the original function (see *e.g.* [Lai 1988]). In other words when F is convex and well-behaved one can rely on the bi-conjugate to study the original function. In our context this implies that $F(\pi) = F^{**}(\pi)$ for all $\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$. To compute $F^{**}(\pi)$ we will need a notion of derivative in the space of measures as described in the next definition:

Definition 4.2.2 (Fréchet differentiable). *Let $\mathcal{X}, \mathcal{Y}$ be compact Hausdorff spaces. A function $F : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is is Fréchet differentiable at π if there exists $\nabla F(\pi) \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$ such that for any $\varepsilon \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ as $t \to 0$:*

$$F(\pi + t\varepsilon) = F(\pi) + t \int \nabla F(\pi) d\varepsilon + o(t) \tag{4.28}$$

Application to the Gromov-Wasserstein problem Let $\mathcal{X} \subset \mathbb{R}^p, \mathcal{Y} \subset \mathbb{R}^q$ be compact spaces and $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. Since $\mathcal{X}, \mathcal{Y}$ are compact and the norms are continuous the GW distance is finite and the function F defined in (4.26) is proper, convex and l.s.c. By application of the Fenchel-Moreau theorem we have:

$$\sup_{\pi \in \Pi(\mu, \nu)} F(\pi) = \sup_{\pi \in \Pi(\mu, \nu)} F^{**}(\pi) = \sup_{\pi \in \Pi(\mu, \nu)} \sup_{h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \int h(\mathbf{x}, \mathbf{y}) \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) - F^*(h) \tag{4.29}$$

In the following we denote by $\mathbf{V}_\pi = \int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y})$. We can prove that we can solve the dual problem by parametrizing h by a linear application $\mathbf{P} \in \mathbb{R}^{q \times p}$ as $h(\mathbf{x}, \mathbf{y}) = \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$. Using this form the problem becomes much simpler as we only need to optimize on a finite dimensional space instead of maximizing over all continuous function.

Lemma 4.2.4. *If π^* is a solution of the primal problem $\sup_{\pi \in \Pi(\mu, \nu)} F(\pi)$ then there exists $\mathbf{P} \in \mathbb{R}^{q \times p}$ and $h^* \in C(\mathcal{X} \times \mathcal{Y})$ of the form $h^*(\mathbf{x}, \mathbf{y}) = \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$ such that (π^*, h^*) is a solution of the dual problem (4.26). Moreover when h^* is in such form we have $F^*(h^*) = \frac{1}{8} \|\mathbf{P}\|_{\mathcal{F}}^2$.*

A proof can be found in Section 6.2.9. The previous Lemma 4.2.4 can be used to prove Theorem 4.2.5. Indeed with previous notations h can be written in the form $h(\mathbf{x}, \mathbf{y}) = \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$. Plugging the calculus of the conjugate $F^*(h) = \frac{1}{8} \|\mathbf{P}\|_{\mathcal{F}}^2$ into (4.29) gives the desired result.

Regularity of (sqGW) **optimal plans** The Fenchel dual problem is more difficult to analyse than the dual problem of the inner product case. Indeed the cost $c : (\mathbf{x}, \mathbf{y}) \to \langle \mathbf{Px}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$ hardly relates to a "standard" linear OT problem. Especially it does not statifies the Twist condition (see Chapter 2) in general and an approach with convex tools is trickier due to the term $\|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$. We will show in the following that if $\nabla u \circ \mathbf{P}$ is "not too far" from an isometry then it defines an optimal coupling. We note $\Gamma_0(\mathbb{R}, \mathbb{R})$ the set of derivable convex functions from $\mathbb{R}$ to $\mathbb{R}$ and we define the following set:

Definition 4.2.3. *We define $\mathcal{H}(\mu, \nu)$ the set of push-forward between μ and ν defined by:*

$$\mathcal{H}(\mu, \nu) = \{T \# \mu = \nu \, | \exists f \in \Gamma_0(\mathbb{R}, \mathbb{R}) \, , \|T(\mathbf{x})\|_2^2 = f'(\|\mathbf{x}\|_2^2) \, \mu \, a.e\} \tag{4.30}$$

When $p = q$ this set encompasses all the linear push forward of the form $T = c\,\mathbf{O}$ where $c > 0, \mathbf{O} \in \mathcal{O}(p)$. It turns out that this condition is also sufficient when considering only linear push-forward (see Lemma 6.2.11).

We have the following sufficient condition if we look at Monge map of the form $\nabla u \circ \mathbf{P}$ with u convex and and $\mathbf{P} \in \mathbb{R}^{q \times p}$:

Proposition 4.2.4. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact subset of respectively $\mathbb{R}^p$ and $\mathbb{R}^q$ with $p \geq q$. Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. Assume that μ is regular w.r.t. the Lebesgue measure on $\mathbb{R}^p$ and that $\mathbb{E}_{X \sim \mu}[X] = 0$ and $\mathbb{E}_{Y \sim \nu}[Y] = 0$ without loss of generality.*

Let $(\pi^*, \mathbf{P}^*)$ be optimal solution of (dual-sqGW). If $\mathbf{P}^*$ is surjective there exists $u : \mathbb{R}^q \to \mathbb{R}$ convex such that $\nabla u \circ \mathbf{P}^*$ pushes μ forward to ν. Moreover if $\nabla u \circ \mathbf{P}^* \in \mathcal{H}(\mu, \nu)$ then the coupling $\gamma = (id \times \nabla u \circ \mathbf{P}^*) \# \mu$ is optimal for (sqGW).

Proof. Let $(\mathbf{P}^*, \pi^*)$ be optimal solution of (dual-sqGW) with $\mathbf{P}^*$ surjective. We have seen previously that $\mathbf{P}^* \# \mu$ is regular with respect to the Lebesgue measure on $\mathbb{R}^q$ using Proposition 4.2.1 such that there exists $u : \mathbb{R}^q \to \mathbb{R}$ convex such that $\nabla u \circ \mathbf{P}^*$ pushes μ forward to ν. Moreover:

$$\int \langle \mathbf{P}^*\mathbf{x}, y \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y}) \overset{1}{\leq} \int (u(\mathbf{P}^*\mathbf{x}) + u^*(\mathbf{y})) \, \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y}) + \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y})$$

$$= \int u(\mathbf{P}^*\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\mathbf{y}) \mathrm{d}\nu(\mathbf{y}) + \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y})$$

$$\overset{2}{=} \int u(\mathbf{P}^*\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\mathbf{y}) \mathrm{d}(\nabla u \circ \mathbf{P}^* \# \mu)(\mathbf{y}) + \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y})$$

$$= \int u(\mathbf{P}^*\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) + \int u^*(\nabla u(\mathbf{P}^*\mathbf{x})) \mathrm{d}\mu(\mathbf{x}) + \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y})$$

$$\overset{3}{=} \int \langle \mathbf{P}^*\mathbf{x}, \nabla u(\mathbf{P}^*\mathbf{x}) \rangle_q \mathrm{d}\mu(x) + \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y})$$

In (1) we used convexity of u, in (2) we used that $\nabla u \circ \mathbf{P}^*$ is a push-forward and in (3) we used $u^*(\nabla u(\mathbf{x})) + u(\mathbf{x}) = \langle \mathbf{x}, \nabla u(\mathbf{x}) \rangle$. Now suppose that $\nabla u \circ \mathbf{P}^* \in \mathcal{H}(\mu, \nu)$. Then there exists $f \in \Gamma_0(\mathbb{R}, \mathbb{R}) \, , \|T(\mathbf{x})\|_2^2 = f'(\|\mathbf{x}\|_2^2) \, \mu$ a.e.

Moreover we have $\int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y}) \leq \int f(\|\mathbf{x}\|_2^2) + f^*(\|\mathbf{y}\|_2^2) \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y})$ by Young's inequality (see

Memo 4.2.2). In this way:

$$\int \langle \mathbf{P}^*\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi^*(\mathbf{x}, \mathbf{y}) \leq \int \langle \mathbf{P}^*\mathbf{x}, \nabla u(\mathbf{P}^*\mathbf{x}) \rangle_q \mathrm{d}\mu(\mathbf{x}) + \int f^*(\|\mathbf{x}\|_2^2) \mathrm{d}\mu(\mathbf{x}) + \int f(\|\mathbf{y}\|_2^2) \mathrm{d}\nu(\mathbf{y})$$

$$\overset{5}{=} \int \langle \mathbf{P}^*\mathbf{x}, \nabla u(\mathbf{P}^*\mathbf{x}) \rangle_q \mathrm{d}\mu(\mathbf{x}) + \int f^*(\|\nabla u(\mathbf{P}^*\mathbf{x})\|_2^2) + f(\|\mathbf{x}\|_2^2) \mathrm{d}\mu(\mathbf{x})$$

$$\overset{6}{=} \int \langle \mathbf{P}^*\mathbf{x}, \nabla u(\mathbf{P}^*\mathbf{x}) \rangle_q \mathrm{d}\mu(\mathbf{x}) + \int \|\nabla u(\mathbf{P}^*\mathbf{x})\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\mu(x)$$

$$\overset{7}{=} \int \langle \mathbf{P}^*\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\gamma(\mathbf{x}, \mathbf{y})$$

In (5) we used that $\nabla u \circ \mathbf{P}^*$ is a push-forward. In (6) we used that f satisfies $\|\nabla u(\mathbf{P}^*\mathbf{x})\|_2^2 \in \partial f(\|\mathbf{x}\|_2^2) = f'(\|\mathbf{x}\|_2^2)$ by definition of $\mathcal{H}(\mu, \nu)$ which implies $f^*(\|\nabla u(\mathbf{P}^*\mathbf{x})\|_2^2) + f(\|\mathbf{x}\|_2^2) = \|\nabla u(\mathbf{P}^*\mathbf{x})\|_2^2 \|\mathbf{x}\|_2^2$ (see Memo 2.1.2). In (7) γ is defined as $\gamma = (id \times \nabla u \circ \mathbf{P}^*)\#\mu$. Overall $(\mathbf{P}^*, \gamma)$ is optimal for (dual-sqGW). Using Theorem 4.2.5 this proves that γ is optimal for (sqGW) which concludes the proof.

$\square$

The condition $\nabla u \circ \mathbf{P} \in \mathcal{H}(\mu, \nu)$ is quite strong compared to the conditions of Theorem 4.2.3. For example in the case where ∇u is linear only orthogonal transformations are admissible (modulo a scaling). We believe that this result can be improved and we leave this study for further works. In general, and without any furter assumption on μ and ν, we postulate that there might be degenerate cases in which μ is regular but there is no deterministic optimal couplings.

4.2.4 Optimization and numerical experiments

In this section we provide numerical solutions for Gromov-Wasserstein problems in Euclidean spaces. We will rely on the equivalent formulation defined in Theorem 4.2.1 and Theorem 4.2.5. We consider two discrete probability measures $\mu = \frac{1}{n} \sum_{i=1}^n a_i \delta_{\mathbf{x}_i} \in \mathcal{P}(\mathbb{R}^p)$ and $\nu = \sum_{i=1}^m b_j \delta_{\mathbf{y}_j} \in \mathcal{P}(\mathbb{R}^q)$ with $\mathbf{a} \in \Sigma_n, \mathbf{b} \in \Sigma_m$.

Numerical solution for (innerGW)   We note $\mathbf{X} = (\mathbf{x}_i)_{i=1}^n \in \mathbb{R}^{n \times p}, \mathbf{Y} = (\mathbf{y}_j)_{j=1}^m \in \mathbb{R}^{m \times q}$. As seen in Theorem 4.2.1 computing (innerGW) can be achieved by solving $\sup_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \sup_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \langle \mathbf{X}\mathbf{P}^T\mathbf{Y}^T, \boldsymbol{\pi} \rangle_{\mathcal{F}}$ or equivalently:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a}, \mathbf{b})} \min_{\|\mathbf{P}\|_{\mathcal{F}} = \sqrt{p}} \langle -\mathbf{X}\mathbf{P}^T\mathbf{Y}^T, \boldsymbol{\pi} \rangle_{\mathcal{F}} \tag{4.33}$$

The resulting problem is convex *w.r.t.* $\mathbf{P}$ and π (but not jointly convex). We propose to solve equation (4.33) using Block Coordinate Descent (BCD) which alternates between minimizing *w.r.t.* $\mathbf{P}$ and $\boldsymbol{\pi}$.

Memo 4.2.2 (Young's Inequality). *Let $a, b \in \mathbb{R}_+ \times \mathbb{R}_+$ and p, q real numbers greater than 1 with $\frac{1}{p} + \frac{1}{q} = 1$. Then:*

$$ab \leq \frac{a}{p} + \frac{b}{q} \tag{4.31}$$

More generally if f is a convex function and f^* is its Legendre transform then:

$$ab \leq f(a) + f^*(b) \tag{4.32}$$

which is a consequence of the celebrated Fenchel-Young inequality (see Memo 2.1.2)

Algorithm 7 Gromov-Wasserstein with inner products

1: Require $\mu = \frac{1}{n} \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i} \in \mathcal{P}(\mathbb{R}^p)$ and $\nu = \sum_{i=1}^{m} b_j \delta_{\mathbf{y}_j} \in \mathcal{P}(\mathbb{R}^q)$

2: Set $\mathbf{X} = (\mathbf{x}_i)_{i=1}^{n} \in \mathbb{R}^{n \times p}, \mathbf{Y} = (\mathbf{y}_j)_{j=1}^{m} \in \mathbb{R}^{m \times q}$

3: Initialize $\boldsymbol{\pi} = \mathbf{a}\mathbf{b}^T$.

4: **while** not converged **do**

5:      $\mathbf{P} \leftarrow \frac{\sqrt{d}}{\|\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X}\|_{\mathcal{F}}} \mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X}$ // (maximize (MaxOT) w.r.t. $\mathbf{P}$)

6:      $\boldsymbol{\pi} \leftarrow \arg\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \langle -\mathbf{X}\mathbf{P}^T\mathbf{Y}^T, \boldsymbol{\pi}\rangle_{\mathcal{F}}$ // (maximize (MaxOT) w.r.t. $\boldsymbol{\pi} \sim$ linear OT)

7: **end while**

8: return $(\pi, \mathbf{P})$

Interestingly enough the minimization of $\boldsymbol{\pi}$ with $\mathbf{P}$ fixed is a linear transportation problem with ground cost $-\mathbf{X}\mathbf{P}^T\mathbf{Y}^T$ which can be computed using standard solvers (see Chapter 2).

The minimization *w.r.t.* $\mathbf{P}$ with $\boldsymbol{\pi}$ fixed reads $\sup_{\|\mathbf{P}\|_{\mathcal{F}}=\sqrt{p}} \langle \mathbf{X}\mathbf{P}^T\mathbf{Y}^T, \boldsymbol{\pi}\rangle_{\mathcal{F}}$ and has a closed-form solution based on Lemma 4.2.2. More precisely it reads:

$$\mathbf{P} \leftarrow \frac{\sqrt{d}}{\|\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X}\|_{\mathcal{F}}} \mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X} \tag{4.34}$$

This procedure is presented in Algorithm 7. The complexity is driven by the linear OT problem which is $O(n^3 \log(n))$ when $n = m$. The complexity for computing $\mathbf{P}$ at each iteration is $O(mn(q+p) + pq)$ with $O(mn(q+p))$ for the computing $\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X}$ and $O(pq)$ for $\|\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X}\|_{\mathcal{F}}$. The overall complexity when $n = m$ is then $O(n^3 \log(n) + n^2(q+p) + pq)$.

Numerical solution for (sqGW)   We note $\mathbf{x} = \mathrm{Diag}(\mathbf{X}\mathbf{X}^T) \in \mathbb{R}^n$, $\mathbf{y} = \mathrm{Diag}(\mathbf{Y}\mathbf{Y}^T) \in \mathbb{R}^m$. As seen in Theorem 4.2.5 computing (sqGW) can be achieved by first substracting the mean of the measures and then by solving:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \min_{\mathbf{P} \in \mathbb{R}^{q \times p}} \langle -(\mathbf{X}\mathbf{P}^T\mathbf{Y}^T + \mathbf{x}\mathbf{y}^T), \boldsymbol{\pi}\rangle_{\mathcal{F}} + \frac{1}{8}\|\mathbf{P}\|_{\mathcal{F}}^2 \tag{4.35}$$

This problem also is convex *w.r.t.* $\boldsymbol{\pi}$ and $\mathbf{P}$ (but not jointly convex) and we can rely on a BCD procedure to find a local minimum. The minimization *w.r.t.* to $\boldsymbol{\pi}$ with $\mathbf{P}$ fixed is a linear OT problem with ground cost $-(\mathbf{X}\mathbf{P}^T\mathbf{Y}^T + \mathbf{x}\mathbf{y}^T)$ and the minimization *w.r.t.* $\mathbf{P}$ with $\boldsymbol{\pi}$ fixed reads:

$$\min_{\mathbf{P} \in \mathbb{R}^{q \times p}} \langle -\mathbf{X}\mathbf{P}^T\mathbf{Y}^T, \boldsymbol{\pi}\rangle_{\mathcal{F}} + \frac{1}{8}\|\mathbf{P}\|_{\mathcal{F}}^2 \stackrel{def}{=} \min_{\mathbf{P} \in \mathbb{R}^{q \times p}} G(\mathbf{P}) \tag{4.36}$$

which is a unconstrained QP which can be solved in closed-form. Indeed the gradient reads $\nabla G(\mathbf{P}) = -\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X} + \frac{1}{4}\mathbf{P}$ and by first order condition a solution can be found at $\nabla G(\mathbf{P}) = 0$. In this way the iteration for $\mathbf{P}$ are:

$$\mathbf{P} \leftarrow 4\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X} \tag{4.37}$$

This procedure is presented in Algorithm 8. Computing all the terms $\|\mathbf{x}^{(i)}\|_2^2 \|\mathbf{y}^{(j)}\|_2^2$ where $\mathbf{x}^{(i)}$ is the i-th row of $\mathbf{X}$ has a naive complexity of $O(np + mq + mn)$ and one needs also to compute $4\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X}$ at a $O(mn(q+p))$ price. The overall complexity is then $O(n^3 \log(n) + (p+q)n^2 + n^2 + n(p+q)) = O(n^3 \log(n) + (p+q)n^2)$ when $n = m$.

Runtimes comparison   We perform a comparison between runtimes of GW using different algorithmic solutions. We consider the Gromov-Wasserstein problem between 10 realizations of one source 2D random probability measures of $n \in \{10, ..., 30000\}$ points and one target 3D random measures with $m = 100$

Algorithm 8 Gromov-Wasserstein with squared Euclidean distances

---

1: Require $\mu = \frac{1}{n} \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i} \in \mathcal{P}(\mathbb{R}^p)$ and $\nu = \sum_{i=1}^{m} b_j \delta_{\mathbf{y}_j} \in \mathcal{P}(\mathbb{R}^q)$

2: Subtract the mean: $\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{n} \sum_k \mathbf{x}_k$, $\mathbf{y}_j \leftarrow \mathbf{y}_j - \frac{1}{m} \sum_k \mathbf{y}_k$.

3: Set $\mathbf{X} = (\mathbf{x}_i)_{i=1}^{n} \in \mathbb{R}^{n \times p}, \mathbf{Y} = (\mathbf{y}_j)_{j=1}^{m} \in \mathbb{R}^{m \times q}, \mathbf{x} = \text{Diag}(\mathbf{X}\mathbf{X}^T) \in \mathbb{R}^n, \mathbf{y} = \text{Diag}(\mathbf{Y}\mathbf{Y}^T) \in \mathbb{R}^m$

4: Initialize $\boldsymbol{\pi} = \mathbf{a}\mathbf{b}^T$.

5: **while** not converged **do**

6: $\quad \mathbf{P} \leftarrow 4\mathbf{Y}^T \boldsymbol{\pi}^T \mathbf{X}$ // (maximize (dual-sqGW) *w.r.t.* $\mathbf{P}$)

7: $\quad \boldsymbol{\pi} \leftarrow \arg\min_{\boldsymbol{\pi} \in \Pi(\mathbf{a},\mathbf{b})} \langle -(\mathbf{X}\mathbf{P}^T\mathbf{Y}^T + \mathbf{x}\mathbf{y}^T), \boldsymbol{\pi} \rangle_{\mathcal{F}}$ // (maximize (dual-sqGW) *w.r.t.* $\boldsymbol{\pi} \sim$ linear OT)

8: **end while**

9: return $(\pi, \mathbf{P})$
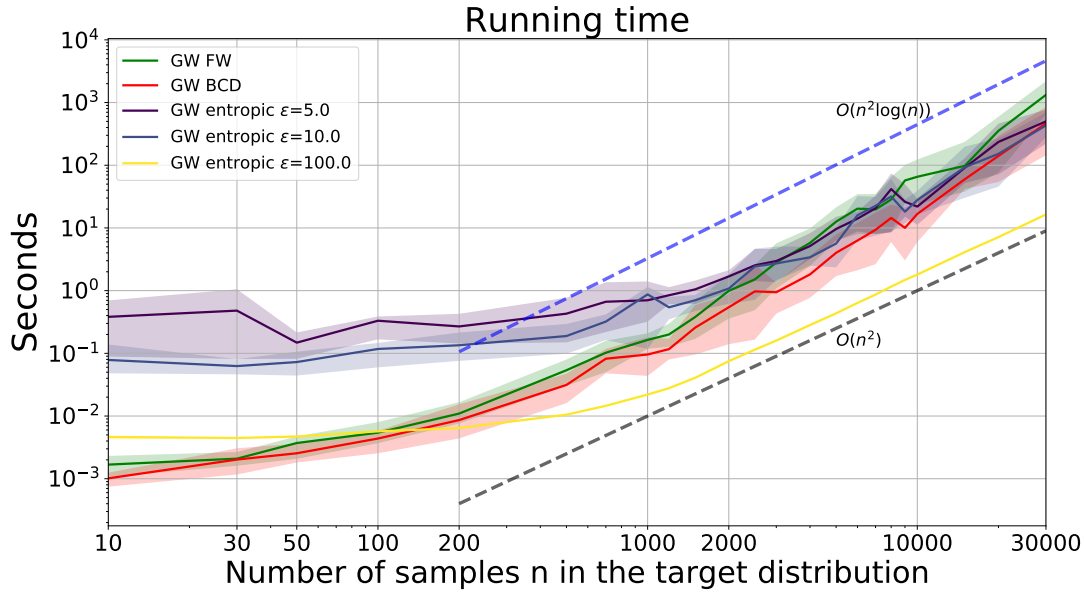
Figure 4.7: Runtimes comparison of *GW* using the BCD approach Algorithm 8. (GW BCD), the Frank-Wolfe approach (GW FW) and entropic-*GW* between two random distributions whose source points vary from 10 to 30000 in log-log scale. The time does not include the calculation of the pair-to-pair distances but only the time of the different loops. The variance is computed among the different 10 realizations.

points. We compute the Gromov-Wasserstein distance using squared Euclidean distances as $c_{\mathcal{X}}, c_{\mathcal{Y}}$. We compare the timings of the Frank-Wolfe algorithm (see Chapter 3), the entropic regularized approach with $\varepsilon \in \{5, 10, 100\}$ (see Chapter 2), and the BCD approach Algorithm 8 using the same initialization of $\boldsymbol{\pi} = \mathbf{a}\mathbf{b}^T$ for all methods. The result is depicted in Figure 4.7. Please note that the timings are calculated without taking into account the time needed for computing the matrices $\mathbf{C}_1, \mathbf{C}_2$ or $\mathbf{X}, \mathbf{Y}$ but are only based on the loops of the different algorithms. The entropic-*GW* could not be computed with $\varepsilon \leq 5$ due to overflows in the Sinkhorn algorithm. Another difficulty of computing entropic-*GW* is the fact that the regularization parameter ε is inversely proportional to the gradient step [Peyré 2016]. Thus only high value of ε are computable in reasonable time and for $\varepsilon < 5$ we observe that the convergence is very slow even on this simple example. This often implies that only very blurred solutions can be computed
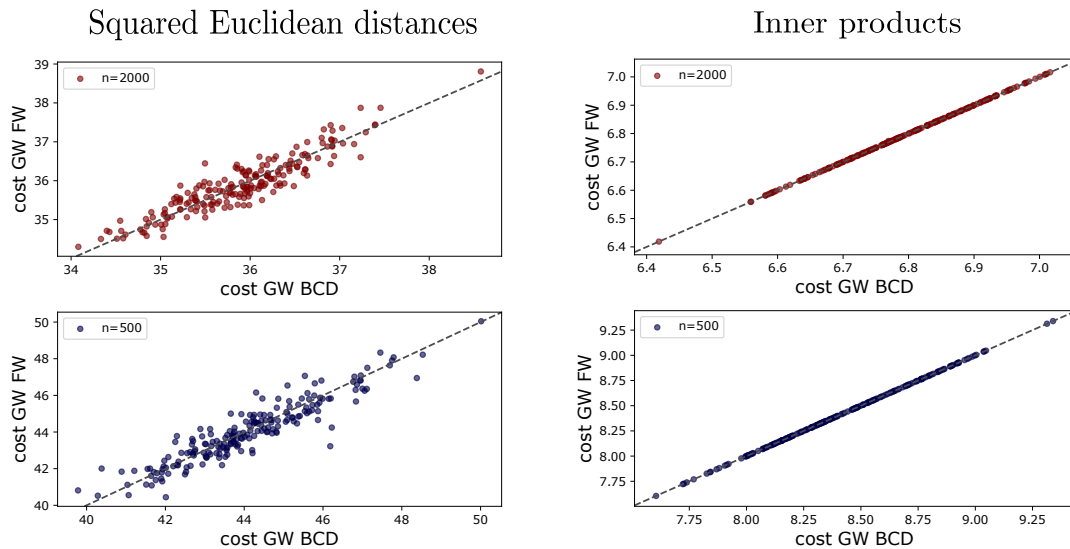
Figure 4.8: Comparison of the FW solver and the BCD solver for computing the Gromov-Wasserstein distance. **(left)** $c_\mathcal{X}, c_\mathcal{Y}$ are squared Euclidean distances. **(right)** $c_\mathcal{X}, c_\mathcal{Y}$ are inner products. The x-axis is the *GW* cost obtained by the BCD approach and the y-axis the *GW* cost obtained with the FW algorithm for respectively **(top)** $n = 2000$ samples **(bottom)** $n = 500$ samples in each distribution. The dashed line represents the diagonal $y = x$.

which explains the low variance of entropic-*GW* with $\varepsilon = 100$. We can see in Figure 4.7 that the BCD approach is a little bit faster than the FW approach, suggesting that the BCD may converge faster to a local minimum than FW, which is, of course, data dependant. Overall both methods are equally rapid on this example.

Costs comparison In this experiment we compared the ability of the BCD approach to find a better solution than the FW approach. We consider the Gromov-Wasserstein problem with both inner product similarities and squared Euclidean distances. We compute 200 distances using both algorithms where each distance is calculated by: (1) We draw $n \in \{500, 2000\}$ samples from two 10 dimensional normal distributions (2) We associate to these points random weights $(\mathbf{a}, \mathbf{b}) \in \Sigma_n \times \Sigma_n$ (3) We initialize the algorithms with the same random coupling matrix $\boldsymbol{\pi}$. The initialization is computed by sampling a random matrix with positive entries and by scaling it with the Sinkhorn algorithm in order to have the prescribed marginals $\mathbf{a}, \mathbf{b}$. Results are depicted in Figure 4.8. We plot the *GW* cost obtained by BCD approach *vs* the cost obtained by the FW approach after convergence of each algorithm. As seen in Figure 4.8 there is no strong differences between FW approach and the BCD approach when squared Euclidean distances are considered (left part of the figure): the *FW* algorithm finds a better solution in 51% of the cases when $n = 2000$ and 56% for $n = 500$. Surprisingly both algorithms seem to lead to the same solution when inner product similarities are used (right part of the figure). Indeed in 99% of the cases the costs are identical up to $1e - 7$.

Conclusion on the experiments As seen in these experiments, there are no significant differences between the FW approach and the BCD procedure. It is not surprising for the runtimes comparison since both methods have a theoretical cubic complexity. The cost comparison experiment may also suggest that both BCD procedure and FW are equivalent in the case of inner product similarities, that is the

iterations of the BCD are the same than the iterations of the FW. Further studies could be conducted to examine the high-dimensional setting where we postulate that the BCD approach may lead to better results that the FW procedure, *i.e.* may converge faster a produce a better solution.

4.2.5 The Gromov-Monge problem in Euclidean spaces

As seen in Section 4.1.2 the Gromov-Wasserstein problem is also very related the Gromov-Monge problem [Mémoli 2018] which is defined in Euclidean spaces for $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ as:

Definition 4.2.4 (Gromov-Monge). *The Gromov-Monge problem aims at finding:*

$$GM_2^2(\mu, \nu) = \inf_{T\#\mu=\nu} J(T) = \inf_{T\#\mu=\nu} \int \left(\|\mathbf{x} - \mathbf{x}'\|_2^2 - \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2 \right)^2 d\mu(\mathbf{x}) d\mu(\mathbf{x}') \quad \text{(GM)}$$

With the understanding that $GM_2(\mu, \nu) = +\infty$ when $\{T : \mathbb{R}^p \to \mathbb{R}^q | T\#\mu = \nu\} = \emptyset$

The problem (GM) is the exact counterpart for the Gromov-Wasserstein distance of the Monge problem for the Wasserstein distance. If we note $GW_2(\mu, \nu)$ the Gromov-Wasserstein distance with squared Euclidean costs then we have $GW_2(\mu, \nu) \leq GM_2(\mu, \nu)$ but in general both problems are not equivalent (see [Mémoli 2018]). Note that when problem 1 holds, *i.e.* when the Gromov-Wasserstein problem admits an optimal transport plan supported on a deterministic map then both (GM) and (sqGW) are equivalent. Moreover we have seen previously in Theorem 4.1.2 that when μ and ν are discrete probability measures with the same number of atoms and uniform weights then (GM) is also equivalent to the Gromov-Wasserstein problem (sqGW) so that that $GW_2(\mu, \nu) = GM_2(\mu, \nu)$.

We propose to study further the Gromov-Monge problem in this section. Especially we consider the special case of Gaussian measures $\mu = \mathcal{N}(\mathbf{m}_\nu, \boldsymbol{\Sigma}_\nu), \nu = \mathcal{N}(\mathbf{m}_\mu, \boldsymbol{\Sigma}_\mu)$. It is motivated by the linear OT theory where, when $p = q$, there is a close form solution for W_2. In this case the optimal Monge map is *linear* and given by [Takatsu 2011] $T : \mathbf{x} \to \mathbf{m}_\nu + \mathbf{A}(\mathbf{x} - \mathbf{m}_\nu)$ where:

$$\mathbf{A} = \boldsymbol{\Sigma}_\mu^{-1/2} (\boldsymbol{\Sigma}_\mu^{1/2} \boldsymbol{\Sigma}_\nu \boldsymbol{\Sigma}_\mu^{1/2})^{\frac{1}{2}} \boldsymbol{\Sigma}_\mu^{-1/2} \tag{4.38}$$

Can we derive the same type of result for the Gromov-Monge geometry? We will prove that when restricted to *linear* push-forward and in the special case $p = q$ the problem admits also a close form expression. In this way we consider the following *linear Gromov-Monge* problem:

$$LGM_2^2(\mu, \nu) = \inf_{\substack{T\#\mu=\nu \\ \text{T is linear}}} J(T) = \inf_{\substack{T\#\mu=\nu \\ \text{T is linear}}} \int \left(\|\mathbf{x} - \mathbf{x}'\|_2^2 - \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2 \right)^2 d\mu(\mathbf{x}) d\mu(\mathbf{x}') \quad \text{(LGM)}$$

We recall that $\mathbb{V}_p(\mathbb{R}^q)$ is the Stiefel manifod defined by $\mathbb{V}_p(\mathbb{R}^q) = \{\mathbf{B} \in \mathbb{R}^{q\times p} | \mathbf{B}^T \mathbf{B} = \mathbf{I_p}\}$. The main result of this section is the following theorem:

Theorem 4.2.6. *Let $\mu = \mathcal{N}(0, \boldsymbol{\Sigma}_\nu) \in \mathcal{P}(\mathbb{R}^p), \nu = \mathcal{N}(0, \boldsymbol{\Sigma}_\mu) \in \mathcal{P}(\mathbb{R}^q)$ centered without loss of generality. Let $\boldsymbol{\Sigma}_\mu = \mathbf{V}_\mu \mathbf{D}_\mu \mathbf{V}_\mu^\top, \boldsymbol{\Sigma}_\nu = \mathbf{V}_\nu \mathbf{D}_\nu \mathbf{V}_\nu^\top$ be the diagonalizations of the covariance matrices such that eigenvalues of $\mathbf{D}_\mu$ and $\mathbf{D}_\nu$ are ordered nondecreasing.*

When $p \neq q$ we have:

$$LGM_2^2(\mu, \nu) = 4(tr(\boldsymbol{\Sigma}_\mu) - tr(\boldsymbol{\Sigma}_\nu))^2 + 8(tr(\boldsymbol{\Sigma}_\mu \boldsymbol{\Sigma}_\mu) + tr(\boldsymbol{\Sigma}_\nu \boldsymbol{\Sigma}_\nu)) + 16 \min_{\mathbf{B} \in \mathbb{V}_p(\mathbb{R}^q)} -tr(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B})$$

$$\tag{4.39}$$

> *When $p = q$, an optimal linear Monge map is given by $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where:*
>
> $$\mathbf{A} = \mathbf{V}_\nu \mathbf{D}_\nu^{1/2} \mathbf{D}_\mu^{-1/2} \mathbf{V}_\mu^\top = \mathbf{\Sigma}_\nu^{1/2} \mathbf{V}_\nu \mathbf{V}_\mu^\top \mathbf{\Sigma}_\mu^{-1/2} \qquad (4.40)$$
>
> *so that:*
>
> $$LGM_2^2(\mu, \nu) = 4(tr(\mathbf{\Sigma}_\mu) - tr(\mathbf{\Sigma}_\nu))^2 + 8(tr(\mathbf{\Sigma}_\mu \mathbf{\Sigma}_\mu) + tr(\mathbf{\Sigma}_\nu \mathbf{\Sigma}_\nu)) - 16tr(\mathbf{D}_\mu \mathbf{D}_\nu) \qquad (4.41)$$

Sketch of proof. We can show that when considering only linear push-forward the problem (GM) can be recast as a Orthogonality constrained Quadratic Program (QPOC) which, when $p = q$, admits a close form. This is made possible thanks to the Gaussian assumption which allows to compute the 4-th order moments of the distributions using Isserlis theorem [Isserlis 1918] which prove that 4-th order moments can be computed using the 2-nd order ones. We give the full proof in Section 6.2.11. □

Geometric interpretations of *LGM*   Interestingly enough the optimal linear map can be related, *inter alia*, to the optimal linear map of the classical Monge problem. In the following we consider $p = q$ so that we have using Theorem 4.2.6:

$$LGM_2^2(\mu, \nu) = 4(\text{tr}(\mathbf{\Sigma}_\mu) - \text{tr}(\mathbf{\Sigma}_\nu))^2 + 8(\text{tr}(\mathbf{\Sigma}_\mu \mathbf{\Sigma}_\mu) + \text{tr}(\mathbf{\Sigma}_\nu \mathbf{\Sigma}_\nu)) - 16\text{tr}(\mathbf{D}_\mu \mathbf{D}_\nu) \qquad (4.42)$$

wich corresponds to $\mathbf{A} = \mathbf{V}_\nu \mathbf{D}_\nu^{1/2} \mathbf{D}_\mu^{-1/2} \mathbf{V}_\mu^\top = \mathbf{\Sigma}_\nu^{1/2} \mathbf{V}_\nu \mathbf{V}_\mu^\top \mathbf{\Sigma}_\mu^{-1/2}$.

- **Case $\mathbf{\Sigma}_\mu = \mathbf{\Sigma}_\nu$.** When the covariances are equals, *i.e.* $\mathbf{\Sigma}_\mu = \mathbf{\Sigma}_\nu$ we can conclude that $LGM_2(\mu, \nu) = 0 + 8(\text{tr}(\mathbf{\Sigma}_\mu \mathbf{\Sigma}_\mu) + \text{tr}(\mathbf{\Sigma}_\mu \mathbf{\Sigma}_\mu)) - 16\text{tr}(\mathbf{D}_\mu \mathbf{D}_\mu) = 16\text{tr}(\mathbf{\Sigma}_\mu \mathbf{\Sigma}_\mu) - 16\text{tr}(\mathbf{\Sigma}_\mu \mathbf{\Sigma}_\mu) = 0$ which corresponds to $\mathbf{A} = \mathbf{I}$. This implies that an optimal way for transferring the masses so that, on average, the pair-to-pair distances are preserved is by the means of the identity mapping.

- **Case $\mathbf{\Sigma}_\mu = k\, \mathbf{\Sigma}_\nu$.** If there is a scaling factor between the covariances $\mathbf{\Sigma}_\mu = k\mathbf{\Sigma}_\nu$ with $k > 0$ then the optimal map scales uniformly the first measure to preserve the distances so that $\mathbf{A} = \frac{1}{\sqrt{k}}\mathbf{I}$. With a little calculus one can check that $LGM_2(\mu, \nu) = 4(k - 1)^2 \left(\text{tr}(\mathbf{\Sigma}_\nu)^2 + 2\text{tr}(\mathbf{\Sigma}_\nu \mathbf{\Sigma}_\nu)\right)$ so that $LGM_2(\mu, \nu)$ is minimal, equal to zero, when $k = 1$ which corresponds to the previous case. When k increases above 1 the distance increases quadratically and when $k \in]0, 1]$ the distances decreases as k goes to 1.

- **Rotation.** Another interesting case is when we rotate the samples of the distribution and we compute the linear Gromov-Monge between the original distribution and its rotated counterpart. This case corresponds to $\mathbf{\Sigma}_\nu = \mathbf{O}\mathbf{\Sigma}_\mu \mathbf{O}^T$ where $\mathbf{O} \in \mathcal{O}(p)$ and we can check easily that $LGM_2(\mu, \nu) = 0$ with an optimal map given the rotation. This behavior is intuitive since the Gromov-Monge problem with linear map is invariant by rotations.

- **Commuting covariances.** Finally when the covariances commute *i.e.* $\mathbf{\Sigma}_\mu \mathbf{\Sigma}_\nu = \mathbf{\Sigma}_\nu \mathbf{\Sigma}_\mu$ we can relate with the linear transportation. In this situation both are simultaneously diagonalizable and eigenspaces coincide. We recall that the optimal linear map for the Wasserstein distance with $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$ is $\mathbf{A}_W = \mathbf{\Sigma}_\mu^{-1/2}(\mathbf{\Sigma}_\mu^{1/2} \mathbf{\Sigma}_\nu \mathbf{\Sigma}_\mu^{1/2})^{\frac{1}{2}} \mathbf{\Sigma}_\mu^{-1/2}$ which reduces to $\mathbf{A}_W = \mathbf{\Sigma}_\nu^{1/2} \mathbf{\Sigma}_\mu^{-1/2}$ when covariances are commuting. Moreover since matrices share the same eigenspaces then we can take $\mathbf{V}_\mu = \mathbf{V}_\nu$ for the linear Gromov-Monge. In this way $\mathbf{A}$ reduces to $\mathbf{A} = \mathbf{\Sigma}_\nu^{1/2} \mathbf{\Sigma}_\mu^{-1/2} = \mathbf{A}_W$. This proves that when the covariances commute the optimal map of Wasserstein is an optimal map for the linear Gromov-Monge.

Figure 4.9: Example of linear Gromov-Monge mapping estimation between empirical distributions. **(left)** 2D source and target distributions. **(middle)** resulting linear mapping of Wasserstein $\mathbf{A}_W$. **(right)** resulting linear mapping using the linear Gromov-Monge mapping $\mathbf{A}$. Note that in this case the mapped samples are arbitrary rotated.

An illustration of the map $\mathbf{A}$ is given in Figure.4.9 where we compute LGM between two 2D empirical distributions, that we consider Gaussian in first approximation. The optimal map $\mathbf{A}$ gives a different behaviour than the Wasserstein map $\mathbf{A}_W$ which seems to better grasp the transformation of the samples. It is somehow natural since the Gromov-Monge cost is less rigid than the Wasserstein one and only forces the samples to be isometrically distributed on average. However note that the target samples are arbitrary rotated for the case of Gromov-Monge since LGM is invariant by rotation of the samples.

Solving the problem when $p \neq q$

As described in Theorem 4.2.6 the situation is more delicate when $p \neq q$, *i.e.* when the Gaussian measures are not part of the same ground Euclidean space. In this case there is no close form anymore and one needs to solve the following Quadratic optimization with Orthogonality Constraints (QPOC) problem:

$$\min_{\mathbf{B} \in \mathbb{V}_p(\mathbb{R}^q)} -\mathrm{tr}(\mathbf{B}^\top \mathbf{D}_\nu \mathbf{B} \mathbf{D}_\mu) \stackrel{def}{=} \min_{\mathbf{B} \in \mathbb{V}_p(\mathbb{R}^q)} F(\mathbf{B}) \qquad \text{(QPOC)}$$

This problem is a particular special case of optimizing a smooth function over the Stiefel manifold, which is non-convex in general but for which various methods have been proposed over the years (see *e.g* [Jiang 2014, Wen 2010, Abrudan 2008, Absil 2009]). A standard approach for solving these types of problems is to leverage the Riemannian structure of $\mathbb{V}_p(\mathbb{R}^q)$ to use a gradient descent method on the manifold. Informally, the idea is to start with a arbitrary point $\mathbf{B}_0 \in \mathbb{V}_p(\mathbb{R}^q)$ then iteratively move in a search direction $D(\mathbf{B}_0)$ defined by a tangent vector while staying on $\mathbb{V}_p(\mathbb{R}^q)$ until a critical point is found. In general this procedure require to compute the so-called *exponential map* which is quite difficult. A cheaper alternative can be found into a *retraction* map which approximates the exponential map (see *e.g.* [Liu 2016] and references therein). We propose to solve (QPOC) using the Geoopt library [Kochurov 2020]. We illustrate the resulting optimal map in Figure 4.10 where we compute the LGM problem between two empirical distributions. The source distribution is 3 dimensional while the target distribution is 2 dimensional. We observe that the optimal Gromov-Monge map manages to capture the overall transformation of the source distribution.
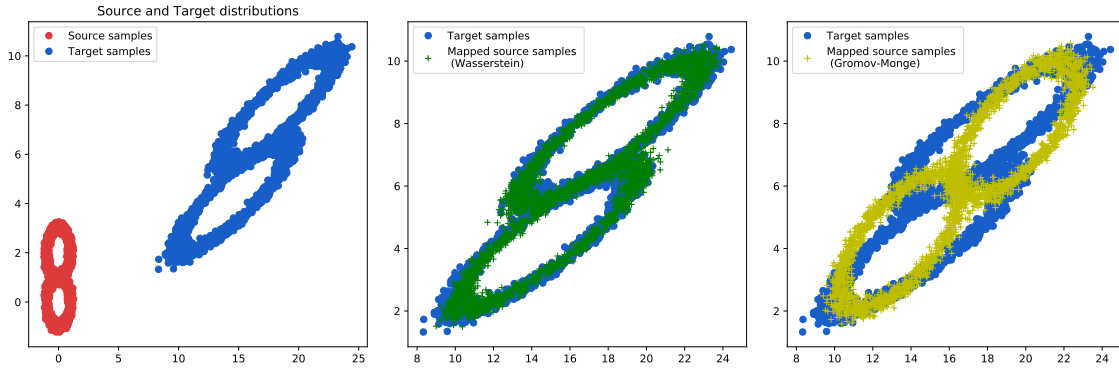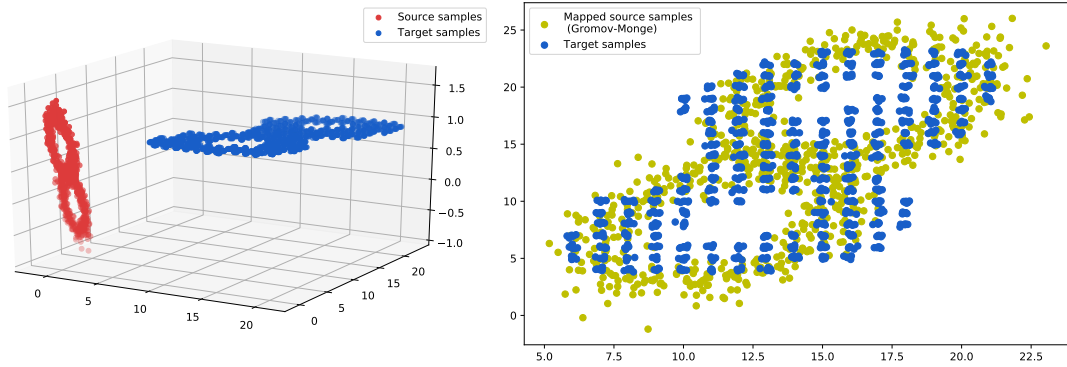
Figure 4.10:   Example of linear Gromov-Monge mapping estimation between two empirical distributions. **(left)** 3D source and 2D target distributions. **(right)** resulting linear mapping using the linear Gromov-Monge mapping **A**.

4.3   Conclusion: perspectives and open questions

In our opinion the previous results raise interesting questions and pave the path for a deeper understanding of the Gromov-Wasserstein and Gromov-Monge geometry in Euclidean spaces. The works on the regularity of optimal couplings are, to the best of our knowledge, the first work on this subject and we believe that these results could be improved in order to bridge the theoretical gap between Wasserstein and Gromov-Wasserstein.

Gromov-Wasserstein in Euclidean spaces   In the inner product case we have seen that among the set of solutions of (MaxOT) the couples with $\mathbf{P}^*$ surjective are of particular interest. For example is it possible under some mild conditions on μ, ν to prove that there always exists an optimal couple $(\pi^*, \mathbf{P}^*)$ of (MaxOT) such that $\mathbf{P}^*$ is surjective? If this result holds then it would imply that, if the source measure is regular, the Gromov-Wasserstein problem with inner products is equivalent to its Monge counterpart:

$$\inf_{T\#\mu=\nu} \int (\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle T(\mathbf{x}), T(\mathbf{x}') \rangle_q)^2 \mathrm{d}\mu(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}') \tag{4.43}$$

The equivalence between Gromov-Monge and Gromov-Wasserstein problems would be a nice generalization of the Brenier's theorem in the case of the Gromov-Wasserstein geometry. Related to this problem the unicity of the couples $(u, \mathbf{P})$ is also an interesting further work. More precisely can we find suitable conditions on μ, ν so that there is a unique $\nabla u \circ \mathbf{P}$ which pushes μ forward to ν optimally? When the support of the measures are not "too symmetric" it seems reasonable that all the solutions are somehow unique modulo a class of isometries such as rotations. As described in this section the Brenier's theorem states that there is a unique optimal push forward in the case of Wasserstein geometry. This property allows to derive close form expression for the Gaussian case and we believe that similar reasoning could be made if the unicity (up to some rotations) holds for (MaxOT). Finally can we extend the previous results about the inner case for kernel similarities? More precisely when $c_{\mathcal{X}}(x, x') = k_{\mathcal{X}}(x, x')$ defines a kernel (same for $c_{\mathcal{Y}}$) we know from the kernel trick that there exists an inner product space $\mathcal{V}_{\mathcal{X}}$ and a feature map $\phi : \mathcal{X} \to \mathcal{V}_{\mathcal{X}}$ such that $c_{\mathcal{X}}(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}_{\mathcal{X}}}$. In this case can we find similar results regarding the regularity of the optimal transport plans of the Gromov-Wasserstein distance?

The squared Euclidean distance case seems to be more delicate to handle as it does not echo to classical OT costs. We believe that the dual problem could be used to find a close form expression for 1D discrete

probability measures with possibly different number of atoms and non-uniform weights, as done for the inner product case. This would be an interesting stronger result than the case considered in Section 4.1.2. We postulate that, in this case, an optimal coupling is also given by a monotone rearrangement which is increasing or decreasing. For general probability measures assessing the regularity of optimal transport plans seems more complicated. The sufficient condition of Proposition 4.2.4 may be improved with a necessary condition in order to characterize optimal couplings under some hypothesis on μ, ν. Another interesting approach for the squared Euclidean distance case would be to consider optimal couplings when the target measure is a small perturbation of the source measure. More precisely we believe the following holds:

Problem 2. *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$. Let $\nu = (id + \varepsilon u)\#\mu \in \mathcal{P}(\mathbb{R}^p)$ with $u : \mathbb{R}^p \to \mathbb{R}^p$ be a small perturbation of μ. Then the map $\gamma = (id \times T)\#\mu$ where $T = id + \varepsilon u$ is optimal for:*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int \left(\|\mathbf{x} - \mathbf{x}'\|_2^2 - \|\mathbf{y} - \mathbf{y}'\|_2^2 \right)^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \tag{4.44}$$

CO-Optimal Transport

> "Tu as tort de lire les journaux; ça te congestionne."
>
> – André Gide, *Les Faux-Monnayeurs*

Contents

Summary of the contributions

This chapter is based on the paper [Redko 2020] and addresses the problem of optimal transport on incomparable spaces. The original formulation of the optimal transport problem relies on the existence of a cost function between the samples of the two distributions, which makes it impractical for comparing data distributions supported on different topological spaces. To circumvent this limitation, we propose a novel OT problem, named COOT for CO-Optimal Transport, that aims to simultaneously optimize two transport maps between both samples and features. This is different from other approaches that either discard the individual features by focusing on pairwise distances (e.g. Gromov-Wasserstein) or need to model explicitly the relations between the features. COOT leads to interpretable correspondences between both samples and feature representations and holds metric properties. We provide a thorough theoretical analysis of our framework and establish rich connections with the Gromov-Wasserstein distance. We demonstrate its versatility with two machine learning applications in heterogeneous domain adaptation and co-clustering/data summarization, where COOT leads to performance improvements over the competing state-of-the-art methods.

5.1   Introduction

The problem of comparing two sets of samples arises in many fields in machine learning, such as manifold alignment [Cui 2014], image registration [Haker 2001], unsupervised word and sentence translation [Rapp 1995] among others. When correspondences between the sets are known *a priori*, one can align them with a global transformation of the features, *e.g*, with the widely used *Procrustes analysis* [Gower 2004, Goodall 1991]. For unknown correspondences, other popular alternatives to this method include correspondence free manifold alignment procedure [Wang 2009], soft assignment coupled with a Procrustes matching [Rangarajan 1997a] or Iterative closest point and its variants for 3D shapes [Besl 1992, Yang 2020]. When one models the considered sets of samples as empirical probability distributions, the optimal transport framework provides a solution to find, without supervision, a soft-correspondence map between them given by an *optimal coupling*. OT-based approaches have been used with success in numerous applications such as embeddings' alignments [Alvarez-Melis 2019, Grave 2019] and Domain Adaptation (DA) [Courty 2017] to name a few. However, one important limit of using OT for such tasks is that the two sets are assumed to lie in the same space so that the cost between samples across them can be computed. This major drawback does not allow OT to handle correspondences' estimation across heterogeneous spaces, preventing its application in problems such as, for instance, heterogeneous DA (HDA). To circumvent this restriction, one may rely on the Gromov-Wasserstein distance [Memoli 2011]: a non-convex quadratic OT problem that finds the correspondences between two sets of samples based on their pairwise intra-domain similarity (or distance) matrices. Such an approach was successfully applied to sets of samples that do not lie in the same Euclidean space, *e.g* for shapes [Solomon 2016], word embeddings [Alvarez-Melis 2018a] and HDA [Yan 2018] mentioned previously. One important limit of GW is that it finds the samples' correspondences but discards the relations between the features by considering pairwise similarities only. Another line of works [Alvarez-Melis 2019, Grave 2019] considers the problem of matching sets of points with respect to a global transformations of the features, usually modeled by a linear transformation such as a rotation. These approches differ from the work proposed here where we consider instead a probabilistic coupling of the features as described below.

In this Chapter, we propose a novel OT approach called CO-Optimal transport (COOT) that simultaneously infers the correspondences between the samples *and* the features of two arbitrary sets. Our new formulation includes GW as a special case, and has an extra-advantage of working with raw data directly without needing to compute, store and choose computationally demanding similarity measures

required for the latter. Moreover, COOT provides a meaningful mapping between both instances and features across the two datasets thus having the virtue of being interpretable. We thoroughly analyze the proposed problem, derive an optimization procedure for it and highlight several insightful links to other approaches. On the practical side, we provide evidence of its versatility in machine learning by putting forward two applications in HDA and co-clustering where our approach achieves state-of-the-art results.

The rest of this chapter is organized as follows. We introduce the COOT problem in Section 5.2, states its mathematical properties in Section 5.3 and give an optimization routine for solving it efficiently in Section 5.4. In Section 5.5, we show how COOT is related to other OT-based distances and recover efficient solvers for some of them in particular cases. Finally, in Section 5.6.1 and Section 5.6.2, we present an experimental study providing highly competitive results in HDA and co-clustering compared to several baselines.

5.2 CO-Optimal transport optimization problem

We consider two datasets represented by matrices $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ and $\mathbf{X}' = [\mathbf{x}'_1, \ldots, \mathbf{x}'_{n'}]^T \in \mathbb{R}^{n' \times d'}$, where in general we assume that $n \neq n'$ and $d \neq d'$. In what follows, the rows of the datasets are denoted as *samples* and its columns as *features*. We endow the samples $(\mathbf{x}_i)_{i \in [\![n]\!]}$ and $(\mathbf{x}'_i)_{i \in [\![n']\!]}$ with weights $\mathbf{w} = [w_1, \ldots, w_n]^\top \in \Sigma_n$ and $\mathbf{w}' = [w'_1, \ldots, w'_{n'}]^\top \in \Sigma_{n'}$ that both lie in the simplex so as to define empirical distributions supported on $(\mathbf{x}_i)_{i \in [\![n]\!]}$ and $(\mathbf{x}'_i)_{i \in [\![n']\!]}$. In addition to these distributions, we similarly associate weights given by vectors $\mathbf{v} \in \Sigma_d$ and $\mathbf{v}' \in \Sigma_{d'}$ with features. Note that when no additional information is available about the data, all the weights' vectors can be set as uniform.

We define the CO-Optimal Transport problem as follows:

$$\min_{\substack{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}') \\ \boldsymbol{\pi}^v \in \Pi(\mathbf{v}, \mathbf{v}')}} \quad \sum_{i,j,k,l} L(X_{i,k}, X'_{j,l}) \boldsymbol{\pi}^s_{i,j} \boldsymbol{\pi}^v_{k,l} = \min_{\substack{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}') \\ \boldsymbol{\pi}^v \in \Pi(\mathbf{v}, \mathbf{v}')}} \langle \mathbf{L}(\mathbf{X}, \mathbf{X}') \otimes \boldsymbol{\pi}^s, \boldsymbol{\pi}^v \rangle_{\mathcal{F}} \qquad \text{(COOT)}$$

where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a divergence measure between 1D variables, $\mathbf{L}(\mathbf{X}, \mathbf{X}')$ is the $d \times d' \times n \times n'$ tensor of all pairwise divergences between the elements of $\mathbf{X}$ and $\mathbf{X}'$, and $\Pi(\cdot, \cdot)$ is the set of linear transport constraints.

Note that problem (COOT) seeks for a simultaneous transport $\boldsymbol{\pi}^s$ between samples and a transport $\boldsymbol{\pi}^v$ between features across distributions. In the following, we write $\text{COOT}(\mathbf{X}, \mathbf{X}', \mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}')$ (or $\text{COOT}(\mathbf{X}, \mathbf{X}')$ when it is clear from the context) to denote the objective value of the optimization problem (COOT).

Entropic regularization Equation (COOT) can be also extended to the entropic regularized case favoured in the OT community for remedying the heavy computation burden of OT and reducing its sample complexity [Cuturi 2013, Altschuler 2017, Genevay 2019]. This leads to the following problem:

$$\min_{\substack{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}') \\ \boldsymbol{\pi}^v \in \Pi(\mathbf{v}, \mathbf{v}')}} \langle \mathbf{L}(\mathbf{X}, \mathbf{X}') \otimes \boldsymbol{\pi}^s, \boldsymbol{\pi}^v \rangle_{\mathcal{F}} + \Omega(\boldsymbol{\pi}^s, \boldsymbol{\pi}^v) \qquad (5.1)$$

where for $\varepsilon_1, \varepsilon_2 > 0$, the regularization term writes as $\Omega(\boldsymbol{\pi}^s, \boldsymbol{\pi}^v) = \varepsilon_1 H(\boldsymbol{\pi}^s | \mathbf{w}\mathbf{w}'^T) + \varepsilon_2 H(\boldsymbol{\pi}^v | \mathbf{v}\mathbf{v}'^T)$ with $H(\boldsymbol{\pi}^s | \mathbf{w}\mathbf{w}'^T) = \sum_{i,j} \log(\frac{\pi^s_{i,j}}{w_i w'_j}) \pi^s_{i,j}$ being the relative entropy. Note that similarly to OT [Cuturi 2013] and GW [Peyré 2016], adding the regularization term can lead to a more robust estimation of the transport matrices but prevents them from being sparse.
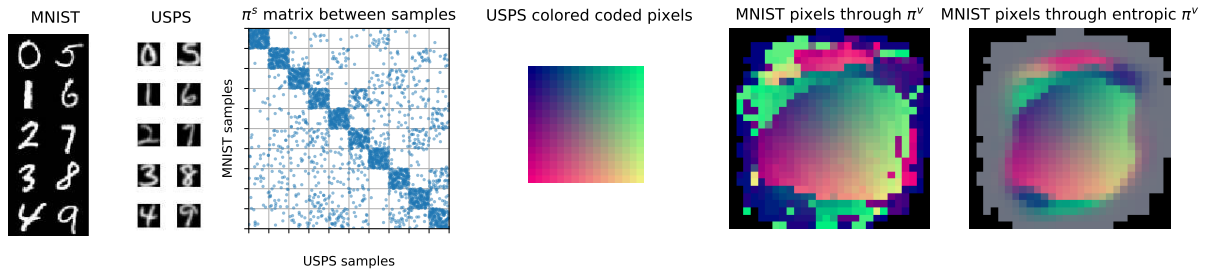
Figure 5.1: Illustration of COOT between MNIST and USPS datasets. **(left)** samples from MNIST and USPS data sets; **(center left)** Transport matrix π^s between samples sorted by class; **(center)** USPS image with pixels colored *w.r.t..* their 2D position; **(center right)** transported colors on MNIST image using π^v, black pixels correspond to non-informative MNIST pixels always at 0; **(right)** transported colors on MNIST image using π^v with entropic regularization.

5.2.1 Illustration of COOT

In order to illustrate our proposed COOT method and to explain the intuition behind it, we solve the optimization problem (COOT) using the algorithm described in Section 5.4 between two classical digit recognition datasets: MNIST and USPS. We choose these particular datasets for our illustration as they contain images of different resolutions (USPS is 16×16 and MNIST is 28×28) that belong to the same classes (digits between 0 and 9). Additionally, the digits are also slightly differently centered as illustrated on the examples in the left part of Figure 5.1. Altogether, this means that without specific pre-processing, the images do not lie in the same topological space and thus cannot be compared directly using conventional distances. We randomly select 300 images per class in each dataset, normalize magnitudes of pixels to $[0, 1]$ and consider digit images as *samples* while each pixel acts as a *feature* leading to 256 and 784 features for USPS and MNIST respectively. We use uniform weights for $\mathbf{w}, \mathbf{w}'$ and normalize average values of each pixel for $\mathbf{v}, \mathbf{v}'$ in order to discard non-informative ones that are always equal to 0.

The result of solving problem (COOT) is reported in Figure 5.1. In the center-left part, we provide the coupling π^s between the samples, *i.e* the different images, sorted by class and observe that 67% of mappings occur between the samples from the same class as indicated by block diagonal structure of the coupling matrix. The coupling π^v, in its turn, describes the relations between the features, *i.e* the pixels, in both domains. To visualize it, we color-code the pixels of the source USPS image and use π^v to transport the colors on a target MNIST image so that its pixels are defined as convex combinations of colors from the former with coefficients given by π^v. The corresponding results are shown in the right part of Figure 5.1 for both the original COOT and its entropic regularized counterpart. From these two images, we can observe that colored pixels appear only in the central areas and exhibit a strong spatial coherency despite the fact that the geometric structure of the image is totally unknown to the optimization problem, as each pixel is treated as an independent feature. COOT has recovered a meaningful spatial transformation between the two datasets in a completely unsupervised way, different from trivial rescaling of images that one may expect when aligning USPS digits occupying the full image space and MNIST digits lying in the middle of it.

For further evidence Figures 5.2 and 5.3 illustrate different images of both datasets obtained by transporting pixels from USPS (*resp.* MNIST) to MNIST (*resp.* USPS) using the optimal coupling π^v. Notably the case USPS $\rightarrow$ MNIST show that transporting the pixel through π^v leads to a better spatial coherency than a simple rescaling of the image.
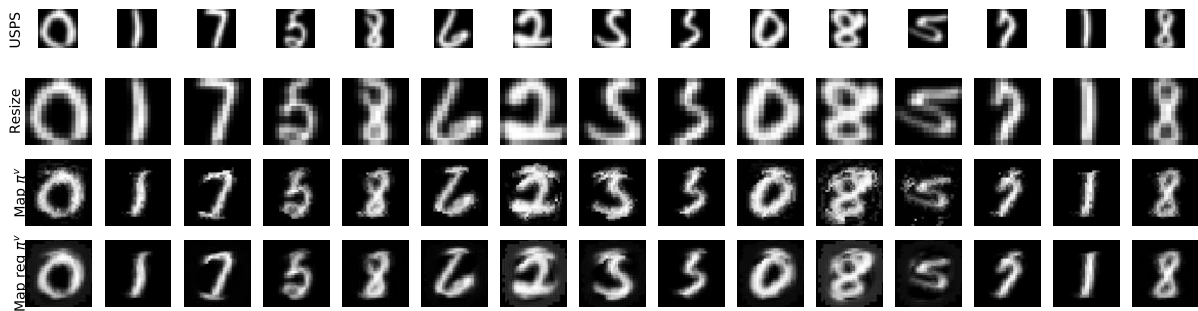
Figure 5.2: Linear mapping from USPS to MNIST using $\boldsymbol{\pi}^v$. **(First row)** Original USPS samples, **(Second row)** Samples resized to target resolution, **(Third row)** Samples mapped using $\boldsymbol{\pi}^v$, **(Fourth row)** Samples mapped using $\boldsymbol{\pi}^v$ with entropic regularization.
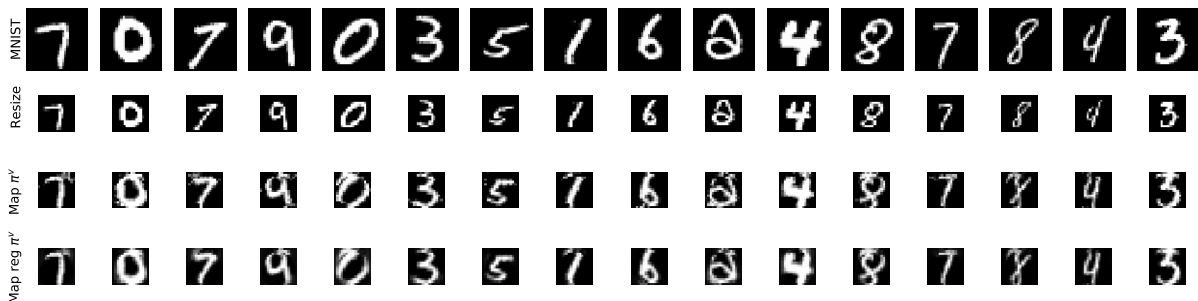
Figure 5.3: Linear mapping from MNIST to USPS using $\boldsymbol{\pi}^v$. **(First row)** Original MNIST samples, **(Second row)** Samples resized to target resolution, **(Third row)** Samples mapped using $\boldsymbol{\pi}^v$, **(Fourth row)** Samples mapped using $\boldsymbol{\pi}^v$ with entropic regularization.

5.3 Properties of COOT

COOT as a billinear program COOT is a special case of a Quadratic Program (QP) with linear constraints called a Bilinear Program (BP). More precisely, it is an indefinite BP problem [Gallo 1977]. It was proved (*e.g* in in [Pardalos 1987, Horst 1996]) that there exists an optimal solution lying on extremal points of the polytopes $\Pi(\mathbf{w}, \mathbf{w}')$ and $\Pi(\mathbf{v}, \mathbf{v}')$. When $n = n', d = d'$ and weights $\mathbf{w} = \mathbf{w}' = \frac{\mathbf{1}_n}{n}, \mathbf{v} = \mathbf{v}' = \frac{\mathbf{1}_d}{d}$ are uniform, Birkhoff's theorem [Birkhoff 1946] states that the set of extremal points of $\Pi(\frac{\mathbf{1}_n}{n}, \frac{\mathbf{1}_n}{n})$ and $\Pi(\frac{\mathbf{1}_d}{d}, \frac{\mathbf{1}_d}{d})$ are the set of permutation matrices so that there exists an optimal solution $(\boldsymbol{\pi}_*^s, \boldsymbol{\pi}_*^v)$ which transport maps are supported on two permutations $\sigma_*^s, \sigma_*^v \in \mathbb{S}_n \times \mathbb{S}_d$.

The BP problem is also related to the Bilinear Assignment Problem (BAP) where $\boldsymbol{\pi}^s$ and $\boldsymbol{\pi}^v$ are searched in the set of permutation matrices. The latter was shown to be NP-hard if $d = O(\sqrt[r]{n})$ for fixed r and solvable in polynomial time if $d = O(\sqrt{\log(n)})$ [Custic 2016]. In this case, we look for the best permutations of the rows and columns of our datasets that lead to the smallest cost. COOT provides a tight convex relaxation of the BAP by 1) relaxing the constraint set of permutations into the convex set of doubly stochastic matrices and 2) ensuring that two problems are equivalent, *i.e.*, one can always find a pair of permutations that minimizes (COOT), as explained in the paragraph above.

Finding a meaningful similarity measure between datasets is useful in many machine learning tasks as pointed out, *e.g* in [Alvarez-Melis 2020]. Interestingly enough, COOT induces a notion of distance between datasets $\mathbf{X}$ and $\mathbf{X}'$. More precisely it vanishes *iff* they are the same up to a permutation of rows

---

Algorithm 9 BCD for COOT

1: $\pi^s_{(0)} \leftarrow \mathbf{w}\mathbf{w}'^T, \pi^v_{(0)} \leftarrow \mathbf{v}\mathbf{v}'^T, k \leftarrow 0$

2: **while** $k <$ maxIt **and** $err > 0$ **do**

3:     $\boldsymbol{\pi}^s_{(k)} \leftarrow \arg\min_{\boldsymbol{\pi}^s \in \Pi(\mathbf{w},\mathbf{w}')} \langle \mathbf{L}(\mathbf{X},\mathbf{X}') \otimes \boldsymbol{\pi}^v_{(k-1)}, \boldsymbol{\pi}^s \rangle_{\mathcal{F}}$ // linear OT problem on the samples

4:     $\boldsymbol{\pi}^v_{(k)} \leftarrow \arg\min_{\boldsymbol{\pi}^v \in \Pi(\mathbf{v},\mathbf{v}')} \langle \mathbf{L}(\mathbf{X},\mathbf{X}') \otimes \boldsymbol{\pi}^s_{(k-1)}, \boldsymbol{\pi}^v \rangle_{\mathcal{F}}$ // linear OT problem on the features

5:     $err \leftarrow ||\boldsymbol{\pi}^v_{(k-1)} - \boldsymbol{\pi}^v_{(k)}||_F$

6:     $k \leftarrow k+1$

7: **end while**

---

and columns as established below:

Proposition 5.3.1 (COOT is a distance). *Suppose $L = |\cdot|^p, p \geq 1$, $n = n', d = d'$ and that the weights $\mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}'$ are uniform. Then $COOT(\mathbf{X},\mathbf{X}') = 0$ iff there exists a permutation of the samples $\sigma_1 \in S_n$ and of the features $\sigma_2 \in S_d$, s.t, $\forall i, k$ $\mathbf{X}_{i,k} = \mathbf{X}'_{\sigma_1(i),\sigma_2(k)}$. Moreover, it is symmetric and satisfies the triangular inequality as long as L satisfies the triangle inequality, i.e., $COOT(\mathbf{X},\mathbf{X}'') \leq COOT(\mathbf{X},\mathbf{X}') + COOT(\mathbf{X}',\mathbf{X}'')$.*

Note that in the general case when $n \neq n', d \neq d'$, positivity and triangle inequality still hold but $COOT(\mathbf{X},\mathbf{X}') > 0$. The proof can be found in Section 6.3. Interestingly, our result generalizes the metric property proved in [Faliszewski 2019] for the election isomorphism problem with this latter result being valid only for the BAP case (for a discussion on the connection between COOT and the work of [Faliszewski 2019], see Section 6.3.7). Finally, we note that this metric property means that COOT can be used as a divergence in a large number of potential applications as, for instance, in generative learning [Bunne 2019].

5.4 Optimization algorithm and complexity

Even though solving COOT exactly is NP-hard, in practice computing a solution can be done rather efficiently. To this end, we propose to use Block Coordinate Descent (BCD) that consists in iteratively solving the problem for $\boldsymbol{\pi}^s$ or $\boldsymbol{\pi}^v$ with the other kept fixed. Interestingly, this boils down to solving at each step a linear OT problem that requires $O(n^3 \log(n))$ operations with a network simplex algorithm as detailed in the pseudo-code given in Algorithm 9. This approach, also known as the "mountain climbing procedure" [Konno 1976a] in the BP literature, was proved to decrease the loss at each iteration and so to converge within a finite number of iterations [Horst 1996]. We also note that at each iteration one needs to compute the equivalent cost matrix $L(\mathbf{X},\mathbf{X}') \otimes \boldsymbol{\pi}^{(\cdot)}$ which has a complexity of $O(ndn'd')$. However, one can reduce it using Proposition 1 from [Peyré 2016] for the case when L is the squared Euclidean distance $|\cdot|^2$ or the Kullback-Leibler divergence. In this case, the overall computational complexity becomes $O(\min\{(n+n')dd' + n'^2n; (d+d')nn' + d'^2d\})$. We refer the interested reader to Section 6.3.2 for further details.

Finally, we can use the same BCD procedure for the entropic regularized version of COOT (5.1) where each iteration an entropic regularized OT problem can be solved efficiently using Sinkhorn's algorithm [Cuturi 2013] with several possible improvements [Altschuler 2017, Altschuler 2019, Alaya 2019]. Note that this procedure can be easily adapted in the same way to include unbalanced OT problems [Chizat 2017] as well.

5.5 Relation with other OT distances

5.5.1 Gromov-Wasserstein

The COOT problem is defined for arbitrary matrices $\mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{X}' \in \mathbb{R}^{n' \times d'}$ and so can be readily used to compare pairwise similarity matrices between the samples $\mathbf{C} = (c(\mathbf{x}_i, \mathbf{x}_j)_{i,j}) \in \mathbb{R}^{n \times n}, \mathbf{C}' = (c'(\mathbf{x}'_k, \mathbf{x}'_l))_{k,l} \in \mathbb{R}^{n' \times n'}$ for some c, c'. To avoid redundancy, we use the term "similarity" for both similarity and distance functions in what follows. This situation arises in applications dealing with relational data, *e.g*, in a graph context (see Chapter 3) or *e.g.* in deep metric alignment [Ezuz 2017]. These problems have been successfully tackled using the Gromov-Wasserstein distance (see Chapter 2). We recall that given $\mathbf{C} \in \mathbb{R}^{n \times n}$ and $\mathbf{C}' \in \mathbb{R}^{n' \times n'}$, the GW distance is defined by:

$$GW(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}') = \min_{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}')} \langle \mathbf{L}(\mathbf{C}, \mathbf{C}') \otimes \boldsymbol{\pi}^s, \boldsymbol{\pi}^s \rangle_{\mathcal{F}} \tag{5.2}$$

As suggested by the similar objective functions and constraints, GW and COOT are linked in multiple ways. Below, we explicit the link between GW and COOT using a reduction of a concave QP to an associated BP problem established in [Konno 1976b] and show that they are equivalent when working with squared Euclidean distance matrices $\mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{C}' \in \mathbb{R}^{n' \times n'}$ or with inner product similarities (see Chapter 4). More precisely, this latter equivalence follows from [Konno 1976b] where it was shown that a concave QP can be solved by a reduction to an associated BP problem:

Theorem 5.5.1 (Adapted from [Konno 1976b]). *If* $\mathbf{Q}$ *is a negative definite matrix then problems:*

$$\begin{aligned} \min_{\mathbf{x}} f(\mathbf{x}) &= \mathbf{c}^T x + \tfrac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x} \\ s.t. \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \geq 0 \end{aligned} \tag{5.3}$$

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{y}} g(\mathbf{x}, \mathbf{y}) &= \tfrac{1}{2}\mathbf{c}^T \mathbf{x} + \tfrac{1}{2}\mathbf{c}^T \mathbf{y} + \tfrac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{y} \\ s.t. \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{A}\mathbf{y} = \mathbf{b}, \ \mathbf{x}, \mathbf{y} \geq 0 \end{aligned} \tag{5.4}$$

are equivalent. More precisely, if $\mathbf{x}^*$ *is an optimal solution for* (5.3), *then* $(\mathbf{x}^*, \mathbf{x}^*)$ *is a solution for* (5.4) *and if* $(\mathbf{x}^*, \mathbf{y}^*)$ *is optimal for* (5.4), *then both* $\mathbf{x}^*$ *and* $\mathbf{y}^*$ *are optimal for* (5.3).

Using this principle one can link GW with the COOT problem when working on intra domain similarity matrices $\mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{C}' \in \mathbb{R}^{n' \times n'}$ thanks to the next proposition:

Proposition 5.5.1. *Let* $L = |\cdot|^2$ *and suppose that* $\mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{C}' \in \mathbb{R}^{n' \times n'}$ *are squared Euclidean distance matrices such that* $\mathbf{C} = \mathbf{x}\mathbf{1}_n^T + \mathbf{1}_n\mathbf{x}^T - 2\mathbf{X}\mathbf{X}^T, \mathbf{C}' = \mathbf{x}'\mathbf{1}_{n'}^T + \mathbf{1}_{n'}\mathbf{x}'^T - 2\mathbf{X}'\mathbf{X}'^T$ *with* $\mathbf{x} = diag(\mathbf{X}\mathbf{X}^T), \mathbf{x}' = diag(\mathbf{X}'\mathbf{X}'^T)$. *Then, the GW problem can be written as a concave quadratic program (QP) which Hessian reads* $\mathbf{Q} = -4 * \mathbf{X}\mathbf{X}^T \otimes_K \mathbf{X}'\mathbf{X}'^T$.

   If $\mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{C}' \in \mathbb{R}^{n' \times n'}$ *are inner products similarities, i.e. such that* $\mathbf{C} = \mathbf{X}\mathbf{X}^T, \mathbf{C}' = \mathbf{X}'\mathbf{X}'^T$ *then the GW is also a concave quadratic program (QP) which Hessian reads* $\mathbf{Q} = -2 * \mathbf{X}\mathbf{X}^T \otimes_K \mathbf{X}'\mathbf{X}'^T$.

When working with arbitrary similarity matrices, COOT provides a lower-bound for GW and using Proposition 5.5.1 we can prove that both problems become equivalent for the cases of squared Euclidean distances and inner product similarities.

Proposition 5.5.2. *Let* $\mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{C}' \in \mathbb{R}^{n' \times n'}$ *be any symmetric matrices, then:*

$$COOT(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}', \mathbf{w}, \mathbf{w}') \leq GW(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}').$$

Algorithm 10 DC Algorithm for solving COOT and GW with squared Euclidean matrices or inner product similarities

---

1: $\boldsymbol{\pi}_{(0)} \leftarrow \mathbf{w}\mathbf{w}'^T, k \leftarrow 0$

2: **while** $k < \text{maxIt}$ **and** $err > 0$ **do**

3:      $\boldsymbol{\pi}_{(k)} \leftarrow \arg\min_{\boldsymbol{\pi} \in \Pi(\mathbf{w}, \mathbf{w}')} \langle \mathbf{L}(\mathbf{C}, \mathbf{C}') \otimes \boldsymbol{\pi}_{(k-1)}, \boldsymbol{\pi} \rangle_{\mathcal{F}}$ // linear OT problem

4:      $err \leftarrow ||\boldsymbol{\pi}_{(k-1)} - \boldsymbol{\pi}_{(k)}||_F$

5:      $k \leftarrow k + 1$

6: **end while**

7: **return** $\boldsymbol{\pi}_{(k)}$ for GW and $(\boldsymbol{\pi}_{(k)}, \boldsymbol{\pi}_{(k)})$ for COOT

The converse is also true under the hypothesis of Proposition 5.5.1. In this case, if $(\boldsymbol{\pi}_*^s, \boldsymbol{\pi}_*^v)$ is an optimal solution of (COOT)*, then both $\boldsymbol{\pi}_*^s, \boldsymbol{\pi}_*^v$ are solutions of* (5.2)*. Conversely, if $\boldsymbol{\pi}_*^s$ is an optimal solution of* (5.2)*, then $(\boldsymbol{\pi}_*^s, \boldsymbol{\pi}_*^s)$ is an optimal solution for* (COOT) *.*

Equivalence of algorithms   Under the hypothesis of Proposition 5.5.1 we know that there exists an optimal solution for the COOT problem of the form $(\boldsymbol{\pi}_*, \boldsymbol{\pi}_*)$, where $\boldsymbol{\pi}_*$ is an optimal solution of the GW problem. This gives a conceptually very simple fixed-point procedure where one only iterates over one coupling in order to compute a optimal solution of GW as described in Algorithm 10. Interestingly enough, the iterations of the fixed point method are exactly equivalent to the Frank Wolfe procedure described in Chapter 3, since, in the concave setting, the line search step can be fixed to 1 [Maron 2018] (see Section 6.3.6 for more details). Also note that the steps of Algorithm 10 are iterations of Difference of Convex Algorithm (DCA) [Tao 2005, Yuille 2003] where the concave function is approximated at each iteration by its linear majorization. When applying the same procedure for entropic regularized COOT, the resulting DCA also recovers exactly the projected gradients iterations proposed in [Peyré 2016] for solving the entropic regularized version of GW.

We would like to stress out that COOT is much more than a generalization of GW and that is for multiple reasons. First, it can be used on raw data without requiring to choose or compute the similarity matrices, that can be costly, for instance, when dealing with shortest path distances in graphs, and to store them ($O(n^2 + n'^2)$ overhead). Second, it can take into account additional information given by feature weights $\mathbf{v}, \mathbf{v}'$ and provides an interpretable mapping between them across two heterogeneous datasets. Finally, contrary to GW, COOT is not invariant neither to feature rotations nor to the change of signs leading to a more informative samples' coupling when compared to GW in some applications. One such example is given in the previous MNIST-USPS transfer task (Figure 5.1) for which the coupling matrix obtained via GW (given in Figure 5.4) exhibits important flaws in respecting class memberships when aligning samples.

5.5.2   Relation with Invariant OT and hierarchical approaches

In [Alvarez-Melis 2019], authors consider a scenario where the OT problem is used to align measures supported on sets of points for which meaningful pairwise distances are hard or impossible to calculate. This may happen, for instance, due to some latent transformation that have been applied to the features. The main underlying idea of their approach is to find an assignment of the points and to calculate a transformation to match the features. More precisely, for two datasets $\mathbf{X}, \mathbf{X}'$, with same feature space $\mathbb{R}^d$,
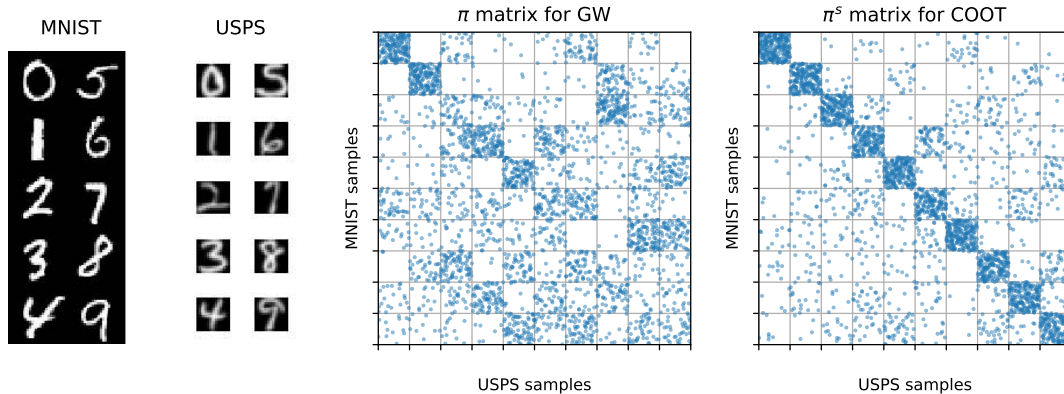
Figure 5.4: Comparison between the coupling matrices obtained via GW and COOT on MNIST-USPS.

the corresponding objective function is:

$$InvOT(\mathbf{X}, \mathbf{X}') = \min_{\boldsymbol{\pi} \in \Pi(\mathbf{w}, \mathbf{w}')} \min_{\mathbf{P} \in F_d} \langle \mathbf{M_P}, \boldsymbol{\pi} \rangle_{\mathcal{F}} \tag{5.5}$$

where $M_{\mathbf{P}}(i,j) = \|\mathbf{x}_i - \mathbf{P}\mathbf{x}'_j\|_2^2$ and F_d is a space of matrices $F_d = \{\mathbf{P} \in \mathbb{R}^{d \times d} | \ \|\mathbf{P}\|_{\mathcal{F}} = \sqrt{d}\}$. As noted by the authors in Lemma 4.3, equation (5.5) can be related to the GW problem when $\mathbf{C}, \mathbf{C}'$ are calculated using inner product similarities and when $\mathbf{X}'$ is $\mathbf{w}'$-whitened, *i.e* $\mathbf{X}'^T \text{diag}(\mathbf{w}')\mathbf{X}' = \mathbf{I_{d'}}$. In this case, author show that GW and InvOT are equivalent, namely a solution of GW is a solution of *InvOT* and conversely. Since the GW problem with cosine similarities is actually concave we have proven COOT and GW are also equivalent in this case which proves the following proposition:

Proposition 5.5.3. *Using previous notations, $L = |\cdot|^2$, $d = d'$, and inner product similarities $\mathbf{C} = \mathbf{X}\mathbf{X}^T, \mathbf{C}' = \mathbf{X}'\mathbf{X}'^T$. Suppose that $\mathbf{X}'$ is $\mathbf{w}'$-whitened, i.e $\mathbf{X}'^T \text{diag}(\mathbf{w}')\mathbf{X}' = \mathbf{I_{d'}}$. Then, $InvOT(\mathbf{X}, \mathbf{X}')$, $COOT(\mathbf{C}, \mathbf{C}')$ and $GW(\mathbf{C}, \mathbf{C}')$ are equivalent, namely any optimal coupling of one of this problem is a solution to others problems.*

Another way of proving this result is to consider the Theorem 4.2.1 of Chapter 4 where we proved that the Gromov-Wasserstein distance, when considering inner product similarities, is equivalent to the problem:

$$\min_{\boldsymbol{\pi} \in \Pi(\mathbf{w}, \mathbf{w}')} \min_{\mathbf{P} \in F_d} \sum_{i=1, j=1}^{n, n'} \langle \mathbf{x}_i, \mathbf{P}\mathbf{x}'_j \rangle \pi_{ij} \tag{MaxOT}$$

When $\mathbf{X}'$ is $\mathbf{w}'$-whitened we can check easily by developing the terms in $M_{\mathbf{P}}(i,j) = \|\mathbf{x}_i - \mathbf{P}\mathbf{x}'_j\|_2^2$ that *InvOT* is equivalent to (MaxOT) (see Chapter 4). InvOT was further used as a building block for aligning clustered datasets in [Lee 2019] where the authors applied it as a divergence measure between the clusters, thus leading to an approach different from ours. Finally, in [Yurochkin 2019] the authors proposed a hierarchical OT distance as an OT problem with costs defined based on precomputed Wasserstein distances but with no global features' mapping, contrary to COOT that optimises two couplings of the features and the samples simultaneously.

5.5.3 Election isomorphism

The election isomorphism problem mentioned earlier has recently been introduced in [Faliszewski 2019] to compare two elections given by preference orders for candidates of voters. The authors express their

problem quite similarly to COOT and also seek for correspondences between voters and candidates across two elections where the preferences of each voter are known (which is unrealistic in modern democracies). They focus on the setting where both elections have exactly the same number of voters $n = n'$ and candidates $d = d'$ and search for an optimal permutation *via* a Linear Integer Program. It is interesting to see that their problem consisting in aligning voters using the Spearman distance is actually equivalent to solving COOT with $L = |\cdot|$ on voters preferences. To that extent, COOT is a more general approach as it is applicable for general loss functions L, contrary to the Spearman distance used in [Faliszewski 2019], and generalizes to the cases where $n \neq n'$ and $m \neq m'$. A more detailed comparison is given in Section 6.3.7.

5.6 Experiments

In the next section, we highlight two possible applications of COOT in a machine learning context: HDA and co-clustering. We consider these two particular tasks because 1) OT-based methods are considered as a strong baseline in DA; 2) COOT is a natural match for co-clustering as it allows for soft assignments of data samples and features to co-clusters.

5.6.1 Heterogeneous domain adaptation

In a classification context, the problem of domain adaptation (DA) arises whenever one has to perform classification on a set of data $\mathbf{X}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ (usually called the target domain) but has only few or no labelled data associated. Given a source domain $\mathbf{X}_s = \{\mathbf{x}_i^s\}_{i=1}^{N_s}$ with associated labels $\mathbf{Y}_s = \{\mathbf{y}_i^s\}_{i=1}^{N_s}$, one would like to leverage on this knowledge to train a classifier in the target domain. Unfortunately, direct use of the source information usually leads to poor results because of the discrepancy between source and target distributions. Among others, several works, e.g. [Courty 2017], use OT to perform this adaptation. However, in the case where the data do not belong to the same metric space ($\mathbf{X}_s \in \mathbb{R}^{N_s \times d}$ and $\mathbf{X}_t \in \mathbb{R}^{N_t \times d'}$ with $d \neq d'$), the problem is getting harder as domains probability distributions can not be anymore compared or aligned in a straightforward way. This instance of the DA problem, known as *Heterogeneous Domain Adaptation* (HDA), has received less attention in the literature, partly due to the lack of appropriate divergence measures that can be used in such context. State-of-the-art HDA methods include Canonical Correlation Analysis [Yeh 2014] and its kernelized version and a more recent approach based on the Gromov-Wasserstein discrepancy [Yan 2018]. Usually, one considers a *semi-supervised* variant of the problem, where one has access to a small number n_t of labelled samples per class in the target domain, because the *unsupervised* problem ($n_t = 0$) is much more difficult. We investigate here the use of COOT for both *semi-supervised* HDA, where one has access to a small number n_t of labelled samples per class in the target domain and *unsupervised* HDA with $n_t = 0$.

Solving HDA with COOT In order to solve the HDA problem, we compute $\text{COOT}(\mathbf{X}_s, \mathbf{X}_t)$ between the two domains and use the $\boldsymbol{\pi}^s$ matrix providing a transport/correspondence between samples (as illustrated in Figure 5.1) to estimate the labels in the target domain via label propagation [Redko 2019]. Assuming uniform sample weights and one-hot encoded labels, a class prediction $\hat{\mathbf{Y}}_t$ in the target domain samples can be obtained by computing $\hat{\mathbf{Y}}_t = \boldsymbol{\pi}^s \mathbf{Y}_s$. When labelled target samples are available, we further prevent source samples to be mapped to target samples from a different class by adding a high cost in the cost matrix for every such source sample as suggested in [Courty 2017, Section 4.2].

Domains	No-adaptation baseline	CCA	KCCA	EGW	SGW	COOT
C→W	69.12±4.82	11.47±3.78	66.76±4.40	11.35±1.93	_78.88_±3.90	**83.47**±2.60
W→C	83.00±3.95	19.59±7.71	76.76±4.70	11.00±1.05	_92.41_±2.18	**93.65**±1.80
W→W	82.18±3.63	14.76±3.15	78.94±3.94	10.18±1.64	_93.12_±3.14	**93.94**±1.84
W→A	84.29±3.35	17.00±12.41	78.94±6.13	7.24±2.78	_93.41_±2.18	**94.71**±1.49
A→C	_83.71_±1.82	15.29±3.88	76.35±4.07	9.82±1.37	80.53±6.80	**89.53**±2.34
A→W	81.88±3.69	12.59±2.92	81.41±3.93	12.65±1.21	_87.18_±5.23	**92.06**±1.73
A→A	_84.18_±3.45	13.88±2.88	80.65±3.03	14.29±4.23	82.76±6.63	**92.12**±1.79
C→C	67.47±3.72	13.59±4.33	60.76±4.38	11.71±1.91	_77.59_±4.90	**83.35**±2.31
C→A	66.18±4.47	13.71±6.15	63.35±4.32	11.82±2.58	_75.94_±5.58	**82.41**±2.79
Mean	78.00±7.43	14.65±2.29	73.77±7.47	11.12±1.86	_84.65_±6.62	**89.47**±4.74
p-value	<.001	<.001	<.001	<.001	<.001	-

Table 5.1: **Semi-supervised HDA** for $n_t = 3$ from Decaf to GoogleNet task.

Datasets We choose to test our method on the classical Caltech-Office dataset [Saenko 2010], which is dedicated to object recognition in images from several domains. Those domains exhibit variability in term of presence/absence of background, lightning conditions, image quality, that as such induce distribution shifts between the domains. Among the available domains, we select the following three: *Amazon* (A), the *Caltech-256* image collection (C) and *Webcam* (W). Ten overlapping classes between the domains are used and two different deep feature representations of image in each domain are obtained using two different neural networks, namely: the Decaf [Donahue 2014] and GoogleNet [Szegedy 2015] neural network architectures. In both cases, we extract the image representations as the activations of the last fully-connected layer, yielding respectively sparse 4096 and 1024 dimensional vectors. The heterogeneity comes from these two very different representations.

Competing methods and experimental settings We evaluate COOT on *Amazon* (A), *Caltech-256* (C) and *Webcam* (W) domains from Caltech-Office dataset [Saenko 2010] with 10 overlapping classes between the domains and two different deep feature representations obtained for images from each domain using the Decaf [Donahue 2014] and GoogleNet [Szegedy 2015] neural network architectures. In both cases, we extract the image representations as the activations of the last fully-connected layer, yielding respectively sparse 4096 and 1024 dimensional vectors. The heterogeneity comes from these two very different representations. We consider 4 baselines: CCA, its kernalized version KCCA [Yeh 2014] with a Gaussian kernel which width parameter is set to the inverse of the dimension of the input vector, EGW representing the entropic version of GW and SGW [Yan 2018] that incorporates labelled target data into two regularization terms. For EGW and SGW, the entropic regularization term was set to 0.1, and the two other regularization hyperparameters for the semi-supervised case to $\lambda = 10^{-5}$ and $\gamma = 10^{-2}$ as done in [Yan 2018]. We use COOT with entropic regularization on the feature mapping, with parameter $\varepsilon_2 = 1$ in all experiments. For all OT methods, we use label propagation to obtain target labels as the maximum entry of $\hat{\mathbf{Y}}_t$ in each row. For all non-OT methods, classification was conducted with a k-nn classifier with $k = 3$. We run the experiment in a semi-supervised setting with $n_t = 3$, *i.e.*, 3 samples per class were labelled in the target domain. The baseline score is the result of classification by only considering labelled samples in the target domain as the training set. For each pair of domains, we selected 20 samples per class to form the learning sets. We run this random selection process 10 times and consider the mean accuracy of the different runs as a performance measure.

Domains	CCA	KCCA	EGW	COOT
C→W	14.20±8.60	21.30±15.64	10.55±1.97	**25.50**±11.76
W→C	13.35±3.70	18.60±9.44	10.60±0.94	**35.40**±14.61
W→W	10.95±2.36	13.25±6.34	10.25±2.26	**37.10**±14.57
W→A	14.25±8.14	23.00±22.95	9.50±2.47	**34.25**±13.03
A→C	11.40±3.23	11.50±9.23	11.35±1.38	**17**.40±8.86
A→W	19.65±17.85	28.35±26.13	11.60±1.30	**30.95**±18.19
A→A	11.75±1.82	14.20±4.78	13.10±2.35	**42.85**±17.65
C→C	12.00±4.69	14.95±6.79	12.90±1.46	**42.85**±18.44
C→A	15.35±6.30	23.35±17.61	12.95±2.63	**33.25**±15.93
Mean	13.66±2.55	18.72±5.33	11.42±1.24	**33.28**±7.61
p-value	<.001	<.001	<.001	-

Table 5.2: **Unsupervised HDA** for $n_t = 0$ from Decaf to GoogleNet task.

Results   We first provide in Table 5.1 the results for the semi-supervised case where we perform adaptation from Decaf to GoogleNet features. Note that we report the results in the opposite direction in Section 6.3.8 for $n_t \in \{0, 1, 3, 5\}$. From it, we see that COOT surpasses all the other state-of-the-art methods in terms of mean accuracy. This result is confirmed by a p-value lower than 0.001 on a pairwise method comparison with COOT in a Wilcoxon signed rank test. SGW provides the second best result, while CCA and EGW have a less than average performance. Finally, KCCA performs better than the two latter methods, but still fails most of the time to surpass the no-adaptation baseline score given by a classifier learned on the available labelled target data. Results for the unsupervised case can be found in Table 5.2. This setting is rarely considered in the literature as unsupervised HDA is regarded as a very difficult problem. In this table, we do not provide scores for the no-adaptation baseline and SGW, as they require labelled data.

As one can expect, most of the methods fail in obtaining good classification accuracies in this setting, despite having access to discriminant feature representations. Yet, COOT succeeds in providing a meaningful mapping in some cases. The overall superior performance of COOT highlights its strengths and underlines the limits of other HDA methods. First, COOT does not depend on approximating empirical quantities from the data, contrary to CCA and KCCA that rely on the estimation of the cross-covariance matrix that is known to be flawed for high-dimensional data with few samples [Song 2016]. Second, COOT takes into account the features of the raw data that are more informative than the pairwise distances used in EGW. Finally, COOT avoids the sign invariance issue discussed previously that hinders GW's capability to recover classes without supervision as illustrated for the MNIST-USPS problem before.

5.6.2   Co-clustering and data summarization

While clustering methods present an important discovery tool for data analysis, one of their main limations is to completely discard the potential relationships that may exist between the features that describe the data samples. For instance, in recommendation systems, where each user is described in terms of his or her preferences for some product, clustering algorithms may benefit from the knowledge about the correlation between different products revealing their probability of being recommended to the same users.

Data set	$n \times d$	$g \times m$	Overlapping	Proportions
D1	600×300	3×3	[+]	Equal
D2	600×300	3×3	[+]	Unequal
D3	300×200	2×4	[++]	Equal
D4	300×300	5×4	[++]	Unequal

Table 5.3: Size ($n \times d$), number of co-clusters ($g \times m$), degree of overlapping ([+] for well-separated and [++] for ill-separated co-clusters) and the proportions of co-clusters for simulated data sets.

This idea is the cornerstone of *co-clustering* [Hartigan 1972] where the goal is to perform clustering of both samples and features simultaneously. More precisely given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the number of samples (rows) and features (columns) clusters denoted by $g \leq n$ and $m \leq d$, respectively, we seek to find $\mathbf{X}_c \in \mathbb{R}^{g \times m}$ that summarizes $\mathbf{X}$ in the best way possible.

COOT-clustering We look for $\mathbf{X}_c$ which is as close as possible to the original $\mathbf{X}$ *w.r.t.* COOT by solving:

$$\min_{\mathbf{X}_c} \text{COOT}(\mathbf{X}, \mathbf{X}_c) = \min_{\boldsymbol{\pi}^s, \boldsymbol{\pi}^v, \mathbf{X}_c} \langle \mathbf{L}(\mathbf{X}, \mathbf{X}_c) \otimes \boldsymbol{\pi}^s, \boldsymbol{\pi}^v \rangle_{\mathcal{F}} \tag{5.6}$$

with potentially entropic regularization. More precisely, we set $\mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}'$ as uniform, initialize $\mathbf{X}_c$ with random values and apply the BCD algorithm over $(\boldsymbol{\pi}^s, \boldsymbol{\pi}^v, \mathbf{X}_c)$ by alternating between the following steps:

1. Obtain $\boldsymbol{\pi}^s$ and $\boldsymbol{\pi}^v$ by solving $\text{COOT}(\mathbf{X}, \mathbf{X}_c)$

2. Set $\mathbf{X}_c$ to $gm\boldsymbol{\pi}^{s\top}\mathbf{X}\boldsymbol{\pi}^v$.

This second step of the procedure is a least-square estimation when $L = |\cdot|^2$ and corresponds to minimizing the COOT objective *w.r.t.* $\mathbf{X}_c$. In practice, we observed that few iterations of this procedure are enough to ensure the convergence. Once solved, we use the soft assignments provided by coupling matrices $\boldsymbol{\pi}^s \in \mathbb{R}^{n \times g}, \boldsymbol{\pi}^v \in \mathbb{R}^{d \times m}$ to partition data points and features to clusters by taking the index of the maximum element in each row of $\boldsymbol{\pi}^s$ and $\boldsymbol{\pi}^v$, respectively.

Simulated data We follow [Laclau 2017] where four scenarios with different number of co-clusters, degrees of separation and sizes were considered (for details, see the supplementary materials). We choose to evaluate COOT on simulated data as it provides us with the ground-truth for feature clusters that are often unavailable for real-world data sets. As in [Laclau 2017], we use the same co-clustering baselines including ITCC [Dhillon 2003], Double K-Means (DKM) [Rocci 2008], Orthogonal Nonnegative Matrix Tri-Factorizations (ONTMF) [Ding 2006], the Gaussian Latent Block Models (GLBM) [Nadif 2008] and Residual Bayesian Co-Clustering (RBC) [Shan 2010] as well as the K-means and NMF run on both modes of the data matrix, as clustering baseline. The performance of all methods is measured using the co-clustering error (CCE) defined as follows [Patrikainen 2006]:

$$\text{CCE}((\mathbf{z}, \mathbf{w}), (\hat{\mathbf{z}}, \hat{\mathbf{w}})) = e(\mathbf{z}, \hat{\mathbf{z}}) + e(\mathbf{w}, \hat{\mathbf{w}}) - e(\mathbf{z}, \hat{\mathbf{z}}) \times e(\mathbf{w}, \hat{\mathbf{w}}) \tag{5.7}$$

where $\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$ are the partitions of samples and features estimated by the algorithm; $\mathbf{z}$ and $\mathbf{w}$ are the true partitions and $e(\mathbf{z}, \hat{\mathbf{z}})$ (resp. $e(\mathbf{w}, \hat{\mathbf{w}})$) denotes the error rate, i.e., the proportion of misclassified instances (resp. features).

For all configurations, we generate 100 data sets and present the mean and standard deviation of the CCE over all sets for all baselines in Table 5.4. Table 5.3 below summarizes the characteristics of the simulated data sets used in our experiment.

Data set	Algorithms									
	K-means	NMF	DKM	Tri-NMF	GLBM	ITCC	RBC	CCOT	CCOT-GW	COOT
D1	$.018 \pm .003$	$.042 \pm .037$	$.025 \pm .048$	$.082 \pm .063$	$.021 \pm .011$	$.021 \pm .001$	$.017 \pm .045$	$.018 \pm .013$	$.004 \pm .002$	**0**
D2	$.072 \pm .044$	$.083 \pm .063$	$.038 \pm .000$	$.052 \pm .065$	$.032 \pm .041$	$.047 \pm .042$	$.039 \pm .052$	$.023 \pm .036$	$.011 \pm .056$	**$.009 \pm 0.04$**
D3	–	–	$.310 \pm .000$	–	$.262 \pm .022$	$.241 \pm .031$	–	$.031 \pm .027$	**$.008 \pm .001$**	$.04 \pm .05$
D4	$.126 \pm .038$	–	$.145 \pm .082$	–	$.115 \pm .047$	$.121 \pm .075$	$.102 \pm .071$	$.093 \pm .032$	$.079 \pm .031$	**0.068 ± 0.04**

Table 5.4: Mean ($\pm$ standard-deviation) of the co-clustering error (CCE) obtained for all configurations. "-" indicates that the algorithm cannot find a partition with the requested number of co-clusters. All the baselines results (first 9 columns) are from [Laclau 2017].
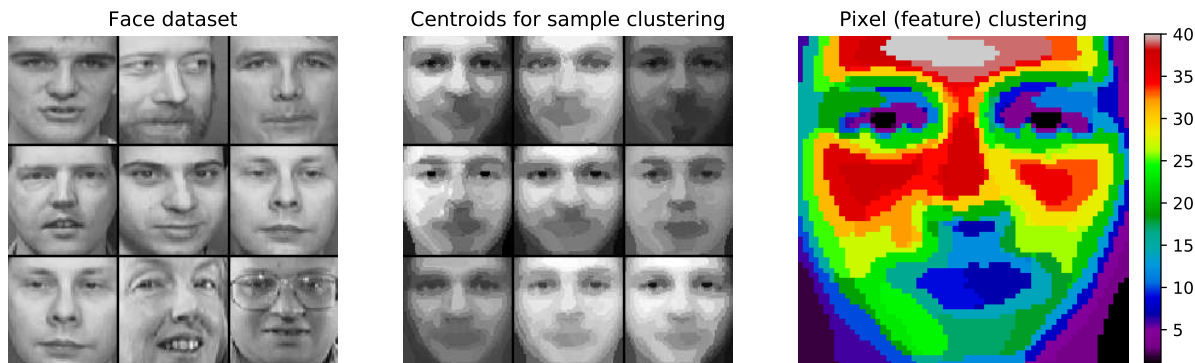
Figure 5.5: Co-clustering with COOT on the Olivetti faces dataset. **(left)** Example images from the dataset, **(center)** centroids estimated by COOT **(right)** clustering of the pixels estimated by COOT where each color represents a cluster.

Based on these results, we see that our algorithm outperforms all the other baselines on D1, D2 and D4 data sets, while being behind (CCOT-GW) proposed by [Laclau 2017] on D3. This result is rather strong as our method relies on the original data matrix, while (CCOT-GW) relies on its kernel representation and thus benefits from the non-linear information captured by it. Finally, we note that while both competing methods rely on OT, they remain very different as (CCOT-GW) approach is based on detecting the positions and the number of jumps in the scaling vectors of GW entropic regularized solution, while our method relies on coupling matrices to obtain the partitions.

Olivetti Face dataset As a first application of COOT for the co-clustering problem on real data, we propose to run the algorithm on the well known Olivetti faces dataset [Samaria 1994].

We take 400 images normalized between 0 and 1 and run our algorithm with $g = 9$ image clusters and $m = 40$ feature (pixel) clusters. As before, we consider the empirical distributions supported on images and features, respectively. The resulting reconstructed image's clusters are given in Figure 5.5 and the pixel clusters are illustrated in its rightmost part. We can see that despite the high variability in the data set, we still manage to recover detailed centroids, whereas L2-based clustering such as standard NMF or k-means based on ℓ_2 norm cost function are known to provide blurry estimates in this case. Finally, as in the MNIST-USPS example, COOT recovers spatially localized pixel clusters with no prior information about the pixel relations.

MovieLens We now evaluate our approach on the benchmark MovieLens-100K[1] data set that provides 100,000 user-movie ratings, on a scale of one to five, collected from 943 users on 1682 movies. The

---

[1] https://grouplens.org/datasets/movielens/100k/

M1	M20
Shawshank Redemption (1994)	Police Story 4: Project S (Chao ji ji hua) (1993)
Schindler's List (1993)	Eye of Vichy, The (Oeil de Vichy, L') (1993)
Casablanca (1942)	Promise, The (Versprechen, Das) (1994)
Rear Window (1954)	To Cross the Rubicon (1991)
Usual Suspects, The (1995)	Daens (1992)

Table 5.5:   Top 5 of movies in clusters M1 and M20. Average rating of the top 5 rated movies in M1 is 4.42, while for the M20 it is 1.

main goal of our algorithm here is to summarize the initial data matrix so that $\mathbf{X}_c$ reveals the blocks (co-clusters) of movies and users that share similar tastes. We set the number of user and film clusters to $g = 10$ and $m = 20$, respectively as in [Banerjee 2007].

The obtained results provide the first movie cluster consisting of films with high ratings (3.92 on average), while the last movie cluster includes movies with very low ratings (1.92 on average). Among those, we show the 5 best/worst rated movies in those two clusters in Table 5.5. Overall, our algorithm manages to find a coherent co-clustering structure in MovieLens-100K and obtains results similar to those provided in [Laclau 2017, Banerjee 2007].

5.7   Conclusion

In this chapter, we presented a novel variant of the optimal transport problem which aims at comparing distributions supported in different spaces. To this end, two optimal transport maps, one acting on the sample space, and the other on the feature space, are optimized to connect the two heterogenous distributions. We show that this novel problem has connections with bilinear assignment and provide algorithms to solve it. We demonstrate its usefulness and versatility on two difficult machine learning problems: heterogeneous domain adaptation and co-clustering/data summarization, where promising results were obtained.

Numerous follow-up of this work are expected. Beyond the potential applications of the method in various contexts, such as *e.g.* statistical matching, data analysis or even losses in deep learning settings, one immediate and intriguing question lies into the generalization of this framework in the continuous setting, and the potential connections to duality theory. This might lead to stochastic optimization schemes enabling large scale solvers for this problem.

Proofs of claims and additional results

Contents

6.1 Proofs and additional results of Chapter 3

This section contains all the proofs of the claims and additional results of the Chapter 3. In the following we will denote by H_q the Wasserstein loss and by J_q the Gromov-Wasserstein loss. More precisely, with notation of Chapter 3:

$$H_q(\pi) = \int d(a,b)^q \mathrm{d}\pi(a,b) \tag{6.1}$$

$$J_q(d_{\mathcal{X}}, d_{\mathcal{Y}}, \pi) = \int\int L(x,y,x',y)^q \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') = \int\int |d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')|^q \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \tag{6.2}$$

$$E_{p,q,\alpha}(\pi) = \int\int \left((1-\alpha)d(a,b)^q + \alpha|d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')|^q\right)^p \mathrm{d}\pi((x,a),(y,b))\mathrm{d}\pi((x',a'),(y',b')) \tag{6.3}$$

We note $P_i\#\pi$ the projection on the marginal i of π.

6.1.1 Additional results – Comparison with W and GW

Cross validation results During the nested cross validation, we divided the dataset into 10 and use 9 folds for training, where α is chosen within $[0,1]$ *via* a 10-CV cross-validation, 1 fold for testing, with the best value of α (with the best average accuracy on the 10-CV) previously selected. The experiment is repeated 10 times for each dataset except for MUTAG and PTC where it is repeated 50 times. Table 6.1 and 6.2 report the average number of time α was chose within $]0, ...1[$ without 0 and 1 corresponding to the Wasserstein and Gromov-Wasserstein distances respectively. Results suggests that both structure and feature pieces of information are necessary as α is consistently selected inside $]0, ...1[$ except for PTC and COX2.

Table 6.1: Percentage of α chosen in $]0, ..., 1[$ compared to $\{0, 1\}$ for discrete labeled graphs

DISCRETE ATTR.	MUTAG	NCI1	PTC
FGW RAW SP	100%	100%	98%
FGW WL H=2 SP	100%	100%	88%
FGW WL H=4 SP	100%	100%	88%

Table 6.2: Percentage of α chosen in $]0, ..., 1[$ compared to $\{0, 1\}$ for vector attributed graphs

VECTOR ATTRIBUTES	BZR	COX2	CUNEIFORM	ENZYMES	PROTEIN	SYNTHETIC
FGW SP	100 %	90%	100%	100%	100%	100%

Nested CV results We report in tables 6.4 and 6.3 the average classification accuracies of the nested classification procedure by taking W and GW instead of FGW (*i.e* by taking $\alpha = 0, 1$). Best result for each dataset is in bold. A (*) is added when best score does not yield to a significative improvement

compared to the second best score. The significance is based on a Wilcoxon signed rank test between the best method and the second one.

Results illustrates that FGW encompasses the two cases of W and GW, as scores of FGW are usually greater or equal on every dataset than scores of both W and GW and when it is not the case the difference is not statistically significant.

Table 6.3: Average classification accuracy on the graph datasets with discrete attributes.

Discrete attr.	MUTAG	NCI1	PTC-MR
FGW raw sp	83.26±10.30	72.82±1.46	55.71±6.74
FGW wl h=2 sp	86.42±7.81	85.82±1.16	63.20±7.68
FGW wl h=4 sp	**88.42±5.67**	**86.42±1.63***	65.31±7.90
W raw sp	79.36±3.49	70.5±4.63	54.79±5.76
W wl h=2 sp	87.78±8.64	85.83±1.75	63.90±7.66
W wl h=4 sp	87.15±8.23	86.42±1.64	**66.28±6.95***
GW sp	82.73±9.59	73.40±2.80	54.45± 6.89

Table 6.4: Average classification accuracy on the graph datasets with vector attributes.

Vector attributes	BZR	COX2	CUNEIFORM	ENZYMES	PROTEIN	SYNTHETIC
FGW sp	85.12±4.15	**77.23±4.86***	**76.67±7.04**	71.00±6.76	74.55±2.74	**100.00±0.00**
W	**85.36±4.87***	77.23±3.16	61.48±10.23	**71.16±6.32***	**75.98± 1.97***	34.07±11.33
GW sp	82.92±6.72	77.65±5.88	50.66±8.91	23.66±3.63	71.96± 2.40	41.66±4.28

Timings   In this paragraph we provide some timings for the discrete attributed datasets. Table 6.5 displays the average timing for computing FGW between two pair of graphs.

Table 6.5: Average timings for the computation of FGW between two pairs of graph

Discrete attr.	MUTAG	NCI1	PTC-MR
FGW	2.5 ms	7.3 ms	3.7 ms

6.1.2   Proof of Proposition 3.5.1 – Comparison between FGW, GW and W

We recall the proposition:

Proposition. *Comparison between FGW, GW and W.*

- *The following inequalities hold:*

$$FGW_{\alpha,p,q}(\mu,\nu) \geq (1-\alpha)W_{pq}(\mu_A,\nu_B)^q \tag{6.4}$$

$$FGW_{\alpha,p,q}(\mu,\nu) \geq \alpha GW_{pq}(\mu_X,\nu_Y)^q \tag{6.5}$$

- Let us suppose that the structure spaces $(\mathcal{X}, d_{\mathcal{X}}), (\mathcal{Y}, d_{\mathcal{Y}})$ are part of a single ground space $(\mathcal{Z}, d_{\mathcal{Z}})$ (i.e. $\mathcal{X}, \mathcal{Y} \subset \mathcal{Z}$ and $d_{\mathcal{X}} = d_{\mathcal{Y}} = d_{\mathcal{Z}}$). We consider the Wasserstein distance between μ and ν for the distance on $\mathcal{Z} \times \Omega$: $\tilde{d}((x,a),(y,b)) = (1-\alpha)d(a,b) + \alpha d_{\mathcal{Z}}(x,y)$. Then:

$$FGW_{\alpha,p,1}(\mu,\nu)(\mu,\nu) \leq 2W_p(\mu,\nu). \tag{6.6}$$

Proof. For the two inequalities (6.4) and (6.5) let π be an optimal coupling for the Fused Gromov-Wasserstein distance between μ and ν. Clearly:

$$FGW_{\alpha,p,q}(\mu,\nu) = \Big(\int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \big((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\big)^p \, \mathrm{d}\pi((x,a),(y,b))\mathrm{d}\pi((x',a'),(y',b')) \Big)^{\frac{1}{p}}$$

$$\geq \Big(\int\limits_{\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega} (1-\alpha)^p d(a,b)^{pq} \, \mathrm{d}\pi((x,a),(y,b)) \Big)^{\frac{1}{p}} = (1-\alpha)\Big(\int\limits_{\Omega \times \Omega} d(a,b)^{pq}\mathrm{d}P_{2,4}\#\pi(a,b) \Big)^{\frac{1}{p}}$$

Since $\pi \in \Pi(\mu,\nu)$ the coupling $P_{2,4}\#\pi$ is in $\Pi(\mu_A, \nu_B)$. So by suboptimality:

$$FGW_{\alpha,p,q}(\mu,\nu) \geq (1-\alpha)(W_{pq}(\mu_A, \nu_B))^q$$

which proves equation (6.4). Same reasoning is used for equation (6.5).

For the last inequality (6.6) let $\pi \in \Pi(\mu,\nu)$ be any admissible coupling. By suboptimality:

$$FGW_{\alpha,p,1}(\mu,\nu) \leq \Big(\int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \big((1-\alpha)d(a,b) + \alpha|d_{\mathcal{Z}}(x,x') - d_{\mathcal{Z}}(y,y')|\big)^p \mathrm{d}\pi((x,a),(y,b))\mathrm{d}\pi((x',a'),(y',b')) \Big)^{\frac{1}{p}}$$

$$\overset{(*)}{\leq} \Big(\int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \big((1-\alpha)d(a,b) + \alpha d_{\mathcal{Z}}(x,y) + \alpha d_{\mathcal{Z}}(x',y')\big)^p \mathrm{d}\pi((x,a),(y,b))\mathrm{d}\pi((x',a'),(y',b')) \Big)^{\frac{1}{p}}$$

$$\leq \Big(\int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \big((1-\alpha)d(a,b) + \alpha d_{\mathcal{Z}}(x,y) + (1-\alpha)d(a',b') + \alpha d_{\mathcal{Z}}(x',y')\big)^p \mathrm{d}\pi((x,a),(y,b))\mathrm{d}\pi((x',a'),(y',b')) \Big)^{\frac{1}{p}}$$

$$\overset{(**)}{\leq} 2\Big(\int\limits_{\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega} \big((1-\alpha)d(a,b) + \alpha d_{\mathcal{Z}}(x,y)\big)^p \mathrm{d}\pi((x,a),(y,b)) \Big)^{\frac{1}{p}}$$

(*) is the triangle inequality of $d_{\mathcal{Z}}$ and (**) Minkowski inequality. Since this inequality is true for any admissible coupling π we can apply it with the optimal coupling for the Wasserstein distance defined in the proposition and the claim follows.

$\square$

6.1.3 Proof of Theorem 3.5.1 – Metric properties of FGW

We recall the theorem:

Theorem (Metric properties). *Let $p, q \geq 1$, $\alpha \in]0,1[$ and $(\mu,\nu) \in \mathbb{S}_{pq}(\Omega) \times \mathbb{S}_{pq}(\Omega)$. The functional $\pi \to E_{p,q,\alpha}(\pi)$ always achieves an infimum π^* in $\Pi(\mu,\nu)$ s.t. $FGW_{\alpha,p,q}(\mu,\nu) = E_{p,q,\alpha}(\pi^*) < +\infty$. Moreover:*

- *$FGW_{\alpha,p,q}$ is symmetric and, for $q = 1$, satisfies the triangle inequality. For $q \geq 2$, the triangular inequality is relaxed by a factor 2^{q-1}.*

- *For $\alpha \in]0,1[$, $FGW_{\alpha,p,q}(\mu,\nu) = 0$ if an only if there exists a bijective function $\phi = (\phi_1, \phi_2)$: $supp(\mu) \rightarrow supp(\nu)$ such that:*

$$\phi\#\mu = \nu$$

$$\forall (x,a) \in supp(\mu) \ , \phi_2(x,a) = a$$

$$\forall (x,a),(x',a') \in supp(\mu)^2, \ d_{\mathcal{X}}(x,x') = d_{\mathcal{Y}}(\phi_1(x,a), \phi_1(x',a'))$$

- *If (μ,ν) are generalized labeled graphs then $FGW_{\alpha,p,q}(\mu,\nu) = 0$ if and only if $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \nu)$ are (II)-strongly isomorphic.*

We propose to prove the theorem point by point: first the existence, then the the triangle inequality statement and finally the equality relation.

Proposition 6.1.1 (Existence of an optimal coupling for the *FGW* distance.). *For $p,q \geq 1$, $\pi \rightarrow E_{p,q,\alpha}(\pi)$ always achieves a infimum π^* in $\Pi(\mu,\nu)$ such that $FGW_{\alpha,p,q}(\mu,\nu) = E_{p,q,\alpha}(\pi^*) < +\infty$.*

Proof. Since $\mathcal{X} \times \Omega$ and $\mathcal{Y} \times \Omega$ are Polish spaces we known that $\Pi(\mu,\nu) \subset \mathcal{P}(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)$ is compact (Theorem 1.7 in [Santambrogio 2015]), so by applying Weierstrass theorem we can conclude that the infimum is attained at some $\pi^* \in \Pi(\mu,\nu)$ if $\pi \rightarrow E_{p,q,\alpha}(\pi)$ is l.s.c.

We will use Lemma 2.2.1 to prove that the functionnal is l.s.c. on $\Pi(\mu,\nu)$. If we consider $W = \mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega$ which is a a metric space endowed with the distance $d_{\mathcal{X}} \otimes d \otimes d_{\mathcal{Y}} \otimes d$ and $f((w = (x,a,y,b), w' = (x',a',y',b')) = ((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q)^p$ then f is l.s.c. by continuity of d, $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$. With the previous reasoning we can conclude that the infimum is attained.

Finally finiteness come from:

$$\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \left((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\right)^p d\pi((x,a),(y,b)) \, d\pi((x',a'),(y',b'))$$

$$\overset{*}{\leq} \int 2^{p-1}(1-\alpha)d(a,b)^{qp} d\mu_A(a) d\nu(b) + \int 2^{p-1}\alpha L(x,y,x',y')^{qp} d\mu_X(x) d\mu_X(x') d\nu_Y(y') d\nu_Y(y') \quad (6.7)$$

$$\overset{**}{<} +\infty$$

where in (*) we used equation (2.44) in Memo 2.2.2 and in (**) that μ, ν are in $\mathbb{S}_{pq}(\Omega)$. $\square$

Proposition 6.1.2 (Symmetry and triangle inequality.). *$FGW_{\alpha,p,q}$ is symmetric and for $q = 1$ satisfies the triangle inequality. For $q \geq 2$ the triangular inequality is relaxed by a factor 2^{q-1}*

To prove this result we will use the following lemma:

Lemma 6.1.1. *Let $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu),(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \beta),(\mathcal{Z} \times \Omega, d_{\mathcal{Z}}, \nu) \in \mathbb{S}(\Omega)^3$. For $(x,a),(x',a') \in (\mathcal{X} \times \Omega)^2$, $(y,b),(y',b') \in (\mathcal{Y} \times \Omega)^2$ and $(z,c),(z',c') \in (\mathcal{Z} \times \Omega)^2$ we have:*

$$L(x,z,x',z')^q \leq 2^{q-1}(L(x,y,x',y')^q + L(y,z,y',z')^q) \quad (6.8)$$

$$d(a,c)^q \leq 2^{q-1}(d(a,b)^q + d(b,c)^q) \quad (6.9)$$

Proof. Direct consequence of equation (2.44) in Memo 2.2.2 and triangle inequalities of $d, d_{\mathcal{X}}, d_{\mathcal{Y}}, d_{\mathcal{Z}}$. $\square$

Proof of Proposition 6.1.2. To prove the triangle inequality of $FGW_{\alpha,p,q}$ distance for arbitrary measures we will use the Gluing lemma which stresses the existence of couplings with a prescribed structure. Let $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu),(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \beta),(\mathcal{Z} \times \Omega, d_{\mathcal{Z}}, \nu) \in \mathbb{S}(\Omega)^3$.

Let $\pi_1 \in \Pi(\mu, \beta)$ and $\pi_2 \in \Pi(\beta, \nu)$ be optimal transportation plans for the Fused Gromov-Wasserstein distance between μ, β and β, ν respectively. By the Gluing Lemma (see [Villani 2008] and Lemma 5.3.2 in [Ambrosio 2005]) there exists a probability measure $\pi \in \mathcal{P}\big((\mathcal{X} \times \Omega) \times (\mathcal{Y} \times \Omega) \times (\mathcal{Z} \times \Omega)\big)$ with marginals π_1 on $(\mathcal{X} \times \Omega) \times (\mathcal{Y} \times \Omega)$ and π_2 on $(\mathcal{Y} \times \Omega) \times (\mathcal{Z} \times \Omega)$. Let π_3 be the marginal of π on $(\mathcal{X} \times \Omega) \times (\mathcal{Z} \times \Omega)$. By construction $\pi_3 \in \Pi(\mu, \nu)$. So by suboptimality of π_3:

$$FGW_{\alpha,p,q}(d_{\mathcal{X}}, d_{\mathcal{Z}}, \mu, \nu) \leq \left(\int_{(\mathcal{X} \times \Omega \times \mathcal{Z} \times \Omega)^2} \big((1-\alpha)d(a,c)^q + \alpha L(x,z,x',z')^q\big)^p d\pi_3((x,a),(z,c)) d\pi_3((x',a'),(z',c')) \right)^{\frac{1}{p}}$$

$$= \left(\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega \times \mathcal{Z} \times \Omega)^2} \big((1-\alpha)d(a,c)^q + \alpha L(x,z,x',z')^q\big)^p d\pi((x,a),(y,b),(z,c)) d\pi((x',a'),(y',b'),(z',c')) \right)^{\frac{1}{p}}$$

$$\overset{(*)}{\leq} 2^{q-1} \left(\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega \times \mathcal{Z} \times \Omega)^2} \big((1-\alpha)d(a,b)^q + (1-\alpha)d(b,c)^q + \alpha L(x,y,x',y')^q + \alpha L(y,z,y',z')^q\big)^p \right.$$

$$\left. d\pi((x,a),(y,b),(z,c)) d\pi((x',a'),(y',b'),(z',c')) \right)^{\frac{1}{p}}$$

$$\overset{(**)}{\leq} 2^{q-1} \left(\left(\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega \times \mathcal{Z} \times \Omega)^2} \big((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\big)^p d\pi((x,a),(y,b),(z,c)) d\pi((x',a'),(y',b'),(z',c')) \right)^{\frac{1}{p}} \right.$$

$$\left. + \left(\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega \times \mathcal{Z} \times \Omega)^2} \big((1-\alpha)d(b,c)^q + \alpha L(y,z,y',z')^q\big)^p d\pi((x,a),(y,b),(z,c)) d\pi((x',a'),(y',b'),(z',c')) \right)^{\frac{1}{p}} \right)$$

$$= 2^{q-1} \left(\left(\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \big((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q\big)^p d\pi_1((x,a),(y,b)) d\pi_1((x',a'),(y',b')) \right)^{\frac{1}{p}} \right.$$

$$\left. + \left(\int_{(\mathcal{Y} \times \Omega \times \mathcal{Z} \times \Omega)^2} \big((1-\alpha)d(b,c)^q + \alpha L(y,z,y',z')^q\big)^p d\pi_2((y,b),(z,c)) d\pi_2((y',b'),(z',c')) \right)^{\frac{1}{p}} \right)$$

$$= 2^{q-1}(FGW_{\alpha,p,q}(\mu, \beta) + FGW_{\alpha,p,q}(\beta, \nu))$$

with (*) comes from (6.8) and (6.9) and (**) is Minkowski inequality. So when $q = 1$, $FGW_{\alpha,p,q}$ satisfies the triangle inequality and when $q > 1$, $FGW_{\alpha,p,q}$ satisfies a relaxed triangle inequality so that it defines a semi-metric as described previously. $\square$

Proposition 6.1.3 (Equality relation.)**.** *For* $\alpha \in\,]0,1[$, $FGW_{\alpha,p,q}(\mu, \nu) = 0$ *if an only if there exists a bijective function* $\phi = (\phi_1, \phi_2) : supp(\mu) \to supp(\nu)$ *such that:*

$$\phi \# \mu = \nu \tag{6.10}$$

$$\forall (x,a) \in supp(\mu) , \phi_2(x,a) = a \tag{6.11}$$

$$\forall (x,a), (x',a') \in supp(\mu)^2, \ d_{\mathcal{X}}(x,x') = d_{\mathcal{Y}}(\phi_1(x,a), \phi_1(x',a')) \tag{6.12}$$

Moreover if (μ, ν) *are generalized labeled graphs then* $FGW_{\alpha,p,q}(\mu, \nu) = 0$ *if and only if* $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu)$ *and* $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \nu)$ *are (II)-strongly isomorphic.*

Proof. For the first point, let us assume that there exists a function ϕ verifying (6.10), (6.11) and (6.12).

We consider the map $\pi = (I_d \times \phi)\#\mu \in \Pi(\mu, \nu)$. We note $\phi = (\phi_1, \phi_2)$. Then:

$$
\begin{aligned}
E_{p,q,\alpha}(\pi) &= \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \left((1-\alpha)d(a,b)^q + \alpha L((x,y,x',y')^q \right)^p d\pi((x,a),(y,b))d\pi((x',a'),(y',b')) \\
&= \int_{(\mathcal{X} \times \Omega)^2} \left((1-\alpha)d(a,\phi_2(x,a))^q + \alpha L\big((x,\phi_1(x,a),x',\phi_1(x',a'))^q \big) \right)^p d\mu(x,a)d\mu(x',a') \\
&= \int_{(\mathcal{X} \times \Omega)^2} \left((1-\alpha)d(a,\phi_2(x,a))^q + \alpha|d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(\phi_1(x,a),\phi_1(x',a'))|^q \right)^p d\mu(x,a)d\mu(x',a') \\
&= 0
\end{aligned}
$$

$$(6.13)$$

Conversely suppose that $FGW_{\alpha,p,q}(\mu,\nu) = 0$. To prove the existence of a map $\phi : \mathrm{supp}(\mu) \to \mathrm{supp}(\nu)$ verifying (6.10), (6.11) and (6.12) we will use the Gromov-Wasserstein properties. We are looking for a vanishing Gromov-Wassersein distance between the spaces $\mathcal{X} \times \Omega$ and $\mathcal{Y} \times \Omega$ equipped with our two measures μ and ν.

More precisely, we define for $\big((x,a),(y,b),(x',a'),(y',b')\big) \in (\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2$ and $\beta \in\,]0,1[$:

$$
d_{\mathcal{X} \times \Omega}\big((x,a),(x',a')\big) = (1-\beta)d_{\mathcal{X}}(x,x') + \beta d(a,a')
$$

and

$$
d_{\mathcal{Y} \times \Omega}\big((y,b),(y',b')\big) = (1-\beta)d_{\mathcal{Y}}(y,y') + \beta d(b,b')
$$

We will prove that $d_{GW,p}(d_{\mathcal{X} \times \Omega}, d_{\mathcal{Y} \times \Omega}, \mu, \nu) = 0$. To show that we will bound the Gromov cost with the metrics $d_{\mathcal{X} \times \Omega}, d_{\mathcal{Y} \times \Omega}$ by the Gromov cost with the metrics $d_{\mathcal{X}}, d_{\mathcal{Y}}$ and a Wasserstein cost.

Let $\pi \in \Pi(\mu, \nu)$ be any admissible transportation plan. Then for $n \geq 1$:

$$
\begin{aligned}
J_n(d_{\mathcal{X} \times \Omega}, d_{\mathcal{Y} \times \Omega}, \pi) &\overset{def}{=} \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} L(x,y,x',y')^n d\pi((x,a),(y,b))d\pi((x',a'),(y',b')) \\
&= \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} |(1-\beta)(d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')) + \beta(d(a,a') - d(b,b'))|^n d\pi((x,a),(y,b))d\pi((x',a'),(y',b')) \\
&\leq \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} (1-\beta)|d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y')|^n d\pi((x,a),(y,b))\, d\pi((x',a'),(y',b')) \\
&\quad + \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \beta|d(a,a') - d(b,b')|^n d\pi((x,a),(y,b))d\pi((x',a'),(y',b'))
\end{aligned}
$$

using Jensen inequality with convexity of $t \to t^n$ and subadditivity of $|.|$. We note $(*)$ the first term above and $(**)$ the second term above. By the triangle inequality property of d we have:

$(**) \leq \beta \int_{\big(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} (d(a,b) + d(a',b'))^n d\pi((x,a),(y,b))\, d\pi((x',a'),(y',b')) \overset{def}{=} \beta M_n(\pi)$ such that we have shown:

$$
\forall \pi \in \Pi(\mu,\nu), \forall n \geq 1, J_n(d_{\mathcal{X} \times \Omega}, d_{\mathcal{Y} \times \Omega}, \pi) \leq (1-\beta)J_n(d_{\mathcal{X}}, d_{\mathcal{Y}}, \pi) + \beta M_n(\pi) \tag{6.14}
$$

Now let π_* be an optimal coupling for $FGW_{\alpha,p,q}$ between μ and ν. By hypothesis $FGW_{\alpha,p,q}(\mu,\nu) = 0$ so that:

$$J_{qp}(d_{\mathcal{X}}, d_{\mathcal{Y}}, \pi*) = 0 \tag{6.15}$$

and:

$$H_{qp}(\pi^*) = 0. \tag{6.16}$$

Then $\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)} d(a, b)^{qp} d\pi^*((x, a), (y, b)) = 0$ which implies that d is zero π^* a.e. so that $\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)} d(a, b)^m d\pi^*((x, a), (y, b)) = 0$ for any $m \in \mathbb{N}^*$. In this way:

$$M_{qp}(\pi^*) = \beta \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \sum_h \binom{qp}{h} d(a, b)^h d(a', b')^{qp-h} d\pi^*((x, a), (y, b)) d\pi^*((x', a'), (y', b'))$$

$$= \beta \sum_h \binom{qp}{h} \Big(\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)} d(a, b)^h d\pi^*((x, a), (y, b)) \Big) \Big(\int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)} d(a', b')^{qp-h} d\pi^*((x', a'), (y', b')) \Big) = 0$$

Using equation (6.14) we have shown

$$J_{qp}(d_{\mathcal{X} \times \Omega}, d_{\mathcal{Y} \times \Omega}, \pi_*) = 0$$

which implies that $d_{GW,p}(d_{\mathcal{X} \times \Omega}, d_{\mathcal{Y} \times \Omega}, \mu, \nu) = 0$ for the coupling π^*.

Thanks to the Gromov-Wasserstein properties (see Chapter 2) this states the existence of an isometry between $\text{supp}(\mu)$ and $\text{supp}(\nu)$. So there exists a surjective function $\phi = (\phi_1, \phi_2) : \text{supp}(\mu) \to \text{supp}(\nu)$ which verifies **P.1** and:

$$\forall((x, a), (x', a')) \in (\text{supp}(\mu))^2, \ d_{\mathcal{X} \times \Omega}((x, a), (x', a')) = d_{\mathcal{Y} \times \Omega}(\phi(x, a), \phi(x', a')) \tag{6.17}$$

or equivalently:

$$\forall((x, a), (x', a')) \in (\text{supp}(\mu))^2, \ (1-\beta)d_{\mathcal{X}}(x, x') + \beta d(a, a') = (1-\beta)d_{\mathcal{Y}}(\phi_1(x, a), \phi_1(x', a')) + \beta d(\phi_2(x, a), \phi_2(x', a')) \tag{6.18}$$

In particular π^* is concentrated on $\{(x, y) = \phi(x, y)\}$ or equivalently $\pi^* = (I_d \times \phi) \# \mu$. Injecting π^* in (6.16) leads to:

$$H_{qp}(\pi^*) = \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)} d(a, b)^{qp} d\pi^*((x, a), (y, b)) = \int_{\mathcal{X} \times \Omega} d(a, \phi_2(x, a))^{qp} d\mu(x, a) = 0 \tag{6.19}$$

Which implies:

$$\forall(x, a) \in \text{supp}(\mu), \phi_2(x, a) = a \tag{6.20}$$

Moreover, using the equality (6.18) we can conclude that:

$$\forall(x, a)(x', a') \in \text{supp}(\mu)^2, \ d_{\mathcal{X}}(x, x') = d_{\mathcal{Y}}(\phi_1(x, a), \phi_1(x', a')) \tag{6.21}$$

In this way f verifies all the properties (6.10),(6.11),(6.12).

Moreover suppose that μ and ν are generalized labeled graphs. In this case there exists $\ell_f : \mathcal{X} \to \Omega$ surjective such that $\mu = (id \times \ell_f) \# \mu_X$. Then (6.21) implies that:

$$\forall(x, x') \in \text{supp}(\mu_X)^2, \ d_{\mathcal{X}}(x, x') = d_{\mathcal{Y}}(\phi_1(x, \ell_f(x)), \phi_1(x', \ell_f(x'))) \tag{6.22}$$

We define $I : \mathrm{supp}(\mu_X) \to \mathrm{supp}(\mu_Y)$ such that $I(x) = \phi_1(x, \ell_f(x))$. Then we have by (6.22) $d_{\mathcal{X}}(x, x') = d_{\mathcal{Y}}(I(x), I(x'))$ for $(x, x') \in \mathrm{supp}(\mu_X)^2$. Overall we have $\phi(x, a) = (I(x), a)$ for all $(x, a) \in \mathrm{supp}(\mu)$. Also since $\phi \# \mu = \nu$ we have $I \# \mu_X = \nu_Y$.

Moreover I is a surjective function. Indeed let $y \in \mathrm{supp}(\nu_Y)$. Let $b \in \mathrm{supp}(\nu_B)$ such that $(y, b) \in \mathrm{supp}(\nu)$. By surjectivity of ϕ there exists $(x, a) \in \mathrm{supp}(\mu)$ such that $(y, b) = \phi(x, a) = (I(x), a)$ so that $y = I(x)$.

Overall ϕ satisfies all **P.1**, **P.2** and **P.3** if μ and ν are generalized labeled graphs. The converse is also true using the reasoning in (6.13). $\qquad\square$

6.1.4 Proof of Theorem 3.5.2 – Convergence and concentration inequality.

We recall the theorem:

Theorem. *Convergence of finite samples and a concentration inequality Let $p \geq 1$. We have:*

$$\lim_{n \to \infty} FGW_{\alpha, p, 1}(\mu_n, \mu) = 0$$

Moreover, suppose that $s > d_p^*(\mu)$. Then there exists a constant C that does not depend on n such that:

$$\mathbb{E}[FGW_{\alpha, p, 1}(\mu_n, \mu)] \leq C n^{-\frac{1}{s}}.$$

The expectation is taken over the i.i.d samples (x_i, a_i). A particular case of this inequality is when $\alpha = 1$ so that we can use the result above to derive a concentration result for the Gromov-Wasserstein distance. More precisely, if $\nu_n = \frac{1}{n} \sum_i \delta_{x_i}$ denotes the empirical measure of $\nu \in \mathcal{P}(\mathcal{X})$ and if $s' > d_p^*(\nu)$ we have:

$$\mathbb{E}[GW_p(\nu_n, \nu)] \leq C' n^{-\frac{1}{s'}}.$$

Proof. The proof of the convergence in FGW derives directly from the weak convergence of the empirical measure and Lemma 2.2.1. Moreover, since μ_n and μ are both in the same ground space, we have:

$$FGW_{\alpha, p, 1}(\mu_n, \mu) \leq 2W_p(\mu_n, \mu) \implies \mathbb{E}[FGW_{\alpha, p, 1}(\mu_n, \mu)] \leq 2\mathbb{E}[W_p(\mu_n, \mu)].$$

We can directly apply theorem 1 in [Weed 2017] to state the inequality. $\qquad\square$

6.1.5 Proof of proposition 3.5.2 – Interpolation properties between GW and W

We recall the proposition:

Proposition (Interpolation properties.). *As α tends to zero, one recovers the Wasserstein distance between the features information and as α goes to one, one recovers the Gromov-Wasserstein distance between the structure information:*

$$\lim_{\alpha \to 0} FGW_{\alpha, p, q}(\mu, \nu) = (W_{pq}(\mu_A, \nu_B))^q$$

$$\lim_{\alpha \to 1} FGW_{\alpha, p, q}(\mu, \nu) = (GW_{pq}(\mu_X, \nu_Y))^q$$

Proof. Let $\pi_{OT} \in \Pi(\mu_A, \nu_B)$ be an optimal coupling for the pq-Wasserstein distance between μ_A and ν_B. We can use the same Gluing lemma (lemma 5.3.2 in [Ambrosio 2005]) to construct:

$$\rho \in \mathcal{P}(\overbrace{\mathcal{X} \times \underbrace{\Omega \times \Omega}_{\pi_{OT}}}^{\mu} \times \overbrace{\mathcal{Y}}^{\nu})$$

such that $\rho \in \Pi(\mu, \nu)$ and $P_{2,3}\#\rho = \pi_{OT}$.

Moreover we have:

$$\int\limits_{\Omega \times \Omega} d(a,b)^{pq} d\pi_{OT}(a,b) = \int\limits_{\mathcal{X} \times \Omega \times \Omega \times \mathcal{Y}} d(a,b)^{pq} d\rho(x,a,b,y) \tag{6.23}$$

Let $\alpha \geq 0$ and π_α optimal plan for the Fused Gromov-Wasserstein distance between μ and ν.

We can deduce that:

$$FGW_{\alpha,p,q}(\mu,\nu)^p - (1-\alpha)^p W_{pq}(\mu_A, \nu_B)^{pq}$$

$$= \int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \left((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q \right)^p d\pi_\alpha((x,a),(y,b)) d\pi_\alpha((x',a'),(y',b')) - \int\limits_{\Omega \times \Omega} (1-\alpha)^p d(a,b)^{pq} d\pi_{OT}(a,b)$$

$$\overset{(*)}{\leq} \int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \left((1-\alpha)d(a,b)^q + \alpha L(x,y,x',y')^q \right)^p d\rho(x,a,b,y) d\rho(x',a',b',y') - \int\limits_{\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega} (1-\alpha)^p d(a,b)^{pq} d\rho(x,a,b,y)$$

$$= (1-\alpha)^p \int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} d(a,b)^{pq} d\rho(x,a,b,y) d\rho(x',a',b',y') - (1-\alpha)^p \int\limits_{\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega} d(a,b)^{pq} d\rho(x,a,b,y)$$

$$+ \sum_{k=0}^{p-1} \binom{p}{k} (1-\alpha)^k \alpha^{p-k} \int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} d(a,b)^{qk} L(x,y,x',y')^{q(p-k)} d\rho(x,a,b,y) d\rho(x',a',b',y')$$

$$= \sum_{k=0}^{p-1} \binom{p}{k} (1-\alpha)^k \alpha^{p-k} \int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} d(a,b)^{qk} L(x,y,x',y')^{q(p-k)} d\rho(x,a,b,y) d\rho(x',a',b',y').$$

We note $H_k = \int\limits_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} d(a,b)^{qk} L(x,y,x',y')^{q(p-k)} d\rho(x,a,b,y) d\rho(x',a',b',y')$.

Using (6.4) we have shown that:

$$(1-\alpha)(W_{pq}(\mu_A, \nu_B))^q \leq FGW_{\alpha,p,q}(\mu,\nu) \leq \left((1-\alpha)^p (W_{pq}(\mu_A, \nu_B))^{pq} + \sum_{k=0}^{p-1} \binom{p}{k} (1-\alpha)^k \alpha^{p-k} H_k \right)^{\frac{1}{p}}$$

So $\lim\limits_{\alpha \to 0} FGW_{\alpha,p,q}(\mu,\nu) = (W_{pq}(\mu_A, \nu_B))^q$.

For the case $\alpha \to 1$ we rather consider $\pi_{GW} \in \Pi(\mu_X, \nu_Y)$ an optimal coupling for the pq-Gromov-Wasserstein distance between μ_X and ν_Y and we construct

$$\gamma \in \mathcal{P}(\overbrace{\Omega \times \underbrace{\mathcal{X} \times \mathcal{Y}}_{\pi_{GW}} \times \Omega}^{\mu \quad \nu})$$

such that $\gamma \in \Pi(\mu, \nu)$ and $P_{2,3}\#\rho = \pi_{GW}$. In the same way as previous reasoning we can derive:

$$\alpha(GW_{pq}(\mu_X, \nu_Y))^q \leq FGW_{\alpha,p,q}(\mu,\nu) \leq \left(\alpha^p (GW_{pq}(\mu_X, \nu_Y))^{pq} + \sum_{k=0}^{p-1} \binom{p}{k} (1-\alpha)^{p-k} \alpha^k J_k \right)^{\frac{1}{p}} \tag{6.24}$$

with $J_k = \int d(a,b)^{q(p-k)} L(x,y,x',y')^{qk} d\rho(x,a,b,y) d\rho(x',a',b',y')$. In this way $\lim\limits_{\alpha \to 1} FGW_{\alpha,p,q}(\mu,\nu) = (GW_{pq}(\mu_X, \nu_Y))^q$. $\qquad\square$

6.1.6   Proof of Theorem 3.5.3 – Constant speed geodesic.

We recall the theorem:

Theorem (Constant speed geodesic.). *Let $p \geq 1$ and $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu_0)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \mu_1)$ in $\mathbb{S}_p(\mathbb{R}^d)$. Let π^* be an optimal coupling for the Fused Gromov-Wasserstein distance between μ_0, μ_1 and $t \in [0, 1]$. We equip $\mathbb{R}^d$ with ℓ_m norm for $m \geq 1$.*

We define $\eta_t : \mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega \to \mathcal{X} \times \mathcal{Y} \times \Omega$ such that:

$$\forall (x, a), (y, b) \in \mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega, \ \eta_t(x, \mathbf{a}, y, \mathbf{b}) = (x, y, (1 - t)\mathbf{a} + t\mathbf{b})$$

Then:

$$(\mathcal{X} \times \mathcal{Y} \times \Omega, (1 - t)d_{\mathcal{X}} \oplus td_{\mathcal{Y}}, \mu_t = \eta_t \# \pi^*)_{t \in [0,1]}$$

is a constant speed geodesic connecting $(\mathcal{X} \times \Omega, d_{\mathcal{X}}, \mu_0)$ and $(\mathcal{Y} \times \Omega, d_{\mathcal{Y}}, \mu_1)$ in the metric space $(\mathbb{S}_p(\mathbb{R}^d), FGW_{\alpha,p,1})$.

Proof. We note $\mathbb{S}_t = (\mathcal{X} \times \mathcal{Y} \times \Omega, d_t, \mu_t = \eta_t \# \pi^*)_{t \in [0,1]}$ where $d_t = (1 - t)d_{\mathcal{X}} \oplus td_{\mathcal{Y}}$. Let $\|.\|$ be any ℓ_m norm for $m \geq 1$. It suffices to prove:

$$FGW_{\alpha,p,1}(\mu_t, \mu_s) \leq |t - s| FGW_{\alpha,p,1}(\mu_0, \mu_1) \tag{6.25}$$

To do so we consider $\Delta_s^t \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \Omega \times \mathcal{X} \times \mathcal{Y} \times \Omega)$ defined by $\Delta_s^t = (\eta_t \times \eta_s) \# \pi^* \in \Pi(\mu_t, \mu_s)$ and the following "diagonal" coupling:

$$d\gamma_s^t((x, y), \mathbf{a}, (x'', y''), \mathbf{b}) = d\Delta_s^t((x, y), \mathbf{a}, (x'', y''), \mathbf{b})d\delta_{(x_0, x_1)}(x_0'', x_1'') \tag{6.26}$$

Then $\gamma_s^t \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \Omega \times \mathcal{X} \times \mathcal{Y} \times \Omega)$ and since $\Delta_s^t \in \Pi(\mu_t, \mu_s)$ then $\gamma_s^t \in \Pi(\mu_t, \mu_s)$ So by suboptimality:

$$FGW_{\alpha,p,1}(\mu_t, \mu_s)^p \leq \int_{(\mathcal{X} \times \mathcal{Y} \times \Omega \times \mathcal{X} \times \mathcal{Y} \times \Omega)^2} \left((1 - \alpha)d(a, b) + \alpha|d_t[(x, y), (x', y')] - d_s[(x'', y''), (x''', y''')]| \right)^p$$

$$d\gamma_s^t(x, y, \mathbf{a}, x'', y'', \mathbf{b})d\gamma_s^t(x', y', \mathbf{a}', x''', y''', \mathbf{b}')$$

$$= \int_{(\mathcal{X} \times \mathcal{Y} \times \Omega \times \mathcal{X} \times \mathcal{Y} \times \Omega)^2} \left((1 - \alpha)d(\mathbf{a}, \mathbf{b}) + \alpha|d_t[(x, y), (x', y')] - d_s[(x, y), (x', y')]| \right)^p$$

$$d\Delta_s^t(x, y, \mathbf{a}, x, y, \mathbf{b})d\Delta_s^t(x', y', \mathbf{a}', x', y', \mathbf{b}')$$

$$= \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \left((1 - \alpha)\|(1 - t)\mathbf{a} + t\mathbf{b} - (1 - s)\mathbf{a} - s\mathbf{b}\| + \alpha|(1 - t)d_{\mathcal{X}}(x, x') + td_{\mathcal{Y}}(y, y') - (1 - s)d_{\mathcal{X}}(x, x') + sd_{\mathcal{Y}}(y, y')| \right)^p$$

$$d\pi^*(x, a, y, b)d\pi^*(x', \mathbf{a}', y', \mathbf{b}')$$

$$= |t - s|^p \int_{(\mathcal{X} \times \Omega \times \mathcal{Y} \times \Omega)^2} \left((1 - \alpha)\|\mathbf{a} - \mathbf{b}\| + \alpha|d_{\mathcal{X}}(x, x') - d_{\mathcal{Y}}(y, y')| \right)^p d\pi^*(x, \mathbf{a}, y, \mathbf{b})d\pi^*(x', \mathbf{a}', y', \mathbf{b}')$$

So $FGW_{\alpha,p,1}(\mu_t, \mu_s) \leq |t - s| FGW_{\alpha,p,1}(d_0, d_1, \mu_0, \mu_1)$. $\qquad \square$

6.2   Proofs and additional results of Chapter 4

This Chapter contains all the proofs of the Chapter 4. We recall the following notations:

$$J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi) \overset{def}{=} \int_{\mathcal{X} \times \mathcal{X}} \int_{\mathcal{Y} \times \mathcal{Y}} |c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') - c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}')|^2 d\pi(\mathbf{x}, \mathbf{y})d\pi(\mathbf{x}', \mathbf{y}')$$

$$J(T) \overset{def}{=} \int \left(\|\mathbf{x} - \mathbf{x}'\|_2^2 - \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2 \right)^2 \mathrm{d}\mu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}')$$

$$LGM_2^2(\mu,\nu) \overset{def}{=} \inf_{\substack{T\#\mu=\nu \\ \mathrm{T\ is\ linear}}} J(T) = \inf_{\substack{T\#\mu=\nu \\ \mathrm{T\ is\ linear}}} \int \left(\|\mathbf{x} - \mathbf{x}'\|_2^2 - \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2 \right)^2 \mathrm{d}\mu(\mathbf{x})\mathrm{d}\mu(\mathbf{x}')$$

6.2.1 Proof of Theorem 4.1.1 – New special of the QAP

We recall the theorem:

Theorem (A new special case for the Quadratic Assignment Problem). *For real numbers $x_1 < \cdots < x_n$ and $y_1 < \cdots < y_n$,*

$$\min_{\sigma \in S_n} \sum_{i,j} -(x_i - x_j)^2 (y_{\sigma(i)} - y_{\sigma(j)})^2 \tag{6.27}$$

is achieved either by the identity permutation $\sigma(i) = i$ (Id) or the anti-identity permutation $\sigma(i) = n+1-i$ (anti $-$ Id). In other words:

$$\exists \sigma \in \{Id, anti - Id\}, \ \sigma \in \arg\min_{\sigma \in S_n} \sum_{i,j} -(x_i - x_j)^2 (y_{\sigma(i)} - y_{\sigma(j)})^2 \tag{6.28}$$

Let us note $\mathcal{I} = \{\mathbf{x}, \mathbf{y} \in \mathbb{R}^n \times \mathbb{R}^n | x_1 < x_2 < \cdots < x_n \ , y_1 < y_2 < \cdots < y_n\}$. We consider for $\mathbf{x}, \mathbf{y} \in \mathcal{I}$:

$$\max_{\sigma \in S_n} Z(\mathbf{x}, \mathbf{y}, \sigma) = \max_{\sigma \in S_n} \sum_{i,j} (x_i - x_j)^2 (y_{\sigma(i)} - y_{\sigma(j)})^2 \tag{6.29}$$

The original problem is equivalent to maximizing $Z(\mathbf{x}, \mathbf{y}, \sigma)$ over S_n. Given $\mathbf{x}, \mathbf{y} \in \mathcal{I}$, we define $X \overset{\text{def}}{=} \sum_i x_i$ and $Y \overset{\text{def}}{=} \sum_i y_i$. Then:

$$\max_{\sigma \in S_n} Z(\mathbf{x}, \mathbf{y}, \sigma) = \max_{\sigma \in S_n} \sum_{i,j} (x_i - x_j)^2 (y_{\sigma(i)} - y_{\sigma(j)})^2$$

$$= \max_{\sigma \in S_n} \sum_{i,j} (x_i^2 + x_j^2)(y_{\sigma(i)}^2 + y_{\sigma(j)}^2) - 2\sum_{i,j} x_i x_j (y_{\sigma(i)}^2 + y_{\sigma(j)}^2) - 2\sum_{i,j} y_{\sigma(i)} y_{\sigma(j)} (x_i^2 + x_j^2)$$

$$+ 4\sum_{i,j} x_i x_j y_{\sigma(i)} y_{\sigma(j)}$$

$$= \max_{\sigma \in S_n} 2n \sum_i x_i^2 y_{\sigma(i)}^2 - 2\sum_{i,j} x_i x_j (y_{\sigma(i)}^2 + y_{\sigma(j)}^2) - 2\sum_{i,j} y_{\sigma(i)} y_{\sigma(j)} (x_i^2 + x_j^2)$$

$$+ 4\sum_{i,j} x_i x_j y_{\sigma(i)} y_{\sigma(j)} + 2(\sum_i x_i^2)(\sum_i y_i^2)$$

$$= \max_{\sigma \in S_n} 2n \sum_i x_i^2 y_{\sigma(i)}^2 - 4X \sum_i x_i y_{\sigma(i)}^2 - 4Y \sum_i x_i^2 y_{\sigma(i)} + 4\sum_{i,j} x_i x_j y_{\sigma(i)} y_{\sigma(j)} + 2(\sum_i x_i^2)(\sum_i y_i^2)$$

$$\overset{(*)}{=} Cte + 2\big(\max_{\sigma \in S_n} \sum_i n x_i^2 y_{\sigma(i)}^2 - 2\sum_i (X x_i y_{\sigma(i)}^2 + Y x_i^2 y_{\sigma(i)}) + 2(\sum_i x_i y_{\sigma(i)})^2 \big)$$

where in (*) we defined $Cte \overset{\text{def}}{=} 2(\sum_i x_i^2)(\sum_i y_i^2)$ the term that does not depend on σ. Overall we have:

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{I}, \ \arg\max_{\sigma \in S_n} Z(\mathbf{x}, \mathbf{y}, \sigma) = \arg\max_{\sigma \in S_n} \sum_i n x_i^2 y_{\sigma(i)}^2 - 2\sum_i (X x_i y_{\sigma(i)}^2 + Y x_i^2 y_{\sigma(i)}) + 2(\sum_i x_i y_{\sigma(i)})^2 \tag{6.30}$$

Since Z is invariant by translation of $\mathbf{x}, \mathbf{y}$ we can suppose without loss of generality that $X = Y = 0$. We consider the set $\mathcal{D} = \{\mathbf{x}, \mathbf{y} \in \mathbb{R}^n \times \mathbb{R}^n | x_1 < x_2 < \cdots < x_n \ , y_1 < y_2 < \cdots < y_n, \ \sum_i x_i = \sum_j y_j = 0\}$. We want to find for $\mathbf{x}, \mathbf{y} \in \mathcal{D}$:

$$\max_{\sigma \in S_n} n \sum_i x_i^2 y_{\sigma(i)}^2 + 2 \left(\sum_i x_i y_{\sigma(i)} \right)^2 \overset{def}{=} \max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma) \tag{QAP}$$

We have the following result:

Lemma 6.2.1. *Let* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ *and consider the problem:*

$$\max_{\boldsymbol{\pi} \in DS} \sum_{ijkl} (x_i^2 y_j^2 + 2x_i y_j x_k y_l) \pi_{ij} \pi_{kl} \tag{QP}$$

where DS *is the set of doubly stochastic matrices. Then* (QP) *and* (QAP) *are equivalent. More precisely if* σ^* *is an optimal solution of* (QAP) *then* $\boldsymbol{\pi}_{\sigma^*}$ *defined by* $\pi_{\sigma^*}(i,j) = 1$ *if* $j = \sigma^*(i)$ *else* 0 *for all* $(i,j) \in [\![n]\!]^2$ *is an optimal solution of* (QP) *and if* $\boldsymbol{\pi}^*$ *is an optimal solution of* (QP) *then it is supported on a permutation* σ^* *which is an optimal solution of* (QAP).

Proof. The problem (QAP) can be rewritten as:

$$
\begin{aligned}
\max_{\substack{P_{ij} \in \{0,1\} \\ \forall j \sum_i P_{ij}=1 \\ \forall i \sum_j P_{ij}=1}} n \sum_{ij} x_i^2 y_j^2 P_{ij} + 2 \left(\sum_{i,j} x_i y_j P_{ij} \right)^2 &= \max_{\substack{P_{ij} \in \{0,1\} \\ \forall j \sum_i P_{ij}=1 \\ \forall i \sum_j P_{ij}=1}} n \sum_{ij} x_i^2 y_j^2 P_{ij} + 2 \sum_{ijkl} x_i x_k y_j y_l P_{ij} P_{kl} \\
\stackrel{*}{=} \max_{\substack{P_{ij} \in \{0,1\} \\ \forall j \sum_i P_{ij}=1 \\ \forall i \sum_j P_{ij}=1}} \sum_{ijkl} x_i^2 y_j^2 P_{ij} P_{kl} + 2 \sum_{ijkl} x_i y_j x_k y_l P_{ij} P_{kl} &= \max_{\substack{P_{ij} \in \{0,1\} \\ \forall j \sum_i P_{ij}=1 \\ \forall i \sum_j P_{ij}=1}} \sum_{ijkl} (x_i^2 y_j^2 + 2x_i y_j x_k y_l) P_{ij} P_{kl}
\end{aligned}
\tag{6.31}
$$

In (*) we used $\sum_{k,l} P_{k,l} = n$. We consider the following relaxation of (6.31) as:

$$\max_{\boldsymbol{\pi} \in DS} \sum_{ijkl} (x_i^2 y_j^2 + 2x_i y_j x_k y_l) \pi_{ij} \pi_{kl} \tag{6.32}$$

which is a maximization of a convex function. More preclisely it is quadratic programming problem which Hessian is positive semi-definite $\mathbf{x}\mathbf{x}^T \otimes_K \mathbf{y}\mathbf{y}^T$. Since the problem is a maximization of a convex function an optimal solution $\boldsymbol{\pi}^*$ of (QP) lies necassarily in the extremal points of DS [Rockafellar 1970] such that both (QP) and (QAP) are equivalent: if $\boldsymbol{\pi}^*$ is an optimal solution it is necessarily supported on a $\sigma^* \in S_n$ such that σ^* is an optimal solution of (QAP) and if $\sigma^* \in S_n$ is an optimal solution of (QAP) then $\boldsymbol{\pi}^*$ defined by $\pi_{ij}^* = 1$ if $j = \sigma^*(i)$ else 0 for all $(i,j) \in [\![n]\!]^2$ is an optimal solution of (QP). $\square$

Lemma 6.2.2. *Let* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$. *For* $\sigma \in S_n$ *we note* $C(\mathbf{x}, \mathbf{y}, \sigma) = \sum_i x_i y_{\sigma(i)}$. *Let* $\boldsymbol{\pi}^*$ *an optimal solution of* (QP) *with* σ^* *the permutation associated to* $\boldsymbol{\pi}^*$.
If $C(\mathbf{x}, \mathbf{y}, \sigma^*) > 0$ *then* $\boldsymbol{\pi}^* = \mathbf{I_n}$ *is the identiy and if* $C(\mathbf{x}, \mathbf{y}, \sigma^*) < 0$ *then* $\boldsymbol{\pi}^* = \mathbf{J_n}$ *is the anti-identity.*

To prove this result we will rely on the following theorem which gives necessary conditions for being an optimal solution of (QP):

Theorem (Theorem 1.12 in [Murty 1988]). *Consider the following (QP):*

$$
\begin{aligned}
\min_{\mathbf{x}} f(\mathbf{x}) \quad &= \mathbf{c}\mathbf{x} + \mathbf{x}^T \mathbf{Q}\mathbf{x} \\
s.t. \quad &\mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \geq 0
\end{aligned}
\tag{6.33}
$$

Then if $\mathbf{x}_*$ *is an optimal solution of* (6.33) *it is an optimal solution of the following (LP):*

$$
\begin{aligned}
\min_{\mathbf{x}} f(\mathbf{x}) \quad &= (\mathbf{c} + \mathbf{x}_*^T \mathbf{Q})\mathbf{x} \\
s.t. \quad &\mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \geq 0
\end{aligned}
\tag{6.34}
$$

Proof. Of Lemma 6.2.2. Applying Theorem 6.2.1 in our case gives that if $\boldsymbol{\pi}^*$ is a solution of (QP) it necessarily a solution of the following (LP):

$$\max_{\boldsymbol{\pi} \in DS} \sum_{ijkl} (x_i^2 y_j^2 + 2x_i y_j x_k y_l) \pi_{ij}^* \pi_{kl} = n \sum_{ij} x_i^2 y_j^2 \pi_{ij}^* + \max_{\boldsymbol{\pi} \in DS} 2(\sum_{ij} x_i y_j \pi_{ij}^*)(\sum_{kl} x_k y_l \pi_{kl}) \tag{6.35}$$

Since $\boldsymbol{\pi}^*$ is supported on a permutation σ^* this gives:

$$n \sum_i x_i^2 y_{\sigma^*(i)}^2 + \max_{\boldsymbol{\pi} \in DS} C(\mathbf{x}, \mathbf{y}, \sigma^*) \sum_{kl} x_k y_l \pi_{kl} \tag{LP}$$

where $C(\mathbf{x}, \mathbf{y}, \sigma^*) = 2 \left(\sum_i x_i y_{\sigma^*(i)} \right)$.

- If $C(\mathbf{x}, \mathbf{y}, \sigma^*) > 0$ then this (LP) has a unique solution which is the identity $\boldsymbol{\pi}^* = \mathbf{I_n}$. This is a consequence of the Rearrangement Inequality (see Memo 6.2.1) which states that for all permutations $\sum_i x_i y_{\sigma(i)} < \sum_i x_i y_i$ (since x_i and y_j are distinct). Using the fact that an optimal solution of (LP) is supported on a permutation concludes.

- If $C(\mathbf{x}, \mathbf{y}, \sigma^*) < 0$ then the anti-identity is the unique optimum with the same reasoning since $\sum_i x_i y_{n+1-i} < \sum_i x_i y_{\sigma(i)}$ for all permutation because of Rearrangement Inequality.

$\square$

Using both results we can prove the following proposition which is the main ingredient to prove Theorem 4.1.1:

Proposition 6.2.1. *Let* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ *and* σ^* *a solution of* (QAP) *i.e.* $\sigma^* \in \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$. *For* $\sigma \in S_n$ *we note* $C(\mathbf{x}, \mathbf{y}, \sigma) = \sum_i x_i y_{\sigma(i)}$.

If $C(\mathbf{x}, \mathbf{y}, \sigma^*) > 0$ *then* σ^* *is the identiy permutation* $\sigma^*(i) = i$ *and if* $C(\mathbf{x}, \mathbf{y}, \sigma^*) < 0$ *then* σ^* *is the anti-identity permutation* $\sigma^*(i) = n + 1 - i$ *for all* $i \in [\![n]\!]$.

Proof. Let σ^* be an optimal solution of (QAP) and $\boldsymbol{\pi}^*$ defined by $\pi_{ij}^* = 1$ if $j = \sigma^*(i)$ else 0. By Lemma 6.2.1 we know that $\boldsymbol{\pi}^*$ is an optimal solution of (QP). Consider the case $C(\mathbf{x}, \mathbf{y}, \sigma^*) > 0$. Suppose that σ^* is not the identity, then $\boldsymbol{\pi}^* \neq \mathbf{I_n}$ which is not possible by Lemma 6.2.2 since $\boldsymbol{\pi}^*$ is an optimal solution of (QP). Same applies for $C(\mathbf{x}, \mathbf{y}, \sigma^*) < 0$ and the anti-identity.

$\square$

To state that (QAP) admits the identity or the anti-identity as optimal permutations we will rely on the previous proposition and on the continuity of the loss g:

Lemma 6.2.3 (Continuity of g). *Let* $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ *fixed. There exists* $\varepsilon_{x,y} > 0$ *such that for all* $\|\mathbf{h}\| < \varepsilon_{x,y}$ *we have:*

$$\arg\max_{\sigma \in S_n} g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \sigma) \subset \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma) \tag{6.37}$$

Memo 6.2.1 (Rearrangement Inequality). *Let* $x_1 \leq \cdots \leq x_n$, $y_1 \leq \cdots \leq y_n$ *then we have:*

$$\forall \sigma \in S_n, \sum_i x_i y_{n+1-i} \leq \sum_i x_i y_{\sigma(i)} \leq \sum_i x_i y_i \tag{6.36}$$

If the numbers are different then the lower bound (resp upper bound) is attained only for the permutation which reverses the order (resp for the identiy permutation)

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, $\sigma^* \in \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$ and τ any permutation in S_n such that $\tau \notin \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$. Then we have $g(\mathbf{x}, \mathbf{y}, \sigma^*) > g(\mathbf{x}, \mathbf{y}, \tau)$. Let $\eta = g(\mathbf{x}, \mathbf{y}, \sigma^*) - g(\mathbf{x}, \mathbf{y}, \tau) > 0$. For all permutation β we have that $g(., \mathbf{y}, \beta)$ is continuous. In this way:

$$\forall \beta \in S_n, \exists \varepsilon_{\mathbf{x},\mathbf{y}}(\beta, \sigma^*, \tau) > 0, \ \forall \|\mathbf{h}\| < \varepsilon_{\mathbf{x},\mathbf{y}}(\beta, \sigma^*, \tau), \ |g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \beta) - g(\mathbf{x}, \mathbf{y}, \beta)| < \frac{\eta}{4} \tag{6.38}$$

Let $\mathbf{h} \in \mathbb{R}^n$ such that $\|\mathbf{h}\| < \min_{(\beta,\sigma,\tau') \in (S_n)^3} \varepsilon_{\mathbf{x},\mathbf{y}}(\beta, \sigma, \tau')$. By (6.38) applied to σ^* and τ:

$$\begin{aligned} g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \sigma^*) - g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \tau) &= g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \sigma^*) - g(\mathbf{x}, \mathbf{y}, \sigma^*) \\ &\quad + g(\mathbf{x}, \mathbf{y}, \sigma^*) - g(\mathbf{x}, \mathbf{y}, \tau) + g(\mathbf{x}, \mathbf{y}, \tau) - g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \tau) \\ &> -\frac{\eta}{4} + \eta - \frac{\eta}{4} \\ &= \frac{\eta}{2} > 0 \end{aligned} \tag{6.39}$$

So that $g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \sigma^*) > g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \tau)$ and in this way $\tau \notin \arg\max_{\sigma \in S_n} g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \sigma)$ because σ^* leads to a striclty better cost. Overall we have proven that for any permutation τ, if $\tau \notin \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$ and $\|\mathbf{h}\| < \min_{(\beta,\sigma,\tau') \in (S_n)^3} \varepsilon_{\mathbf{x},\mathbf{y}}(\beta, \sigma, \tau')$ then $\tau \notin \arg\max_{\sigma \in S_n} g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \sigma)$ which proves that $\arg\max_{\sigma \in S_n} g(\mathbf{x} + \mathbf{h}, \mathbf{y}, \sigma) \subset \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$. $\square$

Using the previous lemma we can now prove the following result:

Lemma 6.2.4. *Let $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ fixed. There exists $\boldsymbol{\varepsilon}_0 \in \mathbb{R}^n$ such that:*

$$\begin{aligned} \arg\max_{\sigma \in S_n} g(\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y}, \sigma) &\subset \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma) \\ \arg\max_{\sigma \in S_n} g(\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y}, \sigma) &\subset \{Id, anti - Id\} \end{aligned} \tag{6.40}$$

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathcal{D}$. We consider $\boldsymbol{\varepsilon}_0 = (\zeta, -\zeta, 0, \dots, 0)$ with $\zeta > 0$ small enough to ensure $\zeta < \frac{x_2 - x_1}{2}$ and $\|\boldsymbol{\varepsilon}_0\| < \varepsilon_{x,y}$ defined in Lemma 6.2.3. We have $\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y} \in \mathcal{D}$ since $\sum_i (x_i + \varepsilon_0(i)) = \sum_i x_i + \zeta - \zeta = 0$ and $x_1 + \varepsilon_0(1) < \cdots < x_N + \varepsilon_0(N)$ since $x_1 + \zeta < x_2 - \zeta$.

Let $\sigma^*_{\boldsymbol{\varepsilon}_0} \in \arg\max_{\sigma \in S_n} g(\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y}, \sigma)$. By Lemma 6.2.3 we have $\sigma^*_{\boldsymbol{\varepsilon}_0} \in \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$.
Moreover we have $C(\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y}, \sigma^*_{\boldsymbol{\varepsilon}_0}) = \sum_i x_i y_{\sigma^*_{\boldsymbol{\varepsilon}_0}(i)} + \zeta(y_{\sigma^*_{\boldsymbol{\varepsilon}_0}(0)} - y_{\sigma^*_{\boldsymbol{\varepsilon}_0}(1)})$.

- If $\sum_i x_i y_{\sigma^*_{\boldsymbol{\varepsilon}_0}(i)} = 0$ then $C(\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y}, \sigma^*_{\boldsymbol{\varepsilon}_0}) = \zeta(y_{\sigma^*_{\boldsymbol{\varepsilon}_0}(0)} - y_{\sigma^*_{\boldsymbol{\varepsilon}_0}(1)}) \neq 0$ since all y_i are distinct. We can apply Proposition 6.2.1 with $\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y} \in \mathcal{D}$ to conclude that $\sigma^*_{\boldsymbol{\varepsilon}_0}$ is wether the identity or the anti-identity.

- If $\sum_i x_i y_{\sigma^*_{\boldsymbol{\varepsilon}_0}(i)} \neq 0$ then $\sigma^*_{\boldsymbol{\varepsilon}_0} \in \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$ and $C(\mathbf{x}, \mathbf{y}, \sigma^*_{\boldsymbol{\varepsilon}_0}) \neq 0$ so by Proposition 6.2.1 with $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ we can conclude that $\sigma^*_{\boldsymbol{\varepsilon}_0}$ is wether the identity or the anti-identity.

$\square$

Corollary 6.2.1 (Theorem 4.1.1 is valid). *Let $\mathbf{x}, \mathbf{y} \in \mathcal{D}$. The identity or the anti-identity is an optimal solution of* (QAP)

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathcal{D}$. We consider $\boldsymbol{\varepsilon}_0$ defined in Lemma 6.2.4 and $\sigma^*_{\boldsymbol{\varepsilon}_0} \in \arg\max_{\sigma \in S_n} g(\mathbf{x} + \boldsymbol{\varepsilon}_0, \mathbf{y}, \sigma)$. Then by Lemma 6.2.4 $\sigma^*_{\boldsymbol{\varepsilon}_0}$ is wether the identity or the anti-identity. Moreover by Lemma 6.2.4 $\sigma^*_{\boldsymbol{\varepsilon}_0} \in \arg\max_{\sigma \in S_n} g(\mathbf{x}, \mathbf{y}, \sigma)$ so it is an optimal solution of (QAP). This concludes that the identity or the anti-identity is an optimal solution of (QAP) which proves Theorem 4.1.1. $\square$

6.2.2 Proof of Theorem 4.1.2 – Equivalence between GM and GW in the discrete case

This paragraph aims at proving the equivalence between GM and GW. We recall the theorem:

Theorem (Equivalence between GW and GM for discrete measures). *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$, $\nu \in \mathcal{P}(\mathbb{R}^q)$ be discrete probability measures with same number of atoms and uniform weights, i.e. $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}, \nu = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{y}_i}$ with $\mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^q$. For $\mathbf{x} \in \mathbb{R}^p$ we note $\|\mathbf{x}\|_{2,p} = \sqrt{\sum_{i=1}^p |x_i|^2}$ the ℓ_2 norm on $\mathbb{R}^p$ (same for $\mathbb{R}^q$). Let $c_\mathcal{X}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_{2,p}^2$, $c_\mathcal{Y}(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_{2,q}^2$. Then:*

$$GW_2(c_\mathcal{X}, c_\mathcal{Y}, \mu, \nu) = GM_2(c_\mathcal{X}, c_\mathcal{Y}, \mu, \nu) \tag{6.41}$$

Moreover if $p = q = 1$, i.e. $c_\mathcal{X}(x, x') = c_\mathcal{Y}(x, x') = |x - x'|^2$ for $x, x' \in \mathbb{R}$, and if $x_1 < \cdots < x_n$ and $y_1 < \cdots < y_n$ the optimal values are achieved by considering either the identity or the anti-identity permutation.

Proof. The proof is essentially based on theoretical results from [Maron 2018] and on Theorem 4.1.1. In [Maron 2018] authors consider the minimizing energy problem $\min_{\mathbf{X} \in \Pi_n} - \text{tr}(\mathbf{B}\mathbf{X}^T\mathbf{A}\mathbf{X})$ where Π_n the set of permutation matrices. In fact, the GM problem defined in this chapter is equivalent to $\min_{\mathbf{X} \in \Pi_n} -\text{tr}(\mathbf{B}\mathbf{X}^T\mathbf{A}\mathbf{X})$ by considering $\mathbf{A} = (\|\mathbf{x}_i - \mathbf{x}_j\|_{2,p}^2)_{i,j}$ and $\mathbf{B} = (\|\mathbf{y}_i - \mathbf{y}_j\|_{2,q}^2)_{i,j}$.

To tackle this problem authors propose to minimize $-\text{tr}(\mathbf{B}\mathbf{X}^T\mathbf{A}\mathbf{X})$ over the set of doubly stochastic matrices (which is the convex-hull of Π_n):

$$DS = \{\mathbf{X} \in \mathbb{R}^{n \times n} \text{ s.t. } \mathbf{X}\mathbf{1}_n = \mathbf{X}^T\mathbf{1}_n = \mathbf{1}_n \text{ , } \mathbf{X} \geq 0\}$$

Minimizing $-\text{tr}(\mathbf{B}\mathbf{X}^T\mathbf{A}\mathbf{X})$ over DS is equivalent to solving the GW distance when $a_i = b_j = \frac{1}{n}$. The paper proves that when both $\mathbf{A}$ and $\mathbf{B}$ are conditionally positive (or negative) definite of order 1 then the relaxation leads to the same optimum so that the minimum over DS is the same as the minimum over Π_n [Maron 2018, Theorem 1]. Yet $\mathbf{A}$ and $\mathbf{B}$ defined previously satisfy this property (see examples under Definition 2 in [Maron 2018]) and so GW and GM coincide.

Moreover when $p = q = 1$ and when the sample are sorted we can apply Theorem 4.1.1 to prove that an optimal permutation of the GM problem is found whether at the identity or the anti-identity permutation which concludes the proof.

$\square$

6.2.3 Computing GW in the 1d case

We recall the result:

Lemma. *The GM_2 and GW_2 costs in 1D with same numbers of atoms and uniform weights can be computed in $O(n)$.*

Proof. As seen in Theorem 4.1.2 finding the optimal permutation σ^* can be found in $O(n\log(n))$. Moreover

the final costs can be written using binomial expansion:

$$
\begin{aligned}
\sum_{i,j} \left((x_i - x_j)^2 - (y_{\sigma^*(i)} - y_{\sigma^*(j)})^2 \right)^2 = {}& 2n \sum_i x_i^4 - 8 \sum_i x_i^3 \sum_k x_k + 6(\sum_i x_i^2)^2 \\
& + 2n \sum_i y_i^4 - 8 \sum_i y_i^3 \sum_k y_k + 6(\sum_i y_i^2)^2 \\
& - 4(\sum_i x_i)^2 (\sum_k y_k)^2 \\
& - 4n \sum_i x_i^2 y_{\sigma^*(i)}^2 + 8 \sum_i ((\sum_k x_k) x_i y_{\sigma^*(i)}^2 + (\sum_k y_k) x_i^2 y_{\sigma^*(i)}) \\
& - 8(\sum_i x_i y_{\sigma^*(i)})^2
\end{aligned}
\tag{6.42}
$$

which can be computed in $O(n)$ operations. $\qquad\square$

6.2.4 Proof of Theorem 4.1.3 – Properties of SGW

We recall the theorem:

Theorem (Properties of SGW).
- *For all Δ, SGW_Δ and $RISGW$ are translation invariant. $RISGW$ is also rotational invariant when $p = q$, more precisely if $\mathbf{Q} \in \mathcal{O}(p)$ is an orthogonal matrix, $RISGW(\mathbf{Q}\#\mu, \nu) = RISGW(\mu, \nu)$ (same for any $\mathbf{Q}' \in \mathcal{O}(q)$ applied on ν).*

- *SGW and $RISGW$ are pseudo-distances on $\mathcal{P}(\mathbb{R}^p)$, i.e. they are symmetric, satisfy the triangle inequality and $SGW(\mu, \mu) = RISGW(\mu, \mu) = 0$.*

- *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ be probability distributions with compact supports. If $SGW(\mu, \nu) = 0$ then μ and ν are isomorphic for the distance induced by the ℓ_1 norm on $\mathbb{R}^p$, i.e. $d(x, x') = \sum_{i=1}^p |x_i - x_i'|$ for $(x, x') \in \mathbb{R}^p \times \mathbb{R}^p$. In particular this implies:*

$$
SGW(\mu, \nu) = 0 \implies GW_2(d, d, \mu, \nu) = 0
\tag{6.43}
$$

The invariance by translation is clear since the costs are invariant by translation of the support of the measures. The pseudo-distances properties are straightforward thanks to the properties of GW. For the invariance by rotation if $p = q$ then $\mathbb{V}_p(\mathbb{R}^p)$ is bijective with $\mathcal{O}(p)$ so for $Q \in \mathcal{O}(p)$:

$$
\begin{aligned}
RISGW(Q\#\mu, \nu) &= \min_{\Delta \in \mathbb{V}_p(\mathbb{R}^p)} SGW_\Delta(Q\#\mu, \nu) \\
&= \min_{\Delta \in \mathcal{O}(p)} SGW_\Delta(Q\#\mu, \nu) \\
&= \min_{\Delta \in \mathcal{O}(p)} \mathbb{E}_{\theta \sim \lambda_{q-1}} [GW(d^2, P_\theta\#(\Delta Q\#\mu), P_\theta\#\nu)] \\
&= \min_{\Delta' \in \mathcal{O}(p)} \mathbb{E}_{\theta \sim \lambda_{q-1}} [GW(d^2, P_\theta\#\Delta'\#\mu, P_\theta\#\nu)] \\
&= RISGW(\mu, \nu)
\end{aligned}
\tag{6.44}
$$

On the other side for ν a change of formula on theta gives the result.

For the case $SGW = 0 \implies GW = 0$ it will be a consequence of the following theorem:

> **Theorem 6.2.1.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ be probability distributions such that μ, ν have compact supports. If for almost all $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$, $P_{\boldsymbol{\theta}} \# \mu$ and $P_{\boldsymbol{\theta}} \# \nu$ are isomoprhic then μ and ν are isomoprhic. In other words if for almost all $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$ we have:*
>
> $$\exists T_{\boldsymbol{\theta}} : supp(P_{\boldsymbol{\theta}} \# \mu) \subset \mathbb{R} \mapsto supp(P_{\boldsymbol{\theta}} \# \nu) \subset \mathbb{R}, \ surjective \ s.t. \ T_{\boldsymbol{\theta}} \#(P_{\boldsymbol{\theta}} \# \mu) = P_{\boldsymbol{\theta}} \# \nu$$
> $$\forall x, x' \in supp(P_{\boldsymbol{\theta}} \# \mu), |T_{\boldsymbol{\theta}}(x) - T_{\boldsymbol{\theta}}(x')| = |x - x'| \tag{6.45}$$
>
> *Then there exists a measure preserving isometry f between $supp(\mu)$ and $supp(\nu)$. More precisely we have $f \# \mu = \nu$ and:*
> $$\forall \mathbf{x}, \mathbf{x}' \in supp(\mu), \|f(\mathbf{x}) - f(\mathbf{x}')\|_1 = \|\mathbf{x} - \mathbf{x}'\|_1 \tag{6.46}$$

To prove this theorem we will exhibit the isometry. This result can be put in light of Cramer–Wold theorem [Cramér 1936] which states that a probability measure is uniquely determined by the totality of its one-dimensional projections. Equivalently, if we consider two probability measures so that the one-dimensional measures resulting from the projections over all the hypersphere are equal then the measures are equal. The equality relation is replaced in our theorem by the isomoprhism relation.

The proof is divided into four parts. In the first one, we construct an "almost orthogonal" basis on which measures are isomorphic. Building upon this result we define a sequence of functions from $supp(\mu)$ to $supp(\nu)$ and show that it has a convergent subsequence. We conclude by proving that the limit of the subsequence is actually a good candidate for being the isometry we are looking for. In the following $\|.\|_1$ denotes the ℓ_1 norm, $\|.\|_2$ denotes the ℓ_2 norm and $p \geq 2$. We recall that $\mathcal{F}_\mu$ is the Fourier transform of μ.

We consider the following $\mathcal{Q}_{\boldsymbol{\theta}}$ property for $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$:

$$\exists T_{\boldsymbol{\theta}} : supp(P_{\boldsymbol{\theta}} \# \mu) \subset \mathbb{R} \mapsto supp(P_{\boldsymbol{\theta}} \# \nu) \subset \mathbb{R}, \ surjective \ s.t. \ T_{\boldsymbol{\theta}} \#(P_{\boldsymbol{\theta}} \# \mu) = P_{\boldsymbol{\theta}} \# \nu$$
$$\forall x, x' \in supp(P_{\boldsymbol{\theta}} \# \mu), |T_{\boldsymbol{\theta}}(x) - T_{\boldsymbol{\theta}}(x')| = |x - x'| \tag{$\mathcal{Q}_{\boldsymbol{\theta}}$}$$

Informally if we have the $\mathcal{Q}_{\boldsymbol{\theta}}$ property for $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$ it implies that μ and ν are isomorphic on the 1D line given by the projection *w.r.t.* $\boldsymbol{\theta}$. We have the following lemma:

Lemma 6.2.5. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ and suppose that $\mathcal{Q}_{\boldsymbol{\theta}}$ holds for almost all $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$. Let $n > p - 1$. There exists a basis $(\mathbf{e}_1(n), ..., \mathbf{e}_p(n))$ of $\mathbb{R}^p$ part of the following spaces:*

$$\mathcal{B}_p^n \stackrel{def}{=} \{(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p) \in (\mathbb{S}^{p-1})^p \ s.t. \ |\langle \boldsymbol{\theta}_i, \boldsymbol{\theta}_j \rangle| < \frac{1}{n}\} \tag{6.47}$$

and

$$Q \stackrel{def}{=} \{(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p) \in (\mathbb{S}^{p-1})^p \ s.t. \ \forall i \in \{1, ..., p\}, \mathcal{Q}_{\boldsymbol{\theta}_i}\} \tag{6.48}$$

Proof. We want to construct a basis $(\mathbf{e}_1, ..., \mathbf{e}_p)$ as orthogonal as possible such that for all i we have $\mathcal{Q}_{\mathbf{e}_i}$.

We note $\lambda_{p-1}^{\otimes p}$ the product measure $\lambda_{p-1} \otimes ... \otimes \lambda_{p-1}$ where λ_{p-1} is the uniform measure on the sphere. $\mathcal{B}_p^n$ is an open set as inverse image by a continuous function of an open set. Then $\lambda_{p-1}^{\otimes p}(\mathcal{B}_p^n) > 0$. Moreover, since for almost all $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$ we have $\mathcal{Q}_{\boldsymbol{\theta}}$ then $\lambda_{p-1}^{\otimes p}(Q) > 0$ and so $\lambda_{p-1}^{\otimes p}(\mathcal{B}_p^n \cap Q) > 0$.

In this way we can consider $(\mathbf{e}_1(n), ..., \mathbf{e}_p(n)) \in \mathcal{B}_p^n \cap Q$. If $n > p - 1$ the Gram matrix of $(\mathbf{e}_1(n), ..., \mathbf{e}_p(n))$ is strictly diagonal dominant, thus invertible, such that $(\mathbf{e}_1(n), ..., \mathbf{e}_p(n))$ is a basis. Note that we can not consider directly an orthogonal basis since the set of all orthogonal basis has measure zero. $\qquad \square$

We now express all the vectors and inner products in this new almost orthogonal basis as expressed in the following lemma:

Lemma 6.2.6. *Let $n > p - 1$ and a basis $(\mathbf{e}_1(n), ..., \mathbf{e}_p(n))$ as defined in Lemma 6.2.5. Then all $\mathbf{x} \in \mathbb{R}^p$ can be written as:*

$$\mathbf{x} = \sum_{i=1}^{p} [\langle \mathbf{x}, \mathbf{e}_i(n) \rangle + R(\mathbf{x}, \mathbf{e}_i(n))] \mathbf{e}_i(n) \tag{6.49}$$

where $|R(\mathbf{x}, \mathbf{e}_i(n))| = o(\frac{1}{n})$. Moreover for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^p \times \mathbb{R}^p$:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{p} \langle \mathbf{x}, \mathbf{e}_i(n) \rangle \langle \mathbf{y}, \mathbf{e}_i(n) \rangle + \tilde{R}(\mathbf{x}, \mathbf{y}) \tag{6.50}$$

where $|\tilde{R}(\mathbf{x}, \mathbf{y})| = o(\frac{1}{n})$.

Proof. In the following x_i denotes the i-th coordinate of a vector $\mathbf{x}$ in the standard basis, *i.e.* a vector writes $\mathbf{x} = (x_1, \ldots, x_p)$. For $\mathbf{x} \in \mathbb{R}^p$, we can write in the new basis $\mathbf{x} = \sum_{i=1}^{p} [\langle \mathbf{x}, \mathbf{e}_i(n) \rangle + R(\mathbf{x}, \mathbf{e}_i(n))] \mathbf{e}_i$ with $R(\mathbf{x}, \mathbf{e}_i(n)) \stackrel{def}{=} x_i - \langle \mathbf{x}, \mathbf{e}_i(n) \rangle$. We have also $|R(\mathbf{x}, \mathbf{e}_i(n))| = o(\frac{1}{n})$. Indeed,

$$\mathbf{x} = \sum_{i=1}^{p} x_i \mathbf{e}_i \implies \forall j, \langle \mathbf{x}, \mathbf{e}_j \rangle = \sum_{i=1}^{p} x_i \langle \mathbf{e}_i, \mathbf{e}_j \rangle \implies x_j - \langle \mathbf{x}, \mathbf{e}_j \rangle = \sum_{i \neq j} x_i \langle \mathbf{e}_i, \mathbf{e}_j \rangle$$

$$\implies |R(\mathbf{x}, \mathbf{e}_j(n))| = |\sum_{i \neq j} x_i \langle \mathbf{e}_i, \mathbf{e}_j \rangle| \implies |R(\mathbf{x}, \mathbf{e}_j(n))| \leq \frac{1}{n} \sum_{i \neq j} |x_i|$$

Also in the same way for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p \times \mathbb{R}^p$ we can rewrite their inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{p} \langle \mathbf{x}, \mathbf{e}_i(n) \rangle \langle \mathbf{y}, \mathbf{e}_i(n) \rangle + \tilde{R}(\mathbf{x}, \mathbf{y}) \tag{6.51}$$

with:

$$\tilde{R}(\mathbf{x}, \mathbf{y}) \stackrel{def}{=} \langle \mathbf{x}, \mathbf{y} \rangle - \sum_{i=1}^{p} \langle \mathbf{x}, \mathbf{e}_i(n) \rangle \langle \mathbf{y}, \mathbf{e}_i(n) \rangle$$

$$= \sum_{i \neq j} \langle \mathbf{x}, \mathbf{e}_i(n) \rangle \langle \mathbf{y}, \mathbf{e}_i(n) \rangle \langle \mathbf{e}_j(n), \mathbf{e}_i(n) \rangle + \sum_{i,j} \langle \mathbf{x}, \mathbf{e}_i(n) \rangle R(\mathbf{y}, \mathbf{e}_j(n)) \langle \mathbf{e}_j(n), \mathbf{e}_i(n) \rangle$$

$$+ \sum_{i,j} \langle \mathbf{y}, \mathbf{e}_j(n) \rangle R(\mathbf{x}, \mathbf{e}_i(n)) \langle \mathbf{e}_j(n), \mathbf{e}_i(n) \rangle + \sum_{i,j} R(\mathbf{x}, \mathbf{e}_j(n)) R(\mathbf{y}, \mathbf{e}_i(n)) \langle \mathbf{e}_j(n), \mathbf{e}_i(n) \rangle$$

and with the same calculus than for R we have $|\tilde{R}(\mathbf{x}, \mathbf{y})| = o(\frac{1}{n})$. $\square$

Proposition 6.2.2. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ and suppose that $\mathcal{Q}_{\boldsymbol{\theta}}$ holds for almost all $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$ and that ν has compact support. There exists a sequence $(f_n)_{n \in \mathbb{N}}$ from $supp(\mu)$ to $supp(\nu)$ uniformly bounded which satisfies:*

$$\forall n \in \mathbb{N}, \forall \mathbf{x}, \mathbf{x}' \in supp(\mu)^2, \ \left| \|f_n(\mathbf{x}) - f_n(\mathbf{x}')\|_1 - \|\mathbf{x} - \mathbf{x}'\|_1 \right| = o\left(\frac{1}{n}\right) \tag{6.52}$$

$$\forall n \in \mathbb{N}, \forall \mathbf{s} \in \mathbb{R}^p, \ |\mathcal{F}_{f_n \# \mu}(\mathbf{s}) - \mathcal{F}_\nu(\mathbf{s})| = o\left(\frac{1}{n}\right) \tag{6.53}$$

Proof. In the following x_i denotes the i-th coordinate of a vector $\mathbf{x}$ in the standard basis, *i.e.* a vector writes $\mathbf{x} = (x_1, \ldots, x_p)$. We define:

$$\forall n > p - 1, \ \forall \mathbf{x} \in \text{supp}(\mu), \ f_n(\mathbf{x}) = (T_{\mathbf{e}_1(n)}(\langle \mathbf{x}, \mathbf{e}_1(n) \rangle), ..., T_{\mathbf{e}_p(n)}(\langle \mathbf{x}, \mathbf{e}_p(n) \rangle)) \tag{6.54}$$

where $(\mathbf{e}_k(n))_{k \in [\![p]\!]}$ is the almost orthogonal basis define in Lemma 6.2.5, and $T_{\mathbf{e}_k(n)}$ is defined from $(\mathcal{Q}_{\boldsymbol{\theta}})$ since we have $\mathcal{Q}_{\mathbf{e}_k(n)}$ for all k. It is clear from the definition that $f_n(\mathbf{x}) \in \text{supp}(\nu)$. Moreover for $\mathbf{x}, \mathbf{x}' \in \text{supp}(\mu)$:

$$\|f_n(\mathbf{x}) - f_n(\mathbf{x}')\|_1 = \sum_{k=1}^{p} |T_{\mathbf{e}_k(n)}(\langle \mathbf{x}, \mathbf{e}_k(n) \rangle) - T_{\mathbf{e}_k(n)}(\langle \mathbf{x}', \mathbf{e}_k(n) \rangle)| \overset{(*)}{=} \sum_{k=1}^{p} |\langle \mathbf{x}, \mathbf{e}_k(n) \rangle - \langle \mathbf{x}', \mathbf{e}_k(n) \rangle|$$

$$= \sum_{k=1}^{p} |\langle \mathbf{x} - \mathbf{x}', \mathbf{e}_k(n) \rangle|$$

where in (*) we used that $T_{\mathbf{e}_k(n)}$ is an isometry since we have $\mathcal{Q}_{\mathbf{e}_k(n)}$ and $\langle \mathbf{x}, \mathbf{e}_k(n) \rangle \in \text{supp}(P_{\mathbf{e}_k(n)} \# \mu)$ (idem for $\mathbf{x}'$). In this way:

$$\left| \|f_n(\mathbf{x}) - f_n(\mathbf{x}')\|_1 - \|\mathbf{x} - \mathbf{x}'\|_1 \right| = \left| \sum_{k=1}^{p} |\langle \mathbf{x} - \mathbf{x}', \mathbf{e}_k(n) \rangle| - |x_k - x_k'| \right| \le \sum_{k=1}^{p} \left| |\langle \mathbf{x} - \mathbf{x}', \mathbf{e}_k(n) \rangle| - |x_k - x_k'| \right|$$

$$\overset{*}{\le} \sum_{k=1}^{p} |\langle \mathbf{x} - \mathbf{x}', \mathbf{e}_k(n) \rangle - (x_k - x_k')| = \sum_{k=1}^{p} |R(\mathbf{x} - \mathbf{x}', \mathbf{e}_k(n))| = o(\frac{1}{n})$$

where in (*) the second triangular inequality $||x| - |y|| \le |x - y|$. Hence:

$$\left| \|f_n(\mathbf{x}) - f_n(\mathbf{x}')\|_1 - \|\mathbf{x} - \mathbf{x}'\|_1 \right| = o(\frac{1}{n}) \tag{6.55}$$

Moreover we have by definition of the Fourier transform, for $s \in \mathbb{R}^P$,

$$\mathcal{F}_{f_n \# \mu}(\mathbf{s}) = \int e^{-2i\pi \langle \mathbf{s}, f_n(\mathbf{x}) \rangle} d\mu(\mathbf{x}) = \int e^{-2i\pi \sum_{k=1}^{p} s_k T_{\mathbf{e}_k(n)}(\langle \mathbf{x}, \mathbf{e}_k(n) \rangle)} d\mu(\mathbf{x})$$

Moreover using $(\mathcal{Q}_{\boldsymbol{\theta}})$ we have $\mathcal{F}_{T_{\mathbf{e}_k(n)} \# (P_{\mathbf{e}_k(n)} \# \mu)}(t) = \mathcal{F}_{P_{\mathbf{e}_k(n)} \# \nu}(t)$ for all $k \in \{1, ..., p\}$, and any real $t \in \mathbb{R}$. This implies $\int e^{-2i\pi t . T_{\mathbf{e}_k(n)}(\langle \mathbf{e}_k(n), \mathbf{x} \rangle)} d\mu(\mathbf{x}) = \int e^{-2i\pi t \langle \mathbf{e}_k(n), \mathbf{y} \rangle} d\nu(\mathbf{y})$. So by applying this results for $t = s_k$ we have:

$$\int e^{-2i\pi s_k T_{\mathbf{e}_k(n)}(\langle \mathbf{x}, \mathbf{e}_k(n) \rangle)} d\mu(\mathbf{x}) = \int e^{-2i\pi s_k \langle \mathbf{e}_k(n), y \rangle} d\nu(\mathbf{y}) \tag{6.56}$$

Combining both results:

$$\mathcal{F}_{f_n \# \mu}(\mathbf{s}) = \int e^{-2i\pi \sum_{k=1}^{p} s_k \langle \mathbf{e}_k(n), \mathbf{y} \rangle} d\nu(\mathbf{y}) \tag{6.57}$$

We can now bound $|\mathcal{F}_{f_n \# \mu}(\mathbf{s}) - \mathcal{F}_{\nu}(\mathbf{s})|$ as:

$$|\mathcal{F}_{f_n \# \mu}(\mathbf{s}) - \mathcal{F}_{\nu}(\mathbf{s})| = |\mathcal{F}_{f_n \# \mu}(\mathbf{s}) - \int e^{-2i\pi \langle \mathbf{s}, \mathbf{y} \rangle} d\nu(\mathbf{y})|$$

$$\overset{*}{=} |\mathcal{F}_{f_n \# \mu}(\mathbf{s}) - \int e^{-2i\pi [\sum_{k=1}^{p} \langle \mathbf{s}, \mathbf{e}_k(n) \rangle \langle \mathbf{e}_k(n), \mathbf{y} \rangle + \tilde{R}(\mathbf{s}, \mathbf{y})]} d\nu(\mathbf{y})|$$

$$\overset{**}{=} |\int e^{-2i\pi \sum_{k=1}^{p} s_k \langle \mathbf{e}_k(n), \mathbf{y} \rangle} d\nu(\mathbf{y}) - \int e^{-2i\pi \tilde{R}(\mathbf{s}, \mathbf{y})} e^{-2i\pi \sum_{k=1}^{p} \langle \mathbf{s}, \mathbf{e}_k(n) \rangle \langle \mathbf{e}_k(n), \mathbf{y} \rangle} d\nu(\mathbf{y})|$$

where in (*) we used the expression in the new base of the inner product $\langle \mathbf{s}, \mathbf{y} \rangle$ seen in Lemma 6.2.6, in

(**) we used (6.57). By injecting the expression of s_k *w.r.t.* the new base we have:

$$|\mathcal{F}_{f_n\#\mu}(\mathbf{s}) - \mathcal{F}_\nu(\mathbf{s})| \leq |\int e^{-2i\pi\sum_{k=1}^p(\langle\mathbf{s},\mathbf{e}_k(n)\rangle+R(\mathbf{s},\mathbf{e}_k(n)))\langle\mathbf{e}_k(n),\mathbf{y}\rangle}d\nu(\mathbf{y}) - \int e^{-2i\pi\tilde{R}(\mathbf{s},\mathbf{y})}e^{-2i\pi\sum_{k=1}^p\langle s,\mathbf{e}_k(n)\rangle\langle\mathbf{e}_k(n),y\rangle}d\nu(\mathbf{y})|$$

$$= |\int e^{-2i\pi\sum_{k=1}^p\langle\mathbf{s},\mathbf{e}_k(n)\rangle\langle\mathbf{e}_k(n),\mathbf{y}\rangle}(e^{-2i\pi\sum_{k=1}^p R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),\mathbf{y}\rangle} - e^{-2i\pi\tilde{R}(\mathbf{s},\mathbf{y})})d\nu(\mathbf{y})|$$

$$\leq \int |e^{-2i\pi\sum_{k=1}^p R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),\mathbf{y}\rangle} - e^{-2i\pi\tilde{R}(\mathbf{s},\mathbf{y})}|d\nu(\mathbf{y})$$

$$= \int |e^{-2i\pi\tilde{R}(\mathbf{s},\mathbf{y})}(e^{-2i\pi(\sum_{k=1}^p R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),y\rangle-\tilde{R}(\mathbf{s},\mathbf{y}))} - 1)|d\nu(\mathbf{y})$$

$$\leq \int |e^{-2i\pi(\sum_{k=1}^p R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),\mathbf{y}\rangle-\tilde{R}(\mathbf{s},\mathbf{y}))} - 1|d\nu(\mathbf{y})$$

$$= \int |2ie^{-i\pi(\sum_{k=1}^p R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),\mathbf{y}\rangle-\tilde{R}(\mathbf{s},\mathbf{y}))}\sin(\pi(\sum_{k=1}^p R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),\mathbf{y}\rangle - \tilde{R}(\mathbf{s},\mathbf{y}))|d\nu(\mathbf{y})$$

$$\leq \int |\sin(\pi(\sum_{k=1}^p R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),\mathbf{y}\rangle - \tilde{R}(\mathbf{s},\mathbf{y}))|d\nu(\mathbf{y})$$

$$\leq \pi\int(\sum_{k=1}^p |R(\mathbf{s},\mathbf{e}_k(n))\langle\mathbf{e}_k(n),\mathbf{y}\rangle| + |\tilde{R}(\mathbf{s},\mathbf{y})|)d\nu(\mathbf{y})$$

$$\overset{*}{=} o(\frac{1}{n})$$

$$\tag{6.58}$$

in (*) the fact that each term is $o(\frac{1}{n})$. In this way:

$$|\mathcal{F}_{f_n\#\mu}(s) - \mathcal{F}_\nu(s)| = o(\frac{1}{n}) \tag{6.59}$$

Moreover $(f_n)_{n>p-1}$ is also uniformly bounded. To see that we consider $\mathbf{x} \in \text{supp}(\mu)$. We have that for all $k \in [\![p]\!]$ $T_{\mathbf{e}_k(n)}(\langle\mathbf{x},e_k(n)\rangle) \in \text{supp}(P_{\mathbf{e}_k(n)}\#\nu)$ by definition of $T_{\mathbf{e}_k(n)}$. So there exists a $\mathbf{y}_0(\mathbf{x},n,k) \in \text{supp}(\nu)$ such that $T_{\mathbf{e}_k(n)}(\langle\mathbf{x},\mathbf{e}_k(n)\rangle) = \langle\mathbf{y}_0(x,n,k),\mathbf{e}_k(n)\rangle$. In this way $|T_{\mathbf{e}_k(n)}(\langle\mathbf{x},\mathbf{e}_k(n)\rangle)| = |\langle\mathbf{y}_0(x,n,k),\mathbf{e}_k(n)\rangle| \leq \|\mathbf{y}_0(x,n,k)\|_2\|\mathbf{e}_k(n)\|_2$ by Cauchy-Swartz.

Moreover $\|\mathbf{e}_k(n)\|_2 < \sqrt{\frac{1}{n}} \leq \sqrt{\frac{1}{p-1}} \leq 1$ and since ν has compact support then there is a constant M_ν we have $\|\mathbf{y}_0(x,n,k)\|_2 \leq M_\nu$

So we have for $n \in \mathbb{N}$, $\mathbf{x} \in \text{supp}(\mu)$,

$$\|f_n(\mathbf{x})\|_2^2 = \sum_{k=1}^p |T_{\mathbf{e}_k(n)}(\langle\mathbf{x},\mathbf{e}_k(n)\rangle)|^2 \leq pM_\nu$$

Since on $\mathbb{R}^p$ all norms are equivalent this suffices to state the existence of a constant C such that $\forall\mathbf{x} \in \mathbb{R}^p, n \in \mathbb{N}, \|f_n(\mathbf{x})\|_1 \leq C$ so that $(f_n)_{n\in\mathbb{N}}$ is uniformly bounded. Reindexing $(f_n)_{n>p-1}$ gives the desired result.

$$\square$$

We can now prove Theorem 6.2.1.

Proof of Theorem 6.2.1. We consider the sequence $(f_n)_{n\in\mathbb{N}}$ defined in Proposition 6.2.2. We will show that $(f_n)_{n\in\mathbb{N}}$ is equicontinuous. Let $\varepsilon > 0$, using (6.52) there exists a $N \in \mathbb{N}$ such that we have for all $\mathbf{x},\mathbf{x}' \in \text{supp}(\mu)$:

$$\|f_n(\mathbf{x}) - f_n(\mathbf{x}')\|_1 \leq \varepsilon + \|\mathbf{x} - \mathbf{x}'\|_1 \quad \text{for all } n \geq N \tag{6.60}$$

Now let $\delta < \varepsilon$. Suppose that $\|\mathbf{x} - \mathbf{x}'\|_1 < \delta$ then

$$\|f_n(x) - f_n(x')\|_1 < \varepsilon + \delta < 2\varepsilon \quad \text{for all } n \geq N \tag{6.61}$$

Without loss of generality we can reindex $(f_n)_{n \in \mathbb{N}}$ for n large enough $(n \geq N)$ so that $(f_n)_{n \in \mathbb{N}}$ is equicontinuous with the previous argument.

Since $(f_n)_{n \in \mathbb{N}}$ is a uniformly bounded and equicontinuous sequence from the support of μ which is compact to $\mathbb{R}^p$ we can apply Arzela-Ascoli theorem (see Memo 6.2.2) which states that $(f_n)_{n \in \mathbb{N}}$ has a uniformly convergent subsequence. We denote by $(f_{\phi(n)})_n$ this sequence. We have $f_{\phi(n)} \underset{n \to \infty}{\overset{u}{\to}} f$ this sequence.

Moreover equation (6.53) states that for all $\mathbf{s} \in \mathbb{R}^p$, $\mathcal{F}_{f_n \# \mu}(\mathbf{s}) \underset{n \to \infty}{\to} \mathcal{F}_\nu(\mathbf{s})$. In this way $(\mathcal{F}_{f_n \# \mu}(\mathbf{s}))_{n \in \mathbb{N}}$ is a convergent real valued sequence, so every adherence value goes to the same limit, hence $\mathcal{F}_{f_{\phi(n)} \# \mu}(\mathbf{s}) \underset{n \to \infty}{\to} \mathcal{F}_\nu(\mathbf{s})$.

Moreover the function f is a measure preserving isometry from $\operatorname{supp}(\mu)$ to $\operatorname{supp}(\nu)$. Indeed let $\varepsilon_1 > 0, \mathbf{s} \in \mathbb{R}^p$, there exists from previous statements $N_0, N_1 \in \mathbb{N}$ such that for $n \geq N_0, |\mathcal{F}_{f_{\phi(n)} \# \mu}(\mathbf{s}) - \mathcal{F}_\nu(\mathbf{s})| < \varepsilon_1$ and $n \geq N_1, |\mathcal{F}_{f_{\phi(n)} \# \mu}(\mathbf{s}) - \mathcal{F}_{f \# \mu}(\mathbf{s})| < \varepsilon_1$. Let $n \geq \max(N_0, N_1)$

$$
\begin{aligned}
|\mathcal{F}_{f \# \mu}(\mathbf{s}) - \mathcal{F}_\nu(\mathbf{s})| &\leq |\mathcal{F}_{f_{\phi(n)} \# \mu}(\mathbf{s}) - \mathcal{F}_\nu(\mathbf{s})| + |\mathcal{F}_{f_{\phi(n)} \# \mu}(s) - \mathcal{F}_{f \# \mu}(\mathbf{s})| \\
&< 2\varepsilon_1
\end{aligned}
$$

As this result holds for any $\varepsilon_1 > 0$ we have $\mathcal{F}_{f \# \mu}(\mathbf{s}) = \mathcal{F}_\nu(\mathbf{s})$ and by injectivity of the Fourrier transform $f \# \mu = \nu$ such that f is measure preserving.

In the same way for any $\mathbf{x}, \mathbf{x}' \in \operatorname{supp}(\mu), \varepsilon_2 > 0$ and n large enough:

$$
\begin{aligned}
\big| \|f(\mathbf{x}) - f(\mathbf{x}')\|_1 - \|\mathbf{x} - \mathbf{x}'\|_1 \big| &\leq \big| \|f_{\phi(n)}(\mathbf{x}) - f_{\phi(n)}(\mathbf{x}')\|_1 - \|f(\mathbf{x}) - f(\mathbf{x}')\|_1 \big| \\
&\quad + \big| \|f_{\phi(n)}(\mathbf{x}) - f_{\phi(n)}(\mathbf{x}')\|_1 - \|\mathbf{x} - \mathbf{x}'\|_1 \big| < 2\varepsilon_2
\end{aligned}
$$

using $f_{\phi(n)} \underset{n \to \infty}{\overset{u}{\to}} f$ and (6.52). As this result holds true for any $\varepsilon_2 > 0$ we have $\|f(\mathbf{x}) - f(\mathbf{x}')\| = \|\mathbf{x} - \mathbf{x}'\|$ for any $\mathbf{x}, \mathbf{x}' \in \operatorname{supp}(\mu)$ which concludes.

$\square$

Corollary 6.2.2. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$ with compact support. If $SGW(\mu, \nu) = 0$ then μ and ν are isomorphic for the distance induced by the ℓ_1 norm on $\mathbb{R}^p$, i.e. $d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^p |x_i - x_i'|$ for*

Memo 6.2.2. *Let $(\mathcal{X}, d)$ be a compact metric space and $\|.\|$ a norm on $\mathbb{R}^p$. We say taht:*

- *A family $\mathcal{F} \subset C(\mathcal{X}, \mathbb{R}^p)$ is* bounded *means if there exists a positive constant $M < \infty$ such that $\|f(x)\| \leq M$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$*

- *A family $\mathcal{F} \subset C(X, \mathbb{R}^p)$ is* equicontinuous *means if for every $\varepsilon > 0$ there exists $\delta > 0$ (which depends only on ε) such that for $x, y \in \mathcal{X}$:*

$$d(x, y) < \delta \Rightarrow \|f(x) - f(y)\| < \varepsilon \quad \forall f \in \mathcal{F} \tag{6.62}$$

If $(f_n)_{n \in \mathbb{N}}$ is a sequence in $C(X, \mathbb{R}^p)$ that is bounded and equicontinuous then the Arzela-Ascoli states that it has a uniformly convergent subsequence (see Theorem 7.25 [Rudin 1976]) .
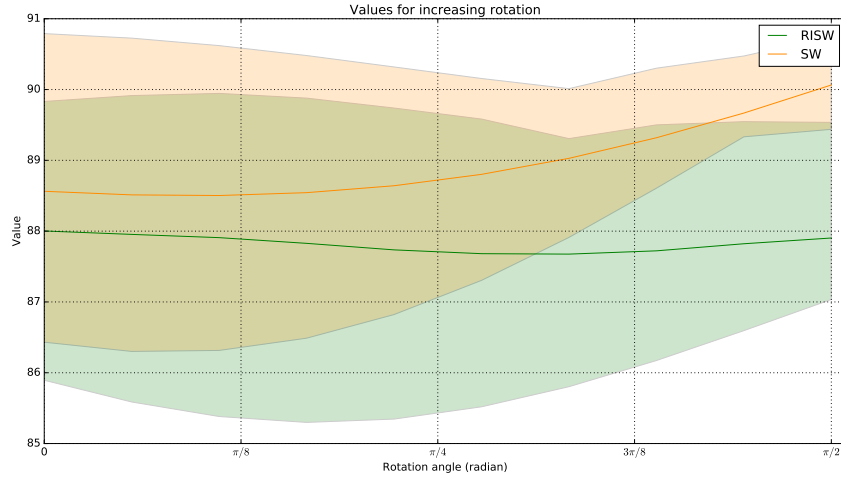
Figure 6.1: Illustration of SW, $RISW$ on spiral datasets for varying rotations on discrete 2D spiral datasets. (left) Examples of spiral distributions for source and target with different rotations. (right) Average value of SW and $RISW$ with $L = 20$ as a function of rotation angle of the target. Colored areas correspond to the 20% and 80% percentiles.

$(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^p \times \mathbb{R}^p$. *In particular this implies:*

$$SGW(\mu, \nu) = 0 \implies GW_2(d, d, \mu, \nu) = 0 \tag{6.63}$$

Proof. If $SGW(\mu, \nu) = 0$ then using the Gromov-Wasserstein properties it implies that for almost all $\boldsymbol{\theta} \in \mathbb{S}^{p-1}$ the projected measures are isomorphic. Moreover since μ, ν have compact support, it is bounded and we can directly apply Theorem 6.2.1 to state the existence of a measure preserving application f as defined in Theorem 6.2.1. We consider the coupling $\pi = (id \times f)\#\mu \in \Pi(\mu, \nu)$ since $f\#\mu = \nu$. Then we have:

$$\int \int |d(\mathbf{x}, \mathbf{x}') - d(\mathbf{y}, \mathbf{y}')|^2 d\pi(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}', \mathbf{y}') = \int \int |d(\mathbf{x}, \mathbf{x}') - d(f(\mathbf{x}), f(\mathbf{x}'))|^2 d\mu(\mathbf{x}) d\mu(\mathbf{x}')$$

$$= \int \int |\|\mathbf{x} - \mathbf{x}'\|_1 - \|f(\mathbf{x}) - f(\mathbf{x}')\|_1|^2 d\mu(\mathbf{x}) d\mu(\mathbf{x}') = 0$$

Since f is an isometry. This directly implies that $GW_2(d, d, \mu, \nu) = 0$. □

6.2.5   Additional results − SW_Δ and $RISW$

Analogously to SGW we can define for the Sliced-Wasserstein distance $SW_\Delta(\mu, \nu)$ for $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^q)$ with $p \neq q$ and its rotational invariant counterpart as:

$$\begin{aligned}
SW_\Delta(\mu, \nu) &= \int_{\mathbf{S}^{q-1}} SW(P_\theta \# \mu_\Delta, P_\theta \# \nu) d\lambda_{q-1}(\theta) \\
RISW(\mu, \nu) &= \min_{\Delta \in \mathbb{V}_q(\mathbb{R}^p)} SW_\Delta(\mu, \nu)
\end{aligned} \tag{6.64}$$

where SW is the Sliced-Wasserstein distance. The complexity for computing SW_Δ between two discrete probability measures with n atoms and uniform weights is $O(Ln(p + q + \log(n)))$ which is exactly the same complexity as SGW_Δ. With these formulations, we can perform the same experiment as for

RISGW on the spiral dataset. The optimisation on the Stiefel manifold is performed using Pymanopt as for SGW. Results are reported in Figure 6.1. As one can see, $RISW$ is rotational invariant on average whereas SW is not. One can also note that, due to the sampling process of the spiral dataset, the variance is quite large. This can be explained by the fact that, unlike SGW, the Sliced-Wasserstein may realign the distributions without taking the rotation into account.

6.2.6 Proof of Lemma 4.2.1 – Reductions of the GW costs for inner products and squared Euclidean distance matrices

We recall the lemma:

Lemma. *Suppose that there exist scalars a, b, c such that $c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = a\|\mathbf{x}\|_2^2 + b\|\mathbf{x}'\|_2^2 + c\langle \mathbf{x}, \mathbf{x}'\rangle_p$ and $c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = a\|\mathbf{y}\|_2^2 + b\|\mathbf{y}'\|_2^2 + c\langle \mathbf{y}, \mathbf{y}'\rangle_q$. Then:*

$$J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi) = C_{\mu,\nu} - 2Z(\pi) \tag{6.65}$$

where $C_{\mu,\nu} = \int c_{\mathcal{X}}^2 \mathrm{d}\mu\mathrm{d}\mu + \int c_{\mathcal{Y}}^2 \mathrm{d}\nu\mathrm{d}\nu - 4ab \int \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 d\mu(\mathbf{x})d\nu(\mathbf{y})$ and:

$$
\begin{aligned}
Z(\pi) = {}&(a^2 + b^2) \int \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + c^2 \|\int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}^2 \\
&+ (a + b)c \int \left[\|\mathbf{x}\|_2^2\langle \mathbb{E}_{Y\sim\nu}[Y], \mathbf{y}\rangle_q + \|\mathbf{y}\|_2^2\langle \mathbb{E}_{X\sim\mu}[X], \mathbf{x}\rangle_p \mathrm{d}\pi(\mathbf{x}, \mathbf{y})\right]
\end{aligned}
\tag{6.66}
$$

Proof. Let $\pi \in \Pi(\mu, \nu)$. We have $J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi) = \int c_{\mathcal{X}}^2 \mathrm{d}\mu\mathrm{d}\mu + \int c_{\mathcal{Y}}^2 \mathrm{d}\nu\mathrm{d}\nu - 2\int c_{\mathcal{X}}c_{\mathcal{Y}}\mathrm{d}\pi\mathrm{d}\pi$. In this way:

$$
\begin{aligned}
\int c_{\mathcal{X}}c_{\mathcal{Y}}\mathrm{d}\pi\mathrm{d}\pi = {}& \int (a\|\mathbf{x}\|_2^2 + b\|\mathbf{x}'\|_2^2 + c\langle \mathbf{x}, \mathbf{x}'\rangle_p)(a\|\mathbf{y}\|_2^2 + b\|\mathbf{y}'\|_2^2 + c\langle \mathbf{y}, \mathbf{y}'\rangle_q)\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \\
= {}& \int \big[(a^2\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 + ab\|\mathbf{x}\|_2^2\|\mathbf{y}'\|_2^2 + ac\|\mathbf{x}\|_2^2\langle \mathbf{y}, \mathbf{y}'\rangle_q) + (ab\|\mathbf{x}'\|_2^2\|\mathbf{y}\|_2^2 + b^2\|\mathbf{x}'\|_2^2\|\mathbf{y}'\|_2^2 + bc\|\mathbf{x}'\|_2^2\langle \mathbf{y}, \mathbf{y}'\rangle_q) \\
& + (ca\langle \mathbf{x}, \mathbf{x}'\rangle_p\|\mathbf{y}\|_2^2 + cb\langle \mathbf{x}, \mathbf{x}'\rangle_p\|\mathbf{y}'\|_2^2 + c^2\langle \mathbf{x}, \mathbf{x}'\rangle_p\langle \mathbf{y}, \mathbf{y}'\rangle_q)\big]\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \\
= {}& (a^2 + b^2) \int \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + 2ab \int \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 d\mu(\mathbf{x})d\nu(\mathbf{y}) + c^2 \int \langle \mathbf{x}, \mathbf{x}'\rangle_p\langle \mathbf{y}, \mathbf{y}'\rangle_q\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \\
& + (a + b)c \int \|\mathbf{x}\|_2^2\langle \int \mathbf{y}\mathrm{d}\pi(\mathbf{x}, \mathbf{y}), \mathbf{y}\rangle_q\mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + (a + b)c \int \|\mathbf{y}\|_2^2\langle \int \mathbf{x}\mathrm{d}\pi(\mathbf{x}, \mathbf{y}), \mathbf{x}\rangle_p\mathrm{d}\pi(\mathbf{x}, \mathbf{y})
\end{aligned}
$$

Moreover:

$$
\begin{aligned}
\int \langle \mathbf{x}, \mathbf{x}'\rangle_p\langle \mathbf{y}, \mathbf{y}'\rangle_q\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') &= \int \mathbf{x}^T\mathbf{x}'\mathbf{y}^T\mathbf{y}'\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') = \int \mathrm{tr}(\mathbf{x}^T\mathbf{x}'\mathbf{y}^T\mathbf{y}')\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \\
&= \int \mathrm{tr}(\mathbf{y}'\mathbf{x}^T\mathbf{x}'\mathbf{y}^T)\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \overset{*}{=} \int \mathrm{tr}(\mathbf{y}'\mathbf{x}'^T\mathbf{x}\mathbf{y}^T)\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\mathrm{d}\pi(\mathbf{x}', \mathbf{y}') \\
&= \mathrm{tr}((\int \mathbf{y}'\mathbf{x}'^T\mathrm{d}\pi(\mathbf{x}', \mathbf{y}')(\int \mathbf{x}\mathbf{y}^T\mathrm{d}\pi(\mathbf{x}, \mathbf{y})))) = \|\int \mathbf{x}\mathbf{y}^T\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}^2 = \|\int \mathbf{y}\mathbf{x}^T\mathrm{d}\pi(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}^2
\end{aligned}
$$

where in (*) we used that $\mathbf{x}'^T\mathbf{x} \in \mathbb{R}$ so it is equal to its transpose. $\square$

6.2.7 Proof of the existence and finiteness (innerGW) and (MaxOT)

In this section we prove that (innerGW) and (MaxOT) are finite and that (MaxOT) always admits a maximizer. More precisely:

Lemma 6.2.7. *Let $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ with $\int \|\mathbf{x}\|_2^4 d\mu(\mathbf{x}) < +\infty, \int \|\mathbf{y}\|_2^4 d\nu(\mathbf{y}) < +\infty$. Both* (innerGW) *and* (MaxOT) *are finite.*

Moreover the set $F_{p,q}$ is a compact subset of $\mathbb{R}^{q \times p}$. The functional $\pi \to \sup_{\mathbf{P} \in F_{p,q}} \int \langle \mathbf{Px}, \mathbf{y} \rangle_q d\pi(\mathbf{x}, \mathbf{y})$ is continuous for the weak convergence of measure. In particular problem (MaxOT) *admits an optimal solution $\pi^* \in \Pi(\mu, \nu)$*

Proof. In this case $\int \|\mathbf{x}\|_2^2 d\mu(\mathbf{x}) < +\infty, \int \|\mathbf{y}\|_2^2 d\nu(\mathbf{y}) < +\infty$ by Hölder's inequality and for all $(\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}') \in \mathcal{X}^2 \times \mathcal{Y}^2$ $\left(\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle \mathbf{y}, \mathbf{y}' \rangle_q \right)^2 \leq 2(\langle \mathbf{x}, \mathbf{x}' \rangle_p^2 + \langle \mathbf{y}, \mathbf{y}' \rangle_q^2) \leq 2(\|\mathbf{x}\|_2^2 \|\mathbf{x}'\|_2^2 + \|\mathbf{y}\|_2^2 \|\mathbf{y}'\|_2^2)$ by Cauchy-Swartz. In particular this implies:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int \int \left(\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle \mathbf{y}, \mathbf{y}' \rangle_q \right)^2 d\pi(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}', \mathbf{y}') \leq \int \int \left(\langle \mathbf{x}, \mathbf{x}' \rangle_p - \langle \mathbf{y}, \mathbf{y}' \rangle_q \right)^2 d\mu(\mathbf{x}) d\mu(\mathbf{x}') d\nu(\mathbf{y}) d\nu(\mathbf{y}')$$

$$\leq 2(\int \|\mathbf{x}\|_2^2 d\mu(\mathbf{x}) \int \|\mathbf{x}'\|_2^2 d\mu(\mathbf{x}') + \int \|\mathbf{y}\|_2^2 d\nu(\mathbf{x}) \int \|\mathbf{y}'\|_2^2 d\nu(\mathbf{y}')) < +\infty$$

Moreover for any $\pi \in \Pi(\mu, \nu), \mathbf{P} \in F_{p,q}$:

$$\int \langle \mathbf{Px}, \mathbf{y} \rangle_q \, d\pi(\mathbf{x}, \mathbf{y}) = \int \langle \mathbf{P}, \mathbf{yx}^T \rangle_{\mathcal{F}} \, d\pi(\mathbf{x}, \mathbf{y}) \leq \int \|\mathbf{P}\|_{\mathcal{F}}^2 \|\mathbf{xy}^T\|_{\mathcal{F}}^2 d\pi(\mathbf{x}, \mathbf{y})$$

$$\leq p \int \|\mathbf{xy}^T\|_{\mathcal{F}}^2 d\pi(\mathbf{x}, \mathbf{y}) = p \int \mathrm{tr}(\mathbf{yx}^T \mathbf{xy}^T) d\pi(\mathbf{x}, \mathbf{y}) = p \int \mathrm{tr}(\mathbf{x}^T \mathbf{xy}^T \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y})$$

$$= p \int \mathrm{tr}(\|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2) d\pi(\mathbf{x}, \mathbf{y}) = p \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 d\pi(\mathbf{x}, \mathbf{y}) \overset{*}{\leq} \frac{p}{2}(\int \|\mathbf{x}\|_2^4 d\mu(\mathbf{x}) + \int \|\mathbf{y}\|_2^4 d\nu(\mathbf{y}))$$

$$< +\infty$$

where in (*) we used Young's inequality (recalled in Memo 4.2.2).

For the compacity, using Borel Lebesgue theorem it suffices to show that $F_{p,q}$ is closed and bounded. It is clearly bounded by $\sqrt{p}$ and closed as the pre-image of the closed set $\{0\}$ by the continuous application $\mathbf{P} \to \sqrt{p} - \|\mathbf{P}\|_{\mathcal{F}}$.

We note $f : \Pi(\mu, \nu) \times F_{p,q} \to \mathbb{R}$ the function $f(\pi, \mathbf{P}) = -\int \langle \mathbf{Px}, \mathbf{y} \rangle_q d\pi(\mathbf{x}, \mathbf{y})$. For any $\pi \in \Pi(\mu, \nu)$, $f(\pi, .)$ is continuous. Indeed suppose that $\mathbf{P}_m \underset{\mathcal{F}}{\to} \mathbf{P}$ where $\underset{\mathcal{F}}{\to}$ denotes the convergence in Frobenius norm,

---

Memo 6.2.3. *Let $\mathcal{X}, \mathcal{Y}$ be topological spaces with $\mathcal{Y}$ compact. If $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous then $g : x \to \inf_y f(x, y)$ is well defined and continuous.*

Proof. Note first that $g(x) > -\infty$, since for every $x \in \mathcal{X}$, $f(x, .) : \mathcal{Y} \to \mathbb{R}$ is continuous on a compact space so it is bounded. To prove the continuity it suffices to show that $g^{-1}(]-\infty, a[)$ and $g^{-1}(]b, +\infty[)$ are open. For the former, if we note $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ the canonical projection then $g^{-1}(]-\infty, a[) = \pi_{\mathcal{X}} \circ f^{-1}(]-\infty, a[)$. By continuity of f we can conclude that $g^{-1}(]-\infty, a[)$ is open. For the latter first observe that $g(x) > b \implies \forall y, f(x, y) > b$ which means that $g(x) > b \implies \forall y, (x, y) \in f^{-1}(]b, +\infty[)$. In particular for $x \in g^{-1}(]b, +\infty[)$ and $y \in \mathcal{Y}$ there exists a neighborhood $U_{(x,y)} \times V_{(x,y)}$ contained in $f^{-1}(]b, +\infty[)$. Since $\mathcal{Y}$ is compact there exists a finite subset $\{(x, y_i)\}$ of all the neighborhoods $U_{(x,y)} \times V_{(x,y)}$ which cover all of $\{x\} \times \mathcal{Y}$. Overall:

$$\{x\} \times \mathcal{Y} \subset (\cap_{i=1}^k U_{(x,y_i)}) \times \mathcal{Y} \subset f^{-1}(]b, +\infty[) \tag{6.67}$$

Hence $g^{-1}(]b, +\infty[) = \cup_{x \in g^{-1}(]b, +\infty[)} \cap_{i=1}^k U_{(x,y_i)})$ which is open so g is continuous. $\square$

then:

$$
\left| \int \langle \mathbf{P}_m \mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) - \int \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q \mathrm{d}\pi(x, y) \right| \leq \int |\langle (\mathbf{P}_m - \mathbf{P})\mathbf{x}, \mathbf{y} \rangle_q| \mathrm{d}\pi(\mathbf{x}, \mathbf{y})
$$

$$
\leq \|\mathbf{P}_m - \mathbf{P}\|_{\mathcal{F}}^2 (\int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y})) \leq \frac{1}{2} \|\mathbf{P}_m - \mathbf{P}\|_{\mathcal{F}}^2 (\int \|x\|_2^4 \mathrm{d}\mu(x) + \int \|y\|_2^4 \mathrm{d}\nu(\mathbf{y})) \underset{m \to +\infty}{\to} 0
$$

In this way, since $F_{p,q}$ is compact, $g : \pi \to \inf_{\mathbf{P} \in F} f(\pi, \mathbf{P})$ is well defined and continuous for the weak convergence of measure (see Memo 6.2.3). Since it is continuous and $\Pi(\mu, \nu)$ is compact we can applied the Weirstrass' theorem to state that $\inf_{\pi \in \Pi(\mu, \nu)} g(\pi)$ exists which concludes the proof.

$\square$

6.2.8   Proof of Lemma 4.2.3 – Reduction of the GW cost

We recall the lemma:

Lemma. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact subset of respectively $\mathbb{R}^p$ and $\mathbb{R}^q$. Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. We can assume without loss of generality that $\mathbb{E}_{X \sim \mu}[X] = 0$ and $\mathbb{E}_{Y \sim \nu}[Y] = 0$. In this case (sqGW) is equivalent to:*

$$
\sup_{\pi \in \Pi(\mu, \nu)} \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + 2 \| \int \mathbf{y}\mathbf{x}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \|_{\mathcal{F}}^2
$$

Indeed equation (sqGW) corresponds to the case $a = 1, b = 1, c = -2$ of Lemma 4.2.1 and so is equivalent to:

$$
\sup_{\pi \in \Pi(\mu, \nu)} 2 \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \mathrm{d}\pi(x, y) + 4 \| \int \mathbf{x}\mathbf{y}^T \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \|_{\mathcal{F}}^2
$$

$$
- 4 \int \left[\|\mathbf{x}\|_2^2 \langle \mathbb{E}_{Y \sim \nu}[Y], y \rangle_{d'} + \|\mathbf{y}\|_2^2 \langle \mathbb{E}_{X \sim \mu}[X], x \rangle_d \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) \right]
$$

(6.68)

We can reduce this problem using the translation invariance of the GW cost since $\|\mathbf{x} + \mathbf{t} - (\mathbf{x}' + \mathbf{t})\|_2^2 = \|\mathbf{x} - \mathbf{x}'\|_2^2$ for all $\mathbf{x}, \mathbf{x}'$ (same for $\mathbf{y}$). More precisely we can rely on the following lemmas:

Lemma 6.2.8. *Let $f : \mathbb{R}^p \to \mathbb{R}^p$ and $f' : \mathbb{R}^q \to \mathbb{R}^q$ Borel and $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ Then*

$$
\Pi\left(f\#\mu, f'\#\nu\right) = \{(f \times f')\#\pi | \pi \in \Pi(\mu, \nu)\}
$$

Proof. This is a straightforward extension of the Lemma 6 in [Paty 2019]. $\square$

Based on previous Lemma 6.2.8 we have:

Lemma 6.2.9 (Translation Invariance). *Let $\mu \in \mathcal{P}(\mathbb{R}^p), \nu \in \mathcal{P}(\mathbb{R}^q)$ and $c_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2^2, c_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = \|\mathbf{y} - \mathbf{y}'\|_2^2$. Let $\mathbf{t}, \mathbf{t}' \in \mathbb{R}^p \times \mathbb{R}^q$ and $f_{\mathbf{t}}(\mathbf{x}) = \mathbf{x} + \mathbf{t}, f'_{\mathbf{t}'} = \mathbf{y} + \mathbf{t}'$ be translations. Then:*

$$
\sup_{\pi \in \Pi(\mu, \nu)} J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi) = \sup_{\pi \in \Pi(f_{\mathbf{t}}\#\mu, f'_{\mathbf{t}'}\#\nu)} J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi)
$$

(6.69)

Proof. Using Lemma 6.2.8 we have:

$$
\sup_{\pi \in \Pi(f_{\mathbf{t}}\#\mu, f'_{\mathbf{t}'}\#\nu)} J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi) = \sup_{\pi \in \Pi(\mu, \nu)} J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, (f_{\mathbf{t}} \times f'_{\mathbf{t}'})\#\pi) \overset{*}{=} \sup_{\pi \in \Pi(\mu, \nu)} J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi)
$$

where in (*) we used the invariance of the cost $J_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \pi)$ with respect to translations. $\square$

Using this property we can assume without loss of generality μ and ν centered, *i.e*, $\mathbb{E}_{X \sim \mu}[X] = 0$ and $\mathbb{E}_{Y \sim \nu}[Y] = 0$. By plugging this condition into $Z(\pi)$ of Lemma 4.2.1 we have the desired result.

6.2.9 Proof of Lemma 4.2.4

We recall the lemma:

Lemma. *If π^* is a solution of the primal problem $\sup_{\pi \in \Pi(\mu,\nu)} F(\pi)$ then there exists $\mathbf{P} \in \mathbb{R}^{q \times p}$ and $h^* \in C(\mathcal{X} \times \mathcal{Y})$ of the form $h(\mathbf{x}, \mathbf{y}) = \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$ such that (π^*, h^*) is a solution of the dual problem (4.26). Moreover when h is in such form we have $F^*(h) = \frac{1}{8} \|\mathbf{P}\|_{\mathcal{F}}^2$.*

To prove this result we will need the following calculus:

Lemma 6.2.10. *With previous notations, the Fréchet derivative of F reads:*

$$\nabla F(\pi) = (\mathbf{x}, \mathbf{y}) \to 4 \langle \mathbf{V}_\pi \mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 \tag{6.70}$$

Proof.

$$F(\pi + t\varepsilon) = 2\|\mathbf{V}_\pi\|_{\mathcal{F}}^2 + \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 d\pi(x,y) + \|t\mathbf{V}_\varepsilon\|_{\mathcal{F}}^2 + t \int (4 \langle \mathbf{V}_\pi, \mathbf{y}\mathbf{x}^T \rangle_{\mathcal{F}} + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2) d\varepsilon(\mathbf{x}, \mathbf{y})$$

$$= F(\pi) + t \int (4 \langle \mathbf{V}_\pi, \mathbf{y}\mathbf{x}^T \rangle_{\mathcal{F}} + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2) d\varepsilon(x,y) + o_{t \to 0}(t) \tag{6.71}$$

for $\varepsilon \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$. Hence $\nabla F(\pi) : (x,y) \to 4 \langle \mathbf{V}_\pi, \mathbf{y}\mathbf{x}^T \rangle_{\mathcal{F}} + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 = 4 \langle \mathbf{V}_\pi \mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$ □

Proof of Lemma 4.2.4 – Parametrization of the dual problem. If π^* is a maximizer of the primal problem then we know that $h^* = \nabla F(\pi^*)$ is a maximizer of $\sup_{h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \sup_{\pi \in \Pi(\mu,\nu)} \int h(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}) - F^*(h)$. The calulus in Lemma 6.2.10) implies that $\nabla F(\pi^*) = (\mathbf{x}, \mathbf{y}) \to 4 \langle \mathbf{V}_{\pi^*} \mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$. Setting $\mathbf{P} = 4\mathbf{V}_{\pi^*} \in \mathbb{R}^{q \times p}$ concludes for the first point.

For the second point we have by definition $F^*(h) = \sup_{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})} \int h(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}) - F(\pi)$. We note $G(\pi) = \int h(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}) - F(\pi)$. Then $\nabla G(\pi) = h - \nabla F(\pi) = (\mathbf{x}, \mathbf{y}) \to h(\mathbf{x}, \mathbf{y}) - 4 \langle \mathbf{V}_\pi \mathbf{x}, \mathbf{y} \rangle_q - \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2$ by Lemma 6.2.10. Using the fact that h is parametrized by a linear application we have $\nabla G(\pi) = (\mathbf{x}, \mathbf{y}) \to \langle (\mathbf{P} - 4\mathbf{V}_\pi)\mathbf{x}, \mathbf{y} \rangle_q$. Then for $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$:

$$\nabla G(\pi)(\mathbf{x}, \mathbf{y}) = 0 \iff \langle (\mathbf{P} - 4\mathbf{V}_\pi)\mathbf{x}, \mathbf{y} \rangle_q = 0 \tag{6.72}$$

We write $\mathbf{P} = \sum_i \lambda_i \mathbf{v}_i \mathbf{u}_i^T$ the SVD of $\mathbf{P}$. We note $\gamma_\mathbf{P} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ the measure $\gamma_\mathbf{P} = \frac{1}{4} \sum_i \lambda_i \delta_{(\mathbf{u}_i, \mathbf{v}_i)}$ (note that we do not need that $\gamma_\mathbf{P}$ be a probability measure). Then $\int \mathbf{y}\mathbf{x}^T d\gamma_\mathbf{P}(\mathbf{x}, \mathbf{y}) = \frac{1}{4} \sum_i \lambda_i \mathbf{v}_i \mathbf{u}_i^T = \frac{1}{4} \mathbf{P}$. By previous calculus $\nabla G(\gamma_\mathbf{P}) = 0$ so that $\gamma_\mathbf{P}$ satisfies the first order condition and is a solution to $\sup_{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})} \int h(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}, \mathbf{y}) - F(\pi)$. Overall:

$$F^*(h) = \int h(\mathbf{x}, \mathbf{y}) d\gamma_\mathbf{P}(\mathbf{x}, \mathbf{y}) - F(\gamma_\mathbf{P})$$

$$= \int \langle \mathbf{P}\mathbf{x}, \mathbf{y} \rangle_q + \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 d\gamma_\mathbf{P}(\mathbf{x}, \mathbf{y}) - \int \|\mathbf{x}\|_2^2 \|\mathbf{y}\|_2^2 d\gamma_\mathbf{P}(\mathbf{x}, \mathbf{y}) - 2\| \int \mathbf{y}\mathbf{x}^T d\gamma_\mathbf{P}(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}^2 \tag{6.73}$$

$$= \langle \mathbf{P}, \int \mathbf{y}\mathbf{x}^T d\gamma_\mathbf{P}(x,y) \rangle_{\mathcal{F}} - 2\| \int \mathbf{y}\mathbf{x}^T d\gamma_\mathbf{P}(\mathbf{x}, \mathbf{y})\|_{\mathcal{F}}^2 = \frac{1}{4} \|\mathbf{P}\|_{\mathcal{F}}^2 - \frac{1}{8} \|\mathbf{P}\|_{\mathcal{F}}^2 = \frac{1}{8} \|\mathbf{P}\|_{\mathcal{F}}^2$$

□

6.2.10 Characterization of $\mathcal{H}(\mu, \nu)$ for linear push-forward

We have the following result:

Lemma 6.2.11. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^p) \times \mathcal{P}(\mathbb{R}^p)$. If T is a linear symmetric push-forward of μ to ν then:*

$$T \in \mathcal{H}(\mu, \nu) \iff T = \lambda \mathbf{O} \tag{6.74}$$

where $\mathbf{O} \in \mathcal{O}(p)$ is an orthogonal matrix and $\lambda > 0$.

Proof. We note $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$. Then we have $\|\mathbf{A}\mathbf{x}\|_2^2 = \mathbf{x}^T \mathbf{A}^T \mathbf{A}\mathbf{x}$, we can write $\mathbf{A}^T\mathbf{A} = \mathbf{PDP}^T$ where $\mathbf{D}$ diagonal and $\mathbf{P}$ orthogonal, then it is equivalent to find a $g = f'$ such that:

$$\|\mathbf{D}^{1/2}\mathbf{v}\|_2^2 = \sum_i |\beta_i v_i|^2 = g(\|\mathbf{v}\|_2^2) \tag{6.75}$$

(with $\beta_i \geq 0$ and we note $\mathbf{v} = \mathbf{P}^T\mathbf{x}$). If there is i such that $\beta_i = 0$ (the first one without loss of generality) then for all $t \geq 0$ we can take $\mathbf{v} = (\sqrt{t}, 0, .., 0)$ then $g(\|\mathbf{v}\|_2^2) = g(t) = 0$ which is not possible unless $\mathbf{A} = 0$. In this way $\beta_i > 0$ for all i. Let $\mathbf{e}_i$ be one eigenvector of $\mathbf{A}^T\mathbf{A}$ then $\|\mathbf{A}\mathbf{e}_i\|_2^2 = \mathbf{e}_i^T\mathbf{A}^T\mathbf{A}\mathbf{e}_i = \lambda_i = g(\|\mathbf{e}_i\|_2^2) = g(1)$. Overall $\beta_i = g(1) = \lambda$ for all i. So $\mathbf{A}^T\mathbf{A} = \lambda_0 \mathbf{I_p}$. Overall it implies that $\mathbf{A}$ can be written as $\mathbf{A} = \lambda\mathbf{O}$ with $\mathbf{O}$ an orthogonal matrix. Conversely if $T = \lambda\mathbf{O}$ then $\|T(\mathbf{x})\|_2^2 = c^2\|\mathbf{x}\|_2^2$. We can take the function $f(t) = \frac{1}{2}c^2t^2$ which is convex and continuous from $\mathbb{R}$ to $\mathbb{R}$. $\square$

6.2.11 Proof of Theorem 4.2.6 – Closed form expression of the linear Gromov-Monge problem between Gaussian measures

We recall the theorem:

Theorem. *Let $\mu = \mathcal{N}(0, \Sigma_\nu) \in \mathcal{P}(\mathbb{R}^p), \nu = \mathcal{N}(0, \Sigma_\mu) \in \mathcal{P}(\mathbb{R}^q)$ centered without loss of generality. Let $\Sigma_\mu = \mathbf{V}_\mu \mathbf{D}_\mu \mathbf{V}_\mu^\top, \Sigma_\nu = \mathbf{V}_\nu \mathbf{D}_\nu \mathbf{V}_\nu^\top$ be the diagonalizations of the covariance matrices such that eigenvalues of $\mathbf{D}_\mu$ and $\mathbf{D}_\nu$ are ordered nondecreasing. When $p \neq q$ we have:*

$$LGM_2^2(\mu, \nu) = 4(tr(\Sigma_\mu) - tr(\Sigma_\nu))^2 + 8(tr(\Sigma_\mu\Sigma_\mu) + tr(\Sigma_\nu\Sigma_\nu)) + 16 \min_{\mathbf{B} \in \mathbb{V}_p(\mathbb{R}^q)} -tr(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B})$$

When $p = q$, an optimal linear Monge map is given by $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where:

$$\mathbf{A} = \mathbf{V}_\nu \mathbf{D}_\nu^{1/2} \mathbf{D}_\mu^{-1/2} \mathbf{V}_\mu^\top = \Sigma_\nu^{1/2} \mathbf{V}_\nu \mathbf{V}_\mu^\top \Sigma_\mu^{-1/2}$$

so that:

$$LGM_2^2(\mu, \nu) = 4(tr(\Sigma_\mu) - tr(\Sigma_\nu))^2 + 8(tr(\Sigma_\mu\Sigma_\mu) + tr(\Sigma_\nu\Sigma_\nu)) - 16tr(\mathbf{D}_\mu\mathbf{D}_\nu)$$

In order to prove Theorem 4.2.6 we will rely on the the following result:

Proposition 6.2.3 (Proposition 3.1 in [Anstreicher 1998]). *Consider $\Sigma_1, \Sigma_2 \in \mathbb{R}^{p \times p} \times \mathbb{R}^{p \times p}$ two symmetric matrices and the following (QQP):*

$$\min_{\mathbf{BB}^T = \mathbf{I}} tr(\Sigma_1 \mathbf{B} \Sigma_2 \mathbf{B}^T) \tag{6.76}$$

Let $\Sigma_1 = \mathbf{V}_1 \mathbf{D}_1 \mathbf{V}_1^T$ and $\Sigma_2 = \mathbf{V}_2 \mathbf{D}_2 \mathbf{V}_2^T$ be the orthogonal diagonalizations of Σ_1, Σ_2 where the eigenvalues of $\mathbf{D}_1$ are ordered nonincreasing the eigenvalues in $\mathbf{D}_2$ are ordered nondecreasing. An optimal solution of (6.76) is found at $\mathbf{B}^* = \mathbf{V}_1 \mathbf{V}_2^T$ and the optimal value of (6.76) is $tr(\mathbf{D}_1\mathbf{D}_2)$.

We will prove that when considering only linear push-forward we can recast the Gromov-Monge problem in the form of Proposition 6.2.3. We have the following result:

Lemma 6.2.12. *Let $\mu = \mathcal{N}(0, \boldsymbol{\Sigma}_\nu) \in \mathcal{P}(\mathbb{R}^p), \nu = \mathcal{N}(0, \boldsymbol{\Sigma}_\mu) \in \mathcal{P}(\mathbb{R}^q)$ centered without loss of generality. For a linear push-forward $T\#\mu = \nu$ we note $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$. The Gromov-Monge problem is equivalent to:*

$$\min_{\substack{\mathbf{A} \in \mathbb{R}^{q \times p} \\ \mathbf{A}\boldsymbol{\Sigma}_\mu \mathbf{A}^\top = \boldsymbol{\Sigma}_\nu}} \langle \mathbf{M}, \mathbf{A}^\top \mathbf{A} \rangle_{\mathcal{F}} \tag{6.77}$$

where:

$$\mathbf{M} = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[-2(\mathbf{x}^\top \mathbf{x}).\mathbf{x}\mathbf{x}^\top - 2(\mathbf{x'}^\top \mathbf{x}').\mathbf{x}\mathbf{x}^\top - 4(\mathbf{x}^\top \mathbf{x}').\mathbf{x}'\mathbf{x}^\top] \tag{6.78}$$

Proof. We have:

$$\min_{\substack{T\#\mu=\nu \\ T \text{ linear}}} J(T) = \min_{\substack{T\#\mu=\nu \\ T \text{ linear}}} \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[(\|\mathbf{x} - \mathbf{x}'\|_2^2 - \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2)^2]$$

$$= \min_{\substack{T\#\mu=\nu \\ T \text{ linear}}} \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x} - \mathbf{x}'\|^4] + \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|T(\mathbf{x}) - T(\mathbf{x}')\|_2^4] - 2\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x} - \mathbf{x}'\|_2^2 \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2]$$

$$= \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x} - \mathbf{x}'\|_2^4] + \mathbb{E}_{\mathbf{y}, \mathbf{y}' \sim \nu}[\|\mathbf{y} - \mathbf{y}'\|_2^4] + 2\min_{\substack{T\#\mu=\nu \\ T \text{ linear}}} J_1(T)$$

With $J_1(T) \stackrel{def}{=} -\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x} - \mathbf{x}'\|_2^2 \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2]$. In this way the problem is equivalent to minimizing $J_1(T)$. Since $T\#\mu = \nu$:

$$J_1(T) = -\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x} - \mathbf{x}'\|_2^2 \|T(\mathbf{x}) - T(\mathbf{x}')\|_2^2]$$

$$= -\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[(\|\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2 - 2\langle \mathbf{x}, \mathbf{x}' \rangle_p)(\|T(\mathbf{x})\|_2^2 + \|T(\mathbf{x}')\|^2 - 2\langle T(\mathbf{x}), T(\mathbf{x}') \rangle_q))]$$

$$= -\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x}\|_2^2 \|T(\mathbf{x})\|_2^2 + \|\mathbf{x}'\|_2^2 \|T(\mathbf{x}')\|_2^2] - \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x}'\|_2^2 \|T(\mathbf{x})\|_2^2 + \|\mathbf{x}\|_2^2 \|T(\mathbf{x}')\|_2^2]$$

$$+ 2\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle \mathbf{x}, \mathbf{x}' \rangle \|T(\mathbf{x})\|_2^2] + 2\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle \mathbf{x}, \mathbf{x}' \rangle_p \|T(\mathbf{x}')\|_2^2] - 4\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle \mathbf{x}, \mathbf{x}' \rangle_p \langle T(\mathbf{x}), T(\mathbf{x}') \rangle_q]$$

$$+ 2\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle T(\mathbf{x}), T(\mathbf{x}') \rangle_q \|\mathbf{x}\|_2^2] + 2\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle T(\mathbf{x}), T(\mathbf{x}') \rangle_q \|\mathbf{x}'\|_2^2]$$

$$= -2\mathbb{E}_{\mathbf{x} \sim \mu}[\|\mathbf{x}\|_2^2 \|T(\mathbf{x})\|_2^2] - 2\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\|\mathbf{x}'\|_2^2 \|T(\mathbf{x})\|_2^2] + 4\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle \mathbf{x}, \mathbf{x}' \rangle_p \|T(\mathbf{x})\|_2^2]$$

$$- 4\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle \mathbf{x}, \mathbf{x}' \rangle_p \langle T(\mathbf{x}), T(\mathbf{x}') \rangle_q] + 4\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\langle T(\mathbf{x}), T(\mathbf{x}') \rangle_q \|\mathbf{x}\|_2^2]$$

Since T is linear it can be written in the form $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$. The push-forward constraint in this case reads [Flamary 2019]:

$$\mathbf{A}\boldsymbol{\Sigma}_\mu \mathbf{A}^\top = \boldsymbol{\Sigma}_\nu \tag{6.79}$$

When A is symmetric positive definite this equation admits a unique solution which is the optimal linear map for the Wasserstein problem. However here we have no result about the regularity of $\mathbf{A}$. Plugging $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$ into $J_1(T)$ and writing the push-forward condition gives the equivalent problem:

$$\min_{\mathbf{A} \in \mathbb{R}^{q \times p}, \mathbf{A}\boldsymbol{\Sigma}_\mu \mathbf{A}^\top = \boldsymbol{\Sigma}_\nu} J_2(\mathbf{A}) \tag{6.80}$$

where

$$J_2(\mathbf{A}) = -2\mathbb{E}_{\mathbf{x} \sim \mu}[\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}] - 2\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\mathbf{x'}^\top \mathbf{x}'\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}] + 4\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\mathbf{x}^\top \mathbf{x}'\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}] \tag{6.81}$$

$$- 4\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\mathbf{x}^\top \mathbf{x}'\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}'] + 4\mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim \mu}[\mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}'\mathbf{x'}^\top \mathbf{x}] \tag{6.82}$$

Using the property of the trace of a matrix and linearity of the inner product, one can reformulate J_2 as

$$J_2(\mathbf{A}) = \langle \mathbf{M}, \mathbf{A}^\top \mathbf{A} \rangle_\mathcal{F} \tag{6.83}$$

where

$$\mathbf{M} = \underset{\mathbf{x}, \mathbf{x}' \sim \mu}{\mathbb{E}} [-2(\mathbf{x}^\top \mathbf{x}).\mathbf{x} \mathbf{x}^\top - 2(\mathbf{x}'^\top \mathbf{x}').\mathbf{x} \mathbf{x}^\top + 4(\mathbf{x}^\top \mathbf{x}').\mathbf{x} \mathbf{x}^\top - 4(\mathbf{x}^\top \mathbf{x}').\mathbf{x}' \mathbf{x}^\top + 4(\mathbf{x}^\top \mathbf{x}).\mathbf{x}' \mathbf{x}^\top] \tag{6.84}$$

Indeed for all terms we used the following reasoning:

$$\underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathbf{x}^\top \mathbf{x} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x}] \overset{*}{=} \underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathrm{tr}(\mathbf{x}^\top \mathbf{x} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x})] \overset{**}{=} \underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathbf{x}^\top \mathbf{x} \mathrm{tr}(\mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x})] \overset{***}{=} \underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathbf{x}^\top \mathbf{x} \mathrm{tr}(\mathbf{x} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A})]$$

$$= \underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathrm{tr}(\mathbf{x}^\top \mathbf{x} \mathbf{x} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A})] = \mathrm{tr}(\underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathbf{x}^\top \mathbf{x} \mathbf{x} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}]) = \mathrm{tr}(\underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathbf{x}^\top \mathbf{x} \mathbf{x} \mathbf{x}^\top] \mathbf{A}^\top \mathbf{A})$$

$$= \langle \underset{\mathbf{x} \sim \mu}{\mathbb{E}} [\mathbf{x}^\top \mathbf{x} \mathbf{x} \mathbf{x}^\top], \mathbf{A}^\top \mathbf{A} \rangle_\mathcal{F} \tag{6.85}$$

where in (*) we used that $\mathbf{x}^\top \mathbf{x} \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x} \in \mathbb{R}$, in (**) that $\mathbf{x}^\top \mathbf{x} \in \mathbb{R}$ and in (***) the cyclical permutation invariance of tr. Moreover $\underset{\mathbf{x}, \mathbf{x}' \sim \mu}{\mathbb{E}} [(\mathbf{x}^\top \mathbf{x}).\mathbf{x}' \mathbf{x}^\top], \underset{\mathbf{x}, \mathbf{x}' \sim \mu}{\mathbb{E}} (\mathbf{x}^\top \mathbf{x}').\mathbf{x} \mathbf{x}^\top] = 0$ since Gaussians are centered. Hence:

$$\mathbf{M} = \underset{\mathbf{x}, \mathbf{x}' \sim \mu}{\mathbb{E}} [-2(\mathbf{x}^\top \mathbf{x}).\mathbf{x} \mathbf{x}^\top - 2(\mathbf{x}'^\top \mathbf{x}').\mathbf{x} \mathbf{x}^\top - 4(\mathbf{x}^\top \mathbf{x}').\mathbf{x}' \mathbf{x}^\top] \tag{6.86}$$

$$\square$$

We can go further on this calculus by using Isserlis' theorem recalled here:

Theorem (Isserlis [Isserlis 1918]). *If (X_1, X_2, X_3, X_4) is a zero-mean multivariate normal random vector then:*

$$\mathbb{E}[X_1 X_2 X_3 X_4] = \mathbb{E}[X_1 X_2]\mathbb{E}[X_3 X_4] + \mathbb{E}[X_1 X_3]\mathbb{E}[X_2 X_4] + \mathbb{E}[X_1 X_4]\mathbb{E}[X_2 X_3] \tag{6.87}$$

Applying this theorem to $\mathbf{M}$ in (6.77) gives the following lemma:

Lemma 6.2.13. *Let $\mu = \mathcal{N}(0, \mathbf{\Sigma}_\nu) \in \mathcal{P}(\mathbb{R}^p), \nu = \mathcal{N}(0, \mathbf{\Sigma}_\mu) \in \mathcal{P}(\mathbb{R}^q)$ centered without loss of generality. Then:*

$$\min_{\substack{\mathbf{A} \in \mathbb{R}^{q \times p} \\ \mathbf{A} \mathbf{\Sigma}_\mu \mathbf{A}^\top = \mathbf{\Sigma}_\nu}} \langle \mathbf{M}, \mathbf{A}^\top \mathbf{A} \rangle_\mathcal{F} = \min_{\substack{\mathbf{A} \in \mathbb{R}^{q \times p} \\ \mathbf{A} \mathbf{\Sigma}_\mu \mathbf{A}^\top = \mathbf{\Sigma}_\nu}} \langle -4tr(\mathbf{\Sigma}_\mu).\mathbf{\Sigma}_\mu - 8\mathbf{\Sigma}_\mu^\top \mathbf{\Sigma}_\mu, \mathbf{A}^\top \mathbf{A} \rangle_\mathcal{F} \tag{6.88}$$

Proof. We will use Isserlis' theorem to calculate $\mathbf{M}$ in (6.77). We have for i, j:

$$(\underset{\mathbf{x} \sim \mu}{\mathbb{E}} [(\mathbf{x}^\top \mathbf{x}).\mathbf{x} \mathbf{x}^\top])_{i,j} = \underset{\mathbf{x}}{\mathbb{E}} [\sum_k x_k x_k x_i x_j] = \sum_k \underset{x}{\mathbb{E}} [x_k x_k x_i x_j]$$

$$\overset{*}{=} \sum_k \underset{x}{\mathbb{E}} [x_k x_k] \underset{\mathbf{x}}{\mathbb{E}} [x_i x_j] + \underset{\mathbf{x}}{\mathbb{E}} [x_k x_i] \underset{\mathbf{x}}{\mathbb{E}} [x_k x_j] + \underset{\mathbf{x}}{\mathbb{E}} [x_k x_j] \underset{\mathbf{x}}{\mathbb{E}} [x_k x_i]$$

$$= \sum_k \mathbf{\Sigma}_\mu^{k,k} \mathbf{\Sigma}_\mu^{i,j} + \mathbf{\Sigma}_\mu^{k,i} \mathbf{\Sigma}_\mu^{k,j} + \mathbf{\Sigma}_\mu^{k,j} \mathbf{\Sigma}_\mu^{k,i} \tag{6.89}$$

$$= \sum_k \mathbf{\Sigma}_\mu^{k,k} \mathbf{\Sigma}_\mu^{i,j} + 2\mathbf{\Sigma}_\mu^{k,i} \mathbf{\Sigma}_\mu^{k,j}$$

$$= \mathbf{\Sigma}_\mu^{i,j} \mathrm{tr}(\mathbf{\Sigma}_\mu) + 2(\mathbf{\Sigma}_\mu^\top \mathbf{\Sigma}_\mu)_{i,j}$$

In (*) we used Isserlis' theorem. Hence $\mathbb{E}_{\mathbf{x}\sim\mu}[(\mathbf{x}^\top\mathbf{x}).\mathbf{x}\mathbf{x}^\top] = \text{tr}(\boldsymbol{\Sigma}_\mu).\boldsymbol{\Sigma}_\mu + 2\boldsymbol{\Sigma}_\mu^\top\boldsymbol{\Sigma}_\mu$. In the same way:

$$
\begin{aligned}
(\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[(\mathbf{x}'^\top\mathbf{x}').\mathbf{x}\mathbf{x}^\top])_{i,j} &= \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[\sum_k x_k' x_k' x_i x_j] = \sum_k \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_k' x_i x_j] \\
&= \sum_k \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_k'] \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_i x_j] + \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_i] \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_j] + \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_j] \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_i] \\
&\overset{*}{=} \sum_k \boldsymbol{\Sigma}_\mu^{k,k} \boldsymbol{\Sigma}_\mu^{i,j} = \text{tr}(\boldsymbol{\Sigma}_\mu)\boldsymbol{\Sigma}_\mu^{i,j}
\end{aligned}
\tag{6.90}
$$

In (*) we used the independence of $\mathbf{x}$ and $\mathbf{x}'$ and the fact that Gaussians are centered. Finally,

$$
\begin{aligned}
(\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[(\mathbf{x}^\top\mathbf{x}').\mathbf{x}'\mathbf{x}^\top])_{i,j} &= \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[\sum_k x_k x_k' x_i' x_j] = \sum_k \mathop{\mathbb{E}}_{x,x'}[x_k' x_k x_i' x_j] \\
&= \sum_k \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_k] \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_i' x_j] + \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_i'] \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k x_j] + \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{x}'}[x_k' x_j] \mathop{\mathbb{E}}_{x}[x_k x_i'] \\
&= \sum_k \boldsymbol{\Sigma}_\mu^{k,i} \boldsymbol{\Sigma}_\mu^{k,j} = (\boldsymbol{\Sigma}_\mu^\top\boldsymbol{\Sigma}_\mu)_{i,j}
\end{aligned}
\tag{6.91}
$$

Overall we want to solve the following optimization problem:

$$
\min_{\mathbf{A},\mathbf{A}\boldsymbol{\Sigma}_\mu\mathbf{A}^\top=\boldsymbol{\Sigma}_\nu} \langle\mathbf{M},\mathbf{A}^\top\mathbf{A}\rangle_\mathcal{F}
\tag{6.92}
$$

where:

$$
\mathbf{M} = -4\text{tr}(\boldsymbol{\Sigma}_\mu).\boldsymbol{\Sigma}_\mu - 8\boldsymbol{\Sigma}_\mu^\top\boldsymbol{\Sigma}_\mu
\tag{6.93}
$$

$\square$

We can now prove the main result:

Of Theorem 4.2.6. With previous notations and calculus, we consider the following change of variable:

$$
\mathbf{A} = \mathbf{V}_\nu\mathbf{D}_\nu^{1/2}\mathbf{B}\mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top
\tag{6.94}
$$

then the pushforward equality becomes:

$$
\mathbf{B}\mathbf{B}^\top = \mathbf{I}_q \; i.e. \; \mathbf{B} \in \mathbb{V}_p(\mathbb{R}^q)
\tag{6.95}
$$

Indeed:

$$
\begin{aligned}
\mathbf{A}\boldsymbol{\Sigma}_\mu\mathbf{A}^\top &= (\mathbf{V}_\nu\mathbf{D}_\nu^{1/2}\mathbf{B}\mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top)(\mathbf{V}_\mu\mathbf{D}_\mu\mathbf{V}_\mu^\top)\mathbf{A}^\top \\
&= \mathbf{V}_\nu\mathbf{D}_\nu^{1/2}\mathbf{B}\mathbf{D}_\mu^{1/2}\mathbf{V}_\mu^\top\mathbf{A}^\top \\
&= \mathbf{V}_\nu\mathbf{D}_\nu^{1/2}\mathbf{B}\mathbf{D}_\mu^{1/2}\mathbf{V}_\mu^\top(\mathbf{V}_\mu\mathbf{D}_\mu^{-1/2}\mathbf{B}^\top\mathbf{D}_\nu^{1/2}\mathbf{V}_\nu^\top) \\
&= \mathbf{V}_\nu\mathbf{D}_\nu^{1/2}\mathbf{B}\mathbf{B}^\top\mathbf{D}_\nu^{1/2}\mathbf{V}_\nu^\top \\
&= \boldsymbol{\Sigma}_\nu \iff \mathbf{V}_\nu\mathbf{D}_\nu^{1/2}\mathbf{B}\mathbf{B}^\top\mathbf{D}_\nu^{1/2}\mathbf{V}_\nu^\top = \mathbf{V}_\nu\mathbf{D}_\nu\mathbf{V}_\nu^\top \\
&\iff \mathbf{B}\mathbf{B}^\top = \mathbf{I}_q \iff \mathbf{B} \in \mathbb{V}_p(\mathbb{R}^q)
\end{aligned}
\tag{6.96}
$$

With this change of variable the criterion becomes:

$$
\begin{aligned}
J_3(\mathbf{B}) &= \langle\mathbf{M},(\mathbf{V}_\mu\mathbf{D}_\mu^{-1/2}\mathbf{B}^\top\mathbf{D}_\nu^{1/2}\mathbf{V}_\nu^\top)(\mathbf{V}_\nu\mathbf{D}_\nu^{1/2}\mathbf{B}\mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top)\rangle_\mathcal{F} \\
&= \langle\mathbf{M},\mathbf{V}_\mu\mathbf{D}_\mu^{-1/2}\mathbf{B}^\top\mathbf{D}_\nu\mathbf{B}\mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top\rangle_\mathcal{F} \\
&= \langle\mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top\mathbf{M}\mathbf{V}_\mu\mathbf{D}_\mu^{-1/2},\mathbf{B}^\top\mathbf{D}_\nu\mathbf{B}\rangle_\mathcal{F} \\
&= \langle\tilde{\mathbf{M}},\mathbf{B}^\top\mathbf{D}_\nu\mathbf{B}\rangle_\mathcal{F}
\end{aligned}
\tag{6.97}
$$

with

$$
\begin{aligned}
\tilde{\mathbf{M}} &= \mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top \mathbf{M}\mathbf{V}_\mu \mathbf{D}_\mu^{-1/2} \\
&= \mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top \left(-4\mathrm{tr}(\boldsymbol{\Sigma}_\mu).\boldsymbol{\Sigma}_\mu - 8\boldsymbol{\Sigma}_\mu^\top \boldsymbol{\Sigma}_\mu\right)\mathbf{V}_\mu \mathbf{D}_\mu^{-1/2} \\
&= \mathbf{D}_\mu^{-1/2}\left(-4\mathrm{tr}(\boldsymbol{\Sigma}_\mu)\mathbf{D}_\mu\right)\mathbf{D}_\mu^{-1/2} + \mathbf{D}_\mu^{-1/2}(-8\mathbf{D}_\mu^2)\mathbf{D}_\mu^{-1/2} \\
&= -4\mathrm{tr}(\boldsymbol{\Sigma}_\mu).\mathbf{I}_p - 8\mathbf{D}_\mu
\end{aligned} \tag{6.98}
$$

Overall we have the following optimization problem:

$$
\begin{aligned}
\min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}_q} \langle \tilde{\mathbf{M}}, \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B}\rangle_\mathcal{F} &= \min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}} \mathrm{tr}\left((-4\mathrm{tr}(\boldsymbol{\Sigma}_\mu).\mathbf{I}_p - 8\mathbf{D}_\mu)\mathbf{B}^\top \mathbf{D}_\nu \mathbf{B}\right) \\
&= \min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}_q} -4\mathrm{tr}(\boldsymbol{\Sigma}_\mu).\mathrm{tr}\left(\mathbf{B}^\top \mathbf{D}_\nu \mathbf{B}\right) - 8\mathrm{tr}(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B}) \\
&= -4\mathrm{tr}(\boldsymbol{\Sigma}_\mu).\mathrm{tr}\left(\boldsymbol{\Sigma}_\nu\right) + 8\min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}_q} -\mathrm{tr}(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B})
\end{aligned} \tag{6.99}
$$

Note that for the final cost we need also to compute:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mu}[\|\mathbf{x}-\mathbf{x}'\|_2^4] &= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mu}[(\|\mathbf{x}\|_2^2 - 2\langle\mathbf{x},\mathbf{x}'\rangle_p + \|\mathbf{x}'\|_2^2)(\|\mathbf{x}\|_2^2 - 2\langle\mathbf{x},\mathbf{x}'\rangle_p + \|\mathbf{x}'\|_2^2)] \\
&= \mathrm{tr}(\boldsymbol{\Sigma}_\mu)^2 + 2\mathrm{tr}(\boldsymbol{\Sigma}_\mu^\top \boldsymbol{\Sigma}_\mu) + \mathrm{tr}(\boldsymbol{\Sigma}_\mu)^2 + 4\mathrm{tr}(\boldsymbol{\Sigma}_\mu^\top \boldsymbol{\Sigma}_\mu) + \mathrm{tr}(\boldsymbol{\Sigma}_\mu)^2 + \mathrm{tr}(\boldsymbol{\Sigma}_\mu)^2 + 2\mathrm{tr}(\boldsymbol{\Sigma}_\mu^\top \boldsymbol{\Sigma}_\mu) \\
&= 4\mathrm{tr}(\boldsymbol{\Sigma}_\mu)^2 + 8\mathrm{tr}(\boldsymbol{\Sigma}_\mu^\top \boldsymbol{\Sigma}_\mu)
\end{aligned} \tag{6.100}
$$

Overall we have:

$$
\begin{aligned}
\min_{\substack{T\#\mu=\nu \\ \text{T linear}}} J(T) &= 4\mathrm{tr}(\boldsymbol{\Sigma}_\mu)^2 + 8\mathrm{tr}(\boldsymbol{\Sigma}_\mu^\top \boldsymbol{\Sigma}_\mu) + 4\mathrm{tr}(\boldsymbol{\Sigma}_\nu)^2 + 8\mathrm{tr}(\boldsymbol{\Sigma}_\nu^\top \boldsymbol{\Sigma}_\nu) \\
&\quad + 2\left(-4\mathrm{tr}(\boldsymbol{\Sigma}_\mu).\mathrm{tr}\left(\boldsymbol{\Sigma}_\nu\right) + 8\min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}_q} -\mathrm{tr}(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B})\right) \\
&= 4\mathrm{tr}(\boldsymbol{\Sigma}_\mu)^2 + 8\mathrm{tr}(\boldsymbol{\Sigma}_\mu^\top \boldsymbol{\Sigma}_\mu) + 4\mathrm{tr}(\boldsymbol{\Sigma}_\nu)^2 + 8\mathrm{tr}(\boldsymbol{\Sigma}_\nu^\top \boldsymbol{\Sigma}_\nu) - 8\mathrm{tr}(\boldsymbol{\Sigma}_\mu).\mathrm{tr}\left(\boldsymbol{\Sigma}_\nu\right) \\
&\quad + 16\min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}} -\mathrm{tr}(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B}) \\
&= 4(\mathrm{tr}(\boldsymbol{\Sigma}_\mu) - \mathrm{tr}(\boldsymbol{\Sigma}_\nu))^2 + 8(\mathrm{tr}(\boldsymbol{\Sigma}_\mu^T \boldsymbol{\Sigma}_\mu) + \mathrm{tr}(\boldsymbol{\Sigma}_\nu^T \boldsymbol{\Sigma}_\nu)) + 16\min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}_q} -\mathrm{tr}(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B})
\end{aligned}
$$

Since the covariances are symmetric it gives (4.39).

We can use Proposition 6.2.3 to solve $\min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}_q} -\mathrm{tr}(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B})$. Indeed $-\mathbf{D}_\mu$ is already diagonal which values are nonincreasing and $\mathbf{D}_\nu$ is already diagonal which values are nondecreasing. In this way when $p=q$ $\min_{\mathbf{B},\mathbf{B}\mathbf{B}^\top=\mathbf{I}_q} -\mathrm{tr}(\mathbf{D}_\mu \mathbf{B}^\top \mathbf{D}_\nu \mathbf{B})$ can be solved in close form with $\mathbf{B}=\mathbf{I}_q$ which corresponds to $\mathbf{A} = \mathbf{V}_\nu \mathbf{D}_\nu^{1/2}\mathbf{D}_\mu^{-1/2}\mathbf{V}_\mu^\top = \boldsymbol{\Sigma}_\nu^{1/2}\mathbf{V}_\nu \mathbf{V}_\mu^\top \boldsymbol{\Sigma}_\mu^{-1/2}$. □

6.3 Proofs and additional results of Chapter 5

This section contains all the proofs of the claims and additional results of the Chapter 5. We recall the notations of the Chapter. Two datasets are represented by matrices $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ and $\mathbf{X}' = [\mathbf{x}'_1,\ldots,\mathbf{x}'_{n'}]^T \in \mathbb{R}^{n'\times d'}$. The rows of the datasets are denoted as *samples* and their columns as *features*. Let $\mu = \sum_{i=1}^n w_i\delta_{\mathbf{x}_i}$ and $\mu' = \sum_{i=1}^{n'} w'_i\delta_{\mathbf{x}'_i}$ be two empirical distributions related to the samples, where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{x}'_i \in \mathbb{R}^{d'}$. We refer in the following to $\mathbf{w} = [w_1,\ldots,w_n]^\top$ and $\mathbf{w}' = [w'_1,\ldots,w'_{n'}]^\top$ as to sample weights vectors that both lie in the simplex ($\mathbf{w} \in \Delta_n$ and $\mathbf{w}' \in \Delta_{n'}$). In addition to them, we also introduce weights for the features that are stored on vectors $\mathbf{v} \in \Delta_d$ and $\mathbf{v}' \in \Delta_{d'}$. Finally, we let vec denote the column-stacking operator.

6.3.1 Proof of Proposition 5.3.1 – COOT is a distance

We recall the proposition:

Proposition (COOT is a distance). *Suppose $L = |\cdot|^p, p \geq 1$, $n = n', d = d'$ and that the weights $\mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}'$ are uniform. Then $COOT(\mathbf{X}, \mathbf{X}') = 0$ iff there exists a permutation of the samples $\sigma_1 \in S_n$ and of the features $\sigma_2 \in S_d$, s.t, $\forall i, k\ \mathbf{X}_{i,k} = \mathbf{X}'_{\sigma_1(i),\sigma_2(k)}$. Moreover, it is symmetric and satisfies the triangular inequality as long as L satisfies the triangle inequality, i.e., $COOT(\mathbf{X}, \mathbf{X}'') \leq COOT(\mathbf{X}, \mathbf{X}') + COOT(\mathbf{X}', \mathbf{X}'')$.*

Proof. The symmetry follows from the definition of COOT. To prove the triangle inequality of COOT for arbitrary measures, we will use the gluing lemma (see [Villani 2008]) which states the existence of couplings with a prescribed structure. Let $\mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{X}' \in \mathbb{R}^{n' \times d'}, \mathbf{X}'' \in \mathbb{R}^{n'' \times d''}$ associated with $\mathbf{w} \in \Sigma_n, \mathbf{v} \in \Sigma_d, \mathbf{w}' \in \Sigma'_n, \mathbf{v}' \in \Sigma'_d, \mathbf{w}'' \in \Sigma''_n, \mathbf{v}'' \in \Sigma''_d$. Without loss of generality, we can suppose in the proof that all weights are different from zeros (otherwise we can consider $\tilde{w}_i = w_i$ if $w_i > 0$ and $\tilde{w}_i = 1$ if $w_i = 0$ see proof of Proposition 2.2 in [Peyré 2019])

Let $(\boldsymbol{\pi}_1^s, \boldsymbol{\pi}_1^v)$ and $(\boldsymbol{\pi}_2^s, \boldsymbol{\pi}_2^v)$ be two couples of optimal solutions for the COOT problems associated with $COOT(\mathbf{X}, \mathbf{X}', \mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}')$ and $COOT(\mathbf{X}', \mathbf{X}'', \mathbf{w}', \mathbf{w}'', \mathbf{v}', \mathbf{v}'')$ respectively.

We define:
$$S_1 = \boldsymbol{\pi}_1^s \mathrm{diag}\left(\frac{1}{\mathbf{w}'}\right) \boldsymbol{\pi}_2^s, \quad S_2 = \boldsymbol{\pi}_1^v \mathrm{diag}\left(\frac{1}{\mathbf{v}'}\right) \boldsymbol{\pi}_2^v$$

Then, it is easy to check that $S_1 \in \Pi(\mathbf{w}, \mathbf{w}'')$ and $S_2 \in \Pi(\mathbf{v}, \mathbf{v}'')$ (see *e.g* Proposition 2.2 in [Peyré 2019]). We now show the following:

$$\mathrm{COOT}(\mathbf{X}, \mathbf{X}'', \mathbf{w}, \mathbf{w}'', \mathbf{v}, \mathbf{v}'') \overset{*}{\leq} \langle \mathbf{L}(\mathbf{X}, \mathbf{X}'') \otimes S_1, S_2 \rangle = \langle \mathbf{L}(\mathbf{X}, \mathbf{X}'') \otimes [\boldsymbol{\pi}_1^s \mathrm{diag}(\frac{1}{\mathbf{w}'}) \boldsymbol{\pi}_2^s], [\boldsymbol{\pi}_1^v \mathrm{diag}(\frac{1}{\mathbf{v}'}) \boldsymbol{\pi}_2^v] \rangle$$

$$\overset{**}{\leq} \langle [\mathbf{L}(\mathbf{X}, \mathbf{X}') + \mathbf{L}(\mathbf{X}', \mathbf{X}'')] \otimes [\boldsymbol{\pi}_1^s \mathrm{diag}(\frac{1}{\mathbf{w}'}) \boldsymbol{\pi}_2^s], [\boldsymbol{\pi}_1^v \mathrm{diag}(\frac{1}{\mathbf{v}'}) \boldsymbol{\pi}_2^v] \rangle$$

$$= \langle \mathbf{L}(\mathbf{X}, \mathbf{X}') \otimes [\boldsymbol{\pi}_1^s \mathrm{diag}(\frac{1}{\mathbf{w}'}) \boldsymbol{\pi}_2^s], [\boldsymbol{\pi}_1^v \mathrm{diag}(\frac{1}{\mathbf{v}'}) \boldsymbol{\pi}_2^v] \rangle + \langle \mathbf{L}(\mathbf{X}', \mathbf{X}'') \otimes [\boldsymbol{\pi}_1^s \mathrm{diag}(\frac{1}{\mathbf{w}'}) \boldsymbol{\pi}_2^s], [\boldsymbol{\pi}_1^v \mathrm{diag}(\frac{1}{\mathbf{v}'}) \boldsymbol{\pi}_2^v] \rangle,$$

where in (*) we used the suboptimality of S_1, S_2 and in (**) the fact that L satisfies the triangle inequality. Now note that:

$$\langle \mathbf{L}(\mathbf{X}, \mathbf{X}') \otimes [\boldsymbol{\pi}_1^s \mathrm{diag}(\frac{1}{\mathbf{w}'}) \boldsymbol{\pi}_2^s], [\boldsymbol{\pi}_1^v \mathrm{diag}(\frac{1}{\mathbf{v}'}) \boldsymbol{\pi}_2^v] \rangle + \langle \mathbf{L}(\mathbf{X}', \mathbf{X}'') \otimes [\boldsymbol{\pi}_1^s \mathrm{diag}(\frac{1}{\mathbf{w}'}) \boldsymbol{\pi}_2^s], [\boldsymbol{\pi}_1^v \mathrm{diag}(\frac{1}{\mathbf{v}'}) \boldsymbol{\pi}_2^v] \rangle$$

$$= \sum_{i,j,k,l,e,o} L(X_{i,k}, X'_{e,o}) \frac{\boldsymbol{\pi}_{1 i,e}^s \boldsymbol{\pi}_{2 e,j}^s}{w'_e} \frac{\boldsymbol{\pi}_{1 k,o}^v \boldsymbol{\pi}_{2 o,l}^v}{v'_o} + \sum_{i,j,k,l,e,o} L(X'_{e,o}, X''_{j,l}) \frac{\boldsymbol{\pi}_{1 i,e}^s \boldsymbol{\pi}_{2 e,j}^s}{w'_e} \frac{\boldsymbol{\pi}_{1 k,o}^v \boldsymbol{\pi}_{2 o,l}^v}{v'_o}$$

$$\overset{*}{=} \sum_{i,k,e,o} L(X_{i,k}, X'_{e,o}) \boldsymbol{\pi}_{1 i,e}^s \boldsymbol{\pi}_{1 k,o}^v + \sum_{l,j,e,o} L(X'_{e,o}, X''_{j,l}) \boldsymbol{\pi}_{2 e,j}^s \boldsymbol{\pi}_{2 o,l}^v$$

where in (*) we used:

$$\sum_j \frac{\boldsymbol{\pi}_{2 e,j}^s}{w'_e} = 1, \quad \sum_l \frac{\boldsymbol{\pi}_{2 o,l}^v}{v'_o} = 1, \quad \sum_i \frac{\boldsymbol{\pi}_{1 i,e}^s}{w'_e} = 1, \quad \sum_k \frac{\boldsymbol{\pi}_{1 k,o}^v}{v'_o} = 1$$

Overall, from the definition of $\boldsymbol{\pi}_1^s, \boldsymbol{\pi}_1^v$ and $\boldsymbol{\pi}_2^s, \boldsymbol{\pi}_2^v$ we have:

$$\mathrm{COOT}(\mathbf{X}, \mathbf{X}'', \mathbf{w}, \mathbf{w}'', \mathbf{v}, \mathbf{v}'') \leq \mathrm{COOT}(\mathbf{X}, \mathbf{X}', \mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}') + \mathrm{COOT}(\mathbf{X}', \mathbf{X}'', \mathbf{w}', \mathbf{w}'', \mathbf{v}', \mathbf{v}'').$$

For the identity of indiscernibles, suppose that $n = n', d = d'$ and that the weights $\mathbf{w}, \mathbf{w}', \mathbf{v}, \mathbf{v}'$ are uniform. Suppose that there exists a permutation of the samples $\sigma_1 \in S_n$ and of the features $\sigma_2 \in S_d$,

$s.t \ \forall i, k \in [\![n]\!] \times [\![d]\!]$, $\mathbf{X}_{i,k} = \mathbf{X}'_{\sigma_1(i),\sigma_2(k)}$. We define the couplings π^s, π^v supported on the graphs of the permutations σ_1, σ_2 respectively, *i.e* $\pi^s = (id \times \sigma_1)$ and $\pi^v = (id \times \sigma_2)$. These couplings have the prescribed marginals and lead to a zero cost hence are optimal.

Conversely, as described in the chapter, there always exists an optimal solution of (COOT) which lies on extremal points of the polytopes $\Pi(\mathbf{w}, \mathbf{w}')$ and $\Pi(\mathbf{v}, \mathbf{v}')$. When $n = n', d = d'$ and uniform weights are used, Birkhoff's theorem [Birkhoff 1946] states that the set of extremal points of $\Pi(\frac{\mathbf{1}_n}{n}, \frac{\mathbf{1}_n}{n})$ and $\Pi(\frac{\mathbf{1}_d}{d}, \frac{\mathbf{1}_d}{d})$ are the set of permutation matrices so there exists an optimal solution $(\boldsymbol{\pi}^s_*, \boldsymbol{\pi}^v_*)$ supported on σ^s_*, σ^v_* respectively with $\sigma^s_*, \sigma^v_* \in \mathbb{S}_n \times \mathbb{S}_d$. Then, if $\text{COOT}(\mathbf{X}, \mathbf{X}') = 0$, it implies that $\sum_{i,k} L(X_{i,k}, X'_{\sigma^s_*(i),\sigma^v_*(k)}) = 0$. If $L = |\cdot|^p$ then $X_{i,k} = X'_{\sigma^s_*(i),\sigma^v_*(k)}$ which gives the desired result. If $n \neq n', d \neq d'$ the COOT cost is always strictly positive as there exists a strictly positive element outside the diagonal. $\qquad\square$

6.3.2   Complexity of computing the value of COOT

We recall the result:

Lemma. *The overall computational complexity of computing the value of COOT when $L = |.|^2$ is $O(\min\{(n + n')dd' + n'^2 n; (d + d')nn' + d'^2 d\})$.*

Proof. As mentionned in [Peyré 2016], if L can be written as $L(a, b) = f(a) + f(b) - h_1(a)h_2(b)$ then we have that

$$\mathbf{L}(\mathbf{X}, \mathbf{X}') \otimes \boldsymbol{\pi}^s = \mathbf{C}_{\mathbf{X},\mathbf{X}'} - h_1(\mathbf{X})\boldsymbol{\pi}^s h_2(\mathbf{X}')^T,$$

where $\mathbf{C}_{\mathbf{X},\mathbf{X}'} = \mathbf{X}\mathbf{w}\mathbb{1}_{n'}^T + \mathbb{1}_n \mathbf{w}'^T \mathbf{X}'^T$ so that the latter can be computed in $O(ndd' + n'dd') = O((n + n')dd')$. To compute the final cost, we must also calculate the inner product with $\boldsymbol{\pi}^v$ that can be done in $O(n'^2 n)$ making the complexity of $\langle \mathbf{L}(\mathbf{X}, \mathbf{X}') \otimes \boldsymbol{\pi}^s, \boldsymbol{\pi}^v \rangle$ equal to $O((n + n')dd' + n'^2 n)$.

Finally, as the cost is symmetric *w.r.t* $\boldsymbol{\pi}^s, \boldsymbol{\pi}^v$, we obtain the overall complexity of $O(\min\{(n + n')dd' + n'^2 n; (d + d')nn' + d'^2 d\})$. $\qquad\square$

6.3.3   Proofs of theorem 5.5.1 – Equivalence between QAP and BAP

As pointed in [Konno 1976b], we can relate the solutions of a QAP and a BAP. We will prove the equivalent following result (maximization version of theorem 5.5.1):

Theorem. *If $\mathbf{Q}$ is a positive semi-definite matrix, then problems:*

$$
\begin{aligned}
\max_{\mathbf{x}} f(\mathbf{x}) \quad &= \mathbf{c}^T \mathbf{x} + \tfrac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x} \\
s.t. \quad &\mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \geq 0
\end{aligned}
\tag{6.101}
$$

$$
\begin{aligned}
\max_{\mathbf{x},\mathbf{y}} g(\mathbf{x}, \mathbf{y}) \quad &= \tfrac{1}{2}\mathbf{c}^T \mathbf{x} + \tfrac{1}{2}\mathbf{c}^T \mathbf{y} + \tfrac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{y} \\
s.t. \quad &\mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{A}\mathbf{y} = \mathbf{b}, \ \mathbf{x}, \mathbf{y} \geq 0
\end{aligned}
\tag{6.102}
$$

are equivalent. More precisely, if $\mathbf{x}^*$ is an optimal solution for (6.101)*, then $(\mathbf{x}^*, \mathbf{x}^*)$ is a solution for* (6.102) *and if $(\mathbf{x}^*, \mathbf{y}^*)$ is optimal for* (6.102)*, then both $\mathbf{x}^*$ and $\mathbf{y}^*$ are optimal for* (6.101)*.*

Proof. This proof follows the proof of Theorem 2.2 in [Konno 1976b]. Let $\mathbf{z}^*$ be optimal for (6.101) and $(\mathbf{x}^*, \mathbf{y}^*)$ be optimal for (6.102). Then, by definition, for all $\mathbf{x}$ satisfying the constraints of (6.101), $f(\mathbf{z}^*) \geq f(\mathbf{x})$. In particular, $f(\mathbf{z}^*) \geq f(\mathbf{x}^*) = g(\mathbf{x}^*, \mathbf{x}^*)$ and $f(\mathbf{z}^*) \geq f(\mathbf{y}^*) = g(\mathbf{y}^*, \mathbf{y}^*)$. Also, $g(\mathbf{x}^*, \mathbf{y}^*) \geq \max_{\mathbf{x},\mathbf{x} \ \text{s.t} \ \mathbf{A}\mathbf{x}=\mathbf{b},\mathbf{x}\geq 0} g(\mathbf{x}, \mathbf{x}) = f(\mathbf{z}^*)$.

To prove the theorem, it suffices to prove that

$$f(\mathbf{y}^*) = f(\mathbf{x}^*) = g(\mathbf{x}^*, \mathbf{y}^*) \tag{6.103}$$

since, in this case, $g(\mathbf{x}^*, \mathbf{y}^*) = f(\mathbf{x}^*) \geq f(\mathbf{z}^*)$ and $g(\mathbf{x}^*, \mathbf{y}^*) = f(\mathbf{y}^*) \geq f(\mathbf{z}^*)$.

Let us prove (6.103). Since $(\mathbf{x}^*, \mathbf{y}^*)$ is optimal, we have:

$$0 \leq g(\mathbf{x}^*, \mathbf{y}^*) - g(\mathbf{x}^*, \mathbf{x}^*) = \frac{1}{2}\mathbf{c}^T(\mathbf{y}^* - \mathbf{x}^*) + \frac{1}{2}\mathbf{x}^{*T}\mathbf{Q}(\mathbf{y}^* - \mathbf{x}^*)$$

$$0 \leq g(\mathbf{x}^*, \mathbf{y}^*) - g(\mathbf{y}^*, \mathbf{y}^*) = \frac{1}{2}\mathbf{c}^T(\mathbf{x}^* - \mathbf{y}^*) + \frac{1}{2}\mathbf{y}^{*T}\mathbf{Q}(\mathbf{x}^* - \mathbf{y}^*).$$

By adding these inequalities we obtain:

$$(\mathbf{x}^* - \mathbf{y}^*)^T \mathbf{Q}(\mathbf{x}^* - \mathbf{y}^*) \leq 0.$$

Since $\mathbf{Q}$ is positive semi-definite, this implies that $\mathbf{Q}(\mathbf{x}^* - \mathbf{y}^*) = 0$. So, using previous inequalities, we have $\mathbf{c}^T(\mathbf{x}^* - \mathbf{y}^*) = 0$, hence $g(\mathbf{x}^*, \mathbf{y}^*) = g(\mathbf{x}^*, \mathbf{x}^*) = g(\mathbf{y}^*, \mathbf{y}^*)$ as required.

Note also that this result holds when we add a constant term to the cost function. □

6.3.4 Proofs of Proposition 5.5.1 – Concavity of GW

We recall the proposition:

Proposition. *Let* $L = |\cdot|^2$ *and suppose that* $\mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{C}' \in \mathbb{R}^{n' \times n'}$ *are squared Euclidean distance matrices such that* $\mathbf{C} = \mathbf{x}\mathbf{1}_n^T + \mathbf{1}_n\mathbf{x}^T - 2\mathbf{X}\mathbf{X}^T, \mathbf{C}' = \mathbf{x}'\mathbf{1}_{n'}^T + \mathbf{1}_{n'}\mathbf{x}'^T - 2\mathbf{X}'\mathbf{X}'^T$ *with* $\mathbf{x} = diag(\mathbf{X}\mathbf{X}^T), \mathbf{x}' = diag(\mathbf{X}'\mathbf{X}'^T)$. *Then, the GW problem can be written as a concave quadratic program (QP) which Hessian reads* $\mathbf{Q} = -4 * \mathbf{X}\mathbf{X}^T \otimes_K \mathbf{X}'\mathbf{X}'^T$.

If $\mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{C}' \in \mathbb{R}^{n' \times n'}$ *are inner products similarities, i.e. such that* $\mathbf{C} = \mathbf{X}\mathbf{X}^T, \mathbf{C}' = \mathbf{X}'\mathbf{X}'^T$ *then the GW is also a concave quadratic program (QP) which Hessian reads* $\mathbf{Q} = -2 * \mathbf{X}\mathbf{X}^T \otimes_K \mathbf{X}'\mathbf{X}'^T$.

This result is a consequence of the following lemma.

Lemma 6.3.1. *When* $\mathbf{C}, \mathbf{C}'$ *are squared Euclidean distance matrices as defined previously, the GW problem can be formulated as:*

$$GW(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}') = \min_{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}')} -4 vec(\mathbf{M})^T vec(\boldsymbol{\pi}^s) - 8 vec(\boldsymbol{\pi}^s)^T \mathbf{Q} vec(\boldsymbol{\pi}^s) + Cte$$

with

$$\mathbf{M} = \mathbf{x}\mathbf{x}'^T - 2\mathbf{x}\mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T - 2\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T \ and \ \mathbf{Q} = \mathbf{X}\mathbf{X}^T \otimes_K \mathbf{X}'\mathbf{X}'^T,$$

$$Cte = \sum_i \|\mathbf{x}_i - \mathbf{x}_j\|_2^4 \mathbf{w}_i \mathbf{w}_j + \sum_i \|\mathbf{x}'_i - \mathbf{x}'_j\|_2^4 \mathbf{w}'_i \mathbf{w}'_j - 4\mathbf{w}^T \mathbf{x}\mathbf{w}'^T \mathbf{x}'$$

When $\mathbf{C}, \mathbf{C}'$ *are inner product similarities as defined in Proposition 5.5.1, the GW problem can be formulated as:*

$$GW(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}') = \min_{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}')} -4 vec(\boldsymbol{\pi}^s)^T \mathbf{Q} vec(\boldsymbol{\pi}^s) + Cte'$$

Proof. Using the results in [Peyré 2016] for $L = |\cdot|^2$, we have $\mathbf{L}(\mathbf{C}, \mathbf{C}') \otimes \boldsymbol{\pi}^s = c_{\mathbf{C}, \mathbf{C}'} - 2\mathbf{C}\boldsymbol{\pi}^s\mathbf{C}'$ with $c_{\mathbf{C}, \mathbf{C}'} = (\mathbf{C})^2\mathbf{w}\mathbf{1}_{n'}^T + \mathbf{1}_n\mathbf{w}'^T(\mathbf{C}')^2$, where $(\mathbf{C})^2 = (\mathbf{C}_{i,j}^2)$ is applied element-wise.

We now have that:

$$\langle \mathbf{C}\boldsymbol{\pi}^s\mathbf{C}', \boldsymbol{\pi}^s \rangle = \text{tr}\left[\boldsymbol{\pi}^{sT}(\mathbf{x}\mathbf{1}_n^T + \mathbf{1}_n\mathbf{x}^T - 2\mathbf{X}\mathbf{X}^T)\boldsymbol{\pi}^s(\mathbf{x}'\mathbf{1}_{n'}^T + \mathbf{1}_{n'}\mathbf{x}'^T - 2\mathbf{X}'\mathbf{X}'^T)\right]$$

$$= \text{tr}\left[(\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{1}_n^T + \mathbf{w}'\mathbf{x}^T - 2\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T)(\boldsymbol{\pi}^s\mathbf{x}'\mathbf{1}_{n'}^T + \mathbf{w}\mathbf{x}'^T - 2\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T)\right]$$

$$= \text{tr}\left[\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{x}'\mathbf{1}_{n'}^T + \boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{x}'^T - 2\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T + \mathbf{w}'\mathbf{x}^T\boldsymbol{\pi}^s\mathbf{x}'\mathbf{1}_{n'}^T + \mathbf{w}'\mathbf{x}^T\mathbf{w}\mathbf{x}'^T - 2\mathbf{w}'\mathbf{x}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T \right.$$
$$\left. - 2\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{x}'\mathbf{1}_{n'}^T - 2\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T + 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T\right]$$

$$\overset{*}{=} \text{tr}\left[\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T(\mathbf{x}'\mathbf{1}_{n'}^T + \mathbf{1}_{n'}\mathbf{x}'^T) + \boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{x}'^T + \mathbf{w}'\mathbf{x}^T\mathbf{w}\mathbf{x}'^T - 2\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T - 2\mathbf{w}'\mathbf{x}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T \right.$$
$$\left. - 2\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{x}'\mathbf{1}_{n'}^T - 2\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T + 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T\right],$$

where in (*) we used:

$$\text{tr}(\mathbf{w}'\mathbf{x}^T\boldsymbol{\pi}^s\mathbf{x}'\mathbf{1}_{n'}^T) = \text{tr}(\mathbf{x}'\mathbf{1}_{n'}^T\mathbf{w}'\mathbf{x}^T\boldsymbol{\pi}^s) = \text{tr}(\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{1}_{n'}\mathbf{x}'^T).$$

Moreover, since:

$$\text{tr}(\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{x}'\mathbf{1}_{n'}^T) = \text{tr}(\mathbf{1}_{n'}^T\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{x}') = \text{tr}(\mathbf{w}^T\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{x}') = \text{tr}(\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T)$$

and $\text{tr}(\mathbf{w}'\mathbf{x}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T) = \text{tr}(\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T)$, we can simplify the last expression to obtain:

$$\begin{aligned}\langle \mathbf{C}\boldsymbol{\pi}^s\mathbf{C}', \boldsymbol{\pi}^s\rangle = \text{tr}\big[&\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T(\mathbf{x}'\mathbf{1}_{n'}^T + \mathbf{1}_{n'}\mathbf{x}'^T) + \boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{x}'^T + \mathbf{w}'\mathbf{x}^T\mathbf{w}\mathbf{x}'^T \\ & - 4\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T - 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T + 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T\big].\end{aligned}$$

Finally, we have that

$$\begin{aligned}\langle \mathbf{C}\boldsymbol{\pi}^s\mathbf{C}', \boldsymbol{\pi}^s\rangle &= \text{tr}\big[\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{x}'\mathbf{1}_{n'}^T + \boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{1}_{n'}\mathbf{x}'^T + \boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{x}'^T \\ & \quad + \mathbf{w}'\mathbf{x}^T\mathbf{w}\mathbf{x}'^T - 4\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T - 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T + 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T\big] \\ &= \text{tr}\big[2\mathbf{w}'\mathbf{x}^T\mathbf{w}\mathbf{x}'^T + 2\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{x}'^T - 4\boldsymbol{\pi}^{sT}\mathbf{x}\mathbf{w}'^T\mathbf{X}'\mathbf{X}'^T - 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T + 4\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T\big] \\ &= 2\mathbf{w}^T\mathbf{x}\mathbf{w}'^T\mathbf{x}' + 2\langle \mathbf{x}\mathbf{x}'^T - 2\mathbf{x}\mathbf{w}^T\mathbf{X}'\mathbf{X}'^T - 2\mathbf{X}\mathbf{X}^T\mathbf{w}\mathbf{x}'^T, \boldsymbol{\pi}^s\rangle + 4\text{tr}(\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T).\end{aligned}$$

The term $2\mathbf{w}^T\mathbf{x}\mathbf{w}'^T\mathbf{x}'$ is constant since it does not depend on the coupling. Also, we can verify that $c_{\mathbf{C},\mathbf{C}'}$ does not depend on $\boldsymbol{\pi}^s$ as follows:

$$\langle c_{\mathbf{C},\mathbf{C}'}, \boldsymbol{\pi}^s\rangle = \sum_i \|\mathbf{x}_i - \mathbf{x}_j\|_2^4 \mathbf{w}_i\mathbf{w}_j + \sum_i \|\mathbf{x}_i' - \mathbf{x}_j'\|_2^4 \mathbf{w}_i'\mathbf{w}_j'$$

implying that:

$$\langle c_{\mathbf{C},\mathbf{C}'} - 2\mathbf{C}\boldsymbol{\pi}^s\mathbf{C}', \boldsymbol{\pi}^s\rangle = Cte - 4\langle \mathbf{x}\mathbf{x}'^T - 2\mathbf{x}\mathbf{w}^T\mathbf{X}'\mathbf{X}'^T - 2\mathbf{X}^T\mathbf{X}\mathbf{w}\mathbf{x}'^T, \boldsymbol{\pi}^s\rangle - 8\text{tr}(\boldsymbol{\pi}^{sT}\mathbf{X}\mathbf{X}^T\boldsymbol{\pi}^s\mathbf{X}'\mathbf{X}'^T).$$

We can rewrite this equation as stated in the proposition using the vec operator.

Using a standard QP form $\mathbf{c}^T\mathbf{x} + \frac{1}{2}\mathbf{x}\mathbf{Q}'\mathbf{x}^T$ with $\mathbf{c} = -4\text{vec}(\mathbf{M})$ and $\mathbf{Q}' = -4\mathbf{X}\mathbf{X}^T \otimes_K \mathbf{X}'\mathbf{X}'^T$ we see that the Hessian is negative semi-definite as the opposite of a Kronecker product of positive semi-definite matrices $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}'\mathbf{X}'^T$.

The inner product case is the same calculus, only the constant term changes and $\mathbf{M} = 0$. Note that both calculus were also made in Chapter 4, and precisely in lemma 4.2.1. $\square$

6.3.5 Proof of Proposition 5.5.2 – Equality between COOT and GW in the concave regime

We recall the proposition:

Proposition. *Let* $\mathbf{C} \in \mathbb{R}^{n\times n}, \mathbf{C}' \in \mathbb{R}^{n'\times n'}$ *be any symmetric matrices, then:*

$$COOT(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}', \mathbf{w}, \mathbf{w}') \le GW(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}').$$

The converse is also true under the hypothesis of Proposition 5.5.1. In this case, if $(\boldsymbol{\pi}_*^s, \boldsymbol{\pi}_*^v)$ *is an optimal solution of COOT, then both* $\boldsymbol{\pi}_*^s, \boldsymbol{\pi}_*^v$ *are solutions of GW. Conversely, if* $\boldsymbol{\pi}_*^s$ *is an optimal solution of GW, then* $(\boldsymbol{\pi}_*^s, \boldsymbol{\pi}_*^s)$ *is an optimal solution for COOT.*

Proof. The inequality follows from the fact that any optimal solution of the GW problem is an admissible solution for the COOT problem, hence the inequality is true by suboptimality of this optimal solution.

For the equality part, by following the same calculus as in the proof of lemma 6.3.1, we can verify that:

$$
\mathrm{COOT}(\mathbf{C}, \mathbf{C}', \mathbf{w}, \mathbf{w}', \mathbf{w}, \mathbf{w}') = \min_{\boldsymbol{\pi}^s \in \Pi(\mathbf{w}, \mathbf{w}')} -2\mathrm{vec}(\mathbf{M})^T \mathrm{vec}(\boldsymbol{\pi}^s)
$$
$$
- 2\mathrm{vec}(\mathbf{M})^T \mathrm{vec}(\boldsymbol{\pi}^v) - 8\mathrm{vec}(\boldsymbol{\pi}^s)^T \mathbf{Q} \mathrm{vec}(\boldsymbol{\pi}^v) + Cte,
$$

with $\mathbf{M}, \mathbf{Q}$ as defined in lemma 6.3.1. Since $\mathbf{Q}$ is negative semi-definite, we can apply Theorem 5.5.1 to prove that both problems are equivalent and lead to the same cost and that every optimal solution of GW is an optimal solution of COOT and vice versa. Same applies for the inner product case. □

6.3.6   Equivalence of DC algorithm and Frank-Wolfe algorithm for GW

We recall the result:

Proposition. *When $\mathbf{X} = \mathbf{C}$, $\mathbf{X}' = \mathbf{C}'$ are squared Euclidean distance matrices or inner product similarities the iterations of Algorithm 10 are the same as the iteration of the FW procedure defined in Chapter 3 for solving GW (provided that the initialization is the same).*

Proof. Using Proposition 5.5.2, we know that when $\mathbf{X} = \mathbf{C}$, $\mathbf{X}' = \mathbf{C}'$ are squared Euclidean distance matrices or inner product similarities, then there is an optimal solution of the form $(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*)$. In this case, we can set $\boldsymbol{\pi}^s_{(k)} = \boldsymbol{\pi}^v_{(k)}$ during the iterations of Algorithm 9 to obtain an optimal solution for both COOT and GW. This reduces to Algorithm 10 that corresponds to a DC algorithm where the quadratic form is replaced by its linear upper bound.

Below, we prove that this DC algorithm for solving GW problems is equivalent to the Frank-Wolfe (FW) based algorithm presented in Chapter 3 and recalled in Algorithm 11 when $L = |\cdot|^2$ and for squared Euclidean distance matrices $\mathbf{C}', \mathbf{C}''$.

Algorithm 11 FW Algorithm for GW (see Chapter 3)

1: **Input:** maxIt, thd
2: $\pi^{(0)} \leftarrow \mathbf{w}\mathbf{w}'^{\top}$
3: **while** $k <$ maxIt **and** $err >$ thd **do**
4:      $\mathbf{G} \leftarrow$ Gradient from equation (5.2) *w.r.t.* $\boldsymbol{\pi}^s_{(k-1)}$
5:      $\tilde{\boldsymbol{\pi}}^s_{(k)} \leftarrow OT(\mathbf{w}, \mathbf{w}', \mathbf{G})$
6:      $\mathbf{z}_k(\tau) \leftarrow \boldsymbol{\pi}^s_{(k-1)} + \tau(\tilde{\boldsymbol{\pi}}^s_{(k)} - \boldsymbol{\pi}^s_{(k-1)})$ for $\tau \in (0, 1)$
7:      $\tau^{(k)} \leftarrow \underset{\tau \in (0,1)}{\mathrm{argmin}} \langle \mathbf{L}(\mathbf{C}, \mathbf{C}') \otimes \mathbf{z}_k(\tau), \mathbf{z}_k(\tau) \rangle$
8:      $\boldsymbol{\pi}^s_{(k)} \leftarrow (1 - \tau^{(k)})\boldsymbol{\pi}^s_{(k-1)} + \tau^{(k)}\tilde{\boldsymbol{\pi}}^s_{(k)}$
9:      $err \leftarrow ||\boldsymbol{\pi}^s_{(k-1)} - \boldsymbol{\pi}^s_{(k)}||_F$
10:      $k \leftarrow k + 1$
11: **end while**

The cases when $L = |\cdot|^2$ and $\mathbf{C}, \mathbf{C}'$ are squared Euclidean distance matrices or inner product similarities have interesting implications in practice, since in this case the resulting GW problem is a concave QP (as explained in this Chapter and shown in Lemma 6.3.1). In [Maron 2018], the authors investigated the solution to QP with *conditionally concave energies* using a FW algorithm and showed that in this case the line-search step of the FW is always 1. Moreover, as shown in Proposition 6.3.1, the GW problem

can be written as a concave QP with concave energy and is minimizing *a fortiori* a conditionally concave energy. Consequently, the line-search step of the FW algorithm proposed in Chapter 3 and described in Algorithm.11 always leads to an optimal line-search step of 1. In this case, the Algorithm.11 is equivalent to Algorithm.12 goven below, since $\tau^{(k)} = 1$ for all k.

Algorithm 12 FW Algorithm for GW with squared Euclidean distance matrices or inner product similarities

---

1: **Input:** maxIt, thd
2: $\pi^{(0)} \leftarrow \mathbf{w}\mathbf{w}'^{\top}$
3: **while** $k < \text{maxIt}$ **and** $err > \text{thd}$ **do**
4:     $\mathbf{G} \leftarrow$ Gradient from equation (5.2) *w.r.t.* $\boldsymbol{\pi}^s_{(k-1)}$
5:     $\boldsymbol{\pi}^s_{(k)} \leftarrow OT(\mathbf{w}, \mathbf{w}', \mathbf{G})$
6:     $err \leftarrow ||\boldsymbol{\pi}^s_{(k-1)} - \boldsymbol{\pi}^s_{(k)}||_F$
7:     $k \leftarrow k + 1$
8: **end while**

Finally, by noticing that in the step 3 of Algorithm 12 the gradient of (5.2) *w.r.t* $\boldsymbol{\pi}^s_{(k-1)}$ is $2L(\mathbf{C}, \mathbf{C}') \otimes \boldsymbol{\pi}^s_{(k-1)}$, which gives the same OT solution as for the OT problem in step 3 of Algorithm 10, we can conclude that the iterations of both algorithms are equivalent. □

6.3.7 Additional results – Relation with election isomorphism problem

This section shows that COOT approach can be used to solve the election isomorphism problem defined in [Faliszewski 2019] as follows: let $E = (C, V)$ and $E' = (C', V')$ be two elections, where $C = \{c_1, \ldots, c_m\}$ (resp. C') denotes a set of candidates and $V = (v_1, \ldots, v_n)$ (resp. V') denotes a set of voters, where each voter v_i has a preference order, also denoted by v_i. The two elections $E = (C, V)$ and $E' = (C', V')$, where abs $C = $ abs C', $V = (v_1, \ldots, v_n)$, and $V' = (v'_1, \ldots, v'_n)$, are said to be isomorphic if there exists a bijection $\sigma : C \to C'$ and a permutation $\nu \in S_n$ such that $\sigma(v_i) = v'_{\nu(i)}$ for all $i \in [n]$. The authors further propose a distance underlying this problem defined as follows:

$$\text{d-ID}(E, E') = \min_{\nu \in S_n} \min_{\sigma \in \Pi(C, C')} \sum_{i=1}^{n} d\left(\sigma(v_i), v'_{\nu(i)}\right),$$

where S_n denotes the set of all permutations over $\{1, \ldots, n\}$, $\Pi(C, C')$ is a set of bijections and d is an arbitrary distance between preference orders. The authors of [Faliszewski 2019] compute d-ID(E, E') in practice by expressing it as the following Integer Linear Programming problem over the tensor $\mathbf{P}_{ijkl} = M_{ij} N_{kl}$ where $\mathbf{M} \in \mathbb{R}^{m \times m}$, $\mathbf{N} \in \mathbb{R}^{n \times n}$

$$\min_{\mathbf{P}, \mathbf{N}, \mathbf{M}} \sum_{i,j,k,l} P_{k,l,i,j} |\text{pos}_{v_i}(c_k) - \text{pos}_{v'_j}(c'_l)|$$

$$\text{s.t.} \quad (\mathbf{N}\mathbf{1}_n)_k = 1, \ \forall k, (\mathbf{N}^{\top}\mathbf{1}_n)_l = 1, \ \forall l \tag{6.104}$$

$$(\mathbf{M}\mathbf{1}_m)_i = 1, \ \forall i, (\mathbf{M}^{\top}\mathbf{1}_m)_j = 1, \ \forall j$$

$$P_{kl} \leq N_{k,l}, P_{i,j,k,l} \leq M_{i,j}, \ \forall i, j, k, l$$

$$\sum_{i,k} P_{i,j,k,l} = 1, \ \forall j, l \tag{6.105}$$

where $\text{pos}_{v_i}(c_k)$ denotes the position of candidate c_k in the preference order of voter v_i. Let us now define two matrices $\mathbf{X}$ and $\mathbf{X}'$ such that $\mathbf{X}_{i,k} = \text{pos}_{v_i}(c_k)$ and $\mathbf{X}'_{j,l} = \text{pos}_{v'_j}(c'_l)$ and denote by $\boldsymbol{\pi}^s_*, \boldsymbol{\pi}^v_*$ a

minimizer of $\text{COOT}(\mathbf{X}, \mathbf{X}', \mathbf{1}_n/n, \mathbf{1}_n/n, \mathbf{1}_m/m, \mathbf{1}_m/m)$ with $L = |\cdot|$ and by $\mathbf{N}^*, \mathbf{M}^*$ the minimizers of problem (6.104), respectively.

As shown in the chapter, there exists an optimal solution for $\text{COOT}(\mathbf{X}, \mathbf{X}')$ given by permutation matrices as solutions of the Monge-Kantorovich problems for uniform distributions supported on the same number of elements. Then, one may show that the solution of the two problems coincide modulo a multiplicative factor, *i.e.*, $\boldsymbol{\pi}_*^s = \frac{1}{n}\mathbf{N}^*$ and $\boldsymbol{\pi}_*^v = \frac{1}{m}\mathbf{M}^*$ are optimal since $\text{abs}\, C = \text{abs}\, C'$ and $\text{abs}\, V = \text{abs}\, V'$. For $\boldsymbol{\pi}_*^s$ (the same reasoning holds for $\boldsymbol{\pi}_*^v$ as well), we have that

$$
(\boldsymbol{\pi}_*^s)_{ij} = \left\{ \begin{array}{ll} \frac{1}{n}, & j = \nu_i^* \\ 0, & \text{otherwise.} \end{array} \right.
$$

where ν_i^* is a permutation of voters in the two sets. The only difference between the two solutions $\boldsymbol{\pi}_*^s$ and $\mathbf{N}^*$ thus stems from marginal constraints (6.104). To conclude, we note that COOT is a more general approach as it is applicable for general loss functions L, contrary to the Spearman distance used in [Faliszewski 2019], and generalizes to the cases where $n \neq n'$ and $m \neq m'$.

6.3.8 Additional results – Complementary results for the HDA experiment

Here, we present the results for the heterogeneous domain adaptation experiment not included in section 5.6.1. Table 6.6 follows the same experimental protocol as in the chapter but shows the two cases where $n_t = 1$ and $n_t = 5$. Table 6.7 and Table 6.8 contain the results for the adaptation from GoogleNet to Decaf features, in a semi-supervised and unsupervised scenarios, respectively Overall, the results are coherent with those from the chapter: in both settings, when $n_t = 5$, one can see that the performance differences between SGW and COOT is rather significant.

	Decaf → GoogleNet					
Domains	Baseline	CCA	KCCA	EGW	SGW	COOT
			$n_t = 1$			
C→W	30.47±6.90	13.37±7.23	29.21±13.14	10.21±1.31	<u>66.95</u>±7.61	**77.74**±4.80
W→C	26.53±7.75	16.26±5.18	40.68±12.02	10.11±0.84	<u>80.16</u>±4.78	**87.89**±2.65
W→W	30.63±7.78	13.42±1.38	36.74±8.38	8.68±2.36	<u>78.32</u>±5.86	**89.11**±2.78
W→A	30.21±7.51	12.47±2.99	39.11±6.85	9.42±2.90	<u>80.00</u>±3.24	**89.05**±2.84
A→C	41.89±6.59	12.79±2.95	28.84±6.24	9.89±1.17	<u>72.00</u>±8.91	**84.21**±3.92
A→W	39.84±4.27	19.95±23.40	38.16±19.30	12.32±1.56	<u>75.84</u>±7.37	**89.42**±4.24
A→A	42.68±8.36	15.21±7.36	38.26±16.99	13.63±2.93	<u>75.53</u>±6.25	**91.84**±2.48
C→C	28.58±7.40	18.37±17.81	35.11±17.96	11.05±1.63	<u>61.21</u>±8.43	**78.11**±5.77
C→A	31.63±4.25	15.11±5.10	33.84±9.10	11.84±1.67	<u>66.26</u>±7.95	**82.11**±2.58
Mean	33.61±5.77	15.22±2.44	35.55±3.98	10.80±1.47	<u>72.92</u>±6.37	**85.50**±4.89
			$n_t = 5$			
C→W	74.27±5.53	14.53±7.37	73.27±4.99	11.40±1.13	<u>84.00</u>±3.99	**85.53**±2.67
W→C	90.27±2.67	21.13±6.85	85.00±3.44	10.60±1.05	**95.20**±2.84	<u>94.53</u>±1.83
W→W	90.93±2.50	15.80±3.27	90.67±2.95	9.80±2.60	**95.40**±2.47	<u>94.93</u>±2.70
W→A	90.47±2.92	16.67±4.85	87.93±2.47	9.80±2.68	<u>95.40</u>±1.53	**95.80**±2.15
A→C	<u>88.33</u>±2.33	15.73±4.64	83.13±2.84	10.40±1.89	84.47±5.81	**91.47**±1.45
A→W	<u>88.40</u>±3.17	13.60±6.25	87.27±2.82	11.87±2.40	87.87±4.66	**93.00**±1.96
A→A	86.20±3.08	14.07±2.93	87.00±3.48	14.07±1.65	<u>89.80</u>±2.58	**92.20**±1.69
C→C	75.93±4.83	13.13±2.98	70.47±3.45	11.13±1.52	**85.73**±3.54	<u>84.60</u>±2.32
C→A	73.47±3.62	15.47±6.50	74.13±5.42	11.20±2.47	<u>85.07</u>±3.26	**87.20**±1.78
Mean	84.25±7.01	15.57±2.25	82.10±7.03	11.14±1.23	<u>89.21</u>±4.64	**91.03**±3.97

Table 6.6: **Semi-supervised Heterogeneous Domain Adaptation** results for adaptation from Decaf to GoogleNet representations with different values of n_t.

	GoogleNet → Decaf					
Domains	Baseline	CCA	KCCA	EGW	SGW	COOT
$n_t = 1$						
C→A	31.16±6.87	12.16±2.78	33.32±2.47	7.00±2.11	<u>77.16</u>±8.00	**83.26**±5.00
C→C	30.42±3.73	13.74±5.29	32.58±9.98	12.47±2.81	<u>76.63</u>±8.31	**86.21**±3.26
W→A	37.68±4.04	15.79±3.71	34.58±5.71	14.32±1.77	<u>86.68</u>±1.90	**89.95**±3.43
A→C	35.95±3.89	15.32±8.18	40.16±17.54	13.21±3.49	<u>87.89</u>±4.03	**90.68**±7.54
A→A	36.89±4.73	13.84±2.47	34.84±10.44	13.16±1.56	<u>89.79</u>±3.93	**94.68**±2.21
W→W	32.05±4.63	19.89±11.82	36.26±21.98	10.00±2.59	<u>84.21</u>±4.55	**90.42**±2.66
W→C	32.68±5.56	21.53±21.01	33.79±22.72	11.47±3.03	<u>86.26</u>±3.41	**89.53**±1.92
A→W	33.84±4.75	16.00±7.74	39.32±18.94	11.00±4.01	<u>87.21</u>±3.67	**91.53**±5.85
C→W	32.32±7.76	15.58±7.72	34.05±15.96	12.89±2.52	<u>81.84</u>±3.51	**84.84**±5.71
Mean	33.67±2.45	15.98±2.81	35.43±2.50	11.73±2.08	<u>84.19</u>±4.43	**89.01**±3.38
$n_t = 3$						
C→A	76.35±4.15	17.47±3.45	73.94±4.53	7.41±2.27	<u>88.24</u>±2.23	**89.88**±0.94
C→C	78.94±3.61	18.18±3.44	69.94±3.51	14.18±3.16	<u>89.71</u>±2.25	**91.06**±1.91
W→A	85.41±3.25	19.29±3.10	80.59±3.82	14.24±2.72	<u>94.76</u>±1.45	**95.29**±2.35
A→C	89.53±4.05	23.18±7.17	80.59±6.30	13.88±2.69	<u>93.76</u>±2.72	**94.76**±1.83
A→A	89.76±1.92	17.00±3.11	83.71±3.30	14.41±2.28	<u>93.29</u>±2.09	**95.53**±1.45
W→W	86.65±5.07	21.88±4.78	84.65±3.67	9.94±2.37	**94.88**±1.79	<u>94.53</u>±1.66
W→C	88.94±5.02	22.59±9.23	80.06±5.65	13.65±3.15	**96.18**±1.15	<u>95.29</u>±2.91
A→W	90.29±1.35	22.35±7.00	87.88±2.53	13.88±3.60	<u>94.53</u>±1.54	**95.35**±1.59
C→W	78.59±3.44	22.53±13.42	80.12±2.95	11.59±3.25	<u>89.29</u>±1.86	**89.59**±2.22
Mean	84.94±5.19	20.50±2.34	80.16±5.12	12.58±2.31	<u>92.74</u>±2.72	**93.48**±2.38
$n_t = 5$						
C→A	84.20±2.65	18.60±3.75	84.33±2.33	6.40±1.27	**92.13**±2.61	<u>91.93</u>±2.05
C→C	85.33±2.76	21.80±5.91	78.60±2.74	13.47±2.00	<u>91.33</u>±2.48	**92.27**±2.67
W→A	95.13±2.29	31.00±9.67	91.93±2.82	14.67±1.40	<u>96.13</u>±2.04	**96.40**±1.84
A→C	91.67±2.60	21.80±4.35	85.33±3.27	13.40±3.63	**95.47**±1.51	<u>94.87</u>±1.27
A→A	93.20±1.57	23.33±4.66	89.67±1.98	13.27±2.10	**95.33**±1.07	<u>95.00</u>±1.37
W→W	95.00±2.33	23.80±5.48	92.13±1.78	11.20±2.58	<u>96.47</u>±1.93	**96.67**±1.37
W→C	95.67±1.50	28.27±9.71	87.67±3.79	14.27±3.19	**97.67**±1.31	<u>96.93</u>±2.25
A→W	92.13±2.36	22.67±3.94	89.20±3.14	11.67±2.50	<u>93.60</u>±1.40	**94.27**±2.11
C→W	84.00±3.45	20.40±4.31	82.53±3.56	11.07±3.70	<u>90.20</u>±2.23	**92.40**±1.69
Mean	90.70±4.57	23.52±3.64	86.82±4.26	12.16±2.37	<u>94.26</u>±2.42	**94.53**±1.85

Table 6.7: **Semi-supervised Heterogeneous Domain Adaptation** results for adaptation from GoogleNet to Decaf representations with different values of n_t.

GoogleNet → Decaf				
Domains	CCA	KCCA	EGW	COOT
C→A	11.30±4.04	<u>14.60</u>±8.12	8.20±2.69	**25.10**±11.52
C→C	13.35±4.32	<u>17.75</u>±10.16	11.90±2.99	**37.20**±14.07
W→A	14.55±10.68	<u>25.05</u>±24.73	14.55±2.05	**39.75**±17.29
A→C	13.80±6.51	<u>20.70</u>±17.94	16.00±2.44	**30.25**±18.71
A→A	16.90±10.45	<u>28.95</u>±30.62	12.70±1.79	**41.65**±16.66
W→W	14.50±6.72	<u>24.05</u>±19.35	9.55±1.77	**36.85**±9.20
W→C	13.15±4.98	<u>14.80</u>±8.79	11.40±2.65	**30.95**±17.18
A→W	10.85±4.62	<u>14.40</u>±12.36	12.70±2.99	**40.85**±16.21
C→W	18.25±14.02	<u>25.90</u>±25.40	11.30±3.87	**34.05**±13.82
Mean	14.07±2.25	<u>20.69</u>±5.22	12.03±2.23	**35.18**±5.24

Table 6.8: **Unsupervised Heterogeneous Domain Adaptation** results for adaptation from GoogleNet to Decaf representations.

Conclusion

> Le soleil est noyé. - C'est le soir - dans le port
> Le navire bercé sur ses câbles, s'endort
> — Tristan Corbière, *Les Amours jaunes*

Contents

7.1   Overview of the contributions

This thesis presents a set of optimal transport tools for dealing with probability distributions on *incomparable spaces*, or equivalently probability distributions whose supports do not lie in a common metric space. We explained how this problem occurs *e.g.* when one needs to consider some structural knowledge about the data or when the data come from heterogeneous sources. As a first instance of probability distributions on incomparable spaces we studied the setting of *structured data* such as labeled graphs, times series or any "relational" data whose structure can be modelled through a notion of *cost* or *similarity*. We showed how to describe them as a probability distributions and how to compare them using the so-called Fused Gromov-Wasserstein distance which builds upon Wasserstein and the Gromov-Wasserstein distances. This new optimal transport distance was successfully applied in a graph context where it finds applications for the classification, clustering or summarization of labeled graphs.

As the main building block of FGW, the Gromov-Wasserstein distance is a central notion of this thesis. We attempted to bridge the gap between the understanding of the Wasserstein distance and GW where we consider the special case of Euclidean spaces. This setting allows us to derive a *sliced* approach, based on the first closed-form expression for GW between 1D probability distributions. We called it Sliced Gromov-Wasserstein, *akin* to the Sliced Wasserstein distance which has recently found many applications in machine learning. The Euclidean setting reveals to be also a good starting point for analysing the regularity of GW optimal transport plans and, as such, we partially answered the following question: can we find guarantees on the probability measures so that an optimal transport plan for GW can be expressed through a deterministic function? This question, central in linear optimal transport theory, can be tackled with the celebrated Brenier's theorem in the context of the Wasserstein distance but was still quite under-addressed when dealing with GW.

Although GW is a powerful tool for comparing probability distributions on incomparable spaces it is limited in its ability to find correspondences between the *features* of the samples of these distributions.

This limitation originates from the method *itself* which discards the feature information by focusing only on pair-to-pair distance matrices. To circumvent this constraint we proposed a novel optimal transport distance which both finds the correspondences between the samples *and* the features of the distributions. This work is based on the CO-Optimal transport framework which computes two optimal transport plans directly on the raw data unlike *GW* which requires pre-computed pair-to-pair distance or similarity matrices. We showed that it is particularly suited for problems such as Heterogeneous Domain Adaptation and Co-clustering. In the light of the previous results we drew interesting connections between COOT and the *GW* distance the latter being a special case of the former in the *concave regime* and with data described by distance or similarity matrices.

Preprints & papers published during this thesis

- [Vayer 2020b] **T. Vayer**, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. *Fused Gromov-Wasserstein distance for structured objects.* Journal. In: Algorithms. (2020)

- [Vayer 2019a] **T. Vayer**, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. *Optimal Transport for structured data with application on graphs.* International conference. In: International Conference on Machine Learning (ICML). (2019)

- [Vayer 2019b] **T. Vayer**, R. Flamary, R. Tavenard, L. Chapel and N. Courty. *Sliced Gromov-Wasserstein.* International conference. In: Advances in Neural Information Processing Systems (NeurIPS). (2019)

- **T. Vayer**, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. *Fused Gromov Wasserstein distance.* National conference. In: Conférence sur l'Apprentissage automatique (CAp). (2018)

- **T. Vayer**, L. Chapel, R. Flamary, R. Tavenard, and N. Courty. *Transport Optimal pour les Signaux sur Graphes.* National conference. In: GRETSI. (2019)

- [Redko 2020] I. Redko, **T. Vayer**, R. Flamary and N. Courty. *CO-Optimal Transport.* In: arXiv:2002.03731. Submitted to NeurIPS (2020)

- [Vayer 2020a] **T. Vayer**, L. Chapel, N. Courty, R. Flamary, Y. Soullard and R. Tavenard. *Time Series Alignment with Global Invariances* . In: arXiv:2002.03848. (2020)

7.2   Perspective for further works

There are many possible extensions and improvements of the works developed here. To conclude we discuss potential limitations and further works of the methods proposed in this thesis from the OT perspective to the machine learning point of view.

What are the improvements and limitations of the different results of this thesis from the OT perspective?   The work about *GW* on Euclidean spaces suggests, in our opinion, interesting further works. The slicing approach for *GW* finds *e.g.* natural extensions inspired by the Wasserstein distance. An immediate one is the max-sliced approach [Deshpande 2019] where only the projection which maximizes the loss is drawn. This setting is directly applicable in our case without changing the theoretical properties of *SGW* stated in Theorem 4.1.3. Moreover, one limitation of *SGW* in the formulation that we proposed is the map Δ which has to be chosen or optimized when one wants to compute *SGW* when dimensions

differ or in order to retrieve the invariants of GW. The max-sliced approach could be a interesting remedy for this situation. We can for example draw one line in each space without requiring any map Δ but with the price of losing the divergence property (*i.e.* $max - SGW(\mu, \mu)$ could be different from zero). Another potential direction would be to consider many lines in each space and then find their correspondences by using *e.g.* Wasserstein which could define a proper metric *w.r.t.* isomorphisms but which will come with a cubic complexity. Interestingly enough this approach would exactly be the COOT distance but with replacing the matching of the feature by a sliced Wasserstein. In any case there are room for improvements in the way of defining SGW when dimensions of the support of the probability measures differ.

Regarding the theoretical aspects of SGW another interesting line of research, in our opinion, would be to draw connections with the statistical properties of the Sliced Wasserstein distance which is known to behave well in terms of convergence of finite samples. We believe that similar type of studies for the Sliced Gromov-Wasserstein could be promising.

From a computational OT point of view the CO-Optimal Transport framework seems also to be quite suited for large scale datasets. Indeed since the BCD procedure used for solving COOT is based on alternating linear optimal transport problems another line of works could be to rely on the dual or semi dual formulations of linear OT, used for large scale OT, in order to compute COOT for large scale datasets whose samples are in incomparable spaces. Finally from a theoretical point of view a continuous formulation of COOT could also be of interest for studying the properties of GW, and especially the regularity of its optimal plans. Indeed we have proven that an optimal solution for COOT can be found *via* permutation matrices, whose continuous counterparts are deterministic push-forwards. As such, can we find a natural extension for COOT which preserves this property in the continuous setting? Also if the problem is concave, such as when GW is considered with squared Euclidean distances, can we preserve the result stating that COOT and GW are equivalent so that an optimal solution of the first is optimal solution for the second? In this way can we conclude that there exists an optimal map for GW that is supported on a Monge map when the problem is concave? We believe, enthusiastically, that these questions worth investigating.

How other fields can contribute to the frameworks developed in this thesis? One of the most challenging improvement of the work developed in this thesis is maybe to lighten the computational complexity of calculating FGW which is driven by the complexity of the GW distance. In this form, the computational complexity FGW renders this framework limited to small to medium scale *scenarii* and, in particular, is inapplicable for very large graphs. Moreover one can not hope to directly rely on the works about GW on Euclidean spaces in order to derive tractable formulations since the structures of labeled graphs are usually highly non-Euclidean. However it worth pointing out that approximating $\mathbf{C}_1, \mathbf{C}_2$ with Euclidean matrices can be done in several ways [Borg 2005, Glunt 1990, Alfakih 1999, Liberti 2014]. We believe that a kind of *no free lunch theorem* applies here: if we faithfully model the structure of the data it is likely to result in a more precise notion of distance between structured data but also in a more challenging optimization problem than if we approximate it for better scalability. Needless to say, since GW is at the backbone of our method, any further improvements in computational efficiency for non convex Quadratic programs, and by way of consequence for GW, would directly benefit to FGW. In this way progress in graph matching could definitely contribute to derive efficient algorithms for solving FGW as both are intrinsically related. Another interesting study would be regarding the design of the matrices $\mathbf{C}_1, \mathbf{C}_2$ representative of the structures for FGW. We considered in this thesis that these matrices are given, such as shortest-path matrices, yet, one could argue that this is a strong prior on the structures of the graphs and that no *perfect* choice arises in practice. The problem of finding "good" structure matrices

is very related with the metric learning field [Bellet 2013] and further appealing works for FGW could be to go into this direction and to add $\mathbf{C}_1, \mathbf{C}_2$ into the learning process. Related to this problem we believe that very insightful connections between FGW and the field of signal graph processing [Ortega 2018] can be drawn and that both can benefit from each other. As one example: can we build upon the spectral tools from signal graph processing in order to find adequate measures of structure $\mathbf{C}_1, \mathbf{C}_2$ of the graphs?

How do the above-mentioned optimal transport tools can, or can not, be used *for* machine learning? As mentioned throughout this thesis, having a adequate measure for comparing data lying on incomparable spaces can be useful for a wide range of machine learning tasks. For applications involving structured data we proposed the FGW distance based on optimal transport. Note, however, that many ways of comparing structured data have been proposed in the literature: from the design of dedicated kernels (see [Kriege 2020] for a comprehensive survey on this topic) to other choices of distances between graphs [Bento 2019, Wills 2020] including the popular graph-edit distance [Willett 1998, Raymond 2002]. More recently, end-to-end approaches [Li 2019, Bai 2020, Riba 2018, Sun 2020] attempt to *learn* a similarity measure function between graphs based on graph neural networks, *i.e.* to learn a neural network based function that takes two graphs as input and outputs the desired similarity. The question of finding a "good" similarity measure for structured data is far from being closed and is naturally dependent of the application. The choice of FGW can be motivated by its metric properties which allow detecting whether two graphs are isomorphic and, as such, could be used as an alternative to the graph-edit distance. However, despite its appealing properties, one could question the optimal transport framework on which this method is based. As described in this thesis the coupling matrices are used to find a probabilistic matching between *all* the nodes of two graphs. Conversely, and depending on the application, one might want to match only a *small* portion of the nodes of two graphs which is not possible using the FGW framework which considers *global* matchings. In this way, in a scenario where only the *local* structure is important, the FGW machinery appears to be quite disproportionate, and some kernel methods which are built on local structures may be more suited. The use of the FGW distance for structured data such as time series is also questionable. Indeed the set of couplings *itself* is constrained so that a matching "back in time" is made possible. On the contrary, it may be more interesting to force that the points which are matched to the target sequence at time t can only depend on the source sequence up to time t as done *e.g.* using Dynamic Time Warping distances [Cuturi 2017, Sakoe 1978]. To remedy to this problem the recent casual optimal transport framework [Lassalle 2018, Veraguas 2016, Acciaio 2020, Zalashko 2017] may be more suited for these type of problems. Apart from these limitations the FGW framework could be useful for other applications than these mentioned in this thesis. As a first instance it could be used as a way for learning graph neural networks (GNN). Given a set of graphs we could for example minimize a triplet loss [Chechik 2010] or a graph similarity score [Bai 2019] which impose that the distances in the embedding of the GNN are close to the FGW distances so as to force the GNN to produce similar embeddings for similar graphs. Moreover the properties of the Fréchet barycenter with FGW could be elaborated, especially the projection on a smaller graph using FGW. A perspective on this topic would be to draw connections with graph signal processing and more standard coarsening procedures (see [Loukas 2019] and references therein) by questioning the ability of the FGW projection to reduce the graph without altering too much its spectral properties.

Finally the results of this thesis suggests that COOT may be suited for Heterogeneous Domain Adaptation. An interesting further work on this topic would be to see if we can confirm these empirical results with theoretical guarantees such as bounds for Heterogeneous Domain Adaptation by relying on the COOT framework. For example can we prove that the adaptation is easier when the datasets are

close from each other *w.r.t.* the COOT distance than if they are far from each other? To the best of our knowledge the question of finding bounds for HDA problems is still quite under-addressed [Zhou 2019] and COOT may lead to promising further works on this topic.

All these considered we hope that the works proposed in this thesis will pave the path for positive and interesting studies on various and interdisciplinary topics in machine learning and that it will also contribute to the richness of the optimal transport theory, which is far from being extinguished.

Bibliography

[Abrudan 2008] T. E. Abrudan, J. Eriksson et V. Koivunen. *Steepest Descent Algorithms for Optimization Under Unitary Matrix Constraint*. IEEE Transactions on Signal Processing, vol. 56, no. 3, pages 1134–1147, 2008. (Cited on page 97.)

[Absil 2009] P-A Absil, Robert Mahony et Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. Princeton University Press, 2009. (Cited on pages 74, 75 and 97.)

[Acciaio 2020] Beatrice Acciaio, Julio Backhoff-Veraguas et Junchao Jia. *Cournot-Nash equilibrium and optimal transport in a dynamic setting*, 2020. (Cited on page 162.)

[Agueh 2011] Martial Agueh et Guillaume Carlier. *Barycenters in the Wasserstein space*. SIAM Journal on Mathematical Analysis, vol. 43, no. 2, pages 904–924, 2011. (Cited on page 25.)

[Alaya 2019] Mokhtar Z. Alaya, Maxime Berar, Gilles Gasso et Alain Rakotomamonjy. *Screening Sinkhorn Algorithm for Regularized Optimal Transport*. In Advances in Neural Information Processing Systems 32, pages 12169–12179. Curran Associates, Inc., 2019. (Cited on page 106.)

[Alfakih 1999] Abdo Y. Alfakih, Amir Khandani et Henry Wolkowicz. *Solving Euclidean Distance Matrix Completion Problems Via Semidefinite Programming*. Computational Optimization and Applications, vol. 12, no. 1, pages 13–30, Janvier 1999. (Cited on page 161.)

[Altschuler 2017] Jason Altschuler, Jonathan Weed et Philippe Rigollet. *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*. In Advances in Neural Information Processing Systems, pages 1961–1971, 2017. (Cited on pages 3, 1, 22, 103 and 106.)

[Altschuler 2019] Jason Altschuler, Francis Bach, Alessandro Rudi et Jonathan Niles-Weed. *Massively scalable Sinkhorn distances via the Nyström method*. In Advances in Neural Information Processing Systems 32, pages 4427–4437. Curran Associates, Inc., 2019. (Cited on page 106.)

[Alvarez-Melis 2018a] David Alvarez-Melis et Tommi Jaakkola. *Gromov-Wasserstein Alignment of Word Embedding Spaces*. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. (Cited on pages 39 and 102.)

[Alvarez-Melis 2018b] David Alvarez-Melis, Tommi S. Jaakkola et Stefanie Jegelka. *Structured Optimal Transport*. In AISTATS, 2018. (Cited on pages 5, 3 and 43.)

[Alvarez-Melis 2019] David Alvarez-Melis, Stefanie Jegelka et Tommi S. Jaakkola. *Towards Optimal Transport with Global Invariances*. In AISTATS, volume 89, pages 1870–1879, 2019. (Cited on pages 33, 82, 102 and 108.)

[Alvarez-Melis 2020] David Alvarez-Melis et Nicolò Fusi. *Geometric Dataset Distances via Optimal Transport*, 2020. (Cited on page 105.)

[Ambrosio 2005] L. Ambrosio, N. Gigli et G. Savare. Gradient flows: In metric spaces and in the space of probability measures. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2005. (Cited on pages 28, 33, 122 and 125.)

[Andoni 2015] Alexandr Andoni, Assaf Naor et Ofer Neiman. *Snowflake universality of Wasserstein spaces*. Annales Scientifiques de l'Ecole Normale Superieure, vol. 51, 09 2015. (Cited on page 12.)

[Anstreicher 1998] Kurt Anstreicher et Henry Wolkowicz. *On Lagrangian Relaxation of Quadratic Matrix Constraints*. SIAM Journal on Matrix Analysis and Applications, vol. 22, 07 1998. (Cited on pages 88 and 144.)

[Arjovsky 2017] M. Arjovsky, S. Chintala et L. Bottou. *Wasserstein Generative Adversarial Networks*. In International Conference on Machine Learning, volume 70, pages 214–223, 2017. (Cited on pages 3, 1, 15, 24, 43 and 70.)

[Bach 2007] F. Bach et Z. Harchaoui. *Image Classification with Segmentation Graph Kernels*. In CVPR, volume 00, pages 1–8, 06 2007. (Cited on pages 5, 3 and 42.)

[Bai 2019] Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun et Wei Wang. *SimGNN: A Neural Network Approach to Fast Graph Similarity Computation*. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, page 384–392, New York, NY, USA, 2019. Association for Computing Machinery. (Cited on page 162.)

[Bai 2020] Yunsheng Bai, Hao Ding, Ken Gu, Yizhou Sun et Wei Wang. *Learning-Based Efficient Graph Similarity Computation via Multi-Scale Convolutional Set Matching*. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pages 3219–3226, 04 2020. (Cited on page 162.)

[Bakir 2007] Gükhan H. Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar et S. V. N. Vishwanathan. Predicting structured data (neural information processing). The MIT Press, 2007. (Cited on page 42.)

[Banerjee 2007] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu et Dharmendra S. Modha. *A Generalized Maximum Entropy Approach to Bregman Co-clustering and Matrix Approximation*. Journal of Machine Learning Research, vol. 8, pages 1919–1986, 2007. (Cited on page 115.)

[Barbe 2020] Amélie Barbe, Marc Sebban, Paulo Gonçalves, Pierre Borgnat et Rémi Gribonval. *Graph Diffusion Wasserstein Distances*. In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Ghent, Belgium, Septembre 2020. (Cited on page 57.)

[Battaglia 2018] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li et R. Pascanu. *Relational inductive biases, deep learning, and graph networks*. ArXiv e-prints, Juin 2018. (Cited on pages 4, 2, 3 and 42.)

[Bellet 2013] Aurélien Bellet, Amaury Habrard et Marc Sebban. *A Survey on Metric Learning for Feature Vectors and Structured Data*, 2013. (Cited on page 162.)

[Benamou 2000] Jean-David Benamou et Yann Brenier. *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem.* Numerische Mathematik, vol. 84, no. 3, pages 375–393, Jan 2000. (Cited on page 65.)

[Benamou 2015] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna et Gabriel Peyré. *Iterative bregman projections for regularized transportation problems.* SIAM Journal on Scientific Computing, vol. 37, no. 2, pages A1111–A1138, 2015. (Cited on pages 22, 23 and 26.)

[Bento 2019] José Bento et Stratis Ioannidis. *A family of tractable graph metrics.* Applied Network Science, vol. 4, no. 1, page 107, Novembre 2019. (Cited on page 162.)

[Berg 2005] Alexander Berg, Tamara Berg et Jitendra Malik. *Shape Matching and Object Recognition Using Low Distortion Correspondences.* volume 1, pages 26–33, 01 2005. (Cited on page 35.)

[Bernard 2018] Florian Bernard, Christian Theobalt et Michael Moeller. *DS*: Tighter Lifting-Free Convex Relaxations for Quadratic Matching Problems.* pages 4310–4319, 06 2018. (Cited on page 35.)

[Bertsimas 1997] D. Bertsimas et J.N. Tsitsiklis. Introduction to linear optimization. Athena Scientific, 1997. (Cited on page 20.)

[Besl 1992] Paul J. Besl et Neil D. McKay. *A Method for Registration of 3-D Shapes.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, pages 239–256, 1992. (Cited on page 102.)

[Bigot 2019a] Jérémie Bigot, Elsa Cazelles et Nicolas Papadakis. *Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration.* Information and Inference, vol. 8, no. 4, pages 719–755, 2019. (Cited on page 27.)

[Bigot 2019b] Jérémie Bigot, Elsa Cazelles et Nicolas Papadakis. *Penalization of Barycenters in the Wasserstein Space.* SIAM Journal on Mathematical Analysis, vol. 51, no. 3, pages 2261–2285, 2019. (Cited on pages 23 and 27.)

[Billingsley 1999] Patrick Billingsley. Convergence of probability measures. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second édition, 1999. A Wiley-Interscience Publication. (Cited on page 28.)

[Birkhoff 1946] Garrett Birkhoff. *Tres observaciones sobre el algebra lineal.* Univ. Nac. Tucumán Rev. Ser. A, 1946. (Cited on pages 20, 105 and 150.)

[Blondel 2020] Mathieu Blondel, André F.T. Martins et Vlad Niculae. *Learning with Fenchel-Young losses.* Journal of Machine Learning Research, vol. 21, no. 35, pages 1–69, 2020. (Cited on pages 5 and 3.)

[Bokhari 1981] S. H. Bokhari. *On the Mapping Problem.* IEEE Trans. Comput., vol. 30, no. 3, page 207–214, Mars 1981. (Cited on page 35.)

[Bonneel 2011] Nicolas Bonneel, Michiel van de Panne, Sylvain Paris et Wolfgang Heidrich. *Displacement Interpolation Using Lagrangian Mass Transport.* In Proceedings of the 2011 SIGGRAPH Asia Conference, SA '11, pages 158:1–158:12, New York, NY, USA, 2011. ACM. (Cited on page 65.)

[Bonneel 2015] Nicolas Bonneel, Julien Rabin, Gabriel Peyré et Hanspeter Pfister. *Sliced and Radon Wasserstein Barycenters of Measures.* Journal of Mathematical Imaging and Vision, vol. 1, no. 51, pages 22–45, 2015. (Cited on pages 25, 27, 71 and 73.)

[Bonneel 2019] Nicolas Bonneel et David Coeurjolly. *SPOT: Sliced Partial Optimal Transport*. ACM Trans. Graph., vol. 38, no. 4, Juillet 2019. (Cited on page 25.)

[Bonnotte 2013] Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, 2013. (Cited on pages 24 and 71.)

[Borg 2005] I. Borg et P.J.F. Groenen. Modern Multidimensional Scaling: Theory and Applications. Springer, 2005. (Cited on page 161.)

[Borgwardt 2005] Karsten M. Borgwardt et Hans-Peter Kriegel. *Shortest-Path Kernels on Graphs*. In ICDM, ICDM '05, pages 74–81, Washington, DC, USA, 2005. IEEE Computer Society. (Cited on pages 51 and 52.)

[Bourgain 1986] Jean Bourgain. *The metrical interpretation of superreflexivity in banach spaces*. Israel Journal of Mathematics, vol. 56, pages 222–230, 1986. (Cited on page 12.)

[Brenier 1991] Yann Brenier. *Polar factorization and monotone rearrangement of vector-valued functions*. Communications on Pure and Applied Mathematics, vol. 44, no. 4, pages 375–417, 1991. (Cited on pages 9, 8 and 16.)

[Bronstein 2017] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam et Pierre Vandergheynst. *Geometric deep learning: going beyond euclidean data*. IEEE Signal Processing Magazine, vol. 34, no. 4, pages 18–42, 2017. (Cited on page 42.)

[Bunne 2019] Charlotte Bunne, David Alvarez-Melis, Andreas Krause et Stefanie Jegelka. *Learning Generative Models across Incomparable Spaces*. In International Conference on Machine Learning, volume 97, 2019. (Cited on pages 39, 70, 76, 77 and 106.)

[Burago 2001] D. Burago, I.D. Burago, I.U.D. Burago, I.U.D. Burago, J.D. Burago, Y. Burago, Û.D. Burago, D. Burago, I.U.D. Burago, S. Ivanov et others. A Course in Metric Geometry. Crm Proceedings & Lecture Notes. American Mathematical Society, 2001. (Cited on page 33.)

[Bures 1969] Donald Bures. *An Extension of Kakutani's Theorem on Infinite Product Measures to the Tensor Product of Semifinite w*-Algebras*. Transactions of the American Mathematical Society, vol. 135, pages 199–212, 1969. (Cited on page 18.)

[Burkard 1996] Rainer E. Burkard, Bettina Klinz et Rüdiger Rudolf. *Perspectives of Monge Properties in Optimization*. Discrete Appl. Math., vol. 70, no. 2, page 95–161, Septembre 1996. (Cited on page 20.)

[Burkard 1999] Rainer E. Burkard et Eranda Çela. Linear assignment problems and extensions, pages 75–149. Springer US, Boston, MA, 1999. (Cited on page 19.)

[Béthune 2020] Louis Béthune, Yacouba Kaloga, Pierre Borgnat, Aurélien Garivier et Amaury Habrard. *Hierarchical and Unsupervised Graph Representation Learning with Loukas's Coarsening*, 2020. (Cited on page 57.)

[Caetano 2009] T. S. Caetano, J. J. McAuley, L. Cheng, Q. V. Le et A. J. Smola. *Learning Graph Matching*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 6, pages 1048–1058, 2009. (Cited on page 35.)

[Cang 2020] Zixuan Cang et Qing Nie. *Inferring spatial and signaling relationships between cells from single cell transcriptomic data.* Nature Communications, vol. 11, no. 1, page 2084, Avril 2020. (Cited on page 57.)

[Carlier 2015] Guillaume Carlier, Adam Oberman et Edouard Oudet. *Numerical methods for matching for teams and Wasserstein barycenters.* ESAIM: Mathematical Modelling and Numerical Analysis, vol. 49, no. 6, pages 1621–1642, Novembre 2015. (Cited on page 26.)

[Carrière 2017] Mathieu Carrière, Marco Cuturi et Steve Oudot. *Sliced Wasserstein Kernel for Persistence Diagrams.* In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 664–673. JMLR.org, 2017. (Cited on page 25.)

[Çela 2011] Eranda Çela, Nina S. Schmuck, Shmuel Wimer et Gerhard J. Woeginger. *The Wiener maximum quadratic assignment problem.* Discrete Optimization, vol. 8, pages 411–416, 2011. (Cited on page 35.)

[Çela 2018] Eranda Çela, Vladimir Deineko et Gerhard J. Woeginger. *New special cases of the Quadratic Assignment Problem with diagonally structured coefficient matrices.* European journal of operational research, vol. 267, no. 3, pages 818–834, 2018. (Cited on page 35.)

[Çela 2013] Eranda Çela. The quadratic assignment problem: Theory and algorithms, volume 1. Springer Science & Business Media, 2013. (Cited on page 35.)

[Çela 2015] Eranda Çela, Vladimir G. Deineko et Gerhard J. Woeginger. *Well-solvable cases of the QAP with block-structured matrices.* Discrete applied mathematics, vol. 186, pages 56–65, 2015. (Cited on page 35.)

[Charlier 2018] Benjamin Charlier, Jean Feydy et Joan Glaunes. *Kernel Operations on the GPU, with autodiff, without memory overflows.* https://github.com/getkeops/keops, 2018. (Cited on page 75.)

[Chechik 2010] Gal Chechik, Varun Sharma, Uri Shalit et Samy Bengio. *Large Scale Online Learning of Image Similarity Through Ranking.* Journal of Machine Learning Research, vol. 11, pages 1109–1135, 2010. (Cited on page 162.)

[Chen 2018] Xinlei Chen, Li-Jia Li, Li Fei-Fei et Abhinav Gupta. *Iterative Visual Reasoning Beyond Convolutions.* In CVPR, pages 7239–7248. IEEE Computer Society, 2018. (Cited on pages 5 and 3.)

[Chizat 2017] Lenaïc Chizat. *Unbalanced Optimal Transport : Models, Numerical Methods, Applications.* Thesiss, PSL Research University, Novembre 2017. (Cited on pages 25, 27 and 106.)

[Chizat 2018] Lénaïc Chizat et Francis Bach. *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport.* In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi et R. Garnett, editeurs, Advances in Neural Information Processing Systems 31, pages 3036–3046. Curran Associates, Inc., 2018. (Cited on page 65.)

[Chowdhury 2019a] Samir Chowdhury et Facundo Mémoli. *The Gromov–Wasserstein distance between networks and stable network invariants.* Information and Inference: A Journal of the IMA, vol. 8, no. 4, pages 757–787, 2019. (Cited on pages 27, 28, 32 and 38.)

[Chowdhury 2019b] Samir Chowdhury et Tom Needham. *Gromov-Wasserstein Averaging in a Riemannian Framework*, 2019. (Cited on page 38.)

[Cohen 2016] Taco Cohen et Max Welling. *Group Equivariant Convolutional Networks.* In Maria Florina Balcan et Kilian Q. Weinberger, editeurs, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on pages 5 and 3.)

[Collins 2002] Michael Collins. *Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms.* In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, page 1–8, USA, 2002. Association for Computational Linguistics. (Cited on pages 5 and 3.)

[Courty 2017] Nicolas Courty, Rémi Flamary, Devis Tuia et Alain Rakotomamonjy. *Optimal transport for domain adaptation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 9, pages 1853–1865, 2017. (Cited on pages 3, 5, 1, 24, 43, 70, 102 and 110.)

[Cramér 1936] H. Cramér et H. Wold. *Some Theorems on Distribution Functions.* Journal of the London Mathematical Society, vol. s1-11, no. 4, pages 290–294, 1936. (Cited on page 134.)

[Csiszar 1975] I. Csiszar. *I-Divergence Geometry of Probability Distributions and Minimization Problems.* The Annals of Probability, vol. 3, no. 1, pages 146–158, 1975. (Cited on pages 6 and 4.)

[Cui 2014] Zhen Cui, Hong Chang, Shiguang Shan et Xilin Chen. *Generalized Unsupervised Manifold Alignment.* In NIPS, pages 2429–2437. 2014. (Cited on page 102.)

[Custic 2016] Ante Custic, Vladyslav Sokol, Abraham Punnen et Binay Bhattacharya. *The Bilinear Assignment Problem: Complexity and polynomially solvable special cases.* Mathematical Programming, vol. 166, 2016. (Cited on page 105.)

[Cuturi 2013] Marco Cuturi. *Sinkhorn distances: Lightspeed computation of optimal transport.* In NIPS, pages 2292–2300, 2013. (Cited on pages 3, 1, 21, 103 and 106.)

[Cuturi 2014] Marco Cuturi et Arnaud Doucet. *Fast Computation of Wasserstein Barycenters.* In Eric P. Xing et Tony Jebara, editeurs, Proceedings of the 31st International Conference on Machine Learning, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR. (Cited on pages 26 and 49.)

[Cuturi 2017] Marco Cuturi et Mathieu Blondel. *Soft-DTW: a Differentiable Loss Function for Time-Series.* In Proceedings of the ICML, volume 70, pages 894–903. PMLR, 06–11 Aug 2017. (Cited on pages 42 and 162.)

[Cuturi 2018] Marco Cuturi et Gabriel Peyré. *Semi-dual Regularized Optimal Transport.* SIAM Review, vol. 60, pages 941–965, 2018. (Cited on page 26.)

[Dantzig 1997] George B. Dantzig et Mukund N. Thapa. Linear programming 1: Introduction. Springer-Verlag, Berlin, Heidelberg, 1997. (Cited on pages 19 and 20.)

[Day 1985] William HE Day. *Optimal algorithms for comparing trees with labeled leaves.* Journal of classification, vol. 2, no. 1, pages 7–28, 1985. (Cited on pages 4, 3 and 42.)

[De Condorcet 1781] Nicolas De Condorcet. *Sur les déblais et les remblais*, 1781. (Cited on pages 3 and 1.)

[Debnath 1991] Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman et Corwin Hansch. *Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity.* Journal of Medicinal Chemistry, vol. 34, no. 2, pages 786–797, 1991. (Cited on page 51.)

[Defferrard 2016] Michaël Defferrard, Xavier Bresson et Pierre Vandergheynst. *Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering.* In NIPS, pages 3844–3852. 2016. (Cited on page 42.)

[Demetci 2020] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble et Ritambhara Singh. *Gromov-Wasserstein optimal transport to align single-cell multi-omics data.* bioRxiv, 2020. (Cited on page 39.)

[Deshpande 2018] Ishan Deshpande, Ziyu Zhang et Alexander G. Schwing. *Generative Modeling Using the Sliced Wasserstein Distance.* In IEEE Conference on Computer Vision and Pattern Recognition, pages 3483–3491, 2018. (Cited on pages 24, 71 and 77.)

[Deshpande 2019] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth et Alexander G. Schwing. *Max-Sliced Wasserstein Distance and Its Use for GANs.* In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. (Cited on pages 24, 74 and 160.)

[Dhillon 2003] Inderjit S. Dhillon, Subramanyam Mallela et Dharmendra S. Modha. *Information-theoretic Co-clustering.* In SIGKDD, pages 89–98, 2003. (Cited on page 113.)

[Ding 2006] C. Ding, T. Li, W. Peng et H. Park. *Orthogonal Nonnegative Matrix Tri-factorizations for Clustering.* In Proceedings ACM SIGKDD, pages 126–135, 2006. (Cited on page 113.)

[Donahue 2014] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng et T. Darrell. *DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.* In ICML, 2014. (Cited on page 111.)

[Dudley 1969] R. M. Dudley. *The Speed of Mean Glivenko-Cantelli Convergence.* Ann. Math. Statist., vol. 40, no. 1, pages 40–50, 02 1969. (Cited on page 18.)

[Dym 2017] Nadav Dym, Haggai Maron et Yaron Lipman. *DS++: A Flexible, Scalable and Provably Tight Relaxation for Matching Problems.* ACM Trans. Graph., vol. 36, no. 6, Novembre 2017. (Cited on page 35.)

[Dvzeroski 2001] Savso Dvzeroski, Luc De Raedt et Kurt Driessens. *Relational Reinforcement Learning.* Machine Learning, vol. 43, no. 1, pages 7–52, Apr 2001. (Cited on pages 5 and 3.)

[Elshafei 1977] Alwalid N. Elshafei. *Hospital Layout as a Quadratic Assignment Problem.* Operational Research Quarterly (1970-1977), vol. 28, no. 1, pages 167–179, 1977. (Cited on page 35.)

[Ezuz 2017] Danielle Ezuz, Justin Solomon, Vladimir G. Kim et Mirela Ben-Chen. *GWCNN: A Metric Alignment Layer for Deep Shape Analysis.* Computer Graphics Forum, vol. 36, no. 5, pages 49–57, 2017. (Cited on pages 39, 70 and 107.)

[Faliszewski 2019] P. Faliszewski, P. Skowron, A. Slinko, S. Szufa et N. Talmon. *How Similar Are Two Elections?* In AAAI, pages 1909–1916, 2019. (Cited on pages 106, 109, 110, 154 and 155.)

[Fatras 2020] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval et Nicolas Courty. *Learning with minibatch Wasserstein : asymptotic and gradient properties*. In Silvia Chiappa et Roberto Calandra, editeurs, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, pages 2131–2141, Online, 26–28 Aug 2020. PMLR. (Cited on pages 19 and 24.)

[Feragen 2013] Aasa Feragen, Niklas Kasenburg, Jens Petersen, Marleen de Bruijne et Karsten Borgwardt. *Scalable kernels for graphs with continuous attributes*. In Advances in Neural Information Processing Systems 26, pages 216–224. 2013. (Cited on pages 51 and 52.)

[Ferradans 2014] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré et Jean-François Aujol. *Regularized discrete optimal transport*. SIAM Journal on Imaging Sciences, vol. 7, no. 3, pages 1853–1882, 2014. (Cited on pages 25 and 47.)

[Fey 2020] Matthias Fey, Jan E. Lenssen, Christopher Morris, Jonathan Masci et Nils M. Kriege. *Deep Graph Matching Consensus*. In International Conference on Learning Representations, 2020. (Cited on page 39.)

[Feydy 2019] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve et Gabriel Peyré. *Interpolating between Optimal Transport and MMD using Sinkhorn Divergences*. In Kamalika Chaudhuri et Masashi Sugiyama, editeurs, Proceedings of Machine Learning Research, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 16–18 Apr 2019. (Cited on page 23.)

[Figalli 2010] Alessio Figalli. *Regularity of optimal transport maps [after Ma-Trudinger-Wang and Loeper]*. In Séminaire Bourbaki : volume 2008/2009 exposés 997-1011 - Avec table par noms d'auteurs de 1848/49 à 2008/09, numéro 332 de Astérisque, pages 341–368. Société mathématique de France, 2010. talk:1009. (Cited on page 16.)

[Flamary 2014] R. Flamary, N.. Courty, D. Tuia et A. Rakotomamonjy. *Optimal transport with Laplacian regularization: Applications to domain adaptation and shape matching*. 2014. (Cited on page 47.)

[Flamary 2017] R'emi Flamary et Nicolas Courty. *POT Python Optimal Transport library*, 2017. (Cited on pages 49 and 76.)

[Flamary 2019] Rémi Flamary, Karim Lounici et André Ferrari. *Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation*, 2019. (Cited on pages 18 and 145.)

[Frisch 2002] Uriel Frisch, Sabino Matarrese, Roya Mohayaee et Andrei Sobolevski. *A reconstruction of the initial conditions of the Universe by optimal mass transportation*. Nature, vol. 417, pages 260–2, 06 2002. (Cited on pages 3 and 1.)

[Frogner 2015] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya et Tomaso A Poggio. *Learning with a Wasserstein Loss*. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama et R. Garnett, editeurs, Advances in Neural Information Processing Systems 28, pages 2053–2061. Curran Associates, Inc., 2015. (Cited on pages 3 and 1.)

[Frogner 2019] Charlie Frogner, Farzaneh Mirzazadeh et Justin Solomon. *Learning Embeddings into Entropic Wasserstein Spaces*. ICLR, 2019. (Cited on page 12.)

[Gallo 1977] Giorgio Gallo et Aydin Ülkücü. *Bilinear Programming: An Exact Algorithm.* Mathematical Programming, vol. 12, pages 173–194, 1977. (Cited on page 105.)

[Gangbo 1996] Wilfrid Gangbo et Robert J. McCann. *The geometry of optimal transportation.* Acta Math., vol. 177, no. 2, pages 113–161, 1996. (Cited on page 16.)

[Gärtner 2003] Thomas Gärtner, Peter Flach et Stefan Wrobel. *On graph kernels: Hardness results and efficient alternatives.* In IN: CONFERENCE ON LEARNING THEORY, pages 129–143, 2003. (Cited on page 52.)

[Gelfand 2005] Natasha Gelfand, Niloy J. Mitra, Leonidas J. Guibas et Helmut Pottmann. *Robust Global Registration.* In Mathieu Desbrun et Helmut Pottmann, editeurs, Eurographics Symposium on Geometry Processing 2005. The Eurographics Association, 2005. (Cited on page 37.)

[Genevay 2016] Aude Genevay, Marco Cuturi, Gabriel Peyré et Francis Bach. *Stochastic Optimization for Large-scale Optimal Transport.* In Advances in Neural Information Processing Systems, pages 3440–3448, 2016. (Cited on pages 3, 1 and 23.)

[Genevay 2018] Aude Genevay, Gabriel Peyre et Marco Cuturi. *Learning Generative Models with Sinkhorn Divergences.* In Amos Storkey et Fernando Perez-Cruz, editeurs, Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. (Cited on pages 3, 1 and 23.)

[Genevay 2019] Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi et Gabriel Peyré. *Sample Complexity of Sinkhorn Divergences.* In ICML, pages 1574–1583, 2019. (Cited on pages 23 and 103.)

[Gens 2014] Robert Gens et Pedro M Domingos. *Deep Symmetry Networks.* In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence et K. Q. Weinberger, editeurs, Advances in Neural Information Processing Systems 27, pages 2537–2545. Curran Associates, Inc., 2014. (Cited on pages 5 and 3.)

[Geoffrion 1976] A. M. Geoffrion et G. W. Graves. *Scheduling Parallel Production Lines with Changeover Costs: Practical Application of a Quadratic Assignment/LP Approach.* Operations Research, vol. 24, no. 4, pages 595–610, 1976. (Cited on page 35.)

[Givens 1984] Clark R. Givens et Rae Michael Shortt. *A class of Wasserstein metrics for probability distributions.* Michigan Math. J., vol. 31, no. 2, pages 231–240, 1984. (Cited on page 17.)

[Glunt 1990] W. Glunt, T. L. Hayden, S. Hong et J. Wells. *An Alternating Projection Algorithm for Computing the Nearest Euclidean Distance Matrix.* SIAM J. Matrix Anal. Appl., vol. 11, no. 4, page 589–600, Septembre 1990. (Cited on page 161.)

[Goldfeld 2020] Ziv Goldfeld et Kristjan Greenewald. *Gaussian-Smoothed Optimal Transport: Metric Structure and Statistical Efficiency.* In Silvia Chiappa et Roberto Calandra, editeurs, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, pages 3327–3337, Online, 26–28 Aug 2020. PMLR. (Cited on page 19.)

[Goodall 1991] Colin Goodall. *Procrustes Methods in the Statistical Analysis of Shape.* Journal of the Royal Statistical Society: Series B (Methodological), vol. 53, no. 2, pages 285–321, 1991. (Cited on page 102.)

[Gordaliza 2019] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice et Jean-Michel Loubes. *Obtaining Fairness using Optimal Transport Theory*. In Kamalika Chaudhuri et Ruslan Salakhutdinov, editeurs, Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 2357–2365, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (Cited on pages 3, 1 and 25.)

[Gower 2004] John C. Gower et Garmt B. Dijksterhuis. Procrustes problems, volume 30 of *Oxford Statistical Science Series*. Oxford University Press, 2004. (Cited on page 102.)

[Gramfort 2015] A. Gramfort, G. Peyré et M. Cuturi. *Fast Optimal Transport Averaging of Neuroimaging Data*. In Sebastien Ourselin, Daniel C. Alexander, Carl-Fredrik Westin et M. Jorge Cardoso, editeurs, Information Processing in Medical Imaging, pages 261–272, Cham, 2015. Springer International Publishing. (Cited on page 25.)

[Grave 2019] Edouard Grave, Armand Joulin et Quentin Berthet. *Unsupervised Alignment of Embeddings with Wasserstein Procrustes*. In AISTATS, pages 1880–1890, 2019. (Cited on pages 33 and 102.)

[Gretton 2007] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf et Alex J. Smola. *A Kernel Method for the Two-Sample-Problem*. In B. Schölkopf, J. C. Platt et T. Hoffman, editeurs, Advances in Neural Information Processing Systems 19, pages 513–520. MIT Press, 2007. (Cited on pages 6 and 4.)

[Gromov 1999] M. Gromov, J. Lafontaine et P. Pansu. Metric Structures for Riemannian and Non-Riemannian Spaces:. Progress in Mathematics - Birkhäuser. Birkhäuser, 1999. (Cited on page 33.)

[Gulrajani 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin et Aaron C Courville. *Improved Training of Wasserstein GANs*. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, editeurs, Advances in Neural Information Processing Systems 30, pages 5767–5777. Curran Associates, Inc., 2017. (Cited on page 24.)

[Haker 2001] S. Haker et A. Tannenbaum. *Optimal mass transport and image registration*. In Proceedings IEEE Workshop on Variational and Level Set Methods in Computer Vision, pages 29–36, 2001. (Cited on pages 3, 1 and 102.)

[Hartigan 1972] J. A. Hartigan. *Direct Clustering of a Data Matrix*. Journal of the American Statistical Association, vol. 67, no. 337, pages 123–129, 1972. (Cited on page 113.)

[Hinton 2012] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath et Brian Kingsbury. *Deep Neural Networks for Acoustic Modeling in Speech Recognition*. Signal Processing Magazine, 2012. (Cited on pages 5 and 3.)

[Hjort 2010] N. Hjort, C. Holmes, P. Mueller et S. Walker. Bayesian nonparametrics: Principles and practice. Cambridge University Press, Cambridge, UK, 2010. (Cited on pages 5 and 3.)

[Horn 1991] Roger A. Horn et Charles R. Johnson. Topics in matrix analysis. Cambridge University Press, 1991. (Cited on page 88.)

[Horst 1996] R. Horst et H. Tuy. Global optimization: Deterministic approaches. Springer Berlin Heidelberg, 1996. (Cited on pages 105 and 106.)

[Huang 2016] G. Huang, C. Guo, M. Kusner, Y. Sun, F. Sha et K. Weinberger. *Supervised Word Mover's Distance*. In Advances in Neural Information Processing Systems, pages 4862–4870, 2016. (Cited on pages 43 and 70.)

[Hull 1994] J. J. Hull. *A database for handwritten text recognition research*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 5, pages 550–554, 1994. (Cited on pages 5 and 3.)

[Isserlis 1918] L. Isserlis. *On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables*. Biometrika, vol. 12, no. 1-2, pages 134–139, 11 1918. (Cited on pages 96 and 146.)

[Jaggi 2013] Martin Jaggi. *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization*. In Sanjoy Dasgupta et David McAllester, editeurs, Proceedings of the 30th International Conference on Machine Learning, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. (Cited on pages 47 and 48.)

[Jianbo Shi 2000] Jianbo Shi et J. Malik. *Normalized cuts and image segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pages 888–905, 2000. (Cited on pages 5 and 3.)

[Jiang 2014] Bo Jiang et Yu-Hong Dai. *A Framework of Constraint Preserving Update Schemes for Optimization onStiefel Manifold*. Mathematical Programming, 09 2014. (Cited on page 97.)

[Kantorovich 1942] L. Kantorovich. *On the translocation of masses*. C.R. (Doklady) Acad. Sci. URSS (N.S.), vol. 37, pages 199–201, 1942. (Cited on page 8.)

[Karcher 2014] Hermann Karcher. *Riemannian Center of Mass and so called karcher mean*, 2014. (Cited on page 25.)

[Kersting 2016] Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel et Marion Neumann. *Benchmark Data Sets for Graph Kernels*, 2016. (Cited on page 51.)

[Kipf 2016] Thomas N. Kipf et Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. CoRR, vol. abs/1609.02907, 2016. (Cited on page 42.)

[Kochurov 2020] Max Kochurov, Rasul Karimov et Serge Kozlukov. *Geoopt: Riemannian Optimization in PyTorch*, 2020. (Cited on page 97.)

[Kolouri 2016] Soheil Kolouri, Yang Zou et Gustavo K. Rohde. *Sliced Wasserstein Kernels for Probability Distributions*. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. (Cited on pages 24 and 71.)

[Kolouri 2017] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev et G. K. Rohde. *Optimal Mass Transport: Signal processing and machine-learning applications*. IEEE Signal Processing Magazine, vol. 34, no. 4, pages 43–59, 2017. (Cited on pages 3 and 1.)

[Kolouri 2019a] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau et Gustavo Rohde. *Generalized Sliced Wasserstein Distances*. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox et R. Garnett, editeurs, Advances in Neural Information Processing Systems 32, pages 261–272. Curran Associates, Inc., 2019. (Cited on pages 24 and 25.)

[Kolouri 2019b] Soheil Kolouri, Phillip E. Pope, Charles E. Martin et Gustavo K. Rohde. *Sliced Wasserstein Auto-Encoders*. In International Conference on Learning Representations, 2019. (Cited on page 71.)

[Kondor 2018] Risi Kondor et Shubhendu Trivedi. *On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups*, 2018. (Cited on pages 5 and 3.)

[Konno 1976a] Hiroshi Konno. *A Cutting Plane Algorithm for Solving Bilinear Programs*. Math. Program., vol. 11, no. 1, page 14–27, 1976. (Cited on page 106.)

[Konno 1976b] Hiroshi Konno. *Maximization of A convex quadratic function under linear constraints*. Mathematical Programming, vol. 11, no. 1, pages 117–127, 1976. (Cited on pages 107 and 150.)

[Koopmans 1957] Tjalling Koopmans et Martin J. Beckmann. *Assignment Problems and the Location of Economic Activities*. Econometrica: journal of the Econometric Society, no. 53–76, 1957. (Cited on pages 34 and 71.)

[Korba 2018] Anna Korba, Alexandre Garcia et Florence d'Alché Buc. *A Structured Prediction Approach for Label Ranking*. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi et R. Garnett, editeurs, Advances in Neural Information Processing Systems 31, pages 8994–9004. Curran Associates, Inc., 2018. (Cited on pages 5 and 3.)

[Kriege 2016] Nils M. Kriege, Pierre-Louis Giscard et Richard C. Wilson. *On Valid Optimal Assignment Kernels and Applications to Graph Classification*. CoRR, vol. abs/1606.01141, 2016. (Cited on pages 4, 3, 42, 51 and 52.)

[Kriege 2018] Nils Kriege, Matthias Fey, Denis Fisseler, Petra Mutzel et Frank Weichert. *Recognizing Cuneiform Signs Using Graph Based Methods*. In International Workshop on Cost-Sensitive Learning (COST), 2018. (Cited on pages 51 and 53.)

[Kriege 2020] Nils M. Kriege, Fredrik D. Johansson et Christopher Morris. *A survey on graph kernels*. Applied Network Science, vol. 5, no. 1, page 6, Janvier 2020. (Cited on page 162.)

[Ktena 2017] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker et Daniel Rueckert. *Distance metric learning using graph convolutional networks: Application to functional brain networks*. In MICCAI, pages 469–477, 2017. (Cited on pages 4, 3 and 42.)

[Kusner 2015] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin et Kilian Q. Weinberger. *From Word Embeddings to Document Distances*. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, page 957–966. JMLR.org, 2015. (Cited on pages 3, 7, 1 and 5.)

[Kwon 2020] Oh-Hyun Kwon et Kwan-Liu Ma. *A Deep Generative Model for Graph Layout*. IEEE Transactions on Visualization and Computer Graphics, vol. 26, 01 2020. (Cited on page 39.)

[Laclau 2017] Charlotte Laclau, Ievgen Redko, Basarab Matei, Younès Bennani et Vincent Brault. *Co-clustering through Optimal Transport*. In ICML, pages 1955–1964, 2017. (Cited on pages 113, 114 and 115.)

[Lacoste-Julien 2016] Simon Lacoste-Julien. *Convergence rate of Frank-Wolfe for non-convex objectives*. arXiv preprint arXiv:1607.00345, 2016. (Cited on page 48.)

[Lafferty 2001] John D. Lafferty, Andrew McCallum et Fernando C. N. Pereira. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. (Cited on pages 5 and 3.)

[Lai 1988] Hang-Chin Lai et Lai-Jui Lin. *The Fenchel-Moreau Theorem for Set Functions*. Proceedings of the American Mathematical Society, vol. 103, no. 1, pages 85–90, 1988. (Cited on page 89.)

[Lai 2014] Rongjie Lai et Hongkai Zhao. *Multi-scale Non-Rigid Point Cloud Registration Using Robust Sliced-Wasserstein Distance via Laplace-Beltrami Eigenmap*. SIAM Journal on Imaging Sciences, vol. 10, 06 2014. (Cited on page 74.)

[Laporte 1988] Gilbert Laporte et Hélène Mercure. *Balancing hydraulic turbine runners: A quadratic assignment problem*. European Journal of Operational Research, vol. 35, no. 3, pages 378 – 381, 1988. (Cited on page 35.)

[Lassalle 2018] Rémi Lassalle. *Causal transport plans and their Monge–Kantorovich problems*. Stochastic Analysis and Applications, vol. 36, no. 3, pages 452–484, 2018. (Cited on page 162.)

[LeCun 2010] Yann LeCun et Corinna Cortes. *MNIST handwritten digit database*. 2010. (Cited on pages 5 and 3.)

[Lee 2019] John Lee, Max Dabagia, Eva Dyer et Christopher Rozell. *Hierarchical Optimal Transport for Multimodal Distribution Alignment*. In NeurIPS, pages 13474–13484. Curran Associates, Inc., 2019. (Cited on page 109.)

[Lei 2020] Jing Lei. *Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces*. arXiv: Statistics Theory, pages 767–798, 2020. (Cited on page 19.)

[Li 2019] Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals et Pushmeet Kohli. *Graph Matching Networks for Learning the Similarity of Graph Structured Objects*. In Kamalika Chaudhuri et Ruslan Salakhutdinov, editeurs, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, pages 3835–3845. PMLR, 2019. (Cited on page 162.)

[Liberti 2014] Leo Liberti, Carlile Lavor, Nelson Maculan et Antonio Mucherino. *Euclidean Distance Geometry and Applications*. SIAM Review, vol. 56, no. 1, pages 3–69, 2014. (Cited on page 161.)

[Lin 2020] Tianyi Lin, Zeyu Zheng, Elynn Y. Chen, Marco Cuturi et Michael I. Jordan. *On Projection Robust Optimal Transport: Sample Complexity and Model Misspecification*, 2020. (Cited on pages 19 and 24.)

[Liu 2016] Huikang Liu, Weijie Wu et Anthony Man-Cho So. *Quadratic Optimization with Orthogonality Constraints: Explicit Lojasiewicz Exponent and Linear Convergence of Line-Search Methods*. In Maria Florina Balcan et Kilian Q. Weinberger, editeurs, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, pages 1158–1167, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on page 97.)

[Liutkus 2019] Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus et Fabian-Robert Stöter. *Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and*

Diffusions. In Kamalika Chaudhuri et Ruslan Salakhutdinov, editeurs, Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 4104–4113, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (Cited on page 71.)

[Loiola 2007] Eliane Loiola, Nair Abreu, Paulo Boaventura-Netto, Peter Hahn et Tania Querido. *A survey of the quadratic assignment problem*. European Journal of Operational Research, vol. 176, pages 657–690, 2007. (Cited on pages 34, 35 and 47.)

[Loukas 2019] Andreas Loukas. *Graph Reduction with Spectral and Cut Guarantees*. Journal of Machine Learning Research, vol. 20, no. 116, pages 1–42, 2019. (Cited on pages 54 and 162.)

[Luss 2007] Ronny Luss et Alexandre d'Aspremont. *Support Vector Machine Classification with Indefinite Kernels*. In Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS'07, pages 953–960, 2007. (Cited on page 52.)

[Lyu 2020] Boyang Lyu, Thao Pham, Giles Blaney, Zachary Haga, Sergio Fantini et Shuchin Aeron. *Domain Adaptation for Robust Workload Classification using fNIRS*, 2020. (Cited on page 57.)

[Lyzinski 2016] V. Lyzinski, D. E. Fishkind, M. Fiori, J. T. Vogelstein, C. E. Priebe et G. Sapiro. *Graph Matching: Relax at Your Own Risk*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pages 60–73, 2016. (Cited on page 35.)

[Lévy 2018] Bruno Lévy et Erica L. Schwindt. *Notions of optimal transport theory and how to implement them on a computer*. Computers & Graphics, vol. 72, pages 135 – 148, 2018. (Cited on page 25.)

[Ma 2005] Xinan Ma, Neil Trudinger et Xu-Jia Wang. *Regularity of Potential Functions of the Optimal Transportation Problem*. Archive for Rational Mechanics and Analysis, vol. 177, pages 151–183, 01 2005. (Cited on page 16.)

[Maclaurin 2015] Dougal Maclaurin, David Duvenaud et Ryan P Adams. *Autograd: Effortless gradients in numpy*. In ICML 2015 AutoML Workshop, 2015. (Cited on page 75.)

[Maron 2018] Haggai Maron et Yaron Lipman. *(Probably) Concave Graph Matching*. In Advances in Neural Information Processing Systems, pages 408–418, 2018. (Cited on pages 35, 72, 108, 132 and 153.)

[McCann 1995] Robert J. McCann. *Existence and uniqueness of monotone measure-preserving maps*. Duke Math. J., vol. 80, no. 2, pages 309–323, 11 1995. (Cited on page 83.)

[McCann 1997] Robert J. McCann. *A Convexity Principle for Interacting Gases*. Advances in Mathematics, vol. 128, no. 1, pages 153–179, jun 1997. (Cited on pages 17, 18 and 27.)

[Meghwanshi 2018] Mayank Meghwanshi, Pratik Jawanpuria, Anoop Kunchukuttan, Hiroyuki Kasai et Bamdev Mishra. *McTorch, a manifold optimization library for deep learning*. arXiv preprint arXiv:1810.01811, 2018. (Cited on page 75.)

[Memoli 2011] Facundo Memoli. *Gromov Wasserstein Distances and the Metric Approach to Object Matching*. Foundations of Computational Mathematics, pages 1–71, 2011. (Cited on pages 3, 2, 7, 27, 33, 37, 39, 43 and 102.)

[Mémoli 2018] Facundo Mémoli et Tom Needham. *Gromov-Monge quasi-metrics and distance distributions*. arXiv:1810.09646, 2018. (Cited on pages 72 and 95.)

[Mensch 2018] Arthur Mensch et Mathieu Blondel. *Differentiable Dynamic Programming for Structured Prediction and Attention*. In Jennifer Dy et Andreas Krause, editeurs, Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 3462–3471, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. (Cited on pages 5 and 3.)

[Mikolov 2013a] Tomas Mikolov, Kai Chen, Greg S. Corrado et Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*, 2013. (Cited on pages 5 and 3.)

[Mikolov 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado et Jeff Dean. *Distributed Representations of Words and Phrases and their Compositionality*. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani et K. Q. Weinberger, editeurs, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc., 2013. (Cited on pages 5 and 3.)

[Monge 1781] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. Histoire de l'Académie Royale des Sciences, pages 666–704, 1781. (Cited on pages 3, 1, 7 and 20.)

[Murty 1988] K.G. Murty. Linear Complementarity, Linear and Nonlinear Programming. Sigma series in applied mathematics. Heldermann, 1988. (Cited on page 129.)

[Muzellec 2018] Boris Muzellec et Marco Cuturi. *Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions*. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi et R. Garnett, editeurs, Advances in Neural Information Processing Systems 31, pages 10237–10248. Curran Associates, Inc., 2018. (Cited on page 18.)

[Nadif 2008] M. Nadif et G. Govaert. *Algorithms for Model-based Block Gaussian Clustering*. In DMIN'08, the 2008 International Conference on Data Mining, 2008. (Cited on page 113.)

[Nadjahi 2019] Kimia Nadjahi, Alain Durmus, Umut Simsekli et Roland Badeau. *Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance*. In Advances in Neural Information Processing Systems 32, pages 250–260. Curran Associates, Inc., 2019. (Cited on page 24.)

[Nadjahi 2020] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour et Umut Simsekli. *Statistical and Topological Properties of Sliced Probability Divergences*, 03 2020. (Cited on page 24.)

[Nenna 2016] Luca Nenna. *Numerical Methods for Multi-Marginal Optimal Transportation*. PhD thesis, 12 2016. (Cited on pages 23 and 25.)

[Neumann 2016] Marion Neumann, Roman Garnett, Christian Bauckhage et Kristian Kersting. *Propagation kernels: efficient graph kernels from propagated information*. Machine Learning, vol. 102, no. 2, pages 209–245, Feb 2016. (Cited on page 52.)

[Niepert 2016] Mathias Niepert, Mohamed Ahmed et Konstantin Kutzkov. *Learning Convolutional Neural Networks for Graphs*. In ICML, volume 48 of *Proceedings of Machine Learning Research*, pages 2014–2023, New York, New York, USA, 2016. PMLR. (Cited on page 52.)

[Nikolentzos 2017] Giannis Nikolentzos, Polykarpos Meladianos et Michalis Vazirgiannis. *Matching Node Embeddings for Graph Similarity*. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., pages 2429–2435, 2017. (Cited on page 43.)

[Niles-Weed 2019] Jonathan Niles-Weed et Philippe Rigollet. *Estimation of Wasserstein distances in the Spiked Transport Model*. 09 2019. (Cited on page 19.)

[Nowicki 2001] Krzysztof Nowicki et Tom A B Snijders. *Estimation and prediction for stochastic block-structures*. Journal of the American statistical association, vol. 96, no. 455, pages 1077–1087, 2001. (Cited on pages 56 and 57.)

[Ortega 2018] A. Ortega, P. Frossard, J. Kova**v**cević, J. M. F. Moura et P. Vandergheynst. *Graph Signal Processing: Overview, Challenges, and Applications*. Proceedings of the IEEE, vol. 106, no. 5, pages 808–828, 2018. (Cited on page 162.)

[Pardalos 1987] Panos M. Pardalos et J. Ben Rosen, editeurs. Bilinear programming methods for nonconvex quadratic problems, pages 75–83. Springer Berlin Heidelberg, 1987. (Cited on page 105.)

[Paszke 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga et Adam Lerer. *Automatic differentiation in pytorch*. 2017. (Cited on page 75.)

[Patrikainen 2006] A. Patrikainen et M. Meila. *Comparing subspace clusterings*. IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 7, pages 902–916, 2006. (Cited on page 113.)

[Paty 2019] François-Pierre Paty et Marco Cuturi. *Subspace Robust Wasserstein Distances*. In Kamalika Chaudhuri et Ruslan Salakhutdinov, editeurs, Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 5072–5081, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (Cited on pages 24, 25, 74 and 142.)

[Pearl 1986] J Pearl. *Fusion, Propagation, and Structuring in Belief Networks*. Artif. Intell., vol. 29, no. 3, pages 241–288, Septembre 1986. (Cited on pages 5 and 3.)

[Pearl 2009] Judea Pearl. Causality: Models, reasoning and inference. Cambridge University Press, New York, NY, USA, 2nd édition, 2009. (Cited on pages 5 and 3.)

[Pennington 2014] Jeffrey Pennington, Richard Socher et Christopher Manning. *GloVe: Global Vectors for Word Representation*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, Octobre 2014. Association for Computational Linguistics. (Cited on pages 5 and 3.)

[Petersen 2012] K. B. Petersen et M. S. Pedersen. *The Matrix Cookbook*, nov 2012. Version 20121115. (Cited on page 36.)

[Peyré 2016] Gabriel Peyré, Marco Cuturi et Justin Solomon. *Gromov-Wasserstein averaging of kernel and distance matrices*. In ICML, pages 2664–2672, 2016. (Cited on pages 23, 34, 35, 36, 37, 38, 43, 47, 48, 49, 62, 71, 76, 93, 103, 106, 108, 150 and 151.)

[Peyré 2019] Gabriel Peyré et Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, vol. 11, pages 355–607, 2019. (Cited on pages 7, 11, 16, 20, 22, 24, 25, 43, 73 and 149.)

[Puy 2020] Gilles Puy, Alexandre Boulch et Renaud Marlet. *FLOT: Scene Flow on Point Clouds Guided by Optimal Transport*, 2020. (Cited on page 57.)

[Rabin 2011] Julien Rabin, Gabriel Peyré, Julie Delon et Marc Bernot. *Wasserstein barycenter and its application to texture mixing.* In International Conference on Scale Space and Variational Methods in Computer Vision, pages 435–446. Springer, 2011. (Cited on pages 24 and 71.)

[Rabin 2012] Julien Rabin, Gabriel Peyré, Julie Delon et Marc Bernot. *Wasserstein Barycenter and Its Application to Texture Mixing.* In Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein et Michael M. Bronstein, editeurs, Scale Space and Variational Methods in Computer Vision, pages 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. (Cited on page 25.)

[Rangarajan 1997a] Anand Rangarajan, Haili Chui et Fred L. Bookstein. *The softassign Procrustes matching algorithm.* In Information Processing in Medical Imaging, pages 29–42, 1997. (Cited on page 102.)

[Rangarajan 1997b] Anand Rangarajan, Alan L Yuille, Steven Gold et Eric Mjolsness. *A Convergence Proof for the Softassign Quadratic Assignment Algorithm.* In M. C. Mozer, M. I. Jordan et T. Petsche, editeurs, Advances in Neural Information Processing Systems 9, pages 620–626. MIT Press, 1997. (Cited on page 37.)

[Rangarajan 1999] Anand Rangarajan, A Vuille et Eric Mjolsness. *Convergence Properties of the Softassign Quadratic Assignment Algorithm.* Neural computation, vol. 11, pages 1455–74, 09 1999. (Cited on page 37.)

[Rapp 1995] Reinhard Rapp. *Identifying Word Translations in Non-Parallel Texts.* In ACL, pages 320–322, 1995. (Cited on page 102.)

[Raymond 2002] John Raymond, Eleanor Gardiner et Peter Willett. *RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs.* Comput. J., vol. 45, pages 631–644, 04 2002. (Cited on page 162.)

[Redko 2019] I. Redko, N. Courty, R. Flamary et D. Tuia. *Optimal Transport for Multi-source Domain Adaptation under Target Shift.* In International Conference on Artificial Intelligence and Statistics (AISTAT), 2019. (Cited on page 110.)

[Redko 2020] Ievgen Redko, Titouan Vayer, Rémi Flamary et Nicolas Courty. *CO-Optimal Transport*, 2020. (Cited on pages 10, 5, 102 and 160.)

[Riba 2018] P. Riba, A. Fischer, J. Llados et A. Fornés. *Learning Graph Distances with Message Passing Neural Networks.* In 2018 24th International Conference on Pattern Recognition (ICPR), pages 2239–2244, 2018. (Cited on page 162.)

[Rocci 2008] R. Rocci et M. Vichi. *Two-mode multi-partitioning.* Computational Statistics and Data Analysis, vol. 52, no. 4, pages 1984–2003, 2008. (Cited on page 113.)

[Rockafellar 1970] R. Tyrrell Rockafellar. Convex analysis. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970. (Cited on pages 13 and 129.)

[Rubner 1998] Y. Rubner, C. Tomasi et L.J. Guibas. *A metric for distributions with applications to image databases.* In ICCV, pages 59–66, 1998. (Cited on pages 3 and 1.)

[Rubner 2000] Yossi Rubner, Carlo Tomasi et Leonidas J. Guibas. *The Earth Mover's Distance as a Metric for Image Retrieval*. International Journal of Computer Vision, vol. 40, no. 2, pages 99–121, Novembre 2000. (Cited on pages 7 and 5.)

[Rudin 1976] W. Rudin. Principles of Mathematical Analysis. International series in pure and applied mathematics. McGraw-Hill, 1976. (Cited on page 138.)

[Rustamov 2013] Raif M Rustamov, Maks Ovsjanikov, Omri Azencot, Mirela Ben-Chen, Frédéric Chazal et Leonidas Guibas. *Map-based exploration of intrinsic shape differences and variability*. ACM Transactions on Graphics (TOG), vol. 32, no. 4, page 72, 2013. (Cited on page 76.)

[Saenko 2010] K. Saenko, B. Kulis, M. Fritz et T. Darrell. *Adapting Visual Category Models to New Domains*. In ECCV, LNCS, pages 213–226, 2010. (Cited on page 111.)

[Sakoe 1978] Hiroaki Sakoe et Seibi Chiba. *Dynamic programming algorithm optimization for spoken word recognition*. ieeeassp, vol. 26, no. 1, pages 43–49, 1978. (Cited on page 162.)

[Samaria 1994] Ferdinando S Samaria et Andy C Harter. *Parameterisation of a stochastic model for human face identification*. In Proceedings of 1994 IEEE workshop on applications of computer vision, pages 138–142. IEEE, 1994. (Cited on page 114.)

[Santambrogio 2015] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, 2015. (Cited on pages 7, 10, 13, 14, 16, 28, 86, 87, 89 and 121.)

[Sato 2020] Ryoma Sato, Marco Cuturi, Makoto Yamada et Hisashi Kashima. *Fast and Robust Comparison of Probability Measures in Heterogeneous Spaces*, 2020. (Cited on page 37.)

[Schellewald 2001] Christian Schellewald, Stefan Roth et Christoph Schnörr. *Evaluation of Convex Optimization Techniques for the Weighted Graph-Matching Problem in Computer Vision*. In Bernd Radig et Stefan Florczyk, editeurs, Pattern Recognition, pages 361–368, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. (Cited on page 35.)

[Schiebinger 2019] Geoffrey Schiebinger, 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 et Eric S. Lander. *Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming*. Cell, vol. 176, no. 4, pages 928 – 943.e22, 2019. (Cited on pages 3 and 1.)

[Schmitzer 2013] Bernhard Schmitzer et Christoph Schnörr. *Modelling Convex Shape Priors and Matching Based on the Gromov-Wasserstein Distance*. Journal of Mathematical Imaging and Vision, vol. 46, no. 1, pages 143–159, Mai 2013. (Cited on page 39.)

[Schmitzer 2016] Bernhard Schmitzer. *Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems*. SIAM Journal on Scientific Computing, vol. 41, no. 3, pages A1443–A1481, 2016. (Cited on pages 22 and 76.)

[Schrodinger 1931] E. Schrodinger. *Uber die umkehrung der naturgesetze*. Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math., 144, pages 876—879, 1931. (Cited on page 21.)

[Seguy 2018] Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet et Mathieu Blondel. *Large Scale Optimal Transport and Mapping Estimation*. In International Conference on Learning Representations, 2018. (Cited on pages 23 and 24.)

[Shan 2010] Hanhuai Shan et Arindam Banerjee. *Residual Bayesian Co-clustering for Matrix Approximation*. In SDM, pages 223–234, 2010. (Cited on page 113.)

[Shervashidze 2009] Nino Shervashidze, S. V. N. Vishwanathan, Tobias H. Petri, Kurt Mehlhorn et et al. *Efficient graphlet kernels for large graph comparison*, 2009. (Cited on pages 52 and 53.)

[Shervashidze 2011] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn et Karsten M. Borgwardt. *Weisfeiler-Lehman Graph Kernels*. J. Mach. Learn. Res., vol. 12, pages 2539–2561, Novembre 2011. (Cited on page 42.)

[Siglidis 2018] Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis et Michalis Vazirgiannis. *GraKeL: A Graph Kernel Library in Python*. arXiv e-prints, Juin 2018. (Cited on page 52.)

[Sinkhorn 1967] Richard Sinkhorn et Paul Knopp. *Concerning nonnegative matrices and doubly stochastic matrices*. Pacific J. Math., vol. 21, no. 2, pages 343–348, 1967. (Cited on page 22.)

[Solomon 2014] Justin Solomon, Raif Rustamov, Leonidas Guibas et Adrian Butscher. *Wasserstein Propagation for Semi-Supervised Learning*. In Eric P. Xing et Tony Jebara, editeurs, Proceedings of the 31st International Conference on Machine Learning, volume 32 of *Proceedings of Machine Learning Research*, pages 306–314, Bejing, China, 22–24 Jun 2014. PMLR. (Cited on pages 3 and 1.)

[Solomon 2015] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du et Leonidas Guibas. *Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains*. ACM Trans. Graph., vol. 34, no. 4, Juillet 2015. (Cited on pages 25 and 27.)

[Solomon 2016] Justin Solomon, Gabriel Peyré, Vladimir G. Kim et Suvrit Sra. *Entropic metric alignment for correspondence problems*. ACM Transactions on Graphics, vol. 35, no. 4, pages 1–13, 2016. (Cited on pages 35, 39, 76 and 102.)

[Song 2016] Yang Song, Peter J. Schreier, David Ramírez et Tanuj Hasija. *Canonical correlation analysis of high-dimensional data with very small sample support*. Signal Process., vol. 128, pages 449–458, 2016. (Cited on page 112.)

[Srivastava 2015] Sanvesh Srivastava, Volkan Cevher, Quoc Dinh et David Dunson. *WASP: Scalable Bayes via barycenters of subset posteriors*. In Guy Lebanon et S. V. N. Vishwanathan, editeurs, Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, volume 38 of *Proceedings of Machine Learning Research*, pages 912–920, San Diego, California, USA, 09–12 May 2015. PMLR. (Cited on page 25.)

[Sturm 2012] Karl-Theodor Sturm. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*. arXiv e-prints, page arXiv:1208.0434, 2012. (Cited on pages 7, 27, 28, 32, 34, 66 and 80.)

[Sun 2020] Yizhou Sun. *Graph Neural Networks for Graph Search.* In Proceedings of the 3rd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), GRADES-NDA'20, New York, NY, USA, 2020. Association for Computing Machinery. (Cited on page 162.)

[Sutherland 2003] Jeffrey J Sutherland, Lee A O'brien et Donald F Weaver. *Spline-fitting with a genetic algorithm: A method for developing classification structure- activity relationships.* Journal of chemical information and computer sciences, vol. 43, no. 6, pages 1906–1915, 2003. (Cited on page 51.)

[Sutskever 2014] Ilya Sutskever, Oriol Vinyals et Quoc V Le. *Sequence to Sequence Learning with Neural Networks.* In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence et K. Q. Weinberger, editeurs, Advances in Neural Information Processing Systems 27, pages 3104–3112. Curran Associates, Inc., 2014. (Cited on pages 5 and 3.)

[Szegedy 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke et Andrew Rabinovich. *Going deeper with convolutions.* In CVPR, pages 1–9, 2015. (Cited on page 111.)

[Takatsu 2011] Asuka Takatsu. *Wasserstein geometry of Gaussian measures.* Osaka J. Math., vol. 48, no. 4, pages 1005–1026, 12 2011. (Cited on pages 17, 85 and 95.)

[Tao 2005] Pham Dinh Tao*et al. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems.* Annals of operations research, vol. 133, no. 1-4, pages 23–46, 2005. (Cited on page 108.)

[Taskar 2005] Ben Taskar, Vassil Chatalbashev, Daphne Koller et Carlos Guestrin. *Learning Structured Prediction Models: A Large Margin Approach.* In Proceedings of the 22nd International Conference on Machine Learning, ICML '05, page 896–903, New York, NY, USA, 2005. Association for Computing Machinery. (Cited on pages 5 and 3.)

[Thorpe 2017] Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K. Rohde et Dejan Slep**v**cev. *A Transportation L^p Distance for Signal Analysis.* Journal of Mathematical Imaging and Vision, vol. 59, no. 2, pages 187–210, 2017. (Cited on pages 43 and 63.)

[Townsend 2016] James Townsend, Niklas Koep et Sebastian Weichwald. *Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation.* The Journal of Machine Learning Research, vol. 17, no. 1, pages 4755–4759, 2016. (Cited on page 75.)

[Vayer 2019a] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard et Nicolas Courty. *Optimal Transport for structured data with application on graphs.* In International Conference on Machine Learning, volume 97, 2019. (Cited on pages 7, 5, 42 and 160.)

[Vayer 2019b] Titouan Vayer, Rémi Flamary, Nicolas Courty, Romain Tavenard et Laetitia Chapel. *Sliced Gromov-Wasserstein.* In Advances in Neural Information Processing Systems 32, pages 14753–14763. Curran Associates, Inc., 2019. (Cited on pages 9, 5, 70 and 160.)

[Vayer 2020a] Titouan Vayer, Laetitia Chapel, Nicolas Courty, Rémi Flamary, Yann Soullard et Romain Tavenard. *Time Series Alignment with Global Invariances*, 2020. (Cited on pages 7, 5 and 160.)

[Vayer 2020b] Titouan Vayer, Laetitia Chapel, Remi Flamary, Romain Tavenard et Nicolas Courty. *Fused Gromov-Wasserstein Distance for Structured Objects.* Algorithms, vol. 13, no. 9, page 212, Aug 2020. (Cited on pages 7, 5, 42 and 160.)

[Veraguas 2016] Julio Backhoff Veraguas, Mathias Beiglböck, Yiqing Lin et Anastasiia Zalashko. *Causal transport in discrete time and applications*, 2016. (Cited on page 162.)

[Verma 2017] Saurabh Verma et Zhi-Li Zhang. *Hunt For The Unique, Stable, Sparse And Fast Feature Learning On Graphs.* In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, editeurs, NIPS, pages 88–98. Curran Associates, Inc., 2017. (Cited on page 44.)

[Villani 2003] C. Villani. *Topics in Optimal Transportation Theory.* vol. 58, 01 2003. (Cited on pages 10 and 12.)

[Villani 2008] Cédric Villani. Optimal Transport: Old and New. Springer, 2008. (Cited on pages 7, 11, 14, 15, 66, 122 and 149.)

[Vishwanathan 2010] S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor et Karsten M. Borgwardt. *Graph Kernels.* J. Mach. Learn. Res., vol. 11, pages 1201–1242, 2010. (Cited on pages 42 and 52.)

[Wale 2008] Nikil Wale, Ian A. Watson et George Karypis. *Comparison of descriptor spaces for chemical compound retrieval and classification.* Knowledge and Information Systems, vol. 14, no. 3, pages 347–375, Mar 2008. (Cited on page 51.)

[Wang 1987] Yuchung J Wang et George Y Wong. *Stochastic blockmodels for directed graphs.* Journal of the American Statistical Association, vol. 82, no. 397, pages 8–19, 1987. (Cited on pages 56 and 57.)

[Wang 2009] Chang Wang et Sridhar Mahadevan. *Manifold Alignment without Correspondence.* In IJCAI, page 1273–1278, 2009. (Cited on page 102.)

[Wang 2011] Chang Wang et Sridhar Mahadevan. *Heterogeneous Domain Adaptation Using Manifold Alignment.* In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11, page 1541–1546. AAAI Press, 2011. (Cited on pages 5 and 3.)

[Wang 2018] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta et Kaiming He. *Non-local Neural Networks.* 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7794–7803, 2018. (Cited on pages 5 and 3.)

[Weed 2017] Jonathan Weed et Francis Bach. *Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance.* arXiv preprint arXiv:1707.00087, 2017. (Cited on pages 18, 64 and 125.)

[Weed 2019a] Jonathan Weed et Quentin Berthet. *Estimation of smooth densities in Wasserstein distance.* In Alina Beygelzimer et Daniel Hsu, editeurs, Proceedings of the Thirty-Second Conference on Learning Theory, volume 99 of *Proceedings of Machine Learning Research*, pages 3118–3119, Phoenix, USA, 25–28 Jun 2019. PMLR. (Cited on page 19.)

[Weed 2019b] Jonathan Daniel Weed. *Statistical Problems in Transport and Alignment.* PhD thesis, 04 2019. (Cited on page 19.)

[Wen 2010] Zaiwen Wen et Wotao Yin. *A feasible method for optimization with orthogonality constraints*. Mathematical Programming, vol. 142, 12 2010. (Cited on page 97.)

[West 2000] Douglas B. West. Introduction to graph theory. Prentice Hall, 2 édition, September 2000. (Cited on page 46.)

[Willett 1998] Peter Willett, John M. Barnard et Geoffrey M. Downs. *Chemical similarity searching*. J. Chem. Inf. Comput. Sci, pages 983–996, 1998. (Cited on page 162.)

[Wills 2020] Peter Wills et François G. Meyer. *Metrics for graph comparison: A practitioner's guide*. PLOS ONE, vol. 15, no. 2, pages 1–54, 02 2020. (Cited on page 162.)

[Wilson 1969] A. G. Wilson. *The Use of Entropy Maximising Models, in the Theory of Trip Distribution, Mode Split and Route Split*. Journal of Transport Economics and Policy, vol. 3, no. 1, pages 108–126, 1969. (Cited on page 21.)

[Wu 2019] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel et Luc Van Gool. *Sliced Wasserstein Generative Models*. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. (Cited on page 71.)

[Wu 2020] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang et P. S. Yu. *A Comprehensive Survey on Graph Neural Networks*. IEEE Transactions on Neural Networks and Learning Systems, pages 1–21, 2020. (Cited on pages 5, 3 and 42.)

[Xia 2020] Haifeng Xia et Zhengming Ding. *Structure Preserving Generative Cross-Domain Learning*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. (Cited on page 39.)

[Xu 2019a] Hongteng Xu, Dixin Luo et Lawrence Carin. *Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching*. In Advances in Neural Information Processing Systems 32, pages 3052–3062. Curran Associates, Inc., 2019. (Cited on page 39.)

[Xu 2019b] Hongteng Xu, Dixin Luo, Hongyuan Zha et Lawrence Carin Duke. *Gromov-Wasserstein Learning for Graph Matching and Node Embedding*. In Kamalika Chaudhuri et Ruslan Salakhutdinov, editeurs, Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 6932–6941, Long Beach, California, USA, 09–15 Jun 2019. PMLR. (Cited on page 39.)

[Xu 2020] Hongteng Xu, Dixin Luo, Ricardo Henao, Svati Shah et Lawrence Carin. *Learning Autoencoders with Relational Regularization*, 2020. (Cited on page 57.)

[Yan 2018] Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan et Qingyao Wu. *Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation*. In International Joint Conference on Artificial Intelligence, pages 2969–2975, 2018. (Cited on pages 39, 70, 102, 110 and 111.)

[Yanardag 2015] Pinar Yanardag et S.V.N. Vishwanathan. *Deep Graph Kernels*. In ACM, editeur, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1365–1374, 2015. (Cited on pages 4, 3, 42 and 51.)

[Yang 2020] Heng Yang, Jingnan Shi et Luca Carlone. *TEASER: Fast and Certifiable Point Cloud Registration*, 2020. (Cited on page 102.)

[Yeh 2014] Yi-Ren Yeh, Chun-Hao Huang et Yu-Chiang Frank Wang. *Heterogeneous domain adaptation and classification by exploiting the correlation subspace.* IEEE Transactions on Image Processing, vol. 23, no. 5, pages 2009–2018, 2014. (Cited on pages 5, 3, 110 and 111.)

[Yuille 2003] Alan L Yuille et Anand Rangarajan. *The concave-convex procedure.* Neural computation, vol. 15, no. 4, pages 915–936, 2003. (Cited on page 108.)

[Yurochkin 2019] Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh et Justin M. Solomon. *Hierarchical Optimal Transport for Document Representation.* In NeurIPS, pages 1599–1609, 2019. (Cited on page 109.)

[Zalashko 2017] Anastasiia Zalashko. Causal optimal transport : theory and applications. Wien, 2017. (Cited on page 162.)

[Zaslavskiy 2009] M. Zaslavskiy, F. Bach et J. Vert. *A Path Following Algorithm for the Graph Matching Problem.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 12, pages 2227–2242, 2009. (Cited on page 35.)

[Zhang 2018] Ruiyi Zhang, Changyou Chen, Chunyuan Li et Lawrence Carin Duke. *Policy Optimization as Wasserstein Gradient Flows.* In Jennifer Dy et Andreas Krause, editeurs, Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 5741–5750, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. (Cited on page 65.)

[Zhou 2014] Joey Tianyi Zhou, Ivor W.Tsang, Sinno Jialin Pan et Mingkui Tan. *Heterogeneous Domain Adaptation for Multiple Classes.* volume 33 of *Proceedings of Machine Learning Research*, pages 1095–1103, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. (Cited on pages 5 and 3.)

[Zhou 2019] Joey Tianyi Zhou, Ivor W. Tsang, Sinno Jialin Pan et Mingkui Tan. *Multi-class Heterogeneous Domain Adaptation.* Journal of Machine Learning Research, vol. 20, no. 57, pages 1–31, 2019. (Cited on page 163.)

Titre : Une contribution au Transport Optimal sur des espaces incomparables

Mot clés : Transport Optimal, Données Structurées, Wassertein et Gromov-Wasserstein

Résumé : Le Transport Optimal est une théorie permettant de définir des notions géométriques de distance entre des distributions de probabilité et de trouver des correspondances, des relations, entre des ensembles de points. De cette théorie, à la frontière entre les mathématiques et l'optimisation, découle de nombreuses applications en machine learning. Cette thèse propose d'étudier le scénario, complexe, dans lequel les différentes données appartiennent à des espaces *incomparables*. En particulier nous abordons les questions suivantes : comment définir et appliquer le transport optimal entre des graphes, entre des données structurées ? Comment l'adapter lorsque les données sont variées et ne font pas partie d'un même espace métrique ? Cette thèse propose un ensemble d'outils de Transport Optimal pour ces différents cas. Un important volet est notamment consacré à l'étude de la distance de Gromov-Wasserstein dont les propriétés permettent de définir d'intéressants problèmes de transport sur des espaces incomparables. Plus largement, nous analysons les propriétés mathématiques des différents outils proposés, nous établissons des solutions algorithmiques pour les calculer et nous étudions leur applicabilité dans de nombreux *scenarii* de machine learning qui couvrent, notamment, la classification, la simplification, le partitionnement de données structurées, ainsi que l'adaptation de domaines hétérogènes.

Title: A contribution to Optimal Transport on incomparable spaces

Keywords: Optimal Transport, Structured Data, Wasserstein and Gromov-Wasserstein

Abstract: Optimal Transport is a theory that allows to define geometrical notions of distance between probability distributions and to find correspondences, relationships, between sets of points. Many machine learning applications are derived from this theory, at the frontier between mathematics and optimization. This thesis proposes to study the complex scenario in which the different data belong to *incomparable* spaces. In particular we address the following questions: how to define and apply the optimal transport between graphs, between structured data? How can it be adapted when the data are varied and not embedded in the same metric space? This thesis proposes a set of Optimal Transport tools for these different cases. An important part is notably devoted to the study of the Gromov-Wasserstein distance whose properties allow to define interesting transport problems on incomparable spaces. More broadly, we analyze the mathematical properties of the various proposed tools, we establish algorithmic solutions to compute them and we study their applicability in numerous machine learning *scenarii* which cover, in particular, classification, simplification, partitioning of structured data, as well as heterogeneous domain adaptation.