Transport of probability distributions across different Euclidean spaces

Antoine Salmona

université PARIS-SACLAY

Transport of probability distributions across different Euclidean spaces

Transport de mesures de probabilités à travers des espaces euclidiens de dimensions différentes

Thèse de doctorat de l'université Paris-Saclay

Ecole doctorale n° 574, mathématiques Hadamard (EDMH)
Spécialité de doctorat : mathématiques appliquées
Graduate school : Mathématiques. Référent : ENS Paris-Saclay

Thèse préparée dans l'unité de recherche **Centre Borelli** (ENS Paris-Saclay), UMR 9010 CNRS, sous la direction de **Agnès DESOLNEUX**, directrice de recherche à l'ENS Paris-Saclay, et de **Julie DELON**, professeure à l'université Paris-Cité.

Thèse soutenue à Paris-Saclay, le 6 décembre 2023, par

Antoine SALMONA

Composition du jury

Membres du jury avec voix délibérative

Gabriel PEYRE Directeur de recherche, ENS Ulm	Président
Jérémie BIGOT Professeur, Université de Bordeaux	Rapporteur & Examinateur
Laetitia CHAPEL Professeure , Institut Agro Rennes-Angers	Rapportrice & Examinatrice
David PICARD Professeur, École des Ponts ParisTech	Examinateur

THESE DE DOCTORAT
NNT : 2023UPASM033


école
normale
supérieure
paris–saclay


université
PARIS-SACLAY


CENTRE
BORELLI


Université
Paris Cité


LABORATOIRE
MAP5


FM
JH
FONDATION
MATHÉMATIQUE
JACQUES HADAMARD

Remerciements

Tout d'abord, j'aimerais remercier les membres de mon jury. Merci d'une part à Gabriel Peyré et David Picard, pour avoir accepté d'en faire partie, mais aussi à mes rapporteur.rice.s, Laetitia Chapel et Jérémie Bigot, pour leurs rapports détaillés sur mon manuscrit. Merci beaucoup de prendre le temps de vous intéresser à mes travaux.

Ensuite, merci infiniment à mes directrices de thèse, Julie et Agnès. Merci bien évidemment pour la grande qualité de l'encadrement sur le plan scientifique, sans qui cette thèse ne serait probablement même pas l'ombre d'elle-même. Merci pour votre investissement et pour votre aide à tous les niveaux, aussi bien sur les tâches de recherche les plus amusantes que les moins palpitantes. Mais merci aussi pour la qualité humaine de cet encadrement, qui m'a permis de travailler pendant 3 ans dans un environnement idéal et qui a fait que je ressortais de nos réunions toujours super motivé, même dans les moments ou la motivation était le moins au rendez-vous. Merci aussi à Valentin de Bortoli d'avoir participer à cet encadrement sur la partie modèles génératifs. Sans toi, cette thèse ne ressemblerait pas à ce qu'elle est aujourd'hui et je ne serais probablement jamais allé à la Nouvelle-Orléans. Merci beaucoup également à Lucía Bouza d'avoir réalisé les expériences et la démo Ipol de l'article sur la colorisation, cet article n'aurait probablement jamais été terminé sans toi.

J'aimerais aussi remercier tout le MAP5 pour l'ambiance bienveillante que l'on peut trouver au labo. Merci d'abord aux piliers du bureau 725-C1 comme Pierre-Louis, Anton et Zoé, puis aux deux Rémi et à Mariem qui s'en ont allé.es vivre de nouvelles aventures au 8ème, puis enfin à Loïc, Eloi, Guillaume, Herb, Alexander et Adélie, qui ont su combler le vide que les autres ont laissé. Merci aussi aux autres doctorants et post-doc du labo que j'ai eu la chance de côtoyer pendant ces 3 ans, comme Antoine, Mehdi, Sonia, Ivan, Charlie, Ariane, Chabane, Keanu, Ousmane, Florian, Thaïs, Leonard, Bernardin, Charles, Yen, Diala, et bien d'autres encore. Je vous souhaite à tous une bonne fin de thèse pour ceux qui ne l'ont pas déjà terminé, et une bonne continuation pour les autres.

Par ailleurs, j'aimerais aussi remercier mes amis d'avoir toujours été là dans les meilleurs comme dans les moins bons moments. Un immense merci à Mathis, Alban, Raphaël, Marc-Antoine, Luc, Jane, Mathilde, Clément, Olivier, Virgile, Lucile, Mathieu, Ella, Neige, et bien d'autres encore. Merci aussi à Céline d'être toujours là pour moi et de me soutenir dans tous mes projets. Merci également à mes amis de la musique, Slim et Julien, pour leur bienveillance ces derniers mois quand je ne pouvais pas autant m'investir qu'eux. Enfin, je remercie mes parents, mes grandes soeurs et mon grand frère pour leur soutien inconditionnel dans tous mes projets.

Contents

Introduction (French)

De nos jours, le domaine du traitement d'image est intrinsèquement lié aux statistiques et aux probabilités. En effet, l'omniprésence des approches par apprentissage a imposé un point de vue probabiliste de ce qu'est une image. Les images sont considérées comme des réalisations d'un vecteur aléatoire de grande dimension (une dimension par pixel) et les jeux de données sont pensés comme des distributions de probabilité empiriques. En ce sens, de nombreuses tâches en science de l'imagerie impliquent la comparaison plus ou moins directe de distributions de probabilité. C'est bien entendu le cas de nombreuses méthodes de *machine learning*, qui consistent grosso modo à ajuster un modèle paramétrique à un jeu de données fixé en minimisant directement ou indirectement une mesure de dissimilarité entre la distribution paramétrique du modèle et la distribution empirique du jeu de données.

Pour effectuer ces comparaisons entre distributions, certaines applications s'appuient sur la théorie du *transport optimal* qui fournit un cadre mathématique bien défini pour comparer les mesures entre elles. La théorie du transport optimal introduit la notion de *transport* : afin de transformer une mesure en une autre, il faut transporter localement la masse qui la compose à chaque point, conformément à la structure globale de la mesure cible. Un exemple classique illustrant cette notion est donné par le mathématicien Monge (1746-1818) : un ouvrier doit déplacer une grande pile de sable se trouvant sur un chantier de construction afin d'ériger une pile cible de forme désirée. Pour ce faire, il doit déplacer chaque grain de sable de manière à former une nouvelle pile de la forme désirée. Le transport est alors *optimal* s'il minimise un coût *global*, dans ce cas l'effort de l'ouvrier, en utilisant l'information *locale* du coût de transport d'un grain de sable d'un endroit à un autre.

En plus de comparer deux distributions, certaines tâches nécessitent également de transporter une distribution vers une autre. C'est notamment le cas des modèles génératifs qui sont devenus, au cours de la dernière décennie, l'un des sujets de recherche les plus populaires en traitement d'image, voire plus généralement en science des données. Informellement, l'objectif de la modélisation générative est de créer de nouvelles données en utilisant l'information d'un jeu de données fixé. Dans le contexte du traitement d'image, un modèle génératif vise à créer de nouvelles images *synthétiques* qui semblent appartenir à un jeu de données d'images *réelles*. En adoptant le point de vue probabiliste, l'objectif d'un modèle génératif est donc de créer de faux échantillons qui semblent avoir été tirés de la distribution empirique de l'ensemble d'images. Une approche générale pour résoudre cette tâche consiste à approcher la distribution empirique du jeu de données par une mesure paramétrique tout en transportant (pas nécessairement de manière optimale) une distribution simple facile à échantillonner vers cette dernière mesure paramétrique potentiellement très complexe.

Jusqu'à récemment, la plupart des recherches sur le transport de mesures présupposaient que les mesures impliquées évoluaient dans le même espace ambiant. L'avènement des modèles génératifs et l'introduction de distances de transport optimal qui restent pertinentes lorsque les mesures vivent dans des espaces incomparables ont simultanément amené l'idée que les mesures pouvaient également être transportées à travers des espaces qui ne sont pas directement comparables, comme des espaces euclidiens de dimensions différentes par exemple. Dans cette thèse, nous étudions trois problèmes liés au transport de mesures qui vivent dans des espaces euclidiens différents, les deux premiers étant dans le contexte du transport optimal et le dernier étant dans le contexte des modèles génératifs. Plus précisément, le but de cette thèse est triple :

(*i*) étudier le comportement des généralisations communes du transport optimal, dont celle dite de Gromov-Wasserstein, entre des distributions gaussiennes de dimensions différentes.

(*ii*) concevoir une distance de transport optimal entre des mélanges de gaussiennes de dimensions différentes.

(*iii*) étudier l'expressivité des modèles génératifs en relation avec la constante de Lipschitz de la fonction de transport.

Nous décrivons en détails ces trois problèmes ci-dessous.

Transport optimal entre mesures sur espaces euclidiens différents

L'objectif du transport optimal est de comparer des distributions de probabilités entre elles. Il fournit donc des outils mathématiques très utiles pour diverses tâches de traitement d'image (Haker and Tannenbaum, 2001; Rubner et al., 1998; Rabin et al., 2012, 2014) ou plus généralement d'apprentissage (Courty et al., 2016, 2018; Xu et al., 2018), par exemple en traitement du language (Kusner et al., 2015) ou encore pour les modèles génératifs (Arjovsky et al., 2017; Genevay et al., 2018; Tolstikhin et al., 2018). Le succès du transport optimal en science des données est principalement dû à sa capacité à établir des correspondances entre des nuages de points tout en induisant une distance géodésique entre les distributions de probabilité, connue sous le nom de distance de Wasserstein.

Dans son cadre classique, une présupposition implicite du transport optimal est que les deux distributions impliquées vivent dans le même espace ambiant, ou du moins que les deux espaces sont comparables, c'est-à-dire qu'il existe une fonction de coût pertinente pour les comparer. Cependant, cette hypothèse n'est pas forcément vérifiée pour de nombreuses applications. C'est le cas lorsqu'il s'agit de données structurées, telles que des graphes par exemple, ou lorsque les données proviennent de sources hétérogènes, comme en adaptation de domaine hétérogène (Wang and Mahadevan, 2011; Yeh et al., 2014; Liu et al., 2020). Un exemple concret de ce dernier problème est donné par Vayer (2020) : comment adapter un classificateur entraîné sur les images de taille 28×28 pixels provenant du jeu de données MNIST (LeCun et al., 1998) afin qu'il fonctionne bien avec les images de taille 16×16 du jeu de données USPS (Hull, 1994) ? De plus, d'autres tâches nécessitent la conception de fonctions de coût qui doivent être telles que le problème soit invariant par rapport à certaines familles de transformations, telles que les translations et les rotations par exemple, au sens où nous voulons que la distance entre une distribution donnée et une version translatée d'elle-même soit nulle. Même si les distributions impliquées dans ces applications peuvent résider dans le même espace ambiant, il n'est pas facile de concevoir une fonction de coût adéquate.

Pour surmonter ces limitations du transport optimal classique, plusieurs généralisations ont été proposées (Cohen and Guibasm, 1999; Pele and Taskar, 2013; Alvarez-Melis et al., 2019; Cai and Lim, 2022). La plupart d'entre elles impliquent de réaligner les deux mesures en envoyant l'une d'entre elles dans l'espace de l'autre. Alternativement, la généralisation la plus couramment utilisée est peut-être la distance de Gromov-Wasserstein (Mémoli, 2011), qui a récemment suscité un grand intérêt dans la littérature grâce à la flexibilité que cette approche offre. En effet, elle ne nécessite pas de spécifier au préalable un sous-ensemble d'invariances ni de concevoir une fonction de coût pertinente entre les espaces sur lesquels résident les distributions. Cette approche a été appliquée a plusieurs problèmes de d'apprentissage (Mémoli, 2009; Solomon et al., 2016; Alvarez-Melis and Jaakkola, 2018), notamment pour des données structurées (Vayer et al., 2019a; Brogat-Motte et al., 2022) ou encore pour les modèles génératifs (Bunne et al., 2019). Malgré le fait que la distance de Gromov-Wasserstein soit largement utilisée dans la littérature, sa compréhension théorique reste encore limitée. Son comportement sur des distributions 1D a été étudié par Vayer (2020), Beinert et al. (2022) et Dumont et al. (2022). En revanche, son comportement sur des distributions gaussiennes n'avait été que partiellement abordé dans Vayer (2020), et nous y remédions dans cette thèse.

Transport optimal en pratique

Le transport optimal est connu pour être un problème difficile à résoudre numériquement. Entre des distributions discrètes, son calcul implique de résoudre un *programme linéaire* (Dantzig, 1951) qui devient rapidement coûteux dès que le nombre de points est modérément élevé. Entre deux ensembles de n points, son calcul se fait en $O(n^3 \log(n))$ (Seguy et al., 2017), ce qui compromet son utilisation dans des applications impliquant plusieurs dizaines de milliers de points. Pour alléger le coût de calcul du transport optimal, de nombreux travaux ont développé des outils pour résoudre plus efficacement le problème. En particulier, Cuturi (2013) propose de résoudre un problème de transport optimal régularisé à l'aide de l'algorithme de Sinkhorn-Knopp (Sinkhorn and Knopp, 1967), réduisant ainsi le coût du problème à $O(n^2)$. En s'appuyant sur cette idée, diverses améliorations ont été développées pour résoudre le problème de transport optimal régularisé en temps quasi-linéaire (Altschuler et al., 2017, 2018, 2019; Forrow et al., 2019; Scetbon and Cuturi, 2020; Scetbon et al., 2021). Un autre type de méthodes introduites par Rabin et al. (2012) repose sur le fait que le problème de transport optimal entre des distributions 1D se réduit

à un simple problème de tri qui peut être résolu en $O(n\log(n))$. Ces méthodes consistent à calculer une infinité de projections linéaires des deux distributions qui vivent en grande dimension pour obtenir des représentations unidimensionnelles, puis de calculer une distance de Wasserstein moyenne entre ces représentations unidimensionnelles. Alternativement, Delon and Desolneux (2020) et Chen et al. (2018) ont proposé d'abord d'approcher les données par des *mélanges de gaussiennes*, puis de comparer les mélanges obtenus à l'aide d'une distance de transport optimal peu côuteuse en terme de calcul. Le principal avantage de cette dernière approche est que la complexité du problème de transport optimal obtenu ne dépend ni de la dimension, ni du nombre de points, mais seulement du nombre de composantes dans les mélanges, ce qui implique que le coût de calcul de cette approche dépend presque entièrement de la phase d'apprentissage des mélanges. Bien que cette méthode ne puisse probablement pas rivaliser avec les raffinements les plus récent de l'algorithme de Sinkhorn-Knopp en termes d'efficacité pure, elle fournit une distance de transport optimal particulièrement adaptée lorsqu'il existe déjà une sorte de structure de *clusters* dans les données.

Les généralisations du transport optimal à des mesures qui ne vivent pas dans le même espace sont connues pour être encore plus coûteuses en termes de calcul que le transport optimal classique. Par exemple, résoudre le problème de Gromov-Wasserstein implique de résoudre un problème d'optimisation quadratique qui est connu pour être un problème NP-difficile (Burkard et al., 1998). Comme pour le transport optimal classique, plusieurs travaux ont proposé des algorithmes plus rapides qui approchent la distance de Gromov-Wasserstein, reposant par exemple sur de la régularisation (Peyré et al., 2016; Scetbon et al., 2022). Dans cette thèse, nous proposons deux généralisations possibles de la distance proposée par Delon and Desolneux (2020) qui restent pertinentes lorsque les mélanges vivent dans des espaces de dimensions différentes. Nous illustrons l'utilisation de ces distances sur divers problèmes reliés à Gromov-Wasserstein.

Expressivité des modèles génératifs

Les modèles génératifs sont devenus ces dernières années l'un des sujets de recherche les plus populaires en science des données. Ils ont récemment attiré l'attention du grand public avec l'arrivée de plusieurs modèles à grande échelle tels que DALL-E 2 (Ramesh et al., 2022) ou Stable Diffusion (Rombach et al., 2022) qui inondent Internet d'images générées synthétiques. En science des données, les modèles génératifs ont été utilisés dans de nombreuses applications dans divers sous-domaines du *machine learning*, tels que l'augmentation de données (Sandfort et al., 2019; Antoniou et al., 2018), la résolution de problèmes inverses (Ravuri et al., 2021; Ledig et al., 2017) ou la traduction automatique (Isola et al., 2017; Yang et al., 2018). De nombreux modèles génératifs synthétisent des données en transformant un vecteur aléatoire suivant une loi normale multidimensionelle à l'aide d'une fonction déterministe, souvent modélisée par un réseau de neurones. C'est notamment le cas pour deux types de modèles très populaires, les *Variational Autoencoders* (VAEs) (Kingma and Welling, 2014) et les *Generative Adversarial Networks* (GANs) (Goodfellow et al., 2014). Ces modèles consistent donc à transporter (pas nécessairement de manière optimale) une distribution gaussienne vers une distribution complexe en grande dimension, en utilisant un réseau de neurones déterministe en tant que fonction de transport.

L'expressivité des réseaux de neurones profonds est un domaine de recherche très actif. Le théorème d'approximation universelle (Funahashi, 1989; Cybenko, 1989; Hornik et al., 1989) stipule que les réseaux de neurones sont des approximateurs universels, au sens où toute fonction peut théoriquement être approchée avec n'importe quelle précision par un réseau de neurones composé d'une seule couche mais avec potentiellement un nombre infini de neurones. Plus récemment, Hanin (2019) a montré que les réseaux de neurones possédant un nombre fini de neurones sur chaque couche mais avec potentiellement un nombre infini de couches pouvaient eux aussi approcher n'importe quelle fonction continue avec n'importe quelle précision, à condition d'avoir un nombre suffisant de neurones à chaque couche. En pratique, les réseaux de neurones qui ont à la fois un nombre fini de neurones et de couches semblent être beaucoup plus restreints en termes d'expressivité. Une restriction importante réside dans le fait que les réseaux de neurones sont la plupart du temps des applications lipschitziennes, car leurs fonctions d'activation sont généralement lipschitziennes. Cela vient principalement du fait que les réseaux de neurones doivent être différentiables presque partout pour être entraînés avec l'algorithme de *back-propagation* (Rumelhart et al., 1986). De plus, il a été largement observé dans la littérature que la constante de Lipschitz d'un réseau de neurones pouvait presque être utilisée comme une mesure de l'instabilité de son entraînement (Glorot and Bengio, 2010; Szegedy et al., 2013; Pennington et al., 2017). Outre les instabilités d'entraînement, il est également bien connu que les méthodes d'optimisation génériques telles que la descente de gradient

stochastique sont implicitement biaisées (Strand, 1974; Morgan and Bourlard, 1989; Gunasekar et al., 2018), dans le sens où elles ont tendance à converger vers des minimums particuliers. Récemment, Mulayoff et al. (2021) ont montré que lors de l'entraînement d'un réseau de neurones, l'algorithme de descente de gradient stochastique était biaisé vers des fonctions relativement régulières, peu importe l'initialisation. Ainsi, lorsque l'on tente d'approcher une fonction irrégulière avec une grande constante de Lipschitz en utilisant un réseau de neurones, en plus des instabilités d'entraînement, il est probable de converger vers un minimum local correspondant à une fonction plus régulière que la fonction cible.

L'analyse de l'expressivité des modèles génératifs par *deep learning* semble cependant être un domaine de recherche relativement nouveau. Plusieurs travaux se sont concentrés sur le cas où la distribution cible réside sur deux ou plusieurs variétés déconnectées (Khayatkhoei et al., 2018; Mehr et al., 2019; Tanielian et al., 2020). Khayatkhoei et al. (2018) fait la simple observation qu'une discontinuité dans la fonction de transport doit être introduite d'une manière ou d'une autre pour pouvoir transporter une distributions gaussienne vers une distribution qui réside sur plusieurs variétés déconnectées. Dans le contexte des *normalizing flows* (Rezende and Mohamed, 2015), il a été montré que la contrainte d'inversibilité limitait nécessairement l'expressivité du modèle. En effet, les auteurs de Cornish et al. (2020) ont montré que les distributions générées par les *normalizing flows* ont des supports nécessairement homéomorphes au support de la distribution latente. Par conséquent, la constante de Lipschitz du flux inverse doit être arbitrairement grande pour pouvoir approcher correctement les distributions résidant sur des variétés déconnectées (Cornish et al., 2020; Hagemann and Neumayer, 2021; Behrmann et al., 2021). Cependant, ce dernier résultat ne concerne que les réseaux de neurones inversibles et les distributions cibles vivant sur des variétés déconnectées. Dans cette thèse, nous étudions le cas plus général où la distribution cible est multimodale et la fonction de transport est n'importe quel réseau de neurones lipschitzien.

Contributions

Cette thèse couvre l'ensemble des travaux de l'auteur menés sur les axes de recherche sur le *transport optimal avec invariances entre mesures sur espaces euclidiens différents* et sur l'*expressivité des modèles génératifs push-forward*. Un travail supplémentaire de l'auteur (Salmona et al., 2022a) sur la colorisation d'images n'est pas inclus dans ce manuscrit. Au cours des trois années de doctorat qui ont été consacrées à la préparation de cette thèse, l'auteur a rédigé les articles scientifiques suivants :

> (Salmona et al., 2021). Antoine Salmona, Julie Delon et Agnès Desolneux. Gromov-Wasserstein distances between Gaussian distributions. *Journal of Applied Probability*[1], 2021.
>
> (Salmona et al., 2022b). Antoine Salmona, Valentin de Bortoli, Julie Delon et Agnès Desolneux. Can Push-forward Generative Models Fit Multimodal Distributions? *Advances in Neural Information Processing*[2], 2022.
>
> (Salmona et al., 2023). Antoine Salmona, Julie Delon et Agnès Desolneux. Gromov-Wassertein-like Distances in the Gaussian Mixture Models Space. Preprint, 2023.
>
> (Salmona et al., 2022a). Antoine Salmona, Lucía Bouza et Julie Delon. DeOldify: A Review and Implementation of an Automatic Colorization Method. *Image Processing On Line*[3], 2022.

Nous donnons des détails sur les contributions de chaque chapitre dans ce qui suit.

Chapitre 2

Ce chapitre introduit les fondements mathématiques de la théorie du transport optimal. Nous présentons également brièvement les solveurs numériques les plus couramment utilisés dans la littérature pour résoudre le problème du transport optimal en pratique. Pour deux distributions de probabilité μ et ν, respectivement sur des espaces $\mathcal{X}$ et $\mathcal{Y}$, et étant donnée une fonction $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ appelée *coût*, le transport optimal, dans sa forme la plus classique, vise à résoudre le problème d'optimisation suivant :

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\pi(x,y) \ ,$$

[1] https://www.cambridge.org/core/journals/journal-of-applied-probability/article/
[2] https://proceedings.neurips.cc/paper-files/paper/2022
[3] https://www.ipol.im/pub/art/2022

où $\Pi(\mu,\nu)$ est l'ensemble des mesures sur $\mathcal{X}\times\mathcal{Y}$ de marginales μ et ν. Lorsque $\mathcal{X}$ et $\mathcal{Y}$ sont égaux et euclidiens, le choix du coût $c_p(x,y)=\|x-y\|^p$, avec $p\geq 1$ et où $\|.\|$ est la norme euclidienne, induit une distance entre les distributions de probabilité qui ont leur moment d'ordre p fini appelée la *distance de Wasserstein* W_p.

Chapitre 3

Dans ce chapitre, nous introduisons la généralisation la plus commune du transport optimal entre mesures qui vivent dans des espaces non comparables, c'est-à-dire lorsque il n'existe pas de manière évidente de concevoir une fonction de coût $c:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$ pour comparer les espaces $\mathcal{X}$ et $\mathcal{Y}$. Il s'agit de la *distance de Gromov-Wasserstein* (Mémoli, 2011). Nous présentons également les solveurs numériques les plus couramment utilisés dans la littérature pour résoudre ce problème. Entre deux distributions μ et ν, respectivement sur des espaces $\mathcal{X}$ et $\mathcal{Y}$, la distance de Gromov-Wasserstein d'ordre $p\geq 1$ s'écrit :

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

où $c_{\mathcal{X}}:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ et $c_{\mathcal{Y}}:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ sont deux fonctions mesurables également appelées *coûts*. GW_p définit une pseudométrique sur l'espace des "réseaux mesurés" (Chowdhury and Mémoli, 2019), c'est-à-dire les triplets de la forme $(\mathcal{X},c_{\mathcal{X}},\mu)$. Nous introduisons également les autres formulations récentes du transport optimal entre mesures qui vivent dans des espaces euclidiens de dimensions différentes proposées par (Alvarez-Melis et al., 2019) et (Cai and Lim, 2022). Nous définissons une nouvelle formulation que nous appelons EW_2, pour *Embedded Wasserstein distance*. Entre deux mesures vivant respectivement dans $\mathbb{R}^d$ et $\mathbb{R}^{d'}$, elle s'ecrit de la manière suivante,

$$EW_2(\mu,\nu)=\inf\left\{\inf_{\phi\in\mathrm{Isom}_{d'}(\mathbb{R}^d)}W_2(\mu,\phi_{\#}\nu),\inf_{\psi\in\mathrm{Isom}_d(\mathbb{R}^{d'})}W_2(\psi_{\#}\mu,\nu)\right\},$$

où, pour $r\geq 1$ et $s\geq 1$, $\mathrm{Isom}_s(\mathbb{R}^r)$ est l'ensemble des isométries de $\mathbb{R}^s$ à valeur dans $\mathbb{R}^r$. Nous montrons que cela définit une pseudométrique sur l'espace de mesures de dimensions arbitraires avec des moments d'ordre 2 finis.

Chapitre 4

Dans ce chapitre, qui est principalement une reproduction de (Salmona et al., 2021), nous étudions le comportement de la distance de Gromov-Wasserstein d'ordre 2 entre deux distributions gaussiennes $\mu=\mathrm{N}(m_0,\Sigma_0)$ et $\nu=\mathrm{N}(m_1,\Sigma_1)$ qui vivent respectivement dans $\mathbb{R}^d$ et $\mathbb{R}^{d'}$, avec d' potentiellement non égal à d. Nous étudions notamment les cas où $c_{\mathcal{X}}$ et $c_{\mathcal{Y}}$ sont, soit les distances euclidiennes au carré sur $\mathbb{R}^d$ et $\mathbb{R}^{d'}$, soit les produits scalaires sur $\mathbb{R}^d$ et $\mathbb{R}^{d'}$. Tout d'abord, nous commenccons par étudier le cas des distances euclidiennes au carré. En utilisant un résultat technique de Vayer (2020), nous montrons que le problème GW_2 avec coûts quadratiques, que nous appelons (GW_2-Q), admet une formulation probabiliste équivalente[4], qui s'écrit de la manière suivante :

$$\sup_{X\sim T_{0\#}\mu,Y\sim T_{1\#}\nu}\sum_{i,j}\mathrm{Cov}(X_i^2,Y_j^2)+2\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2\,, \tag{1.1}$$

où $X=(X_1,X_2,\ldots,X_d)^T$, $Y=(Y_1,Y_2,\ldots,Y_{d'})^T$, $\|.\|_{\mathcal{F}}$ est la norme de Frobenius, $T_0:x\mapsto P_0^T(x-m_0)$ et $T_1:y\mapsto P_1^T(y-m_1)$, et où (P_0,D_0) et (P_1,D_1) sont les diagonalisations respectives de Σ_0 et Σ_1 qui trient les valeurs propres dans l'ordre décroissant. Cette formulation met en évidence que le problème (GW_2-Q) est difficile à résoudre sans autres hypothèses sur le plan de transport π, car résoudre ce problème implique de connaître la règle probabiliste qui lie les co-moments d'ordre 4 aux co-moments d'ordre 2 de π. Ainsi, nous dérivons d'abord une borne inférieure sur (GW_2-Q) en optimisant séparément les deux termes de (1.1). Ensuite, nous dérivons une borne supérieure en restreignant le problème à des plans de transport qui sont eux-mêmes gaussiens. Dans ce cas, la règle qui lie les co-moments d'ordre 4 aux co-moments d'ordre 2 de π est donnée par le théorème d'Isserlis (Isserlis, 1918). Il en découle que le problème restreint GW_2, que nous appelons (GW_2-QG), est équivalent au problème suivant :

$$\sup_{X\sim T_{0\#}\mu,Y\sim T_{1\#}\nu}\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2\,. \tag{1.2}$$

[4]Deux problèmes d'optimisation sont équivalents si les solutions de l'un sont directement obtenues à partir des solutions de l'autre, et inversement.

Nous montrons ensuite que (GW_2-QG) admet des solutions analytiques de la forme $(\mathrm{Id}_d, T)_{\#}\mu$ avec T affine tel que pour tout $x \in \mathbb{R}^d$:

$$T(x) = m_1 + P_1 \begin{pmatrix} \widetilde{\mathrm{Id}}_{d'} D_1^{\frac{1}{2}} {D_0^{(d')}}^{-\frac{1}{2}} & 0_{d',d-d'} \end{pmatrix} P_0^T (x - m_0) \ ,$$

où $D_0^{(d')}$ est la matrice de taille $d' \times d'$ formée avec les d' premières lignes et colonnes de D_0, et $\widetilde{\mathrm{Id}}_{d'}$ est n'importe quelle matrice de la forme $\mathrm{diag}((\pm 1)_{1 \leq i \leq d'})$. Nous montrons que ces solutions sont liées avec l'Analyse en Composantes Principales (ACP). Entre deux mesures centrées $\bar{\mu}$ et $\bar{\nu}$, nous montrons ensuite que les solutions décrites ci-dessus sont également des solutions du problème de Gromov-Wasserstein pour le choix de produits scalaires comme fonctions de coût (GW_2-IP), puisque ce dernier problème est également équivalent au problème (1.2). La distance de Gromov-Wasserstein admet alors une expression simple dans ce cas :

$$GW_2^2(\langle . \rangle_d, \langle . \rangle_{d'}, \bar{\mu}, \bar{\nu}) = \|\Sigma_0\|_{\mathcal{F}}^2 + \|\Sigma_1\|_{\mathcal{F}}^2 - 2\mathrm{tr}(D_0^{(d')} D_1) \ .$$

Nous comparons ensuite aux autres formulations du transport optimal entre les mesures qui vivent dans des espaces non comparables introduites dans le chapitre précédent. Nous montrons que les solutions présentées ci-dessus sont également solutions du problème de l'*Embedded Wasserstein* et des problèmes étudiés par Alvarez-Melis et al. (2019). Enfin, nous montrons que la distance de transport optimal proposée par Cai and Lim (2022) a un comportement différent des autres formulations étudiées dans ce chapitre.

Chapitre 5

Ce chapitre, qui est principalement une reproduction de (Salmona et al., 2023), propose deux généralisations de type Gromov de la distance entre mélanges de gaussiennes proposée par Delon and Desolneux (2020). Plus précisément, Delon and Desolneux (2020) ont proposé la distance suivante, qui consiste à restreindre le problème de Wasserstein aux plans de transport qui sont eux-mêmes des mélanges de gaussiennes,

$$MW_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu) \cap GMM_\infty(\mathbb{R}^{2d})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) \ ,$$

où $GMM_\infty(\mathbb{R}^{2d})$ est l'ensemble des mélanges de gaussiennes sur $\mathbb{R}^d$ avec un nombre fini de composantes. Un résultat clé de Delon and Desolneux (2020) est que MW_2 peut être réécrit comme un problème de transport optimal discret à petite échelle. Plus précisement, entre deux mélanges $\mu = \sum_k^K a_k \mu_k$ et $\nu = \sum_l^L b_l \nu_l$ sur $\mathbb{R}^d$,

$$MW_2^2(\mu, \nu) = \inf_{\omega \in \Pi(a,b)} \sum_{k,l} \omega_{k,l} W_2^2(\mu_k, \nu_l) \ ,$$

où $a = (a_1, \ldots, a_K)^T$ et $b = (b_1, \ldots, b_L)^T$. Cette dernière formulation rend MW_2 facilement calculable en pratique, car la distance W_2 entre gaussiennes a une formule analytique simple. Les plans optimaux ω^* et π^* sont ensuite liés par la relation suivante, pour tout $x, y \in \mathbb{R}^d$:

$$\pi^*(x, y) = \sum_{k,l} \omega_{k,l}^* p_{\mu_k}(x) \delta_{y = T_{W_2}^{k,l}(x)} \ , \tag{1.3}$$

où p_{μ_k} est la densité de μ_k et $T_{W_2}^{k,l}$ est le transport qui minimise W_2 entre μ_k et ν_l. Dans ce chapitre, nous proposons une première généralisation de type Gromov-Wasserstein de MW_2 que nous appelons MGW_2 qui s'écrit, entre deux mélanges $\mu = \sum_{k=1}^K a_k \mu_k$ et $\nu = \sum_{l=1}^L b_l \nu_l$ respectivement sur $\mathbb{R}^d$ et $\mathbb{R}^{d'}$,

$$MGW_2^2(\mu, \nu) = \inf_{\omega \in \Pi(a,b)} \sum_{k,l,i,j} |W_2^2(\mu_k, \mu_i) - W_2^2(\nu_l, \nu_j)|^2 \omega_{k,l} \omega_{i,j} \ .$$

Nous montrons que cela définit une pseudométrique sur l'ensemble des mélanges de gaussiennes avec un nombre fini de composantes et de dimension arbitraire. Cependant, cette distance n'admet pas, à notre connaissance, une formulation continue équivalente simple comme c'était le cas pour MW_2. Par conséquent, la dérivation d'un plan de transport entre deux nuages de points avec MGW_2 n'est pas évidente. Une solution pourrait être de définir un plan π^* par analogie avec MW_2, en utilisant une formule similaire à (1.3). Cependant, cela impliquerait de connaître la transformation isométrique qui a

été implicitement appliquée à l'une des deux mesures lors du calcul de la distance. C'est pourquoi nous introduisons une autre généralisation de MW_2, que nous appelons MEW_2, qui s'exprime comme suit :

$$MEW_2^2(\mu,\nu) = \inf\left\{ \inf_{\phi\in \mathrm{Isom}_{d'}(\mathbb{R}^d)} MW_2(\mu,\phi_\#\nu), \inf_{\psi\in \mathrm{Isom}_{d}(\mathbb{R}^{d'})} MW_2(\psi_\#\mu,\nu)\right\} .$$

Contrairement à MGW_2, cette formulation permet de dériver directement un plan de transport entre deux nuages de points car elle explicite la transformation isométrique. Si l'on suppose sans perte de généralité que $d \geq d'$, un plan π^* optimal pour MEW_2 est alors obtenu en modifiant $T^{k,l}_{W_2}$ dans (1.3) en $\phi^{-1*} \circ T^{k,l}_{W_2}$, où ϕ^{-1*} est l'inverse de la transformation isométrique ϕ^* restreint à $\phi^*(\mathbb{R}^{d'})$. Nous concevons ensuite un plan de transport π pour MGW_2 par analogie avec MEW_2. Enfin, nous illustrons l'utilisation pratique de MGW_2 et MEW_2 sur des problèmes à moyenne et grande échelle tels que la correspondance de formes et le transfert de couleur sur des images hyperspectrales.

Chapitre 6

Ce chapitre introduit les concepts de base des modèles génératifs et présente les principaux modèles utilisés en traitement d'image. Nous mettons en évidence l'existence de deux types principaux de modèles génératifs que nous appelons respectivement les modèles *push-forward* et les modèles *push-forward indirects*.

Dans les modèles *push-forward*, la distribution générée ν_θ est de la forme $\nu_\theta = g_{\theta\#}\mu_{d'}$, où $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$ est une loi normale multidimensionelle et g_θ est un réseau de neurones déterministe. Dans les modèles "push-forward indirects", la distribution générée ν_θ est de la forme $\nu_\theta = G_{\theta\#}\mu_{d(K+1)}$ mais cette fois-ci G_θ est une fonction déterministe qui représente K itérations d'une dynamique de Monte-Carlo. La distribution latente $\mu_{d(K+1)}$ correspond à la concaténation de tous les bruits gaussiens ajoutés pendant la dynamique. Les principales différences par rapport aux modèles *push-forward* sont que l'optimisation n'est pas directement effectuée sur la correspondance *push-forward* G_θ elle-même, mais sur une fonction auxiliaire, et que l'espace latent est de dimension bien supérieure à celle de l'espace ambiant.

Chapitre 7

Dans ce chapitre, qui reprend en grande partie Salmona et al. (2022b), nous étudions l'expressivité des modèles *push-forward* par rapport à la constante de Lipschitz du réseau de neurones utilisé pendant le processus de génération. Plus précisément, nous montrons que pour n'importe quelle fonction lipschitzienne $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ et pour n'importe quel borelien A de $\mathbb{R}^d$,

$$\mathrm{Lip}(g)(g_\#\mu_{d'})^+(\partial\mathsf{A}) \geq \varphi(\Phi^{-1}(g_\#\mu_{d'}(\mathsf{A}))) , \tag{1.4}$$

où $(g_\#\mu_{d'})^+(\partial\mathsf{A})$ désigne la $(g_\#\mu_{d'})$-*surface* de la frontière de A, qui peut informellement être vue comme une mesure de la masse que $g_\#\mu_{d'}$ a sur l'hyper-surface $\partial\mathsf{A}$, où $\varphi(x) = (2\pi)^{-1/2}\exp[-x^2/2]$ est la densité de $\mathrm{N}(0,1)$, et $\Phi(x) = \int_{-\infty}^{x}\varphi(t)\mathrm{d}t$. Ce résultat est principalement une conséquence de l'*inégalité isopérimétrique gaussienne* (Sudakov and Tsirelson, 1978), qui implique que pour n'importe quel borelien A de $\mathbb{R}^d$,

$$\mu_{d'}^+(\partial\mathsf{A}) \geq \varphi(\Phi^{-1}(\mu_{d'}(\mathsf{A}))) .$$

L'inégalité (1.4) peut être utilisée d'abord pour déterminer une borne inférieure sur la constante de Lipschitz des fonctions g qui transportent $\mu_{d'}$ vers une distribution donnée ν. Par exemple, lorsque $\nu = \lambda\mathrm{N}(m_1, \sigma^2\,\mathrm{Id}_d) + (1-\lambda)\mathrm{N}(m_2, \sigma^2\,\mathrm{Id}_d)$ est un mélange de deux gaussiennes avec $m_1, m_2 \in \mathbb{R}^d$, $\sigma > 0$ et $\lambda \in (0,1)$, on peut montrer à partir de (1.4) que les fonctions g qui transportent $\mu_{d'}$ vers ν vérifient nécessairement

$$\mathrm{Lip}(g) \geq \sigma\exp\left[\|m_2 - m_1\|^2/(8\sigma^2) - (\Phi^{-1}(\lambda))^2/2\right] .$$

Cela illustre que lorsque ν est multimodale, les fonctions g qui transportent $\mu_{d'}$ vers ν ont nécessairement de grandes constantes de Lipschitz. Deuxièmement, nous utilisons (1.4) pour dériver des bornes inférieures sur la distance en variation totale et la divergence de Kullback-Leibler entre, d'une part la mesure $g_\#\mu_{d'}$, et d'autre part une distribution cible ν fixée. Étant donné que contraindre les constantes de Lipschitz des réseaux de neurones est un moyen courant de stabiliser les différents modèles, cela met en évidence qu'il existe un compromis entre la capacité des modèles *push-forward* à approcher des distributions multimodales et la stabilité de leur entraînement. Nous validons nos résultats sur des images et des données 1D, et nous montrons empiriquement que les modèles à diffusion récemment introduits par Song and Ermon (2019) et Ho et al. (2020) ne semblent pas souffrir de telles limitations.

Notations

We define in the following some of the notations that will be used throughout the thesis.

Linear algebra

- $\langle x, x' \rangle_d$ stands for the Euclidean inner product in $\mathbb{R}^d$ between x and x'. We will denote $\langle x, x' \rangle$ when there is no ambiguity about the dimension.
- $\|x\|_{\mathbb{R}^d}$ stands for the Euclidean norm of $x \in \mathbb{R}^d$. We will denote $\|x\|$ when there is no ambiguity about the dimension.
- $\mathrm{rk}(M)$ stands for the rank of a matrix M.
- the notation $\mathrm{tr}(M)$ denotes the trace of a matrix M.
- $\|M\|_{\mathcal{F}}$ stands for the Frobenius norm of a matrix M, i.e. $\|M\|_{\mathcal{F}} = \sqrt{\mathrm{tr}(M^T M)}$.
- $\|M\|_*$ stands for the nuclear norm of a matrix M, i.e. $\|M\|_* = \mathrm{tr}((M^T M)^{\frac{1}{2}})$.
- the notation $\boldsymbol{\sigma}(M)$ denotes the vector of singular values of the matrix M.
- Id_d is the identity matrix of size $d \times d$.
- $\widetilde{\mathrm{Id}}_d$ stands for any matrix of size $d \times d$ of the form $\mathrm{diag}((\pm 1)_{1 \leq i \leq d})$
- Suppose $d \geq d'$. For any matrix M of size $d \times d$, we denote $M^{(d')}$ the submatrix of size $d' \times d'$ containing the d' first rows and the d' first columns of A.
- Let $r \leq d$ and $s \leq d'$. For any matrix M of size $r \times s$, we denote $M^{[d,d']}$ the matrix of size of the form $\begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix}$. When $d = d'$, we will denote $M^{[d]}$.
- For any $x \in \mathbb{R}^d$, $\mathrm{diag}(x)$ denotes the matrix of size $d \times d$ with diagonal vector x.
- We denote $\mathbb{S}^d$ the set of symmetric matrices of size $d \times d$, $\mathbb{S}^d_+$ the set of symmetric positive semi-definite matrices, and $\mathbb{S}^d_{++}$ the set of symmetric positive definite matrices.
- $\mathbb{1}_{d',d} = (1)_{\substack{1 \leq i \leq d' \\ 1 \leq j \leq d}}$ denotes the matrix of ones with d' rows and d columns.
- $\mathbb{O}(\mathbb{R}^d)$ denotes the set of orthogonal matrices of size $d \times d$.
- $\mathbb{V}_{d'}(\mathbb{R}^d)$ denotes the Stiefel manifold, i.e the set of rectangular matrices P of size $d \times d'$ such that $P^T P = \mathrm{Id}_{d'}$.

Measure theory

- $\mathsf{A} \cup \mathsf{B}$ denotes the union of the sets A and B. When the two sets don't intersect, we will denote $\mathsf{A} \sqcup \mathsf{B}$.
- The notation $X \sim \mu$ means that X is a random variable with probability distribution μ.
- If μ is a positive measure on $\mathcal{X}$ and $\phi : \mathcal{X} \to \mathcal{Y}$ is a mapping $\phi_{\#}\mu$ stands for the push-forward measure of μ by T, i.e. the measure on $\mathcal{Y}$ such that for any measurable set A of $\mathcal{Y}$, $\phi_{\#}\mu(\mathsf{A}) = \mu(\phi^{-1}(\mathsf{A}))$.
- If μ is a positive measure on $\mathcal{X}$, $\mathrm{supp}(\mu)$ denotes its support, i.e. the subset of $\mathcal{X}$ defined as $\mathrm{supp}(\mu) = \{x \in \mathcal{X} \mid \text{for all open set } N_x \text{ such that } x \in N_x,\ \mu(N_x) > 0\}$.
- If X and Y are random vectors on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, we denote $\mathrm{Cov}(X, Y)$ the matrix of size $d \times d'$ of the form $\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T\right]$.
- For any positive measure μ, we denote $\bar{\mu}$ its associated centered measure, i.e. the measure such that if $X \sim \mu$, we have $X - \mathbb{E}_{X \sim \mu}[X] \sim \bar{\mu}$.

- For any $m \in \mathbb{R}^d$ and any $\Sigma \in \mathbb{S}_+^d$, we denote $\mathrm{N}(m, \Sigma)$ the Gaussian measure of mean m and covariance matrix Σ.
- For $x \in \mathcal{X}$, δ_x denotes the Dirac distribution at x.
- $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel sigma-field on $\mathbb{R}^d$.
- $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions on $\mathcal{X}$
- For $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, $\Pi(\mu, \nu)$ denotes the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.
- Δ_n denotes the probability simplex of $\mathbb{R}^n$, i.e. $\Delta_n = \{a \in \mathbb{R}_+^n \,:\, \sum_{k=1}^{n} a_k = 1\}$.

Operators

- $\oslash$ denotes the entrywise division.
- $\otimes$ denotes the tensor-matrix product.
- $\otimes_K$ denotes the Kronecker product, i.e. if A is a matrix of size $p \times q$ of the form $A = [a_{i,j}]_{\substack{1 \le i \le p \\ 1 \le j \le q}}$ and B is a matrix of size $r \times s$, $A \otimes_K B$ is the matrix of size $pr \times qs$ of the form,

$$\begin{pmatrix} a_{1,1}B & \dots & a_{1,q}B \\ \vdots & \ddots & \vdots \\ a_{p,1}B & \dots & a_{p,q}B \end{pmatrix} .$$

- $\oplus_K$ denotes the Kronecker sum, i.e. if A is a matrix of size $r \times r$ and B is a matrix of size $s \times s$, $A \oplus_K B$ is the matrix of size $rs \times rs$ of the form $A \otimes_K \mathrm{Id}_s + \mathrm{Id}_r \otimes_K B$.
- For any mapping $T : \mathbb{R}^{d'} \to \mathbb{R}^d$, $J[T](x) \in \mathbb{R}^{d \times d'}$ denotes the Jacobian matrix of T at $x \in \mathbb{R}^d$, i.e. the matrix obtained by stacking the gradients in x of each coordinate of T.

Introduction

Contents

Nowadays, modern imaging science is intrinsically linked to statistics and probability. Indeed, the ubiquity of machine learning approaches has imposed a probabilistic point of view of what an image is. Images are thought as realizations of a highly-dimensional random vector (one dimension per pixel) and datasets are seen as empirical probability distributions. To that extent, many imaging science tasks involve more or less direct comparisons of probability distributions at some point. This is of course the case for numerous machine learning methods that roughly consist in fitting a parametric model to a given dataset by minimizing directly or indirectly a dissimilarity measure between the parametric distribution of the model and the empirical distribution of the dataset.

To perform these comparisons between distributions, some applications rely on the theory of Optimal Transport (OT) that provides a well-defined mathematical framework to compare measures. The theory of optimal transport introduces the notion of *transport*: in order to transform a measure into another, one must locally transport the mass that composes it at each point accordingly to retrieve the global structure of the target measure. A classic example that illustrates this notion is given by the French mathematican Monge (1746-1818): a worker has to move a large pile of sand lying on a construction site in order to erect a target pile of a desired shape. To do so, he must move each grain of sand in the right way to form a new pile of the desired shape. The transport is then *optimal* if it minimizes a *global* cost, in this case the effort of the worker, which is done using the *local* information of the cost of transporting one grain of sand from one location to another.

In addition to comparing two distributions, some tasks require also transporting one distribution towards another. This is notably the case of Generative Modeling (GM) that has become over the last decade one of the most popular research topic in imaging science, or more generally in data science. Informally, the goal of generative modeling is to create new data using the information of a given dataset. In the context of imaging science, a generative model aims at creating new *synthetic* images that seems to belong to a given dataset of *real* images. Retrieving the probablistic point of view, the goal of a generative model is thus to create fake samples that seem to have been drawn from the empirical distribution of the image dataset. A general approach to solve this task is to approximate the empirical distribution of the dataset by a parametric measure while transporting (not necessarily in an optimal manner) a simple easy-to-sample distribution towards this latter potentially highly complex parametric measure.

Until recently, most research on transport of measures presupposed that the measures involved were living in the same ground space. The breakthrough of generative modeling and the introduction of OT distances that stay meaningful when the measures involved live in incomparable spaces have at the same time brought to light the idea that measures could also be transported across spaces that are not directly comparable, as for instance Euclidean spaces of different dimensions. In this thesis, we study three problems related to the transport of measures lying on different Euclidean spaces, the first two being in the context of optimal transport and the last one being in the context of generative modeling. More precisely, the purpose of this thesis is threefold:

(i) studying the behavior of the common generalizations of optimal transport, including the so-called Gromov-Wasserstein distance, between Gaussian distributions in incomparable spaces.

(*ii*) designing a computationally efficient and scalable OT distance between Gaussian mixtures possibly living in different Euclidean spaces.

(*iii*) studying the expressivity of generative models relatively to the Lipschitz constant of the mapping which transports the easy-to-sample distribution towards the complex target distribution.

Before going into the details of this thesis, we give more details about these three problems in what follows.

1.1 Optimal transport between measures on different Euclidean spaces

The goal of optimal transport theory is to design meaningful ways to compare probability distributions. It provides thus very useful mathematical tools for diverse imaging sciences and machine learning tasks including image registration (Haker and Tannenbaum, 2001), image retrieval (Rubner et al., 1998), image processing (Rabin et al., 2012, 2014), domain adaptation (Courty et al., 2016), embedding learning (Courty et al., 2018; Xu et al., 2018), natural language processing (Kusner et al., 2015) and generative modeling (Arjovsky et al., 2017; Genevay et al., 2018; Tolstikhin et al., 2018). The sucess of optimal transport in data science is mainly due to its ability to draw correspondences between sets of points while inducing a geodesic distance between probability distributions, known as the Wasserstein distance.

In its classic setting, an implicit prerequisite of optimal transport is that the two distributions involved lie on the same ground space, or at least that the two spaces are comparable, i.e. there exists a relevant cost function to compare them. However, this assumption may not hold for many applications. This is often the case when dealing with structured data, as graphs for instance, or when the data come from heterogeneous sources, as in the case of heterogeneous domain adaptation (Wang and Mahadevan, 2011; Yeh et al., 2014; Liu et al., 2020). An illustrative concrete application of this latter problem is given by Vayer (2020): how to adapt a classifier trained on the 28×28 digit images of the MNIST dataset (LeCun et al., 1998) in order that it works well on the 16×16 digit images of the USPS dataset (Hull, 1994)? Moreover, some other tasks such as shape matching or word embedding require designing cost functions such that the problem is invariant to some families of invariances, such as translations and rotations for example, in the sense that we want the distance between a given distribution and a translated and rotated version of itself to be zero. Even if the distributions involved in these applications may live in the same ground space, it is not straightforward to design an adequate cost function.

To overcome these limitations of classic optimal transport, several generalizations have been proposed (Cohen and Guibasm, 1999; Pele and Taskar, 2013; Alvarez-Melis et al., 2019; Cai and Lim, 2022). Most of them involve realigning the two measures by sending one of them into the space of the other. Alternatively, the most commonly used generalization is perhaps the Gromov-Wasserstein distance (Mémoli, 2011) which has recently received high interest thanks to the flexibility this approach offers. Indeed, it only requires modeling topological aspects of the distributions within each domain to compare them without having to specify first a subset of invariances nor to design a relevant cost function between the spaces the distributions lie on. This approach has been applied to shape matching (Mémoli, 2009) or more generally to correspondence problems (Solomon et al., 2016), word embedding (Alvarez-Melis and Jaakkola, 2018), graph classification (Vayer et al., 2019a), graph prediction (Brogat-Motte et al., 2022), and generative modeling (Bunne et al., 2019). Despite the fact that the Gromov-Wasserstein distance is widely used in the literature, its theoretical understanding remains still nascent. Its behavior on one-dimensional distributions has been studied by Vayer (2020), Beinert et al. (2022) and Dumont et al. (2022). However, its behavior on Gaussian distributions had only been partially studied in Vayer (2020), and we remedy this in this thesis.

1.2 Optimal transport in practice

Optimal transport is known to be a computationally challenging problem. Between discrete distributions, its computation involves solving a *Linear Program* (LP) (Dantzig, 1951) that rapidly becomes costly as soon as the number of points is moderately large. Between two sets of n points, its computation is done in $O(n^3 \log(n))$ (Seguy et al., 2017), which compromises its usability for settings with more than a few tens of thousand of points. To lighten OT computational cost, a large number of works have developped efficient computational tools in order to solve OT problems. In particular, Cuturi (2013)

proposes to solve a regularized OT problem using the Sinkhorn-Knopp algorithm (Sinkhorn and Knopp, 1967), reducing the cost of the problem to $O(n^2)$. Building on this idea, various refinements have been developed to solve the regularized OT problem in near-linear time (Altschuler et al., 2017, 2018, 2019; Forrow et al., 2019; Scetbon and Cuturi, 2020; Scetbon et al., 2021). Another type of methods introduced by Rabin et al. (2012) and known as sliced methods, rely on the fact that the optimal transport problem between one-dimensional distributions reduces itself to a simple sorting problem which can be solved in $O(n\log(n))$. These consist in computing infinitely many linear projections of the high-dimensional distributions to one-dimensional representations and then computing an average Wasserstein distance between these one-dimensional representations. Alternatively, Delon and Desolneux (2020) and Chen et al. (2018) have proposed to first approximate the data by *Gaussian Mixture Models* (GMMs), and to compare the obtained GMMs using a computationally effective composite OT distance. The main benefit of this latter approach is that the complexity of the composite OT problem does not depend of the dimension nor of the number of points but only of the number of components in the GMMs, implying that the computational cost of this approach comes almost entirely from the fitting of the GMMs. Although this method probably does not compete with the fastest recent refinements of the Sinkhorn-Knopp algorithm in terms of pure computational cost, it provides a relatively scalable and computationally effective OT distance that is particularly suited when there already exists a kind of clustering structure in the data.

The generalizations of optimal transport to measures that are not living in the same ground space are known to be even more computationally costly than classic optimal transport. For instance, solving the Gromov-Wasserstein problem involves solving a *Quadratic Assignment Problem* (QAP) which is known to be a NP-hard problem (Burkard et al., 1998). As for classic OT, several works have proposed faster algorithms that approximate the Gromov-Wasserstein distance, building for instance either on regularization (Peyré et al., 2016; Scetbon et al., 2022) or on sliced mechanisms (Vayer et al., 2019b). In this thesis, we propose two possible generalizations of the distance proposed by Delon and Desolneux (2020) that stay relevant when the GMMs are living in spaces of different dimensions and we show that these OT distances can be used to solve relatively efficiently Gromov-Wasserstein related tasks.

1.3 Expressivity of generative models

Generative modeling has become over the last years one of the most popular research topics in imaging science and more generally in data science. It has recently caught the general audience's attention with the arrival of several large-scale models such as DALL-E 2 (Ramesh et al., 2022), or Stable Diffusion (Rombach et al., 2022) that flood the internet with synthetic generated images. In the data science community, generative models have been used in numerous applications in various machine learning subfields, such as data augmentation (Sandfort et al., 2019; Antoniou et al., 2018), solving inverse problems (Ravuri et al., 2021; Ledig et al., 2017) or machine translation (Isola et al., 2017; Yang et al., 2018). Many generative models synthesize data by transforming a standard Gaussian random variable using a deterministic mapping, often modelized by a neural network. This is notably the case for two very popular types of models, the *Variational Autoencoders* (VAEs) (Kingma and Welling, 2014) and the *Generative Adversarial Networks* (GANs) (Goodfellow et al., 2014). Such models consists thus in transporting (not necessarily in an optimal manner) a standard Gaussian distribution towards a highly complex high-dimensional distribution, using a deterministic neural network as transport map.

The expressivity of deep neural networks is an active research field on its own. The universal approximation theorem (Funahashi, 1989; Cybenko, 1989; Hornik et al., 1989) states that shallow neural networks are universal approximators, in the sense that any mapping can theoretically be approximated with any precision by a neural network composed of one single layer but with a pottentially infinite number of neurons. More recently, Hanin (2019) has shown that deep neural networks with finite numbers of neurons on each layer but with a potentially infinite number of layers could approximate any continuous mapping with any precision as long as it has a sufficient number of neurons at each layer. In practice, deep neural networks with finite number of neurons and layers seems to be a lot more restrained in terms of expressivity. An important restriction is that finite deep neural networks are most of the time Lipschitz mappings by design, since their activation functions are generally Lipschitz. This is mainly due to the fact that deep neural networks have to be differentiable almost everywhere in order to be trained using the backpropagation algorithm (Rumelhart et al., 1986). More critically, it has been widely observed in the literature that the Lipschitz constant of a neural network could almost be used as a measurement of the instability of its training (Glorot and Bengio, 2010; Szegedy et al., 2013; Pennington et al., 2017). Beside training instabilities, it is also well known that generic optimization methods such as Stochastic Gradient

Descent (SGD) are implicity biased (Strand, 1974; Morgan and Bourlard, 1989; Gunasekar et al., 2018) in the sense that they tends to converge to particular minima. Recently, Mulayoff et al. (2021) have shown that when training a neural network, SGD was biased towards relatively regular functions, regardless of the initialization. Thus when trying to approximate an irregular function with large Lipschitz constant with a neural network, in addition of training instabilites, it is likely to converge towards a local minimum corresponding to a more regular function than the target one.

Analysing the expressivity of deep generative models seems to be however a relatively new field of research. Several works have focused on the case where the target distribution lies on two or more disconnected manifolds (Khayatkhoei et al., 2018; Mehr et al., 2019; Tanielian et al., 2020). Khayatkhoei et al. (2018) has made the simple observation that a discontinuity in the transport mapping must be somehow introduced in order to be able to transport correctly a Gaussian distributions towards a distribution which lies on disconnected manifolds. In the context of normalizing flows (Rezende and Mohamed, 2015), it has been shown that the invertibility constraint limits the expressivity of the model. Indeed, Cornish et al. (2020) show that distributions generated by invertible normalizing flows have a support which is necessarily homeomorphic to the support of the latent distribution. As an outcome, the Lipschitz constant of the inverse flow has to approach infinity to correctly approximate distributions lying on disconnected manifolds (Cornish et al., 2020; Hagemann and Neumayer, 2021; Behrmann et al., 2021). However, this latter result concerns only invertible neural networks and disconnected target distributions. In this thesis, we study the more general case where the target distribution is multimodal and the transport map is any Lipschitz neural network.

1.4 Contributions

This thesis covers all the author's work conducted on the lines of research of *optimal transport with invariances between measures possibly on different Euclidean spaces* and *expressivity of deep push-forward generative models*. Additional work (Salmona et al., 2022a) of the author on image colorization is not included in this manuscript. During the three years of doctoral studies that went into the preparation of this thesis, the author has written the following scientific papers:

(Salmona et al., 2021). Antoine Salmona, Julie Delon and Agnès Desolneux. Gromov-Wasserstein distances between Gaussian distributions. *Journal of Applied Probability*[5], 2021.

(Salmona et al., 2022b). Antoine Salmona, Valentin de Bortoli, Julie Delon and Agnès Desolneux. Can Push-forward Generative Models Fit Multimodal Distributions? *Advances in Neural Information Processing*[6], 2022.

(Salmona et al., 2023). Antoine Salmona, Julie Delon and Agnès Desolneux. Gromov-Wassertein-like Distances in the Gaussian Mixture Models Space. Preprint, 2023.

(Salmona et al., 2022a). Antoine Salmona, Lucía Bouza and Julie Delon. DeOldify: A Review and Implementation of an Automatic Colorization Method. *Image Processing On Line*[7], 2022.

We give details on the contributions of each chapter in what follows.

Chapter 2

This chapter introduces the mathematical foundations of optimal transport theory. We also briefly introduce the common numerical solvers that are used in the literature to solve optimal transport problems in practice. For two probability distributions μ and ν respectively on some spaces $\mathcal{X}$ and $\mathcal{Y}$, and given a function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ called *cost*, optimal transport in its most classic form, aims at solving the following optimization problem,

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\pi(x,y) \ ,$$

where $\Pi(\mu,\nu)$ is the set of measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. When $\mathcal{X}$ and $\mathcal{Y}$ are equal and Euclidean spaces, the choice of cost $c_p(x,y) = \|x-y\|^p$, with $p \geq 1$ and $\|.\|$ being the Euclidean norm induces a metric between probability distributions with finite p-th moments, called the *Wasserstein distance* W_p.

[5] https://www.cambridge.org/core/journals/journal-of-applied-probability/article/
[6] https://proceedings.neurips.cc/paper-files/paper/2022
[7] https://www.ipol.im/pub/art/2022

Chapter 3

In this chapter, we introduce the common generalization of optimal transport to measures living in incomparable spaces, i.e. when it is not straightfoward to design a meaningful cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, which is the *Gromov-Wasserstein distance* (Mémoli, 2011). We also introduce the common numerical solvers that are used in the literature to solve this problem. Between two distributions μ and ν, respectively on some spaces $\mathcal{X}$ and $\mathcal{Y}$, the Gromov-Wasserstein distance of order $p \geq 1$ reads as

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

where $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are two measurable functions also called *costs*. GW_p defines a pseudometric on the space of network measure spaces Chowdhury and Mémoli (2019), i.e. the triplets of the form $(\mathcal{X}, c_{\mathcal{X}}, \mu)$. We also introduce the other recent formulations of optimal transport to measures living in Euclidean spaces of different dimensions proposed by Alvarez-Melis et al. (2019) and Cai and Lim (2022). Building on these latters, we define a new formulation that we call EW_2 for *embedded Wasserstein distance*. Between two measures living respectively in $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, this reads as

$$EW_2(\mu, \nu) = \inf \left\{ \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} W_2(\mu, \phi_{\#}\nu), \inf_{\psi \in \mathrm{Isom}_d(\mathbb{R}^{d'})} W_2(\psi_{\#}\mu, \nu) \right\} ,$$

where for $r \geq 1$ and $s \geq 1$, $\mathrm{Isom}_s(\mathbb{R}^r)$ denotes the set of isometries from $\mathbb{R}^s$ to $\mathbb{R}^r$. We show that this defines a pseudometric on the space of measures of abitrary dimensions with finite order 2 moments.

Chapter 4

In this chapter, which is mostly a reproduction of (Salmona et al., 2021), we study the behavior of the Gromov-Wasserstein distance of order 2 between two Gaussian distributions $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$ living respectively in $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ with d' possibly not equal to d. We focus on the cases where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are either the squared Euclidean distances on respectively $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ or the inner products on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. First, we start by studying the squared Euclidean case. Building on a technical result of Vayer (2020), we show that the GW_2 problem with quadratic costs, that we call (GW_2-Q), admits an equivalent[8] probabilistic formulation which reads as,

$$\sup_{X \sim T_{0\#}\mu, Y \sim T_{1\#}\nu} \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + 2\|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2 , \tag{1.5}$$

where $X = (X_1, X_2, \ldots, X_d)^T$, $Y = (Y_1, Y_2, \ldots, Y_{d'})^T$, $\|.\|_{\mathcal{F}}$ is the Frobenius norm, and where $T_0 : x \mapsto P_0^T(x - m_0)$ and $T_1 : y \mapsto P_1^T(y - m_1)$ where (P_0, D_0) and (P_1, D_1) are the respective diagonalizations of Σ_0 and Σ_1 that sort the eigenvalues in non-increasing order. This formulation highlights that the (GW_2-Q) problem is hard to solve without further assumptions on the coupling π, because it would requires to know the probabilistic rule that links the co-moments of order 4 to the co-moments of order 2 of π. Thus, we derive first a lower bound on (GW_2-Q) by optimizing the two terms of (1.5) separately. Then, we derive an upper bound by constraining the set of admissible couplings to transportation plans that are themselves Gaussian. In that case, the rule that links the co-moments of order 4 to the co-moments of order 2 of π is given by the Isserlis theorem (Isserlis, 1918). It follows, that the restricted GW_2 problem, that we call (GW_2-QG), is equivalent to the following problem,

$$\sup_{X \sim T_{0\#}\mu, Y \sim T_{1\#}\nu} \|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2 . \tag{1.6}$$

We show then that (GW_2-QG) admits some closed forms solutions of the form $(\mathrm{Id}_d, T)_{\#}\mu$ with T affine such that for all $x \in \mathbb{R}^d$,

$$T(x) = m_1 + P_1 \begin{pmatrix} \widetilde{\mathrm{Id}}_{d'} D_1^{\frac{1}{2}} D_0^{(d')^{-\frac{1}{2}}} & 0_{d',d-d'} \end{pmatrix} P_0^T(x - m_0) ,$$

where $D_0^{(d')}$ is the matrix of size $d' \times d'$ that is formed with the d' first rows and columns of D_0, and $\widetilde{\mathrm{Id}}_{d'}$ is any matrix of the form $\mathrm{diag}((\pm 1)_{1 \leq i \leq d'})$. We show that these solutions share close connections

[8] We say that two optimization problems are equivalent if the solutions of one are readily obtained from the solutions of the other, and vice-versa.

with *Principal Component Analysis* (PCA). Between centered measures $\bar{\mu}$ and $\bar{\nu}$, we then show that the solutions described above are also solutions of the Gromov-Wasserstein problem for the choice of inner-product as cost functions (GW_2-IP), since this latter problem is also equivalent to Problem (1.6). The Gromov-Wasserstein distance has then a nice closed form expression in that case:

$$GW_2^2(\langle . \rangle_d, \langle . \rangle_{d'}, \bar{\mu}, \bar{\nu}) = \|\Sigma_0\|_{\mathcal{F}}^2 + \|\Sigma_1\|_{\mathcal{F}}^2 - 2\mathrm{tr}(D_0^{(d')} D_1) \ .$$

We then compare to the other formulations of optimal transport between measures on incomparable spaces introduced in the previous chapter. We show that the solutions presented above are also solutions of the embedded Wasserstein problem and of the problems studied by Alvarez-Melis et al. (2019). Finally, we show that the OT distance proposed by Cai and Lim (2022) admits a different behavior than the other formulations studied in this chapter.

Chapter 5

This chapter, which is mostly a reproduction of (Salmona et al., 2023), proposes two Gromov-type generalizations of the distance between GMMs proposed by Delon and Desolneux (2020). More precisely, Delon and Desolneux (2020) have proposed the so-called *Mixture Wasserstein* distance (MW) between GMMs, by restricting the set of admissible couplings in the Wasserstein distance to transportation plans that are temselves GMMs, i.e.

$$MW_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu) \cap GMM_\infty(\mathbb{R}^{2d})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) \ ,$$

where $GMM_\infty(\mathbb{R}^{2d})$ is the set of all finite Gaussian mixtures on $\mathbb{R}^d$. One key result of Delon and Desolneux (2020) is that MW_2 can be rewritten as a small-scale discrete optimal transport problem. Between two GMMs $\mu = \sum_k^K a_k \mu_k$ and $\nu = \sum_l^L b_l \nu_l$ on $\mathbb{R}^d$, this reads as,

$$MW_2^2(\mu, \nu) = \inf_{\omega \in \Pi(a,b)} \sum_{k,l} \omega_{k,l} W_2^2(\mu_k, \nu_l) \ ,$$

where $a = (a_1, \dots, a_K)^T$ and $b = (b_1, \dots, b_L)^T$. This latter formulation makes MW_2 easily computable in practice, since the W_2 distance between Gaussian distributions has a simple closed form. The optimal plans ω^* and π^* are then linked by the following relation for all $x, y \in \mathbb{R}^d$

$$\pi^*(x, y) = \sum_{k,l} \omega_{k,l}^* p_{\mu_k}(x) \delta_{y = T_{W_2}^{k,l}(x)} \ , \tag{1.7}$$

where p_{μ_k} is the density of μ_k and $T_{W_2}^{k,l}$ is the W_2 transport map between μ_k and ν_l. In this chapter, we propose a first Gromov generalization of MW_2 that we call MGW_2 for *Mixture Gromov Wasserstein* distance, that is defined between two GMMs $\mu = \sum_{k=1}^K a_k \mu_k$ and $\nu = \sum_{l=1}^L b_l \nu_l$, respectively on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, as

$$MGW_2^2(\mu, \nu) = \inf_{\omega \in \Pi(a,b)} \sum_{k,l,i,j} |W_2^2(\mu_k, \mu_i) - W_2^2(\nu_l, \nu_j)|^2 \omega_{k,l} \omega_{i,j} \ .$$

We show that this defines a pseudometric on the spaces of all finite GMMs in any dimension. However, this OT distance doesn't admit, to the best of our knowledge, a simple equivalent continuous formulation as this was the case for MW_2. As an outcome, the derivation of an assignment between clouds of points with MGW_2 is not straightfoward. A possible solution could be to define a plan π^* by analogy with MW_2, using a similar formula to (1.7). Yet, this would imply to know the isometric transformation that have been implicitly applied to one of the two measures during the derivation of the distance. Thus is why we introduce another generalization of MW_2 that we call MEW_2 for *Mixture Embedded Wasserstein* distance, that reads as

$$MEW_2^2(\mu, \nu) = \inf \left\{ \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} MW_2(\mu, \phi_\# \nu), \inf_{\psi \in \mathrm{Isom}_d(\mathbb{R}^{d'})} MW_2(\psi_\# \mu, \nu) \right\} \ .$$

As opposed to MGW_2, this formulation allows to derive directly an assignment between clouds of points because it expicits the isometric transformation. If we suppose without any loss of generality that $d \geq d'$, an optimal plan π^* for MEW_2 is thus obtained by replacing $T_{W_2}^{k,l}$ in (1.7) by $\phi^{-1*} \circ T_{W_2}^{k,l}$, where ϕ^{-1*} is the inverse of the optimal ϕ^* restricted to $\phi^*(\mathbb{R}^{d'})$. We design then an assignment for MGW_2 by analogy with MEW_2. Finally, we illustrate the pratical use of MGW_2 and MEW_2 on medium-to-large scale problems such as shape matching and hyperspectral image color transfer.

Chapter 6

This chapter introduces the basic concepts of generative modeling as well as the most commonly used generative models in imaging science. We highlight that there exists two main types of generative models that we call respectively *push-forward models* and *indirect push-forward models*. In push-forward models, the generated distribution ν_θ is of the form $\nu_\theta = g_{\theta\#}\mu_{d'}$ with $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$ is the standard Gaussian distribution and g_θ is a deterministic neural network. In indirect push-forward models, the generated distribution ν_θ is of the form $\nu_\theta = G_{\theta\#}\mu_{d(K+1)}$, but this time G_θ is a deterministic mapping that corresponds to K iterations of a Monte-Carlo dynamics. The latent distribution $\mu_{d(K+1)}$ corresponds to the concatenation of all Gaussian noises added during the dynamics. The main differences with push-forward models are that optimization is not directly performed on the push-forward mapping G_θ itself but on an auxilary function, and that the latent space is of much more larger dimension than the ambient space.

Chapter 7

In this chapter, which is mostly a reproduction of Salmona et al. (2022b), we study the expressivity of push-forward models relatively to the Lipschitz constant of the neural network that is used for generation. More precisely, we show that for any Lipschitz mapping $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ and for any Borel set A of $\mathbb{R}^d$,

$$\mathrm{Lip}(g)(g_\#\mu_{d'})^+(\partial\mathsf{A}) \geq \varphi(\Phi^{-1}(g_\#\mu_{d'}(\mathsf{A}))) \;, \tag{1.8}$$

where $(g_\#\mu_{d'})^+(\partial\mathsf{A})$ denotes the $(g_\#\mu_{d'})$-*surface area* of the border of A which is informally a measure of the mass that $g_\#\mu_{d'}$ has on the hypersurface $\partial\mathsf{A}$, and $\varphi(x) = (2\pi)^{-1/2}\exp[-x^2/2]$ is the density function of $\mathrm{N}(0,1)$, and $\Phi(x) = \int_{-\infty}^{x}\varphi(t)\mathrm{d}t$. This result is mainly a consequence of the *Gaussian isoperimetric inequality* (Sudakov and Tsirelson, 1978), that states that for any Borel set A of $\mathbb{R}^d$,

$$\mu_{d'}^+(\partial\mathsf{A}) \geq \varphi(\Phi^{-1}(\mu_{d'}(\mathsf{A}))) \;.$$

Inequality (1.8) can be used first to determine a Lower bound on the Lipschitz constant of the mappings g that push $\mu_{d'}$ into a given distribution ν. For instance, when $\nu = \lambda\mathrm{N}(m_1, \sigma^2\,\mathrm{Id}_d) + (1-\lambda)\mathrm{N}(m_2, \sigma^2\,\mathrm{Id}_d)$ is a bimodal Gaussian mixture with $m_1, m_2 \in \mathbb{R}^d$, $\sigma > 0$ and $\lambda \in (0,1)$, one can show from (1.8) that the mappings g that pushes $\mu_{d'}$ into ν necessarily verify

$$\mathrm{Lip}(g) \geq \sigma\exp\left[\|m_2 - m_1\|^2/(8\sigma^2) - (\Phi^{-1}(\lambda))^2/2\right] \;.$$

This illustrates that when ν is multimodal, the mappings g that push $\mu_{d'}$ into ν have necessarily large Lipschitz constants. Secondly, we use (1.8) to derive lower bounds on the total variation distance and the Kullback-Leibler divergence between the push-forward measure $g_\#\mu_{d'}$ and a given target distribution ν for a given mapping g such that the Lipschitz constant of g is not large enough for reaching ν. Since constraining the Lipschitz constants of neural networks is a common way to stabilize generative models, this highlights that there is a trade-off between the ability of push-forward models to approximate multimodal distributions and the stability of their training. We validate our findings on one-dimensional and image datasets and empirically show that the recently introduced diffusion models (Song and Ermon, 2019; Ho et al., 2020) do not suffer of such limitation.

1.5 Outline of the thesis

The rest of this thesis is divided in two parts. Part I covers all the author's work on *optimal transport with invariances between measures possibly on different Euclidean spaces* and is organized as follow.

In Chapter 2, we expose the mathematical background of optimal transport. We present the fundamental concepts and results of classic optimal transport theory as well as the most commonly used numerical solvers in the literature.

In Chapter 3, We introduce the common generalization of optimal transport to measures that live in incomparable spaces, i.e. the Gromov-Wasserstein distance. We also introduce the common numerical solvers used to solve this problem. In the second part of the chapter, we introduce the other formulations that have been recently proposed in the literature, and we define a new formulation that we call embedded Wasserstein distance. This chapter contains some result of Salmona et al. (2023).

In Chapter 4, we study the behavior of the Gromov-Wasserstein distance between Gaussian distributions, for both choices of squared Euclidean distances and inner-products as cost functions. We also compare with the other formulations presented in the previous chapter. This chapter is mostly based on the work (Salmona et al., 2021) but also contains some results of Salmona et al. (2023).

In Chapter 5, we introduce two Gromov-Wasserstein related OT distances between GMMs possibly living in different dimensions and we show that both can be used to relatively efficiently solve Gromov-Wasserstein-related tasks. This chapter is based on the work (Salmona et al., 2023).

Part II covers all the author's work on the *expressivity of deep push-forward generative models* and is organized as follow.

In Chapter 6, we expose the basic concepts of generative modeling, and we introduce the most commonly used generative models in imaging science. We highlight that there exist two main categories of generative models that we call push-forward generative models and indirect push-forward generative model.

In Chapter 7, we study the expressivity of push-forward generative models relatively to the Lipchitz constant of the generative network. We show that for push-forward generative models, there exists a trade-off between their expressivity and the stability of their training. We also empirically show that indirect push-forward models seem not to suffer of the same limitation. This chapter is mostly based on the work (Salmona et al., 2022b).

Part I

Optimal transport with invariances between measures possibly on different Euclidean spaces

Chapter 2

Generalities about optimal tranport

Contents

In this chapter, we present in short the classic concepts and results of the OT theory. We also briefly present the different classic solvers that exist in the literature. We refer to Santambrogio (2015) for a complete general reference on OT theory, Villani (2008) for a more mathematically oriented reference and Peyré and Cuturi (2019) for a numerically oriented reference.

2.1 The classic optimal transport problem

Before presenting the foundations of OT theory, we shortly introduce some important mathematical notions that we will use throughout the thesis.

Polish spaces. We say that a space $\mathcal{X}$ is Polish if it is a separable complete metrizable space, i.e. if it contains a countable dense subset (separability) and it can be endowed with a metric $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that $(\mathcal{X}, d_{\mathcal{X}})$ is complete, in the sense that every Cauchy sequence[1] in $\mathcal{X}$ converges in $\mathcal{X}$. A basic example of Polish space is $\mathbb{R}^d$ with the usual Euclidean metric for any $d \geq 1$. This relatively general notion is the only prerequisite on the ground space to be able to develop the theory of optimal transport. Note that in all the thesis, we say that a function $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a "metric" or a "distance" when it verifies all the properties of a metric (symmetry, separability, non-negativity, triangle inequality, and finiteness), whereas we use the terms "discrepancies" or "OT distances" in a less rigorous manner to qualify functions (issued from the OT theory) which act as "measures of dissimilarity" between probability distributions but not necessarily verify all the axioms of a metric.

Measures and histograms. Let $\mathcal{X}$ be a Polish space, we write $\mathcal{P}(\mathcal{X})$ the set of Borel probability measures on $\mathcal{X}$. This set includes both continuous and discrete probability measures. Discrete probability measures can be written as $\Sigma_i^n a_k \delta_{x_k}$, with δ_{x_k} being the Dirac distribution at position $x_k \in \mathcal{X}$ and $a = (a_1, \ldots, a_n)^T \in \mathbb{R}_+^n$ being an histogram, i.e. an element of the *probability simplex* of $\mathbb{R}^n$

$$\Delta_n = \{a \in \mathbb{R}_+^n \ : \ \sum_{k=1}^{n} a_k = 1\} \ .$$

To avoid degeneracy issues where locations with no mass are accounted for, we will assume when considering discrete probability measures that the elements of a are all positive.

[1] A Cauchy sequence $\{x_k\}_{k\in\mathbb{N}}$ of $(\mathcal{X}, d_{\mathcal{X}})$ is a sequence such that for any $\varepsilon > 0$, there exists a positive integer K such that for any $k, k' > K$, $d_{\mathcal{X}}(x_k, x'_k) \leq \varepsilon$.

2.1.1 The Monge problem

The OT problem has been historically introduced by Gaspard Monge in 1781. It can be described as the following least effort problem: given two probability distributions μ and ν, how can we transport the mass of μ towards ν so that the overall effort of transferring this mass is minimized? To formalize this problem, we need to introduce the notions of *push-forward measure* and *cost* which respectively translate the notions of *transport* and *effort*.

Push-forward measure. Let μ be a Borel probability measure on a space $\mathcal{X}$ and let $T : \mathcal{X} \to \mathcal{Y}$ be a mapping between $\mathcal{X}$ and another space $\mathcal{Y}$. We call *push-forward measure* and we denote $T_{\#}\mu$ the probability measure defined such that for any Borel set A of $\mathcal{Y}$, $T_{\#}\mu(\mathsf{A}) = \mu(T^{-1}(\mathsf{A}))$. When $\nu = T_{\#}\mu$, we say that T *pushes* μ into ν. Finally, when μ is a discrete measure of the form $\sum_{k=1}^{m} a_k \delta_{x_k}$, the push-forward measure $T_{\#}\mu$ is of the form $\sum_{k=1}^{m} a_k \delta_{T(x_k)}$.

Cost functions and matrices. Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. A cost function can be any positive lower semi-continuous mapping $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. When $\mathcal{Y} = \mathcal{X}$, $d_{\mathcal{X}}^p$ with $p \geq 1$ is a classic example of cost function. Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ and given a m-tuple $\{x_k\}_k^m$ of elements of $\mathcal{X}$ and a n-tuple $\{y_l\}_l^n$ of elements of $\mathcal{Y}$, one can construct a cost matrix $C \in \mathbb{R}_+^{m \times n}$ as $C = (c(x_k, y_l))_{k,l}$. Thus, given two sets of respectively m and n points, a cost matrix can be any positive matrix of size $m \times n$.

The Monge problem. Now we are ready to introduce the Monge problem (Monge, 1781). Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. Given a *source* measure $\mu \in \mathcal{P}(\mathcal{X})$ and a *target* measure $\nu \in \mathcal{P}(\mathcal{Y})$, and given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, it aims at solving the following optimization problem,

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

Thus, we want to find a mapping T which pushes μ into ν while minimizing a *global* cost defined as the continuous sum of all the local costs corresponding to the cost of transporting the mass at position x towards $T(x)$. When T is solution of Problem (MP), we say that T is a *Monge map* and we denote it T_{OT}.

The Monge problem between discrete measures. When both measures μ and ν are discrete and of the form $\sum_{k=1}^{m} a_k \delta_{x_k}$ and $\sum_{l=1}^{n} b_l \delta_{y_l}$, the Monge problem seeks a map that associates to each point x_k a single point y_l and which transports the mass of μ toward the mass of ν. Thus, by mass conservation, the push-forward condition translates into

$$\text{for all } 1 \leq l \leq n \quad \sum_{k : T(x_k) = y_l} a_k = b_l \; .$$

It is also possible to encode T using indices $\sigma : [\![1, m]\!] \to [\![1, n]\!]$ so the mass conservation is written

$$\text{for all } 1 \leq l \leq n, \quad \sum_{k \in \sigma^{-1}(l)} a_k = b_l \; ,$$

where $\sigma^{-1}(l)$ is the preimage set of l. In the special case where $m = n$ and both distributions are uniform, i.e. for any k and l, $a_k = b_l = \frac{1}{n}$, the Monge problem can then be rewritten in an equivalent way as an *optimal assignment problem*,

$$\min_{\sigma \in \mathrm{Perm}(n)} \frac{1}{n} \sum_{k=1}^{n} C_{k,\sigma(k)} \; ,$$

where $C \in \mathbb{R}^{n \times n}$ is a given cost matrix and $\mathrm{Perm}(n)$ is the set of all permutations of $[\![1, n]\!]$. If σ^* is solution of this latter problem, the map $T : \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_n\}$ such that $T(x_k) = y_{\sigma(k)}$ for all $1 \leq k \leq m$ is then solution of the Monge problem. Note that the optimal assignment problem may have several optimal solutions as it is the case for instance if all the points are equidistant. In contrast, when $m \neq n$, they may not exists any feasible[2] map T. This happens when the histograms a and b are not compatible, which is always the case when the target measure has more points than the source measure, i.e. when $m < n$.

[2] A feasible element is any element of the set on which optimization is performed.

2.1.2 Kantorovich relaxation

The Monge problem is not always relevant to studying discrete measures, such as those found in practical problems since there exists cases where the set of feasible mappings T is empty. Moreover this latter set is *non-convex* as well as $\mathrm{Perm}(n)$. Therefore, the Monge problem and the optimal assignment problems are non-convex optimization problems which makes them difficult to solve. One major breakthrough in OT theory is due to Russian mathematician Leonid Kantorovich in 1942. The key idea of Kantorovich is to relax the deterministic nature of the mapping which sends the mass at position x to a given target position $T(x)$. Thus, Kantorovich's formulation allows to split the mass at position x to several target positions. This flexibility is encoded using, instead of a push-forward mapping that pushes μ into ν, a *coupling* measure which associates the two measures.

Coupling measures and matrices. Let $\mathcal{X}$ and $\mathcal{Y}$ be two spaces and let μ and ν be two measures on $\mathcal{X}$ and $\mathcal{Y}$. We call coupling any measure π on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν, i.e. such that $P_{\mathcal{X}\#}\pi = \mu$ and $P_{\mathcal{Y}\#}\pi = \nu$, where for every couple (x, y) in $\mathcal{X} \times \mathcal{Y}$, $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$ are the projections on respectively $\mathcal{X}$ and $\mathcal{Y}$. We will denote $\Pi(\mu, \nu)$ the set of couplings associated with μ and ν. When these latter measures are discrete, $\mu = \sum_{k=1}^{m} a_k \delta_{x_k}$ and $\sum_{l=1}^{n} b_l \delta_{y_l}$ the coupling measure is also discrete and of the form

$$\pi = \sum_{k,l} \omega_{k,l} \delta_{(x_k, y_l)} \ , \tag{2.1}$$

and it is possible to parametrize the set $\Pi(\mu, \nu)$ as the set of measures of the form (2.1) and such that for all $1 \leq k \leq m$ and all $1 \leq l \leq n$, $\sum_k \omega_{k,l} = b_l$ and $\sum_l \omega_{k,l} = a_k$. We will call coupling matrix any $\omega \in \mathbb{R}^{m \times n}$ such that its coefficients verify these latter conditions and we will denote $\Pi(a, b)$ the set of coupling matrices associated with histograms a and b. Note that the constraints on the coefficients are often rewritten under the compact form of the two following constraints: $\omega \mathbb{1}_n = a$ and $\omega^T \mathbb{1}_m = b$, where $\mathbb{1}_n = (1)_{1 \leq k \leq n}$.

Kantorovich formulation. The Kantorovich problem (Kantorovich, 1942) is the modern classic formulation of the OT problem. Given two measures μ and ν on two Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ and given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, it aims at solving the following optimization problem

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

The resulting cost, potentially infinite without further assumptions, corresponds to the minimal cost of pushing μ into ν while authorizing to split the mass μ has at a given position x. The transport is done according the optimal coupling measure π that we call then the *optimal transport plan* betwen μ and ν. Unlike the set of maps which push μ into ν, the set $\Pi(\mu, \nu)$ is always non-empty since it at least contains the product measure $\mu \otimes \nu$, and so Problem (KP) is always feasible.

Kantorovich formulation between discrete measures. Between discrete distributions $\sum_{k=1}^{m} a_k \delta_{x_k}$ and $\sum_{l}^{n} b_l \delta_{y_l}$, the Kantorovich problem translates into

$$\inf_{\omega \in \Pi(a,b)} \sum_{k,l} C_{k,l} \omega_{k,l} \ , \tag{2.2}$$

where $\omega = (\omega_{k,l})_{k,l}$ is a matrix of size $m \times n$. This latter problem is a Linear Program (Dantzig, 1951) with possibly more than one optimal solution. A good illustration of this latter problem is given by the following *resource allocation problem* (Hitchcock, 1941): suppose that an operator runs m warehouses and n factories. Each warehouse is indexed with an integer k and contains a quantity a_k of a given ressource that is needed to run properly the factories. Each factory is indexed with an integer l and must possess a quantity b_l of that ressource to run properly. The operator wants to transport the ressource from the different warehouses to the different factories. To do so, he will use a transportation company which charges $a \times C_{k,l}$ to move a quantity a of ressource from the k-th warehouse to the l-th factory (the price is proportional to the quantity transported). In order to minimize its total cost, the operator can solve Problem (2.2). Indeed, solving this latter problem provides a coupling matrix ω of size $m \times n$ such that for all $1 \leq k \leq m$ and all $1 \leq l \leq n$, $\omega_{k,l}$ indicates the quantity of ressource to transport from the k-th warehouse to l-th factory to minimize the total cost.

Relationship between Kantorovich and Monge problems. Solving the Kantorovich problem can provide a solution of the Monge problem. For instance, in the case where the measure are uniform discrete distributions both composed of n points, solving the Kantorovich problem provides a coupling matrix ω^* such that $n \times \omega^*$ is a permutation matrix[3] of $[\![1, n]\!]$ that minimizes the optimal assignment problem, see Peyré and Cuturi (2019) for more details. More generally, if there is an optimal coupling solution of Problem (KP) of the form $\pi = (\mathrm{Id}_{\mathcal{X}}, T)_{\#}\mu$ with $\mathrm{Id}_{\mathcal{X}}$ being the identity operator on $\mathcal{X}$ and $T : \mathcal{X} \to \mathcal{Y}$ being any deterministic mapping which pushes μ into ν, then T is solution of Problem (MP).

Existence of solutions. When $\mathcal{X}$ and $\mathcal{Y}$ are compact metric spaces, one can show relatively easily - see Santambrogio (2015) for details - that $\Pi(\mu, \nu)$ is compact. One can show then that Problem (KP) admits at least one solution: because c is lower semi-continuous, the functional $\pi \mapsto \int c(x, y)\mathrm{d}\pi(x, y)$ is also lower semi-continuous and so we can directly use the Weierstrass theorem which states that a lower semi-continuous function reaches its infinimum on a compact set. When $\mathcal{X}$ and $\mathcal{Y}$ are not compact but Polish, it is still possible to show that $\Pi(\mu, \nu)$ is compact but we need to use more advanced tools of measure theory. Indeed, it can be shown that any sequence in $\Pi(\mu, \nu)$ is tight[4] and then one can deduce that $\Pi(\mu, \nu)$ is compact using the Prokhorov theorem which states that the condition of every sequence of $\Pi(\mu, \nu)$ being tight is equivalent to its compactness. In a nutshell, the Kantorovich problem always admits a solution and the infinimum in (KP) can be replaced by a minimum.

Convexity. In addition to being compact, $\Pi(\mu, \nu)$ is also a convex set. Indeed, for any π and π' in $\Pi(\mu, \nu)$, observe that every linear combinaison of the form $t\pi + (1 - t)\pi'$ with $t \in [0, 1]$ is also in $\Pi(\mu, \nu)$ since the projection operators are linear:

$$P_{\mathcal{X}\#}(t\pi + (1 - t)\pi') = tP_{\mathcal{X}\#}\pi + (1 - t)P_{\mathcal{X}\#}\pi' = t\mu + (1 - t)\mu = \mu \ ,$$

and the same goes for $P_{\mathcal{Y}}$. Therefore the Kantorovich problem is a linear optimization problem under convex constraints and so all the tools of convex optimization, in particular duality, can be used to solve Problem (KP).

Probabilistic interpretation. Kantorovich's problem can be reinterpreted through the prism of random variables. Indeed, Problem (KP) is equivalent to

$$\inf_{X \sim \mu, Y \sim \nu} \mathbb{E}[c(X, Y)] \ , \tag{2.3}$$

where the notation $X \sim \mu$ means that X is a random variable with probability μ. The law of the couple (X, Y) is then $\pi \in \Pi(\mu, \nu)$.

2.1.3 Dual formulation

The Kantorovich problem (KP) is a convex optimzation problem with constraints. Therefore it can be naturally paired with a so-called dual problem, which is a constrained concave maximization problem. By strong duality, this dual problem admits the same optimal value that the primal problem (KP). Let $\mathcal{C}_b(\mathcal{X})$ denotes the set of continuous bounded functions from $\mathcal{X}$ to $\mathbb{R}$.

Dual Kantorovich problem. The dual problem of Problem (KP) can be expressed as follow:

$$\sup_{(f,g) \in \mathcal{R}(c)} \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\nu(y) \ , \tag{DP}$$

where

$$\mathcal{R}(c) = \{(f, g) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}) \ : \ \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}, \ f(x) + g(y) \leq c(x, y)\} \ .$$

The dual variable f and g are often referred to as *Kantorovich potentials*. The derivation of this dual problem is not trivial and requires an important property of optimal couplings: if a coupling π is

[3] This is a consequence of the fact that the optimum of a linear program is reached at an extremal point of the feasible set (see Bertsimas and Tsitsiklis (1997, Theorem 2.7)) and of the Birkhoff theorem (Birkhoff, 1946) that states that the set of extremal points of $\Pi(\mathbb{1}_n, \mathbb{1}_n)$ is exactly the set of permutation matrices of $[\![1, n]\!]$.

[4] A sequence $(\mu_k)_{k \in \mathbb{N}}$ is tight if for every $\varepsilon > 0$, there is a compact set K such that $\mu_k(\mathsf{K}) > 1 - \varepsilon$ for all k.

optimal for Problem (KP), then its support[5] $\mathrm{supp}(\pi)$ is a *c-cyclically monotone* set, i.e. for any n-tuple $\{(x_k, y_k)\}_{k=1}^n$ of elements of $\mathrm{supp}(\pi)$ and any permutation σ of $[\![1, n]\!]$, we have

$$\sum_{k=1}^{n} c(x_k, y_k) \leq \sum_{k=1}^{n} c(x_k, y_{\sigma(k)}) \ .$$

We refer to (Villani, 2008, Theorem 5.10) for the full proof of the derivation of the dual problem. Note that the Kantorovich potentials f and g can be interpreted as the Lagrange multipliers associated with the constraint of π being in $\Pi(\mu, \nu)$. They are therefore continous bounded functions since the space of measures is in duality with this latter space. Observe also that the strong duality condition of optimality allows to locate the support of the optimal coupling π, as

$$\mathrm{supp}(\pi) \subset \{(x, y) \in \mathcal{X} \times \mathcal{Y} \ : \ f(x) + g(y) = c(x, y)\} \ .$$

Dual problems for discrete distributions. When μ and ν are discrete and of the form $\sum_{k=1}^m a_k \delta_{x_k}$ and $\sum_{l=1}^n b_l \delta_{y_l}$, one can derive the following dual problems of Problem (2.2),

$$\sup_{(\alpha,\beta)\in \mathrm{R}(C)} \alpha^T a + \beta^T b \ , \tag{2.4}$$

with $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ and $\beta = (\beta_1, \ldots, \beta_n)^T$ and where,

$$\mathrm{R}(C) = \{(\alpha, \beta) \in \mathbb{R}^m \times \mathbb{R}^n \ : \ \text{for all } 1 \leq k \leq m \text{ and all } 1 \leq l \leq n, \ \alpha_k + \beta_l \leq C_{k,l}\} \ .$$

The derivation of this latter problem is a lot more straightfoward than in the general case and is a direct consequence of the more general result of the strong duality for linear programs (Bertsimas and Tsitsiklis, 1997, Theorem 4.4). The proof consists roughly in writing the Lagrangian associated with the primal problem, see (Peyré and Cuturi, 2019, Proposition 2.4) for details. The Kantorovich potentials α and β can be interpreted as *prices*, as opposed with c which can be interpreted as a *cost*. Indeed, one can illustrate the dual problem retrieving the previous example of warehouses and factories: imagine that the operator decides to subcontract the previous transportation problem to a company that basically buys the ressource from the warehouses and sell it to the factories. This company has to set a price at which it is willing to buy the quantity a_k of ressource of the k-th warehouse and a price at which it wishes to sell the quantity b_l of ressource needed by the l-th factory. Thus, if the company also applies a scheme of pricing proportional to the quantity of ressource, it can set the price of buying a quantity a of ressource from the k-th warehouse as $a \times |\alpha_k|$ with α_k being negative (since the company actually spends money when buying the ressource) and the price of selling a quantity b of ressource to the l-th warehouse as $b \times \beta_l$, so it costs in total $\alpha^T a + \beta^T b$ to the operator to move all the ressource from all the warehouses to all the factories. As opposed to the operator which wants to minimize its total cost, the company wants to maximize its gains, and so wants to find α and β such that $\alpha^T a + \beta^T b$ is maximal. Yet, in order to be competitive, the company has to set its prices such that it is cheaper for the operator to subcontract with it rather than to use the previous transportation company. Thus, it has to set its prices such that for all $1 \leq k \leq m$ and all $1 \leq l \leq n$, $\alpha_k + \beta_l \leq C_{k,l}$.

c-transforms and $\bar{c}$-transforms. Consider any dual feasible pair $(f, g) \in \mathcal{R}(c)$. For this given $f : \mathcal{X} \to \mathbb{R}$, observe that there is no better solution for $g : \mathcal{Y} \to \mathbb{R}$ than the following function $f^c : \mathcal{Y} \to \mathbb{R}$ called the *c-transform* of f and defined for all $y \in \mathcal{Y}$ as

$$f^c(y) = \inf_{x\in\mathcal{X}} \left(c(x, y) - f(x)\right) \ .$$

Indeed, it is easy to see that $(f, f^c) \in \mathcal{R}(c)$ and that is the function such that the constraint is saturated. Alternatively, for a given g, there is no better solution for f than the $\bar{c}$-transform $g^{\bar{c}} : \mathcal{X} \to \mathbb{R}$ of g defined for all $x \in \mathcal{X}$ as

$$g^{\bar{c}}(x) = \inf_{y\in\mathcal{Y}} \left(c(x, y) - g(y)\right) .$$

Moreover we say that a function $h : \mathcal{Y} \to \mathbb{R}$ is *$\bar{c}$-concave* if there exists $f : \mathcal{X} \to \mathbb{R}$ such that $h = f^c$ and we say that a function $h' : \mathcal{X} \to \mathbb{R}$ is c-concave if there exists $g : \mathcal{Y} \to \mathbb{R}$ such that $h' = g^{\bar{c}}$. Note that

[5] The support $\mathrm{supp}(\mu)$ of a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ is defined as the smallest closed Borel set A of $\mathcal{X} \times \mathcal{Y}$ such that $\mu(\mathsf{A}) = 1$, or equivalently, $\mathrm{supp}(\mu) = \{x \in \mathcal{X} \ : \$ there exits N_x open such that $x \in N_x$ and $\mu(N_x) > 0\}$.

when $\mathcal{X} = \mathcal{Y}$ and c is symmetric, the distinction between c and $\bar{c}$ can be omitted. Using the c-transform of f, one can rewrite Problem (DP) as the following single variable constrained optimization problem.

$$\sup_{f \ c\text{-concave}} \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} f^c(y)\mathrm{d}\nu(y) \ . \tag{2.5}$$

2.2 The Wasserstein distance

2.2.1 Metric properties

The most common scenario in many OT applications is when $\mathcal{X} = \mathcal{Y}$. In that case, a natural choice of cost is to set $c = d_{\mathcal{X}}^p$ where $p \geq 1$ and $d_{\mathcal{X}}$ is the metric associated with $\mathcal{X}$. This choice defines the so-called Wasserstein distance of order p

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

Metric properties of Wasserstein distance. It can be shown - see Villani (2008, definition 6.1) for details - that W_p satisfies all the axioms of a metric on $\mathcal{P}(\mathcal{X})$, i.e. if μ and ν and ξ are three probability measures on $\mathcal{X}$,

(i) $W_p(\mu, \nu)$ is symmetric and non-negative.

(ii) $W_p(\mu, \nu) = 0$ if and only if $\mu = \nu$.

(iii) W_p satisfies the triangle inequality, i.e.

$$W_p(\mu, \nu) \leq W_p(\mu, \xi) + W_p(\xi, \nu) \ .$$

However, without further assumptions on μ and ν, $W_p(\mu, \nu)$ is not a metric in the strict sense since it can be infinite. To complete its construction, it is natural to restrict W_p on a subset of $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ where it takes finite values. One can thus define the *Wasserstein space* as

$$\mathcal{W}_p(\mathcal{X}) = \{\mu \in \mathcal{P}(\mathcal{X}) \ : \ \textstyle\int_{\mathcal{X}} d_{\mathcal{X}}(x_0, x)^p \mathrm{d}\mu(x) < +\infty\} \ ,$$

where x_0 is an arbitrary element of $\mathcal{X}$. Note that this space does not depend on the choice of the point x_0. Finally, W_p defines a (finite) metric on $\mathcal{W}_p(\mathcal{X})$. The space $\mathcal{W}_p(\mathcal{X})$ endowed with the topology induced by W_p has a nice geodesic structure, in the sense that given an optimal transport coupling $\pi \in \Pi(\mu_0, \mu_1)$, the parametric curve

$$(\mu_t)_{t\in[0,1]} = \{\mu_t \in \mathcal{W}_p(\mathcal{X}) \ : \ \mu_t = P_{t\#}\pi, \text{ with } P_t = (1-t)x + ty \text{ and } t \in [0, 1]\} \ ,$$

is a constant speed geodesic between μ_0 and μ_1, i.e for every s and t in $[0, 1]$, we have $W_p(\mu_s, \mu_t) = |s - t|W_p(\mu_0, \mu_1)$. The interpolated measures of the form μ_t are often called *Wasserstein barycenters* in the literature. This formulation can be easily extended to more than two probability distributions.

Weak convergence of measures. Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence of probability measures on a Polish space $\mathcal{X}$. We say that $(\mu_k)_{k\in\mathbb{N}}$ converges weakly towards μ in $\mathcal{X}$ if for all continuous and bounded functions $h : \mathcal{X} \to \mathbb{R}$,

$$\int_{\mathcal{X}} h\mathrm{d}\mu_k \xrightarrow[k\to+\infty]{} \int_{\mathcal{X}} h\mathrm{d}\mu \ .$$

Villani (2003) has shown that for any sequence $(\mu_k)_{k\in\mathbb{N}}$ such that $\int d_{\mathcal{X}}^p(x_0, x)\mathrm{d}\mu_k(x) \to \int d_{\mathcal{X}}^p(x_0, x)\mathrm{d}\mu(x)$, the Wasserstein distance *metrizes* the weak convergence of $(\mu_k)_{k\in\mathbb{N}}$, i.e. $(\mu_k)_{k\in\mathbb{N}}$ converges weakly towards μ if and only if $W_p(\mu_k, \mu) \to 0$.

Wasserstein spaces are Polish spaces. An interesting fact shown by Bolley (2008) is that Wasserstein spaces are themselves separable complete metric spaces when endowed with W_p as metric. Therefore, for a given Polish space $\mathcal{X}$, its associated Wasserstein space $\mathcal{W}_p(\mathcal{X})$ of order p is also a Polish space.

Euclidean case. A notable particular case is when $\mathcal{X}$ is Euclidean, typically $\mathbb{R}^d$. In that case $d_{\mathcal{X}} = \|.\|$ is the Euclidean norm and

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

and the Wasserstein space of order p is defined as the set of measures on $\mathbb{R}^d$ with finite p-th order moment, i.e.

$$\mathcal{W}_p(\mathbb{R}^d) = \{\mu\in\mathcal{P}(\mathbb{R}^d) \ : \ \textstyle\int_{\mathbb{R}^d} \|x\|^p \mathrm{d}\mu(x) < +\infty\} \ .$$

2.2.2 Wasserstein distance with quadratic cost

An important particular case is when $\mathcal{X} = \mathbb{R}^d$ is Euclidean and $p = 2$. This case has been analyzed in depth in the 1980s-90s by the works of Knott and Smith (1984), Cuesta and Matrán (1989), Rüschendorf and Rachev (1990), Brenier (1991) that have resulted in the *Brenier theorem* which is one major breakthrough in the OT theory.

Brenier Theorem. The Brenier theorem (Brenier, 1991) can be stated as follows. If μ and ν are in $\mathcal{W}_2(\mathbb{R}^d)$ and at least one of the two measures, say μ, admits a density with respect of the Lesbegue measure, the Wasserstein problem

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \mathrm{d}\pi(x,y) \ , \tag{2.6}$$

admits a unique solution π^* of the form $(\mathrm{Id}_d, T_{\mathrm{OT}})_{\#}\mu$, with Id_d being the identity mapping on $\mathbb{R}^d$ and where T_{OT} is solution of the Monge problem (MP). Furthermore, T - which is often called *Brenier map* in that case - is uniquely defined as the gradient of a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, where ϕ is the unique (up to an additive constant) convex function such that $\nu = (\nabla\phi)_{\#}\mu$. Finally, ϕ is related to the dual Kantorovich potential f as for all $x\in\mathbb{R}^d$, $\phi(x) = \frac{\|x\|^2}{2} - f(x)$.

Note that McCann (1995) established a version with weaker assumptions that is that if μ vanishes on all Borel sets with Hausdorff dimension[6] $d-1$, then there exists a unique convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\nu = (\nabla\phi)_{\#}\mu$ even if μ and ν have infinite second order moments. In the case where μ and ν have finite second order moments, $\nabla\phi$ is then solution of the Monge problem (MP). Another possible generalization of Brenier theorem is to consider continuous costs c more general than the squared Euclidean distance that satisfies the so-called *Twist condition* (Villani, 2008; McCann and Guillen, 2011), i.e. such that c is differentiable with respect to x at every point and the map $y \mapsto \nabla_x c(x_0, y)$ is injective for every x_0. In particular, if c is of the form $c(x,y) = h(x-y)$ with h strictly convex for every couple (x,y), then it satisfies the Twist condition.

Monge-Ampère equation. When μ and ν both admit densities p_μ and p_ν with respect of the Lesbegue measure, one can reformulate the condition $\nu = T_{\#}\mu$, assuming T is smooth and bijective, using the change-of-variable formula. The condition $\nu = T_{\#}\mu$ translates into

$$\text{for all } x\in\mathbb{R}^d, \ p_\mu(x) = |\det(J[T](x))| p_\nu(T(x)) \ ,$$

where $J[T](x)\in\mathbb{R}^{d\times d}$ is the Jacobian matrix of T at x, i.e. the matrix obtained by stacking the gradients in x of each coordinate of T. Using Brenier theorem, we get that the unique convex function $\nabla\phi$ such that $(\mathrm{Id}_d, \nabla\phi)_{\#}\mu$ is solution of the 2-Wasserstein problem (2.6) verifies

$$\text{for all } x\in\mathbb{R}^d, \ \det(\partial^2\phi(x)) p_\nu(T(x)) = p_\mu(x) \ , \tag{2.7}$$

where $\partial^2\phi(x)$ is the Hessian of ϕ at x. This latter equation is a Monge-Ampère type equation and is particularly useful to study the regularity of Brenier maps as well as the regularity of Kantorovich potentials. We refer to Caffarelli (2003) and Figalli (2009) as review papers on this topic.

[6] The Hausdorff dimension of a Borel set A is the smallest real number d such that the Hausdorff measure of order d of A is null, i.e. $d = \inf\{d\in\mathbb{R}_+ \ : \ \mathcal{H}^d(\mathsf{A}) = 0\}$, see Ambrosio et al. (2000) for more details.

Translations. A nice property of the Wasserstein distance of order 2 is that it is possible to factor out translations. This means that if $\mathbb{E}_{X\sim\mu}[X] = m_\mu$ and $\mathbb{E}_{Y\sim\nu}[Y] = m_\nu$ with m_μ and m_ν being in $\mathbb{R}^d$, then

$$W_2^2(\mu,\nu) = \|m_\mu - m_\nu\|^2 + W_2^2(\bar{\mu},\bar{\nu}) \;, \tag{2.8}$$

where $\bar{\mu}$ and $\bar{\nu}$ are the centered measures associated with μ and ν, i.e. the measures such that if $X \sim \mu$ (respectively $Y \sim \nu$), then $X - \mathbb{E}_{X\sim\mu}[X] \sim \bar{\mu}$ (respectively $Y - \mathbb{E}_{Y\sim\nu}[Y] \sim \bar{\nu}$).

Equivalent formulation of the W_2 problem. Observe that when developping the 2-Wasserstein problem (2.6), we get

$$\inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} \|y\|^2 \mathrm{d}\nu(y) - 2\int_{\mathbb{R}^d\times\mathbb{R}^d} x^T y \mathrm{d}\pi(x,y) \right) .$$

Since $\int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x)$ and $\int_{\mathbb{R}^d} \|y\|^2 \mathrm{d}\nu(y)$ do not depend on π, it follows that the problem is equivalent to

$$\sup_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} x^T y \mathrm{d}\pi(x,y) \;.$$

This latter problem can be thought as an OT problem with cost $c(x,y) = -x^T y$ that can takes negative values.

2.2.3 Earth mover's distance

Another case of interest introduced under the name of the *Earth mover's distance* by Rubner et al. (2000) is when $p = 1$ with $\mathcal{X}$ not necessarily being Euclidean. In that case the Wasserstein problem of order 1 reads as

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}(x,y) \mathrm{d}\pi(x,y) \;. \tag{2.9}$$

There doesn't exist any result analogous to the Brenier theorem for this latter problem, and so the optimal coupling is in general not unique.

Dual formulation. The dual problem of Problem (2.9) has an interesting formulation deriving from the formulation (2.5) of the dual Kantorovich problem with c-transforms:

$$\sup_{\mathrm{Lip}(f)\leq 1} \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) - \int_{\mathcal{X}} f(x)\mathrm{d}\nu(x) \;, \tag{2.10}$$

where the Lipschitz constant $\mathrm{Lip}(f)$ of a mapping $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$\mathrm{Lip}(f) = \sup\{ \tfrac{|f(x)-f(y)|}{d_{\mathcal{X}}(x,y)} \;:\; (x,y) \in \mathcal{X},\; x \neq y \} \;.$$

The key observation to derive this formulation from (2.5) is that if $f : \mathcal{X} \to \mathbb{R}$ is c-concave, its Lipschitz constant $\mathrm{Lip}(f)$ is necessarily smaller than 1. Indeed, for all $x, y \in \mathcal{X}$, we have

$$\begin{aligned} |f(x) - f(y)| &= |\inf_{z\in\mathcal{X}} (d_{\mathcal{X}}(x,z) - g(z)) + \inf_{z\in\mathcal{X}} (d_{\mathcal{X}}(y,z) - g(z))\,| \\ &\leq \sup_{z\in\mathcal{X}} |d_{\mathcal{X}}(x,z) - d_{\mathcal{X}}(y,z)| \\ &\leq d_{\mathcal{X}}(x,y) \;, \end{aligned}$$

where the first equality follows from the definition of the c-transform, the next inequality follows from the identity $|\inf f - \inf g| \leq \sup |f - g|$, and the last from the triangle inequality. From the fact that $\mathrm{Lip}(f) \leq 1$, one can deduce with further calculations, see Peyré and Cuturi (2019, Proposition 6.1), that $f^c = -f$, which gives (2.10). This latter formulation plays an important role in generative modeling, since it is the core theoretical component of the *Wasserstein Generative Adversarial Networks* (WGANs) (Arjovsky et al., 2017) and the *Wasserstein Autoencoders* (WAEs) (Tolstikhin et al., 2018).

2.2.4 Particular cases: one-dimensional and Gaussian distributions

There are two important particular cases of distributions on which Wasserstein problems admit closed-form solutions: when μ and ν are on $\mathbb{R}$ and when μ and ν are Gaussian distributions.

Optimal transport on the real line. For a measure μ on $\mathbb{R}$, one can define its *cumulative distribution function* $F_\mu : \mathbb{R} \to [0,1]$ for all $x \in \mathbb{R}$ as

$$F_\mu(x) = \mu((-\infty, x]) \ .$$

One can also define its pseudoinverse called the *generalized quantile function* of μ for all $r \in [0,1]$ as

$$F_\mu^{-1}(r) = \inf\{x \in \mathbb{R} \ : \ F_\mu(x) \geq r\} \ .$$

Then, for all $p \geq 1$,

$$W_p^p(\mu, \nu) = \int_0^1 |F_\mu^{-1}(r) - F_\nu^{-1}(r)|^p \mathrm{d}r \ .$$

Furthermore the optimal coupling π^* is of the form $(\mathrm{Id}, T_{\mathrm{OT}})_\# \mu$, where Id is the identity operator on $\mathbb{R}$ and where T_{OT} is defined as,

$$T_{\mathrm{OT}} = F_\nu^{-1} \circ F_\mu \ . \tag{2.11}$$

We refer to Santambrogio (2015, Chapter 2) for a detailed survey of the properties of optimal transport on the real line. Note that T is an non-decreasing function. Therefore, the notion of gradient of a convex function in the Brenier theorem corresponds to a generalization of T being non-decreasing in higher dimension. Moreover, by analogy with (2.11), one can define a generalization of the generalized quantile function of a distribution μ on $\mathbb{R}^d$ as the Monge map between a reference distribution on $\mathbb{R}^d$ - typically the uniform distribution on the unit cube or the standard Gaussian distribution $\mathrm{N}(0, \mathrm{Id}_d)$ - and μ, see Carlier et al. (2016). When μ and ν are two discrete distributions of the forms $\mu = \sum_{k=1}^m \frac{1}{m} \delta_{x_k}$ and $\nu = \sum_{l=1}^n \frac{1}{n} \delta_{y_l}$, this corresponds to sorting $x_1 \leq \cdots \leq x_m$ and $y_1 \leq \cdots \leq y_n$ and sending as much mass as possible from x_1 to y_1, then sending the remaining mass to y_2 (and so on if it remains some mass), then repeating this procedure for x_2 and so on until no more mass is left. Thus the Wasserstein distance between discrete distributions on the real line can be solved using simple sorting algorithms.

Optimal transport between Gaussian distributions. Another important case where the Wasserstein distance has a closed-form is when $\mathcal{X} = \mathbb{R}^d$, $p = 2$ and μ and ν are Gaussian distributions. Indeed if $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$, with m_0 and m_1 in $\mathbb{R}^d$ and Σ_0 and Σ_1 in $\mathbb{S}_+^d$, where $\mathbb{S}_+^d$ denotes the set of symmetric Positive Semi-Definite (PSD) matrices, the Wasserstein distance of order 2 is written

$$W_2^2(\mu, \nu) = \|m_0 - m_1\|^2 + \mathrm{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_1^{\frac{1}{2}} \Sigma_0 \Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \ ,$$

where for any PSD matrix $A \in \mathbb{S}_+^d$, $A^{\frac{1}{2}}$ is the unique PSD square root of A. Note that the rightmost term defines itself a metric between PSD matrices often referred to as the *Bures distance* Bures (1969). Therefore the Wassertein distance of order 2 between Gaussian distributions is often called the *Bures-Wasserstein* distance and defines a metric on the space of Gaussian distributions on $\mathbb{R}^d$ that we denote here $\mathcal{N}(\mathbb{R}^d)$[7]. Note that when Σ_0 and Σ_1 commute, the Bures distance coincides with the Hellinger distance

$$d_{\mathrm{H}}(\Sigma_0, \Sigma_1) = \|\Sigma_0^{\frac{1}{2}} - \Sigma_1^{\frac{1}{2}}\|_{\mathcal{F}} \ ,$$

where $\|.\|_{\mathcal{F}}$ is the Frobenius norm between matrices of size $d \times d$. When Σ_0 is non-singular, the optimal coupling π is of the form $(\mathrm{Id}_d, T_{\mathrm{OT}})_\# \mu$ and the Monge map T_{OT} between μ and ν turns out to be affine and defined for all $x \in \mathbb{R}^d$ as

$$T_{\mathrm{OT}}(x) = m_1 + \Sigma_0^{-\frac{1}{2}} (\Sigma_0^{\frac{1}{2}} \Sigma_1 \Sigma_0^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_0^{-\frac{1}{2}} (x - m_0) \ . \tag{2.12}$$

These results have been known since Dowson and Landau (1982) and have been proved in several different ways. One key point in the proof is to observe that the optimal transport plan is Gaussian: observe first that by sucessively factoring out the translations as in (2.8) and using the probabilistic formulation (2.3), we get that

$$W_2^2(\mu, \nu) = \|m_0 - m_1\|^2 + \inf_{X \sim \bar{\mu}, \ Y \sim \bar{\nu}} \mathbb{E}[\|X - Y\|^2]$$

[7] Note that $\mathcal{N}(\mathbb{R}^d)$ includes the degenerate Gaussian distributions, as for instance the Dirac distributions. The interior set of $\mathcal{N}(\mathbb{R}^d)$ is the set of all non-degenerate Gaussian distributions and is often referred to as the *Bures manifold* in the literature.

$$= \|m_0 - m_1\|^2 + \mathbb{E}_{X\sim\bar{\mu}}[\|X\|^2] + \mathbb{E}_{Y\sim\bar{\nu}}[\|Y\|^2] - 2 \sup_{X\sim\bar{\mu},\ Y\sim\bar{\nu}} \operatorname{tr}(\mathbb{E}[XY^T]) \ ,$$

and so the problem is equivalent to find the law of the couple of random variables (X, Y) such that the trace of the cross-covariance matrix $\mathbb{E}[XY^T]$ is maximal. Yet, for any feasible value A of this cross-covariance matrix, there exists a Gaussian coupling π of (X, Y) such that $\mathbb{E}_{(X,Y)\sim\pi}[XY^T] = A$, and so there exists a Gaussian coupling π^* such that $\operatorname{tr}(\mathbb{E}_{(X,Y)\sim\pi^*}[XY^T])$ is maximal.

2.3 Solving OT in practice

Apart from the special cases mentioned above and some other additional simple cases, it is in general not possible to solve analytically the OT problem. However, when μ and ν are discrete, it is possible to solve it numerically. To complete our general introduction to classic optimal transport, we briefly present here the different numerical solvers that are commonly used in the literature. We refer to Peyré and Cuturi (2019, Chapters 3 and 4) for a complete introduction on this topic. In all this section, we suppose than μ and ν are discrete and on the form $\mu = \sum_{k=1}^m a_k \delta_{x_k}$ and $\nu = \sum_{l=1}^n b_l \delta_{y_l}$.

Linear programing. In order to solve Kantorovich problem in the discrete case (2.2), one can rely on classic algorithms for solving linear programs (Dantzig, 1951). Among them, the reference algorithms to solve (2.2) are the *network simplexes* (Cunningham, 1976). These algorithms rely on the dual problem (2.4) that is

$$\min_{\substack{(\alpha,\beta)\in\mathbb{R}^m\times\mathbb{R}^n \\ \text{for all } 1\leq k,l\leq m,n,\ \alpha_k+\beta_l\leq C_{k,l}}} \alpha^T a + \beta^T b \ . \tag{2.13}$$

This type of algorithms consists in searching an optimal couple of Kantorovich solutions (α, β) among the extremal feasible points, i.e. the couples (α, β) such that $\alpha_k + \beta_l = C_{k,l}$ for the indices k and l such that $\omega_{k,l} > 0$, where ω is the primal variable. This leverages one fundamental result of linear programing which states that the optimum of a linear program is reached at an extremal point of the feasible set, see Bertsimas and Tsitsiklis (1997, Theorem 2.7). When $m = n$, the most efficient algorithm of this type has a complexity $O(n^3 \log(n))$. There exist alternative algoritms for solving this problem which include interior points, as dual ascent methods (Kuhn, 1955) for instance, but these methods do not perform as well as the network simplexes on this particular type of linear program. In the special case of the optimal assignment problem where $m = n$ and both histograms are uniform, one can use the *Auction algorithm* (Bertsekas and Eckstein, 1988), whose most effective refinement has a cubic complexity $O(n^3)$.

The failure of alternate optimization. A common type of algorithms to solve optimization problems with two variables as (2.13) consists in alternating the optimization of each variable. One could be tempted to design an alternate optimization algorithm using the discrete versions of the c and $\bar{c}$-transforms: given a feasible couple (α, β), one can define $\alpha^C \in \mathbb{R}^n$ and $\beta^{\bar{C}} \in \mathbb{R}^m$ such that for all $1 \leq k \leq m$ and all $1 \leq l \leq n$,

$$\begin{cases} (\alpha^C)_l = \min\limits_{1\leq k\leq m} C_{k,l} - \alpha \\ (\beta^{\bar{C}})_k = \min\limits_{1\leq l\leq n} C_{k,l} - \beta \ , \end{cases}$$

and for a given feasible couple (α, β) we have

$$\alpha^T a + \beta^T b \leq \alpha^T a + (\alpha^C)^T b \leq (\alpha^{C\bar{C}})^T a + (\alpha^C)^T b \ .$$

Thus, one could design an algorithm by applying successively C and $\bar{C}$ transforms, yet this doesn't work because $\alpha^{C\bar{C}C} = \alpha$, and so we would quickly reach a stationary regime. This behavior is a classic behavior of alternating optimization schemes on non-smooth problems and a typical way to cope with it is to introduce regularization, which motivates the use of entropic optimal transport.

Entropic regularization. Solving the OT problem (2.13) with a network simplex algorithm remains costly since it has a complexity of $O(n^3 \log(n))$, and it is not possible to solve it directly with an alternate maximization scheme as we have seen above. An idea that has been made very popular in the OT community by (Cuturi, 2013) is to penalize the entropy of the coupling ω and thus to solve the following regularized problem

$$\inf_{\omega\in\Pi(a,b)} \sum_{k,l} C_{k,l}\omega_{k,l} - \varepsilon H(\omega) \ , \tag{ε-KP}$$

where

$$H(\omega) = -\sum_{k,l} \omega_{k,l}(\log(\omega_{k,l}) - 1) \ .$$

Solving Problem (ε-KP) instead of (KP) has several important advantages: it turns the optimal transport problem into a strongly-convex minimization problem with a unique solution, and the minimization of the regularized problem can be solved using a simple alternate minimization scheme with simple matrix-vector products as iterations, which makes it particularly suited for GPU implementation. The algorithm used to solve Problem (ε-KP) is the Sinkhorn-Knopp matrix scaling algorithm. It leverages a result of Sinkhorn and Knopp (1967) that states that Problem (ε-KP) admits a unique solution ω^* that is of the form

$$\omega^* = \text{diag}(u)K\text{diag}(v) \ ,$$

with $u \in \mathbb{R}_+^m$, $v \in \mathbb{R}_+^n$ and $K = \exp[-\frac{C}{\varepsilon}]$ where the exponential is applied entrywise, and where $\text{diag}(u)$ is the diagonal matrix of size $m \times m$ with diagonal u. The Sinkhorn-Knopp algorithm consists in starting from an initial couple $(u^{(0)}, v^{(0)})$, usually, $(\mathbb{1}_m, \mathbb{1}_n)$, and finding $u^{(1)} \in \mathbb{R}^m$ such that the coupling $\omega = \text{diag}(u^{(1)})K\text{diag}(v^{(0)})$ has left-marginal a, i.e. such that $\omega\mathbb{1}_n = a$, then finding $v^{(1)} \in \mathbb{R}^n$ such that the coupling $\omega' = \text{diag}(u^{(1)})K\text{diag}(v^{(1)})$ has right-marginal b, i.e. such that $\omega^T\mathbb{1}_m = b$, and then continuing this alternating optimization scheme until convergence. This leads to the following updates for u and v:

$$\begin{cases} u^{\{i+1\}} = a \oslash Kv^{\{i\}} \\ v^{\{i+1\}} = b \oslash K^T u^{\{i+1\}} \end{cases} ,$$

where $\oslash$ denotes the entrywise division. In a nutshell, this gives Algorithm 1.

Algorithm 1 Sinkhorn-Knopp algorithm for regularized OT problem

Require: a, b, C, $\varepsilon > 0$, $v^{\{0\}} = \mathbb{1}_n$
1: $K \leftarrow \exp[-C/\varepsilon]$
2: **for** $i = 1, \dots, N_{it}$ **do**
3: $\quad u^{\{i\}} \leftarrow a \oslash Kv^{\{i-1\}}$ ▷ Update left scaling
4: $\quad v^{\{i\}} \leftarrow b \oslash K^T u^{\{i\}}$ ▷ Update right scaling
5: **end for**
6: **return** $\omega = \text{diag}(u)K\text{diag}(v)$

In terms of complexity, Altschuler et al. (2017) have shown that when $m = n$ and setting $\tau = \frac{4\log(n)}{\varepsilon}$, the Sinkhorn-Knopp algorithm could produce a coupling ω such that $\sum_{k,l} C_{k,l}\omega_{k,l}$ approximate the optimal value of the unregularized problem (KP) with precision τ in $O(n^2\log(n)\tau^{-3})$ operations. Finally, note that there exist countless refinements, extensions and generalizations of the Sinkhorn-Knopp algorithm. Among them, some recent refinements build notably on low-rank factorizations or approximations of the Kernel matrix K (Solomon et al., 2015; Altschuler et al., 2019; Scetbon and Cuturi, 2020) while others impose a low-rank constraint on the coupling ω (Forrow et al., 2019; Scetbon et al., 2021), which results in very efficient solvers whose complexity depends linearly on the number of points.

Sliced methods. Another type of methods commonly used in the literature to solve OT problems builds on the fact that Wasserstein problems on the real line can be solved using simple sorting algorithms, see Section 2.2.4. This results in the so-called *Sliced Wasserstein* distance (SW) (Rabin et al., 2012) that is defined as follows for all $p \geq 1$ and given μ and ν on $\mathbb{R}^d$,

$$SW_p^p(\mu, \nu) = \int_{\mathsf{S}^{d-1}} W_p^p(P_{\theta\#}\mu, P_{\theta\#}\nu)\mathrm{d}\mathcal{L}^{d-1}(\theta) \ , \qquad (SW_p)$$

where S^{d-1} is the unit-hypersphere on $\mathbb{R}^d$, $\mathcal{L}^{d-1}$ is the uniform measure on S^{d-1}, and P_θ is a projection on θ, i.e. for all $x \in \mathbb{R}^d$, $P_\theta(x) = x^T\theta$. The idea behind the sliced Wasserstein distance is to first, obtain a family of one-dimensional representations for a higher-dimensional probability distribution through linear projections, and then, calculate the distance between two input distributions as a functional of the Wasserstein distance of their one-dimensional representations, i.e. the one-dimensional projected measures. If the Sliced Wasserstein distance can be thought as an approximation of the Wasserstein distance, it can also be thought as another OT distance on its own that has several interesting properties. It has notably been shown that SW_2 defines a metric on $\mathcal{W}_2(\mathbb{R}^d)$ (Bonnotte, 2013) that metrizes the weak

convergence (Nadjahi et al., 2019) and which is equivalent to the Wasserstein distance W_2 for measures with compact supports (Bonnotte, 2013; Nadjahi et al., 2020). In practice, (SW_p) is approximated using a Monte-Carlo method that corresponds to choose randomly L projection directions on S^{d-1} and to compute $\frac{1}{L}\sum_i^L W_p^p(P_{\theta_i\#}\mu, P_{\theta_i\#}\nu)$. Hence, for discrete probability measures composed of n points, the overall complexity of computing (SW_p) is $O(Ln\log(n))$, which makes it very attractive when dealing with large-scale problems. Finally, one drawback of this type of methods is it doesn't provide directly an optimal coupling between the measures μ and ν.

2.4 Conclusion

In this chapter, we have introduced the main concepts of the classic optimal transport theory. The theory of optimal transport has matured over the years starting from its initial formulation by Monge in 1781, then its rediscovery in the 1940s thanks to the works of Kantorovich and Dantzig, and finally its revisit under new points of view in the 1990s by mathematicians such as Brenier and later in the 2000s with the works of Villani. The introduction of entropic-regularized OT by Cuturi (2013) has yet triggered another revolution of the field, transforming it from a predominantly theoretical domain to an applied field that provides tools very useful to solve a large class of data science problems. In the imaging science field, OT has been used in numerous applications such as image matching (Zhu et al., 2007; Wang et al., 2013; Li et al., 2013), medical imaging (Wang et al., 2010; Gramfort et al., 2015), texture synthesis and style transfer (Leclaire and Rabin, 2021; Gutierrez et al., 2017), or shape registration (Feydy et al., 2017; Su et al., 2015), just to name a few.

In its classic setting, an implicit prerequisite of optimal transport is that the two distributions involved lie on the same ground space, or at least that the two spaces are comparable, i.e. there exists a relevant cost function to compare them. However, this assumption may not hold for many applications. This is often the case when dealing with structured data, as graphs for instance, or when the data come from heterogeneous sources, as in the case of heterogeneous domain adaptation. Some other tasks such as shape matching or word embedding require designing cost functions such that the problem is invariant to some families of invariances, such as translations and rotations for example, in the sense that we want the distance between a given distribution and a translated and rotated version of itself to be null. Even if the distributions involved in these applications may live in the same ground space, it is not straightforward to design an adequate cost function. In the next chapter, we introduce the common generalization of optimal transport to measures living in incomparable spaces, which is known as the Gromov-Wasserstein distance (Mémoli, 2011). We also introduce two other recent formulations proposed by Alvarez-Melis et al. (2019) and Cai and Lim (2022) and we define a new formulation that we call *embedded Wasserstein* discrepancy.

Chapter 3

Optimal transport between measures on incomparable spaces

Contents

In the first part of this chapter, we introduce the theoretical concepts of the Gromov-Wasserstein distance (Mémoli, 2011), that is probably the most common generalization of optimal transport between measures on incomparable spaces. We also briefly present the common numerical solvers used for this problem. In the second part of this chapter, we introduce two other recent formulations proposed by Alvarez-Melis et al. (2019) and Cai and Lim (2022) and we define a new formulation that we call embedded Wasserstein. Parts of this chapter are reproduction of Salmona et al. (2023).

3.1 The Gromov-Wasserstein distance

One intrisic limitation of the classic OT theory introduced above is that it implicitly assumes that the spaces $\mathcal{X}$ and $\mathcal{Y}$ are *comparable*, i.e. that there exists a relevant cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ to compare them. Yet, this assumption is not always verified. For instance, if $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^{d'}$ with $d \neq d'$, the definition of a meaningful cost function $c : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}_+$ is not straightforward. Furthermore, some applications such as shape matching require having an OT distance that is invariant to important families of transformations, such as translations or rotations or more generally to *isometries*. Even if the two distributions involved in these applications do live in the same ground space, it is not straightforward to design a cost function such that the resulting OT distance will be invariant to these families of transformations. The common generalization of the classic optimal transport problem that overcomes these limitations is the *Gromov-Wasserstein* (GW) distance (Mémoli, 2011). The goal of this section is to present the Gromov-Wasserstein problem and its metric properties. We refer to Mémoli (2011), Sturm (2012), and Chowdhury and Mémoli (2019) for further readings on the theoretical properties of the Gromov-Wasserstein problem.

3.1.1 Problem statement

The Gromov-Wasserstein problem. The Gromov-Wasserstein problem (Mémoli, 2011) can be defined as follows: given two Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, two measurable integrable functions $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

and $c_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, it aims at finding for any $p \geq 1$,

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

With a slight abuse of language, we will also call $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ cost functions, even if they can take values in $\mathbb{R}$ and not only in $\mathbb{R}_+$ and they are not necessarily lower semi-continuous. Problem (GW_p) depends on the choice of $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$. When this choice is clear from the context, we will note $GW_p(\mu, \nu)$ instead of $GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu)$. The GW distance is constructed so that if an optimal coupling π assigns x to y and x' to y', then the value of $c_{\mathcal{X}}(x, x')$ should be close to the value of $c_{\mathcal{Y}}(y, y')$. Thus it measures a distortion between the pair of points (x, x') and (y, y') within each space. Since $\mathcal{X}$ and $\mathcal{Y}$ can each be endowed with respective metric $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, a natural choice for $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ is $d^q_{\mathcal{X}}$ and $d^q_{\mathcal{Y}}$ with $q \geq 1$. This case has been studied in depth by Sturm (2012). The general case when $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are not metrics but simply measurable integrable functions which for instance don't verify the triangle inequality has been studied by Chowdhury and Mémoli (2019).

Network measure and metric measure spaces. (GW_p) defines a measure of dissimilarity between *network measure spaces* (Chowdhury and Mémoli, 2019), i.e. the triplets of the form $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ where $\mathcal{X}$ is a Polish space, $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a measurable integrable function, and μ is a probability measure on $\mathcal{X}$. When $c_{\mathcal{X}} = d^q_{\mathcal{X}}$, where $q \geq 1$ and $d_{\mathcal{X}}$ is the metric associated with $\mathcal{X}$, (GW_p) defines a measure of dissimilarity between *metric measure spaces* (Sturm, 2012), i.e. the triplets of the form $(\mathcal{X}, d_{\mathcal{X}}, \mu)$.

The Gromov-Wasserstein problem between discrete distributions. When μ and ν are discrete probability distributions of the form $\mu = \sum_{k=1}^m a_k \delta_{x_k}$ and $\nu = \sum_{l=1}^n b_l \delta_{y_l}$, the Gromov-Wasserstein problem (to a power p) reads as

$$\inf_{\omega \in \Pi(a,b)} \sum_{i,j,k,l} |C^x_{i,k} - C^y_{j,l}|^p \omega_{i,j} \omega_{k,l} \ , \qquad (3.1)$$

where C^x and C^y are matrices of respective sizes $m \times m$ and $n \times n$. This is a non-convex quadratic program (Loiola et al., 2007) that can be seen as a relaxation of the *Quadratic Assignment Problem* (QAP) (Koopmans and Beckmann, 1957), which is known to be in all generality a NP-hard problem. Such problem consists in its most standard form in solving, given two matrices $A = (a_{i,j})_{1 \leq i,j \leq n}$ and $D = (D_{i,j})_{1 \leq i,j \leq n}$,

$$\min_{\sigma \in \mathrm{Perm}(n)} \sum_{i,j} a_{i,j} d_{\sigma(i),\sigma(j)} \ .$$

This latter problem can be illustrated with the *facility location problem*. Given n different facilities and n possible locations such that the distance between the i-th and the j-th facility is $d_{i,j}$, and given a matrix of flows $a_{i,j}$ corresponding for instance to the number of people that have to move everyday from facility i to facility j, we aim to attribute a given location to each facility that minimizes the total sum of distance covered in one day by all people that have to move from one given facility to another. Thus, we want to find the permutation σ such that the total cost $\sum_{i,j} a_{i,j} d_{\sigma(i),\sigma(j)}$ is minimal. The QAP is also intrinsically linked with the *graph matching problem* (Karp et al., 1990) whose goal is to match the edge affinities of two graphs that are represented by symmetric matrices.

Discrete distributions as structured objects. Between discrete distributions $\mu = \sum_{k=1}^m a_k \delta_{x_k}$ and $\nu = \sum_{l=1}^n b_l \delta_{y_l}$ respectively on $\mathcal{X}$ and $\mathcal{Y}$, solving the GW problem (3.1) with ground cost matrices C^x and C^y can informally be thought as comparing the edge affinities of two graphs, whose vertices are respectively the tuples $\{x_k\}_{1 \leq k \leq m}$ and $\{y_l\}_{1 \leq l \leq n}$, and whose edges are encoded by the matrices C^x and C^y. Hence the Gromov-Wasserstein distance compares the *edge affinities* of the two graphs, related to the inherent *structures* of the two distributions μ and ν, while the Wasserstein distance compares *the vertex positions* in the two graphs. This explains why the Gromov-Wasserstein is naturally well-suited to compare structured data. Building on this analysis, the fused Gromov-Wasserstein distance (Vayer et al., 2019a) proposes to use both edge affinities and vertex positions informations by mixing the Gromov-Wasserstein and the Wasserstein problems.

Existence of solutions. As for Problem (KP), one can show that Problem (GW_p) admits at least one solution if $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are continuous. The proof builds on similar arguments as for the existence of solutions of Problem (KP): the main idea is to use the Weierstrass theorem using the fact that $\Pi(\mu,\nu)$ is compact if $\mathcal{X}$ and $\mathcal{Y}$ are Polish. The only fact that needs to be checked is that the functional $J : \pi \mapsto \int\int |c_{\mathcal{X}} - c_{\mathcal{Y}}|^p \mathrm{d}\pi\mathrm{d}\pi$ is lower semi-continuous. This can be done by showing that J can be expressed as a supremum of lower semi-continuous functions, see Vayer (2020, Lemma 2.2.1).

Link with Gromov-Hausdorff distance. The Gromov-Wasserstein problem shares close connections with the Gromov-Hausdorff distance (Gromov, 1981) that informally quantifies how far two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are from being isometric to each other. However computing this distance results in a highly non-convex optimization problem whose global solution is generally untractable. The Gromov-Wasserstein distance can be thought as a "smoothing" of the Gromov-Hausdorff distance as shown in Mémoli (2011).

The Gromov-Monge problem. Analogously to Problem (MP), one can define the Gromov-Monge problem (Mémoli and Needham, 2018) as

$$\inf_{T \,:\, \nu = T_{\#}\mu} \int_{\mathcal{X}} \int_{\mathcal{X}} |c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(T(x),T(x'))|^p \mathrm{d}\mu(x)\mathrm{d}\mu(x') \ .$$

However, as for (MP), there are many cases where there doesn't exist any map such that $\nu = T_{\#}\mu$ and so where the infinimum doesn't actually exist. An active field of research aims at establishing whether there exists a result similar to the Brenier theorem for the Gromov-Wasserstein problem of order 2 in the Euclidean setting, more precisely whether the optimal coupling solution of (GW_p) is supported by the graph of the gradient of a convex function. Dumont et al. (2022) have notably shown an analogous result to the Brenier theorem in the case where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^{d'}$ with $d \geq d'$, when choosing the inner-products as cost functions $c_{\mathbb{R}^d}$ and $c_{\mathbb{R}^{d'}}$, supposing that μ admits a density with respect to the Lesbegue measure on $\mathbb{R}^d$ and supposing that the supports of μ and ν are compact. A similar result has been proved by Vayer (2020) but with stronger assumptions that happen to be difficult to check in practice. The situation seems however to be more tricky in the case where $c_{\mathbb{R}^d} = \|.\|^2_{\mathbb{R}^d}$ and $c_{\mathbb{R}^{d'}} = \|.\|^2_{\mathbb{R}^{d'}}$, where $\|.\|^2_{\mathbb{R}^d}$ denotes the Euclidean norm in $\mathbb{R}^d$. Indeed, in the case where $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$, and where $\mu = \frac{1}{n}\sum_k \delta_{x_k}$ and $\nu = \frac{1}{n}\sum_l \delta_{y_l}$, Beinert et al. (2022) have exhibited a counter-example which suggests there doesn't systematically exist a monotone Monge map solution of the problem in one dimension, and so there doesn't systematically exist a Monge map that takes the form of a gradient of convex function in higher dimension. A procedure to exhibit other counter-examples in one dimension has been proposed in Dumont et al. (2022).

3.1.2 Metric properties of Gromov-Wasserstein distances

Before presenting the metric properties of the Gromov-Wasserstein distance, we need to introduce the notion of isometry, as well as the notions of strong and weak isomorphisms.

Isometries. Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two metric spaces. We say that $\phi : \mathcal{X} \to \mathcal{Y}$ is an isometry if for every couple (x, x') in $\mathcal{X} \times \mathcal{X}$,

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

Note that with this definition, isometries are necessarily injective but not necessarily bijective. When there exists a bijective isometry between two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, we say that these metric spaces are *isometric*. In the Euclidean setting, the Mazur-Ulam theorem (Mazur and Ulam, 1932) states, in the refined version of Baker (1971), that the isometries from $(\mathbb{R}^{d'}, \|.\|_{\mathbb{R}^{d'}})$ to $(\mathbb{R}^d, \|.\|_{\mathbb{R}^d})$ with $d \geq d'$ are necessarily affine.

Strong isomorphisms. When considering *metric measure spaces* of the form $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ instead of simple *metric spaces*, one can enrich the notion of isometry by taking into account the measure μ. This results in the notion of *strong isomorphisms* (Sturm, 2012): we say that two metric measures spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ are strongly isomorphic if there exists a bijective isometry $\phi : \mathrm{supp}(\mu) \to \mathrm{supp}(\nu)$ that pushes μ into ν.

Weak isomorphisms. There exists a similar notion to the notion of strong isomorphisms when considering *network measure spaces* of the form $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ instead of *metric measure spaces*, with $c_{\mathcal{X}}$ not necessarily being a metric: we say that $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ is *weakly isomorphic* (Sturm, 2012) to $(\mathcal{Y}, c_{\mathcal{Y}}, \nu)$ if there exists a third measure network $(\mathcal{Z}, c_{\mathcal{Z}}, \xi)$ such that $\text{supp}(\xi) = \mathcal{Z}$ and there exist two maps $\phi_0 : \mathcal{Z} \to \mathcal{X}$ and $\phi_1 : \mathcal{Z} \to \mathcal{Y}$ such that

(i) for all $(z, z') \in \mathcal{Z} \times \mathcal{Z}$, $c_{\mathcal{Z}}(z, z') = c_{\mathcal{X}}(\phi_0(z), \phi_0(z')) = c_{\mathcal{Y}}(\phi_1(z), \phi_1(z'))$.

(ii) ϕ_0 pushes ξ into μ and ϕ_1 pushes ξ into ν.

When $c_{\mathcal{X}} = d_{\mathcal{X}}^q$ and $c_{\mathcal{Y}} = d_{\mathcal{Y}}^q$ with $q \geq 1$, Sturm (2012) has shown that the notion of weak isomorphism was in fact equivalent to the notion of strong isomorphism, i.e. the network measure spaces $(\mathcal{X}, d_{\mathcal{X}}^q, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}^q, \nu)$ are weakly isomorphic if and only if the metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ are strongly isomorphic.

Metric properties of Gromov-Wasserstein distances. It can be shown, see Chowdhury and Mémoli (2019, Theorem 18) for details, that GW_p satisfies all the axioms of a pseudo-metric on the space of network measure spaces. More precisely, for $(\mathcal{X}, c_{\mathcal{X}}, \mu)$, $(\mathcal{Y}, c_{\mathcal{Y}}, \nu)$ and $(\mathcal{Z}, c_{\mathcal{Z}}, \xi)$, we have

(i) $GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu)$ is symmetric and non-negative.

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

(iii) GW_p satisfies the triangle inequality, i.e.

$$GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) \leq GW_p(c_{\mathcal{X}}, c_{\mathcal{Z}}, \mu, \xi) + GW_p(c_{\mathcal{Z}}, c_{\mathcal{Y}}, \xi, \nu) \, .$$

In the case where $c_{\mathcal{X}} = d_{\mathcal{X}}^q$ and $c_{\mathcal{Y}} = d_{\mathcal{Y}}^q$ with $q \geq 1$, (ii) can be replaced by:

$$GW_p(d_{\mathcal{X}}^q, d_{\mathcal{Y}}^q, \mu, \nu) = 0 \text{ if and only if } (\mathcal{X}, d_{\mathcal{X}}, \mu) \text{ and } (\mathcal{Y}, d_{\mathcal{Y}}, \nu) \text{ are } \textit{strongly isomorphic},$$

and GW_p defines thus a pseudo-metric on the space of metric measure spaces in that case. We refer to Sturm (2012, Lemma 9.2) for the proof of this latter fact. However GW_p is not a metric in the strict sense of the term since: (i) it can possibly take infinite values, (ii) we can have $GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = 0$ and $(\mathcal{X}, c_{\mathcal{X}}, \mu)$ not equal to $(\mathcal{Y}, c_{\mathcal{Y}}, \nu)$. A natural way to remedy (i) is to restrict GW_p to spaces where it can only take finite values. We thus define, for $p \geq 1$,

$$\mathbb{M}_p = \{(\mathcal{X}, c_{\mathcal{X}}, \mu) \, : \, \textstyle\int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{X}}^p(x, x') \mathrm{d}\mu(x) \mathrm{d}\mu(x') < +\infty\} \, ,$$

such that GW_p defines a pseudo-metric on $\mathbb{M}_p$. Finally, GW_p defines a metric on $\mathbb{M}_p$ quotiented by the weak isomorphisms, or equivalently quotiented by the strong-isomorphisms if we restrict GW_p to the metric measure spaces. Interestingly, the space $\mathbb{M}_p$ quotiented by the weak isomorphims has also a nice geodesic structure (Sturm, 2012) when endowed with the topology induced by GW_p. We can therefore defines the notion of *Gromov-Wasserstein barycenters* (Peyré et al., 2016) between network measure spaces, analogously with the notion of Wasserstein barycenters for the Wasserstein distance.

3.1.3 Particular case: one-dimensional distributions

In general, the Gromov-Wasserstein problem cannot be solved analytically. There is however one very particular case where closed-form solutions can be derived. This happens when $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $p = 2$, the ground costs $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are both the Euclidean inner-product on $\mathbb{R}$, i.e $c_{\mathcal{X}}(x, y) = c_{\mathcal{Y}}(x, y) = xy$ for all x and y in $\mathbb{R}$. In that case, Problem (GW_p) reads as

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} \int_{\mathbb{R}\times\mathbb{R}} |xx' - yy'|^2 \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y') \, . \tag{3.2}$$

Vayer (2020) has shown that this latter problem admits two distinct solutions that are respectively of the form $(\mathrm{Id}_{\mathbb{R}}, T_{\mathrm{GW}}^{\uparrow})_{\#}\mu$ and $(\mathrm{Id}_{\mathbb{R}}, T_{\mathrm{GW}}^{\downarrow})_{\#}\mu$ where $T_{\mathrm{GW}}^{\uparrow} = F_{\nu}^{-1} \circ F_{\mu}^{\uparrow}$ and $T_{\mathrm{GW}}^{\downarrow} = F_{\nu}^{-1} \circ F_{\mu}^{\downarrow}$ with F_{ν}^{-1} being the generalized quantile function associated with ν and $F_{\mu}^{\uparrow}$ and $F_{\mu}^{\downarrow}$ are respectively the *cumulative* and *anti-cumulative* distribution function associated with μ, i.e for all $x \in \mathbb{R}$,

$$F_{\mu}^{\uparrow}(x) = \mu((-\infty, x]) \qquad \text{and} \qquad F_{\mu}^{\downarrow}(x) = \mu([-x, +\infty)) \, .$$

The proof of Vayer (2020) use the equivalence between Problem (3.2) and the problem introduced by Alvarez-Melis et al. (2019) that we will present in Section 3.2. We give an alternative proof of this result in Section 3.3.4. Note that this result doesn't hold anymore if we replace the inner-product by the squared distance, i.e. when $c_{\mathcal{X}}(x,y) = c_{\mathcal{Y}}(x,y) = |x-y|^2$ for all x in y in $\mathbb{R}$, as we have seen that counter-examples could be exhibited (Beinert et al., 2022; Dumont et al., 2022) when introducing the Gromov-Monge problem.

3.1.4 Solving GW in practice

In this section we briefly introduce the commonly used numerical solvers for the Gromov-Wasserstein problem. In all what follows μ and ν are discrete probability distributions on $\mathcal{X}$ and $\mathcal{Y}$ of the form $\sum_{k=1}^{m} a_k \delta_{x_k}$ and $\sum_{l=1}^{n} b_l \delta_{y_l}$.

Quadratic programing. The Gromov-wasserstein problem (3.1) between μ and ν reads as

$$\inf_{\omega \in \Pi(a,b)} \sum_{i,j,k,l} |C^x_{i,k} - C^y_{j,l}|^p \omega_{i,j} \omega_{k,l} \ .$$

This is a non-convex quadratic program (QP) which is known to be NP-hard in general (Loiola et al., 2007). It is therefore expected that its approximation is costly. Observe that this latter problem can be rewritten under the form

$$\inf_{\omega \in \Pi(a,b)} \langle L^p(C^x, C^y) \otimes \omega, \omega \rangle_{\mathcal{F}} \ , \tag{3.3}$$

where L is the functional which associates to (C^x, C^y) the 4-th order tensor $(|C^x_{i,k} - C^y_{j,l}|)_{i,j,k,l}$, $\otimes$ is the tensor-matrix product, and $\langle . \rangle_{\mathcal{F}}$ is the Frobenius inner-product between matrices of size $m \times n$. In general, evaluating this objective costs $O(m^2 n^2)$ operations, which prevents solving the problem even in moderate scale settings. In the particular case $p = 2$, Problem (3.3) can be rewritten in the equivalent form

$$\inf_{\omega \in \Pi(a,b)} \langle c_{C^x, C^y}, \omega \rangle_{\mathcal{F}} - 2 \langle C^x \omega (C^y)^T, \omega \rangle_{\mathcal{F}} \ , \tag{3.4}$$

where $c_{C^x, C^y} = (C^x)^2 a \mathbb{1}_n^T + \mathbb{1}_m b^T (C^y)^2$. This objective can in that case be computed using $O(m^2 n + n^2 m)$ operations instead of $O(m^2 n^2)$, see Peyré et al. (2016, Proposition 1), which is already a significant gain in terms of complexity. This latter problem shares strong connections with the *graph matching problem* that in its standard form reads as

$$\sup_{\sigma \in \mathrm{Perm}(n)} \langle C_1 \sigma C_2^T, \sigma \rangle_{\mathcal{F}} \ , \tag{3.5}$$

where C_1 and C_2 are two matrices of size $n \times n$. One way to approximate solutions of (3.5) is to relax the combinatorial nature of the problem, by expanding the constraint set to the convex-hull of $\mathrm{Perm}(n)$, i.e. to exactly $\Pi(\mathbb{1}_n, \mathbb{1}_n)$ (Birkhoff, 1946). Thus, Problem (3.4) can be seen, in the case where $m = n$ and $a = b = \frac{1}{n}$, as the convex relaxation of Problem (3.5), namely as a *soft graph matching problem*. Finally Problem (3.4) can be rewritten in the standard QP form

$$\inf_{\omega \in \Pi(a,b)} \frac{1}{2} \mathrm{vec}^T(\omega) Q \mathrm{vec}(\omega) + \mathrm{vec}^T(c_{C^x, C^y}) \mathrm{vec}(\omega) \ , \tag{3.6}$$

where for a matrix A of size $m \times n$, $\mathrm{vec}(A)$ is the *vector operator*, i.e. the vector in $\mathbb{R}^{mn}$ obtained by stacking the columns of A, and $Q = -4 C^x \otimes_K C^y$ where $\otimes_K$ denotes the Kronecker product.

Algorithmic solution. Problem (3.6) can be thought as a classic OT problem with a non-convex quadratic regularization. The non-convexity of the regularization motivates the use of a *Conditional Gradient algorithm*, also known as the Frank-Wolfe algorithm (Frank et al., 1956). This algorithm consists in first deriving the first-order Taylor approximation of the objective at current estimate ω, that reads as

$$D_\omega GW(\omega) = L(C^x, C^y) \otimes \omega + L((C^x)^T, (C^y)^T) \otimes \omega \ ,$$

then in finding a descent direction by minimizing a classic OT problem with ground cost $D_\omega GW(\omega)$, using the network simplex algorithm, then in updating the estimate of the coupling ω with a line-search that boils down to a constrained minimization of a second degree polynomial function admitting a closed form solution. This gives Algorithm 2. This latter algorithm is known to converge to a local minima with a rate of $O(N^{-\frac{1}{2}})$, where N denotes the number of iterations (Lacoste-Julien, 2016). Yet, its principal bottleneck is the network simplex step that has a complexity of $O(n^3 \log(n))$.

Algorithm 2 Conditional gradient algorithm for non-regularized GW problem

Require: a, b, C^x, C^y, $\omega^{\{0\}} = ab^T$
1: **while** not converged **do**
2: $\quad C \leftarrow D_{\omega^{\{i-1\}}} GW_{a,b,C^x,C^y}(\omega^{\{i-1\}})$ ▷ Derive first-order Taylor approximation
3: $\quad \zeta \leftarrow$ NETWORK-SIMPLEX(a, b, C) ▷ Find a direction by solving a classic OT problem
4: $\quad \omega^{\{i\}} \leftarrow$ LINE-SEARCH$(\omega^{\{i-1\}}, \zeta)$ ▷ Line-search between $\omega^{\{i\}}$ and ζ
5: **end while**
6: **return** ω

Entropic Regularization. Peyré et al. (2016) and Solomon et al. (2016) have proposed to solve an *entropic-regularized* version of Problem (3.1):

$$\inf_{\omega \in \Pi(a,b)} \sum_{i,j,k,l} |C^x_{i,k} - C^y_{j,l}|^p \omega_{i,j} \omega_{k,l} - \varepsilon H(\omega) \ . \tag{ε-GW_p}$$

This latter problem can be solved using a projected gradient descent scheme, where each update consists in deriving the first-order Taylor approximation $D_\omega GW(\omega)$ of the objective at current estimate ω as for the non-regularized problem, then solving a regularized classic OT problem with ground cost $D_\omega GW(\omega)$ using the Sinkhorn-Knopp algorithm. This is summarized in Algorithm 3. Although this algorithm works well in practice and always leads to a converging sequence of ω, there is to the best of our knowledge no theory guaranting the convergence of this algorithm. Its overall theoretical complexity is in $O(n^3)$. Yet recently, Scetbon et al. (2022) has proposed a refinement using low-rank approximations (as for the classic regularized OT problem) of both cost and couplings, which results in a solver that approximates GW with a linear complexity in time and memory.

Algorithm 3 Entropic-regularized GW solver

Require: a, b, C^x, C^y, $\varepsilon > 0$, $\omega^{\{0\}} = ab^T$
1: **for** $i = 1, \dots, N_{it}$ **do**
2: $\quad C \leftarrow D_{\omega^{\{i-1\}}} GW_{a,b,C^x,C^y}(\omega^{\{i-1\}})$ ▷ Derive first-order Taylor approximation
3: $\quad \omega^{\{i\}} \leftarrow$ SINKHORN-KNOPP(a, b, C, ε) ▷ Solve a regularized OT problem using Algorithm 1
4: **end for**
5: **return** ω

Computing a lower bound. Initially, Mémoli (2011) had proposed to optimize the following lower bound instead of Problem (3.1), referred to as the *Third Lower Bound* (TLB),

$$\inf_{\omega \in \Pi(a,b)} \sum_{k,l} W_p^p(\mu_k, \nu_l) \omega_{k,l} \ , \tag{TLB}$$

where $\mu_k = \sum_i a_i \delta_{C^x_{i,k}}$ and $\nu_l = \sum_j b_j \delta_{C^y_{j,l}}$ are discrete probability measures on $\mathbb{R}$. The interest of this lower bound is that it can be solved using only tools of classic optimal transport theory, more precisely by solving consecutively two classic OT problems, where the second is between one-dimensional distributions and thus can be solved using a simple sorting algorithm.

Other solvers. There exist numerous alternative solvers for the Gromov-Wasserstein problem that build either on approximations or alternate formulations of the initial problem (3.1). In the Euclidean setting, one can notably cite the *sliced Gromov-Wasserstein distance* (Vayer et al., 2019b) that builds on similar ideas as the sliced Wasserstein distance (Rabin et al., 2012). Other solvers consist in reducing the size of the GW problem, either through quantization of input measures (Chowdhury et al., 2021), or by recursive clustering approaches (Xu et al., 2019a; Blumberg et al., 2020).

3.2 Other formulations

We introduce here two other OT distances that have been respectively introduced by Alvarez-Melis et al. (2019) and Cai and Lim (2022). In this thesis, we call them respectively *invariant Wasserstein* discrepancy

(IW) and *projection Wasserstein* discrepancy (PW). These two approaches differ from GW mainly in the fact that they both encode explicitely their invariances whereas the invariance to isometries in GW is encoded implicitly. Note also that, in contrast to Gromov-Wasserstein, these two OT distances are only defined in the Euclidean setting.

3.2.1 Invariant Wasserstein discrepancy

We start by the OT distance proposed Alvarez-Melis et al. (2019) that we call *invariant Wasserstein* discrepancy. Initially, Alvarez-Melis et al. (2019) have introduced this OT distance in the setting where μ and ν are both living in the same Euclidean space $\mathbb{R}^d$. Yet, it generalizes well to settings where μ and ν are living in spaces of different dimensions.

Problem statement. Alvarez-Melis et al. (2019) propose to solve the following problem, between two centered measures μ and ν on $\mathbb{R}^d$,

$$IW_2^2(\mathcal{H},\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \inf_{h\in\mathcal{H}} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \|x-h(y)\|^2 \mathrm{d}\pi(x,y) \;, \tag{IW_2}$$

where $\mathcal{H}$ is a class of mappings from $\mathbb{R}^{d'}$ to $\mathbb{R}^d$ encoding the invariance. This is a *non-convex* optimization problem in π and h that becomes convex in π if h is fixed and becomes also convex in h if π is fixed and $\mathcal{H}$ is a convex set.

Equivalent formulations. When d is equal to d', Alvarez-Melis et al. (2019) have notably shown that when ν is such that $\mathbb{E}_{Y\sim\nu}[YY^T] = \mathrm{Id}_d$ and when $\mathcal{H} = \mathcal{H}_1 := \{P\in\mathbb{R}^{d\times d} \;:\; \|P\|_{\mathcal{F}} \leq \sqrt{d}\}$, Problem ($IW_2$) is equivalent to the Gromov-Wasserstein problem (GW_p) of order 2 with inner-product as cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$. Indeed, it can be shown (see Chapter 4 for details) that both problems are equivalent in that case to

$$\sup_{\pi\in\Pi(\mu,\nu)} \left\| \int_{\mathbb{R}^d\times\mathbb{R}^d} xy^T \mathrm{d}\pi(x,y) \right\|_{\mathcal{F}} \;, \tag{$\mathcal{F}$-COV}$$

where for any matrix A of size $d\times d$, $\|A\|_{\mathcal{F}}$ denotes the Frobenius norm, i.e. $\sqrt{\mathrm{tr}(A^TA)}$. Another interesting case is when $\mathcal{H} = \mathcal{H}_2 := \mathbb{O}(\mathbb{R}^d) = \{P\in\mathbb{R}^{d\times d} \;:\; P^TP = \mathrm{Id}_d\}$ is the set of orthogonal matrices of size $d\times d$. In that case, Problem (IW_2) is equivalent to

$$\sup_{\pi\in\Pi(\mu,\nu)} \left\| \int_{\mathbb{R}^d\times\mathbb{R}^d} xy^T \mathrm{d}\pi(x,y) \right\|_{*} \;, \tag{$*$-COV}$$

where for any matrix A of size $d\times d$, $\|A\|_*$ is the nuclear norm of A, i.e. $\|A\|_* = \mathrm{tr}((A^TA)^{\frac{1}{2}})$. Note that both Problems ($\mathcal{F}$-COV) and ($*$-COV) are *non-convex*. These results have been shown by Alvarez-Melis et al. (2019) in the case where μ and ν are discrete but can easily be extended to continuous distributions. Observe that problem ($*$-COV) consists in maximizing the sum of the singular values of the cross-covariance matrix $\int xy^T\mathrm{d}\pi(x,y)$, whereas the Problem ($\mathcal{F}$-COV) consists in maximizing the sum of the squared singular values of the cross-covariance matrix. In general, these two problems are not equivalent despite being structurally similar, as the example of Figure 3.1 illustrates it.

3.2.2 Projection Wasserstein discrepancy

We introduce now the OT distance proposed by (Cai and Lim, 2022), that we call here *projection Wasserstein* discrepancy.

Problem statement. Cai and Lim (2022) have proposed the following OT distance between two measures $\mu\in\mathcal{P}(\mathbb{R}^d)$ and $\nu\in\mathcal{P}(\mathbb{R}^{d'})$, supposing that $d\geq d'$

$$PW_2(\mu,\nu) = \inf_{\phi\in\Gamma_d(\mathbb{R}^{d'})} W_2(\phi_{\#}\mu,\nu) \;, \tag{PW_2}$$

where $\Gamma_d(\mathbb{R}^{d'})$ is the set of affine mappings ϕ from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$ such that for all $x\in\mathbb{R}^d$, $\phi(y) = P^T(x-b)$ where $b\in\mathbb{R}^d$ and where P is in the *Stiefel manifold* (James, 1976), i.e. the set of orthogonal d'-frames

$$\mathbb{V}_{d'}(\mathbb{R}^d) = \{P\in\mathbb{R}^{d\times d'} \;:\; P^TP = \mathrm{Id}_{d'}\} \;. \tag{3.7}$$

The projection Wasserstein discrepancy consists thus in projecting the measure living in the larger space to the smaller space while finding the "best" projection possible.

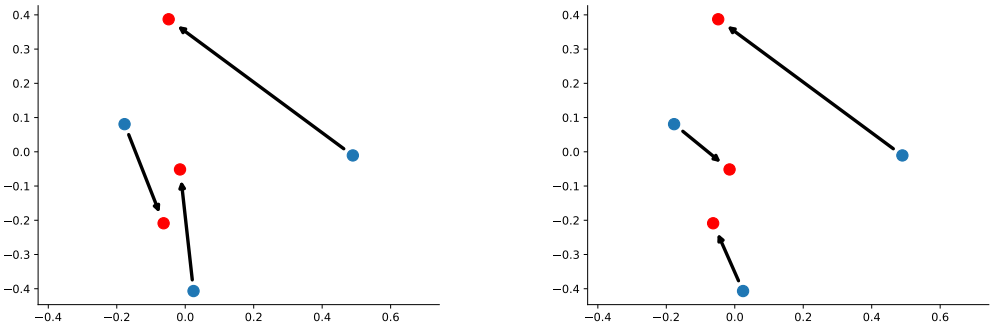

Figure 3.1: Transport plans between two discrete centered distributions on $\mathbb{R}^2$ composed of three points. Left: optimal coupling given by the maximization of Problem ($\mathcal{F}$-COV). Right: optimal coupling given by the maximization of Problem ($*$-COV).

Equivalent formulation. One of the key results of Cai and Lim (2022) is to show that PW_2 has the following equivalent formulation

$$PW_2(\mu,\nu) = \inf_{\xi \in \mathcal{P}^\nu(\mathbb{R}^d)} W_2(\mu,\xi) \;, \tag{3.8}$$

where $\mathcal{P}^\nu(\mathbb{R}^d)$ is the subset of $\mathcal{P}(\mathbb{R}^d)$ defined as

$$\mathcal{P}^\nu(\mathbb{R}^d) = \{\xi \in \mathcal{P}(\mathbb{R}^d) \mid \text{there exists } \phi(x) = P^T(x-b) \text{ with } P \in \mathbb{V}_{d'}(\mathbb{R}^d) \text{ and } b \in \mathbb{R}^d \text{ such that } \phi_\# \xi = \nu\} \;.$$

Note that Problem (3.8) is different to find the "best" affine mapping $\phi : \mathbb{R}^{d'} \to \mathbb{R}^d$ of the form $\phi(y) = Py + b$ with $b \in \mathbb{R}^d$ and $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ for all $y \in \mathbb{R}^{d'}$ since the measure $\phi_\#\nu$ is necessarily degenerate whereas $\mathcal{P}^\nu(\mathbb{R}^d)$ can include measures ξ that are not degenerate.

3.3 Embedded Wasserstein distance

In this section, we define another OT distance between measures living in incomparable spaces that we call *embedded Wasserstein* distance. This OT distance can be seen as the symmetrized mirror construction of the projection Wasserstein discrepancy or as a particular case of the invariant Wasserstein discrepancy. In contrast to the two previously introduced formulations which are not symmetric, we will show that this OT distance defines a pseudometric invariant to isometries.

Motivation. Observe that the equivalence between (IW_2) and ($\mathcal{F}$-COV) generalizes well to cases of different dimensions. Indeed, the equivalence between (IW_2) and ($\mathcal{F}$-COV) or ($*$-COV) holds as soon as it is possible to develop Problem (IW_2) under the following form

$$\int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x) + \int_{\mathbb{R}^{d'}} \|y\|^2 \mathrm{d}\mu(y) - 2 \sup_{\pi \in \Pi(\mu,\nu)} \sup_{h \in \mathcal{H}} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} x h(y)^T \mathrm{d}\pi(x,y) \;,$$

in which the term $\int_{\mathbb{R}^{d'}} \|y\|^2 \mathrm{d}\mu(y)$ doesn't depend on h. In the case where $\mathcal{H} = \{P \in \mathbb{R}^{d \times d'} \;:\; \|P\|_{\mathcal{F}} \leq \sqrt{d'}\}$, this is possible as soon as ν is such that $\mathbb{E}_{Y\sim\nu}[YY^T] = \mathrm{Id}_{d'}$ regardless whether $d \geq d'$ or $d < d'$, see the proof of Corollary 4.3.2 for details. However, generalizing the equivalence between Problem (IW_2) and Problem ($*$-COV) when the measures are living in different dimensions is a bit less immediate. When $d \neq d'$, the natural generalization of $\mathbb{O}(\mathbb{R}^d)$ is the Stiefel manifold $\mathbb{V}_{d'}(\mathbb{R}^d)$, as defined in (3.7). Yet this generalization intrinsically supposes that $d \geq d'$ since it implies that the feasible P are of rank d'. When $d < d'$, there doesn't exist any matrix P of size $d \times d'$ such that $P^T P = \mathrm{Id}_{d'}$. In other words, when $d \geq d'$, there exist *isometries* P from $\mathbb{R}^{d'}$ to $\mathbb{R}^d$, whereas this is not the case when $d < d'$, since any isometry must necessarily be *injective*. Hence, we need to suppose $d \geq d'$ to generalize the equivalence between (IW_2) and ($*$-COV). In that case, observe that Problem (IW_2) corresponds to find an isometric *embedding* of the measure living in the smaller space into the larger space. Informally, this seems to be the mirror construction of the projection Wasserstein distance, that corresponds to project the measure living in the

larger space into the smaller space. This motivates the introduction of the *embedded Wasserstein* distance, which can be seen as the symmetrized mirror construction of the projection Wasserstein discrepancy and which can be defined independently whether $d \geq d'$ or $d < d'$.

Definition 3.3.1. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^{d'})$. For $r \geq 1$ and $s \geq 1$, let us denote $\mathrm{Isom}_s(\mathbb{R}^r)$ the set of all isometries - for the Euclidean norm - from $\mathbb{R}^s$ to $\mathbb{R}^r$. We define*

$$EW_2(\mu,\nu) = \inf\left\{ \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} W_2(\mu, \phi_{\#}\nu), \inf_{\psi \in \mathrm{Isom}_d(\mathbb{R}^{d'})} W_2(\psi_{\#}\mu, \nu) \right\}, \tag{EW_2}$$

with the convention that the infinimum over an empty set is equal to $+\infty$.

Observe that if $d > d'$, the set $\mathrm{Isom}_d(\mathbb{R}^{d'})$ is empty and so $EW_2(\mu,\nu) = \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} W_2(\mu, \phi_{\#}\nu)$. In contrast, if $d < d'$, $\mathrm{Isom}_{d'}(\mathbb{R}^d)$ is empty and so $EW_2(\mu,\nu) = \inf_{\psi \in \mathrm{Isom}_d(\mathbb{R}^{d'})} W_2(\psi_{\#}\mu, \nu)$. When $d = d'$, the two infinimums are equivalent. In all what follows, we will suppose without any loss of generality that $d \geq d'$. More generally, one can define, given two - not necessarily Euclidean - Polish spaces $\mathcal{X}$ and $\mathcal{Y}$ each endowed with respective distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$, and given two measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the following OT distance

$$EW_p(\mu,\nu) = \inf\left\{ \inf_{\phi \in \mathrm{Isom}_{\mathcal{Y}}(\mathcal{X})} W_p(\mu, \phi_{\#}\nu), \inf_{\psi \in \mathrm{Isom}_{\mathcal{X}}(\mathcal{Y})} W_p(\psi_{\#}\mu, \nu) \right\}, \tag{EW_p}$$

where $p \geq 1$ and $\mathrm{Isom}_{\mathcal{Y}}(\mathcal{X})$ (respectively $\mathrm{Isom}_{\mathcal{X}}(\mathcal{Y})$) is the set of all isometries from $\mathcal{Y}$ to $\mathcal{X}$ (respectively from $\mathcal{X}$ to $\mathcal{Y}$), i.e. such that $d_{\mathcal{X}}(\phi(y), \phi(y')) = d_{\mathcal{Y}}(y, y')$ for every $y, y' \in \mathcal{Y}$, and W_p is the Wasserstein distance of order p on $\mathcal{P}(\mathcal{X})$. However there might be cases where both set $\mathrm{Isom}_{\mathcal{Y}}(\mathcal{X})$ and $\mathrm{Isom}_{\mathcal{X}}(\mathcal{Y})$ are empty. In that case, by convention $EW_p(\mu,\nu) = +\infty$.

3.3.1 Links with invariant and projection Wasserstein discrepancies

We give now details on the links between EW_2 and the two other OT distances defined in the previous section. First, observe that EW_2 is the symmetrized mirror construction of PW_2, as a direct consequence of the following result, which is itself a consequence of the Mazur-Ulam theorem (Mazur and Ulam, 1932) which implies that any isometry from $\mathbb{R}^{d'}$ to $\mathbb{R}^d$ - both endowed with the Euclidean norms - is necessarily affine. The proof of this lemma is postponed to Appendix A.

Lemma 3.3.2. *Suppose $d \geq d'$. Then $\phi : \mathbb{R}^{d'} \to \mathbb{R}^d$ is an isometry for the Euclidean norm if and only if there exists $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ and $b \in \mathbb{R}^d$ such that for all $y \in \mathbb{R}^{d'}$, ϕ is of the form*

$$\phi(y) = Py + b\,.$$

Hence, the embedded Wasserstein distance consists in finding an isometry ϕ from the smaller space $\mathbb{R}^{d'}$ to the larger space $\mathbb{R}^d$, which is necessarily of the form $\phi(y) = Py + b$ for all $y \in \mathbb{R}^{d'}$, with $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ and $b \in \mathbb{R}^d$, while the projection Wasserstein discrepancy consists in finding a mapping ψ from the larger space $\mathbb{R}^d$ to the smaller space $\mathbb{R}^{d'}$ of the form $\psi(x) = P^T(x - b)$ for all $x \in \mathbb{R}^d$. This mirror structure between PW_2 and EW_2 is illustrated in Figure 3.2. Now, we show that EW_2 can in fact be seen as a particular case of IW_2.

Proposition 3.3.3. *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ and let suppose $d \geq d'$. Then,*

$$EW_2^2(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d),\ b \in \mathbb{R}^d} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - Py - b\|^2 \mathrm{d}\pi(x,y)\,. \tag{3.9}$$

This result is a consequence of Lemma 3.3.2 and of the following result by Delon et al. (2022).

Lemma 3.3.4 (Delon et al., 2022). *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ with d not necessarily greater than d', and let $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a measurable map. Then $\pi' \in \Pi(\mu, T_{\#}\nu)$ if and only if there is some $\pi \in \Pi(\mu,\nu)$ such that $\pi' = (\mathrm{Id}_d, T)_{\#}\pi$. In particular, if there exist $a, b \geq 0$ such that $\|T(y)\| \leq a + b\|y\|$ for all $y \in \mathbb{R}^{d'}$, then*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - T(y)\|^2 \mathrm{d}\pi(x,y) = \inf_{\pi \in \Pi(\mu, T_{\#}\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - z\|^2 \mathrm{d}\pi(x,z)\,.$$

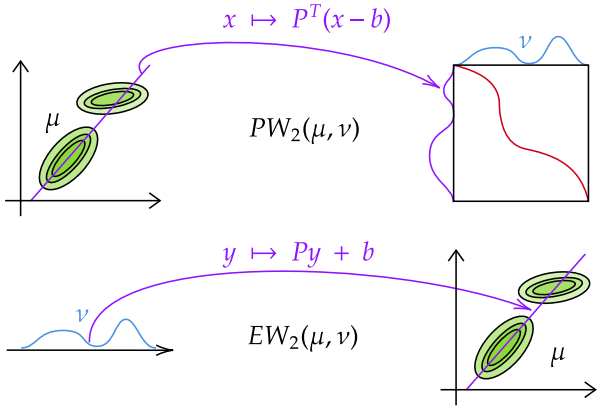

Figure 3.2: Link between PW_2 and EW_2 for two distributions μ and ν respectively on $\mathbb{R}^2$ and $\mathbb{R}$. In PW_2, μ is projected into $\mathbb{R}$ by a mapping of the form $x \mapsto P^T(x-b)$. In EW_2, ν is transformed into a degenerate measure (lying on the purple line) on $\mathbb{R}^2$ with an isometric mapping of the form $y \mapsto Py + b$.

Proof of Proposition 3.3.3. Since we suppose $d \geq d'$, we have

$$EW_2^2(\mu,\nu) = \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} W_2^2(\mu, \phi_\# \nu) \ .$$

Let $\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)$ for the Euclidean norm. Using Lemma 3.3.2, we get that there exists $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ and $b \in \mathbb{R}^d$ such that for all $y \in \mathbb{R}^{d'}$, $\phi(y) = Py + b$. Moreover, we have, using Lemma 3.3.4,

$$\begin{aligned} EW_2^2(\mu,\nu) &= \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} \inf_{\pi \in \Pi(\mu, \phi_\# \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \mathrm{d}\pi(x,y) \\ &= \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^{d'} \times \mathbb{R}^d} \|x-\phi(y)\|^2 \mathrm{d}\pi(x,y) \\ &= \inf_{\pi \in \Pi(\mu,\nu)} \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d),\ b \in \mathbb{R}^d} \int_{\mathbb{R}^{d'} \times \mathbb{R}^d} \|x - Py - b\|^2 \mathrm{d}\pi(x,y) \ , \end{aligned}$$

which concludes the proof. □

Observe that Problem (3.9) is very similar to the $IW_2(\mathbb{V}_{d'}(\mathbb{R}^d), \mu, \nu)$ problem presented in the previous section and differs only by the introduction of the variable $b \in \mathbb{R}^d$ that allows to handle non-centered distributions. Observe also that, as soon as $\mathcal{H}$ is such that for any $h \in \mathcal{H}$, there exists $a, b \geq 0$ such that $\|h(y)\| \leq a + b\|y\|$ for every $y \in \mathbb{R}^{d'}$, one can directly put Problem (IW_2) under the following form using Lemma 3.3.4:

$$IW_2(\mathcal{H}, \mu, \nu) = \inf_{h \in \mathcal{H}} W_2(\mu, h_\# \nu) \ .$$

Thus, we can see from this formulation that PW_2 can also be seen as a particular case of IW_2. Finally, one can show from Proposition 3.3.3 that the infinimum in ϕ in (EW_2) is in fact always achieved.

Corollary 3.3.5. *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ and let suppose $d \geq d'$. Then there exists an optimal isometry $\phi^* : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that $EW_2(\mu,\nu) = W_2(\mu, \phi^*_\# \nu)$.*

This result is a consequence of Lemma 3.3.4 and of the following result, whose proof is postponed to Appendix A.

Lemma 3.3.6. *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ with d not necessarily greater than d'. Let $\bar{\mu}$ and $\bar{\nu}$ denote the centered measures associated to μ and ν and let $\mathfrak{P}$ be any subset of matrices of size $d \times d'$. Then,*

$$\inf_{\pi \in \Pi(\mu,\nu)} \inf_{P \in \mathfrak{P},\ b \in \mathbb{R}^d} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - Py - b\|^2 \mathrm{d}\pi(x,y) = \inf_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \inf_{P \in \mathfrak{P}} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - Py\|^2 \mathrm{d}\pi(x,y) \ .$$

Proof of Corollary 3.3.5. Using Lemma 3.3.6 and Lemma 3.3.4, we have that

$$EW_2^2(\mu,\nu) = \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \inf_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - Py\|^2 \mathrm{d}\pi(x,y)$$

$$= \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} W_2^2(\bar{\mu}, P_\# \bar{\nu}) \ ,$$

where $\bar{\mu}$ and $\bar{\nu}$ are the centered measures associated with μ and ν. Let us denote $J : P \mapsto W_2(\bar{\mu}, P_\# \bar{\nu})$ and let us show that J is continuous. For any P_0 and P_1 in $\mathbb{V}_{d'}(\mathbb{R}^d)$, we have,

$$|J(P_0) - J(P_1)| = |W_2(\bar{\mu}, P_{0\#}\bar{\nu}) - W_2(\bar{\mu}, P_{1\#}\bar{\nu})| \leq W_2(P_{0\#}\bar{\nu}, P_{1\#}\bar{\nu}) \ ,$$

where we used the triangular inequality property of W_2. Furthermore,

$$\begin{aligned} W_2^2(P_{0\#}\bar{\nu}, P_{1\#}\bar{\nu}) &= \inf_{\pi \in \Pi(P_{0\#}\bar{\nu}, P_{1\#}\bar{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) \\ &= \inf_{\pi \in \Pi(\bar{\nu}, \bar{\nu})} \int_{\mathbb{R}^{d'} \times \mathbb{R}^{d'}} \|P_0 x - P_1 y\|^2 \mathrm{d}\pi(x, y) \ , \end{aligned}$$

where we used Lemma 3.3.4 twice. Now observe that the coupling $(\mathrm{Id}_{d'}, \mathrm{Id}_{d'})_\# \bar{\nu}$ is in $\Pi(\bar{\nu}, \bar{\nu})$, so it follows

$$\inf_{\pi \in \Pi(\bar{\nu}, \bar{\nu})} \int_{\mathbb{R}^{d'} \times \mathbb{R}^{d'}} \|P_0 x - P_1 y\|^2 \mathrm{d}\pi(x, y) \leq \int_{\mathbb{R}^{d'}} \|P_0 x - P_1 x\|^2 \mathrm{d}\bar{\nu}(x) \ .$$

Finally, for any $x \in \mathbb{R}^{d'}$, we have

$$\|P_0 x - P_1 x\|^2 \leq \|x\|^2 \sup_{\|z\|=1} \|(P_0 - P_1)z\|^2 \leq \|P_0 - P_1\|_{\mathcal{F}}^2 \|x\|^2 \ ,$$

and so it follows that

$$|J(P_0) - J(P_1)|^2 \leq \|P_0 - P_1\|_{\mathcal{F}}^2 \int_{\mathbb{R}^n} \|x\|^2 \mathrm{d}\bar{\nu} \ .$$

Since ν is in $\mathcal{W}_2(\mathbb{R}^{d'})$, $\bar{\nu}$ is in $\mathcal{W}_2(\mathbb{R}^{d'})$ and so $\int_{\mathbb{R}^{d'}} \|x\|^2 \mathrm{d}\bar{\nu} < +\infty$. It follows that $|J(P_0) - J(P_1)| \longrightarrow 0$ when $\|P_0 - P_1\|_{\mathcal{F}}^2 \longrightarrow 0$ and so J is continuous. Moreover, since $\mathbb{V}_{d'}(\mathbb{R}^d)$ is compact (James, 1976), J has a minimum on $\mathbb{V}_{d'}(\mathbb{R}^d)$ as a result of the classic Weierstrass theorem that states that any real-valued continous function defined on a compact set achieves its infinimum. Thus, there exists P^* such that $EW_2(\mu, \nu) = W_2(\bar{\mu}, P^*_\# \bar{\nu})$ and setting $b^* = \mathbb{E}_{X \sim \mu}[X] - P^* \mathbb{E}_{Y \sim \nu}[Y]$ and $\phi^*(x) = P^* x + b^*$ for all $x \in \mathbb{R}^d$, we get that there exists $\phi^* \in \mathrm{Isom}_{d'}(\mathbb{R}^d)$ such that $EW_2(\mu, \nu) = W_2(\mu, \phi^*_\# \nu)$, which concludes the proof. □

3.3.2 Equivalent formulations of the embedded and projection Wasserstein problems

Here we derive equivalent problems to respectively Problems (EW_2) and (PW_2). We start with (EW_2).

Equivalent problem for embedded Wasserstein distance. We show that Problem (EW_2) is equivalent to Problem ($*$-COV) between the centered measures $\bar{\mu}$ and $\bar{\nu}$. In what follows, for any matrix A of size $r \times s$ with $r \leq d$ and $s \leq d'$, we denote $A^{[d,d']}$ the matrix of size $d \times d'$ of the form

$$A^{[d,d']} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \ .$$

Proposition 3.3.7. *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ and let suppose $d \geq d'$. Problem (EW_2) is equivalent to*

$$\sup_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \sup_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \mathrm{tr}(P^T K_\pi) \ , \tag{3.10}$$

where $\bar{\mu}$ and $\bar{\nu}$ denotes the centered measures associated with μ and ν and $K_\pi = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x, y)$. Furthermore, solving this latter problem in P for a fixed π, (3.10) reduces to

$$\sup_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \left\| \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x, y) \right\|_* \ ,$$

and this is achieved at

$$P^*_\pi = U_\pi \, \mathrm{Id}_{d'}^{[d,d']} \, V_\pi^T \ ,$$

where $U_\pi \in \mathbb{O}(\mathbb{R}^d)$ and $V_\pi \in \mathbb{O}(\mathbb{R}^{d'})$ are the left and right orthogonal matrices associated with the Singular Value Decomposition (SVD) of K_π.

This is a consequence of the following technical result, whose proof is postponed to Appendix A.

Lemma 3.3.8. *Let K be a matrix of size $d \times d'$ with Singular Value Decomposition (SVD) $K = U_K \Sigma_K V_K^T$ and let $\mathfrak{P}$ be any compact set of matrices of size $d \times d'$. Then,*

$$\sup_{P \in \mathfrak{P}} \operatorname{tr}(P^T K) = \max_{P \in \mathfrak{P}} \operatorname{tr}(\Sigma_P^T \Sigma_K) \ ,$$

where $\Sigma_P = \operatorname{diag}^{[d,d']}(\boldsymbol{\sigma}(P))$ with $\boldsymbol{\sigma}(P) \in \mathbb{R}_+^{d'}$ denoting the vector of singular values of P. Furthermore it is achieved at P of the form,

$$P = U_K \Sigma_P V_K^T \ .$$

Proof of Proposition 3.3.7. Using Lemmas 3.3.6, (3.9) can be rewritten

$$\begin{aligned} EW_2^2(\mu,\nu) &= \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \inf_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - Py\|^2 \mathrm{d}\pi(x,y) \\ &= \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \inf_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \left(\|x\|^2 + \|Py\|^2 - 2\langle x, Py \rangle \right) \mathrm{d}\pi(x,y) \ . \end{aligned}$$

Since for all $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$, $\|Py\|$ doesn't depend on P, we get that the problem is equivalent to

$$\sup_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \sup_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \langle x, Py \rangle \mathrm{d}\pi(x,y) \ .$$

Now observe that for all $\pi \in \Pi(\bar{\mu},\bar{\nu})$,

$$\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \langle x, Py \rangle \mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \operatorname{tr}(xy^T P^T) \mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \operatorname{tr}(P^T xy^T) \mathrm{d}\pi(x,y) \ ,$$

where we used the cyclical permutation property of the trace operator. Finally using the linearity of the trace, we get that the problem is equivalent to

$$\sup_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \sup_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \operatorname{tr}\left(P^T \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x,y) \right) \ ,$$

or equivalently,

$$\sup_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \sup_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \left\langle P, \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x,y) \right\rangle \ .$$

Now, using Lemma 3.3.8 and using the fact that if $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$, $\boldsymbol{\sigma}(P) = \mathbb{1}_{d'}$, we get that the problem reduces to

$$\sup_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \left\| \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x,y) \right\|_* \ ,$$

and this is achieved for $P^* = U_\pi \operatorname{Id}_{d'}^{[d,d']} V_\pi^T$, where $U_\pi \in \mathbb{O}(\mathbb{R}^d)$ and $V_\pi \in \mathbb{O}(\mathbb{R}^{d'})$ are respectively the left and right orthogonal matrices of the SVD of $\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x,y)$, which concludes the proof. □

Observe that P_π^* is the projection of K_π on the Stiefel manifold $\mathbb{V}_{d'}(\mathbb{R}^d)$ since $\operatorname{tr}(P^T K_\pi)$ is the Frobenius inner-product between P and K_π and so maximizing the inner-product is equivalent to minimizing the distance between K_π and P since $\|P\|_{\mathcal{F}}$ is necessarily equal to d'.

Equivalent problem for projection Wasserstein discrepancy. To highlight the difference between (EW_2) and (PW_2), we also derive similarly an equivalent problem for (PW_2). Observe that in that case, the mapping ϕ in (PW_2) is not an isometry since it is not injective. As a result, the term that previously depended only on the marginal μ in the developpement of the square of the Euclidean distance will now depend on P. More precisely, this gives the following result.

Proposition 3.3.9. *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ and let suppose $d \geq d'$. Problem (PW_2) is equivalent to*

$$\inf_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \left(\operatorname{tr}(P^T \Sigma_x P) - 2\operatorname{tr}(P^T K_\pi) \right) \ , \tag{3.11}$$

where $\Sigma_x = \int_{\mathbb{R}^d \times \mathbb{R}^d} xx^T d\bar{\mu}(x)$, $K_\pi = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T d\pi(x,y)$, and where $\bar{\mu}$ and $\bar{\nu}$ are the centered measures associated with μ and ν.

Proof of Proposition 3.3.9. First observe that using Lemma 3.3.6, we can consider without any loss generality that μ and ν are centered and omit b. Using Lemma 3.3.4, it follows

$$\begin{aligned} PW_2^2(\mu,\nu) &= \inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \inf_{\pi'\in\Pi(P^T_\#\mu,\nu)} \int_{\mathbb{R}^{d'}\times\mathbb{R}^{d'}} \|z-y\|^2 \mathrm{d}\pi'(z,y) \\ &= \inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^{d}\times\mathbb{R}^{d'}} \|P^Tx-y\|^2 \mathrm{d}\pi(x,y) \\ &= \inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \left(\int_{\mathbb{R}^d} \|P^Tx\|^2 \mathrm{d}\mu(x) + \int_{\mathbb{R}^{d'}} \|y\|^2 \mathrm{d}\nu(y) - 2 \sup_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} (P^Tx)^T y \mathrm{d}\pi(x,y) \right) , \end{aligned}$$

and so the problem is equivalent to

$$\inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \left(\int_{\mathbb{R}^d} \|P^Tx\|^2 \mathrm{d}\mu(x) - 2 \sup_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^{d'}\times\mathbb{R}^{d'}} (P^Tx)^T y \mathrm{d}\pi(x,y) \right) ,$$

which is itself equivalent to (3.11), which concludes the proof. □

Observe that Problem (3.11) can be interpreted as a regularization in P of Problem (3.10). It can also be interpreted as a W_2 problem between ν and a measure μ' which has a different second-order moment than μ.

3.3.3 Metric properties of EW_2

Here we show that EW_2 satisfies all the axioms of a pseudometric on $\bigsqcup_{k\geq 1} \mathcal{W}_2(\mathbb{R}^k)$ that is invariant to isometries for the Euclidean norm.

Theorem 3.3.10. *In the following, $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$. Then,*

(i) EW_2 is symmetric, non-negative and satisfies the triangular inequality, i.e. for any $\xi \in \mathcal{W}_2(\mathbb{R}^{d''})$,

$$EW_2(\mu,\nu) \leq EW_2(\mu,\xi) + EW_2(\xi,\nu) .$$

(ii) $EW_2(\mu,\nu) = 0$ if and only if there exists an isometry $\phi : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that $\nu = \phi_\#\mu$.

Thus EW_2 defines a pseudometric on $\bigsqcup_{k\geq 1} \mathcal{W}_2(\mathbb{R}^k)$.

In the proof of this theorem, we will use the following intermediary result, whose proof is postponed to Appendix A.1.

Lemma 3.3.11. *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ with d not necessarily greater than d'. Let $r \geq \max\{d,d'\}$ and let $\psi \in \mathrm{Isom}_d(\mathbb{R}^r)$. Then, $EW_2(\mu,\nu) = EW_2(\psi_\#\mu,\nu)$.*

Proof of Theorem 3.3.10. First observe that non-negativity is straightforward. Furthermore, observe also that if $d \neq d'$, symmetry is also straightfoward. Now suppose $d = d'$ and observe that that the set $\mathbb{V}_{d'}(\mathbb{R}^d)$ coincides with the set of orthogonal matrices $\mathbb{O}(\mathbb{R}^d)$. Thus we have

$$\begin{aligned} \inf_{\phi\in\mathrm{Isom}_d(\mathbb{R}^d)} W_2(\mu,\phi_\#\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \inf_{P\in\mathbb{O}(\mathbb{R}^d),\ b\in\mathbb{R}^d} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-Py-b\|^2 \mathrm{d}\pi(x,y) \\ &= \inf_{\pi\in\Pi(\mu,\nu)} \inf_{P\in\mathbb{O}(\mathbb{R}^d),\ b\in\mathbb{R}^d} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|P^Tx-y-P^Tb\|^2 \mathrm{d}\pi(x,y) \\ &= \inf_{\psi\in\mathrm{Isom}_d(\mathbb{R}^d)} W_2(\psi_\#\mu,\nu) , \end{aligned}$$

and so EW_2 is also symmetric in that case. Before turning to the proof of the two other points, we recall that the infinimum in ϕ is always achieved, see Corollary 3.3.5.

(i) Now we prove the triangle inequality. Let $r \geq \max\{d,d',d''\}$, $\phi_0 \in \mathrm{Isom}_d(\mathbb{R}^r)$ and for $\xi \in \mathcal{W}_\in(\mathbb{R}^{d''})$, let $\phi_1 \in \arg\min_{\phi\in\mathrm{Isom}_{d''}(\mathbb{R}^r)} W_2(\phi_{0\#}\mu,\phi_\#\xi)$. We have, using first Lemma 3.3.11, then using the triangle inequality property of W_2,

$$EW_2(\mu,\nu) = EW_2(\phi_{0\#}\mu,\nu) = \inf_{\phi\in\mathrm{Isom}_{d'}(\mathbb{R}^r)} W_2(\phi_{0\#}\mu,\phi_\#\nu)$$

$$
\begin{aligned}
&\leq \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^r)} [W_2(\phi_{0\#}\mu, \phi_{1\#}\xi) + W_2(\phi_{1\#}\xi, \phi_{\#}\nu)] \\
&\leq W_2(\phi_{0\#}\mu, \phi_{1\#}\xi) + \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^r)} W_2(\phi_{1\#}\xi, \phi_{\#}\nu) \\
&\leq EW_2(\phi_{0\#}\mu, \xi) + EW_2(\phi_{1\#}\xi, \nu) \ .
\end{aligned}
$$

We conclude then by applying Lemma 3.3.11 on both terms.

(ii) Suppose without any loss of generality that $d \geq d'$ and suppose $EW_2(\mu, \nu) = 0$. Since the infinimum in ϕ is achieved, there exists $\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)$ such that $W_2(\mu, \phi_{\#}\nu) = 0$ and so $\mu = \phi_{\#}\nu$. The reverse implication is obvious.

Finally, observe that if μ and ν have finit order 2 moments, then EW_2 necessarily takes finite values, and so EW_2 defines a pseudometric on $\bigsqcup_{k\geq 1} \mathcal{W}_2(\mathbb{R}^k)$. □

Observe that Lemma 3.3.11 highlights that EW_2 shares close connections with the distance between metric measure spaces introduced in Sturm (2006) that can be defined as follows,

$$
D_p((\mathcal{X}, d_{\mathcal{X}}, \mu), (\mathcal{Y}, d_{\mathcal{Y}}, \nu)) = \inf_{\mathcal{Z}, \psi, \phi} W_p(\psi_{\#}\mu, \phi_{\#}\nu) \ , \tag{3.12}
$$

where $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ are two metric measure spaces, $\mathcal{Z}$ is a third Polish space, and where $\psi : \mathcal{X} \to \mathcal{Z}$ and $\phi : \mathcal{Y} \to \mathcal{Z}$ are two isometric mappings. However it is not clear that the two distances are strictly equivalent because the infinimum in $\mathcal{Z}$ in Equation (3.12) also includes non-Euclidean spaces. However, if we restrict the problem to only Euclidean spaces $\mathcal{Z}$, then Lemma 3.3.11 directly implies that the two distances are equivalent.

3.3.4 Case of equivalence with Gromov-Wasserstein

Finally, we exhibit here a particular case where the EW_2 problem is equivalent with the Gromov-Wasserstein problem with inner-product costs, which is when the smaller space is $\mathbb{R}$. Indeed, When ν is a one-dimensional distribution on $\mathbb{R}$, the EW_2 problem is equivalent to the GW_2 problem with inner-products as cost functions, as it is stated in the following result.

Theorem 3.3.12. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R})$. Then the Gromov-Wasserstein problem with inner-products as cost functions between the centered measures $\bar{\mu}$ and $\bar{\nu}$, i.e.*

$$
\inf_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}} \int_{\mathbb{R}^d \times \mathbb{R}} (\langle x, x' \rangle_d - yy')^2 \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y') \ ,
$$

is equivalent to Problem (EW_2) between μ and ν, i.e.

$$
\inf_{\phi \in \mathrm{Isom}_1(\mathbb{R}^d)} W_2(\mu, \phi_{\#}\nu) \ .
$$

Proof. Since $K_\pi = \int_{\mathbb{R}^d \times \mathbb{R}} xy^T \mathrm{d}\pi(x, y)$ is of size $d \times 1$, it has a unique singular value $\lambda_\pi > 0$, and so one can observe that Problems ($\mathcal{F}$-COV) and ($*$-COV), i.e.

$$
\sup_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \|K_\pi\|_{\mathcal{F}} \qquad \text{and} \qquad \sup_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \|K_\pi\|_* \ ,
$$

that are respectively equivalent to the Gromov-Wasserstein problem with inner-product costs between $\bar{\mu}$ and $\bar{\nu}$, and the Embedded Wasserstein problem between μ and ν, can both be rewritten

$$
\sup_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \lambda_\pi \ ,
$$

implying thus that the two problems are equivalent. □

Observe that this theorem allows to recover almost directly the result of Vayer (2020) on Gromov-Wasserstein between one-dimensional distributions presented in Section 3.1.3. However, Theorem 3.3.12 is a bit more general than the result of Vayer (2020) since μ is not necessarily one-dimensional here but of can be of any arbitrary dimension d.

3.4 Conclusion

In this chapter, we have introduced the common generalization of optimal transport to measures living in incomparable spaces, i.e. the Gromov-Wasserstein distance (Mémoli, 2011), as well as the other recent other formulations proposed respectively by Alvarez-Melis et al. (2019) and Cai and Lim (2022). We also have introduced a new formulation, the embedded Wassertein distance, that can be seen as a generalization of a particular case of the OT distance proposed by Alvarez-Melis et al. (2019) to measures living in Euclidean spaces of different dimensions, and as the mirror construction of the OT distance proposed by Cai and Lim (2022). The relatively recent introduction of the Gromov-Wasserstein distance has generalized OT to cases where there doesn't exist any meaningful ground cost to compare the two spaces $\mathcal{X}$ and $\mathcal{Y}$ the distributions are living in. The property of the GW distance of comparing the inherent structures of the distributions rather than the positions of their atoms makes it naturally well-suited for structured data as graphs. Hence the Gromov-Wasserstein distance has been used over the last past years for numerous graph related applications, such as graph matching (Xu et al., 2019b; Petric Maretic et al., 2019; Vincent-Cuaz et al., 2021), graph classification (Vayer et al., 2019a; Jin et al., 2022) or graph prediction (Brogat-Motte et al., 2022). Thanks to its invariance property to isometric transformations, it has also been used in object matching related applications, such as shape matching (Mémoli, 2009; Schmitzer and Schnörr, 2013), cell alignment (Demetci et al., 2020, 2022), or word embedding (Alvarez-Melis and Jaakkola, 2018).

In contrast to classic OT, the theoretical understandings of the Gromov-Wasserstein problem are still relatively nascent. Even on one-dimensional distributions, the behavior of the Gromov-Wasserstein distance is still not well understood in the case the ground costs $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are the squared Euclidean distances, whereas the Wasserstein distance has a simple closed-form solution in that case. The other known simple particular case where the Wasserstein problem admits a closed-form solution is the case of Gaussian distributions. In the case of Gromov-Wasserstein, little was known prior to our work - to the best of our knowledge - on its behavior on Gaussian distributions. Only Vayer (2020) has derived a closed-form solution for the Gromov-Monge problem restricted to linear Monge maps. In order to bridge the gap between the theoretical understandings of the Gromov-Wasserstein and the Wasserstein distances, the goal of Chapter 4 is therefore to derive a closed-form expression of the Gromov-Wasserstein distance between Gaussian distributions. We also continue to establish links with the three formulations presented above. In Chapter 5, we introduce two new Gromov-Wasserstein-like OT distances between Gaussian mixture models, similarly to the work of Delon and Desolneux (2020) with the Wasserstein distance, and illustrate their pratical uses on several Gromov-related tasks.

Chapter 4

The Gromov-Wasserstein distance between Gaussian distributions

Contents

In this chapter, we study the behavior of the Gromov-Wasserstein distance of order 2 between Gaussian distributions possibly living in different dimensions, as well as the two other formulations presented in Section 3.2. This chapter is mostly a reproduction of Salmona et al. (2021) but also contains some results of Salmona et al. (2023).

4.1 Introduction

Most of the time, OT problems cannot be solved analytically and require the use of numerical solvers. However, there are two notable cases, e.g. between one-dimensional or between Gaussian distributions, where the Wasserstein distance admits a closed-form expression and the associated optimal transport plan admits a nice closed-form solution. Despite being relatively simple, these closed-forms have been proved to be very useful for practitioners to design new OT tools or tools inspired by the OT geometry. For instance, the sliced Wasserstein distance (Rabin et al., 2012) builds on the closed-form of OT on one-dimensional distributions by deriving a family of one-dimensional representations for a higher-dimensional probability distribution through linear projections, and then by calculating the distance between two input distributions as a functional of the Wasserstein distance between their one-dimensional representations. This yields to a computationally effective OT distance with nice properties that has therefore attracted a lot of attention. It has been sucessfully applied to a variety of practical tasks, including generative modeling (Kolouri et al., 2018a; Deshpande et al., 2018; Liutkus et al., 2019) and learning GMMs (Kolouri et al., 2018b). On the other hand, the closed-form expression of the Wasserstein distance between Gaussian distributions has inspired the Frechet Inception Distance (FID) (Heusel et al., 2017) that is the most commonly used tool to assess the results of generative models. Furthermore, analogously to the sliced Wasserstein distance with the one-dimensional distributions, Delon and Desolneux (2020) have proposed a composite OT distance between GMMs that leverages the closed-form expression of the W_2

distance between Gaussian distributions and that can also be used as a relatively computationally efficient alternative to the Wasserstein distance.

Similarly to classic OT, the Gromov-Wasserstein distance of order 2 has also a closed-form expression between one-dimensional distributions when the cost functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are both the inner-product on $\mathbb{R}$. Indeed, Vayer (2020) has shown that the GW problem admits two distinct solutions in that case that correspond in the discrete case to the non-increasing and the non-decreasing rearrangement of the points. If this property also seems to hold in most cases when the costs functions $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are the square Euclidean distances instead of the inner-products (Vayer et al., 2019b), Beinert et al. (2022) and Dumont et al. (2022) have shown that one can construct very specific counter-examples where it is not the case. Still, these results and observations motivate the introduction of the *sliced Gromov-Wasserstein distance* (Vayer et al., 2019b) in a similar way to the sliced Wasserstein distance, that leads to a computationally efficient solver for GW problems in the Euclidean setting. In contrast, little was known before this work on the behavior of the Gromov-Wasserstein distance between Gaussian distributions, possibly living in different dimensions. Vayer (2020) has derived a closed-form expression of the Gromov-Monge problem restricted to linear push-forward mappings when the ground costs $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are the squared Euclidean distances. Vayer (2020) has also derived a closed-form solution in case where the two Gaussian distributions are living in the same dimension. Still, these results concern only a very restricted problem and so there is no guarantee at all that the solution found by Vayer (2020) is also solution to the non-restricted Gromov-Wasserstein problem.

It has been known since Dowson and Landau (1982) that the Wasserstein distance of order 2 between Gaussian distributions admits a closed-form solution that is supported by the graph of an affine mapping. This degeneracy of the optimal coupling is in line with the Brenier theorem (Brenier, 1991) which states that as soon as μ is absolutely continuous with respect with the Lesbegue measure, the W_2 problem admits a unique solution that is supported by the graph of a mapping. In the Gromov-Wasserstein case, Dumont et al. (2022) have recently shown that the GW problem with inner-products as cost functions admits at least one solution that is supported by the graph of a mapping, assuming that the supports of the measures μ and ν are compact. Even though the Gaussian distributions do not satisfy the compactness property of their support, it is thus likely to find a solution of the GW problem between Gaussian distributions that is a degenerate transport plan supported by the graph of a mapping. However, when the cost functions are the squared Euclidean distances, it seems that there doesn't exists any result similar to the Brenier theorem. Hence one cannot intuit the form of the solutions of the GW problem in that case.

Contributions of this chapter. In this chapter, we study the behavior of the Gromov-Wasserstein distance between Gaussian distributions, focusing on the cases where the cost functions are the squared Euclidean distances or the inner-products. We also study the behaviors of the other formulations presented in Section 3.2. More precisely, In Section 4.2 we derive lower and upper bounds of the GW problem with squared Euclidean distances as cost functions, then we show that the latter problem restricted to Gaussian couplings admits closed form solutions that are supported by the graph of an affine mapping. These solutions are closely related to Principal Components Analysis (PCA). We then study the tightness of our bounds and exhibit some particular cases where the solution of the restricted problem is also solution of the general problem. Finally, we discuss the form of the solution in the general case. In Section 4.3, we show that the solutions of the previous restricted problem are also solutions of the GW problem with inner-product as cost functions, but now without any restriction on the set of admissible couplings. We also show that these couplings are also solutions, on the one hand, of the invariant Wasserstein discrepancy (Alvarez-Melis et al., 2019) in the particular case where this latter problem is equivalent to Problem ($\mathcal{F}$-COV), and on the the other hand of the embedded Wasserstein distance. Finally, we derive the expression of the projection Wasserstein discrepancy (Cai and Lim, 2022) between two multivariate Gaussian distributions.

4.2 The quadratic case

In this section, we focus on the Gromov-Wasserstein distance of order 2 between two Gaussian measures $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$, respectively on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, i.e. $m_0 \in \mathbb{R}^d$, $m_1 \in \mathbb{R}^{d'}$, $\Sigma_0 \in \mathbb{S}_+^d$ and $\Sigma_1 \in \mathbb{S}_+^{d'}$. More precisely, our goal is to solve the following optimization problem

$$GW_2^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}}\int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \left|\|x-x'\|_{\mathbb{R}^d}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2\right|^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \;, \qquad (GW_2\text{-Q})$$

where $\|.\|_{\mathbb{R}^d}$ and $\|.\|_{\mathbb{R}^{d'}}$ are the Euclidean norms on respectively $\mathbb{R}^d$ and $\mathbb{R}^{d'}$.

4.2.1 Probabilistic formulation

In this section, we derive an equivalent probalistic formulation of Problem (GW_2-Q) which takes the form of a functional of co-moments of order 2 and 4 of π. Note that this formulation is not specific to Gaussian measures but to all measures with finite order 2 moments. More precisely, we show the following result.

Proposition 4.2.1. *Let $\mu \in \mathcal{W}_2(\mathbb{R}^d)$ with mean vector $m_0 \in \mathbb{R}^d$ and covariance matrix $\Sigma_0 \in \mathbb{S}_+^d$ and $\nu \in \mathcal{W}_2(\mathbb{R}^{d'})$ with mean vector m_1 and covariance matrix $\Sigma_1 \in \mathbb{S}_+^{d'}$. Let P_0, D_0 and P_1, D_1 be respective diagonalizations of $\Sigma_0 (= P_0 D_0 P_0^T)$ and $\Sigma_1 (= P_1 D_1 P_1^T)$. Let us define $T_0 : x \mapsto P_0^T(x - m_0)$ and $T_1 : y \mapsto P_1^T(y - m_1)$. Then Problem ($GW_2$-Q) is equivalent to the following problem*

$$\sup_{X \sim T_{0\#}\mu, Y \sim T_{1\#}\nu} \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + 2\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2 \,, \tag{supCOV}$$

where $X = (X_1, X_2, \dots, X_d)^T$, $Y = (Y_1, Y_2, \dots, Y_{d'})^T$. More precisely, (X, Y) is optimal for (supCOV) if and only if the law of $(T_0^{-1}(X), T_1^{-1}(Y))$ is optimal for (GW_2-Q).

This proposition is a direct consequence of the two following intermediary results.

Lemma 4.2.2. *We denote $\mathbb{O}(\mathbb{R}^d) = \{O \in \mathbb{R}^{d\times d} |\ O^T O = \mathrm{Id}_d\}$ the set of orthogonal matrices of size d. Let μ and ν be two probability measures respectively $\mathcal{W}_2(\mathbb{R}^d)$ and $\mathcal{W}_2(\mathbb{R}^{d'})$. Let $T_d : x \mapsto O_d x + x_d$ and $T_{d'} : y \mapsto O_{d'} y + y_{d'}$ be two affine applications with $x_d \in \mathbb{R}^d$, $O_d \in \mathbb{O}(\mathbb{R}^d)$, $y_{d'} \in \mathbb{R}^{d'}$, and $O_{d'} \in \mathbb{O}(\mathbb{R}^{d'})$. Then for $GW_2(T_{d\#}\mu, T_{d'\#}\nu) = GW_2(\mu, \nu)$.*

Lemma 4.2.3 (Vayer, 2020). *Suppose there exist some scalars a, b, c such that $c_{\mathcal{X}}(x, x') = a\|x\|_{\mathbb{R}^d}^2 + b\|x'\|_{\mathbb{R}^{d'}}^2 + c\langle x, x'\rangle_d$, where $\langle ., .\rangle_d$ denotes the inner product on $\mathbb{R}^d$, and $c_{\mathcal{Y}}(y, y') = a\|y\|_{\mathbb{R}^{d'}}^2 + b\|y'\|_{\mathbb{R}^{d'}}^2 + c\langle y, y'\rangle_{d'}$. Let μ and ν be two probability measures respectively in $\mathcal{W}_2(\mathbb{R}^d)$ and $\mathcal{W}_2(\mathbb{R}^{d'})$. Then,*

$$GW_2^2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = C_{\mu,\nu} - 2 \sup_{\pi \in \Pi(\mu,\nu)} Z(\pi) \,,$$

where $GW_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu)$ is defined as in (GW_p), $C_{\mu,\nu} = \int c_{\mathcal{X}}^2 \mathrm{d}\mu\mathrm{d}\mu + \int c_{\mathcal{Y}}^2 \mathrm{d}\nu\mathrm{d}\nu - 4ab \int \|x\|_{\mathbb{R}^d}^2 \|y\|_{\mathbb{R}^{d'}}^2 \mathrm{d}\mu\mathrm{d}\nu$, and

$$\begin{aligned} Z(\pi) = (a^2 + b^2) \int \|x\|_{\mathbb{R}^d}^2 \|y\|_{\mathbb{R}^{d'}}^2 \mathrm{d}\pi(x,y) + c^2 \left\| \int xy^T \mathrm{d}\pi(x,y) \right\|_{\mathcal{F}}^2 \\ + (a+b)c \int \left(\|x\|_{\mathbb{R}^d}^2 \langle \mathbb{E}_{Y\sim\nu}[Y], y\rangle_{d'} + \|y\|_{\mathbb{R}^{d'}}^2 \langle \mathbb{E}_{X\sim\mu}[X], x\rangle_d \right) \mathrm{d}\pi(x,y) \,. \end{aligned} \tag{4.1}$$

Proof of Proposition 4.2.1. Using Lemma 4.2.2, we can focus without any loss of generality on centered Gaussian measures with diagonal covariance matrices. Thus, defining $T_0 : x \mapsto P_0^T(x - m_0)$ and $T_1 : y \mapsto P_1^T(y - m_1)$ and then applying Lemma 4.2.3 on $GW_2(T_{0\#}\mu, T_{1\#}\nu)$ with $a = 1$, $b = 1$, and $c = -2$ while remarking that the last term in Equation (4.1) is null because $\mathbb{E}_{X\sim T_{0\#}\mu}[X] = 0$ and $\mathbb{E}_{Y\sim T_{1\#}\nu}[Y] = 0$, it follows that Problem (GW_2-Q) is equivalent to

$$\sup_{\pi \in \Pi(T_{0\#}\mu, T_{1\#}\nu)} \int \|x\|_{\mathbb{R}^d}^2 \|y\|_{\mathbb{R}^{d'}}^2 \mathrm{d}\pi(x,y) + 2 \left\| \int xy^T \mathrm{d}\pi(x,y) \right\|_{\mathcal{F}}^2 .$$

Since $T_{0\#}\mu$ and $T_{1\#}\nu$ are centered, it follows that $\int xy^T \mathrm{d}\pi(x,y) = \mathrm{Cov}(X,Y)$ where $X \sim T_{0\#}\mu$ and $Y \sim T_{1\#}\nu$. Furthermore, it can be easily computed that

$$\int \|x\|_{\mathbb{R}^d}^2 \|y\|_{\mathbb{R}^{d'}}^2 \mathrm{d}\pi(x,y) = \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + \sum_{i,j} \mathbb{E}[X_i^2]\mathbb{E}[Y_j^2] \,.$$

Since the second term doesn't depend on π, we get that problem (GW_2-Q) is equivalent to problem (supCOV). □

The left-hand term of (supCOV) is closely related to the sum of symmetric co-kurtosis[1] and so depends on co-moments of order 4 of π. On the other hand, the right-hand term is directly related to the co-moments of order 2 of π. For this reason, Problem (supCOV) is hard to solve because it involves to optimize simultaneously the co-moments of order 2 and 4 of π and so to know the probabilistic rule which links them. This rule is well-known when π is Gaussian thanks to the Isserlis lemma (Isserlis, 1918), but this is not the case in general to the best of our knowledge and there is no *a priori* reason for the solution of Problem (supCOV) to be Gaussian even if the marginals μ and ν are Gaussian.

4.2.2 Study of the general problem

Although Problem (supCOV) is hard to solve because of its dependence on co-moments of order 2 and 4 of π, one can still optimize both terms seperately in order to find a lower bound of $GW_2(\mu,\nu)$. In the rest of the chapter, we suppose for convenience and without any loss of generality that $d \geq d'$.

Proposition 4.2.4. *Suppose without any loss of generality that $d \geq d'$. Let $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ be two Gaussian measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of Σ_0 ($= P_0 D_0 P_0^T$) and Σ_1 ($= P_1 D_1 P_1^T$) that sort the eigenvalues of Σ_0 and Σ_1 in non-increasing order. We suppose that Σ_0 is non-singular. A lower bound for $GW_2(\mu,\nu)$ is then*

$$GW_2^2(\mu,\nu) \geq LGW_2^2(\mu,\nu) \ ,$$

where

$$\begin{aligned} LGW_2^2(\mu,\nu) = 4\left(\mathrm{tr}(D_0) - \mathrm{tr}(D_1)\right)^2 + 4\left(\|D_0\|_{\mathcal{F}} - \|D_1\|_{\mathcal{F}}\right)^2 + 4\|D_0^{(d')} - D_1\|_{\mathcal{F}}^2 \\ + 4\left(\|D_0\|_{\mathcal{F}}^2 - \|D_0^{(d')}\|_{\mathcal{F}}^2\right) \ , \end{aligned} \tag{4.2}$$

where for any matrix A of size $d \times d$, $A^{(d')}$ denotes the submatrix of size $d' \times d'$ containing the d' first rows and the d' first columns of A.

The proof of this proposition is divided in smaller intermediary results. First we recall the Isserlis lemma (Isserlis, 1918), which allows to derive the co-moments of order 4 of a Gaussian distribution as a function of its co-moments of order 2.

Lemma 4.2.5 (Isserlis, 1918). *Let X be a zero-mean Gaussian vector of size d. Then for all 4-tuple of indices i, j, k, l in $[\![1, d]\!]$,*

$$\mathbb{E}\left[X_i X_j X_k X_l\right] = \mathbb{E}\left[X_i X_j\right]\mathbb{E}\left[X_k X_l\right] + \mathbb{E}\left[X_i X_k\right]\mathbb{E}\left[X_j X_l\right] + \mathbb{E}\left[X_i X_l\right]\mathbb{E}\left[X_j X_k\right] \ .$$

Then we derive the two following general optimization lemmas, whose proofs are postponed to Appendix A.2. In all the following, Id_d denotes any diagonal matrix of the form $\mathrm{diag}((\pm 1)_{1 \leq i \leq d})$. We also recall, that for any matrix A of size $r \times s$ with $r \leq d$ and $s \leq d'$, we denote $A^{[d,d']}$ the matrix of size $d \times d'$ of the form

$$A^{[d,d']} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \ .$$

When $d = d'$, we will denote $A^{[d]}$ instead of $A^{[d,d']}$.

Lemma 4.2.6. *Suppose that $d \geq d'$. Let Σ be a symmetric positive semi-definite matrix of size $d + d'$ of the form*

$$\Sigma = \begin{pmatrix} \Sigma_0 & K \\ K^T & \Sigma_1 \end{pmatrix} \ ,$$

with $\Sigma_0 \in \mathbb{S}_{++}^d$, $\Sigma_1 \in \mathbb{S}_+^{d'}$ and K being a rectangular matrix of size $d \times d'$. Let P_0, D_0 and P_1, D_1 be the respective diagonalisations of Σ_0 ($= P_0 D_0 P_0^T$) and Σ_1 ($= P_1 D_1 P_1^T$) that sort the eigenvalues in non-increasing order. Then

$$\max_{K \ : \ \Sigma_1 - K^T \Sigma_0^{-1} K \in \mathbb{S}_+^{d'}} \|K\|_{\mathcal{F}}^2 = \mathrm{tr}(D_0^{(d')} D_1), \tag{4.3}$$

and is achieved at any K^* of the form

$$K^* = P_0 \left(\widetilde{\mathrm{Id}}_{d'} D_0^{(d')^{\frac{1}{2}}} D_1^{\frac{1}{2}}\right)^{[d,d']} P_1^T \ . \tag{opK1}$$

[1] The symmetric co-kurtosis of two random variables X and Y is defined as $\frac{\mathbb{E}[(X-\mathbb{E}[X])^2(Y-\mathbb{E}[Y]^2)]}{\sigma_X^2 \sigma_Y^2}$, where σ_X and σ_Y denote the standard deviations of X and Y.

Lemma 4.2.7. *Suppose that $d \geq d'$. Let Σ be a symmetric positive semi-definite matrix of size $d+d'$ of the form:*

$$\Sigma = \begin{pmatrix} \Sigma_0 & K \\ K^T & \Sigma_1 \end{pmatrix},$$

with $\Sigma_0 \in \mathbb{S}^d_{++}$, $\Sigma_1 \in \mathbb{S}^{d'}_+$ and K being a rectangular matrix of size $d \times d'$. Let $A \in \mathbb{R}^{d' \times d}$ be a matrix with rank 1. Then,

$$\max_{K \,:\, \Sigma_1 - K^T\Sigma_0^{-1}K \in \mathbb{S}^{d'}_+} \operatorname{tr}(KA) = \sqrt{\operatorname{tr}(A\Sigma_0 A^T \Sigma_1)}.$$

In particular, if $\Sigma_0 = \operatorname{diag}(\alpha)$ and $\Sigma_1 = \operatorname{diag}(\beta)$ with $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^{d'}$, then, denoting $\mathbb{1}_{d',d}$ the rectangular matrix of size $d' \times d$ whose coefficients are all equal to 1,

$$\max_{K \,:\, \Sigma_1 - K^T\Sigma_0^{-1}K \in \mathbb{S}^{d'}_+} \operatorname{tr}(K\mathbb{1}_{d',d}) = \sqrt{\operatorname{tr}(\Sigma_0)\operatorname{tr}(\Sigma_1)}\,,$$

and this is achieved at

$$K^* = \frac{\alpha\beta^T}{\sqrt{\operatorname{tr}(\Sigma_0)\operatorname{tr}(\Sigma_1)}}. \tag{opK2}$$

Proof of Proposition 4.2.4. For $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$, we denote P_0, D_0 and P_1, D_1 the respective diagonalizations of Σ_0 and Σ_1 that sort the eigenvalues in non-increasing order. Let $T_0 : x \mapsto P_0^T(x - m_0)$ and $T_1 : y \mapsto P_1^T(y - m_1)$. For $\pi \in \Pi(T_{0\#}\mu, T_{1\#}\nu)$ and $(X,Y) \sim \pi$, we denote Σ_π the covariance matrix of (X,Y) and $\tilde{\Sigma}_\pi$ the covariance matrix of (X^2, Y^2) where X^2 and Y^2 are defined respectively as $(X_i^2)_{1\leq i\leq d}$ and $(Y_j^2)_{1\leq j\leq d'}$. Using Isserlis lemma 4.2.5 to compute $\operatorname{Cov}(X^2, X^2)$ and $\operatorname{Cov}(Y^2, Y^2)$, it follows that Σ_π and $\tilde{\Sigma}_\pi$ are of the form:

$$\Sigma_\pi = \begin{pmatrix} D_0 & K_\pi \\ K_\pi^T & D_1 \end{pmatrix} \quad \text{and} \quad \tilde{\Sigma} = \begin{pmatrix} 2D_0^2 & \tilde{K}_\pi \\ \tilde{K}_\pi^T & 2D_1^2 \end{pmatrix}.$$

In order to find a supremum for each term of (supCOV), we use two necessary conditions for π to be in $\Pi(T_{0\#}\mu, T_{1\#}\nu)$ that are that Σ_π and $\tilde{\Sigma}_\pi$ must be positive semi-definite matrices. To do so, we can use the equivalent conditions that the Schur complements of Σ_π and $\tilde{\Sigma}_\pi$, i.e. $D_1 - K_\pi^T D_0^{-1} K_\pi$ and $2D_1^2 - \frac{1}{2}\tilde{K}_\pi^T D_0^{-2}\tilde{K}_\pi$, must also be positive semi-definite matrices. Remarking that the left-hand term in (supCOV) can be rewritten $\operatorname{tr}(\tilde{K}_\pi \mathbb{1}_{d',d})$, we get the two following inequalities

$$\sup_{X\sim T_{0\#}\mu, Y\sim T_{1\#}\nu} \sum_{i,j} \operatorname{Cov}(X_i^2, Y_j^2) \leq \max_{2D_1^2 - \frac{1}{2}\tilde{K}_\pi^T D_0^{-2}\tilde{K}_\pi \in \mathbb{S}^{d'}_+} \operatorname{tr}(\tilde{K}_\pi \mathbb{1}_{d',d})\,, \tag{4.4}$$

and

$$\sup_{X\sim T_{0\#}\mu, Y\sim T_{1\#}\nu} \|\operatorname{Cov}(X,Y)\|_{\mathcal{F}}^2 \leq \max_{D_1 - K_\pi^T D_0^{-1} K_\pi \in \mathbb{S}^{d'}_+} \|K_\pi\|_{\mathcal{F}}^2\,. \tag{4.5}$$

Applying Lemmas 4.2.6 and 4.2.7 on both right-hand terms, we get on the one hand:

$$\sup_{X\sim T_{0\#}\mu, Y\sim T_{1\#}\nu} \|\operatorname{Cov}(X,Y)\|_{\mathcal{F}}^2 \leq \operatorname{tr}(D_0^{(d')} D_1),$$

and on the other hand:

$$\sup_{X\sim T_{0\#}\mu, Y\sim T_{1\#}\nu} \sum_{i,j} \operatorname{Cov}(X_i^2, Y_j^2) \leq 2\sqrt{\operatorname{tr}(D_0^2)\operatorname{tr}(D_1^2)} = 2\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}.$$

Furthermore, using Lemma 4.2.3, it follows that

$$\begin{aligned} GW_2^2(\mu,\nu) &= C_{\mu,\nu} - 4 \sup_{X\sim T_{0\#}\mu, Y\sim T_{1\#}\nu} \left(\sum_{i,j} \operatorname{Cov}(X_i^2, Y_j^2) + \sum_{i,j} \mathbb{E}\left[X_i^2\right]\mathbb{E}\left[X_j^2\right] + 2\left\|\operatorname{Cov}(X,Y)\right\|_{\mathcal{F}}^2 \right) \\ &\geq C_{\mu,\nu} - 8\sqrt{\operatorname{tr}(D_0^2)\operatorname{tr}(D_1^2)} - 4\operatorname{tr}(D_0)\operatorname{tr}(D_1) - 8\operatorname{tr}(D_0^{(d')} D_1)\,, \end{aligned}$$

where

$$\begin{aligned} C_{\mu,\nu} &= \mathbb{E}_{U\sim \mathrm{N}(0,2D_0)}[\|U\|_{\mathbb{R}^d}^4] + \mathbb{E}_{V\sim \mathrm{N}(0,2D_1)}[\|V\|_{\mathbb{R}^{d'}}^4] - 4\mathbb{E}_{X\sim\mu}[\|X\|_{\mathbb{R}^d}^2]\mathbb{E}_{Y\sim\nu}[\|Y\|_{\mathbb{R}^{d'}}^2] \\ &= 8\operatorname{tr}(D_0^2) + 4(\operatorname{tr}(D_0))^2 + 8\operatorname{tr}(D_1^2) + 4(\operatorname{tr}(D_1))^2 - 4\operatorname{tr}(D_0)\operatorname{tr}(D_1)\,. \end{aligned}$$

Finally

$$\begin{aligned} GW_2^2(\mu,\nu) &\geq 4(\mathrm{tr}(D_0))^2 + 4(\mathrm{tr}(D_1))^2 - 8\mathrm{tr}(D_0)\mathrm{tr}(D_1) + 8\mathrm{tr}(D_0^2) + 8\mathrm{tr}(D_1^2) \\ &\quad - 8\sqrt{\mathrm{tr}(D_0^2)\mathrm{tr}(D_1^2)} - 8\mathrm{tr}(D_0^{(d')}D_1) \\ &= LGW_2^2(\mu,\nu) \ , \end{aligned}$$

which concludes the proof. □

Inequality (4.4) becomes an equality if there exists a plan $\pi \in \Pi(T_{0\#}\mu, T_{1\#}\nu)$ such that for $(X,Y) \sim \pi$, $\sum \mathrm{Cov}(X_i^2, Y_j^2) = 2\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}$. Thanks to Lemma 4.2.7, we know that $\mathrm{tr}(\tilde{K}^* \mathbb{1}_{d',d}) = 2\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}$ for

$$\tilde{K}^* = \frac{2\alpha^2(\beta^2)^T}{\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}} \ ,$$

where $\alpha^2 = (\alpha_1^2, \alpha_2^2, \dots, \alpha_d^2)$ and $\beta^2 = (\beta_1^2, \beta_2^2, \dots, \beta_{d'}^2)$ are the diagonal vectors of D_0^2 and D_1^2. However, it doesn't seems straightforward to exhibit a plan $\pi \in \Pi(T_{0\#}\mu, T_{1\#}\nu)$ such that for $(X,Y) \sim \pi$, $\mathrm{Cov}(X^2, Y^2) = \tilde{K}^*$. An important point to mention is that it can be shown, thanks to Isserlis lemma 4.2.5, that there does not exist any Gaussian plan in $\Pi(T_{0\#}\mu, T_{1\#}\nu)$ with such symmetric co-moments of order 4.

On the other hand, it can be easily seen that inequality (4.5) is in fact an equality since the maximal value of $\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2$ is reached when the law of (X,Y) is Gaussian and $\mathrm{Cov}(X,Y)$ is of the form (opK1). Moreover, the following lemma shows that if $\mathrm{Cov}(X,Y)$ is of the form (opK1), then the law of (X,Y) is necessarily Gaussian and (X,Y) is in general sub-optimal for the left-hand term in (supCOV).

Lemma 4.2.8. *Suppose $d \geq d'$. Let $X \sim \mathrm{N}(0, D_0)$ and $Y \sim \mathrm{N}(0, D_1)$ be two Gaussian vectors of respective size d and d' and with diagonal covariance matrices. If $\mathrm{Cov}(X,Y)$ is of the form*

$$\mathrm{Cov}(X,Y) = \left(\widetilde{\mathrm{Id}}_{d'} D_0^{(d')\frac{1}{2}} D_1^{\frac{1}{2}}\right)^{[d,d']} \ ,$$

then

$$Y = \left(\widetilde{\mathrm{Id}}_{d'} D_1^{\frac{1}{2}} D_0^{(d')-\frac{1}{2}}\right)^{[d',d]} X \ ,$$

which implies that (X,Y) is a Gaussian vector and

$$\sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) = 2\mathrm{tr}(D_0^{(d')} D_1) \ ,$$

where $X = (X_1, X_2, \dots, X_d)^T$ and $Y = (Y_1, Y_2, \dots, Y_{d'})^T$.

Thus, apart from particular cases discussed in Section 4.2.4, there doesn't exist any plan π in $\Pi(T_{0\#}\mu, T_{1\#}\nu)$ with co-moments of order 2 of the form (opK1) and with symmetric co-moments of order 4 of the form (opK2) since the former requires π to be Gaussian and the latter requires generally π not to be Gaussian. However, it is not clear from the proof of Lemma 4.2.6 that the solutions of the form (opK1) are necessarily the only solutions of Problem (4.3). Hence there might exist a plan π which is optimal for both terms of (supCOV) but with co-moments of order 2 of a different form than (opK1). Thus, we cannot conclude whether $GW_2(\mu,\nu) = LGW_2(\mu,\nu)$ or $GW_2(\mu,\nu) > LGW_2(\mu,\nu)$.

4.2.3 Problem restricted to Gaussian couplings

In this section, we study the following problem, where we restrict the set of feasible transport plans to Gaussian couplings.

$$GGW_2^2(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu) \cap \mathcal{N}(\mathbb{R}^{d+d'})} \int\int \left(\|x-x'\|_{\mathbb{R}^d}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2\right)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \ , \qquad (GW_2\text{-QG})$$

where $\mathcal{N}(\mathbb{R}^{d+d'})$ is the set of Gaussian measures on $\mathbb{R}^{d+d'}$.

Since GGW_2 is the Gromov-Wasserstein problem restricted to Gaussian transport plans, it is clear that (GW_2-QG) is an upper bound of (GW_2-Q). Combining this result with Proposition 4.2.4, we get the following immediate but important result.

Theorem 4.2.9. *Let $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$ be two Gaussian measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ with Σ_0 non-singular. Then,*

$$LGW_2(\mu,\nu) \leq GW_2(\mu,\nu) \leq GGW_2(\mu,\nu) \ .$$

We then exhibit in the following a solution of Problem (GW_2-QG), which yields an explicit form for the upper bound $GGW_2(\mu,\nu)$.

Theorem 4.2.10. *Suppose without any loss of generality that $d \geq d'$. Let $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$ be two Gaussian measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of Σ_0 $(= P_0 D_0 P_0^T)$ and Σ_1 $(= P_1 D_1 P_1^T)$ which sort the eigenvalues in non-increasing order. We suppose that μ is not degenerate, i.e. Σ_0 is non-singular. Then Problem (GW_2-QG) admits solutions of the form $\pi^* = (\mathrm{Id}_d, T)_\# \mu$ with T affine of the form*

$$\forall x \in \mathbb{R}^d, \ T(x) = m_1 + P_1 A P_0^T (x - m_0) \ , \tag{4.6}$$

where A is any rectangular matrix of size $d' \times d$ of the form

$$A = \left(\widetilde{\mathrm{Id}}_{d'} D_1^{\frac{1}{2}} D_0^{(d')^{-\frac{1}{2}}} \right)^{[d',d]} \ .$$

Furthermore,

$$GGW_2^2(\mu,\nu) = 4(\mathrm{tr}(D_0) - \mathrm{tr}(D_1))^2 + 8\|D_0^{(d')} - D_1\|_{\mathcal{F}}^2 + 8\left(\|D_0\|_{\mathcal{F}}^2 - \|D_0^{(d')}\|_{\mathcal{F}}^2\right) \ . \tag{4.7}$$

Proof. Since the problem is restricted to Gaussian plans, the left-hand term in Equation (supCOV) can be rewritten $2\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2$ thanks to Isserlis lemma 4.2.5, and so Problem (supCOV) becomes in that case

$$\sup_{X \sim T_{0\#}\mu, Y \sim T_{1\#}\nu} 4\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2 \ .$$

Applying Lemma 4.2.6, we can exhibit optimal Gaussian couplings $\tilde{\pi}^* \in \Pi(T_{0\#}\mu, T_{1\#}\nu)$ with associated covariance matrix $\Sigma_{\tilde{\pi}^*}$ of the form

$$\Sigma_{\tilde{\pi}^*} = \begin{pmatrix} D_0 & K_{\tilde{\pi}^*} \\ K_{\tilde{\pi}^*}^T & D_1 \end{pmatrix} \ ,$$

with

$$K_{\tilde{\pi}^*} = \left(\widetilde{\mathrm{Id}}_{d'} D_0^{(d')^{\frac{1}{2}}} D_1^{\frac{1}{2}} \right)^{[d,d']} \ .$$

Thus, applying Lemma 4.2.8, it follows directly that $\tilde{\pi}^*$ are of the form $(\mathrm{Id}_d, \tilde{T})_\# T_{0\#}\mu$, with $\tilde{T}$ linear and such that for all $x \in \mathbb{R}^d$,

$$\tilde{T}(x) = Ax \ ,$$

with A being any rectangular matrix of size $d' \times d$ of the form

$$A = \left(\widetilde{\mathrm{Id}}_{d'} D_1^{\frac{1}{2}} D_0^{(d')^{-\frac{1}{2}}} \right)^{[d',d]} \ .$$

Then, we can deduce the form of the optimal Gaussian plans $\pi^* \in \Pi(\mu,\nu)$:

$$\pi^* = (T_0^{-1}, T_1^{-1})_\# \tilde{\pi}^* = (T_0^{-1}, T_1^{-1})_\# (\mathrm{Id}_d, \tilde{T})_\# T_{0\#}\mu = (\mathrm{Id}_d, T_{1\#}^{-1}\tilde{T}_\# T_0)_\# \mu = (\mathrm{Id}_d, T)_\# \mu \ ,$$

where T is affine and such that for all $x \in \mathbb{R}^d$,

$$T(x) = T_1^{-1} \circ \tilde{T} \circ T_0(x) = m_1 + P_1 A P_0^T (x - m_0) \ .$$

Moreover, using successively Lemma 4.2.3 then Lemma 4.2.6, it follows that

$$\begin{aligned} GGW_2^2(\mu,\nu) &= C_{\mu,\nu} - 16 \sup_{X \sim T_{0\#}\mu, Y \sim T_{1\#}\nu} \|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2 \\ &= 8\mathrm{tr}(D_0^2) + 4(\mathrm{tr}(D_0))^2 + 8\mathrm{tr}(D_1^2) + 4(\mathrm{tr}(D_1))^2 - 4\mathrm{tr}(D_0)\mathrm{tr}(D_1) - 16\mathrm{tr}(D_0^{(d')} D_1) \\ &= 4(\mathrm{tr}(D_0) - \mathrm{tr}(D_1))^2 + 8\mathrm{tr}\left((D_0^{(d')} - D_1)^2\right) + 8\left(\mathrm{tr}(D_0^2) - \mathrm{tr}((D_0^{(d')})^2)\right) \ , \end{aligned}$$

which concludes the proof. □

Note that it is not clear from the proof of Lemma 4.2.6 that the solutions that we exhibited here are the only solution of Problem (GW_2-QG). Indeed, there might exist other cross-covariance matrices K_π such that $\|K_\pi\|_{\mathcal{F}}$ is maximal but that are not of the form (opK1).

Link with Gromov-Monge restricted to linear mappings. The previous result generalizes Vayer (2020, Theorem 4.2.6), which studies the solutions of the Gromov-Monge problem restricted to linear mappings between Gaussian distributions

$$\inf_{\nu=T_{\#}\mu \,:\, T \text{ linear}} \int\int \left(\|x-x'\|^2_{\mathbb{R}^d} - \|T(x)-T(x')\|^2_{\mathbb{R}^{d'}}\right)^2 \mathrm{d}\mu(x)\mathrm{d}\mu(x') \,. \tag{4.8}$$

Indeed, solutions of (4.8) necessarily provide Gaussian transport plans $\pi = (\mathrm{Id}_d, T)_{\#}\mu$ if T is linear. Conversely, Theorem 4.2.10 shows that restricting the optimal plan to be Gaussian in the GW problem between two Gaussian distributions yields an optimal plan of the form $\pi = (\mathrm{Id}_d, T)_{\#}\mu$ with a linear T, whatever the dimensions d and d' of the two Euclidean spaces.

Link with Principal Component Analysis. We can easily draw connections between GGW_2 and PCA. Indeed, we can remark that the optimal plans can be derived by performing PCA on both distributions μ and ν in order to obtain distributions μ' and ν' with zero mean vectors and diagonal covariance matrices with eigenvalues in non-increasing order ($\tilde{\mu} = T_{0\#}\mu$ and $\tilde{\nu} = T_{1\#}\nu$), then by keeping only the d' first components in μ' and finally by deriving the optimal transport plan solution of the W_2 problem between the obtained truncated distribution and ν'. In other terms, denoting $P_{d'} : \mathbb{R}^d \to \mathbb{R}^{d'}$ the linear mapping that, for $x \in \mathbb{R}^d$ keeps only its d' first coordinates, T_{W_2} the optimal transport map as defined in (2.12) such that $\pi_{W_2} = (\mathrm{Id}_{d'}, T_{W_2})_{\#} P_{d'\#}\tilde{\mu}$ achieves $W_2(P_{d'\#}\tilde{\mu}, \tilde{\nu})$, it follows that the optimal plans π_{GGW_2} that achieve $GGW_2(\mu', \nu')$ are of the form

$$\pi_{GGW_2} = (\mathrm{Id}_d, \widetilde{\mathrm{Id}}_{d'\#} T_{W_2\#} P_{d'})_{\#}\tilde{\mu} \,,$$

where, with an abuse of notation, we write Id_d (resp. $\widetilde{\mathrm{Id}}_{d'}$) the map on $\mathbb{R}^d$ (resp. $\mathbb{R}^{d'}$) represented by the similarly denoted matrix. An example of π_{GGW_2} can be found in Figure 4.1 when $d = 2$ and $d' = 1$.

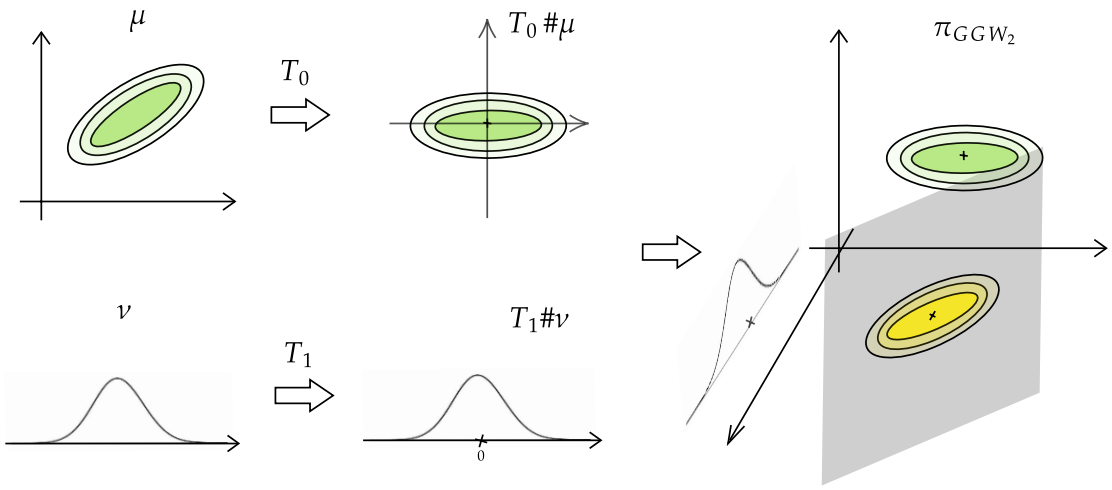

Figure 4.1: An optimal transport plan π_{GGW_2} solution of problem (GW_2-QG) with $d = 2$ and $d' = 1$. In that case, π_{GGW_2} is the degenerate Gaussian distribution supported by the affine plane of equation $y = T_{W_2}(x)$, where T_{W_2} is the classic W_2 optimal transport map, as defined in (2.12), when the distributions are rotated and centered first.

Case of equal dimensions. When $d = d'$, the optimal transport plans π_{GGW_2} that achieve $GGW_2(\mu, \nu)$ are closely related to the optimal transport plan $\pi_{W_2} = (\mathrm{Id}_d, T_{W_2})_{\#} T_{0\#}\mu$. Indeed, a plan π_{GGW_2} can be simply derived by applying the transformations T_0 and T_1 to respectively μ and ν, then by computing π_{W_2} between $T_{0\#}\mu$ and $T_{1\#}\nu$, and finally by applying the inverse transformations T_0^{-1} and T_1^{-1}. In other terms, π_{GGW_2} can be written

$$\pi_{GGW_2} = (\mathrm{Id}_d, T_{1\#}^{-1} \widetilde{\mathrm{Id}}_{d'\#} T_{W_2\#} T_0)_{\#}\mu \,.$$

An example of transport between two Gaussians measures in dimension 2 in Figure 4.2.

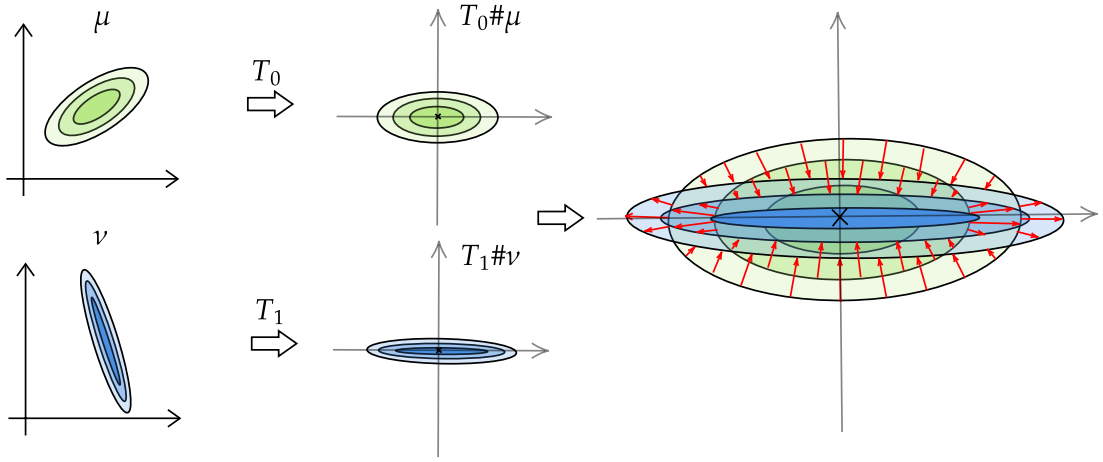

Figure 4.2: Solution of (GW_2-QG) between two Gaussians measures in dimension 2. First the distributions are centered and rotated. Then a classic W_2 transport is applied between the two aligned distributions.

As illustrated in Figure 4.3, the GGW_2 optimal transport map T_{GGW_2} defined in Equation (4.6) is not equivalent to the W_2 optimal transport map T_{W_2} as defined in (2.12) even when the dimensions d and d' are equal. More precisely, if Σ_0 and Σ_1 can be diagonalized in the same orthonormal basis with eigenvalues sorted in the same order, i.e. non-increasing or non-decreasing, then T_{W_2} and T_{GGW_2} are equivalent, see top of Figure 4.3. In contrast, if Σ_0 and Σ_1 can be diagonalized in the same orthonormal basis but their eigenvalues are not sorted in the same order, T_{W_2} and T_{GGW_2} will have very different behaviors, see bottom of Figure 4.3. Between these two extreme cases, we can say that the closer the columns of P_0 will be colinear to the columns of P_1, the more T_{W_2} and T_{GGW_2} will tend to have similar behaviors, see middle of Figure 4.3.

4.2.4 Tightness of the bounds and particular cases

4.2.4.1 Bound on the difference

Proposition 4.2.11. *Suppose without loss of generality that $d \geq d'$. Let $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$, then*

$$GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) \leq 8\|\Sigma_0\|_{\mathcal{F}}\|\Sigma_1\|_{\mathcal{F}}\left(1 - \frac{1}{\sqrt{d}}\right) .$$

To prove this proposition, we will use the following technical result:

Lemma 4.2.12. *Let $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$ be two unit vectors with non-negative coordinates ordered in non-increasing order. Then*

$$u^T v \geq \frac{1}{\sqrt{d}} ,$$

with equality if $u = (\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \dots)^T$ and $v = (1, 0, \dots)^T$.

Proof of Proposition 4.2.11. By subtracting (4.2) from (4.7), it follows that

$$\begin{aligned} GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) &= 8\left(\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}} - \mathrm{tr}(D_0^{(d')}D_1)\right) \\ &= 8\left(\|D_0\|_{\mathcal{F}}\|D_1^{[d]}\|_{\mathcal{F}} - \mathrm{tr}(D_0 D_1^{[d]})\right) . \end{aligned} \tag{4.9}$$

Denoting $\alpha \in \mathbb{R}^d$ and $\beta^{[d]} \in \mathbb{R}^d$ the vectors of eigenvalues of D_0 and $D_1^{[d]}$, it follows that

$$GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) = 8(\|\alpha\|\|\beta^{[d]}\| - \alpha^T\beta^{[d]}) = 8\|\alpha\|\|\beta^{[d]}\|(1 - u^T v) ,$$

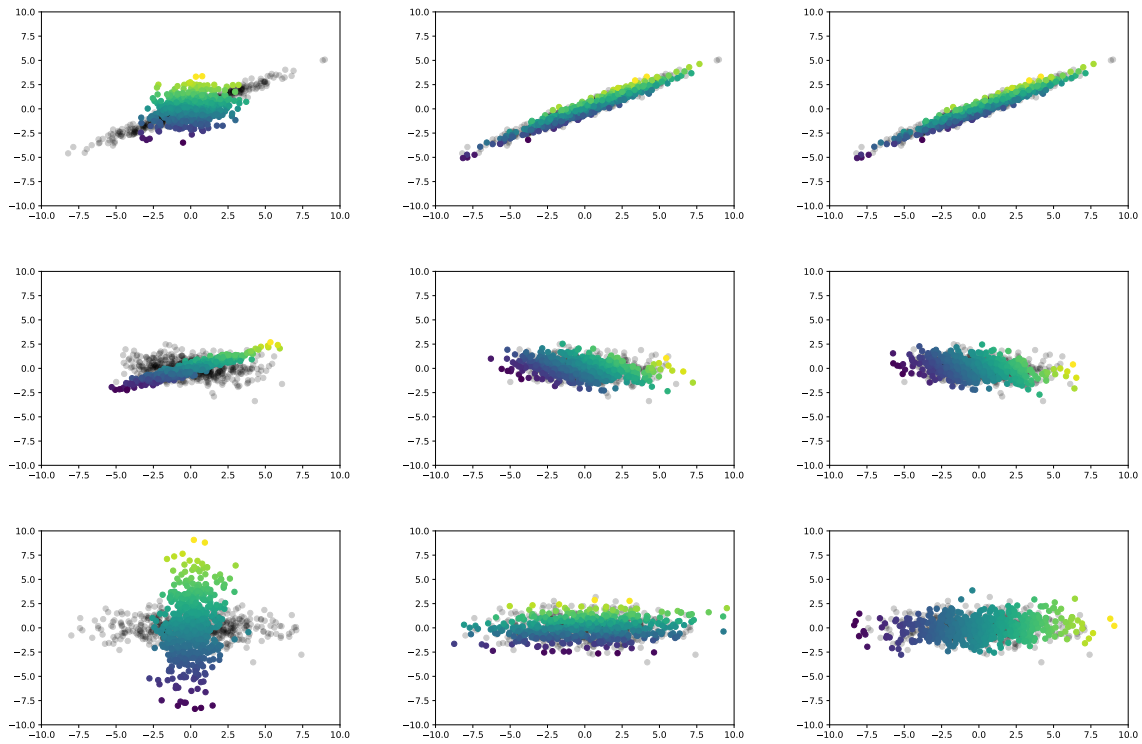

Figure 4.3: Comparison between W_2 and GGW_2 mappings between empirical distributions. Left: 2D source distribution (colored) and target distribution (transparent). Middle: resulting mapping of Wasserstein T_{W_2}. Right: resulting mapping of Gaussian Gromov-Wasserstein T_{GGW_2}. The colors are added in order to visualize where each sample has been sent.

where $u = \frac{\alpha}{\|\alpha\|}$ and $v = \frac{\beta^{[d]}}{\|\beta^{[d]}\|}$. Applying lemma 4.2.12, we get directly that

$$\begin{aligned} GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) &\leq 8\|D_0\|_{\mathcal{F}}\|D_1^{[d]}\|_{\mathcal{F}}\left(1 - \frac{1}{\sqrt{d}}\right). \\ &= 8\|\Sigma_0\|_{\mathcal{F}}\|\Sigma_1\|_{\mathcal{F}}\left(1 - \frac{1}{\sqrt{d}}\right) , \end{aligned}$$

which concludes the proof. □

The difference between $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$ can be seen as the difference between the right and left terms of the Cauchy-Schwarz inequality applied to the two vectors of eigenvalues $\alpha \in \mathbb{R}^d$ and $\beta^{[d]} \in \mathbb{R}^d$. The difference is maximized when the vectors α and $\beta^{[d]}$ are the least colinear possible. This happens when the eigenvalues of D_0 are all equal and $d' = 1$ or ν is degenerate of true dimension 1. On the other hand, this difference is null when α and $\beta^{[d]}$ are colinear. Between those two extremal cases, we can say that the difference between $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$ will be relatively small if the last $d - d'$ eigenvalues D_0 are small compared to the d' first eigenvalues and if the d' first eigenvalues are close to be proportional to the eigenvalues of D_1. An example in the case where $d = 2$ and $d' = 1$ can be found in Figure 4.4.

4.2.4.2 Explicit case

As seen before, the difference between $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$, with $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$, is null when the two vectors of eigenvalues of Σ_0 and Σ_1 - sorted in non-increasing order - are colinear. When we suppose Σ_0 non-singular, this implies that $d = d'$ and that the eigenvalues of Σ_1 are proportional to the eigenvalues of Σ_0 (rescaling).

Proposition 4.2.13. *Suppose $d = d'$. Let $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ two Gaussian measures on $\mathbb{R}^d$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of Σ_0 $(= P_0 D_0 P_0^T)$ and Σ_1 $(= P_1 D_1 P_1^T)$ that*

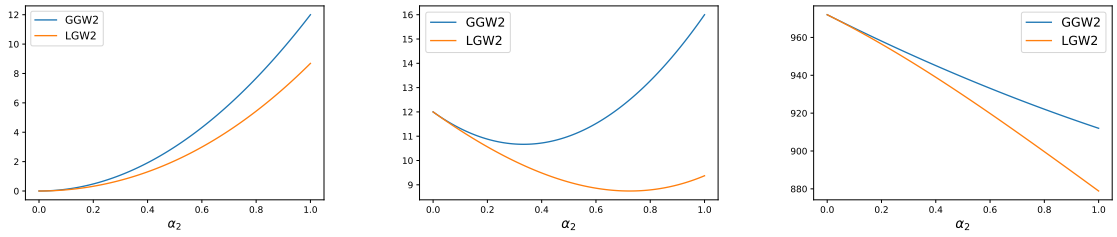

Figure 4.4: plot of $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$ in function of α_2 for $\mu = \mathrm{N}(0, \mathrm{diag}(\alpha))$, $\nu = \mathcal{N}(0, \beta_1^{[d]})$, $\alpha = (\alpha_1, \alpha_2)^T$, for $(\alpha_1, \beta_1^{[d]}) = (1,1)$ (left), $(\alpha_1, \beta_1^{[d]}) = (1,2)$ (middle), $(\alpha_1, \beta_1^{[d]}) = (1,10)$ (right). One can easily compute using (4.7) and (4.2) that $GGW_2^2(\mu,\nu) = 12\alpha_2^2 + 8\alpha_2(\alpha_1 - \beta_1^{[d]}) + 12(\alpha_1 - \beta_1^{[d]})^2$ and $LGW_2^2(\mu,\nu) = 12\alpha_2^2 + 8\alpha_2(\alpha_1 - \beta_1^{[d]}) - 4\sqrt{\alpha_2^2 + \alpha_1^2\beta_1^{[d]}} + 12(\alpha_1 - \beta_1^{[d]})^2 + 8\alpha_1\beta_1^{[d]}$.

sort eigenvalues in non-increasing order. Suppose Σ_0 is non-singular and that there exists a scalar $\lambda \geq 0$ such that $D_1 = \lambda D_0$. In that case, $GW_2^2(\mu,\nu) = GGW_2^2(\mu,\nu) = LGW_2^2(\mu,\nu)$ and Problem (GW_2-Q) admits solutions of the form $(\mathrm{Id}_d, T)_{\#}\mu$ with T such that for all $x \in \mathbb{R}^d$,

$$T(x) = m_1 + \sqrt{\lambda} P_1 \widetilde{\mathrm{Id}}_d P_0^T (x - m_0) \ . \tag{4.10}$$

Moreover, in that case

$$GW_2^2(\mu,\nu) = (\lambda - 1)^2 \left(4(\mathrm{tr}(\Sigma_0))^2 + 8\|\Sigma_0\|_{\mathcal{F}}^2\right) \ . \tag{4.11}$$

Proof. From (4.9), we have

$$GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) = 8\left(\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}} - \mathrm{tr}(D_0 D_1)\right) \ .$$

Denoting $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$ the eigenvalues vectors of D_0 and D_1, it follows that

$$GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) = 8(\|\alpha\|\|\beta\| - \alpha^T\beta) \ .$$

Since there exists $\lambda \geq 0$ such that $D_1 = \lambda D_0$, we have $\beta = \lambda\alpha$, and so $\alpha^T\beta = \|\alpha\|\|\beta\|$. Thus $GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) = 0$ and using Proposition 4.2.9, we get that $GW_2^2(\mu,\nu) = GGW_2^2(\mu,\nu) = LGW_2^2(\mu,\nu)$. We then get (4.10) and (4.11) by simply reinjecting $D_1 = \lambda D_0$ in (4.6) and (4.7). □

This case also includes the more particular case where $d = d' = 1$. In that case $\mu = \mathrm{N}(m_0, \sigma_0^2)$ and $\nu = \mathrm{N}(m_1, \sigma_1^2)$, because σ_1 is always proportional to σ_0.

Corollary 4.2.14. *Let $\mu = \mathrm{N}(m_0, \sigma_0^2)$ and $\nu = \mathrm{N}(m_1, \sigma_1^2)$ be two Gaussian measures on $\mathbb{R}$. Then*

$$GW_2^2(\mu,\nu) = 12\left(\sigma_0^2 - \sigma_1^2\right)^2 ,$$

and the optimal transport plans π^* are of the form $(\mathrm{Id}_{\mathbb{R}}, T)_{\#}\mu$ with T affine of the form, for all $x \in \mathbb{R}$,

$$T(x) = m_1 \pm \frac{\sigma_1}{\sigma_0}(x - m_0) \ . \tag{4.12}$$

Observe that the solution of $W_2^2(\mu,\nu)$ is also solution of $GW_2^2(\mu,\nu)$ in that case. More precisely, the two solutions of the form (4.12) correspond to the mappings $F_\nu^{-1} \circ F_\mu^{\uparrow}$ and $F_\nu^{-1} \circ F_\mu^{\downarrow}$, where $F_\mu^{\uparrow}$ and $F_\nu^{\downarrow}$ denote respectively the *cumulative* and *anti-cumulative* distribution functions of μ, i.e. for all $x \in \mathbb{R}$,

$$F_\mu^{\uparrow}(x) = \mu((-\infty, x]) \qquad \text{and} \qquad F_\mu^{\downarrow}(x) = \mu([-x, +\infty)) \ .$$

The couplings $(\mathrm{Id}_{\mathbb{R}}, F_\nu^{-1} \circ F_\mu^{\uparrow})_{\#}\mu$ and $(\mathrm{Id}_{\mathbb{R}}, F_\nu^{-1} \circ F_\mu^{\downarrow})_{\#}\mu$ are also the two solutions of the GW problem with inner-product as cost functions, as shown in Vayer (2020). This result implies therefore that in case of one-dimensional Gaussian distributions, the GW problem with squared Euclidean distances as cost functions has the same "nice" behavior as in the inner-product case. This is not the case in general when the distributions are not Gaussian, as it has been shown by Beinert et al. (2022) and Dumont et al. (2022).

4.2.4.3 Case of degenerate measures

In all the results of the previous sections, we have supposed Σ_0 non-singular, which means that μ admits a density with respect to the Lesbegue measure. Yet, if Σ_0 is not full rank, one can easily extend the previous results thanks to the following proposition.

Proposition 4.2.15. *Let $\mu = \mathrm{N}(0, D_0)$ and $\nu = \mathrm{N}(0, D_1)$ be two centered Gaussian measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ with diagonal covariance matrices D_0 and D_1 with eigenvalues sorted in non-increasing order. We denote $r = \mathrm{rk}(D_0)$ the rank of D_0 and we suppose that $r < d$. Let $P_r = \begin{pmatrix} \mathrm{Id}_r & 0_{r,d-r} \end{pmatrix}$ be the linear mapping from $\mathbb{R}^d$ to $\mathbb{R}^r$ that keeps only the r first coordinates of the vector of $\mathbb{R}^d$. Then $GW_2^2(\mu,\nu) = GW_2^2(P_{r\#}\mu,\nu)$, $GGW_2^2(\mu,\nu) = GGW_2^2(P_{r\#}\mu,\nu)$, and $LGW_2^2(\mu,\nu) = LGW_2^2(P_{r\#}\mu,\nu)$.*

Proof. For $r < d$, we denote $\Gamma_r(\mathbb{R}^d)$ the set of vectors $x = (x_1, \dots, x_d)^T$ of $\mathbb{R}^d$ such that $x_{r+1} = \cdots = x_d = 0$. For $\pi \in \Pi(\mu,\nu)$, one can remark that for any Borel set $\mathsf{A} \subset \mathbb{R}^d \setminus \Gamma_r(\mathbb{R}^d)$, and any Borel set $\mathsf{B} \subset \mathbb{R}^{d'}$, we have $\pi(\mathsf{A},\mathsf{B}) = 0$ and so

$$\begin{aligned} GW_2^2(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} (\|x-x'\|_{\mathbb{R}^d}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \\ &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Gamma_r(\mathbb{R}^d)\times\mathbb{R}^{d'}} \int_{\Gamma_r(\mathbb{R}^d)\times\mathbb{R}^{d'}} (\|x-x'\|_{\mathbb{R}^d}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \\ &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Gamma_r(\mathbb{R}^d)\times\mathbb{R}^{d'}} \int_{\Gamma_r(\mathbb{R}^d)\times\mathbb{R}^{d'}} (\|P_r(x-x')\|_{\mathbb{R}^r}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \ . \end{aligned}$$

Now, observe that for $\pi \in \Pi(\mu,\nu)$, $(P_r, \mathrm{Id}_{d'})_\#\pi \in \Pi(P_{r\#}\mu,\nu)$. It follows that

$$\begin{aligned} GW_2^2(\mu,\nu) &\leq \inf_{\pi\in\Pi(P_{r\#}\mu,\nu)} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} (\|x-x'\|_{\mathbb{R}^r}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \\ &= GW_2^2(P_{r\#}\mu,\nu). \end{aligned}$$

Conversely, since μ has no mass outside of $\Gamma_r(\mathbb{R}^d)$, $P_{r\#}^T P_{r\#}\mu = \mu$, which implies that for $\pi \in \Pi(P_{r\#}\mu,\nu)$, $(P_r^T, \mathrm{Id}_{d'})_\#\pi \in \Pi(\mu,\nu)$. It follows that

$$\begin{aligned} GW_2^2(P_{r\#}\mu,\nu) &= \inf_{\pi\in\Pi(P_{r\#}\mu,\nu)} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} (\|x-x'\|_{\mathbb{R}^r}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \\ &= \inf_{\pi\in\Pi(P_{r\#}\mu,\nu)} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} (\|P_r^T(x-x')\|_{\mathbb{R}^r}^2 - \|y-y'\|_{\mathbb{R}^{d'}}^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \\ &\leq \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} (\|x-x'\|_{\mathbb{R}^d}^2 - \|y-y'\|_{\mathbb{R}^d i}^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') \\ &\leq GW_2^2(\mu,\nu) \ . \end{aligned}$$

The exact same reasoning can be made in the case of GGW_2. Furthermore, it can be easily seen when looking at (4.2) that $LGW_2^2(\mu,\nu) = LGW_2^2(P_{r\#}\mu,\nu)$. □

Thus, when Σ_0 is not full rank, one can apply Proposition 4.2.15 and consider directly the GW problem between the projected non-degenerate measure $P_{r\#}\mu$ on $\mathbb{R}^r$ and ν and so Theorem 4.2.9 still holds when μ is degenerate.

In the case of GGW_2, an explicit optimal transport plan can still be exhibited. In the following, we denote r_0 and r_1 the ranks of Σ_0 and Σ_1, and we suppose without loss of generality that $r_0 \geq r_1$, but this time not necessarily that $d \geq d'$. Let $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$ be two Gaussian measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, and let (P_0, D_0) and (P_1, D_1) be the respective diagonalizations of Σ_0 $(= P_0 D_0 P_0^T)$ and Σ_1 $(= P_1 D_1 P_1^T)$ that sort the eigenvalues in decreasing order. Optimal transport plans for $GGW_2(\mu,\nu)$ are then of the form $\pi^* = (\mathrm{Id}_d, T)\#\mu$ with T such that for all $x \in \mathbb{R}^d$,

$$T(x) = m_1 + P_1 A P_0^T (x - m_0) \ ,$$

where $A \in \mathbb{R}^{d'\times d}$ is any matrix of the form

$$A = \left(\widetilde{\mathrm{Id}}_{r_1} D_1^{(r_1)\frac{1}{2}} D_0^{(r_1)-\frac{1}{2}} \right)^{[d',d]} \ .$$

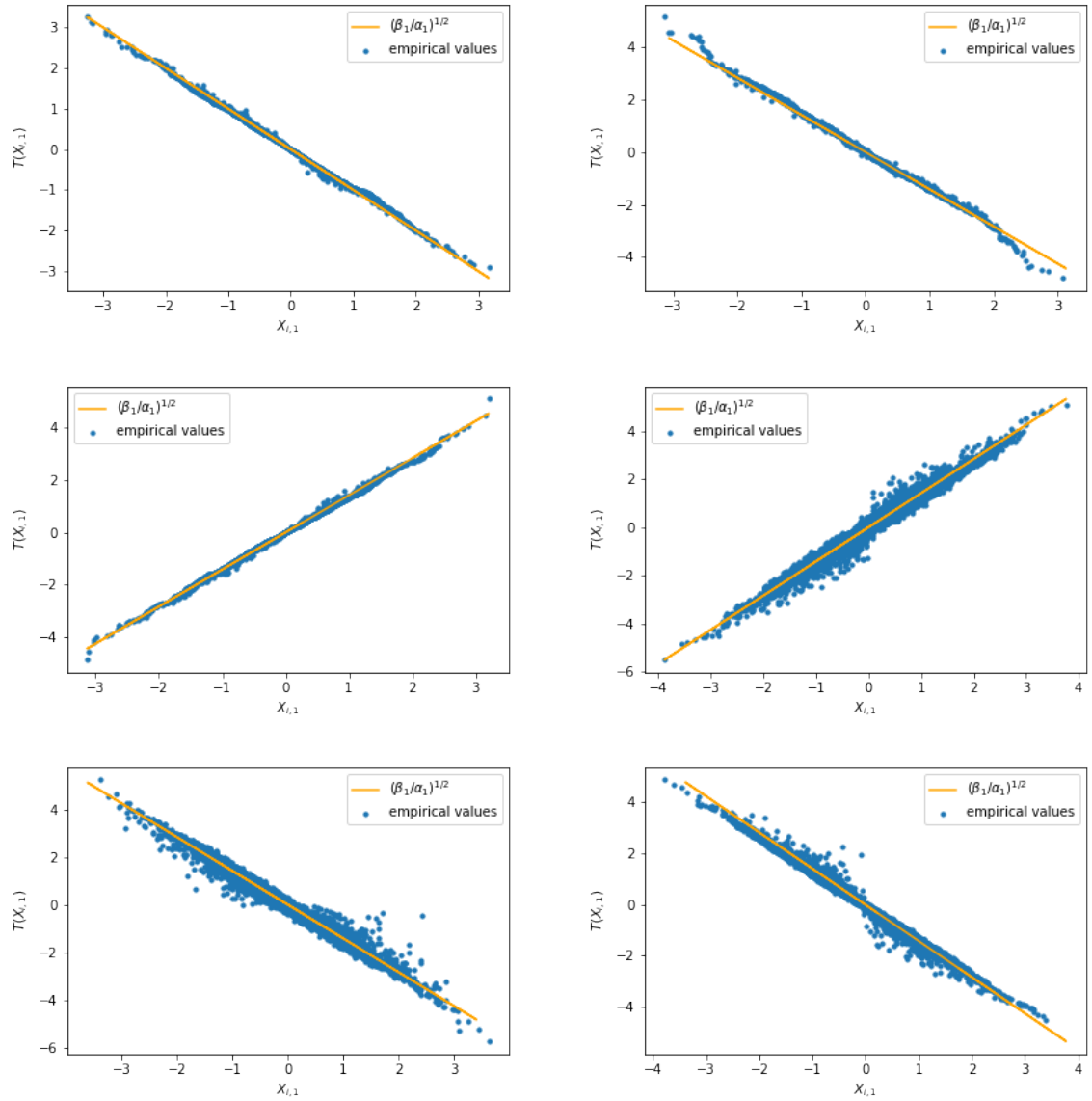

Figure 4.5: Plot of the first coordinate of samples Y_i in function of the first coordinate of their assigned samples X_j (blue dots) and line of equation $y = \pm\sqrt{\beta}x$ (orange line) for $k = 2000$, $\alpha = (1, 0.1)^T$ and $\beta = 2$ (top left), $k = 2000$, $\alpha = (1, 0.1)^T$ and $\beta = (2, 0.3)^T$ (top right), $k = 2000$, $\alpha = (1, 0.1, 0.01)^T$ and $\beta = 2$ (middle left), $k = 7000$, $\alpha = (1, 0.3)$ and $\beta = 2$ (middle right), $k = 7000$, $\alpha = (1, 0.1)^T$, and $\beta = (2, 1)^T$ (bottom left), and $k = 7000$ and $\alpha = (1, 0.3, 0.1)$ and $\beta = 2$ (bottom right).

4.2.5 Behavior of the empirical solutions

In this section, we perform a simple experiment to illustrate the behavior of the empirical solutions of the Gromov Wasserstein problem. In this experiment, we draw independently n samples $(X_j)_{1\leq j\leq n}$ and $(Y_i)_{1\leq i\leq n}$ from respectively $\mu = \mathrm{N}(0, \mathrm{diag}(\alpha))$ and $\nu = \mathrm{N}(0, \mathrm{diag}(\beta))$ with $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^{d'}$. Then we compute the Gromov-Wasserstein distance between the two histograms X and Y using the non-regularized GW solver described in Section 3.1.4. We use for that the implementation provided by the Python Optimal Transport (POT) library[2] (Flamary et al., 2021). In Figure 4.5, we plot the first coordinates of the samples Y_i in function of the first coordinate of the samples X_j they have been assigned to by the solver (blue dots). We draw also the line of equation $y = \pm\sqrt{\beta}x$ in order to compare with the theoretical solution of the Gaussian restricted problem (orange line) for $n = 2000$, $\alpha = (1, 0.1)^T$ and $\beta = 2$ (top left), $n = 2000$, $\alpha = (1, 0.1)^T$ and $\beta = (2, 0.3)^T$ (top right), $n = 2000$, $\alpha = (1, 0.1, 0.01)^T$ and $\beta = 2$ (middle

[2]The package is accessible here: https://pythonot.github.io/.

left), $n = 7000$, $\alpha = (1, 0.3)$ and $\beta = 2$ (middle right), $n = 7000$, $\alpha = (1, 0.1)^T$, and $\beta = (2, 1)^T$ (bottom left), and $n = 7000$ and $\alpha = (1, 0.3, 0.1)$ and $\beta = 2$ (bottom right). Observe that the empirical solution seems to be behaving exactly in the same way as the theoretical solution exhibited in Theorem 4.2.10 as soon as α and β are close to be colinear. However, when α and β are further away from colinearity, determining the behavior of the empirical solution becomes more complex. Solving the GW problem numerically, even approximately, is a particularly hard task, therefore we cannot conclude if the empirical solution does not behave in the same way as the theoretical solution exhibited in Theorem 4.2.10 or if the solver has not converged in these more complex cases. This second assumption seems to be more likely because it seems that increasing the number of points n reduces the gap between the blue dots and the orange line. Thus, we conjecture that in most cases the optimal plan which achieves $GGW_2(\mu, \nu)$ is also solution of the non-restricted problem $GW_2(\mu, \nu)$ and that $GW_2(\mu, \nu) = GGW_2(\mu, \nu)$. However, the situation might be analogous to the one-dimensional case where the Gromov-Wasserstein distance behaves "nicely" with high-probability but we can construct very specific particular cases where this not case (Beinert et al., 2022; Dumont et al., 2022): there might exist very specific configurations of eigenvalues α and β where the solution of the non-restricted problem $GW_2(\mu, \nu)$ is not solution of $GGW_2(\mu, \nu)$. Constructing such counter-examples remains, to the best of our knowledge, an open problem.

4.3 The inner-product case and other formulations

In this section, we study the GW problem between Gaussian distributions, but this time with inner-products as cost functions instead of the squared Euclidean distance as well as the three other formulations discussed in the previous chapter.

4.3.1 The inner-product case

Here, we focus on the Gromov-Wasserstein distance of order 2 between two centered Gaussian measures $\mu = \mathrm{N}(0, \Sigma_0)$ and $\nu = \mathrm{N}(0, \Sigma_1)$, respectively on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. Our goal is therefore to solve the following optimization problem

$$GW_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} |\langle x, x' \rangle_d - \langle y, y' \rangle_{d'}|^2 \, \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y') \, . \tag{GW_2-IP}$$

Note that if μ and ν are not centered, one can still use (GW_2-IP) to compare them by simply centering them beforehand. We show the following main result.

Theorem 4.3.1. *Suppose without any loss of generality that $d \geq d'$. Let $\mu = \mathrm{N}(0, \Sigma_0)$ and $\nu = \mathrm{N}(0, \Sigma_1)$ be two centered Gaussian measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of Σ_0 $(= P_0 D_0 P_0^T)$ and Σ_1 $(= P_1 D_1 P_1^T)$ that sort the eigenvalues in non-increasing order. We suppose that μ is not degenerate, i.e. Σ_0 is non-singular. Then Problem (GW_2-IP) admits solutions of the form $\pi^* = (\mathrm{Id}_d, T)_{\#}\mu$ with $T : \mathbb{R}^d \to \mathbb{R}^{d'}$ being any affine mapping such that for all $x \in d$,*

$$T(x) = P_1 \left(\widetilde{\mathrm{Id}}_{d'} D_1^{\frac{1}{2}} D_0^{(d')^{-\frac{1}{2}}} \right)^{[d',d]} P_0^T x \, .$$

Furthermore,

$$GW_2^2(\mu, \nu) = \|D_0\|_{\mathcal{F}}^2 + \|D_1\|_{\mathcal{F}}^2 - 2\mathrm{tr}(D_0^{(d')} D_1) \, . \tag{4.13}$$

Proof. The proof of this result is a direct consequence of Lemma 4.2.3: indeed, applying this latter lemma with $a = 0$, $b = 0$, and $c = 1$ yields the following equivalent problem,

$$\sup_{\pi \in \Pi(\mu,\nu)} \left\| \int x y^T \mathrm{d}\pi(x, y) \right\|_{\mathcal{F}}^2 .$$

Since μ and ν are centered, it follows that Problem (GW_2-IP) is equivalent to

$$\sup_{X \sim \mu, Y \sim \nu} \|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2 .$$

Applying Lemma 4.2.6, this yields to the solutions exhibited in Lemma 4.2.8. Finally, reinjecting the expression of the optimal $\mathrm{Cov}(X, Y)$ in (4.1) and observing that $\int x^T x x'^T x' \mathrm{d}\mu(x)\mathrm{d}\mu(x') = \|D_0\|_{\mathcal{F}}$ and $\int y^T y y'^T y' \mathrm{d}\mu(y)\mathrm{d}\mu(y') = \|D_1\|_{\mathcal{F}}$, we get (4.13), which concludes the proof. □

This theorem implies that the solutions of Problem (GW_2-QG) between two Gaussian measures $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$ are also solutions of Problem (GW_2-IP) between the associated centered measures $\bar{\mu}$ and $\bar{\nu}$. Note that Problem (GW_2-IP) is not restricted to Gaussian couplings only. Thus, the GW problem between Gaussian distributions with inner-product as cost functions seems to have a much simpler structure than the GW problem with squared Euclidean distance as cost functions, as it was already the case for one-dimensional distributions. Finally, note that, as for Problem (GW_2-QG), the solutions exhibited in Theorem 4.3.1 are not necessarily the only ones.

4.3.2 Invariant Wassertein discrepancy

In this section, we study the behavior of Problem (IW_2) between Gaussian distributions, initially introduced in Alvarez-Melis et al. (2019), when this latter problem is equivalent to

$$\sup_{\pi \in \Pi(\mu,\nu)} \left\| \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x,y) \right\|_{\mathcal{F}} , \tag{$\mathcal{F}$-COV}$$

where we recall that for any matrix A of size $d \times d'$, $\|A\|_{\mathcal{F}}$ denotes the Frobenius norm, i.e. $\sqrt{\mathrm{tr}(A^T A)}$. In all what follows, for any $\pi \in \Pi(\mu,\nu)$, we denote K_π the cross-covariance matrix $\int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x,y)$ of size $d \times d'$ associated with the coupling π. Clearly, Problem ($\mathcal{F}$-COV) is equivalent to Problem (GW_2-IP), as it has been already shown by Alvarez-Melis et al. (2019). We thus directly get that the solutions exhibited in Theorem 4.3.1 are also solutions of Problem ($\mathcal{F}$-COV). When $\mu = \mathrm{N}(0, \Sigma_0)$ and $\nu = \mathrm{N}(0, \mathrm{Id}_{d'})$, this corresponds to choosing the set of invariance $\mathcal{H}$ in the definition of (IW_2) as $\mathcal{H}_1 = \{P \in \mathbb{R}^{d \times d'} \ : \ \|P\|_{\mathcal{F}} \leq \sqrt{d'}\}$ and so this gives the solutions of the $IW_2(\mathcal{H}_1, \mu, \nu)$ problem. This gives therefore the following corollary of Theorem 4.3.1.

Corollary 4.3.2. *Let $\mu = \mathrm{N}(0, \Sigma_0)$ and $\nu = \mathrm{N}(0, \mathrm{Id}_{d'})$ with d not necessarily greater than d'. Let P_0, D_0 be the diagonalization of Σ_0 ($= P_0 D_0 P_0^T$) that sorts its eigenvalues in non-increasing order. We suppose furthermore that μ is not degenerate, i.e. Σ_0 is non-singular. Let $\mathcal{H}_1 = \{P \in \mathbb{R}^{d \times d'} \ : \ \|P\|_{\mathcal{F}} \leq \sqrt{d'}\}$. Then the problem*

$$IW_2(\mathcal{H}_1, \mu, \nu) = \inf_{P \in \mathcal{H}_1} W_2(\mu, P_{\#}\nu) , \tag{4.14}$$

admits as solution any couple (π^*, P^*) with $P^* = \frac{\sqrt{d'}}{\|K_{\pi^*}\|_{\mathcal{F}}} K_{\pi^*}$, and π^* of the form:

(i) if $d \geq d'$, $\pi^* = (\mathrm{Id}_d, T)_{\#}\mu$ with $T : \mathbb{R}^d \to \mathbb{R}^{d'}$ being any affine mapping of the form,

$$T = \left(\widetilde{\mathrm{Id}}_{d'} D_0^{(d')^{-\frac{1}{2}}} \right)^{[d',d]} P_0^T .$$

(ii) if $d \leq d'$, $\pi^* = (T, \mathrm{Id}_{d'})_{\#}\nu$ with $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ being any affine mapping of the form

$$T = P_0 \left(\widetilde{\mathrm{Id}}_d D_0^{\frac{1}{2}} \right)^{[d,d']} .$$

Furthermore, in both cases

$$IW_2^2(\mathcal{H}_1, \mu, \nu) = \mathrm{tr}(D_0) + d' - 2\sqrt{d' \mathrm{tr}(D_0^{(d')})} ,$$

with the convention that $D_0^{(d')} = D_0$ when $d \leq d'$.

To prove this theorem, we will use the following intermediary lemma.

Lemma 4.3.3 (Vayer, 2020). *For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^{d'})$, and given any matrix K of size $d \times d'$, denoting $\langle ., . \rangle_{\mathcal{F}}$ the Frobenius inner-product,*

$$\sup_{P \in \mathcal{H}_1} \langle K, P \rangle_{\mathcal{F}} = \sqrt{d'} \|K\|_{\mathcal{F}} , \tag{4.15}$$

and this supremum is achieved at $P^* = \frac{\sqrt{d'}}{\|K\|_{\mathcal{F}}} K$.

Observe that this lemma yields the closed-form of the projection of any matrix K of size $d \times d'$ on the set $\{P \in \mathbb{R}^{d\times d'} \ : \ \|P\|_{\mathcal{F}} = \sqrt{d'}\}$. Indeed, this projection is defined as the matrix $P \in \mathcal{H}_1$ that minimizes

$$\inf_{\|P\|_{\mathcal{F}}=\sqrt{d'}} \|P - K\|_{\mathcal{F}} \ .$$

Hence, this latter problem is equivalent to

$$\inf_{\|P\|_{\mathcal{F}}=\sqrt{d'}} \|P - K\|_{\mathcal{F}}^2 = \inf_{\|P\|_{\mathcal{F}}=\sqrt{d'}} \left(\|P\|_{\mathcal{F}}^2 + \|K\|_{\mathcal{F}}^2 - 2\langle K, P\rangle_{\mathcal{F}}\right) \ ,$$

Since $\|P\|_{\mathcal{F}}^2$ is necessarily equal to d', the problem is equivalent to Problem (4.15). Now, we turn to the proof of Corollary 4.3.2.

Proof of Corollary 4.3.2. Since $\mathbb{E}_{Y\sim\nu}[YY^T] = \mathrm{Id}_{d'}$, Problem (4.14) can be rewritten, using Lemma 3.3.4, as

$$IW_2^2(\mathcal{H}_1, \mu, \nu) = \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x) + \int_{\mathbb{R}^{d'}} \|y\|^2 \mathrm{d}\nu(y) - 2 \sup_{\pi\in\Pi(\mu,\nu)} \sup_{P\in\mathcal{H}_1} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} x^T P y \mathrm{d}\pi(x, y) \ .$$

First observe that

$$\int x^T P y \mathrm{d}\pi(x, y) = \int \mathrm{tr}(P^T x y^T)\mathrm{d}\pi(x, y) = \mathrm{tr}\left(P^T \int x y^T \mathrm{d}\pi(x, y)\right) \ ,$$

where we successively used cyclical permutation invariance and the linearity properties of the trace operator. Then, by interverting the suprema, we get that the problem is equivalent to

$$\sup_{P\in\mathcal{H}_1} \langle P, K_{\pi^*}\rangle_{\mathcal{F}} \ .$$

Applying Lemma 4.3.3 with $K = K_{\pi^*}$ gives the expression of P^*. Furthermore it follows that

$$IW_2^2(\mathcal{H}_1, \mu, \nu) = \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x) + \int_{\mathbb{R}^{d'}} \|y\|^2 \mathrm{d}\nu(y) - 2\sqrt{d'}\, \|K_{\pi^*}\|_{\mathcal{F}} \ ,$$

Since this latter problem is clearly equivalent with (GW_2-IP), we get point (i) by directly applying Theorem 4.3.1. Point (ii) is obtained also by applying Theorem 4.3.1 but this time by exchanging the role of μ and ν. Finally, we can derive that $\int \|x\|^2 \mathrm{d}\mu(x) = \mathrm{tr}(D_0)$ and $\int \|y\|^2 \mathrm{d}\nu(y) = \mathrm{tr}(\mathrm{Id}_{d'}) = d'$, which concludes the proof. □

4.3.3 Embedded Wasserstein distance

Suppose without any loss of generality that $d \geq d'$. We recall that the EW_2 problem is equivalent to, see Proposition 3.3.7,

$$\sup_{\pi\in\Pi(\bar{\mu},\bar{\nu})} \left\| \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} x y^T \mathrm{d}\pi(x, y) \right\|_* \ , \qquad (*\text{-COV})$$

where for any matrix A of size $d \times d'$, $\|A\|_*$ is the nuclear norm of A, i.e. $\|A\|_* = \mathrm{tr}((A^T A)^{\frac{1}{2}})$ and $\bar{\mu}$ and $\bar{\nu}$ are the centered measures associated with μ and ν. As discussed in Section 3.2, Problems ($\mathcal{F}$-COV) and ($*$-COV) are not equivalent in general. The following result shows that when μ and ν are Gaussian measures, the two problems share in fact some common solutions.

Theorem 4.3.4. *Suppose without any loss of generality that $d \geq d'$. Let $\mu = \mathrm{N}(0, \Sigma_0)$ and $\nu = \mathrm{N}(0, \Sigma_1)$ be two centered Gaussian measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of Σ_0 $(= P_0 D_0 P_0^T)$ and Σ_1 $(= P_1 D_1 P_1^T)$ that sort the eigenvalues in non-increasing order. We suppose that μ is not degenerate, i.e. Σ_0 is non-singular. Then the problem*

$$EW_2(\mu, \nu) = \inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} W_2(\mu, P_{\#}\nu) \ ,$$

admits solutions of the form (π^*, P^*) with P^* of the form $P^* = P_0 \widetilde{\mathrm{Id}}_{d'}^{[d,d']} P_1^T$ and $\pi^* = (\mathrm{Id}_d, T)_{\#}\mu$ with T being any affine map such that for all $x \in \mathbb{R}^d$,

$$T(x) = P_1 \left(\widetilde{\mathrm{Id}}_{d'} D_1^{\frac{1}{2}} D_0^{(d')^{-\frac{1}{2}}}\right)^{[d',d]} P_0^T x \ .$$

In other terms, the solutions of Problem (GW_2-IP) exhibited in Theorem 4.3.1 are also solutions of Problem (EW_2). Furthermore,

$$EW_2^2(\mu,\nu) = \mathrm{tr}(D_0) + \mathrm{tr}(D_1) - 2\mathrm{tr}(D_0^{(d')\frac{1}{2}} D_1^{\frac{1}{2}}) \,.$$

The proof of this theorem is mostly based on the following result whose proof is postponed to Appendix A.2.

Lemma 4.3.5. *Suppose that $d \geq d'$. Let Σ be a symmetric positive semi-definite matrix of size $d+d'$ of the form*

$$\Sigma = \begin{pmatrix} \Sigma_0 & K \\ K^T & \Sigma_1 \end{pmatrix} \,,$$

with $\Sigma_0 \in \mathbb{S}_{++}^d$, $\Sigma_1 \in \mathbb{S}_+^{d'}$ and K is a rectangular matrix of size $d \times d'$. Let P_0, D_0 and P_1, D_1 be the respective diagonalisations of Σ_0 ($= P_0 D_0 P_0^T$) and Σ_1 ($= P_1 D_1 P_1^T$) that sort the eigenvalues in non-increasing order. Then,

$$\max_{K \,:\, \Sigma_1 - K^T\Sigma_0^{-1}K \in \mathbb{S}_+^{d'}} \|K\|_* = \max_{K \,:\, \Sigma_1 - K^T\Sigma_0^{-1}K \in \mathbb{S}_+^{d'}} \max_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \langle P, K\rangle_{\mathcal{F}} = \mathrm{tr}(D_0^{(d')\frac{1}{2}} D_1^{\frac{1}{2}}) \,, \tag{4.16}$$

and it is achieved at any couple (K^*, P^*) of the form $P^* = P_0 \widetilde{\mathrm{Id}}_{d'} P_1^T$ and

$$K^* = P_0 \left(\widetilde{\mathrm{Id}}_{d'} D_0^{(d')\frac{1}{2}} D_1^{\frac{1}{2}} \right)^{[d,d']} P_1^T \,. \tag{4.17}$$

Proof of Theorem 4.3.4. Using Proposition 3.3.7, we get that Problem (EW_2) is equivalent to

$$\sup_{\pi \in \Pi(\mu,\nu)} \sup_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \langle P, K_\pi \rangle_{\mathcal{F}} \,,$$

where $K_\pi = \int xy^T \mathrm{d}\pi(x,y)$. As before, we use the necessary condition for π to be in $\Pi(\mu,\nu)$ that is that the covariance matrix Σ_π of the law π is a PSD matrix, or equivalently that the Schur complement of Σ_π, i.e. $\Sigma_1 - K_\pi^T \Sigma_0^{-1} K_\pi$ is also a PSD matrix. This gives the following inequality:

$$\sup_{\pi \in \Pi(\mu,\nu)} \sup_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \langle P, K_\pi \rangle_{\mathcal{F}} \leq \max_{K \,:\, \Sigma_1 - K^T\Sigma_0^{-1}K \in \mathbb{S}_+^{d'}} \max_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \langle P, K \rangle_{\mathcal{F}} \,.$$

Using Lemma 4.3.5, we get that the right-hand term is equal to $\mathrm{tr}(D_0^{(d')\frac{1}{2}} D_1^{\frac{1}{2}})$ and this is achieved for any couple (P^*, K^*) of the form exhibited above. Now observe that the optimal K^* (4.17) are in fact the same as in (opK1). Thus using Lemma 4.2.8, we can deduce the form of the optimal couples (P^*, π^*) that are solutions of Problem (EW_2). Finally by reinjecting the optimal value in the expression of $EW_2(\mu,\nu)$, we get

$$EW_2^2(\mu,\nu) = \mathrm{tr}(D_0) + \mathrm{tr}(D_1) - 2\mathrm{tr}(D_0^{(d')\frac{1}{2}} D_1^{\frac{1}{2}}) \,,$$

which concludes the proof. □

Supposing that $d' \leq d$, Theorem 4.3.4 implies that between Gaussian distributions, solving Problem (GW_2-IP) is to find an isometric embedding of ν in $\mathbb{R}^d$ which minimize the W_2 distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and the embedded degenerate measure $\tilde{\nu} \in \mathcal{P}(\mathbb{R}^d)$. This is coherent with the observations made in Section 4.2.3. This behavior seems however to be specific to Gaussian measures since maximizing the Frobenius norm of the cross-covariance matrix is in general not equivalent to maximizing its nuclear norm, see Figure 3.1 for a simple discrete example in $\mathbb{R}^2$.

4.3.4 Projection Wasserstein discrepancy

Finally, we study the behavior of Problem (PW_2), initially introduced by Cai and Lim (2022), when μ and ν are Gaussian distribution. We recall that this latter problem is equivalent to, see Proposition 3.3.9,

$$\inf_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \left(\mathrm{tr}(P^T \Sigma_x P) - 2\mathrm{tr}(P^T K_\pi) \right) \,,$$

where $\Sigma_x = \int_{\mathbb{R}^d \times \mathbb{R}^d} xx^T d\bar{\mu}(x)$, $K_\pi = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} xy^T d\pi(x,y)$, and where $\bar{\mu}$ and $\bar{\nu}$ are the centered measures associated with μ and ν. In the following, $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$ and we suppose $d \geq d'$ and that Σ_0 is non-singular. Let $(\alpha_1, \dots, \alpha_d)^T \in \mathbb{R}^d$ and $(\beta_1, \dots, \beta_{d'})^T \in \mathbb{R}^{d'}$ be the respective eigenvalues of Σ_0 and Σ_1 ordered in non-increasing order and let us now denote by $(P_{0\downarrow}, D_{0\downarrow})$ and $(P_{1\downarrow}, D_{1\downarrow})$ the respective diagonalizations of Σ_0 $(= P_{0\downarrow} D_{0\downarrow} P_{0\downarrow}^T)$ and Σ_1 $(= P_{1\downarrow} D_{1\downarrow} P_{1\downarrow}^T)$ which sort the eigenvalues in non-increasing order, i.e. $D_{0\downarrow} = \mathrm{diag}(\alpha_1, \dots, \alpha_d)$ and $D_{1\downarrow} = \mathrm{diag}(\beta_1, \dots, \beta_{d'})$. Let $(P_{0\uparrow}, D_{0\uparrow})$ and $(P_{1\uparrow}, D_{1\uparrow})$ denote the respective diagonalizations of Σ_0 and Σ_1 which sort the eigenvalues in non-decreasing order, i.e. $D_{0\uparrow} = \mathrm{diag}(\alpha_d, \dots, \alpha_1)$ and $D_{1\uparrow} = \mathrm{diag}(\beta_{d'}, \dots, \beta_1)$. We show the following result, whose full proof is postponed to Appendix A.2.7.

Proposition 4.3.6. *Suppose $d \geq d'$. Let $\mu = \mathrm{N}(0, \Sigma_0)$ and $\nu = \mathrm{N}(0, \Sigma_1)$ with $\Sigma_0 \in \mathbb{S}^d_{++}$ and $\Sigma_1 \in \mathbb{S}^{d'}_{+}$. Then,*

$$PW_2(\mu, \nu) = \inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} W_2(P^T_\# \mu, \nu) = \inf_{P \,:\, P^T \Sigma_0 P \Sigma_1 = \Sigma_1 P^T \Sigma_0 P} \|(P^T \Sigma_0 P)^{\frac{1}{2}} - \Sigma_1^{\frac{1}{2}}\|_{\mathcal{F}} \,,$$

Furthermore,

(i) if $\alpha_d > \beta_1$, then

$$PW_2(\mu, \nu) = \|D_{0\uparrow}^{(d')^{\frac{1}{2}}} - D_{1\uparrow}^{\frac{1}{2}}\|_{\mathcal{F}} \,.$$

It is achieved at any (π^*, P^*) of the form $P^* = P_{0\uparrow} \widetilde{\mathrm{Id}}_{d'}^{[d,d']} P_{1\uparrow}^T$ and $\pi^* = (\mathrm{Id}_d, T)_\# \mu$ with T being any affine mapping of the form,

$$T = P_{1\uparrow} \left(\widetilde{\mathrm{Id}}_{d'} D_{1\uparrow}^{\frac{1}{2}} D_{0\uparrow}^{(d')^{-\frac{1}{2}}} \right)^{[d',d]} P_{0\uparrow}^T \,.$$

(ii) if $\alpha_1 < \beta_{d'}$, then

$$PW_2(\mu, \nu) = \|D_{0\downarrow}^{(d')^{\frac{1}{2}}} - D_{1\downarrow}^{\frac{1}{2}}\|_{\mathcal{F}} \,.$$

It is achieved at any (π^*, P^*) of the form $P^* = P_{0\downarrow} \widetilde{\mathrm{Id}}_{d'}^{[d,d']} P_{1\downarrow}^T$ and $\pi^* = (\mathrm{Id}_d, T)_\# \mu$ with T being any linear mapping of the form,

$$T = P_{1\downarrow} \left(\widetilde{\mathrm{Id}}_{d'} D_{1\downarrow}^{\frac{1}{2}} D_{0\downarrow}^{(d')^{-\frac{1}{2}}} \right)^{[d',d]} P_{0\downarrow}^T \,.$$

Sketch of the proof. The proof of this result consists mostly in using the equivalent formulation (3.11) of Problem (PW_2), then using the necessary condition for π to be in $\Pi(\mu, \nu)$ that its associated covariance matrix Σ_π is a PSD matrix, or equivalently that its Schur complement is a PSD matrix, and finally in solving the obtained relaxed problem similarly to Lemma 4.3.5. See Appendix A.2.7 for the full proof. □

Note that this result generalizes Cai and Lim (2022, Example VI.1) that derived the expression of the PW_2 discrepancy between a d-dimensional Gaussian and a one-dimensional Gaussian distributions. The PW_2 problem between Gaussian distributions is thus equivalent to minimize the Hellinger distance between $P^T \Sigma_0 P$ and Σ_1 on the subset of $\mathbb{V}_{d'}(\mathbb{R}^d)$ of matrices P such that $P^T \Sigma_0 P$ and Σ_1 commute. Observe that PW_2 has a different behavior in the case where the eigenvalues of Σ_0 are all greater than the eigenvalues of Σ_1 and in the case where the eigenvalues of Σ_0 are all smaller than the eigenvalues of Σ_1. In the case where the eigenvalues of Σ_0 and Σ_1 are entangled, PW_2 vanishes as soon as there exists a projection $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ such that $P^T \Sigma_0 P$ has the same eigenvalues than Σ_1. Geometrically speaking, this corresponds to finding a d'-dimensional plan in $\mathbb{R}^d$ encoded by P on which the *projection* of μ has the same structure than ν has on $\mathbb{R}^{d'}$. This is a really different behavior from the previously studied OT distances that vanish only when there exists a d'-dimensional plan *that contains entirely* the support of μ and such that μ has the same structure on that plan than ν has on $\mathbb{R}^{d'}$. In other words, P^T is not an *isometric* operator but a projection operator, and so it can transform μ while pushing it into a distributions μ' on $\mathbb{R}^{d'}$ that will have a similar structure to ν whereas in $\mathbb{R}^d$, the embedded measure $P_\# \nu$ will be different from μ. The following example illustrates the difference between PW_2 and the previously studied OT distances.

Example 4.3.7. *Let $\mu = \mathrm{N}(0, \Sigma_0)$ and $\nu = \mathrm{N}(0, \sigma_1^2)$ with*

$$\Sigma_0 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad \sigma_1^2 = \frac{3}{2} \,.$$

Then we have $PW_2(\mu, \nu) = 0$ but $GW_2(\langle . \rangle_2, \langle . \rangle_1, \mu, \nu) > 0$.

Observe indeed that when setting $p^* = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^T$, we have

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{3}{2} ,$$

and so $p^{*T}\Sigma_0 p^* = \sigma_1^2$. Since p^* is clearly in $\mathbb{V}_1(\mathbb{R}^2)$ and $p^{*T}\Sigma_0 p^*$ commutes clearly with σ_1^2 (since there both are scalars), we have

$$PW_2^2(\mu,\nu) \leq |(p^{*T}\Sigma_0 p^*)^{\frac{1}{2}} - \sigma_1|^2 = 0 ,$$

and so $PW_2(\mu,\nu) = 0$. On the other hand,

$$GW_2(\langle.\rangle_2, \langle.\rangle_1, \mu, \nu) = \|\Sigma_0\|_{\mathcal{F}}^2 + \|\sigma_1^2\|^2 - 2\mathrm{tr}(\Sigma_0^{(1)}\sigma_1^2) = \frac{17}{4} > 0 .$$

Note that we have also automatically $IW_2(\mathcal{H}_1,\mu,\nu) > 0$ and $EW_2(\mu,\nu) > 0$ since these latter problems are equivalent to the $GW_2(\langle.\rangle_2, \langle.\rangle_1, \mu, \nu)$ problem. Finally, note that apart from the two cases studied above, we could expect the IW_2 discrepancy, as defined in (IW_2), to behave similarly to the PW_2 discrepancy, since in general Problem (IW_2) reads as

$$IW_2^2(\mathcal{H},\mu,\nu) = \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x) + \inf_{\pi\in\Pi(\mu,\nu)} \inf_{h\in\mathcal{H}} \left(\int_{\mathbb{R}^{d'}} \|h(y)\| \mathrm{d}\nu(y) - 2\int_{\mathbb{R}^d\times\mathbb{R}^{d'}} x h^T(y) \mathrm{d}\pi(x,y) \right) ,$$

with the term $\int_{\mathbb{R}^{d'}} \|h(y)\| \mathrm{d}\nu(y)$ depending on h, and so the problem is in general structurally similar to (3.11).

4.4 Discussion

In this chapter, we have studied the behavior of the Gromov-Wasserstein distance of order 2 between two Gaussian distributions $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$ for costs given by the squared Euclidean distances and the inner-products, as well as the behaviors of the three other formulations of optimal transport on different Euclidean spaces introduced in Chapter 3. We have derived closed-form solutions for the GW_2 problem with inner-products as cost functions and we have shown that these solutions are also solutions of the GW_2 problem with squared Euclidean distances restricted to Gaussian couplings. We also have shown that these solutions are also solutions of the embedded Wasserstein distance and also of the invariant Wasserstein discrepancy in the case this latter is equivalent to ($\mathcal{F}$-COV). We have also empirically shown that these solutions seamed to also be most of the time solutions of the unrestricted GW problem with squared Euclidean distances. Yet, we are not able to prove it formally since it would imply understanding the probabilistic rule that links the symmetric co-moments of order 2 and 4 of an arbitrary (non-Gaussian) coupling in $\Pi(\mu,\nu)$. This is in general, to the best of our knowledge, not known at the exception of few particular cases including the Gaussian distributions thanks to the Isserlis lemma (Isserlis, 1918). The solutions we have exhibited share close links with principal componant analysis since these solutions roughly consist in ordering first the eigenvalues of each covariance matrix, then assigning the largest eigenvalue of Σ_0 to the largest eigenvalue of Σ_1, then doing the same for the second largest eigenvalue and so on until all the eigenvalues of the smallest covariance matrix are assigned. In constrast, the OT distance introduced by Cai and Lim (2022), i.e. the projection Wasserstein distance, presents a very different behavior than the other OT distances studied in this chapter.

A question that arises with regard to our work and in conjunction with the works of Vayer (2020), Beinert et al. (2022) and Dumont et al. (2022) is *what is a good choice of cost functions for the Gromov-Wasserstein problem in Euclidean spaces?* Indeed, the choice of squared Euclidean distances, despite being natural, seems to induce strange behaviors, both on one-dimensional (Beinert et al., 2022) and Gaussian distributions. In contrast, the choice of inner-products as cost functions induce nice properties, both on one-dimensional (Vayer, 2020) and Gaussian distributions, and more generally on distributions that admit densities since Dumont et al. (2022) have shown an analogous result to the Brenier theorem. These properties seem to generalize "naturally" the properties of the W_2 distance in the case the distributions are living in different Euclidean spaces. From a probabilistic point of view, an important difference between the GW_2 problem (GW_2-IP) with inner-products as cost functions and the GW_2 problem (GW_2-Q) with squared Euclidean distances as cost functions is that the former depends only on the co-moments of order 2 associated with the coupling π, whereas the latter depends not only on the co-moments of order

2 but also on the co-moments of order 4. This joint optimization of the co-moments of order 2 and 4 of π might explain why Problem (GW_2-Q) favours sometimes non-natural couplings (Beinert et al., 2022), even if most of the time, Problem (GW_2-Q) seems to behave similarly to Problem (GW_2-Q). The situation might be probably the same for Gaussian distributions. Indeed we have observed that Problem (GW_2-Q) seemed to behave similarly to Problem (GW_2-IP) in our experiments. However we are not able to tell whether the solutions of (GW_2-IP) are always also solutions of (GW_2-Q) or if there exists peculiar situations where this is not the case.

Chapter 5

Gromov-Wasserstein type distances between Gaussian mixture models

Contents

In this chapter, we introduce two Gromov-Wasserstein type OT distances between Gaussian mixture models and we illustrate their pratical use in solving Gromov-Wasserstein related tasks. This chapter is mostly a reproduction of Salmona et al. (2023).

5.1 Introduction

Gaussian Mixture Models (GMMs) have become ubiquitous in modern data science. These models are most of the time used in applied fields to represent probability distributions of real datasets, thanks to their ability to approximate any continuous density when the numbers of components is chosen large enough, including the most complex multimodal ones. Their parameters can also be inferred analytically using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). In imaging science, GMMs have been widely used for various applications, such as image restoration (Zoran and Weiss, 2011; Yu et al., 2011; Feng et al., 2013; Teodoro et al., 2015; Zhang et al., 2017) or texture synthesis (Galerne et al., 2017).

In parallel, optimal transport has also become ubiquitous in modern data science. Indeed, if it had become a predominantly theoretical field in the past, the developpement of efficient numerical solvers has widened the use of OT to various data science problems. In the imaging science field, OT has been used in numerous applications such as image matching (Zhu et al., 2007; Wang et al., 2013; Li et al., 2013), medical imaging (Wang et al., 2010; Gramfort et al., 2015), texture synthesis and style transfer (Leclaire and Rabin, 2021; Gutierrez et al., 2017), or shape registration (Feydy et al., 2017; Su et al., 2015). It has also been used in other machine learning subfields such as domain adaptation (Courty et al., 2016), embedding learning (Courty et al., 2018; Xu et al., 2018), natural language processing (Kusner et al., 2015) and generative modeling (Arjovsky et al., 2017; Genevay et al., 2018; Tolstikhin et al., 2018). These efficient numerical solvers are often based on entropic regularization of the classic OT problem that allows to solve the problem with an alternate minimization scheme using the Sinkhorn-Knopp algorithm (Sinkhorn and Knopp, 1967; Cuturi, 2013). Over the last past years, a large body of works have focused on

speeding up the Sinkhorn-Knopp algorithm, building mostly on diverse low-rank approximations (Solomon et al., 2015; Altschuler et al., 2018, 2019; Forrow et al., 2019; Scetbon and Cuturi, 2020; Scetbon et al., 2021). These approaches have helped to reduce the computational cost of the problem from cubic (for the non-regularized problem) to linear complexity. Another type of commonly used solvers are building on sliced mechanisms (Rabin et al., 2012; Kolouri et al., 2019) that leverage the fact that the OT problem between one-dimentional distributions can be solved using a simple sorting algorithm. These solvers roughly consist in computing infinitely many linear projections of the high-dimensional distributions to one-dimensional distributions and then computing the average of the Wasserstein distances between these one-dimensional representations. Alternatively, Delon and Desolneux (2020) have proposed an OT distance not relying on direct comparison of histograms of points. First, a *Gaussian mixture model* (GMM) is fitted on each distribution, then the two obtained GMMs are compared using a restricted version of the W_2 distance where the admissible transportation couplings π must themselves be GMMs. This OT problem has an equivalent formulation, that had already been proposed by Chen et al. (2018), that allows to solve it by merely calculating W_2 distances between pairwise Gaussian components, which can be done analytically, and then solving a small-scale discrete OT problem. The main benefit of this latter approach is that the complexity of the composite OT problem does not depend on the dimension nor on the number of points but only on the number of components in the GMMs, implying that the computational cost of this approach comes almost entirely from the fitting of the GMMs. Although this method probably doesn't compete with the fastest recent refinements of the Sinkhorn-Knopp algorithm in terms of pure computational cost, it provides a relatively scalable and computationally effective OT distance that is particularly suited when there already exists a kind of clustering structure in the data. This approach has been used for texture synthesis (Leclaire et al., 2022), evaluating generative models (Luzi et al., 2023), or Gaussian Mixture reduction (Zhang and Chen, 2020).

Computationally speaking, the Gromov-Wasserstein problem is known to be much more costly to solve than the classic linear OT problem. Indeed, solving numerically the non-regularized GW problem involves solving a classic OT problem at each iteration. As for linear OT, entropic-regularized solvers have also been proposed (Peyré et al., 2016; Solomon et al., 2016) and involve in that case solving a regularized linear OT problem at each iteration. Recently, Scetbon et al. (2022) have shown that the low-rank approximations used to speed-up the Sinkhorn-Knopp algorithm were particularly suited for the regularized GW problem, resulting in a much more computationally efficient solver. Alternatively, other works have proposed efficient solvers by reducing the size of the GW problem, either through quantization of input measures (Chowdhury et al., 2021), or by recursive clustering approches (Xu et al., 2019a; Blumberg et al., 2020). Specifically to the Euclidean setting, Vayer et al. (2019b) has introduced a solver buiding on a sliced mechanism, and leveraging the observation that the GW problem seems most of the time easy to solve between one-dimensional distributions. In this work, we aim to construct an OT distance between GMMs that is invariant to isometries and that stays relevant between GMMs of different dimensions, in order to design a relatively efficient and scalable solver for Gromov-Wasserstein related problems, especially when there already exists a kind of clustering structure in the data.

Contributions of this chapter. In this chapter, we introduce two Gromov-Wasserstein type OT distances between GMMs. More precisely, we introduce in Section 5.3 a natural Gromov version of the distance introduced by Chen et al. (2018) and Delon and Desolneux (2020), that we call MGW_2 for *Mixture Gromov Wasserstein*. This distance can be used for applications which only require to evaluate how far the distributions are from each other, without having to identify correspondences between points. However, this formulation does not allow to derive directly an optimal transportation plan between the points. To design a way to define such an optimal transportation plan, we define in Section 5.4 another OT distance between GMMs derived from EW_2, that we call MEW_2 for *Mixture Embedded Wasserstein* distance. This latter OT distance is not as computationally efficient as MGW_2 but allows to derive directly an optimal assignment between the points. We also define a transportation plan for MGW_2 by analogy with MEW_2. Finally, in Section 5.5, we show that MGW_2 and MEW_2 can both be used to solve relatively efficiently Gromov-Wasserstein related problems. The proofs of all the lemmas are postponed to Appendix A.3.

5.2 Background: GMMs and Mixture Wasserstein distance

We present here the distance introduced in Delon and Desolneux (2020), as well as various results of this latter paper. We denote $GMM_K(\mathbb{R}^d)$ the set of Gaussian mixtures on $\mathbb{R}^d$ with less than K components,

i.e. the set of measures in $\mathcal{P}(\mathbb{R}^d)$ which can be written

$$\mu = \sum_{k=1}^{K'} a_k \mu_k \ ,$$

where $K' \leq K$, $a = (a_1, \dots, a_{K'})^T$ is in $\Delta_{K'}$, and $\{\mu_k\}_k$ is a family of pairwise distinct Gaussian distributions, each of mean $m_k \in \mathbb{R}^d$ and covariance matrix $\Sigma_k \in \mathbb{S}_+^d$. Again, to avoid degeneracy issues where locations with no mass are accounted for, we will assume that the elements of a are all positive. The set of all finite Gaussian mixture distributions on $\mathbb{R}^d$ is then written

$$GMM_\infty(\mathbb{R}^d) = \bigcup_{K \geq 0} GMM_K(\mathbb{R}^d) \ .$$

Note that the condition that the Gaussian components are pairwise distinct ensures the identifiability of the elements of $GMM_\infty(\mathbb{R}^d)$ (Yakowitz and Spragins, 1968), in the sense that two GMMs $\mu = \sum_k^K a_k \mu_k$ and $\nu = \sum_l^L b_l \nu_l$ are equal if and only if $K = L$, and we can reorder the indices such that for all k, $a_k = b_k$ and $\mu_k = \nu_k$. It can been shown that $GMM_\infty(\mathbb{R}^d)$ is dense in $\mathcal{W}_p(\mathbb{R}^d)$ for the metric W_p, meaning that any measure in $\mathcal{W}_p(\mathbb{R}^d)$ can be approximated with any precision for the distance W_p by a finite Gaussian mixture distribution. Let $\mu \in GMM_K(\mathbb{R}^d)$ and $\nu \in GMM_L(\mathbb{R}^d)$. The Mixture-Wasserstein distance of order 2 is defined as

$$MW_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu) \cap GMM_\infty(\mathbb{R}^{2d})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) \right)^{\frac{1}{2}} \ .$$

As for W_2 with $\mathcal{W}_2(\mathbb{R}^d)$, MW_2 defines a metric on $GMM_\infty(\mathbb{R}^d)$. In general, the transportation plan solution of the W_2 problem is not a Gaussian mixture, thus by restricting the set of admissible couplings, we most of the time have $MW_2(\mu, \nu) > W_2(\mu, \nu)$. It can be shown that the difference between $MW_2(\mu, \nu)$ and $W_2(\mu, \nu)$ is upper-bounded by a term that only depends on the weights and the covariances matrices of the components of the two mixtures. Finally, MW_2 can be written in an equivalent form, which had already been introduced in Chen et al. (2018): if $\mu = \sum_k^K a_k \mu_k$ and $\nu = \sum_l^L b_l \nu_l$, then

$$MW_2^2(\mu, \nu) = \inf_{\omega \in \Pi(a,b)} \sum_{k,l} \omega_{k,l} W_2^2(\mu_k, \nu_l) \ , \tag{MW_2}$$

where $a = (a_1, \dots, a_K)^T$, $b = (b_1, \dots, b_L)^T$. From a computational point of view, this latter formulation reduces the problem to a simple small-scale discrete optimal transport problem since the W_2 distance between Gaussian distributions has a closed form: indeed, recall that if $\mu_k = \mathrm{N}(m_k, \Sigma_k)$ and $\nu_l = \mathrm{N}(m_l, \Sigma_l)$, then

$$W_2^2(\mu_k, \nu_l) = \|m_k - m_l\|^2 + \mathrm{tr}\left(\Sigma_k + \Sigma_l - 2\left(\Sigma_l^{\frac{1}{2}} \Sigma_k \Sigma_l^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \ .$$

5.3 Gromov-Wasserstein distance between GMMs

In this section, we define a Gromov-Wasserstein type distance between Gaussian mixture distributions. This distance is a natural "Gromovization" of Problem (MW_2). Indeed, as it has already been observed in the literature (Chen et al., 2018; Lambert et al., 2022), any Gaussian mixture in dimension d can be identified with a probability distribution on $\mathbb{R}^d \times \mathbb{S}_+^d$, i.e. the product space of means and covariance matrices. Equivalently, a finite Gaussian mixture can be seen as a discrete probability distribution on the space of Gaussian distributions $\mathcal{N}(\mathbb{R}^d)$[1], which has been proven to be a complete metric space when endowed with W_2 (Takatsu, 2010) and is furthermore separable since it is a subspace of $\mathcal{W}_2(\mathbb{R}^d)$ which is itself a separable metric space when endowed with W_2 (Bolley, 2008). Since the theory of optimal transport still applies on measures over non-Euclidean spaces (Villani, 2008), it follows that Problem (MW_2) can formally be thought as a simple OT problem between two discrete measures in $\mathcal{P}(\mathcal{N}(\mathbb{R}^d))$. Thus, one can define directly its Gromov version.

Definition 5.3.1. *Let $\mu = \sum_k a_k \mu_k$ and $\nu = \sum_l b_l \nu_l$ be two Gaussian mixtures respectively on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, we define*

$$MGW_2^2(\mu, \nu) = \inf_{\omega \in \Pi(a,b)} \sum_{i,j,k,l} |W_2^2(\mu_i, \mu_k) - W_2^2(\nu_j, \nu_l)|^2 \omega_{i,j} \omega_{k,l} \ . \tag{MGW_2}$$

[1] $\mathcal{N}(\mathbb{R}^d)$ includes the degenerate Gaussian distributions, as for instance the Dirac distributions.

Unlike MW_2, there is no straightforward equivalent formulation of this latter problem. In particular, it is not clear whether Problem (MGW_2) is equivalent or not to the continuous GW problem between μ and ν - seen as continuous measures on $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ - where the set of admissible couplings is restricted to Gaussian mixture distributions. In the rest of the chapter, we distinguish the distribution $\tilde{\mu} \in \mathcal{P}(\mathcal{N}(\mathbb{R}^d))$ from its associated GMM $\mu \in GMM_\infty(\mathbb{R}^d)$. Thanks to the identifiability property of the set of finite Gaussian mixture, we have that each $\mu \in GMM_\infty(\mathbb{R}^d)$ is associated with a unique discrete distribution $\tilde{\mu} \in \mathcal{P}(\mathcal{N}(\mathbb{R}^d))$ and MGW_2 between μ and ν coincides with GW_2 with squared W_2 as cost functions between the associated measures $\tilde{\mu}$ and $\tilde{\nu}$. More generally, one can define for any metric measure space of the form $(\mathcal{N}(\mathbb{R}^d), W_2^2, \tilde{\mu})$ and $(\mathcal{N}(\mathbb{R}^{d'}), W_2^2, \tilde{\nu})$, the following continuous GW problem,

$$\inf_{\pi \in \Pi(\tilde{\mu}, \tilde{\nu})} \int_{\mathcal{N}(\mathbb{R}^d) \times \mathcal{N}(\mathbb{R}^{d'})} \int_{\mathcal{N}(\mathbb{R}^d) \times \mathcal{N}(\mathbb{R}^{d'})} |W_2^2(\gamma, \gamma') - W_2^2(\zeta, \zeta')|^2 \mathrm{d}\pi(\gamma, \zeta) \mathrm{d}\pi(\gamma', \zeta') \ ,$$

where $\tilde{\mu}$ and $\tilde{\nu}$ can be possibly thought as infinite mixture of Gaussians. However, there is in general no identifiability property for infinite Gaussian mixture and so for a given GMM μ on $\mathbb{R}^d$, they might be more than one associated measure $\tilde{\mu}$ on $\mathcal{N}(\mathbb{R}^d)$. For instance, the standard Normal distribution $\mathrm{N}(0,1)$ can naturally be identified in $\mathcal{P}(\mathcal{N}(\mathbb{R}))$ with the Dirac distribution at $\mathrm{N}(0,1)$, but also with the Normal distribution $\mathrm{N}(0,1/2)$ over the parametrized line $\{\mathrm{N}(\theta, 1/2) \in \mathcal{N}(\mathbb{R}) \ : \ \theta \in \mathbb{R}\}$, or with $\mathrm{N}(0,1)$ over the parametrized line $\{\delta_\theta \in \mathcal{N}(\mathbb{R}) \ : \ \theta \in \mathbb{R}\}$.

5.3.1 Metric properties

Here we study the metric property of MGW_2 that mainly arises from the Gromov-Wasserstein structure of Problem (MGW_2). Indeed, the following result is a direct consequence of the theory developped by Sturm (2012).

Proposition 5.3.2. *In the following, $\mu = \sum_k a_k \mu_k$ and $\nu = \sum_l b_l \nu_l$ are two GMMs respectively in $GMM_K(\mathbb{R}^d)$ and $GMM_L(\mathbb{R}^{d'})$.*

(i) MGW_2 is non-negative and symmetric.

(ii) MGW_2 satisfies the triangle inequality, i.e. for any $\xi \in GMM_S(\mathbb{R}^{d''})$,

$$MGW_2(\mu, \nu) \leq MGW_2(\mu, \xi) + MGW_2(\xi, \nu) \ .$$

(iii) $MGW_2(\mu, \nu) = 0$ if and only if there exists a bijection $\phi : \{\mu_k\}_k \to \{\nu_l\}_l$ such that $\nu = \sum_k a_k \phi(\mu_k)$ and ϕ is an isometry for W_2, i.e. for all k and i smaller than K, $W_2(\phi(\mu_k), \phi(\mu_i)) = W_2(\mu_k, \mu_i)$.

Proof. Takatsu (2010) has shown that the space of Gaussian distributions $\mathcal{N}(\mathbb{R}^d)$ is a complete metric space when endowed with W_2. Moreover, $\mathcal{N}(\mathbb{R}^d)$ is separable since it is a subspace of $\mathcal{W}_2(\mathbb{R}^d)$ which is itself a separable metric space when endowed with W_2 (Bolley, 2008). Thus, $\mathcal{N}(\mathbb{R}^d)$ is Polish and we can directly apply the Gromov-Wasserstein theory developped in Sturm (2012). Let $(\mathcal{N}(\mathbb{R}^d), W_2, \tilde{\mu})$ and $(\mathcal{N}(\mathbb{R}^{d'}), W_2, \tilde{\nu})$ be two metric measure spaces respectively in $\mathbb{M}_4$. Let us define

$$D(\tilde{\mu}, \tilde{\nu}) = \inf_{\pi \in \Pi(\tilde{\mu}, \tilde{\nu})} \int_{\mathcal{N}(\mathbb{R}^d) \times \mathcal{N}(\mathbb{R}^{d'})} \int_{\mathcal{N}(\mathbb{R}^d) \times \mathcal{N}(\mathbb{R}^{d'})} |W_2^2(\gamma, \gamma') - W_2^2(\zeta, \zeta')|^2 \mathrm{d}\pi(\gamma, \zeta) \mathrm{d}\pi(\gamma', \zeta') \ .$$

Applying Sturm (2012, Corollary 9.3), we get that D defines a metric over the space of metric measure spaces of the form $(\mathcal{N}(\mathbb{R}^d), W_2, \tilde{\mu})$ quotiented by the strong isomorphisms, and thus we get directly that D is symmetric, non-negative, satisfies the triangle inequality and $D(\tilde{\mu}, \tilde{\nu}) = 0$ if and only if there exists a bijection $\phi : \mathrm{supp}(\tilde{\mu}) \to \mathrm{supp}(\tilde{\nu})$ such that $\tilde{\nu} = \phi_\# \tilde{\mu}$, where for any γ and γ' in $\mathrm{supp}(\tilde{\mu})$, $W_2(\phi(\gamma), \phi(\gamma')) = W_2(\gamma, \gamma')$. Now observe that if $\mu = \sum_k a_k \mu_k$ and $\nu = \sum_l b_l \nu_l$ are respectively in $GMM_K(\mathbb{R}^d)$ and $GMM_L(\mathbb{R}^{d'})$ and $\tilde{\mu} = \sum_k a_k \delta_{\mu_k}$ and $\tilde{\nu} = \sum_l b_l \delta_{\nu_l}$ are respectively in $\mathcal{P}(\mathcal{N}(\mathbb{R}^d))$ and $\mathcal{P}(\mathcal{N}(\mathbb{R}^{d'}))$, we have

$$\int_{\mathcal{N}(\mathbb{R}^d) \times \mathcal{N}(\mathbb{R}^d)} W_2^4(\gamma, \gamma') \mathrm{d}\tilde{\mu}(\gamma) \mathrm{d}\tilde{\mu}(\gamma') = \sum_{k,i} a_k a_i W_2^4(\mu_k, \mu_i) < +\infty \ ,$$

and

$$\int_{\mathcal{N}(\mathbb{R}^{d'}) \times \mathcal{N}(\mathbb{R}^{d'})} W_2^4(\zeta, \zeta') \mathrm{d}\tilde{\nu}(\zeta) \mathrm{d}\tilde{\nu}(\zeta') = \sum_{l,j} b_l b_j W_2^4(\nu_l, \nu_j) < +\infty \ ,$$

so $(\mathcal{N}(\mathbb{R}^d), W_2, \tilde{\mu})$ and $(\mathcal{N}(\mathbb{R}^{d'}), W_2, \tilde{\nu})$ are both in $\mathbb{M}_4$. Furthermore, we have $MGW_2(\mu, \nu) = D(\tilde{\mu}, \tilde{\nu})$. Hence MGW_2 inherits the metric properties of D, which concludes the proof. □

MGW_2 defines thus a pseudometric on the set of all finite Gaussian mixtures of arbitrary dimensions, i.e. the set,

$$\mathcal{GMM}_\infty = \bigsqcup_{d\geq 1} GMM_\infty(\mathbb{R}^d) \ ,$$

that is invariant to the mappings ϕ that transform a finite Gaussian mixture $\sum_{k=1} a_k\mu_k$ into another finite Gaussian mixture of the form $\sum_{k=1}^K a_k\nu_k$ such that for all k and i smaller than K, $W_2(\nu_k,\nu_i) = W_2(\mu_k,\mu_i)$. A question that arises is: *are all these mappings ϕ associated with mappings T that are isometries for the Euclidean norm and such that $T_\#$ coincides with ϕ?* We can already state the following converse result.

Proposition 5.3.3. *Let $d \geq d'$, and let $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a mapping that is an isometry for the Euclidean norm. Then the mapping $\phi_T : GMM_\infty(\mathbb{R}^{d'}) \to \mathcal{P}(\mathbb{R}^d)$ defined as $\phi_T(\mu) = T_\#\mu$ for all $\mu \in GMM_\infty(\mathbb{R}^{d'})$, is such that for any μ of the form $\Sigma_{k=1}^K a_k\mu_k$, $\phi_T(\mu)$ is in $GMM_\infty(\mathbb{R}^{d'})$ and is of the form $\Sigma_{k=1}^K a_k\nu_k$, with $\{\nu_k\}_{k=1}^K$ being such that, for all k and i smaller than K, $W_2(\nu_k,\nu_i) = W_2(\mu_k,\mu_i)$.*

Proof. First recall that the push-foward measure $T_\#\mu$ with μ on $\mathbb{R}^{d'}$ and $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ is defined as the measure on $\mathbb{R}^d$ such that for every Borel set A of $\mathbb{R}^d$, $T_\#\mu(\mathsf{A}) = \mu(T^{-1}(\mathsf{A}))$. Equivalently, for any measurable map $h : \mathbb{R}^d \to \mathbb{R}$, we have

$$\int_{\mathbb{R}^d} h(x)\mathrm{d}(T_\#\mu)(x) = \int_{\mathbb{R}^{d'}} (h\circ T)(y)\mathrm{d}\mu(y) \ .$$

Now observe that for any finite GMM μ on $\mathbb{R}^{d'}$ of the form $\mu = \sum_k^K a_k\mu_k$, we have

$$\begin{aligned}
\textstyle\int_{\mathbb{R}^{d'}} (h\circ T)(y)\mathrm{d}\mu(y) &= \textstyle\int_{\mathbb{R}^{d'}} (h\circ T)(y)\mathrm{d}\left(\sum_k^K a_k\mu_k(y)\right) \\
&= \textstyle\sum_k^K a_k \int_{\mathbb{R}^{d'}} (h\circ T)(y)\mathrm{d}\mu_k(y) \\
&= \textstyle\sum_k^K a_k \int_{\mathbb{R}^d} h(x)\mathrm{d}(T_\#\mu_k)(x) \\
&= \textstyle\int_{\mathbb{R}^d} h(x)\mathrm{d}\left(\sum_k^K a_k(T_\#\mu_k)(x)\right) \ ,
\end{aligned}$$

and so $T_\#\mu$ is of the form $\sum_k^K a_k(T_\#\mu_k)$ with $T_\#\mu_k$ Gaussian since T is necessarily affine as a consequence of Lemma 3.3.2. Thus, $T_\#\mu$ is in $GMM_\infty(\mathbb{R}^d)$. This proves that ϕ_T takes its values only in $GMM_\infty(\mathbb{R}^d)$ and that $\phi_T(\sum_{k=1}^K a_k\mu_k)$ is of the form $\sum_{k=1} a_k\nu_k$. Now observe that, for every k and i smaller than K,

$$W_2^2(\phi_T(\mu_k),\phi_T(\mu_i)) = \inf_{\pi\in\Pi(T_\#\mu_k,T_\#\mu_i)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2\mathrm{d}\pi(x,y) \ .$$

Using two times succesively Lemma 3.3.4 using the fact that T is an isometry an so for any $y \in \mathbb{R}^{d'}$, $\|T(y)\| = \|y\|$, it follows

$$\inf_{\pi\in\Pi(T_\#\mu_k,T_\#\mu_i)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-x'\|^2\mathrm{d}\pi(x,x') = \inf_{\pi\in\Pi(\mu_k,\mu_i)} \int_{\mathbb{R}^{d'}\times\mathbb{R}^{d'}} \|y-y'\|^2\mathrm{d}\pi(y,y') = W_2(\mu_k,\mu_i) \ ,$$

which concludes the proof. □

Hence, if $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ is an isometry for the Euclidean norm, then MGW_2 is invariant to the mapping $\phi_T : GMM_\infty(\mathbb{R}^{d'}) \to GMM_\infty(\mathbb{R}^d)$ such that for all $\mu \in GMM_\infty(\mathbb{R}^{d'})$, $\phi(\mu) = T_\#\mu$. Yet, in general, there exist mappings $\phi : \mathcal{W}_2(\mathbb{R}^{d'}) \to \mathcal{W}_2(\mathbb{R}^d)$ that are isometries for W_2 and that are not induced by any mapping $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ that is an isometry for the Euclidean norm. This has been proven by Kloeckner (2010). The following counter-example shows that this still holds when considering isometries defined on subspaces of $\mathcal{N}(\mathbb{R}^{d'})$.

Example 5.3.4. *Let $\mathcal{N}_{++}(\mathbb{R})$ be the set of one-dimensional Gaussian distributions with strictly positive mean. Let $\phi : \mathcal{N}_{++}(\mathbb{R}) \to \mathcal{N}_{++}(\mathbb{R})$ be the mapping that swaps the mean and the standard deviation, i.e. such that for any $\gamma = \mathrm{N}(m_\gamma,\sigma_\gamma^2)$ with $m_\gamma > 0$ and $\sigma_\gamma > 0$, $\phi(\gamma) = \mathrm{N}(\sigma_\gamma, m_\gamma^2)$. Then ϕ is an isometry for W_2.*

Observe indeed that for γ and ζ in $\mathcal{N}_{++}(\mathbb{R})$, we have

$$W_2(\phi(\gamma),\phi(\zeta)) = (\sigma_\gamma-\sigma_\zeta)^2 + (m_\gamma-m_\zeta)^2 = W_2(\gamma,\zeta) \ .$$

Thus ϕ is an isometry for W_2, yet ϕ is not induced by any isometry of $\mathbb{R}$. Hence there exist mappings from $GMM_\infty(\mathbb{R}^{d'})$ to $GMM_\infty(\mathbb{R}^d)$ that satisfy the conditions above but which are not induced by isometries for the Euclidean norm from $\mathbb{R}^{d'}$ to $\mathbb{R}^d$. To sum things up, we state a straightforward but important corollary of Proposition 5.3.2 that implies that MGW_2 is invariant to isometries for the Euclidean norm, and whose converse is not true as we have just seen above.

Corollary 5.3.5. *Let $\mu \in GMM_K(\mathbb{R}^d)$ and $\nu \in GMM_L(\mathbb{R}^{d'})$, and let suppose that there exists an isometry $T : \mathbb{R}^{d'} \to \mathbb{R}^d$ for the Euclidean norm such that $\mu = T_\# \nu$. Then $MGW_2(\mu, \nu) = 0$.*

5.3.2 MGW_2 in practice

Using MGW_2 on discrete data distributions. Most applications of optimal transport involve discrete data that can be thought as samples drawn from underlying distributions, which are not GMMs in general. In those applications, we aim to evaluate an OT distance between two distributions of the form $\hat{\mu} = (1/M)\sum_i \delta_{x_i}$ and $\hat{\nu} = (1/N)\sum_j \delta_{y_j}$ where $\{x_i\}_i$ and $\{y_j\}_j$ are families of respectively M and N vectors of $\mathbb{R}^d$ and $\mathbb{R}^{d'}$. Though $\hat{\mu}$ and $\hat{\nu}$ can be thought as mixtures of degenerate Gaussian distributions, evaluating directly $MGW_2(\hat{\mu}, \hat{\nu})$ is not particularly interesting since we have in that case $MGW_2(\hat{\mu}, \hat{\nu}) = GW_2(\|.\|^2, \|.\|^2, \hat{\mu}, \hat{\nu})$. However, we can design a pseudometric $MGW_{K,2}$ between $\hat{\mu}$ and $\hat{\nu}$ by fitting two GMMs μ and ν with K components on $\hat{\mu}$ and $\hat{\nu}$ and then setting $MGW_{K,2}(\hat{\mu}, \hat{\nu}) = MGW_2(\mu, \nu)$. The approximation of $\hat{\mu}$ and $\hat{\nu}$ by μ and ν can be done by maximizing the log-likelihood of the GMMs with the EM algorithm (Dempster et al., 1977). Note that if K is chosen too small, the approximations $\hat{\mu}$ and $\hat{\nu}$ will be of bad quality and we are likely to observe undesirable behaviors, as for instance having $MGW_{K,2}(\hat{\mu}, \hat{\nu}) = 0$ despite $\hat{\mu}$ and $\hat{\nu}$ not being equal up to an isometry. Thus, the choice of K must be a compromise between the quality of the approximation given by the GMM and the computational cost. To illustrate the pratical use of MGW_2 on a simple toy example, we draw 150 samples from the spiral dataset provided in the scikit-learn toolbox[2] (Pedregosa et al., 2011) and we apply rotations with various angles on this dataset. We then fit independently GMMs with 20 components on the initial and the target rotated datasets and we compute MGW_2 between the two obtained GMMs. We also compute GW_2 with inner-product as cost functions, MW_2 using also 20 Gaussian components and W_2. The results can be found in Figure 5.1. As expected, MGW_2 is rotation-invariant as GW_2 which is not the case of MW_2 and W_2.

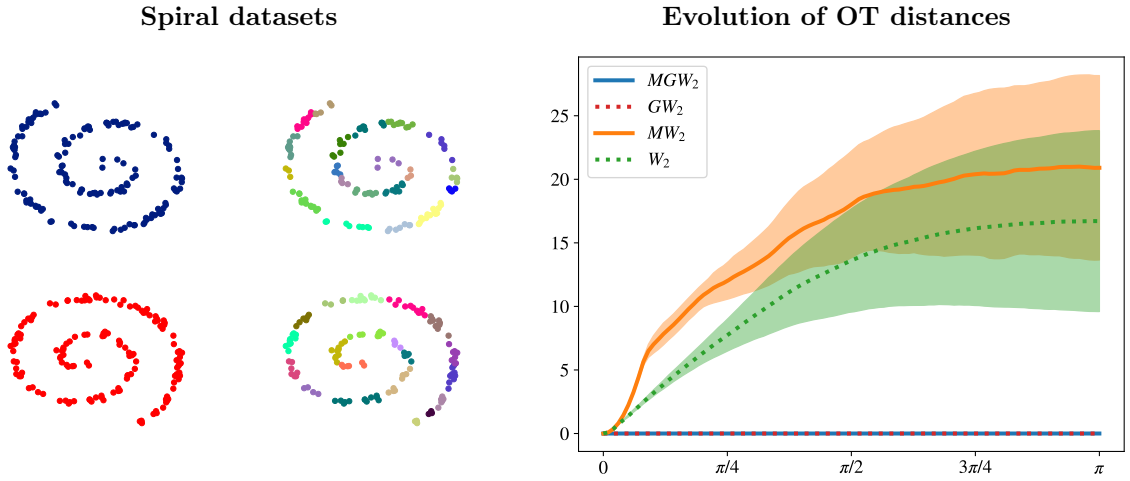

Figure 5.1: Left first column: spiral datasets (in blue and red) composed of 150 points of $\mathbb{R}^2$. The red dataset corresponds to points sampled from the distribution of the blue dataset rotated from π. Left second column: The two corresponding learned GMMs with 20 components via EM algorithm (each color corresponds to a Gaussian component of the GMMs). Right: evolution of MGW_2, GW_2, MW_2, and W_2 between the initial distribution (in blue) and the rotated ones in function of the angle of rotation. Experiments are averaged over 10 runs and the colored bands correspond to $+/-$ the standard deviation. This experiment is inspired from Vayer et al. (2019b).

[2]The package is accessible here: https://scikit-learn.org/stable/.

Difficulty of designing a transportation plan. The MGW_2 problem can be used on discrete data to provide an optimal coupling between the Gaussian components of the two Gaussian mixtures μ and ν that approximate the discrete data distributions $\hat{\mu}$ and $\hat{\nu}$. However, some applications require an optimal coupling between the points that compose $\hat{\mu}$ and $\hat{\nu}$. It is not straightforward to design such a transportation plan associated with the plan that minimizes the MGW_2 problem. More precisely, for two GMMs $\mu = \sum_k a_k \mu_k$ and $\nu = \sum_l b_l \nu_l$, the discrete MW_2 problem between the associated distributions $\tilde{\mu} \in \mathcal{P}(\mathcal{N}(\mathbb{R}^d))$ and $\tilde{\nu} \in \mathcal{P}(\mathcal{N}(\mathbb{R}^d))$ is equivalent to restricting the set of coupling to be GMMs in the continuous W_2 problem between μ and ν. Thus, there exists a direct relationship between the optimal couplings ω^* and π^* associated with these two latter problems. Indeed, when the mu_k are all non-degenerate distributions, we have for any $x, y \in \mathbb{R}^d$,

$$\pi^*(x,y) = \sum_{k,l} \omega^*_{k,l} p_{\mu_k}(x) \delta_{y = T^{k,l}_{W_2}(x)}, \tag{5.1}$$

where for $\mu_k = \mathrm{N}(m_k, \Sigma_k)$, $p_{\mu_k}(x) = (2\pi)^{-\frac{m}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-m_k)^T \Sigma_k^{-1}(x-m_k)\right]$ is the density of μ_k in x, and $T^{k,l}_{W_2}$ is the optimal affine transportation map between μ_k and $\nu_l = \mathrm{N}(m_l, \Sigma_l)$ associated with W_2, i.e. for all $x \in \mathbb{R}^d$,

$$T^{k,l}_{W_2}(x) = m_l + \Sigma_k^{-\frac{1}{2}} (\Sigma_k^{\frac{1}{2}} \Sigma_l \Sigma_k^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_k^{-\frac{1}{2}} (x - m_k) \ .$$

However, in the case of MGW_2, there doesn't exist to the best of our knowledge, any equivalent continuous formulation of the problem and so there is formally no such plan π^* associated with the discrete optimal coupling ω^* that minimizes the MGW_2 problem. Yet, supposing without any loss of generality that $d \geq d'$, and that $\nu \in GMM_L(\mathbb{R}^{d'})$, one could still define by analogy a plan π relatively to ω^* using (5.1) and replacing T_{W_2} by suited transportation maps between Gaussian components. A naive approach would be to simply replace T_{W_2} by the affine transportation map T_{GGW_2} associated with the Gaussian plan that minimizes the GW problems (GW_2-QG) and (GW_2-IP) between μ_k and ν_l that we have exhibited in the previous chapter:

$$T^{k,l}_{GGW_2}(x) = m_l + P_l \left(\widetilde{\mathrm{Id}}_{d'} D_l^{\frac{1}{2}} D_k^{(d')^{-\frac{1}{2}}} \right)^{[d',d]} P_k^T (x - m_k) \ ,$$

where (P_k, D_k) and (P_l, D_l) are the respective diagonalizations of Σ_k $(= P_k D_k P_k^T)$ and Σ_l $(= P_l D_l P_l^T)$ that sort the eigenvalues in non-increasing order. Yet this approach implies that each Gaussian component is transported independently of the others and so offers too many degrees of freedom. Observe indeed that T_{GGW_2} is defined up to $\widetilde{\mathrm{Id}}_{d'}$ that can be any matrix of the form $\mathrm{diag}((\pm 1)_{i \leq d'})$, implying that we have $2^{d'}$ possibilities for each Gaussian component. For each component, since we want that points that are close to each other but associated with different Gaussian components remain close when tranported, we need to determine, relatively to all the other components, which of the $2^{d'}$ possibilities is the correct one. To illustrate this problem, we show in Figure 5.2 a 2-dimensional example where we derive two different transport maps using T_{GGW_2} but each time with a different $\widetilde{\mathrm{Id}}_2$. If the transport on the middle of Figure 5.2 preserves the global structure of the distribution since two points that are close to each other but associated with different Gaussian components remain close when tranported, this is not the case of the transport on the right.

Therefore, in order to design a transport plan π_{MGW_2} associated with the MGW_2 problem, it is in general necessary to determine for each pair of indices k, l, which $\widetilde{\mathrm{Id}}_{d'}$ preserves, relatively to the others, the global structure of the GMM, which becomes a difficult combinatorial problem in itself as soon as the dimension d' is large. Alternatively, a less tedious solution to design such plan would be to derive explicitely the isometric transformation that has been implicitely applied to one of the two measures when solving the MGW_2 problem. This is the idea behind the mixture embedded Wasserstein distance that we introduce in the following section.

5.4 Embedded Wasserstein distance between GMMs

Similarly to Delon and Desolneux (2020), one can define an OT distance derived from EW_2 when μ and ν are GMMs by restricting the set of admissible couplings to be themselves GMMs.

Definition 5.4.1. *Let $\mu \in GMM_K(\mathbb{R}^d)$ and $\nu \in GMM_L(\mathbb{R}^{d'})$ and suppose that $d \geq d'$. We define*

$$MEW_2(\mu,\nu) = \inf \left\{ \inf_{\phi \in \mathrm{Isom}_{d'}(\mathbb{R}^d)} MW_2(\mu, \phi_\# \nu), \inf_{\psi \in \mathrm{Isom}_d(\mathbb{R}^{d'})} MW_2(\psi_\# \mu, \nu) \right\} \ . \tag{5.2}$$

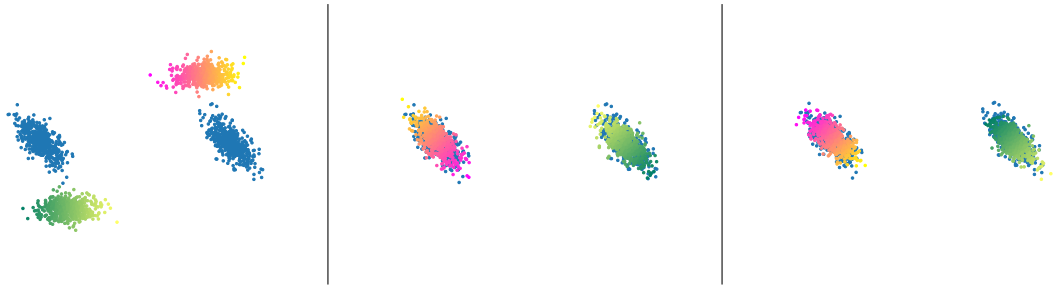

Figure 5.2: Left: two discrete distributions $\hat{\mu}$ (in gradient of colors) and $\hat{\nu}$ (in blue) that have been drawn from two GMMs. The colors have been added to $\hat{\mu}$ in order to visualize the couplings between $\hat{\mu}$ and $\hat{\nu}$. Middle: transport of $\hat{\mu}$ obtained by plugging the discrete plan that minimizes MGW_2 in (5.1), then using T_{GGW_2} with $\widetilde{\mathrm{Id}}_2 = \mathrm{Id}_2$ for all components to transport the points. Right: transport of $\hat{\mu}$ obtained the same way as previously, but with another $\widetilde{\mathrm{Id}}_2$.

As before, one can reformulate this latter problem by observing that the isomorphic mappings for the Euclidean norm are necessarily of the form $Px + b$ with $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ and $b \in \mathbb{R}^d$. Similarly to EW_2, one can show that the infinimum in ϕ is always achieved and that MEW_2 satisfies all the properties of a pseudometric on $\mathcal{GMM}_\infty$ by simply replacing W_2 by MW_2 in the proofs of Corollary 3.3.5 and Theorem 3.3.10. Supposing without any loss of generality that $d \geq d'$ and using the equivalent discrete formulation (MW_2) of the MW_2 problem, we get that for $\mu = \sum_k a_k \mu_k$ and $\nu = \sum_l b_l \nu_l$, the problem is equivalent to

$$\inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \inf_{\omega \in \Pi(a,b)} \sum_{k,l} \omega_{k,l} W_2^2(\mu'_k, P_\# \nu'_l) \ , \tag{MEW_2}$$

where for any $k \leq K$ and $l \leq L$, μ'_k and ν'_l are the Gaussian components respectively associated to the centered GMMs $\bar{\mu}$ and $\bar{\nu}$. Note that μ'_k and ν'_l are not necessarily themselves centered.

5.4.1 Numerical solver

This time, it is not possible to derive analytically the closed form of the optimal P^* for Problem (MEW_2). However, one can still solve the problem numerically using an alternate minimization scheme. Indeed, Problem (MEW_2) is not convex in P and ω, but is convex in ω if P is fixed and is furthermore a simple small-scale discrete OT problem in that case, which motivates the use of an alternating optimization scheme for solving this problem. However, Problem (MEW_2) is not convex in P for a fixed ω because the feasible set, i.e. the Stiefel manifold $\mathbb{V}_{d'}(\mathbb{R}^d)$, is not convex. For a fixed ω, the minimization in P can be done by projected gradient descent (Calamai and Moré, 1987), i.e. for a given iterate $P^{\{i\}}$ and a given ω, the next iterate $P^{\{i+1\}}$ is given by

$$P^{\{i+1\}} = \kappa_{\mathbb{V}_{d'}(\mathbb{R}^d)} \left(P^{\{i\}} - \eta \frac{\partial J_\omega(P^{\{i\}})}{\partial P} \right) \ ,$$

where $\kappa_{\mathbb{V}_{d'}(\mathbb{R}^d)}$ is the projection mapping on the Stiefel manifold, where $\eta > 0$ and where for all matrices P of size $d' \times d$, $J_\omega(P) = \sum_{k,l} \omega_{k,l} W_2(\mu'_k, P_\# \nu'_l)$. As we have seen above in Proposition 3.3.7, for all P of size $d' \times d$, the projection $\kappa_{\mathbb{V}_{d'}(\mathbb{R}^d)}$ can be written

$$\kappa_{\mathbb{V}_{d'}(\mathbb{R}^d)}(P) = U_P \, \mathrm{Id}_{d'}^{[d,d']} \, V_P^T \ ,$$

where $U_P \in \mathbb{O}(\mathbb{R}^d)$ and $V_P \in \mathbb{O}(\mathbb{R}^{d'})$ are respectively the left and right orthogonal matrices associated with the SVD of P. In a nutshell, this yields to Algorithm 4.

When μ and ν are only composed of non-degenerate Gaussian components, one can compute $\partial J_\omega(P)/\partial P$ either by using automatic differentiation (Baydin et al., 2018) or by using the following technical result, whose proof is postponed to Appendix A.3.

Lemma 5.4.2. *Let for any $k \leq K$, $\mu_k = \mathrm{N}(m_{0k}, \Sigma_{0k})$ with $m_{0k} \in \mathbb{R}^d$ and $\Sigma_{0k} \in \mathbb{S}_{++}^d$ and for any $l \leq L$, $\nu_l = \mathrm{N}(m_{1l}, \Sigma_{1l})$ with $m_{1l} \in \mathbb{R}^{d'}$ and $\Sigma_{1l} \in \mathbb{S}_{++}^{d'}$. For any ω in the $K \times L$ simplex, let $J_\omega : \mathbb{R}^{d \times d'} \to \mathbb{R}$ be the functional defined, for all matrix P of size $d \times d'$, by*

$$J_\omega(P) = \sum_{k,l} \omega_{k,l} W_2^2(\mu_k, P_\# \nu_l) \ .$$

Algorithm 4 Mixture embedded Wasserstein solver

Require: $\mu = \sum_k^K a_k \mu_k$, $\nu = \sum_l^L b_l \nu_l$, $P^{\{0\}} \in \mathbb{V}_{d'}(\mathbb{R}^d)$, $\eta > 0$.
while not converged **do**
  $[C]_{k,l} \leftarrow W_2^2(\mu_k, \nu_l)$ for $k = 1, \dots, K$; $l = 1, \dots, L$
  $\omega^{\{i\}} \leftarrow$ NETWORK-SIMPLEX(a, b, C) ▷ Solve a classic OT problem.
  while not converged **do** ▷ Do projected gradient descent on P.
    $A \leftarrow P^{\{i-1\}} - \eta \partial J_{\omega^{\{i\}}}(P^{\{i-1\}})/\partial P$
    $U, \Sigma, V^T \leftarrow \mathrm{SVD}(A)$
    $P^{\{i\}} \leftarrow U \, \mathrm{Id}_{d'}^{[d,d']} \, V^T$
  end while
end while
return ω, P

Then for any full-rank matrix P of size $d \times d'$, we have

$$\frac{\partial J_\omega(P)}{\partial P} = 2 \sum_{k,l} \omega_{k,l} \left[P m_{1l} m_{1l}^T - m_{0k} m_{1l}^T - \Sigma_{0k} P \Sigma_{1l}^{\frac{1}{2}} (\Sigma_{1l}^{\frac{1}{2}} P^T \Sigma_{0k} P \Sigma_{1l}^{\frac{1}{2}})^{-\frac{1}{2}} \Sigma_{1l}^{\frac{1}{2}} \right] .$$

Initialization procedure. Since the problem is non-convex, the solution to which Algorithm 4 converges strongly depends on the initialization of P. It is therefore crucial to design a good initialization procedure. To do so, we propose to use the *annealing* scheme introduced by Alvarez-Melis et al. (2019). More precisely, we propose to set the initial P as the solution of the following iterative procedure. First we solve an entropic-regularized W_2 problem between the two discrete measures $\mu^\circ = \sum_k a_k \delta_{m_{0k}}$ and $\nu^\circ = \sum_k b_l \delta_{m_{1l}}$ with a large value of regularization ε_0 in order to obtain a coupling $\omega^{\{1\}}$. Then we set

$$P^{\{1\}} = \kappa_{\mathbb{V}_{d'}(\mathbb{R}^d)} \left(\textstyle\sum_{k,l} \omega_{k,l}^{\{1\}} m_{0k} m_{1l}^T \right) .$$

We then solve another entropic-regularized W_2 problem, this time between μ° and $P_\#^{\{1\}} \nu^\circ$, using a smaller value of regularization $\varepsilon_1 = \alpha \times \varepsilon_0$ with $\alpha \in (0, 1)$. We obtain thus a new coupling $\omega^{\{2\}}$ and we can then derive $P^{\{2\}}$ as previously. We repeat this procedure N_{it} times until the regularization term $\varepsilon_{N_{it}}$ becomes small enough. This boils down to Algorithm 5.

Algorithm 5 Annealed initialization procedure for mixture embedded Wasserstein

Require: a, b, $\{m_{0k}\}_k^K$, $\{m_{1l}\}_l^L$, $\varepsilon_0 > 0$, $\alpha \in (0, 1)$, $P^{\{0\}} = \mathrm{Id}_{d'}^{[d,d']}$
for $i = 1, \dots, N_{it}$ **do**
  $[C]_{k,l} \leftarrow \|m_{0k} - P^{\{i-1\}} m_{1l}\|^2$
  $\omega^{\{i\}} \leftarrow$ SINKHORN-KNOPP$(a, b, C, \varepsilon_{i-1})$ ▷ Solve a regularized OT problem using Algorithm 1.
  $A \leftarrow \sum_{k,l} \omega_{k,l}^{\{i\}} m_{0k} m_{1l}^T$
  $U, \Sigma, V^T \leftarrow \mathrm{SVD}(A)$
  $P^{\{i\}} = U \, \mathrm{Id}_{d'}^{[d,d']} \, V^T$
  $\varepsilon_i \leftarrow \alpha \varepsilon_{i-1}$ ▷ Annealing scheme.
end for
return P

In practice, we set in all our experiments $\alpha = 0.95$ and $\varepsilon_0 = 1$ as in Alvarez-Melis et al. (2019). Furthermore we observed that in most cases, setting $N_{it} = 10$ was sufficient to obtain a good initialization of P for Algorithm 4.

5.4.2 Transportation plans and transportation maps

Since (MEW_2) has a continous equivalent formulation (5.2), one can derive from any optimal solution (ω^*, P^*) of the former, an optimal solution (π^*, ϕ^*) of the latter. More precisely, we have on the one hand for all $y \in \mathbb{R}^{d'}$, $\phi^*(y) = P^* y + b^*$, where $b^* = \mathbb{E}_{X \sim \mu}[X] - P^* \mathbb{E}_{Y \sim \nu}[Y]$, and on the other hand for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$,

$$\pi^*(x, y) = \sum_{k,l} \omega_{k,l}^* p_{\mu_k}(x) \delta_{y = \psi^* \circ T_{W_2}^{k,l}(x)} , \tag{5.3}$$

where $T^{k,l}_{W_2}(x)$ is the optimal W_2 transport map between μ'_k and $P^*_{\#}\nu'_l$ and $\psi^* : \mathbb{R}^d \to \mathbb{R}^{d'}$ is defined for all $x \in \mathbb{R}^d$ as $\psi^*(x) = P^{*T}(x - b^*)$. As in Delon and Desolneux (2020), it is possible to define a unique assignment of each x by setting for all $x \in \mathbb{R}^d$,

$$T_{\text{mean}}(x) = \mathbb{E}_{(X,Y)\sim\pi^*}[Y|X = x] = \frac{\sum_{k,l} \omega^*_{k,l} p_{\mu_k}(x) \psi^* \circ T^{k,l}_{W_2}(x)}{\sum_k a_k p_{\mu_k}(x) p_{\mu_k}(x)} .$$

Note that T_{mean} is not a Monge map since π^* is not of the form $(\text{Id}_d, T)_{\#}\mu$. In particular, $T_{\text{mean}\#}\mu$ is not equal to ν and $T_{\text{mean}\#}\mu$ is not necessarily the gradient of a convex function. Another possible way to define an assignment proposed by Delon and Desolneux (2020) is to define it as a random assignment for a fixed x, i.e.

$$T_{\text{rand}}(x) = \psi^* \circ T^{k,l}_{W_2}(x) \qquad \text{with probability} \qquad p_{k,l}(x) = \frac{\omega^*_{k,l} p_{\mu_k}(x)}{\sum_i a_i p_{\mu_i}(x)} .$$

Hence, when using MEW_2 to obtain an assignment between two sets $\{x_i\}_i^M$ and $\{y_j\}_j^N$ of respectively M and N vectors of $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, one can compute either $T_{\text{mean}}(x)$ or $T_{\text{rand}}(x)$ for each x_i, and then determine for each x_i which y_j is the closest of $T_{\text{mean}}(x_i)$ - or $T_{\text{rand}}(x_i)$ - using a nearest-neighbor algorithm (Fix and Hodges, 1951).

5.4.3 Improving the MGW_2 method

Inspired by the MEW_2 method presented above, we propose in this section to improve the MGW_2 method by: (i) proposing an annealed scheme similarly to Algorithm 5 in order to reduce the chances of converging to sub-optimal local minima, (ii) designing a transportation plan for MGW_2 similarly to (5.3).

Annealing scheme. Since Problem (MGW_2) is non-convex, we are only guaranteed to converge towards a local minimum when solving it with a classic non-regularized GW solver, e.g. the conditional gradient algorithm presented in Algorithm 2. Furthermore, the convergence towards a particular minimum depends strongly on the initialization of the coupling ω. Since the discrete GW problem in MGW_2 is of very small scale and so not costly in itself, we propose, by anology with MEW_2, to use a similar annealing scheme as in Algorithm 5 to reduce the chance of converging to a sub-optimal local minimum. More precisely, this gives the following algorithm.

Algorithm 6 Annealed mixture Gromov-Wasserstein solver

Require: $\mu = \sum_k^K a_k \mu_k$, $\nu = \sum_l^L b_l \nu_l$, $\alpha \in (0, 1)$, ε_0, $\omega^{\{0\}} = ab^T$
1: $[C^x]_{k,i} \leftarrow W_2^2(\mu_k, \mu_i)$ for $k = 1, \dots, K$, $i = 1, \dots, K$
2: $[C^y]_{l,j} \leftarrow W_2^2(\nu_l, \nu_j)$ for $l = 1, \dots, L$, $j = 1, \dots, L$
3: **for** $n = 1, \dots, N_{it}$ **do**
4: $\quad \omega^{\{n\}} \leftarrow \varepsilon\text{-GW}(a, b, C^x, C^y, \varepsilon_{n-1}, \omega^{\{n-1\}})$ ▷ Solve a regularized GW problem (Algorithm 3).
5: $\quad \varepsilon_n \leftarrow \alpha\varepsilon_{n-1}$ ▷ Annealing scheme.
6: **end for**
7: **return** $\text{GW}(a, b, C^x, C^y, \omega^{\{N_{it}\}})$ ▷ Solve the non-regularized GW problem (Algorithm 2).

As previously, we set in our experiments $\alpha = 0.95$ and $\varepsilon_0 = 1$ as in Alvarez-Melis et al. (2019) and we observed that, in toy cases where we know what the global minimum is, that $N_{it} = 10$ seemed to be a sufficient number of iterations to prevent the algorithm from converging towards a sub-optimal minimum.

Designing a transportation plan. Still by analogy with MEW_2, one can design a transportation plan for MGW_2 by defining a matrix $P_{MGW_2} \in \mathbb{V}_{d'}(\mathbb{R}^d)$ and a vector $b_{MGW_2} \in \mathbb{R}^d$, and then replacing $T_{W_2} \circ \psi^*$ in (5.3) by $T_{W_2} \circ \psi_{MGW_2}$, where for all $x \in \mathbb{R}^d$, $\psi_{MGW_2}(x) = P^T_{MGW_2}(x - b_{MGW_2})$. Given two GMMs $\mu = \sum_k a_k \mu_k$ and $\nu = \sum_l b_l \nu_l$ respectively in $GMM_K(\mathbb{R}^d)$ and $GMM_L(\mathbb{R}^{d'})$ and given the optimal discrete plan ω^* solution of Problem (MGW_2), one can define the matrix P_{MGW_2} as the solution of the following problem

$$\inf_{P \in \mathbb{V}_{d'}(\mathbb{R}^d)} \sum_{k,l} \omega^*_{k,l} W_2^2(\mu'_k, P_{\#}\nu'_l) , \tag{5.4}$$

where μ'_k and ν'_l are the Gaussian component of the centered GMMs $\bar{\mu}$ and $\bar{\nu}$, then we can set $b_{MGW_2} = \mathbb{E}_{X\sim\mu}[X] - P_{MGW_2}\mathbb{E}_{Y\sim\mu}[Y]$. As above, this problem can be solved numerically by performing a projected

gradient descent on P, using either automatic differentiation or Lemma 5.4.2. This is also a non-convex optimization problem since $\mathbb{V}_{d'}(\mathbb{R}^d)$ is non-convex and so the solution given by the projected gradient descent depends on the initialization. We propose thus to initialize with the projection on the Stiefel manifold of the discrete cross-covariance matrix between the means of the Gaussian components, i.e.

$$P_{MGW_2}^{\{0\}} = \kappa_{\mathbb{V}_{d'}(\mathbb{R}^d)}\left(\sum_{k,l} \omega_{k,l}^* m_{0k} m_{1l}^T\right).$$

Finally, using P_{MGW_2} one can define a continous plan π_{MGW_2} associated with the discrete optimal plan ω^* solution of the MGW_2 problem similarly to (5.3). We can therefore use MGW_2 to transport distributions, using as previously either T_{mean} or T_{rand}. We can also, as for MEW_2, use MGW_2 to obtain an assignment between two sets of points.

5.5 Experiments

In what follows, we use MGW_2 and MEW_2 to solve Gromov-Wasserstein related tasks on various datasets. More precisely, we apply first the two methods on simple toy low-dimensional GMMs. Then, we show that both methods can be used to solve relatively efficiently GW related tasks on real datasets in large scale settings involving sometimes several tens of thousands of points. We apply thus our methods to two shape matching problems, then to color transfer on hyperspectral images. In all our experiments, we use the numerical solvers provided by the Python Optimal Transport (POT) package[3] (Flamary et al., 2021) that implements the network-simplex algorithm for the classic linear OT problems, the Sinkhorn-Knopp algorithm for the regularized linear OT problems, as well as the non-regularized and regularized solvers of the GW problem presented in 3.1.4.

5.5.1 Low dimensional GMMs

In Figure 5.3, we use again the example of Figure 5.2 and we derive an optimal transport plan for the MGW_2 problem as described in Section 5.4.3. We also show the plan obtained by solving the EW_2 problem. One can see that with both solutions, the global structure of the distribution is preserved in the sense that points that are closed to each other but in two different Gaussian components have been sent to points that are also close to each other but in different Gaussian components.

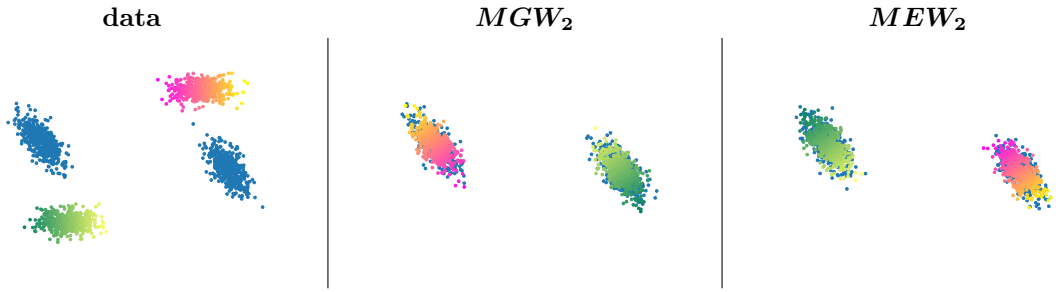

Figure 5.3: Left: two discrete distributions $\hat{\mu}$ (in gradient of colors) and $\hat{\nu}$ (in blue) that have been drawn from two GMMs. The colors have been added to $\hat{\mu}$ in order to visualize the couplings between $\hat{\mu}$ and $\hat{\nu}$. Middle: transport of $\hat{\mu}$ obtained by solving the MGW_2 problem, then deriving $P_{MGW_2} \in \mathbb{V}_2(\mathbb{R}^2)$ by solving Problem (5.4). Right: transport of $\hat{\mu}$ obtained by solving the MEW_2 problem.

5.5.2 Application to shape matching

We apply now our methods on two shape matching problems. The first one consists in reproducing an experiment originally conducted in Rustamov et al. (2013) and presented in Solomon et al. (2016) with the use of entropic-regularized GW, that aims to revover the cyclical nature of a horse's gallop. The second problem consists in drawing correspondences betwen human shaped meshes from the SHREC'19 dataset[4] (Melzi et al., 2019) in the sense that we aim to assign a hand with a hand, a foot with a foot, etc. Note that the goal of these experiments is not to obtain state-of-the-art results in shape-matching, but rather to demonstrate the usability of our methods in moderate-to-large scale settings.

[3]The package is accessible here: https://pythonot.github.io/.
[4]The SHREC'19 dataset is accessible here: http://profs.scienze.univr.it/ marin/shrec19/

Galloping horse sequence. Here we repoduce the experiment of the galloping horse, that has been originally conducted in Rustamov et al. (2013) and presented in Solomon et al. (2016) with the use of entropic-regularized GW. The goal of this experiment is to compute a matrix of pairwise distances between the 45 meshes representing a galloping horse, and then to conduct a Multi-Dimensional Scaling (MDS) (Borg and Groenen, 2005) - which roughly can be thought as a generalization of PCA - of the pairwise distances in order to plot each mesh as a 2-dimensional point. The results can be found in Figure 5.4. As in Solomon et al. (2016), the cyclical nature of the motion is recovered in both cases when MGW_2 or MEW_2 is used to compare the meshes. Each mesh is composed of approximately 9000 vertices and the average time to compute one distance when using the POT implementation of the entropic-regularized GW solver is around 30 minutes which makes the computation of the full pairwise distance matrix impractical, as mentioned in Solomon et al. (2016). In constrast, when using our methods with GMMs with $K = 20$ components, it took us only approximately 10 minutes to compute the full distance matrix using MGW_2, and around one hour using MEW_2, these times including the fitting of all the GMMs.

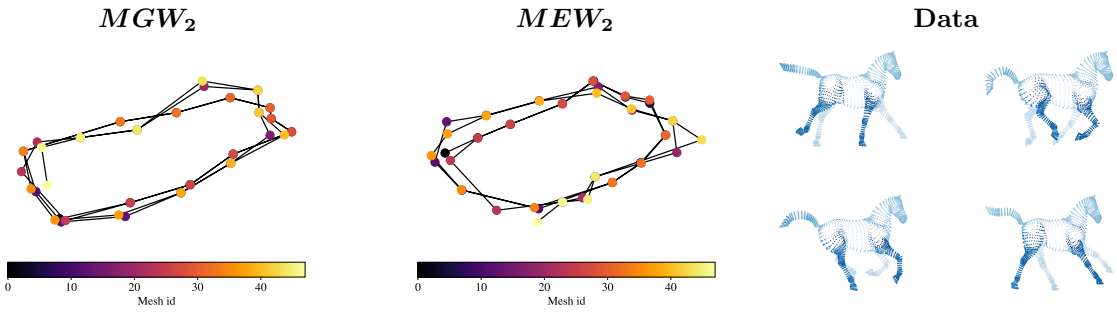

Figure 5.4: MDS on the galloping horse animation using the MGW_2 distance (left), and the MEW_2 distance (middle). Each point corresponds to a given mesh and the meshes are colored in function of their number in the sequence. Right: 4 examples among the 45 meshes that composes the sequence. The computations of both distances have been done by first fitting GMMs with 20 components on each mesh independently.

Local minima. To highlight the importance of using an annealing scheme when deriving MGW_2 or MEW_2, we have reconducted the previous experiment but this time without the annealing schemes described in Algorithm 6 and Algorithm 5. In Figure 5.5, we plot the evolutions of the values of MGW_2^2 and MEW_2^2 between one given fixed mesh and all the others. In both cases, the annealing scheme seems to be useful to prevent the solver to converge towards sub-optimal mininima. However, if the MGW_2 solver seems to often converge to the same optimum regardless the use of the annealing scheme, this is not the case of MEW_2 which, without the annealing initialization procedure (Algorithm 5), converges most of the time to a sub-optimal minimum, so much that the periodical aspect doesn't even appear in that case. Beside to highlight the importance of using a good initialization, this experiment also emphasizes the fact that when solving a GW problem with the non-regularized solver or the entropic solver presented in Section 3.1.4, we are not at all guaranteed to converge towards a global minimum and, more critically, we have in general no ways to know if the solution we converged to is actually optimal or sub-optimal.

Matching human shaped meshes. To demonstrate the usability of our methods in larger scale settings, we use the SHREC'19 dataset that contains human shaped meshes that can sometimes be composed of more than 300000 vertices. Our goal is to draw correspondences between the shapes using only the information of the vertices (the dataset also includes edges). To do so, we first fit independently GMMs with 20 components on each mesh and we derive directly couplings at the scale of the Gaussian components that represents the different parts of the bodies. In such large scale settings, the main bottleneck of the methods in terms of computational time is clearly the fitting of the GMMs that can take at worst 2 minutes for the meshes composed of the highest number of vertices. The results are displayed on Figure 5.6. Observe that in most cases, both methods seem to be able to match correctly the colored parts. Yet in the last row, MEW_2 matches a leg at the left in red to an arm at the right. This probably implies that the method has been trapped in a local minimum despite the annealing initialization procedure. Finally, note that we presented here cases where the methods performed relatively well, but

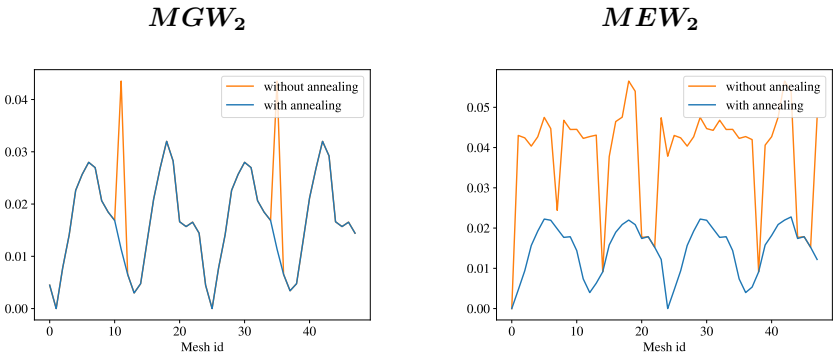

Figure 5.5: Left: Evolution of MGW_2^2 between the second mesh and all the others, using an annealing scheme (Algorithm 6) in blue, and without the annealing scheme in orange. Right: Evolution of MEW_2^2 between the first mesh and all the others, with the annealing initialization procedure (Algorithm 5) in blue, and without in orange. The computation of both distances have been done by first fitting GMMs with 20 components on each mesh independently.

there are cases where MGW_2 or MEW_2 fail to find correct correspondences and exhibit behaviors similar to MEW_2 in the last row, which suggests that the methods converge sometimes to sub-optimal minima despite the annealing schemes.

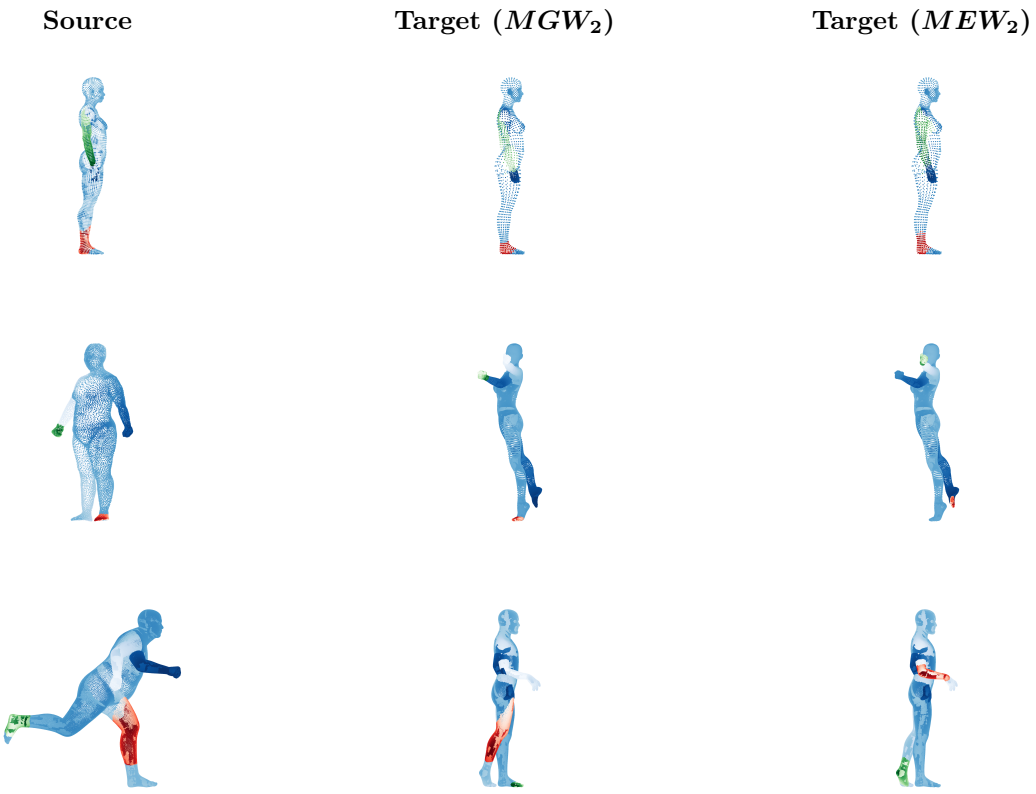

Figure 5.6: Shape matching between human-shaped meshes using MGW_2 (middle) and MEW_2 (right). Each shape on the left column is matched with the shapes on the same row. GMMs with 20 components have been fitted independently on each shape and the points colored in green and red correspond to Gaussian components that are matched together when solving MGW_2 or MEW_2. From left to right and top to bottom, the meshes are composed respectively of 84912, 30300, 75000, 273624, 360678, and 360357 vertices.

5.5.3 Application to hyperspectral image color transfer

The goal here is to reproduce the experiment of color transfer conducted in Delon and Desolneux (2020), but this time using a hyperspectral image, i.e an image with more than 3 color channels. More precisely, we aim to create an RGB image from an hyperspectral image u using the colors of another RGB image v. To do so, we consider images as empirical distributions in the color spaces and we solve a Gromov-Wasserstein problem between the distributions $\hat{\mu} = \frac{1}{M}\sum_k^M \delta_{u_k}$ and $\hat{\nu} = \frac{1}{N}\sum_l^N \delta_{v_l}$, where M and N are the number of pixels in respectively the hyperspectral image and the RGB image we use as color palette, and $\{u_k\}_k^M$ and $\{v_l\}_l^N$ are the values at each pixel, i.e for here all l, $v_l \in \mathbb{R}^3$ and for all k, $u_k \in \mathbb{R}^d$ with $d > 3$. We thus fit two GMMs μ and ν and respectively $\hat{\mu}$ and $\hat{\nu}$ and we use MGW_2 or MEW_2 to derive a mapping $T_{\text{mean}} : \mathbb{R}^d \to \mathbb{R}^3$, as described in Section 5.4.2. We apply this process to a hyperspectral image of 512×512 pixels with 15 channels that are displayed in Figure 5.7 top left. We use as color palettes two paintings by Gauguin and Renoir, displayed in Figure 5.7 top right, that are respectively *Manhana no atua* (top) and *Le déjeuner des canotiers* (bottom). These two images are composed of 1024×768 pixels. The resulting images $T_{\text{mean}}(u)$ are displayed in Figure 5.7 bottom (Gauguin at the left and Renoir at the right). For this experiment, we observed that setting the number of Gaussian components to $K = 15$ was a good compromise between capturing the complexity of the color distributions and obtaining a relatively regular mapping T_{mean}. This experiment shows that MGW_2 and MEW_2 can be used in large scale settings: observe indeed that the color distributions $\hat{\mu}$ and $\hat{\nu}$ are composed respectively of approximatively 300000 and 800000 points, which makes the problem intractable with entropic-GW solvers such as Peyré et al. (2016) or Solomon et al. (2016). Furthermore, note also that $d = 15$ is already a relatively high dimension in the context of Gromov-Wasserstein. In term of computation time, the fitting of the GMM for the hyperspectral image takes aproximatively 15 minutes against one minute for the GMM for the RGB image. The projected gradient descent becomes rather slow in that setting, which makes it preferable to few updates of P at each step of Algorithm 4 for the computation of MEW_2. Finally, for both methods, it takes around 20 minutes to compute the whole RGB image $T_{\text{mean}}(u)$.

5.6 Discussion

In this chapter, we have introduced two new OT distances on the set of Gaussian mixture models, MGW_2 and MEW_2, and we have shown that they both can be used to solve relatively efficiently Gromov-Wasserstein related problems on Euclidean spaces, especially in moderate-to-large scale settings involving several tens of thousands of points. These OT distances are also by design particularly suited to settings where there already exists a kind of clustering structure in the data. This being said, if MEW_2 remains an efficient alternative to the entropic GW solvers proposed by Peyré et al. (2016) and Solomon et al. (2016), we observed that the method was actually slower and perhaps harder to tune than MGW_2 for roughly the same quality of results, and so we believe that MGW_2 is a better choice in practice. This latter distance is part of the families of Gromov-Wasserstein type OT distances that reduce the size of the GW problem by quantization (Chowdhury et al., 2021) or by clustering (Xu et al., 2019a; Blumberg et al., 2020). To the best of our knowledge, however, no such methods specific to the Euclidean case had already been proposed in the literature. MGW_2 could also be easily extended to other type of mixtures as soon as we have an identifiability property between the mixtures and the probability distributions on the space of the distributions that compose the mixtures. If in the Euclidean setting GMMs seem to be versatile enough to represent large classes of concrete and applied problems, an interesting extension on our work could be to consider mixture of distributions on non-Euclidean spaces.

Computationally speaking, the main bottleneck of the method probably comes from the fitting of the GMMs with the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) which can become relatively costly in large scale settings or as soon as the dimension increases. If the EM algorithm remains invariably the classical algorithm for learning GMMs, some recent approaches (Hosseini and Sra, 2020; Sembach et al., 2022; Pasande et al., 2022) have proposed alternative algorithms that seems to outperform it. These approaches are based on Riemannian stochastic optimization, leveraging the rich Riemmanian structure of the set of positive definite matrices. Another interesting alternative that has been shown to outperform the EM algorithm has been proposed by (Kolouri et al., 2018b) and is based on the minimization of the sliced-Wasserstein distance. Integrating this in our method could result thus in an approach fully-based on optimal transport.

Another possible limitation of our work lies in the fact that the MGW_2 solver converges sometimes to sub-optimal local minima. If the annealed procedure introduced in Section 5.4.3 seems to reduce this issue, we generally have no guarantee that the solution we have converged to is optimal. This is not

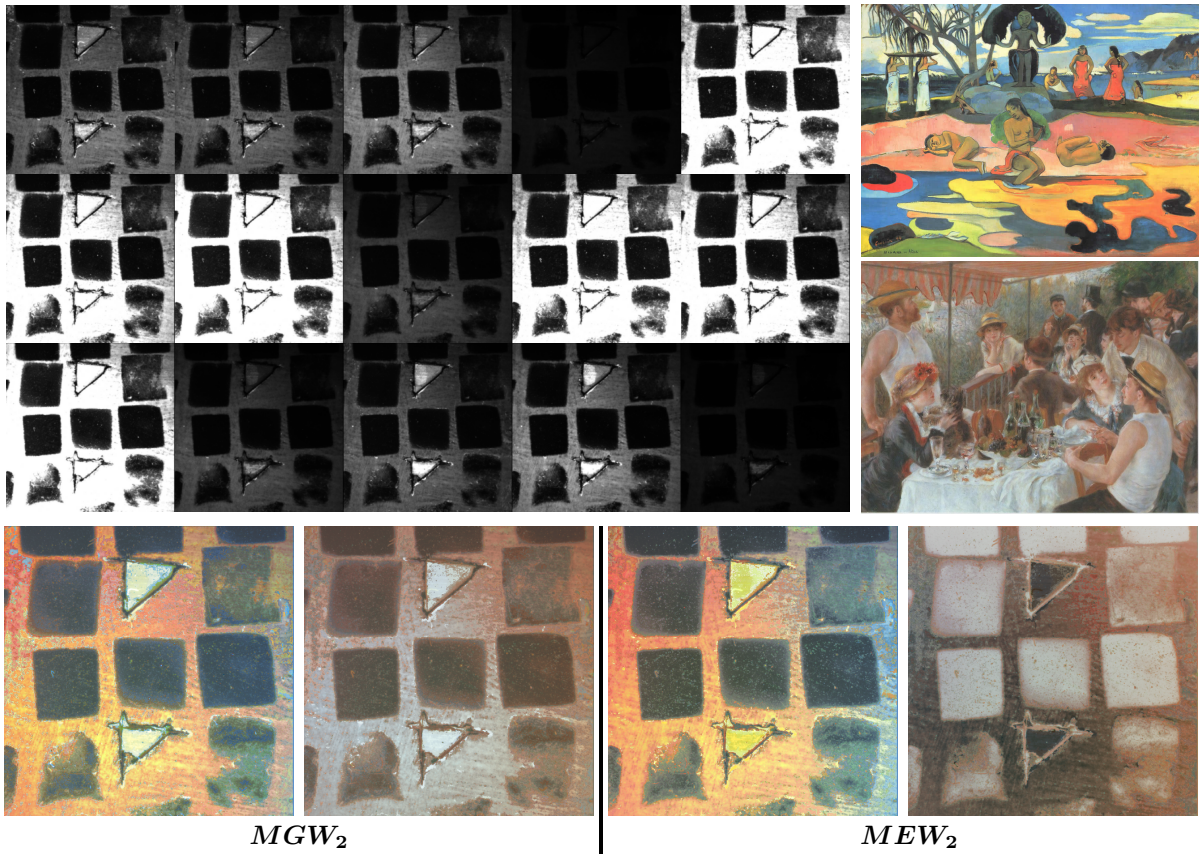

Figure 5.7: Color transfers between a hyperspectral image with 15 channels (top left) and two paintings by Gauguin and Renoir (top right, top to bottom). Bottom line: the obtained RGB images using MGW_2 and MEW_2. For this experiment, we used GMMs with 15 components. Image taken by Francesca Ramacciotti (Alma Mater Studiorum - University of Bologna) and Laure Cazals (supported by the European Commission in the framework of the GoGreen project (GA no. 101060768)).

specific to our method and comes from the gradient descent structure of the classic GW solvers described in Section 3.1.4. Still, when solving the GW problem between GMMs rather than solving it directly between the points, it is likely that we increase the probability of converging towards a sub-optimal local minimum because we inevitably introduce symmetries by simplifying the problem and so we probably increase in the mean time the number of local minima in the GW objective. In the Euclidean setting, the recent work of Ryner et al. (2023) proposes an algorithm for solving the GW problem that is guaranteed to converge toward a global minimum, leveraging the low-rank structure of the cost matrices when the cost functions are the squared Euclidean distances. A future perspective of work could be therefore to study if a similar idea could be applied for solving the MGW_2 problem.

Part II

Expressivity of deep push-forward generative models

Chapter 6

An introduction to generative modeling

Contents

In this chapter, we introduce the basic concepts of generative modeling and we present the most commonly used generative models in imaging science.

6.1 The generative modeling problem

Informally, the goal of generative modeling is to create "fake" data - synthetic images for instance - that look like they belong to a given dataset. This makes thus generative models directly useful for data augmentation (Antoniou et al., 2018; Haradal et al., 2018; Shao et al., 2019; Luo et al., 2020) that consists in artificially increasing the size and diversity of a dataset in order to improve the performances of machine learning methods. Generative modeling has been also sucessfully used to solve a wide range of inverse problems such as image inpaiting (Yu et al., 2018; Yeh et al., 2017; Song et al., 2021), super-resolution (Ledig et al., 2017), image colorization (Nazeri et al., 2018; Saharia et al., 2022), or audio source separation (Subakan and Smaragdis, 2018), just to name a few. Yet, the popularity of generative models can probably be explained more by the impressive aspect of their direct application rather than by their true scientific usefulness. To that extent, recent large scale text-to-image models such as DALL-E 2 (Ramesh et al., 2022) or Stable Diffusion (Rombach et al., 2022) have recently caught general audience's attention thanks to ther ability of generating photorealistic or artistic high definition images from simple text descriptions. In this chapter, we introduce the basic concepts of generative modeling and briefly introduce the models that are used today in the machine learning community. Note that we especially focus on generative models in the context of imaging science in this thesis, but most of the models presented here can also be used in other machine learning subfields, such as natural language processing or audio signal processing.

6.1.1 Mathematical formulation

In the following, we adopt the probabilistic point of view where data are seen as realizations of a random variable following a probability law. We suppose indeed that there exists an unknown underlying probability distribution ν on $\mathbb{R}^d$ such that the n points x_i forming the dataset are actually n samples that

have been drawn independently from ν. In this framework, the task of creating "fake" data mathematically translates into predicting new samples from ν given the information of the n *true* samples.

General approach. A general approach to solve the problem described above is to define a parametric family of distributions $\{\nu_\theta\}_{\theta\in\Theta}$ and to solve the following problem

$$\min_{\theta\in\Theta} D\left(\hat{\nu}, \nu_\theta\right) , \tag{6.1}$$

where D is a measure of dissimilarity between probability distributions and $\hat{\nu} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical data distribution. The idea behind this approach is that if the parametric family $\{\nu_\theta\}_{\theta\in\Theta}$ is well chosen and if $\{x_i\}_i^n$ is a representative sample set of ν, the optimal distribution ν_{θ^*} that minimizes (6.1) will be similar to ν. Hence, a generative model with this approach consists in first solving Problem (6.1) and then sampling from the obtained optimal distribution ν_{θ^*}. Note that in the next section, we will discuss more in details the two assumptions mentioned above. Finally, note also that D can be a distance as well as a divergence, or simply a functional that measures the dissimilarity between probability distributions without verifying any axioms of a metric, which is most of the time the case in practice. Still, these functionals are always more or less directly linked to a distance or a divergence, which can typically be the Wasserstein distance, the Kullback-Leibler or the Jensen-Shannon divergence. For μ and ν being two measures on $\mathbb{R}^d$ with μ absoluty continuous with respect to ν. The Kullback-Leibler divergence is defined as

$$D_{\mathrm{KL}}(\mu||\nu) = \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)\mathrm{d}\mu(x) ,$$

where $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is the Radon–Nikodym derivative of μ with respect to ν, i.e. the unique function $p_{\mu/\nu} : \mathbb{R}^d \to \mathbb{R}$ such that $\mathrm{d}\mu(x) = p_{\mu/\nu}(x)\mathrm{d}\nu(x)$ ν-almost everywhere. One can then define the Jensen-Shannon divergence as

$$D_{\mathrm{JS}}(\mu||\nu) = \tfrac{1}{2}D_{\mathrm{KL}}(\mu||\xi) + \tfrac{1}{2}D_{\mathrm{KL}}(\nu||\xi) ,$$

where $\xi = \frac{1}{2}(\mu+\nu)$ is the mixture distribution of μ and ν.

Sampling from the parametric distribution. Once Problem (6.1) is solved, a question that arises is: *how do we sample from the parametric distribution ν_{θ^*}?* A first approach to do so is to design the family of parametric distributions $\{\nu_\theta\}_{\theta\in\Theta}$ in such a way that for all $\theta\in\Theta$, ν_θ is easy to sample. An important example of this approach is when $\{\nu_\theta\}_{\theta\in\Theta}$ is defined as family of distribution of the form, for all $\theta\in\Theta$,

$$\nu_\theta = g_{\theta\#}\mu_{d'} ,$$

where $g_\theta : \mathbb{R}^{d'} \to \mathbb{R}^d$ is a parametric mapping and $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$ is the standard Gaussian distribution in dimension d'. In that case, we say that the model is a *push-forward generative model* since it consists in deriving a transport map - not necessarily optimal in the sense of Part I - that pushes $\mu_{d'}$ into a distribution close to ν. Alternatively, another approach consists in, rather than designing the family $\{\nu_\theta\}_{\theta\in\Theta}$ specifically to be sampled easily, sampling from ν_{θ^*} using a Markov Chain Monte Carlo (MCMC) method (Robert et al., 1999). Indeed, the generative modeling problem is inherently linked with the MCMC methods since these latters consist in sampling from a target probability distribution ν by designing an ergodic Markov chain that converges towards ν and such that its initial distribution is easy to sample. We refer to Andrieu et al. (2003) for a machine learning oriented introduction on these methods. This approach consists thus in designing a Markov chain that can be thought as a functional G_θ defined on the space of trajectories of a Brownian motion $(B_t)_{t\in\mathfrak{T}}$[1], and such that for a given realization $(z_t)_{t\in\mathfrak{T}}$ of $(B_t)_{t\in\mathfrak{T}}$, $G_\theta((z_t)_{t\in\mathfrak{T}})$ is a sample of ν_θ. We will refer to G_θ as the *generation dynamics* of the generative model.

Supervised variant. The generative modeling problem as described above can be classified as an unsupervised learning problem since it doesn't require in any way the data to be labeled. However, there exists an important supervized variant of the generative modeling problem which is known as *conditional generative modeling*. Given a labeled dataset $\{x_i, y_i\}_i^n$ such that the x_i are n independent realizations of a random variable X with unknown probability ν and the y_i are n independent realizations of a random variable Y, a conditional generative model aims at approximating the regular conditional probability

[1] Note that $\mathfrak{T}$ can be a subset of $\mathbb{N}$ as well as a subset of $\mathbb{R}$.

distributions $\nu_{X|Y}(.|Y = y_i)$[2] for each y_i instead of the probability distribution ν. This supervised variant makes generative modeling particularly useful for solving diverse inverse problems such as image inpainting (Song et al., 2021), super-resolution (Ledig et al., 2017) image colorization (Nazeri et al., 2018), or more generally image-to-image translation (Isola et al., 2017; Saharia et al., 2022).

Generative versus discriminative models. Here we would like to clarify the fact that the meaning of the term *generative model* differs from its use in statistics. Indeed a generative model, in the statistical sense of the term, can be any model of the joint law of the observations (X, Y), where X is the random variable corresponding to the data and Y is the random variable corresponding to the labels. This is opposed to the discriminative models that model the conditional law of the observations X given the labels Y. Hence, the unsupervised generative models in the machine learning sense of the term are also generative models in the statistical sense, whereas the conditional generative models are discriminative models. Conversely, not all generative models in the statistical sense are generative models in the machine learning sense, but many can be used as such. For instance, one can use a Gaussian mixture model to synthesize new data.

6.1.2 Challenges of generative modeling in imaging science

Before introducing the major approaches which are used for generative modeling, we explain in this section what makes generative modeling a challenging problem, focusing mainly on the context of imaging science.

Ill-posed problem. First of all, it should be noted that the problem described above is a very ill-posed problem in the sense that for a given dataset $\{x_i\}_i^n$, there exists an infinite number of probability distributions ν from which the points x_i could have been sampled. The goal of generative modeling is thus to approach a distribution ν that *could be* the underlying law of the dataset, even if we have no ways to verify if this is actually the case and we don't even know if such an underlying law does exist. Furthermore, note also that we can only hope to approach a law ν for which $\{x_i\}_i^n$ is a representative sample set. Thus, in the case where we know the distribution ν from which the x_i have been sampled from but the sample set $\{x_i\}_i^n$ misses significant parts of its support, we could never manage to recover the missing parts of ν.

Choosing the right parametric family of distribution. A fundamental difficulty of generative modeling lies in the choice of the parametric family of distributions $\{\nu_\theta\}_{\theta\in\Theta}$. A first issue we might encounter if this family is not well chosen is of course that we might not be able to minimize D correctly, implying that we will not be able to correctly approach ν and so to generate good quality synthetic samples. More interestingly, another issue we might also encounter would be, at the opposite, to over-minimize D and to obtain a distribution ν_{θ^*} that is similar to the empirical distribution $\hat{\nu} = \frac{1}{n}\sum_i \delta_{x_i}$. This would imply that our model has overfitted the dataset and has no ability to generate synthetic data that are different from the original data $\{x_i\}_i^n$. Thus, one must choose the family $\{\nu_\theta\}_{\theta\in\Theta}$ in a way such that ν_θ can be close of the underlying distribution ν but at the same time cannot be to close to the empirical distribution $\hat{\nu}$.

Approaching a distribution of high dimension. Another major difficulty of generative modeling in the context of imaging science is that it involves high dimensional distributions. Indeed, in the framework described above, images are seen as high dimensional vectors with each pixel corresponding to a dimension. Thus, for a medium-high resolution image of 1024×768 pixels, the dimension of the probability distribution ν we want to approach is already around 800000. Yet, if ν lives in a highly dimensional space, it is likely that it actually lies in a low dimensional sub-manifold. This common assumption in imaging science is known as the *manifold hypothesis* and has been partly validated empirically by Pope et al. (2020). The basic idea behind this hypothesis is that the value of a given pixel in an image is strongly conditioned by the values of all the other pixels. On one hand, the manifold hypothesis mitigates the dimension of the problem since the intrisic dimension of the distribution we want to approach is in reality much lower than the dimension of the ambient space. On the other hand, it complexifies even more the problem since ν is an highly degenerate measure and so its support is of Lesbegue measure zero.

[2]These regular conditional probability distributions are defined as the unique probability measures on $\mathbb{R}^d$ such that, for all Borel set A of $\mathbb{R}^d$, $\nu_{X|Y}(\mathsf{A}|Y = y_i) = \mathbb{E}[\chi_{\mathsf{A}}|Y = y_i]$ where χ_{A} denotes the characteristic function of the set A. In the following, we will note $\nu_{X|Y}(\mathsf{A}|y_i)$ instead of $\nu_{X|Y}(\mathsf{A}|Y = y_i)$.

6.2 Deep generative modeling

The recent emergence of generative modeling in imaging science, or more generally, in data science, is closely linked to the emergence of deep learning approaches in these fields. Indeed, the commonly used generative models in imaging science nowadays all invariably use a deep neural network at some point in the model. In this section, we present briefly some key notions of deep learning, then we introduce the different models that are commonly used in imaging science.

6.2.1 A brief introduction to deep learning

We call deep learning any machine learning method that builds upon a *deep neural network* architecture. In the imaging science context, the deep neural networks that are used are most of the time Convolutional Neural Networks (CNN) (LeCun et al., 1989; Krizhevsky et al., 2012) which are a particular type of feed-forward neural networks.

Feed-forward neural network. Given a L-tuple of dimensions $\{d_l\}_{l=1}^L$ such that $d_1 = d'$ and $d_L = d$, a feed-forward neural network is any parametric mapping $f_\theta : \mathbb{R}^{d'} \to \mathbb{R}^d$ of the form

$$f_\theta = f_{\theta_L}^L \circ \dots f_{\theta_1}^1 \ ,$$

with $\theta = (\theta_1, \dots, \theta_L)$ and where for all $1 \leq l \leq L$, $f_{\theta_l}^l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_{l+1}}$ is of the form, for all $x \in \mathbb{R}^{d_l}$

$$f_{\theta_l}^l(x) = \rho_l(W_l x + b_l) \ ,$$

where $\theta_l = (W_l, b_l)$, $W_l \in \mathbb{R}^{d_{l+1} \times d_l}$, $b_l \in \mathbb{R}^{d_{l+1}}$ and $\rho_l : \mathbb{R} \to \mathbb{R}$ is a non-linear function applied elementwise that is called *activation function*. The W_l and b_l are referred to respectively as *weights* and *bias* and the functions $f_{\theta_l}^l$ are called the *layers* of the neural network. The most commonly used activations functions are the Rectified Linear Unit (ReLU), i.e. $\mathrm{ReLU}(x) = \max\{0, x\}$, the sigmoid function $\mathrm{sigmoid}(x) = (1+\exp[-x])^{-1}$, the softmax function or the hyperbolic tangent function $\tanh(x) = (\exp[2x] - 1)(\exp[2x] + 1)^{-1}$. An important particular case is the *ReLU networks* in which the ρ_l are all ReLU activation functions. In that case, it is well known that f_θ is piecewise-linear with at most $2^{d_2 + \dots d_L}$ pieces (Montufar et al., 2014). A CNN is a particular type of feedforward neural network in which the weights matrices W_l are imposed to be convolution matrices. This convolutional structure is particularly adapted for image processing and allows to considerably reduce the number of parameters at each layers. This allows in practice to encode complex models that would be considerably larger in terms of parameters if this structure were not imposed, which explains why CNNs have been key elements to the success of deep learning in imaging science.

Modern deep neural network architectures. In practice, the modern deep neural networks architectures used in the literature are much more complex than the simple model of feedforward neural networks described above. Indeed, practitioners typically add *normalizations layers* such as batch normalization (Ioffe and Szegedy, 2015), weight normalization (Salimans and Kingma, 2016) or spectral normalization (Miyato et al., 2018). They also commonly add *skip-connections* (Bishop, 1995; Ronneberger et al., 2015; He et al., 2016), i.e. structures where the output at a given layer is re-used sometimes several layers later. More recently, following the breakthrough of transformers (Vaswani et al., 2017) in natural language processing, the *Vision Tranformers* (ViT) networks (Dosovitskiy et al., 2020) have been shown to outperform CNNs for various computer vision tasks. These networks are built on non-local *attention layers* (Wang et al., 2018). Still, modern deep neural networks architecture in imaging science are still roughly composition of functions that can been encoded with some basic operations, such as tensor multiplications, additions and concatenations, and with some non-linear activation functions applied elementwise. Hence, the model of feedforward neural networks stays a relevant simplified model for most of modern neural network architectures used nowadays.

Expressivity of deep neural networks. In machine learning, neural networks are mostly used to approximate functions. One question that arises is: *are neural networks able to approximate any mapping?* Examining the expressivity of deep neural networks is still an active research field. The universal approximation theorem (Funahashi, 1989; Cybenko, 1989; Hornik et al., 1989) states that shallow feedforward neural networks are universal approximators, in the sense that any mapping can

theoretically be approximated with any precision by a neural network composed of one single layer but with a pottentially infinite number of neurons. More recently, Hanin (2019) has shown that deep feedforward neural networks with finite numbers of neurons on each layer but with a potentially an infinite number of layers could approximate any continuous mapping with any precision as long as it has a sufficient number of neurons at each layer. In practice, deep neural Networks seem to be much more limited in terms of expresivity, mainly because of their training.

Learning a deep neural network. A general approach to train a deep neural network to approximate a target mapping f whose values are known on a dataset of n points $\{x_i\}_i^n$, is to define a *loss function* L and to solve the following empirical risk minimization problem

$$\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(f_\theta(x_i), f(x_i)) \; .$$

The typical method to do so is to use a *stochastic gradient descent* algorithm (Robbins and Monro, 1951) or one of its numerous recent refinements (Kingma and Ba, 2015; Ruder, 2016). This algorithm roughly consists in randomly selecting at each iterate a subset of data that we call *minibatch* and computing the gradient of the sum on the minibatch instead of computing the gradient of the whole sum as in traditional gradient descent. In terms of order of magnitude, the number of data n in deep learning is typically larger than 10^6, whereas the number of data in each minibatch can vary from a dozen to a thousand, depending on the application. It is important to note that minimizing the empirical risk is in general not a convex problem, and so one can only hope to converge towards a local minimum (which is not necessarily a global minimum) when training a deep neural network. The *stochastic gradient descent* algorithm involves computing the gradient of f_θ with respect to the parameters $\theta = (W_1, b_1, \dots, W_L, b_l)$. This is possible thanks to the *backpropagation* algorithm (Rumelhart et al., 1986) which roughly consists in retropropagating the gradient through the $f^l_{\theta_l}$, leveraging the Leibniz's chain rule. Observe that the backpropagation algorithm requires all the $f^l_{\theta_l}$, and so all the activation ρ_l, to be differentiable almost everywhere. This is the main reason why all the commonly used activation functions are chosen Lipschitz, which limits in practice the expressivity of deep neural network to Lipschitz mappings.

6.2.2 The most commonly used deep generative models in imaging science

Here we present the three models that are the most commonly used generative models in imaging science. These models are the *Variational Autoencoders* (VAEs) (Kingma and Welling, 2014), the *Generative Adversarial Networks* (GANs) (Goodfellow et al., 2014) and the more recent *Score-based Generative Models* (SGMs) (Song and Ermon, 2019) that are also known as *Denoising Diffusion Probabilistic Models* (DDPMs) (Ho et al., 2020), or simply *diffusion models*. Note that for each of these three models, there exists an extensive number of refinements but we present them here in their vanilla versions.

6.2.2.1 Variational autoencoders

First, we begin with the *Variational Autoencoders* (VAEs) that have been introduced by Kingma and Welling (2014). We refer to Kingma et al. (2019) and Doersch (2016) for two tutorials on VAEs.

Motivation. In this model, we suppose there exists a latent random variable Z in dimension d' and a deterministic mapping $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that the data $\{x_i\}_i^n$ are realizations of the random variable $g(Z)$. This basic idea builds on the manifold hypothesis: since the underlying distribution ν from which the data are sampled lies on a low dimensional manifold, one can encode the data with a variable living in much more smaller dimension than the dimension of the ambient space. Thus, d' is chosen most of the time much smaller than d, although the model theoretically doesn't assume any order relation between d' and d. In theory, as soon as d' is greater than the intrisic dimension of ν, there probably exists an infinite number of low-dimensional representations Z and mappings g such that the data could be realizations of $g(Z)$. The variational autoencoder model proposes to learn a representation in which Z follows a standard gaussian law $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$ and to approximate g by a neural network g_θ that we call *decoder*. Therefore, VAEs are *push-forward generative models* since the parametric distribution ν_θ that approaches ν is of the form $\nu_\theta = g_{\theta\#}\mu_{d'}$.

Training of VAEs: the variational approach. In order to train this model, we would like to maximize its *log-likelihood*, i.e. the following quantity

$$l(\theta, \boldsymbol{x}) = \sum_{i=1}^{n} \log(p_{\nu_\theta}(x_i)) \ ,$$

where $\boldsymbol{x} = (x_1, \dots, x_n)$ and p_{ν_θ} is the density function of ν_θ. A first theoretical problem we encounter here is that as soon as d' is smaller than d, which is most of the time the case in practice, ν_θ is a degenerate distribution - because $\nu_\theta = g_{\theta\#}\mu_{d'}$ - and so p_{ν_θ} doesn't exist. The typical remedy to this problem is to add noise in the model: instead of ν_θ, we maximize the log-likelihood of a noisy version $\nu_\theta^\sigma = \nu_\theta * \mathrm{N}(0, \sigma^2 \, \mathrm{Id}_d)$ where $*$ denotes the convolution operator between measures and the noise level $\sigma > 0$ is an hyper-parameter of the model to be chosen. In order to maximize the log-likelihood, we would like to use the latent structure of our model. To do so we could use the fact that for all $x \in \mathbb{R}^d$,

$$p_{\nu_\theta^\sigma}(x) = \int_{\mathbb{R}^{d'}} p_{\pi_\theta^\sigma}(x, z) \mathrm{d}z \ , \tag{6.2}$$

where $p_{\pi_\theta}^\sigma$ is the density of the joint law of (X^σ, Z), where $X^\sigma \sim \nu_\theta^\sigma$ and $Z \sim \mu_{d'}$. Yet this integral is in practice intractable as soon as the dimension d' is moderately high. A possible solution could be to approximate (6.2) using a Monte-Carlo method but this yields to high variance in the gradient estimates during the maximization of the log-likelihood. A better solution consists in, rather than maximizing the log-likelihood directly, constructing a tractable lower bound and to optimize this lower bound instead. This methods are known in the literature as *variational approaches*. The main idea is to introduce another parametric family of distributions $\{\xi_\lambda\}_{\lambda \in \Lambda}$ with support $\mathbb{R}^{d'}$ and density p_{ξ_λ} and to inject it in the log-likelihood. More precisely, observe that for all $x \in \mathbb{R}^d$

$$\log(p_{\nu_\theta^\sigma}(x)) = \log\left(\int_{\mathbb{R}^{d'}} p_{\pi_\theta^\sigma}(x, z) \mathrm{d}z\right) = \log\left(\int_{\mathbb{R}^{d'}} \frac{p_{\xi_\lambda}(z)}{p_{\xi_\lambda}(z)} p_{\pi_\theta^\sigma}(x, z) \mathrm{d}z\right) \ .$$

Using furthermore the Jensen-inequality, it follows

$$\log(p_{\nu_\theta^\sigma}(x)) \geq \int_{\mathbb{R}^{d'}} \log\left(\frac{p_{\pi_\theta^\sigma}(x, z)}{p_{\xi_\lambda}(z)}\right) p_{\xi_\lambda}(z) \mathrm{d}z = \mathbb{E}_{Z \sim \xi_\lambda}\left[\log\left(\frac{p_{\pi_\theta^\sigma}(x, Z)}{p_{\xi_\lambda}(Z)}\right)\right] \ .$$

This lower bound is known in the literature of variational methods as the *Evidence Lower Bound* (ELBO). With further calculation, using the fact that for all $x \in \mathbb{R}^d$ and all $z \in \mathbb{R}^{d'}$, $p_{\pi_\theta^\sigma}(x, z) = p_{\nu_\theta^\sigma}(x|z) p_{\mu_{d'}}(z)$, where $p_{\nu_\theta^\sigma}(x|z)$ is the density of the conditional law of X^σ given Z and $p_{\mu_{d'}}(z)$ is the density function of $\mu_{d'}$, we get

$$\mathrm{ELBO}(x, \theta, \lambda) = \mathbb{E}_{Z \sim \xi_\lambda}\left[\log(p_{\nu_\theta^\sigma}(x|Z))\right] - D_{\mathrm{KL}}(\xi_\lambda || \mu_{d'}) \ . \tag{6.3}$$

Observe that with our model, for a given $z \in \mathbb{R}^{d'}$, the conditional law of X given z is given by $\nu_{\theta, X|Z}(.|z) = \delta_{g_\theta(z)}$ and so the the conditional law of X^σ reads as $\nu_{\theta, X|Z}^\sigma(.|z) = \mathrm{N}(g_\theta(z), \sigma^2 \, \mathrm{Id}_d)$. Hence maximizing the left-hand term in (6.3) is equivalent to minimizing $(1/2\sigma^2)\mathbb{E}_{Z \sim \xi_\lambda}[\|x - g_\theta(Z)\|^2]$. On the other hand, it is usual for these models to choose, for a given $x \in \mathbb{R}^d$, the parametric family $\{\xi_\lambda\}_{\lambda \in \Lambda}$ as the conditional distributions of the form $\xi_{\lambda, Z|X}(.|x) = \mathrm{N}(f_{1\lambda}(x), \mathrm{diag}(\exp[f_{2\lambda}(x)]))$ where $f_\lambda(x) = (f_{1\lambda}(x), f_{2\lambda}(x)) \in \mathbb{R}^{2d'}$ is the output of another neural network that we call *encoder*. This choice makes the right-hand term of (6.3) also tractable since the Kullback-Leibler divergence between two non-degenerate Gaussian distribution $\mu_0 = \mathrm{N}(m_0, \Sigma_0)$ and $\nu_1 = \mathrm{N}(m_1, \Sigma_1)$ on $\mathbb{R}^{d'}$ admits as closed form

$$D_{\mathrm{KL}}(\mu_0 || \nu_1) = \frac{1}{2}\left[\log\frac{|\Sigma_1|}{|\Sigma_0|} - d' + \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (m_1 - m_0)^T \Sigma_1^{-1}(m_1 - m_0)\right] \ .$$

The training is then done by optimizing simultanously both variables in the following problem using stochastic gradient descent

$$\max_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{E}_{X \sim \hat{\nu}}[\mathrm{ELBO}(X, \theta, \lambda)] \ .$$

Note that it can be problematic that the expectation in the left-hand term in (6.3) depends on λ. A simple way to remedy this, known as the *reparametrization trick*, consists in writing the random variable Z as $f_{1\lambda}(x) + \mathrm{diag}^{\frac{1}{2}}(\exp[f_{2\lambda}(x)])Y$ with $Y \sim \mathrm{N}(0, \mathrm{Id}_{d'})$, and to sample from the random variable Y instead from Z. Finally, the training is summarized in Algorithm 7.

Algorithm 7 Training of VAEs

Require: data distribution $\hat{\nu}$, decoder g_θ, encoder f_λ, minibatch size m, learning rate η,

while not converged **do**
  sample minibatch $\{x_i\}_i^m$ from $\hat{\nu}$
  draw k samples $\{\varepsilon_i\}_i^m$ from $\mathrm{N}(0, \mathrm{Id}_{d'})$
  $z_i \leftarrow f_{1\lambda}(x_i) + \mathrm{diag}^{\frac{1}{2}}(\exp[f_{2\lambda}(x_i)])\varepsilon_i$ for $i = 1, \dots, m$
  $\theta \leftarrow \theta - \frac{\eta}{\sigma^2 m} \sum_{i=1}^m \nabla_\theta(\|x_i - g_\theta(z_i)\|^2)$ ▷ update the decoder
  $\lambda \leftarrow \lambda - \frac{\eta}{m} \sum_{i=1}^m \nabla_\lambda D_{\mathrm{KL}}(\xi_{\lambda,Z|X}(.|x_i)||\mathrm{N}(0, \mathrm{Id}_{d'}))$ ▷ update the encoder
end while

Variant: the Bernoulli-VAE. Note that we described above the Gaussian-VAE model from Kingma and Welling (2014) because we are interested in generating data in $\mathbb{R}^d$. There exists a variant in which the data are booleans, i.e. in $\{0, 1\}$. In that case one can set the conditional law of X given z to follow a Bernoulli law of parameter $g_\theta(z)$ instead of a Gaussian distribution.

Generation process. Given a trained VAE model, i.e a trained encoder/decoder pair (f_λ, g_θ), the generation process is straightforward: sample z from $\mathrm{N}(0, \mathrm{Id}_{d'})$ then give z as input of the decoder to obtain the generated data $g_\theta(z)$. Note that if we have introduced the noisy distribution ν_θ^σ in order to be able to write the log-likelihood, we are in reality interested in sampling from ν_θ and so we don't need ν_θ^σ anymore.

Performances of VAEs. VAEs are known to have some nice properties such as stable training (Tolstikhin et al., 2018) or robustness to outlier data (Dai et al., 2018). Still, in term of quality of results, VAEs are not able to reach the same kind performances than the two other models presented below. A possible explanation proposed by (Dai and Wipf, 2018) is that the ELBO can be maximized in a way such that ν_θ approximates well the target distribution ν but in the mean time such that the aggregated posterior (Makhzani et al., 2015), i.e. the distribution in the latent space defined as

$$\xi_{\lambda,\theta} = \int_{\mathbb{R}^d} \xi_{\lambda,Z|X}(.|x)\mathrm{d}\nu_\theta(x) \,, \tag{6.4}$$

is quite far from $\mathrm{N}(0, \mathrm{Id}_{d'})$. This implies that even if we found a good low dimensional representation of our data, we are still not able to correctly generate new synthetic data because we are not able to correctly sample from this low dimensional representation.

6.2.2.2 Generative Adversarial networks

We present here the vanilla model of *Generative Adversarial Networks* (GANs) that have been introduced by Goodfellow et al. (2014). We refer to Creswell et al. (2018) and Gui et al. (2021) as two review papers on the topic.

Motivation. As for the VAE model, we suppose there exists a latent random variable Z in dimension d' and a deterministic mapping $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that the data $\{x_i\}_i^n$ are realizations of the random variable $g(Z)$. As for the VAE, we aim at finding a low dimensional representation of the data such that Z follows a standard Gaussian law $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$ and we approximate g by a neural network g_θ that is called this time *generator*. Hence generative adversarial networks are also *push-forward generative models* since the parametric distribution ν_θ that approaches ν is also of the form $\nu_\theta = g_{\theta\#}\mu_{d'}$.

Training of GANs: the adversarial approach. The key idea of the GAN model is to train the generator g_θ using an adversarial scheme. For that, we define another neural network $f_\lambda : \mathbb{R}^d \to [0, 1]$ that we call *discriminator* and whose role is to classify if a given input data has been generated by the generator or comes from the dataset. We train both networks g_θ and f_λ simultanously such that the discriminator f_λ becomes gradually better at detecting whether a given data is synthetic or real, and the generator g_θ becomes gradually better at fooling the discriminator. Mathematically, this generator/discriminator dynamics can be created by performing an alternate optimization scheme on the following min-max problem:

$$\min_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathbb{E}_{X\sim\hat{\nu}}[\log(f_\lambda(X))] + \mathbb{E}_{Z\sim\mu_{d'}}[\log(1 - f_\lambda(g_\theta(Z)))] \,. \tag{6.5}$$

The training is done by an alternating minimization on θ and a maximization on λ using respectively stochastic gradient descent and ascent. This can be thought as a two-player min-max game between the generator g_θ and the discriminator f_λ. Thus, following the game theory terminology, we want the training to converge to a local *Nash equilibrum* (Osborne and Rubinstein, 1994) which can be thought as a particular type of *saddle points* where no player can improve its objective given the current state of the other player. Yet, it has been observed that the training of GANs, as initially proposed in Goodfellow et al. (2014), may fail to converge towards a Nash equilibrum (Goodfellow et al., 2014; Goodfellow, 2016; Salimans et al., 2016). Heusel et al. (2017) have suggested that the problem might come from the fact that the generator and the discriminator were optimized with the same learning rate. Indeed, they have proven that, under mild conditions, the training of GANs with the Adam optimizer (Kingma and Ba, 2015) and with different learning rates for the generator and the discriminator converges to a local stationary Nash equilibrum. Interestingly enough, these conditions include a smoothness assumption on the networks g_θ and f_λ that ReLU networks, which are still today the most used in practice, don't verify. Finally the training of GANs is summarized in Algorithm 8.

Algorithm 8 Training of GANs

Require: data distribution $\hat{\nu}$, generator g_θ, discriminator f_λ, minibatch size m, learning rates η_g and η_f, number of update of dicriminator N_f at each iterate

while not converged **do**
  for $j = 1, \dots, N_f$ **do**
    sample minibatch $\{x_i\}_i^m$ from $\hat{\nu}$
    draw m noise samples $\{z_i\}_i^m$ from $\mathrm{N}(0, \mathrm{Id}_{d'})$
    $\lambda \leftarrow \lambda + \frac{\eta_f}{m} \sum_{i=1}^m \nabla_\lambda \left[\log(f_\lambda(x_i)) + \log(1 - f_\lambda(g_\theta(z_i)))\right]$ ▷ update the discriminator
  end for
  draw m noise samples $\{z_i\}_i^m$ from $\mathrm{N}(0, \mathrm{Id}_{d'})$
  $\theta \leftarrow \theta - \frac{\eta_g}{m} \sum_{i=1}^m \nabla_\theta \log(1 - f_\lambda(g_\theta(z_i)))$ ▷ update the generator
end while

Optimal discriminator analysis. A possible theoretical analysis of GANs initially proposed by Goodfellow et al. (2014) is to study the model while assuming that at each update of the generator, we choose the best discriminator possible. Goodfellow et al. (2014) have shown that if we assume that both distributions ν and ν_θ admit densities p_ν and p_{ν_θ} (which is not the case in the most usual pratical setting), the best discriminator possible for a given fixed g_θ is,

$$f_\lambda(x) = \frac{p_\nu(x)}{p_\nu(x) + p_{\nu_\theta}(x)} \, .$$

Plugging this into the min-max objective, this gives that Problem (6.5) is in that case equivalent to minimizing the Jensen-Shannon divergence $D_{\mathrm{JS}}(\nu||\nu_\theta)$ between ν and ν_θ. Building on this analysis, a large number of variants of (6.5) have been proposed in order to replace the Jensen-Shannon divergence by another divergence or distance between probability distributions. Among them, one can cite notably the *Wasserstein GAN* (WGAN) (Arjovsky et al., 2017) model that proposes to learn a real-valued discriminator $f_\lambda : \mathbb{R}^d \to \mathbb{R}$, renamed *critic*, and to solve the following min-max problem,

$$\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \mathbb{E}_{X \sim \hat{\nu}}[f_\lambda(X)] - \mathbb{E}_{Z \sim \mu_{d'}}[f_\lambda(g_\theta(Z))] \, , \tag{6.6}$$

while enforcing the Lipschitz constant $\mathrm{Lip}(f_\lambda)$ of the critic to be smaller than 1. The idea is, building on the same optimal discriminator analysis than before, that (6.6) is in that case equivalent to solving the dual W_1 problem (2.10) as described in Chapter 2 and so learning a WGAN while choosing the optimal discrimator at each iteration is equivalent to minimizing the W_1 distance between ν and ν_θ. If this analysis seems to reveal an interesting theoretical property of GANs by showing that they would implicitly minimize an underlying divergence or distance and has led to the development of numerous variants of the initial model, it might not be as relevant as it seems. Firstly, the assumption of choosing the optimal discriminator is never verified in practice. Furthermore, the large scale study of Lucic et al. (2018) suggests that the difference of performances between the different models built over this analysis could in fact be more explained by the choice of the different hyperparameters rather than by the choice of the objective function. More critically, in the WGAN setting, the studies of Mallasto et al. (2019) and

Stanczuk et al. (2021) suggest that a better approximation of the optimal discriminator would in reality lead to worse perfomances in terms of generation, even if the distance $W_1(\nu, \nu_\theta)$ is theoretically better minimized. A possible explanation could be that it comes from the fact that there is an infinite number of distributions ν from which the data could have been sampled, including the empirical data distribution $\hat{\nu}$. Since the model has only access to the empirical data distribution $\hat{\nu}$ and to an empirical version $\hat{\nu}_\theta$ of the distribution ν_θ, it is possible that a better approximation of the W_1 distance leads to a distribution ν_θ that is close to the minimizer of $W_1(\hat{\nu}, \hat{\nu}_\theta)$, which corresponds in reality to a "k-medians" clustering (Stanczuk et al., 2021) of $\hat{\nu}$. Yet, interestingly, Korotin et al. (2022) have shown that even if WGANs don't approximate well the W_1 distance between ν_θ and ν (when this latter is known), the gradients computed during training can however be used as good estimators of the W_1 gradients, suggesting that the training of these models stays linked with a minimization of a W_1 distance. In short, this analysis reveals an interesting connection between the GAN training and the minimization of an underlying distance (or divergence) but is far from sufficient to explain the GAN training, as the dynamics between the generator and the discriminator is probably at least as important as this underlying distance and because the link with this latter is less direct than it first seems.

Stability of training. Until recently, GANs were considered as the state-of-the-art models for image generation. Still, a common pitfall of such models is the instability of their training (Salimans et al., 2016). This can lead to models that forget significant parts of the support of the target distribution, which is known as *mode collapsing* (Arjovsky and Bottou, 2017; Metz et al., 2017) in the GAN literature. Thus, the success of GANs is largely due to advances in stabilizing their training. First, enforcing the Lipschitz constant of the critic to be smaller than 1 in Wasserstein GANs has been proved, as a by product of the method, to stabilize their training. The same idea has been applied with sucess to other GAN models. Typical methods to constraint the Lipschitz constant of the discriminator are to penalize the Jacobian of the discriminator Gulrajani et al. (2017) or to apply spectral normalization Miyato et al. (2018). It has been then shown by Odena et al. (2018) that constraining the Lipschitz constant of the generator was also important to stabilize the training of GANs. To that extent, state-of-the-art models such as Zhang et al. (2019) or Brock et al. (2019) apply spectral normalization both on generator and discriminator.

The StyleGAN variant. Karras et al. (2019) have introduced an important variant of the model described above by redesigning the generator architecture using ideas coming from the style transfer literature. This has led to the model StyleGAN 3 (Karras et al., 2021) which is the state-of-the-art in generation on the CelebA dataset (Liu et al., 2015). Besides some technical changes, the main structural difference with a classic generator architecture lies in the addition of Gaussian noise after each convolution. Thus, the StyleGAN model can still be thought as a push-forward generative model, but with a latent variable that consists in the concatenation of all the noises added in the generator and with a push-forward map g_θ with a particular structure in which the information of the latent space is progressively added.

6.2.2.3 Diffusion models

We present here the *diffusion models*, also known as *Score-based Generative Models* (SGMs) that have been introduced by Song and Ermon (2019). Note that they also have been introduced almost at the same time by Ho et al. (2020) under the name of *Denoising Diffusion Probabilistic Models* (DDPMs) with a different point of view. The equivalence between the two points of view has been finally established by Song et al. (2020). We refer to this latter paper for an introduction on SGMs.

Motivation: the score-matching point of view. One popular Monte Carlo method for sampling from a probability distribution ν on $\mathbb{R}^d$ with probability density p_ν builds on the *Langevin dynamics*, i.e. the following Stochastic Differential Equation (SDE) which describes an Itô diffusion,

$$\mathrm{d}X_t = \nabla \log p_\nu(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \ X_0 \sim \mathrm{N}(0, \mathrm{Id}_d) \ ,$$

where $(B_t)_{t\in[0,T]}$ is a Brownian motion and $T > 0$. An Euler-Maruyama discretization (Kloeden et al., 1992) of the process $(X_t)_{t\in[0,T]}$ that solves this SDE yields to the following Monte Carlo method, known as the *Unajusted Langevin Algorithm* (ULA),

$$x^{\{k\}} = x^{\{k-1\}} + \alpha_k \nabla_x \log p_\nu(x^{\{k-1\}}) + \sqrt{2\alpha_k} z^{\{k\}} \ , \ z^{\{k\}} \sim \mathrm{N}(0, \mathrm{Id}_d) \ ,$$

for any $1 \leq k \leq K$, with $x^{\{0\}} \sim \mathrm{N}(0, \mathrm{Id}_d)$, and where $\{\alpha_k\}_k^K$ is a non-increasing sequence. Note that this algorithm introduces an error that needs to be corrected using an additional Metropolis-Hastings update

at each step, which yields to the *Metropolis-Adjusted Langevin Algorithm* (MALA) (Besag, 1994). Yet, it has been observed that this error could be often ignored in practice (Chen et al., 2014; Du and Mordatch, 2019; Nijkamp et al., 2020). The term $\nabla \log p_\nu$ is often referred to as the *Stein score*, or simply the score of the probability distribution ν. Thus, if we have access to a good estimator of the score $\nabla \log p_\nu$, we are theoretically able to sample from ν. The basic idea of SGMs is to approximate the Stein score by a neural network. As before, a first theoretical problem that comes up, when generating real data such as images, is that the underlying target distribution has probably no density, because of the manifold hypothesis. The typical remedy is once again to inject noise in the model by estimating the score of a noisy version ν^σ of ν, i.e. $\nu^\sigma = \nu * \mathrm{N}(0, \sigma^2 \mathrm{Id}_d)$. Note that we will never be able to correctly approximate the score $\nabla_x \log p_\sigma$ of ν^σ in the regions where the density p_σ is too small because the model will see almost no data in these areas. Hence, when σ is chosen too small, the algorithm is likely to be initialized in an area of low density and never reach a region where the score is well approximated. On the other hand, when σ is chosen too large, ν^σ doesn't approximate well anymore ν. A natural solution to this problem proposed by Song and Ermon (2019) is to rely on an annealed scheme:

(i) choose an increasing sequence of K noise levels $\{\sigma_k\}_k^K$ such that σ_K is large enough so that ν^{σ_K} is close to a standard Gaussian distribution $\mathrm{N}(0, \mathrm{Id}_d)$ - which implicitly supposes that the data are normalized in a way that their empirical variance is smaller than 1 - and σ_1 is small enough such that ν^{σ_1} is close to ν.

(ii) train a neural network $s_\theta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ on $\{x_i\}_i^n \times \{\sigma_k\}_k^K$ at approximating $\nabla_x \log p_\sigma$ for any noise level between σ_K and σ_1 .

(iii) perform K iterations of the so-called *annealed Langevin dynamics* (Song and Ermon, 2019):

$$x^{\{k\}} = x^{\{k-1\}} + \alpha_k s_\theta(x^{\{k-1\}}, \sigma_{K-k}) + \sqrt{2\alpha_k} z^{\{k\}} \ , \ z^{\{k\}} \sim \mathrm{N}(0, \mathrm{Id}_d) \ ,$$

for any $1 \leq k \leq K$, with $x^{\{0\}} \sim \mathrm{N}(0, \mathrm{Id}_d)$, and where $\{\alpha_k\}_k^K$ is a non-increasing sequence. Note that in practice, the sequence $\{\sigma_k\}_k^K$ is most of the time chosen piecewise constant, in the sense that there exists an integer L such that $\sigma_1 = \cdots = \sigma_L$, then $\sigma_{L+1} = \cdots = \sigma_{2L}$, and so on until reaching K. The idea is that during the early steps of the dynamics, the noise level is large enough so that $\nabla_x \log p_\sigma$ is well approximated everywhere and that we reach the high-density regions before σ decreases to much.

The variational approach. Here we present an approach initially introduced by Sohl-Dickstein et al. (2015) and applied to generative modeling in Ho et al. (2020) under the name of *Denoising Diffusion Probabilistic Models* (DDPMs). The basic idea of DDPMs is to construct a Markov chain $(Y_k)_{k \in [\![0,K]\!]}$ such that $Y_0 \sim \hat{\nu}$ and for any $1 \leq k \leq K$, the conditional law of Y_k given $Y_{k-1} = y_{k-1}$ is $\mathrm{N}(y_{k-1} + \alpha_k f(y_{k-1}), 2\alpha_k \mathrm{Id}_d)$, where $f : \mathbb{R}^d \to \mathbb{R}^d$ and $\{\alpha_k\}_k^K$ is a non-increasing sequence of step sizes chosen such that the law of Y_K is close of a standard Gaussian distribution $\mathrm{N}(0, \mathrm{Id}_d)$ - which also supposes implicitly that the second order moment of $\hat{\nu}$ is smaller than 1. This Markov chain consists in progressively applying noise to the data until having nothing but noise. A generative model can be then obtained by approaching the reverse-time Markov chain associated with $(Y_k)_{k \in [\![0,K]\!]}$. To do so, we define a parametric Markov chain $(X_{\theta,k})_{k \in [\![0,K]\!]}$ such that $X_{\theta,K} \sim \mathrm{N}(0, \mathrm{Id}_d)$ and for any $1 \leq k \leq K$, the conditional law of $X_{\theta(k-1)}$ given $X_{\theta,k} = x_k$ is $\mathrm{N}(g_\theta(x_k, k), 2\alpha_k \mathrm{Id}_d)$ where $g_\theta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is a neural network. We then train our model using a variational approach, i.e. by optimizing an evidence lower bound of the log-likelihood of the model. More precisely, denoting p_{ν_θ} the density of the law of $X_{\theta,0}$, i.e. the parametric distribution ν_θ that approaches the underlying data distribution ν, p_{π_θ} the density of the joint law of the parametric markov chain $(X_{\theta,0}, \ldots, X_{\theta,K})$ and $\xi_{1:T|Y_0}(.|y_0)$ the law of $(Y_1, \ldots, Y_K)$ given the data $Y_0 = y_0$, we have for any $x \in \mathbb{R}^d$,

$$\log(p_{\nu_\theta}(x)) \geq \mathbb{E}_{(Y_1,\ldots,Y_K) \sim \xi_{1:T|Y_0}(.|x)} \left[\frac{p_{\pi_\theta}(x, Y_1, \ldots, Y_K)}{\xi_{1:T|Y_0}(Y_1, \ldots, Y_K|x)} \right] = \mathrm{ELBO}(x, \theta) \ .$$

Once the model is trained, the generation process consists simply in deriving a realization of the Markov chain $(X_{\theta^*,k})_k^K$, i.e. first sampling $x_K \sim \mathrm{N}(0, \mathrm{Id}_d)$, then sampling $x_{K-1} \sim \mathrm{N}(g_{\theta^*}(x_K, K), 2\alpha_k \mathrm{Id}_d)$ and pursuing until obtaining a synthetic data.

Unifying the two approaches: the SDE point of view. An observation made by Song et al. (2020) is that the forward Markov chain $(Y_k)_{k\in[\![0,K]\!]}$ in DDPMs, whose transition rules read as for any $1 \leq k \leq K$,

$$y_k = y_{k-1} + \alpha_k f(y_{k-1}) + \sqrt{2\alpha_k} z_k \ , \ z_k \sim \mathrm{N}(0, \mathrm{Id}_d) \ ,$$

can be thought as an Euler-Maruyama discretization of the process $(Y_t)_{t\in[0,T]}$ that solves the following SDE, which describes an Itô diffusion,

$$\mathrm{d}Y_t = f(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \ Y_0 \sim \hat{\nu} \ . \tag{6.7}$$

Yet, it is well-known since (Anderson, 1982) that, under mild conditions, any process $(Y_t)_{t\in[0,T]}$ that solves an SDE of the form

$$\mathrm{d}Y_t = h(Y_t, t)\mathrm{d}t + r(t)\mathrm{d}B_t \ , Y_0 \sim \nu_0 \ ,$$

where $\nu_0 \in \mathcal{P}(\mathbb{R}^d)$ and with $h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ and $r : \mathbb{R} \to \mathbb{R}$ regular enough, admits a reverse-time process $(X_t)_{t\in[0,T]}$ that solves the following diffusion SDE,

$$\mathrm{d}X_t = [-h(X_t, T-t) + r(T-t)^2 \nabla_x \log_{T-t}(X_t)]\mathrm{d}t + r(T-t)\mathrm{d}B_t \ , \ X_0 \sim \nu_T \ ,$$

where $\nu_T \in \mathcal{P}(\mathbb{R}^d)$ is the law of Y_T and for all $t \in [0,T]$, p_t is the density of the law of Y_t. Hence, the reverse time process $(X_t)_{t\in[0,T]}$ associated with the process $(Y_t)_{t\in[0,T]}$ that solves (6.7) is solution of the following SDE,

$$\mathrm{d}X_t = [-f(X_t) + 2\nabla_x \log p_{T-t}(X_t)]\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \ X_0 \sim \mathrm{N}(0, \mathrm{Id}_d) \ , \tag{6.8}$$

where we made the approximation that the law of Y_T was exactly $\mathrm{N}(0, \mathrm{Id}_d)$. Informally, this implies that in DDPMs, the parametric Markov chain $(X_{\theta k})_{k\in[\![0,K]\!]}$ approximates in reality a discretization of $(X_{T-t})_{t\in[0,T]}$, where $(X_t)_{t\in[0,T]}$ solves (6.8). Hence the neural network g_θ implicitly learns the score of the law of X_{T-t}. On the other hand, the authors of Song et al. (2020) show that the annealed Langevin dynamics, when $\{\sigma_k\}_k^K$ is chosen as a geometric progression and for an adequate choice of step sizes $\{\alpha_k\}_k^K$, can be thought as an Euler-Maruyama discretization of the process $(X_t)_{t\in[0,T]}$ that solves

$$\mathrm{d}X_t = e^{2(T-t)}\nabla_x \log p_{T-t}(X_t)\mathrm{d}t + e^{T-t}\mathrm{d}B_t \ , \ X_0 \sim \mathrm{N}(0, \mathrm{Id}_d) \ ,$$

which is the reverse-time SDE associated with

$$\mathrm{d}Y_t = e^t \mathrm{d}B_t \ , \ Y_0 \sim \hat{\nu} \ .$$

Thus, the two approaches of DDPMs (Ho et al., 2020) and SGMs (Song and Ermon, 2019) are in fact two different point of views of the same model and they only differ in the choice of the SDE, which can be thought as a component of the model. Finally, note that practitioners usually learn diffusion models with the score-matching approach but use most of the time the SDE (6.7) induced by the DDPM model from (Ho et al., 2020).

Training of SGMs: denoising score matching. The training of SGMs builds on the works of Hyvärinen (2005) and Vincent (2011) which draw connections between the task of estimating the score and the denoising task. Indeed, Vincent (2011) has proven that for a given distribution ν on $\mathbb{R}^d$ and a given noise level σ, approximating the score $\nabla_y \log p_\sigma$ of $\nu^\sigma = \nu * \mathrm{N}(0, \sigma^2 \mathrm{Id}_d)$ with a function $s_\theta(y, \sigma)$ could be done by solving the following problem,

$$\min_{\theta\in\Theta} \tfrac{1}{2}\mathbb{E}_{X\sim\hat{\nu}}\left[\mathbb{E}_{Y\sim\mathrm{N}(x,\sigma^2\mathrm{Id}_d)}\left[\left\|s_\theta(Y,\sigma) - \tfrac{x-Y}{\sigma^2}\right\|^2 \middle| X = x\right]\right] .$$

Informally, this can be fastly retrieven using the *Tweedie formula* (Robbins, 1955), that states that for any $y \in \mathbb{R}^d$,

$$\nabla_y \log p_\sigma(y) = \frac{1}{\sigma^2}(\hat{x}(y) - y) \ ,$$

where $\hat{x}(y)$ is the Minimum Mean Square Error (MMSE) estimator of the denoised version of y, i.e.

$$\hat{x}(y) = \mathbb{E}[X|Y = y] = (2\pi\sigma^2)^{-\frac{d}{2}} \int_{\mathbb{R}^d} x \exp[-\|y - x\|^2/2\sigma^2]\mathrm{d}\nu(x) \ .$$

Hence, given a dataset $\{x_i\}_i^n$, one can construct a noisy dataset $\{y_i\}_i^n$ by sampling from $\hat{\nu} * \mathrm{N}(0, \sigma^2 \mathrm{Id}_d)$ and, if we suppose that for any i, x_i is the perfectly denoised version of y_i and so $\hat{x}(y_i) \simeq x_i$, we get that

$$\nabla_y \log p_\sigma(y_i) \simeq \frac{1}{\sigma^2}(x_i - y_i) \, .$$

Finally, the objective is averaged over the empirical noise level distribution $\frac{1}{K}\sum_{k=1}^K \delta_{\sigma_k}$. The training of SGMs is summarized in Algorithm 9.

Algorithm 9 Training of SGMs

Require: data distribution $\hat{\nu}$, noise level distribution $\hat{\zeta}$, minibatch size m, learning rate η
1: **while** not converged **do**
2: sample minibatches $\{x_i\}_i^m$ and $\{\sigma_i\}_i^m$ from $\hat{\nu}$ and $\hat{\zeta}$.
3: draw sample $y_i \sim \mathrm{N}(x_i, \sigma_i^2 \mathrm{Id}_d)$ for $i = 1, \dots, m$
4: $\theta \leftarrow \theta - \frac{\eta}{2m}\sum_i^m \nabla_\theta(\|s_\theta(x_i, \sigma_i) - \frac{x_i - y_i}{\sigma_i^2}\|^2)$
5: **end while**

Score network architecture. One question that arises is: *what kind of network architecture should we use for s_θ?* In particular, can we use any deep neural network that has been pretrained for denoising? If this should work in theory, it is not the case in practice. A possible explanation is that the generation task requires to approximate the score $\nabla_x \log p_\sigma$ for any noise level σ much more accurately than the denoising task. Hence, s_θ should have a much more complex architecture than the ones that are sufficient for the denoising task. Another question is: how to integrate the noise information in the network? Initially, Song and Ermon (2019) have proposed to integrate the noise information relying on *conditional instance normalization* (Dumoulin et al., 2016). Alternatively, Ho et al. (2020) have proposed an architecture composed of two parallel networks, taking respectively x and σ as inputs. The network taking σ as input consists in the composition of a positional encoding (Vaswani et al., 2017) and of a simple feed-forward network. Then, its output is injected at each layer of the other network by simple concatenation. It seems that it is this latter architecture that has been adopted by the community. Finally, note that the architecture of Ho et al. (2020) uses a *U-Net* structure as backbone (Ronneberger et al., 2015), which is a classic neural network architecture in computer vision for neural networks that take values in $\mathbb{R}^d$. It has been observed by Jolicoeur-Martineau et al. (2020) that this U-Net architecture substantially improved the sample quality compared to the previous architecture (Lin et al., 2017; Song and Ermon, 2020) used for denoising score matching.

Computational cost of diffusion models. SGMs are nowadays the state-of-the-art in generative modeling. Indeed, they have been shown to outperform GANs on image synthesis (Dhariwal and Nichol, 2021). Futhermore, SGMs are also responsible of the general public's recent craze for generative models, with the arrival of large-scale text-to-image conditional models such as DALL-E 2 (Ramesh et al., 2022) or Stable Diffusion (Rombach et al., 2022). Beside their impressive results, their training is known to be relatively stable - at least compared to GANs - since it roughly consists in simply training a neural network at denoising for various noise levels. The main downfall of diffusion models is however related to their computational cost, which is heavy not only during the training phase, as with other models, but also during the generation phase because the neural network is used sometimes thousand of times in the generation dynamics and because there is no reduction of dimensions. There exists an extensive literature focusing on speeding the generation dynamics of SGMs (Nichol and Dhariwal, 2021; Watson et al., 2021; San-Roman et al., 2021; Jolicoeur-Martineau et al., 2021; Luhman and Luhman, 2021; De Bortoli et al., 2021). Other methods (Vahdat et al., 2021; Rombach et al., 2022) propose to learn a diffusion model in the latent space of a VAE to combine the benefits of SGMs with the benefits of learning a low dimensional representation of the data.

SGMs as indirect push-forward generative models. As we have seen above, the generation dynamics G_θ in SGMs can be expressed as a SDE and so it takes theorically the form of a function on the space of trajectories of a Brownian motion $(B_t)_{t\in[0,T]}$. Still, the generation is done in practice by achieving an Euler-Maruyama discretization of the SDE and sampling from the obtained discrete Markov Chain. For instance, the generation can be done by performing K iterations of the annealed Langevin

dynamics, which are of the form

$$x^{\{k\}} = x^{\{k-1\}} + \alpha_k s_\theta(x^{\{k-1\}}, \sigma_{K-k}) + \sqrt{2\alpha_k} z^{\{k\}} \ ,$$

where for any $k \geq 0$, $z^{\{k\}} \sim \mathrm{N}(0, \mathrm{Id}_d)$. Denoting for any k, $h_\theta^k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ the map such that for any x and z in $\mathbb{R}^d$,

$$h_\theta^k(x, z) = x + \alpha_k s_\theta(x, \sigma_{K-k}) + \sqrt{2\alpha_k} z \ ,$$

we get that, in practice, the generation dynamics G_θ is a deterministic mapping from $\mathbb{R}^{d(K+1)}$ to $\mathbb{R}$ which can be written as, for any $\boldsymbol{z} = (z_0, \dots, z_K) \in \mathbb{R}^{d(K+1)}$

$$G_\theta(\boldsymbol{z}) = h_\theta^K \left(h_\theta^{K-1} \left(\dots (h_\theta^2 \left(h_\theta^1(z_0, z_1), z_2 \right), \dots), z_{K-1} \right), z_K \right) \ .$$

Hence, the generated distribution ν_θ in SGMs is of the form $\nu_\theta = G_{\theta\#}\mu_{d(K+1)}$. Still, an important difference with push-foward generative models such as VAEs or GANs is that the optimization is not directly performed on G_θ but on an auxiliary function (the score). For this reason, we refer to them as *indirect push-forward generative models*.

6.2.3 Other common models

We shortly present here three other generative models that are also used in imaging science but a bit less commonly than the three previous models. These models are the *Normalizing Flows* (NFs) (Rezende and Mohamed, 2015), the *Autoregressive Models* (ARMs) (Bengio et al., 2000), and the *Energy-Based Models* (EBMs) (LeCun et al., 2006).

6.2.3.1 Normalizing flows

Normalizing Flows (NFs), in their vanilla forms introduced by Rezende and Mohamed (2015), are *push-forward generative models* that approach the underlying data distribution ν by a parametric distribution $\nu_\theta = g_{\theta\#}\mu_d$ where $\mu_d = \mathrm{N}(0, \mathrm{Id}_d)$ is the standard Gaussian distribution in dimension d and where $g_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible mapping. As an outcome of the invertibility of g_θ, the distribution ν_θ admits necessarily a density p_{ν_θ}. The NFs builds then on the *change-of-variable* formula between densities, i.e. for any $x \in \mathbb{R}^d$,

$$p_{\nu_\theta}(x) = p_{\mu_d}(g_\theta^{-1}(x)) |\det(J[g_\theta^{-1}](x))| \ ,$$

where $J[g_\theta^{-1}](x)$ is the Jacobian matrix of g_θ^{-1} in x, i.e. for any $1 \leq i, j \leq d$, $[J[g_\theta^{-1}](x)]_{i,j} = \frac{\partial g_{\theta_i}^{-1}}{\partial x_j}(x)$ Supposing that g_θ is of the form

$$g_\theta = g_\theta^L \circ \dots \circ g_\theta^1 \ ,$$

with for any $1 \leq l \leq L$, $g_\theta^l : \mathbb{R}^d \to \mathbb{R}^d$ being a non-linear invertible map, the density p_{ν_θ} can be rewritten for any $x \in \mathbb{R}^d$ as,

$$p_{\nu_\theta}(x) = p_{\mu_d}(g_\theta^{-1}(x)) \prod_{l=1}^{L} |\det(J[{g^l}_\theta^{-1}](x^{(l)}))| \ .$$

If the g_θ^l are designed such that their Jacobian is tractable, for instance such that the Jacobian matrix is triangular, the model can be simply trained by maximizing the log-likelihood $\sum_{i=1}^n \log(p_{\nu_\theta}(x))$ on the data. Note that, from a practical point of view, the conditions of being invertible and having a tractable Jacobian prevent the use of classical neural network architectures. From a theoretical point of view, it has been shown that the invertibility constraint limits the expressivity of the model (Cornish et al., 2020). One possible remedy to this problem is to inject stochasticity in the model (Cornish et al., 2020; Wu et al., 2020). Finally, we refer to Kobyzev et al. (2020) for a complete introduction on the topic of NFs.

6.2.3.2 Autoregressive models

Autoregressive Models (ARMs) (Bengio et al., 2000) propose to generate images *pixel per pixel*. More precisely, they learn a parametric distribution ν_θ with density p_{ν_θ} of the form, for a given $x = (x^1, \dots, x^d)$ in $\mathbb{R}^d$,

$$p_{\nu_\theta}(x) = \prod_{k=1}^{d^2} p_\theta(x^k | x^1, \dots, x^{k-1}) \ ,$$

where the conditional densities $p_\theta(.|x^1,\dots,x^{k-1})$ are typically Gaussian distributions, each parametrized by a different neural network. The model is then learned by simply maximizing the log-likelihood. In practice, Autoregressive models in imaging science use one single network with masked convolutions to predict the parameters of the different conditional distributions. The most popular autoregressive model in imaging science is PixelCNN (Van Den Oord et al., 2016). Note that, if ARMs are not so commonly used in imaging science compared to VAEs, GANs, or SGMs, the autoregressive structure is however particularly adapted for temporal data, and so ARMs have been very popular in audio synthesis until the introduction of diffusion models. Finally, observe that when the conditional densities are chosen as Gaussian distributions (which is the natural choice when generating real-valued data), ARMs are also *push-forward generative models* since the generation procedure can be rewritten of the form $g_\theta(Z)$ with $Z \sim \mathrm{N}(0, \mathrm{Id}_d)$ and with $g_\theta : \mathbb{R}^d \to \mathbb{R}^d$ being a deterministic mapping whose only specificity is to have a particular autoregressive structure.

6.2.3.3 Energy-based models

Energy-Based Models (EBMs) (LeCun et al., 2006) propose to approach the underlying data distribution ν by a parametric Boltzman distribution ν_θ, i.e. a distribution with density p_{ν_θ} of the form for any $x \in \mathbb{R}^d$

$$p_{\nu_\theta}(x) = \frac{\exp[-f_\theta(x)]}{Z(\theta)} \ ,$$

where $f_\theta : \mathbb{R}^d \to \mathbb{R}$ is a neural network and $Z(\theta) = \int_{\mathbb{R}^d} \exp[-f_\theta(x]\mathrm{d}x$ is called *partition function* and can be thought as a normalization factor to ensure that ν_θ is a probability distribution. An important difficulty for maximizing the log-likelihood of this model is that $Z(\theta)$ is intractable. The typical solution of this problem is to use a contrastive divergence algorithm (Hinton, 2002) which roughly consists in estimating $Z(\theta)$ with a MCMC method. Once the model is trained, the generation is done by MCMC sampling from ν_θ. Observe that Langevin methods such as ULA or MALA are especially adapted to the generation dynamics of EBMs since we have $\nabla_x \log p_{\nu_\theta} = -f_\theta(x)$. Thus, since in practice the generation dynamics is always composed of a finite number of steps, EBMs can be classified, as SGMs, as *indirect push-forward generative models*.

6.2.4 Two stage models

A recent trend in deep generative modeling consists in combining the different models presented above, or more precisely to learn a second generative model in the latent space of a VAE (Van Den Oord et al., 2017; Razavi et al., 2019; Ghosh et al., 2019; Vahdat et al., 2021; Rombach et al., 2022). These types of model are often referred to as *two stage models*. This is mainly motivated by the analysis that the bad performances of VAEs in image synthesis probably come from the fact that the aggregated posterior $\xi_{\lambda,\theta}$, as defined in (6.4), is quite far from a standard Gaussian distribution $\mathrm{N}(0, \mathrm{Id}_{d'})$ at the end of the training. Thus, a low dimensional representation of the data is first learned with the VAE and the second model then learns how to sample from this low dimensional representation. An alternative possibility consists in learning the models simultaneously (Kingma et al., 2016; Chen et al., 2016; Sønderby et al., 2016; Maaløe et al., 2019; Vahdat and Kautz, 2020; Child, 2020) although it is not clear in practice if learning the models simultaneously increases the perfomances compared to learning them successively.

6.2.5 Evaluating the models

An important topic in generative modeling is the question of evaluating the models. Indeed, evaluating generative models such as VAEs, GANs or SGMs is not a trivial problem since these models do not predict likelihood values. If the results of a generative model can be qualitatively evaluated, it is difficult to verify whether a given model correctly approachs the target distribution ν or only produces new synthetic data that could have been sampled from ν but without recovering its underlying structure. Hence a generative model should be evaluated on the two following criterions: (i) the *fidelity* of the results, i.e. the ability of the model to generate images perceptually similar to the data, (ii) the *diversity* of the results, i.e. the ability of the model to capture all the diversity of the dataset and so to cover integraly the support of the underlying distribution ν. Several quantitative measures of performances have been designed in order to assess generative models. The two most popular among them are the *Inception Score* (IS) (Salimans et al., 2016) and the *Frechet Inception Distance* (FID) (Heusel et al., 2017) which are both used specifically to evaluate models trained on Imagenet (Russakovsky et al., 2015). These two measures of performances

are both using the features of InceptionV3 classifier (Szegedy et al., 2016) that has been pre-trained on Imagenet. More precisely, the Inception Score (IS) reads as, given a generated distribution ν_θ

$$IS(\nu_\theta) = \exp\left[\int_{\mathbb{R}^d} D_{\mathrm{KL}}(\omega_{Y|X=x}||\omega_\theta)\mathrm{d}\nu_\theta(x)\right] ,$$

where $\omega_{Y|X=x}$ is the conditional distribution of the labels Y predicted by the InceptionV3 classifier given an input image x and $\omega_\theta = \int_{\mathbb{R}^d} \omega_{Y|X=x}\mathrm{d}\nu_\theta(x)$ is the marginal distribution of the labels. The idea is that, on one hand, if ν_θ covers correctly all the support of the underlying distribution of Imagenet, the generative model should output a diverse set of images from all the different classes and so ω_θ should be uniform. On the other hand, if the model is able to predict correct samples, they should be classified by the InceptionV3 network without too much incertitude and so $\omega_{Y|X=x}$ should be close to a one-hot vector and have low entropy. Thus the Kullback-Leibler divergence, and so the inception score, is maximized when the generative model satisfies both conditions of fidelity and diversity. Alternatively, the Frechet Inception Distance (FID) reads as, given a empirical version $\hat{\nu}_\theta$ of ν_θ

$$FID(\hat{\nu}_\theta, \hat{\nu}) = W_2(\mu_{\nu_\theta}, \mu_{\hat{\nu}}) ,$$

where $\mu_{\hat{\nu}_\theta}$ and $\mu_{\hat{\nu}}$ are two Gaussian distributions that have been fitted on the input responses of an intermediary layer of the InceptionV3 classifier when the input are respectively samples of $\hat{\nu}_\theta$ and $\hat{\nu}$. While these two measures of performances have been proved to be valuable tools, they have some key limitations which come mainly from the fact that they heavily depend on the features of a pre-trained network. Thus, it is unclear how they relate to any classical distance or divergence between probability distributions and how well they transfer to other datasets. More critically, it seems that they are heavily sensitive to particular implementation details (Barratt and Sharma, 2018; Parmar et al., 2022). Another possible measure of performances of generative models is the *Precision and Recall* measure (Sajjadi et al., 2018). Formally, for two probability distributions μ and ν, μ is said to have an attainable *precision* a and *recall* b with respect to ν, if it exists three probability distributions ξ, ξ_μ and ξ_ν such that

$$\mu = a\xi + (1-a)\xi_\mu \qquad \text{and} \qquad \nu = b\xi + (1-b)\xi_\nu .$$

The component ξ_ν denotes the part of ν that is missed by μ, whereas, ξ_μ denotes the noise part of μ. The maximal attainable precision and recall $\bar{a}$ and $\bar{\nu}$ are respectively $\bar{a} = \mu(\mathrm{supp}(\nu))$ and $\bar{b} = \nu(\mathrm{supp}(\mu))$. Applied to generative models, the idea is that the precision a quantifies the fidelity of the model, while the recall b quantifies the diversity.

6.3 Conclusion

In this chapter, we have presented the key concepts of generative modeling as well as the most commonly used models in imaging science. We have shown that most of the models could either be classified as *push-forward generative models* or as *indirect push-forward generative models*. In (direct) push-forward generative models, the generated distribution is of the form $g_{\theta\#}\mu_{d'}$ with $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$ being the standard Gaussian distribution in dimension d' and g_θ being a neural network which is directly optimized during training. In constrast, the generation dynamics G_θ in indirect push-forward generative models takes the form of a Monte Carlo procedure. Since these generation dynamics are necessarily only composed of a finite number of iterations in practice, the generated distribution in indirect push-forward generative models is also of the form $G_{\theta\#}\mu_{d'}$, but this time d' is much larger than d and the optimization is not performed directly on G_θ.

Generative modeling is a challenging problem which has become very popular over the recent years because of its impressive abilities to generate photorealistic images. In just a few years, the performance as well as the models themselves have evolved considerably. Still, these changes stem more from practical progresses than from theoretical advances. Indeed, the theoretical understandings of most of generative models remain relatively nascent. For instance, if analyzing the expressivity of deep neural networks has been an active field since the 90s, little is known however, to the best of our knowledge, on the expressivity of deep generative models, with the exception of several works on GANs that have focused on the case where the target distribution lies on two or more disconnected manifolds (Khayatkhoei et al., 2018; Mehr et al., 2019; Tanielian et al., 2020) and works on normalizing flows that have shown that the invertibility constraints limits the expressivity of the models (Cornish et al., 2020; Hagemann and Neumayer, 2021; Behrmann et al., 2021). In the next chapter, we study more generally the expressivity of *push-forward generative models* and we show that there is a trade-off for these models between expressivity and stability of training.

Chapter 7

Fitting push-forward generative models on multimodal distributions

Contents

In this chapter, we study the expressivity of push-forward generative models relatively to the Lipschitz constant of the generative network when the target distribution is multimodal. This chapter is mostly a reproduction of Salmona et al. (2022b).

7.1 Introduction

Generative modeling has become over the last years one of the most popular research topics in machine learning and computer vision. Beside their direct application (Sandfort et al., 2019; Antoniou et al., 2018), generative models have been used in numerous applications in various machine learning subfields, such as solving inverse problems (Ravuri et al., 2021; Ledig et al., 2017) or machine translation (Isola et al., 2017; Yang et al., 2018). However, most generative modeling methods still lack theoretical understanding and it remains often unclear whether the method approaches correctly the underlying probability distribution ν from which the data have been sampled, or only generates samples that appear to have been drawn from ν without fully recovering the underlying structure of the distribution.

Deep neural networks are most of the time Lipschitz mappings by design, since their activation functions are generally Lipschitz. In the literature, constraining the Lipschitz constant of a neural network is widely used as a way to increase its robustness (Virmaux and Scaman, 2018; Fazlyab et al., 2019), in particular to adversarial attacks (Goodfellow et al., 2015). Common approaches to bound Lipschitz constants of neural networks are spectral normalization (Miyato et al., 2018), adding a gradient penalization in the loss (Gulrajani et al., 2017; Mohajerin Esfahani and Kuhn, 2018), or Jacobian regularization (Pennington et al., 2017). These approaches have been widely used to stabilize the training of GANs, where Lipschitz constraints have been first imposed on discrimators (Arjovsky et al., 2017; Kodali et al., 2017; Fedus et al., 2018), while recent state-of-the-art architectures such as BigGAN (Brock et al., 2019), SAGAN (Zhang et al., 2019) or StyleGAN2 (Karras et al., 2020) also impose similar constraints on the generators through spectral normalization (Brock et al., 2019; Zhang et al., 2019), or Jacobian regularization (Karras et al., 2020). In contrast to GANs, the recent study of Kumar and Poole (2020) shows that the decoder Jacobian in VAEs is implicitly regularized, which limits its Lipschitz constant. A similar implicit regularization

might be operating in the case of normalizing flows (Behrmann et al., 2021), for which it is known that limited Lipschitz constant are necessary to ensure invertibility (Behrmann et al., 2019), and large bi-Lipschitz constants lead to numerical instability (Behrmann et al., 2021).

In the case where the underlying data distribution ν lies on two or more disconnected manifolds, several works (Khayatkhoei et al., 2018; Mehr et al., 2019; Tanielian et al., 2020; Cornish et al., 2020; Hagemann and Neumayer, 2021; Behrmann et al., 2021) have shown that GANs and normalizing flows were unable to correctly fit ν. This could be explained simply by an observation made by (Khayatkhoei et al., 2018): since the support of the distribution ν is a discontinuous set, a discontinuity must somehow been introduced in the transport map that pushes the standard Gaussian distribution $\mu_{d'}$ into ν. More generally, this raises the following question: are *push-forward generative models* able to correctly fit multimodal distributions? Indeed, since on one hand, it is clear in 1D that the mappings that push a standard Gaussian distribution into a given multimodal distribution must necessarily have large Lipchitz constants, and on the other hand, the Lipschitz constant of a neural network can almost be used as a measurement of the instability of its training (Glorot and Bengio, 2010; Szegedy et al., 2013; Pennington et al., 2017), it seems that the ability of push-forward generative models to generate multimodal distributions may be antagonist with the stability of their training. In this chapter, we prove that it is indeed the case, regardless of the dimension of the target measure ν nor the dimension of the standard Gaussian distribution $\mu_{d'}$.

Recently, Dhariwal and Nichol (2021) trained an unconditional *Score-based Generative Model* (SGM) (Song and Ermon, 2019; Ho et al., 2020) on ImageNet (Russakovsky et al., 2015) and achieved state-of-the-art generation. To the best of our knowledge, there is no push-forward generative model capable of reaching this kind of performance on such a complex dataset without explicitly adding any conditional label information in the model, see (Brock et al., 2019) for instance. This suggests that *indirect push-forward generative models*, such as SGMs, might not suffer of the same limitations than push-forward generative models.

Contributions of this chapter. In this chapter, we study the expressivity of direct and indirect push-forward generative models in relation to the Lipschitz constant of the push-forward mapping they learn. More precisely, in Section 7.3, for a Lipschitz function g and a given multimodal probability distribution ν, we formally demonstrate that the Lipschitz constant of g must necessarily be large in order for $g_{\#}\mu_{d'}$ to approximate ν correctly, as it has been already intuitively observed in the literature (Lu et al., 2020; Luise et al., 2020; Khayatkhoei et al., 2018). As a direct consequence, we exhibit lower bounds on $D(g_{\#}\mu_{d'}, \nu)$, where D is the total variation distance or the Kullback-Leibler divergence, with an explicit dependence on the Lipschitz constant $\mathrm{Lip}(g)$ of g, which highlights that there is a fundamental trade-off for (direct) push-forward generative models between expressivity and stability of training. In Section 7.4, we illustrate these theoretical results on several experiments, showing the difficulties of GANs and VAEs to generate multimodal distributions. We compare these models with SGMs and show experimentally that SGMs seem to be able to generate correctly multimodal distributions while keeping the Lipschitz constant of the score network relatively small, suggesting that these models do not suffer of such previously mentioned limitations.

7.2 Related works

Assessing the efficiency of push-forward models is a recurrent and important question in the literature. Sajjadi et al. (2018) and Kynkäänniemi et al. (2019) propose Precison and Recall metrics to assess GANs, aiming to measure simultaneously the mode collapse and the proportion of off-manifold generated samples. Using similar metrics, Tanielian et al. (2020) prove an upper bound on the precision of vanilla GANs (the proportion of generated samples which could have been generated by the target distribution). To overcome this limitation, they simply propose to reject samples associated with large values of the generator Jacobian. The intuition behind this idea is that those samples lie in regions of the space where the discontinuous optimal generator would "jump" between modes and so are off-manifold.

In the context of normalizing flows, it has been shown that the invertibility constraint limits the expressivity of the model. Indeed, Cornish et al. (2020) show that distributions generated by invertible normalizing flows have a support which is necessarily homeomorphic to the support of the latent distribution. As an outcome, the Lipschitz constant of the inverse flow has to approach infinity to correctly approximate distributions lying on disconnected manifolds (Cornish et al., 2020; Hagemann and Neumayer, 2021; Behrmann et al., 2021). To improve the expressivity of normalizing flows, it has been proposed in Cornish et al. (2020) and Wu et al. (2020) to inject stochasticity in the model.

Another line of research focuses on the fact that the model has access to only the empirical distribution $\hat{\nu} = \frac{1}{n}\sum_i \delta_{x_i}$ and not to the true target distribution. For instance Nagarajan et al. (2018) study to what extend GANs only memorize the data. Gulrajani et al. (2018) highlight the fact that common GAN benchmarks prefer training set memorization to a model which imperfectly fits the true distribution but covers more of its support. Related to this, Stéphanovitch et al. (2022) study specifically the Wasserstein GAN case, where the latent distribution is uniform and construct an optimal generator which minimizes the Wasserstein distance of order 1 between the push-forward measure and the empirical distribution, thus deriving a lower bound on the 1-Wasserstein distance. In the same paper, and more related to our work, the authors study the asymptotic case of an infinite number of data and show that most of the time the minimal 1-Wasserstein distance between the push-forward measure and the target distribution remains strictly positive.

7.3 Push-forward measure and Lipschitz mappings

In this section, we study the properties of the push-forward measure $g_{\#}\mu_{d'}$ when $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$ is the standard Gaussian distribution in dimension d' and g is a Lipschitz mapping. In the following, we denote $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field on $\mathbb{R}^d$ and, for any probability measure γ on $\mathbb{R}^d$ and any Borel set $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, we define the γ-surface area of A by

$$\gamma^+(\partial \mathsf{A}) = \liminf_{\varepsilon \to 0^+} \frac{\gamma(\mathsf{A}_\varepsilon) - \gamma(\mathsf{A})}{\varepsilon} \ ,$$

where $\mathsf{A}_\varepsilon = \{x \in \mathbb{R}^d \ : \ \text{there exists } a \in \mathsf{A}, \ \|x - a\| \leq \varepsilon\}$ is the ε-extension of A and $\partial \mathsf{A}$ is the boundary of A. The γ-surface area can be interpreted as the mass of γ on the hypersurface $\partial \mathsf{A}$. Note that the support of γ and A can be sets of intrinsic dimension smaller than d, which is most of the time the case when working with real data which are likely to live on low dimensional manifolds (Pope et al., 2020).

7.3.1 Isoperimetric property of push-forward measures

The main theoretical result of this chapter establishes some properties of push-forward measures depending on the regularity of the push-forward mapping.

Theorem 7.3.1. *Let $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a Lipschitz function with Lipschitz constant $\mathrm{Lip}(g)$. Then for any Borel set $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$,*

$$\mathrm{Lip}(g)(g_{\#}\mu_{d'})^+(\partial \mathsf{A}) \geq \varphi\left(\Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))\right) \ , \tag{7.1}$$

where $\varphi(x) = (2\pi)^{-1/2}\exp[-x^2/2]$ and $\Phi(x) = \int_{-\infty}^{x} \varphi(t)\mathrm{d}t$. In addition, we have that for any $r \geq 0$

$$g_{\#}\mu_{d'}(\mathsf{A}_r) \geq \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))\right) \ . \tag{7.2}$$

For vizualization purpose, the graph of the function $\varphi \circ \Phi^{-1}$ is represented in Figure 7.1.

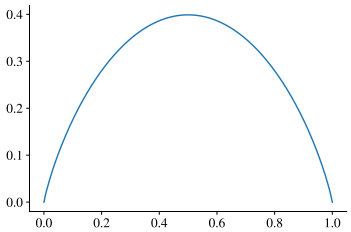

Figure 7.1: Graph of $\varphi \circ \Phi^{-1}$.

Theorem 7.3.1 is mainly a consequence of the Gaussian isoperimetric inequality (Sudakov and Tsirelson, 1978) which can be stated as follows.

Lemma 7.3.2 (Sudakov and Tsirelson, 1978). *Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^{d'})$ and $\mu_{d'} = \mathrm{N}(0, \mathrm{Id}_{d'})$. Then we have*

$$\mu_{d'}^+(\partial \mathsf{A}) \geq \varphi(\Phi^{-1}(\mu_{d'}(\mathsf{A}))) \ ,$$

where $\varphi(x) = (2\pi)^{-1/2}\exp[-x^2/2]$ *and* $\Phi(x) = \int_{-\infty}^{x}\varphi(t)\mathrm{d}t$. *Equivalently, for all* $r \geq 0$,

$$\mu_{d'}(\mathsf{A}_r) \geq \Phi(r + \Phi^{-1}(\mu_{d'}(\mathsf{A}))) \; .$$

Informally, the Gaussian isoperimetric implies that the half-spaces of $\mathbb{R}^{d'}$, i.e. the spaces of the form $\{x \in \mathbb{R}^{d'} \; : \; a^T x \geq 0\}$ with $a \in \mathbb{R}^{d'}$, have minimal $\mu_{d'}$-surface area. Indeed, it can be shown (see proof of Corollary 7.3.3 for details) that for any half-space H of $\mathbb{R}^{d'}$, $\mu_{d'}^{+}(\mathsf{H}) = \varphi(\Phi^{-1}(\mu_{d'}(\mathsf{H})))$. An illustration of the Gaussian isoperimetric inequality can be found in Figure 7.2.

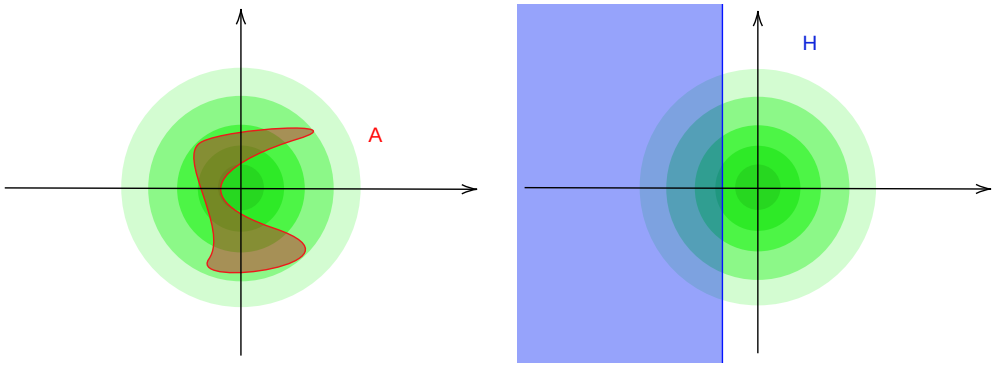

Figure 7.2: The Gaussian isoperimetric inequality on $\mathbb{R}^2$. Supposing $\mu_2(\mathsf{A}) = \mu_2(\mathsf{H})$, Lemma 7.3.2 implies that μ_2 has more mass on $\partial\mathsf{H}$ (in blue, right) than on $\partial\mathsf{A}$ (in red, left).

Now, we are ready to turn to the proof of Theorem 7.3.1.

Proof of Theorem 7.3.1. Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ such that $g_{\#}\mu_{d'}(\mathsf{A}) > 0$ (note that if $g_{\#}\mu_{d'}(\mathsf{A}) = 0$ then the result is trivial). First, we show that for any $\varepsilon > 0$, $g((g^{-1}(\mathsf{A}))_{\varepsilon/\mathrm{Lip}(g)}) \subset \mathsf{A}_\varepsilon$. Let x be in $g((g^{-1}(\mathsf{A}))_{\varepsilon/\mathrm{Lip}(g)})$. There exists $z_1 \in (g^{-1}(\mathsf{A}))_{\varepsilon/\mathrm{Lip}(g)}$ such that $g(z_1) = x$. There also exists $z_2 \in g^{-1}(\mathsf{A})$ such that

$$\|z_1 - z_2\| \leq \frac{\varepsilon}{\mathrm{Lip}(g)} \; .$$

Hence, we have that

$$\|x - a\| \leq \mathrm{Lip}(g)\|z_1 - z_2\| \leq \varepsilon \; ,$$

where $a = g(z_2)$. Since $z_2 \in g^{-1}(\mathsf{A})$, $a \in \mathsf{A}$, and therefore $x \in \mathsf{A}_\varepsilon$. Using this result, the fact that $g_{\#}\mu_{d'}(\mathsf{B}) = \mu_{d'}(g^{-1}(\mathsf{B}))$ and $\mathsf{B} \subset g^{-1}(g(\mathsf{B}))$ for any $\mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$, we have

$$\begin{aligned}\liminf_{\varepsilon\to 0^+}\frac{1}{\varepsilon}(g_{\#}\mu_{d'}(\mathsf{A}_\varepsilon) - g_{\#}\mu_{d'}(\mathsf{A})) &\geq \liminf_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\left(g_{\#}\mu_{d'}(g((g^{-1}(\mathsf{A}))_{\varepsilon/\mathrm{Lip}(g)})) - g_{\#}\mu_{d'}(\mathsf{A})\right) \\ &\geq \liminf_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\left(\mu_{d'}((g^{-1}(\mathsf{A}))_{\varepsilon/\mathrm{Lip}(g)}) - \mu_{d'}(g^{-1}(\mathsf{A}))\right) \; . \end{aligned} \tag{7.3}$$

Using Lemma 7.3.2, we have

$$\mathrm{Lip}(g)\liminf_{\varepsilon\to 0^+}\frac{1}{\varepsilon}\left(\mu_{d'}((g^{-1}(\mathsf{A}))_{\varepsilon/\mathrm{Lip}(g)}) - \mu_{d'}(g^{-1}(\mathsf{A}))\right) \geq \varphi(\Phi^{-1}(\mu_{d'}(g^{-1}(\mathsf{A})))) \; ,$$

Combining this result and (7.3), we get that

$$\mathrm{Lip}(g)(g_{\#}\mu_{d'})^{+}(\partial\mathsf{A}) \geq \varphi(\Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))) \; .$$

In addition, using Lemma 7.3.2, we have for all $r \geq 0$

$$\mu_{d'}((g^{-1}(\mathsf{A}))_{r/\mathrm{Lip}(g)}) \geq \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(\mu_{d'}(g^{-1}(\mathsf{A})))\right) \; .$$

Using this result and that $g((g^{-1}(\mathsf{A}))_{r/\mathrm{Lip}(g)}) \subset \mathsf{A}_r$, we have for any $r \geq 0$

$$g_{\#}\mu_{d'}(\mathsf{A}_r) = \mu_{d'}(g^{-1}(\mathsf{A}_r)) \geq \mu_{d'}((g^{-1}(\mathsf{A}))_{r/\mathrm{Lip}(g)}) \geq \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))\right) \; .$$

□

Note that (7.2) implies (7.1) upon remarking that (7.2) is an equality for $r = 0$, dividing by r and letting $r \to 0$. Theorem 7.3.1 recovers the Gaussian inequality in the case where g is the identity mapping and extends it to all Lipschitz mappings. As the Gaussian inequality, Theorem 7.3.1 is dimension free, in the sense that neither d, nor d', nor the intrinsic dimension of $g(\mathbb{R}^{d'})$ play a role in the lower bounds. In the following section, we are going to use Theorem 7.3.1 to (i) give a lower bound on the Lipschitz constant so that push-forward generative models *exactly* match the data distribution, (ii) give a lower bound on the total variation and the Kullback-Leibler divergence between the push-forward and data distributions which depends on the Lipschitz constant of the model.

7.3.2 Lower bounding the Lipschitz constant of push-forward mappings

Equation (7.1) implies that the Lipschitz constant of g must necessarily be large for $g_{\#}\mu_{d'}$ to be multimodal. It provides indeed a lower bound on the Lipschitz constant of the mappings g which push $\mu_{d'}$ into a given measure ν. In the extreme case where the support of ν is composed of disconnected manifolds, we retrieve that there doesn't exist any Lipschitz mapping which pushes $\mu_{d'}$ into ν since it can be found Borel sets A with null ν -surface area but such that the right-hand term of (7.1) is strictly positive, which occurs when $0 < \nu(\mathsf{A}) < 1$. In the intermediate case where the support of ν is connected but ν is multimodal, the less mass ν has between modes, the larger must be the Lipschitz constant of the mappings which push $\mu_{d'}$ into ν. Indeed, if ν has little mass between its modes, one can find sets A with small ν-surface area and such that $0 < \nu(\mathsf{A}) < 1$. As a toy example, we get an explicit bound on the Lipschitz constant of the mappings which push $\mu_{d'}$ into a mixture of two isotropic Gaussians.

Corollary 7.3.3. *Let $\nu = \lambda \mathrm{N}(m_1, \sigma^2 \mathrm{Id}_d) + (1-\lambda)\mathrm{N}(m_2, \sigma^2 \mathrm{Id}_d)$ with $m_1, m_2 \in \mathbb{R}^d$, $\sigma > 0$ and $\lambda \in (0,1)$. Assume that there exists $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ Lipschitz such that $g_{\#}\mu_{d'} = \nu$. Then*

$$\mathrm{Lip}(g) \geq \sigma \exp\left[\|m_2 - m_1\|^2/(8\sigma^2) - (\Phi^{-1}(\lambda))^2/2\right] .$$

Proof of Corollary 7.3.3. We prove the corollary when $\nu = \lambda \mathrm{N}(-m, \sigma^2 \mathrm{Id}_d) + (1-\lambda)\mathrm{N}(m, \sigma^2 \mathrm{Id}_d)$ since the problem can always be reduced to that case by translation and setting $m = (m_2 - m_1)/2$. Let H be defined by $\mathsf{H} = \{x \in \mathbb{R}^d \,:\, m^T x \geq 0\}$. Note that for any $x \in \partial\mathsf{H}$, $\|x - m\| = \|x + m\|$. Since the problem is invariant by rotation, we can consider without any loss of generality that $m = (\|m\|, 0, \dots, 0)$. In that case, we have $\nu = \nu_1 \otimes \mathrm{N}(0, \sigma^2 \mathrm{Id}_{d-1})$, where $\nu_1 = \lambda \mathrm{N}(-\|m\|, \sigma^2) + (1-\lambda)\mathrm{N}(\|m\|, \sigma^2)$, and $\otimes$ is the tensor product between measures. In this case, we have that $\mathsf{H} = \{x_1 \in \mathbb{R} \,:\, x_1 \geq 0\} \times \mathbb{R}^{d-1}$. Therefore, we have

$$\begin{aligned}
\nu^+(\partial\mathsf{H}) &= \liminf_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(\int_{\mathsf{H}_\varepsilon} p_\nu(x)\mathrm{d}x - \int_{\mathsf{H}} p_\nu(x)\mathrm{d}x \right) , \\
&= \liminf_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \left(\int_{-\varepsilon}^{+\infty} \int_{\mathbb{R}^{d-1}} p_{\nu_1}(x_1)h(y)\mathrm{d}x_1\mathrm{d}y - \int_{0}^{+\infty} \int_{\mathbb{R}^{d-1}} p_{\nu_1}(x_1)h(y)\mathrm{d}x_1\mathrm{d}y \right) ,
\end{aligned}$$

where p_ν and p_{ν_1} are the respective densities of ν and ν_1, and h is the density of $\mathrm{N}(0, \sigma^2 I_{d-1})$. It follows that

$$\begin{aligned}
\nu^+(\partial\mathsf{H}) &= \liminf_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \int_{-\varepsilon}^{0} p_{\nu_1}(x_1) \left(\int_{\mathbb{R}^{d-1}} h(y)\mathrm{d}y \right) \mathrm{d}x_1 \\
&= \liminf_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \int_{-\varepsilon}^{0} p_{\nu_1}(x_1)\mathrm{d}x_1 = p_{\nu_1}(0) = (2\pi\sigma^2)^{-1/2} \exp[-\|m\|^2/(2\sigma^2)] .
\end{aligned}$$

Applying Theorem 7.3.1, we get that

$$\mathrm{Lip}(g) \geq \varphi(\Phi^{-1}(\nu(\mathsf{H})))/\nu^+(\partial\mathsf{H}) .$$

Furthermore, one can derive that

$$\begin{aligned}
\nu(\mathsf{H}) &= \lambda(1 - \Phi(m/\sigma)) + \Phi(m/\sigma)(1-\lambda) \\
&= \lambda(1 - 2\Phi(m/\sigma)) + \Phi(m/\sigma) .
\end{aligned}$$

Observing that $\lambda - \nu(\mathsf{H})$ is an increasing function of λ and $\lambda - \nu(\mathsf{H}) = 0$ if $\lambda = 1/2$, we get that $\lambda \leq \nu(\mathsf{H})$ if $\lambda \leq 1/2$ and $\lambda \geq \nu(\mathsf{H})$ if $\lambda \geq 1/2$. Since $\varphi \circ \Phi^{-1}$ reaches its maximum in $1/2$, it follows that for any $\lambda \in (0,1)$ we have

$$\varphi(\Phi^{-1}(\nu(\mathsf{H}))) \geq \varphi(\Phi^{-1}(\lambda)) ,$$

and thus

$$\begin{aligned}\mathrm{Lip}(g) &\geq (2\pi)^{1/2}\sigma\varphi(\Phi^{-1}(\lambda))\exp[\|m\|^2/(2\sigma^2)] \\ &\geq \sigma\exp[\|m\|^2/(2\sigma^2) - (\Phi^{-1}(\lambda))^2/2]\,,\end{aligned}$$

which concludes the proof. □

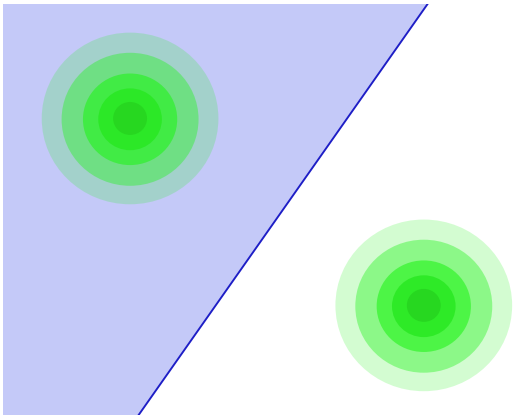

Figure 7.3: The half-space H used in the proof of Corollary 7.3.3 when $d = 2$.

Note that assuming there exists $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that $g_{\#}\mu_{d'} = \nu$ implies $d' \geq d$ since ν covers the whole ambient space and so g must be a surjective mapping. This bound is maximal in the balanced case when $\lambda = 1/2$ since $\Phi^{-1}(\lambda) = 0$ in that case. Otherwise, the more unbalanced the modes are, the smaller the bound is since the two terms in the exponential compensate each other more and more. Extending this corollary to mixtures of more than two Gaussians with different covariance matrices is technically difficult but we could expect a similar exponential growth in the square distance between modes since it depends mainly on the order of magnitude of the local minima of the distribution density.

7.3.2.1 Lipschitz constant of the Brenier map

As a by-product of Theorem 7.3.1, we also get the following result which shows that (in the one-dimensional case) the *Brenier map*, i.e. the optimal transport map for the ℓ_2 cost, minimizes the Lipschitz constant of the push-forward mapping.

Corollary 7.3.4. *Let ν be a probability measure on $\mathbb{R}$ with density w.r.t. the Lesbesgue measure and such that $\mathrm{supp}(\nu) = \mathbb{R}$. Assume that there exists $g : \mathbb{R}^{d'} \to \mathbb{R}$ Lipschitz such that $\nu = g_{\#}\mu_{d'}$. Let us denote $T_{\mathrm{OT}} = \Phi_\nu^{-1} \circ \Phi$ the Monge map between μ_1 and ν, where Φ_ν is the cumulative distribution function of ν. Then we have $\mathrm{Lip}(g) \geq \mathrm{Lip}(T_{\mathrm{OT}})$.*

Proof. Since ν admits a density p_ν with respect to the Lesbegue measure, it follows that Φ_ν is differentiable almost everywhere. Moreover, since $\mathrm{supp}(\nu) = \mathbb{R}$, it follows that $\Phi_\nu : \mathbb{R} \to (0, 1)$ is increasing and therefore is bijective, and so $T_{\mathrm{OT}} = \Phi_\nu^{-1} \circ \Phi$ is also differentiable almost everywhere and bijective, with inverse $T_{\mathrm{OT}}^{-1} = \Phi^{-1} \circ \Phi_\nu$, using (Peyré and Cuturi, 2019, Remark 2.29). Therefore, for any $x \in \mathbb{R}$ we have

$$\begin{aligned}T'_{\mathrm{OT}}(x) &= \frac{\varphi(x)}{p_\nu(T_{\mathrm{OT}}(x))} \\ &= \frac{\varphi(\Phi^{-1}(\Phi_\nu(T_{\mathrm{OT}}(x))))}{p_\nu(T_{\mathrm{OT}}(x))}\,.\end{aligned}$$

Let $y \in \mathbb{R}$. Using Theorem 7.3.1 with $\mathsf{A} = (-\infty, y]$ we get that for any $g : \mathbb{R}^{d'} \to \mathbb{R}$ Lipschitz such that $g_{\#}\mu_{d'} = \nu$,

$$\mathrm{Lip}(g) \geq \sup_{y\in\mathbb{R}} \frac{\varphi(\Phi^{-1}(\Phi_\nu(y)))}{p_\nu(y)}\,,$$

and so, since T_{OT} is bijective

$$\mathrm{Lip}(g) \geq \sup_{x\in\mathbb{R}} |T'_{\mathrm{OT}}(x)|\,,$$

which concludes the proof. □

Interestingly, we can also show that the same result holds in the case where the target measure ν is any non-degenerate Gaussian measure on $\mathbb{R}^d$.

Corollary 7.3.5. *Let $\nu = \mathrm{N}(m, \Sigma)$ with $m \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ non-singular. Let $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ Lipschitz with Lipschitz constant $\mathrm{Lip}(g)$ such that $g_{\#}\mu_{d'} = \nu$. Let T_{OT} the Brenier map between μ_d and ν, i.e. for all $x \in \mathbb{R}^d$,*

$$T_{\mathrm{OT}}(x) = m + \Sigma^{\frac{1}{2}} x \ .$$

Then,

$$\mathrm{Lip}(g) \geq \mathrm{Lip}(T_{\mathrm{OT}}) \ .$$

Proof. First observe that (7.1) can be rewritten in the following way:

$$\mathrm{Lip}(g)\nu^+(\partial \mathsf{A}) \geq \mu_d^+(\partial \mathsf{H}) \ ,$$

where H is any half-space such that $\mu_d(\mathsf{H}) = \nu(\mathsf{A})$. Let us denote $\{\lambda_i\}_i^d$ the eigenvalues of Σ and $\{p_i\}_i^d$ the corresponding unitary eigenvectors in $\mathbb{R}^d$. Let set for any $1 \leq i \leq d$,

$$\mathsf{A}_i = \{x \in \mathbb{R}^d \ : \ p_i^T(x - m) \geq 0\} \ .$$

It is easy to see that for any $1 \leq i \leq d$, $\nu(\mathsf{A}_i) = 1/2$. Since μ_d is invariant by rotation, we set

$$\mathsf{H} = \{x_1 \in \mathbb{R} \ : \ x_1 \leq 0\} \times \mathbb{R}^{d-1} \ ,$$

such that $\mu_d(\mathsf{H}) = 1/2$. Thus it follows, for any $1 \leq i \leq d$,

$$\mathrm{Lip}(g) \geq \sup_i \frac{\mu_d^+(\partial \mathsf{H})}{\nu^+(\partial \mathsf{A}_i)} \ .$$

On one hand we have, denoting $\varphi_{(m,\Sigma)}$ the density of $\mathrm{N}(m, \Sigma)$,

$$\begin{aligned}
\mu_d^+(\mathsf{H}) &= \liminf_{\varepsilon\to 0+} \frac{1}{\varepsilon}\left(\int_{\mathsf{H}_\varepsilon} \varphi_{(0,\mathrm{Id}_d)}(x)\mathrm{d}x - \int_{\mathsf{H}} \varphi_{(0,\mathrm{Id}_d)}(x)\mathrm{d}x\right) \\
&= \liminf_{\varepsilon\to 0+} \frac{1}{\varepsilon}\left(\int_0^\varepsilon \int_{\mathbb{R}^{d-1}} \varphi_{(0,\mathrm{Id}_{d-1})}(y)\varphi(t)\mathrm{d}y\mathrm{d}t\right) \\
&= \liminf_{\varepsilon\to 0+} \frac{1}{\varepsilon}\int_0^\varepsilon \varphi(t)\mathrm{d}t \\
&= \varphi(0) = (2\pi)^{-\frac{1}{2}} \ .
\end{aligned}$$

On the other hand we have for any $1 \leq i \leq d$, denoting $D = \mathrm{diag}\left((\lambda_i)_{i\leq d}\right)$

$$\begin{aligned}
\nu^+(\mathsf{A}_i) &= \liminf_{\varepsilon\to 0+} \frac{1}{\varepsilon}\left(\int_{\mathsf{A}_{i_\varepsilon}} \varphi_{(m,\Sigma)}(x)\mathrm{d}x - \int_{\mathsf{A}_i} \varphi_{(m,\Sigma)}(x)\mathrm{d}x\right) \\
&= \liminf_{\varepsilon\to 0+} \frac{1}{\varepsilon}\left(\int_{P^T(\mathsf{A}_{i_\varepsilon} - m)} \varphi_{(0,D)}(x)\mathrm{d}x - \int_{P^T(\mathsf{A}_i - m)} \varphi_{(0,D)}(x)\mathrm{d}x\right) \\
&= \liminf_{\varepsilon\to 0+} \frac{1}{\varepsilon}\left(\int_0^\varepsilon \int_{\mathbb{R}^{d-1}} \varphi_{(0,D^{(i)})}(y)\varphi_{(0,\lambda_i)}(t)\mathrm{d}y\mathrm{d}t\right) \\
&= \liminf_{\varepsilon\to 0+} \frac{1}{\varepsilon}\int_0^\varepsilon \varphi_{(0,\lambda_i)}(t)\mathrm{d}t \\
&= \varphi_{(0,\lambda_i)}(0) = (2\pi\lambda_i)^{-\frac{1}{2}} \ .
\end{aligned}$$

Thus it follows,

$$\mathrm{Lip}(g) \geq \sup_i \sqrt{\lambda_i} \ .$$

Moreover, since $T_{\mathrm{OT}}(x) = m + \Sigma^{1/2}x$ for all $x \in \mathbb{R}^d$, we have, denoting $|||.|||$ the operator matrix norm, i.e. for any M of size $d \times d$, $|||M||| = \sup_{\|x\|=1} \|Mx\|$,

$$\mathrm{Lip}(T_{\mathrm{OT}}) = \sup_{x\in\mathbb{R}^d} |||J_{T_{\mathrm{OT}}}(x)||| = |||\Sigma^{1/2}||| = \sup_i \sqrt{\lambda_i} \ ,$$

which concludes the proof. □

Apart from these two particular cases where the analytical expression of the Brenier map is known, identifying whether this result holds or not for arbitrary target distribution ν on $\mathbb{R}^d$ with $d > 1$ remains, to the best of our knowledge, an open problem.

7.3.3 Lower bounds on dissimilarity measures between probability distributions

Equation (7.2) provides a bound on the minimal mass the push-forward measure $g_{\#}\mu_{d'}$ can have on a given set when g is fixed with Lipschitz constant $\mathrm{Lip}(g)$. As a consequence, if ν is a distribution such that there exist sets on which ν has less mass than the minimal quantity that $g_{\#}\mu_{d'}$ can reach on those sets given the value of $\mathrm{Lip}(g)$, then $g_{\#}\mu_{d'}$ cannot be equal to ν, implying that most of dissimilarity measures between $g_{\#}\mu_{d'}$ and ν will be automatically strictly positive. In the following, we consider that g and ν are fixed and we derive lower bounds on the total variation distance and the Kullback-Leibler divergence between $g_{\#}\mu_{d'}$ and ν. We recall that the total variation distance between two probability measures on $\mathbb{R}^d$, ν_0, ν_1 is given by

$$D_{\mathrm{TV}}(\nu_0, \nu_1) = \sup\{\nu_0(\mathsf{A}) - \nu_1(\mathsf{A}) \ : \ \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)\} \ .$$

Similarly, we define the Kullback-Leibler divergence between two probability measures on $\mathbb{R}^d$, ν_0, ν_1, using the Donsker-Varadhan representation (Dupuis and Ellis, 2011, Lemma 1.4.3a):

$$D_{\mathrm{KL}}(\nu_0||\nu_1) = \sup\{\textstyle\int_{\mathbb{R}^d} f(x)\mathrm{d}\nu_0(x) - \log\left(\int_{\mathbb{R}^d} \exp[f(x)]\mathrm{d}\nu_1(x)\right) \ : \ f \in \mathfrak{B}(\mathbb{R}^d, \mathbb{R})\} \ ,$$

where $\mathfrak{B}(\mathbb{R}^d, \mathbb{R})$ denotes the set of all bounded mappings from $\mathbb{R}^d$ to $\mathbb{R}$. In the following, we will denote for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and $r > 0$,

$$\begin{aligned} \alpha_g(\mathsf{A}, r) &= \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))\right) \ , \\ \beta_g(\mathsf{A}, r) &= \alpha_g(\mathsf{A}, r) - g_{\#}\mu_{d'}(\mathsf{A}) \ , \end{aligned}$$

where $\alpha_g(\mathsf{A}, r)$ and $\beta_g(\mathsf{A}, r)$ are the lower bounds of $g_{\#}\mu_{d'}(\mathsf{A}_r)$ and $g_{\#}\mu_{d'}(\mathsf{A}_r \backslash \mathsf{A})$ provided by Theorem 7.3.1.

7.3.3.1 Lower bound on the total variation distance

We start by showing the following lower bound on the total variation distance.

Theorem 7.3.6. *Let ν be a probability measure on $\mathbb{R}^d$ and let $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a Lipschitz function. Then,*

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \sup\{\alpha_g(\mathsf{A}, r) - \min\{g_{\#}\mu_{d'}(\mathsf{A}), \nu(\mathsf{A})\} - \nu(\mathsf{A}_r \setminus \mathsf{A}) \ : \ \mathsf{A} \in \mathcal{B}(\mathbb{R}^d), r > 0\} \ . \tag{7.4}$$

Proof. Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and let $r > 0$. We have on one hand

$$\begin{aligned} |g_{\#}\mu_{d'}(\mathsf{A}_r \setminus \mathsf{A})| &\leq |g_{\#}\mu_{d'}(\mathsf{A}_r \setminus \mathsf{A}) - \nu(\mathsf{A}_r \setminus \mathsf{A})| + |\nu(\mathsf{A}_r \setminus \mathsf{A})| \\ &\leq D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) + \nu(\mathsf{A}_r \setminus \mathsf{A}) \ . \end{aligned}$$

Using Theorem 7.3.1, we get

$$|g_{\#}\mu_{d'}(\mathsf{A}_r \setminus \mathsf{A})| = g_{\#}\mu_{d'}(\mathsf{A}_r) - g_{\#}\mu_{d'}(\mathsf{A}) \geq \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))\right) - g_{\#}\mu_{d'}(\mathsf{A}) \ ,$$

and so

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \alpha_g(\mathsf{A}, r) - g_{\#}\mu_{d'}(\mathsf{A}) - \nu(\mathsf{A}_r \setminus \mathsf{A}) \ .$$

On the other hand, we have

$$\begin{aligned} |g_{\#}\mu_{d'}(\mathsf{A}_r)| &\leq |g_{\#}\mu_{d'}(\mathsf{A}_r) - \nu(\mathsf{A}_r)| + |\nu(\mathsf{A}_r)| \\ &\leq D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) + \nu(\mathsf{A}_r \setminus \mathsf{A}) + \nu(\mathsf{A}) \ . \end{aligned}$$

Using Theorem 7.3.1, we get

$$|g_{\#}\mu_{d'}(\mathsf{A}_r)| \geq \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))\right) \ ,$$

and so

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \alpha_g(\mathsf{A}, r) - \nu(\mathsf{A}) - \nu(\mathsf{A}_r \setminus \mathsf{A}) \ ,$$

which concludes the proof. □

Observe that (7.4) always holds but the right-hand term may become negative if the Lipschitz constant of g is large enough. The main idea behind this bound is to find a set A and a real $r > 0$ such that ν has a lot of mass on A but almost no mass on $\mathsf{A}_r \setminus \mathsf{A}$. For instance, if ν is a distribution on two disconnected manifolds M_1 and M_2, an optimal choice for A would either be M_1 or M_2 and the optimal r would be the distance between the two manifolds. Using Theorem 7.3.6, one can derive smaller but more explicit lower bounds only depending on ν and the Lipschitz constant of g. As a first example, we derive an explicit lower bound in the case where ν is a bi-modal distribution on two disconnected manifolds. The proof of the following result is postponed to Appendix B.1.

Corollary 7.3.7. *Let ν be a measure on $\mathbb{R}^d$ on two disconnected manifolds M_1 and M_2 such that $\nu(\mathsf{M}_1) = \lambda$ and $\nu(\mathsf{M}_2) = 1 - \lambda$, with $\lambda \in [1/2, 1)$, and let $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a Lipschitz function. Then,*

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \int_{\Phi^{-1}(\lambda)}^{d(\mathsf{M}_1,\mathsf{M}_2)/2\mathrm{Lip}(g)+\Phi^{-1}(\lambda)} \varphi(t)\mathrm{d}t \ ,$$

where $d(\mathsf{M}_1, \mathsf{M}_2) = \inf\{\|m_1 - m_2\| \ : \ m_1 \in \mathsf{M}_1, m_2 \in \mathsf{M}_2\}$.

As a second example, we also get an explicit lower bound in the case where ν is a mixture of two isotropic Gaussians (the proof is also postponed to Appendix B.1). For simplicity we stick to the balanced case.

Corollary 7.3.8. *Let $\nu = \frac{1}{2}[\mathrm{N}(m_1, \sigma^2 \mathrm{Id}_d) + \mathrm{N}(m_2, \sigma^2 \mathrm{Id}_d)]$ with $m_1, m_2 \in \mathbb{R}^d$ and $\sigma \geq 0$. Let $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a Lipschitz function. Then,*

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \int_0^{\|m_2-m_1\|/4\sigma\mathrm{Lip}(g)} \varphi(t)\mathrm{d}t - \frac{1}{2}\int_{\|m_2-m_1\|(2\sigma-1)/4\sigma^2}^{\|m_2-m_1\|(2\sigma+1)/4\sigma^2} \varphi(t)\mathrm{d}t \ .$$

In both corollaries, the lower bound tends to $1/2$ when the distance between the modes tends to infinity, meaning that $g_{\#}\mu_{d'}$ is far from well approaching ν. Note that the lower bound exhibited in Corollary 7.3.7 is always strictly positive regardless of the value of the Lipschitz constant of g. One can also observe that this latter bound is maximal in the balanced case, when $\lambda = 1/2$, since the standard normal distribution concentrates its mass around 0.

7.3.3.2 Lower bound on the Kullback-Leibler divergence

Now we derive a similar lower bound on the Kullback-Leibler divergence between $g_{\#}\mu_{d'}$ and ν. We consider the Kullback-Leibler divergence since this is a measure of dissimilarity between measures which is bounded and is very sensitive to the mismatch of supports between the generated and the data distributions.

Theorem 7.3.9. *Let ν be a probability measure on $\mathbb{R}^d$ and let $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a Lipschitz function. Then,*

$$D_{\mathrm{KL}}(g_{\#}\mu_{d'}||\nu) \geq \sup\left\{\beta_g(\mathsf{A}, r)\log\left(\frac{\beta_g(\mathsf{A},r)}{\nu(\mathsf{A}_r\setminus\mathsf{A})}\right) + (1-\beta_g(\mathsf{A}, r))\log\left(\frac{1-\beta_g(\mathsf{A},r)}{1-\nu(\mathsf{A}_r\setminus\mathsf{A})}\right) \ : \ \mathsf{A} \in \mathcal{B}(\mathbb{R}^d), r > 0\right\} .$$

Proof. Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $r > 0$ and $\zeta > 0$. We set for any $x \in \mathbb{R}^d$ $f(x) = \zeta\chi_{\mathsf{A}_r\setminus\mathsf{A}}(x)$, where χ_{A} denotes the characteristic function of the set A. Since f is bounded, it follows that

$$\begin{aligned} d_{KL}(g_{\#}\mu_{d'}||\nu) &\geq \int_{\mathbb{R}^d} f(x)dg_{\#}\mu_{d'}(x) - \log\left(\int_{\mathbb{R}^d} e^{f(x)}d\nu(x)\right) \\ &\geq \zeta g_{\#}\mu_{d'}(\mathsf{A}_r \setminus \mathsf{A}) - \log\left(1 + (e^{\zeta} - 1)\nu(\mathsf{A}_r \setminus \mathsf{A})\right) . \end{aligned}$$

Using Theorem 7.3.1, we get

$$g_{\#}\mu_{d'}(\mathsf{A}_r \setminus \mathsf{A}) = g_{\#}\mu_{d'}(\mathsf{A}_r) - g_{\#}\mu_{d'}(\mathsf{A}) \geq \beta_g(\mathsf{A}, r) \ .$$

Thus we get

$$D_{\mathrm{KL}}(g_{\#}\mu_{d'}||\nu) \geq \sup\{J(\zeta, \mathsf{A}, r) \ : \ \zeta \in \mathbb{R}, \mathsf{A} \in \mathcal{B}(\mathbb{R}^d), r > 0\} \ ,$$

where the functional J is defined by

$$J(\zeta, \mathsf{A}, r) = \zeta\beta_g(\mathsf{A}, r) - \log\left(1 + (e^{\zeta} - 1)\nu(\mathsf{A}_r \setminus \mathsf{A})\right) \ .$$

Differentiating J with respect to ζ, we get that

$$\nabla_\zeta J(\zeta, \mathsf{A}, r) = \beta_g(\mathsf{A}, r) - (e^\zeta \nu(\mathsf{A}_r \setminus \mathsf{A}))/(1 + (e^\zeta - 1)\nu(\mathsf{A}_r \setminus \mathsf{A})) \; .$$

Applying the first order condition, we get that:

$$\zeta^* = \log[\beta_g(\mathsf{A}, r)(1 - \nu(\mathsf{A}_r \setminus \mathsf{A}))] - \log[\nu(\mathsf{A}_r \setminus \mathsf{A})(1 - \beta_g(\mathsf{A}, r))] \; .$$

By re-injecting the value of ζ^*, we get

$$\begin{aligned}
\zeta^* \beta_g(\mathsf{A}, r) - \log\left(1 + (e^{\zeta^*} - 1)\nu(\mathsf{A}_r \setminus \mathsf{A})\right) &= \beta_g(\mathsf{A}, r) \log\left(\tfrac{\beta_g(\mathsf{A},r)(1-\nu(\mathsf{A}_r \setminus \mathsf{A}))}{\nu(\mathsf{A}_r \setminus \mathsf{A})(1-\beta_g(\mathsf{A},r))}\right) \\
&\quad - \log\left(\tfrac{1-\nu(\mathsf{A}_r \setminus \mathsf{A})}{1-\beta_g(\mathsf{A},r)}\right) \\
&= \beta_g(\mathsf{A}, r) \log\left(\tfrac{\beta_g(\mathsf{A},r)}{\nu(\mathsf{A}_r \setminus \mathsf{A})}\right) \\
&\quad + (1 - \beta_g(\mathsf{A}, r)) \log\left(\tfrac{1-\beta_g(\mathsf{A},r)}{1-\nu(\mathsf{A}_r \setminus \mathsf{A})}\right) \; ,
\end{aligned}$$

which concludes the proof. □

As above, this bound always holds but the right-hand term becomes negative if $\mathrm{Lip}(g)$ is large enough. As for Theorem 7.3.6, the main idea is to find a set A and a real r such that ν has a lot of mass on A, but ν has almost no mass on $\mathsf{A}_r \setminus \mathsf{A}$. Observe that if $\nu(\mathsf{A}_r \setminus \mathsf{A})$ tends to 0, the left-hand term of the bound tends to infinity. This is coherent with the behavior of the Kullback-Leibler divergence. Similarly to Corollary 7.3.8, we also get an explicit lower bound in the case where ν is a mixture of two isotropic Gaussians. As for Corollary 7.3.8, we stick to the balanced case for simplicity. The proof of the following result is postponed to Appendix B.1.

Corollary 7.3.10. *Let* $\nu = \frac{1}{2}\left[\mathrm{N}(m_1, \sigma^2 \,\mathrm{Id}_d) + \mathrm{N}(m_2, \sigma^2 \,\mathrm{Id}_d)\right]$ *with* $m_1, m_2 \in \mathbb{R}^d$ *and* $\sigma \geq 0$. *Let* $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ *be a Lipschitz function. We denote*

$$\lambda = g_{\#}\mu_{d'}\left(\{(m_2 - m_1)^T (x - (m_2 + m_1)/2) \leq 0 \; : \; x \in \mathbb{R}^d\}\right) \; ,$$

and we suppose without loss of generality, that $\lambda \in (0, 1/2]$. *Then,*

$$D_{\mathrm{KL}}(g_{\#}\mu_{d'}, \nu) \geq A \log\left(\tfrac{A}{B}\right) + (1 - A) \log\left(\tfrac{1-A}{1-B}\right) \; ,$$

where

$$A = \int_{-\Phi^{-1}(1-\lambda)}^{\|m_2 - m_1\|/4\sigma \mathrm{Lip}(g) - \Phi^{-1}(1-\lambda)} \varphi(t)\mathrm{d}t \qquad \textit{and} \qquad B = \frac{1}{2}\int_{\|m_2 - m_1\|(2\sigma-1)/4\sigma^2}^{\|m_2 - m_1\|(2\sigma+1)/4\sigma^2} \varphi(t)\mathrm{d}t \; .$$

Observe that this time, $\mathrm{Lip}(g)$ is no longer the only dependency in g since the bound also depends on the proportion of the modes of $g_{\#}\mu_{d'}$. However, it should be noted that when $g_{\#}\mu_{d'}$ approximates correctly ν, λ is automatically close to $1/2$ and so $\Phi^{-1}(1 - \lambda)$ is small in that case. To conclude, this section, we highlight the fact that, if our results are dimension free in theory, the dimension might be hidden in the distances between modes and the Lipschitz constant of g when working with real datasets. Indeed, the order of magnitude of the Euclidean distance between two samples x_i is likely to increase with the dimension d. As an outcome, the orders of magnitude of the distance between modes and so the Lipschitz constant that g must reach for correct generation probably increases with d also.

7.4 Experiments

In what follows, we illustrate the pratical implications of our results by training GANs, VAEs and SGMs on simple bi-modal distributions. More precisely, we show on one hand that generating multimodal distributions with GANs and VAEs is difficult since for those models, good generation necessarily involves generative networks with large Lipschitz constants. On the other hand, we show that SGMs seem to be able to generate multimodal distributions while keeping the Lipschitz constant of the score network relatively small and thus do not suffer of the same limitation. First, we focus on the univariate case where we can easily assess the Lipschitz constants of the networks. Then we illustrate our results in higher dimensions by training the three models on datasets derived from MNIST (LeCun et al., 1998). In all our experiments, we use the same architecture for the VAE decoder and the GAN generator in order to offer rigorous comparisons of the different models. For score-based modeling, we use architectures with similar numbers of learnable parameters. All details on the experiments and architecture of the networks can be found in Appendix B.3.

7.4.1 Univariate case

First, we train a VAE and a GAN with one-dimensional latent spaces on 50000 independent samples drawn from a balanced mixture of two univariate Gaussians $\nu = \frac{1}{2}[\mathrm{N}(-m, 1) + \mathrm{N}(m, 1)]$ for different values of $m > 0$. We also train a SGM on the same samples.

Histograms of generated distributions. Figure 7.4 shows histograms of generated distributions for $m = 10$ with the three different models. VAE models seem to generate Gaussians modes but interpolate significantly between them, while GANs do not interpolate but fail to retrieve the structure of the target distribution and forget parts of their support, which is known as *mode collapse* and is a common pitfall of such models (Arjovsky and Bottou, 2017; Metz et al., 2017). On the same task, SGMs do not suffer from such shortcomings. In the following section, we will link the interpolation/mode-collapsing properties of these models with their Lipschitz constants.

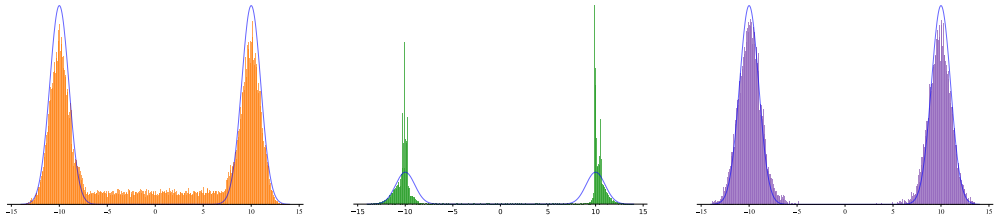

Figure 7.4: Histograms of distributions generated with VAE (left, in orange), GAN (middle, in green), and with SGM (right, in purple) for $m = 10$. The data distribution densities are plotted in blue.

Lipschitz constants and mass between modes. In Figure 7.5 (right), we observe that the GAN generator reaches much larger Lipschitz constants than the VAE decoder. This explains the difference of behaviors between GAN and VAE observed in Figure 7.4, as the mapping learned by the VAE is not stiff enough to concentrate the push forward measure on the two modes. One possible explanation for the interpolating behavior of the VAE is that the Euclidean norm of the Jacobian of the VAE decoder is implicitly regularized during training, as it has been demonstrated in Kumar and Poole (2020). Both GAN and VAE saturate the constraint on $g_{\#}\mu_{d'}([-m/2, m/2])$ provided by Theorem 7.3.1, meaning that the generative networks minimize the amount of mass between modes as much as their Lipschitz constants allow it. Finally, we can observe that the score network is able to keep a relatively small Lipschitz constant compared to the GAN, while managing to interpolate less than the latter. A probable explanation for this follows from the fact that the score network is used multiple time during inference. Hence, the Lipschitz constant of the push-forward mapping (the whole generation dynamic) is likely much larger than the Lipschitz constant of the neural network itself, and so the model is able to push-forward a Gaussian distribution into a multimodal distribution keeping a relatively small Lipschitz constant of the score network. Finally, in Figure 7.5 (left), we observe that when m increases, the Lipschitz constant of the VAE decoder and the GAN generator becomes rapidly much smaller than the value of the lower bound provided by Corollary 7.3.3. This means that for m large enough it is not possible to close the gap between the data distribution and the push-forward distribution. We highlight that this observation does not apply to SGMs since in this setting the network is applied multiple times.

Stability of GAN and mode collapse. Odena et al. (2018) suggested that the magnitude of the norm of the generator jacobian may be causally related to instability and mode collapse. This is why many state-of-the-art GANs apply spectral normalization (Miyato et al., 2018) on their generators. In Figure 7.6 (left), we show that this technique cannot be used when training GANs on multimodal distributions: since spectral normalization constraints the Lipschitz constant of the generator to be smaller than 1, the GAN is trained towards concentrating in one of the modes of the distribution over interpolating massively between them. This has been referred to as *mode dropping* by Khayatkhoei et al. (2018). To complete this analysis, we also train the GAN adding an additional gradient penalty term $10/L^2 \max_{z\sim\mathrm{N}(0,\mathrm{Id}_{d'})}(\|\nabla_z g_\theta(z)\|_2^2 - L)^2$, in the generator loss function, similarly to WP-GAN (Gulrajani et al., 2017), where L is an hyperparameter corresponding to the targeted Lipschitz constant. As expected, we can observe in Figure 7.6 (right), that when $\mathrm{Lip}(g)$ increases, the GAN begin to generate both modes

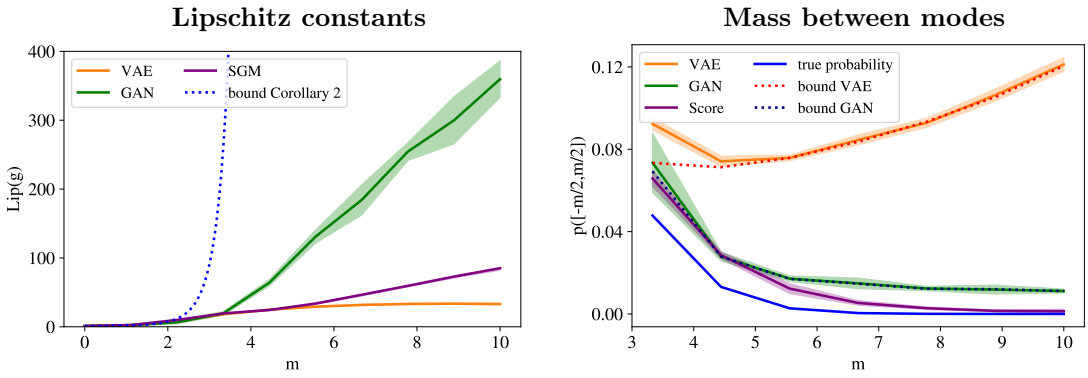

Figure 7.5: Left: evolution of the Lipschitz constants of the three different generative models trained on 50000 samples of $\frac{1}{2}[\mathrm{N}(-m,1)+\mathrm{N}(m,1)]$ in function of m. Right: evolution of the proportion of samples generated by the three models on the interval $[-m/2, m/2]$. We also show on this graph the lower bounds predicted by Theorem 7.3.1 for the VAE and the GAN, as well as the true probability $\nu([-m/2, m/2])$. Experiments are averaged over 10 runs and the colored bands correspond to +/- the standard deviation.

but becomes also more and more prone to mode collapse. This illustrates the fundamental trade-off between expressivity and robustness in push-forward generative models.

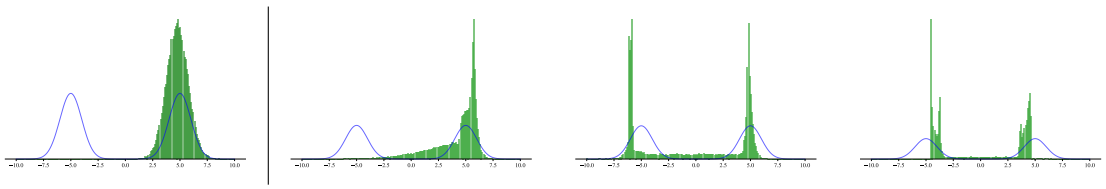

Figure 7.6: Histograms of distributions generated with GANs with spectral normalization applied on the generator (left), and with gradient penalty (right) for $\mathrm{Lip}(g) \approx L = 5$, $\mathrm{Lip}(g) \approx L = 15$ and $\mathrm{Lip}(g) \approx L = 25$. The data distribution densities are plotted in blue.

Influences of generator depth and time of training. In Figure 7.7, we study the effect of increasing the number of layers of the generative network as well as increasing the training time on the value of the Lipschitz constant of the VAE decoder and the GAN generator. In the VAE setting, the Lipschitz constant increases linearly with the depth of the decoder. This is not the case in the GAN setting, where increasing the size of the model seems to dramatically affect its stability. For both models, the Lipschitz constants of the generative network grow with the number of epochs. Yet this growth seems to be logarithmic for the VAE and the GAN seems to becomes more unstable as the number of epochs increases.

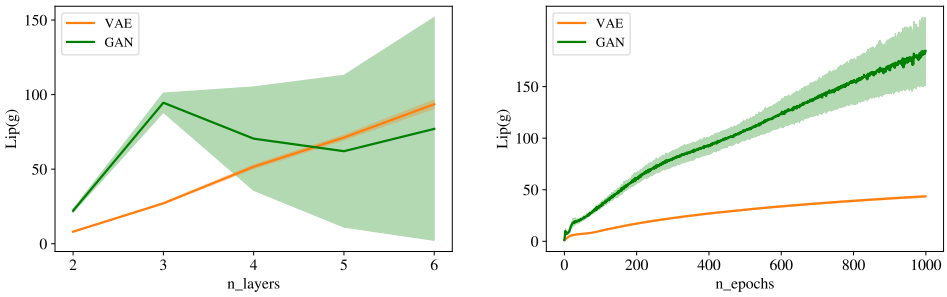

Figure 7.7: Evolution of the Lipschitz constant of the generative network with respect to its number of layers (left) and of the Lipschitz constant in function of the numbers of epochs (right). The experiments are averaged over 10 runs and the colored bands correspond to +/- the standard deviation.

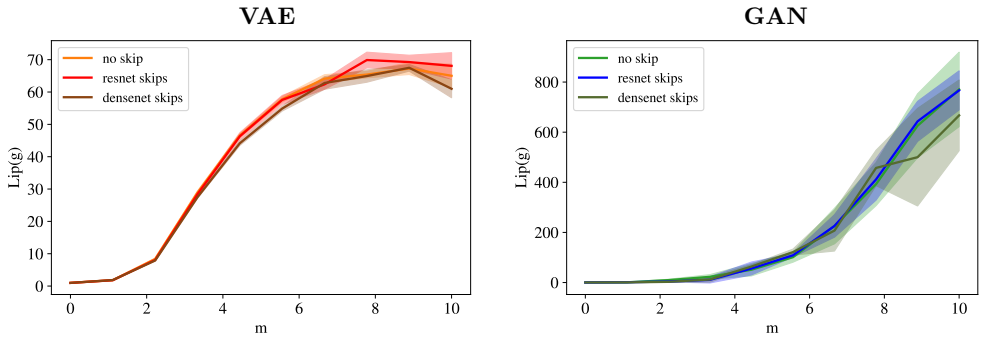

Figure 7.8: Evolution of the Lipschitz constant of the VAE decoder (left) and the GAN generator (right) trained on 50000 samples of $(1/2)[\mathrm{N}(-m,1)+\mathrm{N}(m,1)]$ for 3 different architectures of the generative network: simple feed-forward backbone, backbone with skip-connections of type "resnet", and backbone with skip-connections of type "densenet". Experiments are averaged over 5 runs and the colored bands correspond to +/- the standard deviation.

Influence of generator architecture. Finally, we study in Figure 7.8 the impact of the architecture of the generative network (i.e. the VAE decoder and the GAN generator) on its Lipschitz constant as well as on the training stability of the model by comparing three different architectures: first, we use a simple feed-forward network as precedently, then we add additive skip-connections of type "resnet" (He et al., 2016) to the previous backbone, and last we add concatenation skip-connections of type "densenet" (Huang et al., 2017) instead of additive skip-connections. For both models, it seems that more expressive decoder architectures do not help to reach larger values of Lipschitz constant. However, one can observe that in the GAN setting, even if the model remains certainly too unstable for correct distribution generation, adding additive skip-connections seems to stabilize the training a little since the colored bands are narrower than for the two other models. This suggests that some generator architectures may be better than others at learning mappings with large Lipschitz constants while staying stable.

7.4.2 Experiments on MNIST

We train a VAE, a GAN and a SGM on two datasets derived from MNIST (LeCun et al., 1998): first, two images of two different digits (3 and 7) are chosen and 10000 noisy versions of theses images are drawn with a noise amount of $\sigma = 0.15$, forming a dataset of $n = 20002$ independent samples drawn from a balanced mixture of two Gaussian distributions in dimension $784 = 28 \times 28$. Second, we train the models on the subset of all 3 and 7 of MNIST. We emphasize that our goal is not reach state-of-the-art performance on this problem but rather to illustrate our theoretical results in a moderate dimensional setting.

Mixture of Gaussians. For this experiment, we set the dimension of the latent space in the GAN and the VAE to $784 = 28 \times 28$ since it is the intrinsic dimension of the support of the data distribution. In order to visualize the interpolation between modes, we project the data on the line passing through the mean of each Gaussian, i.e. the two original clean images, and we plot histograms of the one-dimensional projections. In order to understand which bins of data in the histograms correspond to which digit, we train a classifier and we assign a color in function of which digit the data have been classified as. Results can be found in Figure 7.9 top. Moreover, GAN and VAE both fail to generate noisy versions of the images. As in the univariate case, the SGM is able to not interpolate between modes and seem to retrieve the Gaussian structure of the modes. This suggests that while direct push-forward models fail at representing multimodal distributions, considering stacked models with noise input at each step (as in SGM) might help to close the gap between the generated and the data distributions. However SGM does not manage to retrieve the right modes proportions. This is a well-known shortcoming of score-based models which has been studied in (Wenliang and Kanagawa, 2020).

Subset of MNIST. Finally, we train the three different models on the subset of MNIST composed of all 3 and 7 (no Gaussian noise was added). We choose a latent dimension of 20 for the VAE and the GAN. Since the Euclidean distance is not a meaningful metric to compare the different digits of MNIST, we use the deep Wasserstein embedding proposed by Courty et al. (2018): an autoencoder is learned in a

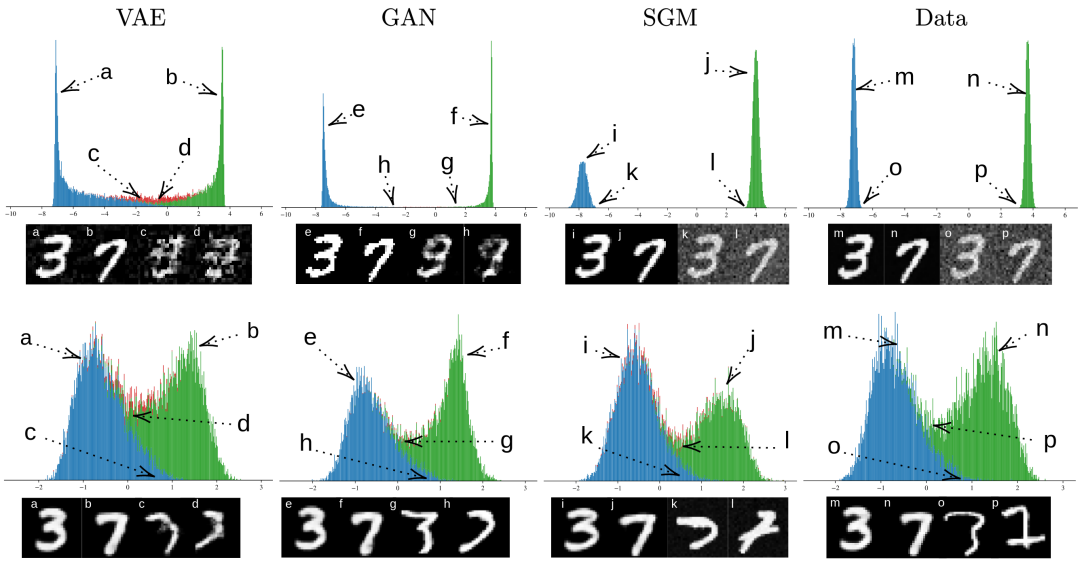

Figure 7.9: mixture of Gaussians (top): histograms of projections on the line passing through the mean of each Gaussian. Subset of MNIST (bottom): histograms of projections on the line passing through the barycenters of all the 3 and 7 in the deep Wasserstein embedding space. Bins of data are colored in blue if they are classified as 3, in green if classified as 7, and in red if classified as another digit.

supervised fashion such that the Euclidean distance in the latent space approximates the Wasserstein distance between pairs of images of MNIST. In the learned Wasserstein space, we project data on the line passing through the Euclidean barycenters of all 3 and 7 and plot histograms of projections, using the same classifier as before. Results can be found in Figure 7.9 (bottom). Note that the distribution does not exhibit strong multimodality features contrary to the mixture of Gaussians settings, see Figure 7.9. As before, the VAE interpolate between modes, the GAN manages to not interpolate but generate a narrower histogram, and the score-based model does not interpolate and seems to recover the structure of the distribution, but doesn't retrieve the right modes proportions. However, we emphasize that all these models seem to perform better than on the previous dataset. A possible explanation of this is that the modes are less separated than in the Gaussian mixtures and therefore the model is easier to train.

7.5 Discussion

In this chapter, given a Lipschitz mapping g and a measure ν, we derived lower bounds on the total variation distance and the Kullback-Leibler divergence between the push-forward measure $g_{\#}\mu_{d'}$ and ν depending on the Lipschitz constant of the mapping g. These bounds indicate how the mass between the modes of the push-forward measure depends on the Lipschitz constant of the push-forward mapping. They highlight the trade-off between the ability of VAEs and GANs to fit multimodal distributions and the stability of their training.

A common assumption in the imaging literature, validated empirically by Pope et al. (2020), is that distributions of natural images live on low dimensional manifolds. Understanding whether these distributions are composed of separated modes or not remains, to the best of our knowledge, an open problem. To that extent, the fact that unsupervised push-forward generative models perform well on datasets such as CelebA (Liu et al., 2015) could possibly be, in regard of our work, an indicator that the data distributions of those datasets are unimodal, or at least not composed of well separated modes.

Several techniques have been proposed in the literature to fit data distributions on disconnected manifolds. Most of them consist in overparametrizing the model, either by using stacked generative networks (Khayatkhoei et al., 2018; Mehr et al., 2019) or by learning a more complex latent distribution than the standard Gaussian (Gurumurthy et al., 2017; Rezende and Mohamed, 2015; Kingma et al., 2016; Luise et al., 2020). Other methods consist in rejecting a posteriori samples associated to large values of the Jacobian generator (Tanielian et al., 2020; Issenhuth et al., 2020). In this work, we empirically showed that score-based models seemed to be able to fit separated manifolds without model overparametrization or additional posterior sample rejection scheme. This suggests that the structure of the generation

dynamic in these models is particularly adapted to (indirectly) learn mappings with large Lipschitz constants. Their good performance on multimodal distributions might follow from the fact that these models do not optimize directly the push-forward mapping itself and/or that noise is injected at each step during the generation process. Hence, a future perspective of work would be to study what are the structural aspects of diffusion models that play a significant role in their expressivity.

A possible limitation of this work is that the bounds derived on the Kullback-Leibler divergence and total variation distance are not tight (see Appendix B.4), mainly because they take no account of the fact that when interpolating, $g_{\#}\mu_{d'}$ has automatically less mass than ν on the modes since a significant amount of its total mass is between them. In future work, we plan to tighten the gap between our bounds and the true distance.

Conclusion and perspectives

Overview of the contributions

In this thesis, we have provided some answers to three problems related to the transport of measures across different Euclidean spaces, the first two being in the context of optimal transport between measures on incomparable spaces and the last one being in the context of generative modeling. More precisely, in Part I, we have first defined an alternative formulation to the Gromov-Wasserstein distance that we have called *embedded Wasserstein* distance. Then we have studied the behavior of the Gromov-Wasserstein and the embedded distances between two Gaussian distributions $\mu = \mathrm{N}(m_0, \Sigma_0)$ and $\nu = \mathrm{N}(m_1, \Sigma_1)$. We have focused on the choices of costs of the squared Euclidean distances and the inner-products. We have derived closed-form solutions for the GW_2 problem with inner-products as cost functions and we have shown these solutions were also solutions of the GW_2 problem with squared Euclidean distances restricted to Gaussian couplings. We also have shown these solutions were also solutions of the embedded Wasserstein distance. Then, we have introduced two new OT distances on the set of Gaussian mixture models, MGW_2 and MEW_2, which can both be thought as generalizations of the distance proposed by Delon and Desolneux (2020) to Gaussian mixture models living in different dimensions. We have shown that these two OT distances can be used to solve relatively efficiently Gromov-Wasserstein related problems on Euclidean spaces, especially in moderate-to-large scale settings involving several tens of thousands of points.

In Part II, we first have shown that most of the generative models that are commonly used in imaging science could either be classified as push-forward generative models or as indirect push-forward generative models. Then, we have studied the expressivity of push-forward generative models relatively to the Lipschitz constant of the generative network when the target distribution is multimodal. More precisely, given a Lipschitz mapping g and a measure ν, we have derived lower bounds on the total variation distance and the Kullback-Leibler divergence between the push-forward measure $g_{\#}\mu_{d'}$ and ν depending on the Lipschitz constant of the mapping g. These bounds indicate how the mass between the modes of the push-forward measure depends on the Lipschitz constant of the push-forward mapping. They highlight the trade-off between the ability of push-forward generative models to fit multimodal distributions and the stability of their training. We have also empirically shown that indirect push-forward generative models such as diffusion models don't seem to suffer of such limitations.

Future perspectives of work

To conclude, we discuss in this section the possible extensions of the works presented in this thesis.

Choice of costs functions for Gromov-Wasserstein. A first perspective of work with regard of the results of Chapter 4 and in conjunction with the works of Vayer (2020), Beinert et al. (2022) and Dumont et al. (2022) on Gromov-Wasserstein between one-dimensional distributions, could be to study more in depth the differences between the choice of squared Euclidean distances and the choice of inner-products as cost functions. Indeed, the choice of squared Euclidean distances, despite being natural, seems to induce strange behaviors, both on one-dimensional (Beinert et al., 2022) and Gaussian distributions. In contrast, the choice of inner-products costs induces nice properties, both on one-dimensional (Vayer, 2020) and Gaussian distributions, and more generally on distributions that admit densities since (Dumont et al., 2022) have shown a result analogous to the Brenier theorem. Hence, a future perspective of work could be to compare the performances of the two distances $GW_2(\langle .\rangle_d, \langle .\rangle_{d'}, \mu, \nu)$ and $GW_2(\|.\|^2_{\mathbb{R}^d}, \|.\|^2_{\mathbb{R}^{d'}}, \mu, \nu)$ for solving Gromov-Wasserstein-related tasks such as shape matching.

Optimization landscape of the Gromov-Wasserstein objective. One potential limitation of the MGW_2 distance proposed in Chapter 5 is that the MGW_2 solver we have proposed might converge sometimes to a suboptimal local minimum. This is not specific to our method and comes from the gradient descent structure of the classic GW solvers. Still, when solving the GW problem between GMMs rather than solving it directly between the points, it is likely that we increase the probability of converging toward a sub-optimal local minimum because we inevitably introduce symmetries by simplifying the problem and so we probably increase in the mean time the number of local minima in the GW objective. A future perspective of work could be thus to analyze the optimization landscape of the GW objective. In particular: *what kind of couplings do the local minima correspond to?* It is likely that these local minima are intrinsically linked with the implicit isometric transformation that is applied to one of the two measures during the computation of the distance.

Globally solving the MGW_2 problem. Recently, Ryner et al. (2023) have proposed an algorithm for solving the GW problem with quadratic cost that is guaranteed to converge toward a global minimum. A future perspective of work could be to study if a similar idea could be applied for solving the MGW_2 problem. The method of Ryner et al. (2023) builds on the low-rank structure of the cost matrices in the quadratic case. In the MGW_2 setting, it is not clear that we even need such a low-rank property for the Wasserstein distance matrices since the problem is of very small scale if the number of components is not chosen too large. More generally, since the discrete GW problem involved in the MGW_2 method is of very small scale, it is possible that we can use global optimization methods which are usually not accessible in medium-to-large scale settings because too computationally expensive.

Expressivity of indirect push-forward generative models. A future perspective of work with regard to the work of Chapter 7 would be to study theoretically the expressivity of indirect push-forward models such as diffusion models. For instance, can we show that the Lipschitz constant of the whole generation dynamics can be larger than the Lipschitz constant of the neural network that approximates the score? Such a question is closely related to the question of *convergence of diffusion models* (De Bortoli et al., 2021; De Bortoli, 2022; Lee et al., 2022). One challenge in analysing the convergence of diffusion models lies in the fact that the approximation of the scores $\nabla_x \log p_\sigma$ of the noisy versions of the target distribution ν by a neural network s_θ introduces errors in the dynamics. Thus, analysing the convergence of diffusion models involves understanding how these errors affect the distribution ν_θ towards which the dynamics converges. In particular, can we generate a distribution ν using a neural network s_θ such that when σ is small, $x \mapsto s_\theta(x, \sigma)$ is much more regular than $\nabla_x \log p_\sigma$?

Generative modeling and optimal transport. A last future perspective of work with regard to the works presented in this thesis would be to investigate more in details the connections between optimal transport and generative modeling. From a pratical point of view, there exists several connections between the two fields. First, an OT distance can of course be used more or less directly as a loss function for the training of a generative model, as it is the case in Wasserstein-GANs (Arjovsky et al., 2017) or Wasserstein Autoencoders (Tolstikhin et al., 2018). Second, generative models such as GANs can be used to approximate Monge maps (González-Sanz et al., 2022; Fan et al., 2022) between two fixed - not necessarily Gaussian - distributions. Furthermore, Rout et al. (2021) have recently proposed a push-forward generative model where the push-forward map g_θ approximates a Monge map between a standard Gaussian distribution and the target distribution ν. The work of Rout et al. (2021) supposes that the dimension of the latent space d' is equal to the dimension of the ambient space d, but we could imagine generative models that approximate optimal transportation maps associated with OT distances that remain meaningful between measures living in different dimensions, as long as d' stays larger than the intrisic dimension of the support of ν. A question that arises is: *Is it interesting for a generative model to approximate an optimal transportation map?* To that extent, Corollary 7.3.4 brings some answers in 1D since it implies that the Monge map between the standard Gaussian distribution $\mu_1 = \mathrm{N}(0, 1)$ and the target distribution ν is the mapping which pushes μ_1 into ν with the smallest Lipschitz constant. Furthermore, Corollary 7.3.4 might also suggest that it is possible that any push-forward generative model in 1D actually approximates this Monge map as a consequence of the stochastic gradient descent algorithm being biased towards regular mappings regardless of the initialization (Mulayoff et al., 2021). However, if we have shown in Corollary 7.3.5 a similar result to Corollary 7.3.4 when ν is a Gaussian distribution on $\mathbb{R}^d$, it is not clear that Corollary 7.3.4 generalizes to arbitrary target distribution ν on $\mathbb{R}^d$. This problem shares close connections with studying the regularity of Brenier maps, which is an active research field on its own (Caffarelli, 1996; Figalli, 2007; Philippis, 2013; Paty et al., 2020). The

key ingredient in this field is the fact that the Brenier maps are gradients of convex potentials which are solutions of the *Monge-Ampère* equation defined in (2.7). If proving a result similar to Corollary 7.3.4 for abitrary target distributions ν on $\mathbb{R}^d$ seems probably too difficult in view of the advances in this field, one could try to study if a similar result to Corollary 7.3.4 holds when ν is a specific distribution on $\mathbb{R}^d$, for instance a Gaussian mixture.

Appendix A

Supplementary materials of Part I

Contents

A.1 Proofs of the claims of Chapter 3

A.1.1 Proof of Lemma 3.3.2

Proof of Lemma 3.3.2. First observe that for $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ and $b \in \mathbb{R}^d$, $y \mapsto Py + b$ is an isometry since we have, for any y and y' in $\mathbb{R}^{d'}$

$$\|Py + b - Py' - b\|^2 = \|P(y - y')\|^2 = (y - y')^T P^T P(y - y') = (y - y')^T (y - y') = \|y - y'\|^2 .$$

The converse is a consequence of the Mazur–Ulam theorem (Mazur and Ulam, 1932) that states - in the version of Baker (1971) - that an isometry from a real normed space to a *strictly convex* normed space, i.e. a normed space where the unit ball is a stricly convex set, is necessarily affine. Since it is easy to show that the unit ball $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$ is a strictly convex set, we get that for all $x \in \mathbb{R}^{d'}$, ϕ is of the form $y \mapsto Py + b$ with P being a matrix of size $d \times d'$, and $b \in \mathbb{R}^d$. Moreover we have for all $y, y' \in \mathbb{R}^{d'}$

$$\|\phi(y) - \phi(y')\|^2 = \|Py - Py'\|^2 = \|P(y - y')\|^2 = (y - y')^T P^T P(y - y') .$$

Since ϕ is an isometry, it follows that $\|y - y'\|^2 = (y - y')^T P^T P(y - y')$ and so $P^T P = \mathrm{Id}_{d'}$, which concludes the proof. □

A.1.2 Proof of Lemma 3.3.6

Proof of Lemma 3.3.6. Denoting $m_0 = \mathbb{E}_{X\sim\mu}[X]$, $m_1 = \mathbb{E}_{Y\sim\nu}[Y]$, $\tilde{x} = x - m_0$, and $\tilde{y} = y - m_1$, we have for any $\pi \in \Pi(\mu, \nu)$,

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \|x - Py - b\|^2 \mathrm{d}\pi(x, y) &= \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \|\tilde{x} - P\tilde{y} - b + m_0 - Pm_1\|^2 \mathrm{d}\pi(x, y) \\ &= \|m_0 - b - Pm_1\|^2 + \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \|\tilde{x} - P\tilde{y}\|^2 \mathrm{d}\pi(x, y) ,\end{aligned}$$

since $\int \langle \tilde{x} - P\tilde{y}, m_0 - b - Pm_1 \rangle \mathrm{d}\pi(x,y) = 0$. Thus it follows,

$$\inf_{\pi \in \Pi(\mu,\nu)} \inf_{P \in \mathfrak{P},\ b \in \mathbb{R}^d} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - Py - b\|^2 \mathrm{d}\pi(x,y)$$
$$= \inf_{P \in \mathfrak{P}} \left(\inf_{b \in \mathbb{R}^d} \|m_0 - Pm_1 - b\|^2 + \inf_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} \|x - Py\|^2 \mathrm{d}\pi(x,y) \right) .$$

Observe now that for any $P \in \mathfrak{P}$, $\|m_0 - Pm_1 - b\|^2 = 0$ if $b = m_0 - Pm_1$, which concludes the proof. □

A.1.3 Proof of Lemma 3.3.8

Proof of Lemma 3.3.8. Note that this lemma can be proven with a proof similar to the one of Alvarez-Melis et al. (2019, Lemma 4.2), using the min-max theorem for singular values. Here we offer an alternative proof based on Lagragian analysis. First observe that the supremum is achieved as a direct consequence of the Weierstrass theorem because $\mathfrak{P}$ is compact and the mapping $P \mapsto \mathrm{tr}(P^T K)$ is continuous. For a given $P \in \mathfrak{P}$, let $U_P \Sigma_P V_P^T$ be the SVD of P. The problem can be rewritten as

$$\max_{P \in \mathfrak{P}} \mathrm{tr}(V_P \Sigma_P^T U_P^T U_K \Sigma_K V_K^T) .$$

Now, let us denote $U = U_P^T U_K$ and $V = V_P^T V_K$. Observe that U is in $\mathbb{O}(\mathbb{R}^d)$ and V is in $\mathbb{O}(\mathbb{R}^{d'})$. Using the cyclical permutation of the trace operator, the problem becomes

$$\max_{P \in \mathfrak{P}} \mathrm{tr}(\Sigma_P^T U \Sigma_K V^T) .$$

Now, for a given fixed Σ_P, we determine which U and V maximize $\mathrm{tr}(\Sigma_P^T U \Sigma_K V^T)$. This problem reads as

$$\max_{U \in \mathbb{O}(\mathbb{R}^d),\ V \in \mathbb{O}(\mathbb{R}^{d'})} \mathrm{tr}(\Sigma_P^T U \Sigma_K V^T) .$$

The Lagrangian of this problem reads as

$$\mathcal{L}(U, V, C_0, C_1) = -\mathrm{tr}(\Sigma_P^T U \Sigma_K V^T) + \mathrm{tr}(C_0(U^T U - \mathrm{Id}_d)) + \mathrm{tr}(C_1(V^T V - \mathrm{Id}_{d'})) ,$$

where $C_0 \in \mathbb{S}^d$ and $C_1 \in \mathbb{S}^{d'}$ are the Lagrange multipliers respectively associated with the constraints $U \in \mathbb{O}(\mathbb{R}^d)$ and $V \in \mathbb{O}(\mathbb{R}^{d'})$. The first order condition gives

$$\begin{cases} \Sigma_P V \Sigma_K^T = 2UC_0 \\ \Sigma_P^T U \Sigma_K = 2VC_1 , \end{cases}$$

or equivalently

$$\begin{cases} U^T \Sigma_P V \Sigma_K^T = 2C_0 \\ \Sigma_P^T U \Sigma_K V^T = 2VC_1 V^T . \end{cases}$$

Since C_0 and C_1 are symmetric matrices (because they are associated with symmetric constraints), we get that both left-hand terms are symmetric. This gives the following conditions

$$\begin{cases} U^T \Sigma_P V \Sigma_K^T = \Sigma_K V^T \Sigma_P^T U \\ \Sigma_P^T U \Sigma_K V^T = V \Sigma_K^T U^T \Sigma_P . \end{cases}$$

Now, observe that when multiplying the first condition at right by $U^T \Sigma_P$ and multiplying the second condition at left by $\Sigma_K V^T$, we get by combining the two conditions

$$\begin{cases} U \Sigma_K V^T \Sigma_P^T \Sigma_P = \Sigma_P \Sigma_P^T U \Sigma_K V^T \\ U^T \Sigma_P V \Sigma_K^T \Sigma_K = \Sigma_K \Sigma_K^T U^T \Sigma_P V^T , \end{cases}$$

or equivalently,

$$\begin{cases} U \Sigma_K V^T D_P = D_P^{[d]} U \Sigma_K V^T \\ U^T \Sigma_P V D_K = D_K^{[d]} U^T \Sigma_P V^T , \end{cases}$$

where $D_P = \mathrm{diag}(\boldsymbol{\sigma}(P))$ and $D_K = \mathrm{diag}(\boldsymbol{\sigma}(K))$. Multiplying the first condition at left by $V\Sigma_K^T U^T$ and the second condition at right by $V\Sigma_P^T U$, this yields to

$$\begin{cases} VD_KV^TD_P = V\Sigma U^T D_P^{[d]} U\Sigma_K V^T \\ D_K^{[d]} U^T D_P^{[d]} U = U^T\Sigma_P V D_K V\Sigma_P^T U \end{cases}.$$

It follows that $VD_KV^TD_P$ and $D_K^{[d]}U^TD_P^{[d]}U$ are symmetric matrices and so VD_KV^T commutes with D_P and $U^TD_P^{[d]}U$ commutes with $D_K^{[d]}$. Thus we can deduce that U and V are permutation matrices. Since the singular values are ordered in non-increasing order, we deduce that the problem is maximized when $U = \mathrm{Id}_d$ and $V = \mathrm{Id}_{d'}$. This implies that $U_P = U_K$ and $V_P = V_K$, which concludes the proof. □

Note that Lemma 3.3.8 is especially useful when the constraint of belonging to the set $\mathfrak{P}$ can be expressed as a constraint on the singular values. Observe that this is the case of $\mathbb{V}_{d'}(\mathbb{R}^d)$ since for all $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$, we have $P^TP = \mathrm{Id}_{d'}$ and so an equivalent condition of belonging in $\mathbb{V}_{d'}(\mathbb{R}^d)$ is that $\boldsymbol{\sigma}(P) = \mathbb{1}_{d'}$.

A.1.4 Proof of Lemma 3.3.11

Proof of Lemma 3.3.11. First, using Lemma 3.3.2, we get that there exists $P_1 \in \mathbb{V}_d(\mathbb{R}^r)$ and $b_1 \in \mathbb{R}^r$ such that for all $x \in \mathbb{R}^d$, $\psi(x) = P_1x + b_1$. Since $r \geq d'$, we have, denoting $\bar{\mu}$, $\overline{\psi_\#\mu}$ and $\bar{\nu}$ the centered measures respectively associated with μ, $\psi_\#\mu$, and ν, and using successively Lemma 3.3.6 and Lemma 3.3.4,

$$\begin{aligned} EW_2^2(\psi_\#\mu,\nu) &= \inf_{\pi\in\Pi(\psi_\#\mu,\nu)} \inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^r),\ b\in\mathbb{R}^r} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} \|z - Py - b\|^2 \mathrm{d}\pi(z,y) \\ &= \inf_{\pi\in\Pi(\overline{\psi_\#\mu},\bar{\nu})} \inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^r)} \int_{\mathbb{R}^r\times\mathbb{R}^{d'}} \|z - Py\|^2 \mathrm{d}\pi(z,y) \\ &= \inf_{\pi\in\Pi(\bar{\mu},\bar{\nu})} \inf_{P\in\mathbb{V}_{d'}(\mathbb{R}^r)} \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} \|P_1x - Py\|^2 \mathrm{d}\pi(x,y) \\ &= \int_{\mathbb{R}^d} \|P_1x\|^2 \mathrm{d}\bar{\mu}(x) + \int_{\mathbb{R}^{d'}} \|Py\|^2 \mathrm{d}\bar{\nu}(y) - 2 \sup_{\pi\in\Pi(\bar{\mu},\bar{\nu})} \sup_{P\in\mathbb{V}_{d'}(\mathbb{R}^r)} \mathrm{tr}(P^TP_1K_\pi) \\ &= \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\bar{\mu}(x) + \int_{\mathbb{R}^{d'}} \|y\|^2 \mathrm{d}\bar{\nu}(y) - 2 \sup_{\pi\in\Pi(\bar{\mu},\bar{\nu})} \sup_{P\in\mathbb{V}_{d'}(\mathbb{R}^r)} \mathrm{tr}(P^TP_1K_\pi)\ , \end{aligned}$$

where $K_\pi = \int_{\mathbb{R}^d\times\mathbb{R}^{d'}} xy^T \mathrm{d}\pi(x,y)$. Applying Proposition 3.3.7, we get

$$\sup_{\pi\in\Pi(\bar{\mu},\bar{\nu})} \sup_{P\in\mathbb{V}_{d'}(\mathbb{R}^r)} \mathrm{tr}(P^TP_1K_\pi) = \sup_{\pi\in\Pi(\bar{\mu},\bar{\nu})} \|P_1K_\pi\|_*\ .$$

Now observe that P_1K_π has the same singular values as K_π since $K_\pi^TP_1^TP_1K_\pi = K_\pi^TK_\pi$. Thus $\|P_1K_\pi\|_* = \|K_\pi\|_*$ and so $EW_2(\psi_\#\mu,\nu) = EW_2(\mu,\nu)$, which concludes the proof. □

A.2 Proofs of the claims of Chapter 4

A.2.1 Proof of Lemma 4.2.2

Proof of Lemma 4.2.2. By simple computation, it follows for all $x, x' \in \mathbb{R}^d$

$$\|T_d(x) - T_d(x')\| = \|O_dx + x_d - O_dx - x_d\| = \|O_d(x - x')\| = \|x - x'\|\ ,$$

since $O_d \in \mathbb{O}(\mathbb{R}^d)$. The same reasoning can be made with T_d', which concludes the proof. □

A.2.2 Proof of Lemma 4.2.6

The proof of Lemma 4.2.6 is inspired from the proof of the closed form of W_2 between Gaussian distributions provided in Givens et al. (1984). Before turning to the actual proof of Lemma 4.2.6, we prove the following technical result, that we will use multiple times thorough the chapter.

Lemma A.2.1. *Suppose that $d \geq d'$. Let Σ be a positive semi-definite (PSD) matrix of size $d + d'$ of the form*

$$\Sigma = \begin{pmatrix} \Sigma_0 & K \\ K^T & \Sigma_1 \end{pmatrix} ,$$

with $\Sigma_0 \in \mathbb{S}^d_{++}$, $\Sigma_1 \in \mathbb{S}^{d'}_+$ and K being a rectangular matrix of size $d \times d'$. Let $S = \Sigma_1 - K^T\Sigma_0^{-1}K$ be the Schur complement of Σ. Then there exists $r \leq d'$ and $B_r \in \mathbb{V}_r(\mathbb{R}^d)$ such that

$$K = \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T ,$$

where $U_r \in \mathbb{V}_r(\mathbb{R}^{d'})$ and Λ_r is a diagonal positive matrix of size r such that

$$\Sigma_1 - S = U_r \Lambda_r^2 U_r^T .$$

Proof. For a given Schur complement $S = \Sigma_1 - K^T\Sigma_0^{-1}K$, we have $K^T\Sigma_0^{-1}K = \Sigma_1 - S$. Since $\Sigma_0 \in \mathbb{S}^d_{++}$, we can deduce that $K^T\Sigma_0^{-1}K \in \mathbb{S}^{d'}_+$ and so that $\Sigma_1 - S \in \mathbb{S}^{d'}_+$. We note r the rank of $K^T\Sigma_0^{-1}K$. One can observe that

$$r \leq d' \leq d ,$$

where the left-hand side inequality follows from the fact that $\mathrm{rk}(AB) \leq \min\{\mathrm{rk}(A), \mathrm{rk}(B)\}$. Then, $\Sigma_1 - S$ can be diagonalized

$$\Sigma_1 - S = K^T\Sigma_0^{-1}K = U\Lambda^2 U^T = U_r\Lambda_r^2 U_r^T , \tag{A.1}$$

with $\Lambda^2 = \mathrm{diag}(\lambda_1^2, ..., \lambda_r^2)^{[d']}$, $\Lambda_r^2 = \mathrm{diag}(\lambda_1^2, ..., \lambda_r^2)$, and $U_r \in \mathbb{V}_r(\mathbb{R}^{d'})$ such that $U = \begin{pmatrix} U_r & U_{d'-r} \end{pmatrix}$. From (A.1), we can deduce that

$$(\Sigma_0^{-\frac{1}{2}} K U_r \Lambda_r^{-1})^T \Sigma_0^{-\frac{1}{2}} K U_r \Lambda_r^{-1} = \mathrm{Id}_r ,$$

where Λ_r is the unique PSD square-root of Λ_r^2. Let us set $B_r = \Sigma_0^{-\frac{1}{2}} K U_r \Lambda_r^{-1}$ such that $B_r \in \mathbb{V}_r(\mathbb{R}^d)$. It follows that

$$KU_r = \Sigma_0^{\frac{1}{2}} B_r \Lambda_r .$$

Moreover, since $U_{d-r}^T K^T \Sigma_0^{-1} K U_{d-r} = 0$ and $\Sigma_0 \in S_d^{++}(\mathbb{R})$, it follows that $KU_{d'-r} = 0$ and so

$$K = KUU^T = KU_rU_r^T = \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T ,$$

which concludes the proof. □

Now we can turn to the proof of Lemma 4.2.6.

Proof of Lemma 4.2.6. We want to maximize $\mathrm{tr}(K^TK)$ with the constraint that Σ is semi-definite positive. Problem (4.3) can be written in the following way

$$\min_{S \in \mathbb{S}^{d'}_+} -\mathrm{tr}(K^TK) ,$$

where S is the Schur complement of Σ, i.e. $S = \Sigma_1 - K^T\Sigma_0^{-1}K$. Now using Lemma A.2.1, we can write $\mathrm{tr}(K^TK)$ as a function of B_r :

$$\begin{aligned} \mathrm{tr}(K^TK) &= \mathrm{tr}(U_r\Lambda_r B_r^T \Sigma_0 B_r \Lambda_r U_r^T) \\ &= \mathrm{tr}(U_r^T U_r \Lambda_r B_r^T \Sigma_0 B_r \Lambda_r) \\ &= \mathrm{tr}(\Lambda_r^2 B_r^T \Sigma_0 B_r) . \end{aligned}$$

Thus, for a given S, the set of K such that $K^T\Sigma_0^{-1}K = \Sigma_1 - S$ is parametrized by B_r. We want to find B_r which maximizes $\mathrm{tr}(K^TK)$ for a given S. This problem can be rewritten in the following way.

$$\min_{B_r \in \mathbb{V}_r(\mathbb{R}^d} -\mathrm{tr}(\Lambda_r^2 B_r^T \Sigma_0 B_r) . \tag{A.2}$$

The rest of the proof is a readaptation of the proof of the Anstreicher and Wolkowicz (2000, Proposition 3.1) when B_r is not a squared matrix. The Lagrangian of problem (A.2) can be written

$$\mathcal{L}(B_r, C) = -\mathrm{tr}(\Lambda_r^2 B_r^T \Sigma_0 B_r) + \mathrm{tr}(C(B_r^T B_r - \mathrm{Id}_r)) ,$$

where $C \in \mathbb{S}^r$ is the Lagrange multiplier associated to the constraint $B_r^T B_r = \mathrm{Id}_r$ (C is symmetric because $B_r^T B_r - \mathrm{Id}_r$ is symmetric). We can then derive the first-order condition

$$-2\Sigma_0 B_r \Lambda_r^2 + 2B_r C = 0 \ ,$$

or equivalently

$$\Sigma_0 B_r \Lambda_r^2 B_r^T = B_r C B_r^T \ .$$

Since C is symmetric, $B_r C B_r^T$ is also symmetric and it follows that $\Sigma_0 B_r \Lambda_r^2 B_r^T$ is symmetric. We can deduce that Σ_0 and $B_r \Lambda_r^2 B_r^T$ commute. Moreover, since Σ_0 and $B_r \Lambda_r^2 B_r^T$ are both symmetric, they can be diagonalized in the same basis. Since $B_r \in \mathbb{V}_r(\mathbb{R}^d)$, it can be thought as the r first vectors of an orthogonal basis of $\mathbb{R}^d$. It means there exists a matrix B_{d-r} such that

$$B_r \Lambda_r^2 B_r^T = B \Lambda_d^2 B^T \ ,$$

where $\Lambda_d^2 = \mathrm{diag}(\lambda_1^2, ..., \lambda_r^2)^{[d]}$ and $B = \begin{pmatrix} B_r & B_{d-r} \end{pmatrix}$. Thus the eigenvalues of $B_r \Lambda_r^2 B_r^T$ are exactly the eigenvalues of Λ_d^2. Since Σ_0 and $B_r \Lambda_r^2 B_r^T$ can be diagonalized in the same basis, we get that $\mathrm{tr}(\Lambda_r^2 B_r^T \Sigma_0 B_r) = \mathrm{tr}(\Sigma_0 B_r \Lambda_r^2 B_r^T) = \mathrm{tr}(D_0 \tilde{\Lambda}_d)$ where $\tilde{\Lambda}_d$ is a diagonal matrix with the same eigenvalues as Λ_d but in a different order. Now, it can be easily seen that the optimal value of (A.2) is reached when B is a permutation matrix which sorts the eigenvalues of Λ_d in non-increasing order.

Thus, for a given S, the maximum value of $\mathrm{tr}(K^T K)$ is $\mathrm{tr}(D_0 \tilde{\Lambda}_d(S))$. We can now establish for which S, $\mathrm{tr}(D_0 \tilde{\Lambda}_m(S))$ is optimal. For a given S, we denote $\lambda_1, ..., \lambda_{d'}$ the eigenvalues of $\Sigma_1 - S$ and $\beta_1, ..., \beta_{d'}$ the eigenvalues of Σ_1, both ordered in non-increasing order. Since $S \in \mathbb{S}_+^{d'}$, the following inequality holds for all $y \in \mathbb{R}^{d'}$,

$$y^T(\Sigma_1 - S)y \leq y^T \Sigma_1 y \ ,$$

ant this inequality still holds when restricted to any subspace of $\mathbb{R}^{d'}$. Using the classic algebra min-max Courant-Fischer theorem (Courant, 1920; Fischer, 1905), we can conclude that for all $1 \leq i \leq d'$,

$$\lambda_i \leq \beta_i \ .$$

Thus, the optimal value of $\mathrm{tr}(D_0 \tilde{\Lambda}_d(S))$ is reached when $S = 0$ and

$$\tilde{\Lambda}_d(0) = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} \ ,$$

and so $\mathrm{tr}(D_0 \tilde{\Lambda}_d(0)) = \mathrm{tr}(D_0^{(d')} D_1)$. Now let $A = (\widetilde{\mathrm{Id}}_{d'} (D_0^{(d')})^{\frac{1}{2}} D_1^{\frac{1}{2}})^{[d,d']}$ with $\widetilde{\mathrm{Id}}_{d'}$ being any matrix of the form $\mathrm{diag}((\pm 1)_{i \leq d'})$. It can be easily verified that $A^T D_0^{-1} A = D_1$ and if $K^* = P_0 A P_1^T$, then

$$K^{*T} \Sigma_0^{-1} K^* = P_1 A^T P_0^T \Sigma_0^{-1} P_0 A P_1^T = P_1 A^T D_0^{-1} A P_1^T = P_1 D_1 P_1^T = \Sigma_1 \ .$$

Furthermore $K^{*T} K^*$ has the same eigenvalues as $A^T A$ and $\mathrm{tr}(A^T A) = \mathrm{tr}(D_0^{(d')} D_1)$, which concludes the proof. □

A.2.3 Proof of Lemma 4.2.7

In order to prove Lemma 4.2.7, we will use the following result of Anstreicher and Wolkowicz (2000).

Lemma A.2.2 (Anstreicher and Wolkowicz, 2000). *Let Σ_0 and Σ_1 be two symmetric matrices of size d. We note $\Sigma_0 = P_0 \Lambda_0 P_0^T$ and $\Sigma_1 = P_1 \Lambda_1 P_1^T$ there respective diagonalization such that the eigenvalues of Λ_0 are sorted in non-increasing order and the eigenvalues of Λ_1 are sorted in increasing order. Then*

$$\min_{P^T P = I_d} \mathrm{tr}(\Sigma_0 P \Sigma_1 P^T) = \mathrm{tr}(\Lambda_0 \Lambda_1) \ ,$$

and it is achieved at $P^* = P_0 P_1^T$.

Now we are ready to turn to the proof of Lemma 4.2.7.

Proof of Lemma 4.2.7. As before, we want to maximize $\mathrm{tr}(KA)$ with the constraint that Σ is semi-definite positive. Problem (4.3) can be written in the following way

$$\min_{S \in \mathbb{S}_+^{d'}} -\mathrm{tr}(KA) \ ,$$

where S is the Schur complement of Σ, i.e. $S = \Sigma_1 - K^T\Sigma_0^{-1}K$. Now using Lemma A.2.1, we can write $\mathrm{tr}(KA)$ as a function of B_r:

$$\mathrm{tr}(KA) = \mathrm{tr}(A^TK^T) = \mathrm{tr}(A^TU_r\Lambda_rB_r^T\Sigma_0^{\frac{1}{2}}) = \mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T) \ .$$

For a given S, the problem of finding the optimal value is parametrized by B_r and is:

$$\min_{B_r\in\mathbb{V}_r(\mathbb{R}^d)} -\mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T) \ .$$

The Lagrangian of this problem reads as

$$\mathcal{L}(B_r, C) = -\mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T) + \mathrm{tr}(C(B_r^TB_r - \mathrm{Id}_r)) \ ,$$

where $C \in \mathbb{S}^r$ is the Lagrangian multiplier associated to the constraint $B_r^TB_r = \mathrm{Id}_r$. We then can derive the first-order condition that reads as

$$-\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_r + 2B_rC = 0 \ ,$$

or equivalently,

$$\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T = 2B_rCB_r^T \ .$$

Since C is symmetric, $2B_rCB_r^T$ is also symmetric and so $\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T \in$ is symmetric. Furthermore, the rank of $\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T$ is equal to 1 because $\mathrm{rk}(A) = 1$ a nd $\mathrm{rk}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T) = 0$ would imply that $\mathrm{tr}(KA) = 0$, which cannot be the maximum value of our problem. So there exists a vector $u_d \in \mathbb{R}^d$ such that

$$\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T = u_du_d^T \ .$$

Then we can reinject the value B_r in the expression:

$$\begin{aligned}\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T &= \Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_r\Lambda_r^{-1}U_r^TK^T\Sigma_0^{-\frac{1}{2}} \\ &= \Sigma_0^{\frac{1}{2}}A^TU_rU_r^TK^T\Sigma_0^{-\frac{1}{2}} \\ &= \Sigma_0^{\frac{1}{2}}A^TK^T\Sigma_0^{-\frac{1}{2}} \ ,\end{aligned}$$

where we used the fact that $K = KUU^T = KU_rU_r^T$ because $KU_{d'-r} = 0$. Thus, we have on one hand,

$$\mathrm{tr}(KA) = \mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TK^T\Sigma_0^{-\frac{1}{2}}) = \mathrm{tr}(u_du_d^T) = u_d^Tu_d \ .$$

On the other hand, we have

$$\begin{aligned}\Sigma_0^{\frac{1}{2}}A^TK^T\Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}A^TK^T\Sigma_0^{-\frac{1}{2}})^T &= \Sigma_0^{\frac{1}{2}}A^TK^T\Sigma_0^{-1}KAD_0^{\frac{1}{2}} \\ &= \Sigma_0^{\frac{1}{2}}A^T(\Sigma_1 - S)A\Sigma_0^{\frac{1}{2}} \\ &= u_du_d^Tu_du_d^T \\ &= u_d^Tu_du_du_d^T \ ,\end{aligned}$$

and thus

$$\mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^T(\Sigma_1 - S)A\Sigma_0^{\frac{1}{2}}) = u_d^Tu_d\mathrm{tr}(u_du_d^T) = (u_d^Tu_d)^2 = (\mathrm{tr}(KA))^2.$$

Then we determine for which S, $\mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^T(\Sigma_1 - S)A\Sigma_0^{\frac{1}{2}})$ is maximum. We have

$$\begin{aligned}\mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^T(\Sigma_1 - S)A\Sigma_0^{\frac{1}{2}}) &= \mathrm{tr}(A\Sigma_0A^T(\Sigma_1 - S)) \\ &= \mathrm{tr}(A\Sigma_0A^T\Sigma_1) - \mathrm{tr}(A\Sigma_0A^TS) \ .\end{aligned}$$

Let $B = A\Sigma_0A^T$. Observe that $B \in \mathbb{S}_+^{d'}$ and is of rank 1. Moreover, since $S \in \mathbb{S}_+^{d'}$, it can be diagonalized, such that $S = PDP^T$. As before, we will first determine the value of $\mathrm{tr}(BS)$ for a given D, then we will determine which D minimizes $\mathrm{tr}(BS)$. Thus, for a given D, we want to find the optimal value of the following problem.

$$\min_{P^TP=\mathrm{Id}_{d'}} \mathrm{tr}(BPDP^T) \ .$$

Since B is symmetric with rank 1, it has only one non null eigenvalue which is equal to its trace. Using Lemma A.2.2, we can deduce therefore that

$$\min_{P^TP=\mathrm{Id}_{d'}} \mathrm{tr}(BPDP^T) = \mathrm{tr}(B)\lambda_{d'},$$

where $\lambda_{d'}$ is the smallest eigenvalue of D. Since $S \in \mathbb{S}_+^{d'}$, the smallest possible value for $\lambda_{d'}$ is 0.

Now, if $\Sigma_0 = \mathrm{diag}(\alpha)$, $\Sigma_1 = \mathrm{diag}(\beta)$, it can be easily seen that $\mathrm{tr}(\mathbb{1}_{d',d}\Sigma_0\mathbb{1}_{d,d'}\Sigma_1) = \mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)$. Thus, if $K = \frac{\alpha\beta^T}{\sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}} = \frac{\Sigma_0\mathbb{1}_{d,d'}\Sigma_1}{\sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}}$, we can observe that

$$\mathrm{tr}(K\mathbb{1}_{d',d}) = \mathrm{tr}(\mathbb{1}_{d',d}K) = \frac{\mathrm{tr}(\mathbb{1}_{d',d}\Sigma_0\mathbb{1}_{d,d'}\Sigma_1)}{\sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}} = \sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}\,,$$

Now we must show that $S = \Sigma_1 - K^T\Sigma_0^{-1}K$ is in $\mathbb{S}_+^{d'}$. To do so, we will show that for all $1 \le i \le d'$, the determinant of the principal minor $S^{(i)}$ is positive. We can derive that

$$S = \Sigma_1 - \frac{\beta\alpha^T\Sigma_0^{-1}\alpha\beta^T}{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)} = \Sigma_1 - \frac{\beta\beta^T\mathrm{tr}(\Sigma_0)}{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)} = \Sigma_1 - \frac{\beta\beta^T}{\mathrm{tr}(\Sigma_1)}\,.$$

Using the matrix determinant lemma, it follows for all $1 \le i \le d'$,

$$\det(S^{(i)}) = \prod_k^i \beta_k \left(1 - \frac{\mathrm{tr}(\Sigma_1^{(i)})}{\mathrm{tr}(\Sigma_1)}\right)\,.$$

Thus, for all $1 \le i < d'$, $\det(S^{(i)}) > 0$, and $\det(S) = 0$. We conclude that S is in $\mathbb{S}_+^{d'}$ and that its smallest eigenvalue of is 0. □

A.2.4 Proof of Lemma 4.2.8

Proof of Lemma 4.2.8. for any $1 \le i \le d$ and any $1 \le j \le d'$, the Cauchy-Schwarz inequality tells us that

$$|\mathrm{Cov}(X_i, Y_j)| \le \sqrt{\mathbb{E}\left[X_i^2\right]\mathbb{E}\left[Y_j^2\right]}\,,$$

with equality if and only if $Y_j = \lambda X_i$ with $\lambda \in \mathbb{R}$. If $\mathrm{Cov}(X,Y)$ is of the form $(\widetilde{\mathrm{Id}}_{d'}(D_0^{(d')})^{\frac{1}{2}}D_1^{\frac{1}{2}})^{[d,d']}$, then for all $1 \le i \le d'$, we have

$$|\mathrm{Cov}(X_i, Y_i)| = \sqrt{\alpha_i\beta_i}\,,$$

where $\alpha_i = \mathbb{E}\left[X_i^2\right]$ and $\beta_i = \mathbb{E}\left[Y_i^2\right]$. Thus, for all $i \le d'$, $Y_i = \lambda_i X_i$ with $\lambda_i \in \mathbb{R}$. Since $X_i \sim \mathrm{N}(0, \alpha_i)$ and $Y_i \sim \mathrm{N}(0, \beta_i)$, it follows that $\lambda_i = \pm\sqrt{\frac{\beta_i}{\alpha_i}}$ and that

$$Y = \left(\widetilde{\mathrm{Id}}_{d'}D_1^{\frac{1}{2}}(D_0^{(d')})^{-\frac{1}{2}}\right)^{[d',d]}X.$$

Since Y depends linearly on X, it follows that (X,Y) is a Gaussian vector. Thus, using Isserlis Lemma 4.2.5, we can compute that

$$\mathrm{Cov}(X_i^2, Y_j^2) = \begin{cases} 2\alpha_i\beta_i & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Thus it follows that $\sum_{i,j}\mathrm{Cov}(X_i^2, Y_j^2) = 2\mathrm{tr}(D_0^{(d')}D_1)$. □

A.2.5 Proof of Lemma 4.2.12

Proof of Lemma 4.2.12. For $d \ge 1$, let Γ_d denote the set of vectors $v = (v_1, \ldots, v_d)$ of $\mathbb{R}^d$ such that $v_1 \ge v_2 \ge \ldots \ge v_d \ge 0$ and $\sum_{i=1}^d v_i^2 = 1$. We want to prove that

$$\forall u, v \in \Gamma_d, \quad \sum_{i=1}^d u_i v_i \ge \frac{1}{\sqrt{d}}\,.$$

We proceed by induction on d. For $d = 1$, it's obviously true since $\Gamma_1 = \{1\}$. Assume now $d > 1$, and the result true for $d-1$. Let $u, v \in \Gamma_d$, then using the result for $(u_2, \dots, u_d)/(\sum_{i=2}^d u_i^2)^{1/2}$ and $(v_2, \dots, v_d)/(\sum_{i=2}^d v_i^2)^{1/2}$ that both belong to Γ_{d-1}, we have

$$\begin{aligned}\sum_{i=1}^d u_i v_i = u_1 v_1 + \sum_{i=2}^d u_i v_i &\geq u_1 v_1 + \frac{1}{\sqrt{d-1}} \left(\sum_{i=2}^d u_i^2\right)^{1/2} \left(\sum_{i=2}^d v_i^2\right)^{1/2} \\ &= u_1 v_1 + \frac{1}{\sqrt{d-1}} \sqrt{1-u_1^2}\sqrt{1-v_1^2} \,.\end{aligned}$$

Now since $u, v \in \Gamma_d$, we have $u_1, v_1 \in [\frac{1}{\sqrt{d}}, 1]$. Let us denote $F(u_1, v_1) = u_1 v_1 + \frac{1}{\sqrt{d-1}}\sqrt{1-u_1^2}\sqrt{1-v_1^2}$. We have for all $v_1 \in [\frac{1}{\sqrt{d}}, 1]$:

$$F(1, v_1) = v_1 \geq \frac{1}{\sqrt{d}} \quad \text{and} \quad F\left(\frac{1}{\sqrt{d}}, v_1\right) = \frac{\sqrt{1-v_1^2} + v_1}{\sqrt{d}} \geq \frac{1 - v_1^2 + v_1}{\sqrt{d}} \geq \frac{1}{\sqrt{d}} \,.$$

And computing the partial derivative of F with respect to u_1, we get

$$\frac{\partial F}{\partial u_1}(u_1, v_1) = v_1 - \frac{u_1\sqrt{1-v_1^2}}{\sqrt{d-1}\sqrt{1-u_1^2}} \,.$$

This is a decreasing function of u_1, with value v_1 at $u_1 = 0$ and value that goes to $-\infty$ when u_1 goes to 1. Therefore the function $F(\cdot, v_1)$ on $[0, 1]$ is first increasing and then decrasing, showing that

$$\forall u_1 \in \left[\frac{1}{\sqrt{d}}, 1\right], \quad F(u_1, v_1) \geq \min\left(F(\frac{1}{\sqrt{d}}, v_1), F(1, v_1)\right) \geq \frac{1}{\sqrt{d}} \,.$$

Finally we thus have proved that

$$\sum_{i=1}^d u_i v_i \geq \frac{1}{\sqrt{d}} \,,$$

and moreover the equality is achieved when the vectors u and v are the vectors $(1, 0, \dots, 0)$ and $(\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}, \dots, \frac{1}{\sqrt{d}})$. □

A.2.6 Proof of Lemma 4.3.5

Before turning to the proof of Lemma 4.3.5, we will prove the following technical results.

Lemma A.2.3. *Let $A \in \mathbb{S}^d$. We denote λ_1 and λ_d its largest and smallest eigenvalues. For all $x \in \mathbb{R}^d$ such that $\|x\| = 1$, we have*

(i) x is an eigenvector of A associated to λ_1 if and only if $x^T A x = \lambda_1$.

(ii) x is an eigenvector of A associated to λ_d if and only if $x^T A x = \lambda_d$.

Proof. Let $x \in \mathbb{R}^d$ such $\|x\| = 1$. Since A is symmetric, there exists $O \in \mathbb{O}(\mathbb{R}^d)$ and $\Lambda = \mathrm{diag}((\lambda_k)_{1 \leq k \leq d})$ such that $x^T A x = x^T O A O^T x$. Denoting z the vector $O^T x$, we get thus

$$x^T A x = z^T \Lambda z = \sum_{k=1}^d \lambda_k z_k^2 \,.$$

Hence it follows that

$$\lambda_d \|z\|^2 \leq x^T A x \leq \lambda_1 \|z\|^2 \,,$$

with equality if and only if z is an eigenvector associated with λ_1 or λ_d. □

Now we are ready to prove Lemma 4.3.5.

Proof of Lemma 4.3.5. First we prove that $\max_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \langle P, K\rangle_{\mathcal{F}} = \|K\|_*$. Using Lemma 3.3.8, we get that

$$\sup_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \langle P, K\rangle_{\mathcal{F}} = \sup_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \mathrm{tr}(P^T K) = \max_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \mathrm{tr}(\Sigma_P^T \Sigma_K) \ ,$$

where $\Sigma_P = \mathrm{diag}^{[d,d']}(\boldsymbol{\sigma}(P))$ and $\Sigma_K = \mathrm{diag}^{[d,d']}(\boldsymbol{\sigma}(K))$. Now observe that for all $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$, $\boldsymbol{\sigma}(P)$ is necessarily equal to $\mathbb{1}_{d'}$ and so $\Sigma_P = \mathrm{Id}_{d'}^{[d,d']}$. Thus, by definition of the singular values, it follows,

$$\max_{P\in\mathbb{V}_{d'}(\mathbb{R}^d)} \mathrm{tr}(\Sigma_P^T \Sigma_K) = \mathrm{tr}(\mathrm{Id}_{d'}^{[d',d]}\, \Sigma_K) = \mathrm{tr}((K^T K)^{\frac{1}{2}}) = \|K\|_* \ ,$$

which proves the left-hand equality in (4.16). Now we prove the right-hand equality. The rest of the proof is inspired from the proof of the closed-form of the W_2 between two Gaussians provided by Givens et al. (1984). We want to solve the following constrained optimization problem

$$\min_{\substack{\Sigma_1 - K^T\Sigma_0^{-1}K \in \mathbb{S}_+^{d'} \\ P\in\mathbb{V}_{d'}(\mathbb{R}^d)}} -2\mathrm{tr}(P^T K) \ .$$

As before, using Lemma A.2.1, we can write $\mathrm{tr}(P^T K)$ as a function of B_r. This gives the following equivalent constrained optimization problem

$$\min_{B_r^T B_r = \mathrm{Id}_r, P^T P = \mathrm{Id}_{d'}} -2\mathrm{tr}(P^T \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T) \ .$$

The Lagrangian of this latter problem reads as

$$\mathcal{L}(B_r, P, C_0, C_1) = -2\mathrm{tr}(P^T \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T) + \mathrm{tr}(C_0(B_r^T B_r - \mathrm{Id}_r)) + \mathrm{tr}(C_1(P^T P - \mathrm{Id}_{d'})) \ ,$$

where $C_0 \in \mathbb{S}^r$ and $C_1 \in \mathbb{S}^{d'}$ are the Lagrange multipliers respectively associated with the constraints $B_r^T B_r = \mathrm{Id}_r$ and $P^T P = \mathrm{Id}_{d'}$. The first order condition gives

$$\begin{cases} \Sigma_0^{\frac{1}{2}} P U_r \Lambda_r = B_r C_0 \\ \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T = P C_1 \ . \end{cases}$$

Since Σ_0, P, U_r, and Λ_r are full rank, $\Sigma_0^{\frac{1}{2}} P U_r \Lambda_r$ is of rank r and so C_0 is also of rank r. Thus we get that

$$B_r = \Sigma_0^{\frac{1}{2}} P U_r \Lambda_r C_0^{-1} \ ,$$

and so

$$B_r^T B_r = \mathrm{Id}_r = C_0^{-1} \Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r C_0^{-1} \ .$$

Thus,

$$C_0 = (\Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r)^{\frac{1}{2}} \ .$$

On the other hand, by reinjecting the expression of B_r in the other first order condition we get

$$P^T \Sigma_0 P U_r \Lambda_r (\Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r)^{-\frac{1}{2}} \Lambda_r U_r^T = C_1 \ .$$

By multiplying this equation by itself we get

$$P^T \Sigma_0 P U_r \Lambda_r^2 U_r^T = C_1^2 \ .$$

Since C_1^2 is symmetric we get that $P^T \Sigma_0 P$ commutes with $U_r \Lambda_r^2 U_r^T$ and so $\Sigma_1 - S$. Moreover, as before we have

$$\begin{aligned} \mathrm{tr}(P^T K) &= \mathrm{tr}(((\Sigma_1 - S)^{\frac{1}{2}} P^T \Sigma_0 P (\Sigma_1 - S)^{\frac{1}{2}})^{\frac{1}{2}}) \\ &= \mathrm{tr}((\Sigma_1 - S)^{\frac{1}{2}} (P^T \Sigma_0 P)^{\frac{1}{2}}) \ . \end{aligned}$$

As before, using the Courant-Fischer min-max theorem (Courant, 1920; Fischer, 1905) to characterize the eigenvalues of $\Sigma_1 - S$, we get that $\mathrm{tr}(P^T K)$ is maximized when $S = 0$ and so the problem is equivalent to the following problem

$$\max_{\substack{P\in\mathbb{V}_{d'}(\mathbb{R}^d) \\ P^T\Sigma_0 P \Sigma_1 = \Sigma_1 P^T \Sigma_0 P}} \mathrm{tr}(\hat{D}_1^{\frac{1}{2}} D_{0,P}^{\frac{1}{2}}) \ , \tag{A.3}$$

where $(\hat{P}_1, \hat{D}_1)$ is any diagonalization of Σ_1 and $D_{0,P} = \hat{P}_1^T P^T \Sigma_0 P \hat{P}_1$. For all $y \in \mathbb{R}^{d'}$ we have

$$\alpha_d \|y\|^2 \leq y^T P^T \Sigma_0 P y \leq \alpha_1 \|y\|^2 \ ,$$

where $\alpha_1, \ldots, \alpha_d$ are the eigenvalues of Σ_0 ordered in non-increasing order. Thus, denoting $\lambda_1, \ldots, \lambda_{d'}$ the eigenvalues of $P^T \Sigma_0 P$, we get that for all $k \leq d'$,

$$\alpha_d \leq \lambda_k \leq \alpha_1 \ .$$

Since we want to maximize $\mathrm{tr}(\hat{D}_1^{\frac{1}{2}} D_{0,P}^{\frac{1}{2}})$, we set the largest eigenvalue λ_1 of $P^T \Sigma_0 P$ to α_1. We denote $y_1 \in \mathbb{R}^{d'}$ the eigenvector associated. We have $y_1 P^T \Sigma_0 P y_1 = \alpha_1$ and $\|P y_1\| = \|y_1\| = 1$ so using Lemma A.2.3, we get that $\|P y_1\|$ is an eigenvector of Σ_0 associated with α_1. Let λ_k and y_k be any other eigenvalue and its associated eigenvector in the orthonormal basis in which $P^T \Sigma_0 P$ is diagonal. We have $y_k^T y_1 = 0$ and so $y_k^T P^T P y_1 = 0$. Thus $P y_k$ is orthogonal to $P y_1$. Since $\|P y_k\| = 1$, we get that $P y_k$ is also an eigenvector of Σ_0 and so it exists $i \leq d-1$ such that $\lambda_k = y_k^T P^T \Sigma_0 P y_k = \alpha_i$. Thus, we conclude that the eigenvalues of the optimal $P^T \Sigma_0 P$ are the d' largest eigenvalues of Σ_0. Moreover, $\mathrm{tr}(\hat{D}_1^{\frac{1}{2}} D_{0,P}^{\frac{1}{2}})$ is clearly maximized when $D_{0,P}$ and $\hat{D}_1$ have their eigenvalues sorted in the same order. We conclude then that setting $D_{0,P} = D_0^{(d')}$ and $\hat{D}_1 = D_1$, where D_0 and D_1 are the diagonal matrices associated with the diagonalizations that sort the eigenvalues in non-increasing order, maximizes (A.3). This proves the right-hand equality of (4.16). Finally, observe that when setting K^* of the form

$$K^* = P_0 (\widetilde{\mathrm{Id}}_{d'} D_0^{(d')^{\frac{1}{2}}} D_1^{\frac{1}{2}})^{[d,d']} P_1^T \ ,$$

K^* is clearly in the feasible set since its the solution of Problem (4.3), and we have

$$\|K\|_* = \mathrm{tr}((K^{*T} K^*)^{\frac{1}{2}}) = \mathrm{tr}((D_0^{(d')} D_1)^{\frac{1}{2}}) = \mathrm{tr}(D_0^{(d')^{\frac{1}{2}}} D_1^{\frac{1}{2}}) \ ,$$

and so K^* is optimal. Furthermore, observe that K^* admits as SVD $P_0((D_0^{(d')})^{\frac{1}{2}} D_1^{\frac{1}{2}})^{[d,d']} \widetilde{\mathrm{Id}}_{d'} P_1^T$. For a given fixed $\widetilde{\mathrm{Id}}_{d'}$, we get using Lemma 3.3.8, that the optimal P^* associated with K^* is $P = P_0 \widetilde{\mathrm{Id}}_{d'}^{[d,d']} P_1^T$, which concludes the proof. □

A.2.7 Proof of Proposition 4.3.6

Proof of Proposition 4.3.6. As before, the proof is inspired from the proof of the closed-form of the W_2 between two Gaussians provided by Givens et al. (1984). This time, we want to solve the following constrained optimization problem

$$\min_{\substack{\Sigma_1 - K^T \Sigma_0^{-1} K \in \mathbb{S}_+^{d'} \\ P \in \mathbb{V}_{d'}(\mathbb{R}^d)}} \mathrm{tr}(P^T \Sigma_0 P) - 2\mathrm{tr}(P^T K) \ .$$

Again, one can, using Lemma A.2.1, write $\mathrm{tr}(P^T K)$ as a function of B_r. This gives the following problem

$$\min_{\substack{B_r^T B_r = \mathrm{Id}_r \\ P^T P = \mathrm{Id}_{d'}}} \mathrm{tr}(P^T \Sigma_0 P) - 2\mathrm{tr}(P^T \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T) \ .$$

The Lagrangian of this latter problem reads as

$$\mathcal{L}(B_r, P, C_0, C_1) = \mathrm{tr}(P^T \Sigma_0 P) - 2\mathrm{tr}(P^T \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T) + \mathrm{tr}(C_0 (B_r^T B_r - \mathrm{Id}_r)) + \mathrm{tr}(C_1 (P^T P - \mathrm{Id}_{d'})) \ ,$$

where $C_0 \in \mathbb{S}^r$ and $C_1 \in \mathbb{S}^{d'}$ are the Lagrange multipliers respectively associated with the constraints on B_r and P. The first order condition gives

$$\begin{cases} \Sigma_0^{\frac{1}{2}} P U_r \Lambda_r = B_r C_0 \\ \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^T = P C_1 + \Sigma_0 P \ . \end{cases}$$

Since Σ_0, P, U_r, and Λ_r are full rank, $\Sigma_0^{\frac{1}{2}} P U_r \Lambda_r$ is of rank r and so C_0 is also of rank r. Thus we get that

$$B_r = \Sigma_0^{\frac{1}{2}} P U_r \Lambda_r C_0^{-1} \ ,$$

and so

$$B_r^T B_r = \mathrm{Id}_r = C_0^{-1} \Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r C_0^{-1} .$$

Thus

$$C_0 = (\Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r)^{\frac{1}{2}} .$$

On the other hand, by reinjecting the expression of B_r in the other first order condition we get

$$P^T \Sigma_0 P U_r \Lambda_r (\Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r)^{-\frac{1}{2}} \Lambda_r U_r^T = C_1 + P^T \Sigma_0 P .$$

By multiplying this equation by itself, we get

$$P^T \Sigma_0 P U_r \Lambda_r^2 U_r^T = (C_1 + P^T \Sigma_0 P)^2 .$$

Since $(C_1 + P^T \Sigma_0 P)^2$ is symmetric, we get that $P^T \Sigma_0 P$ commutes with $U_r \Lambda_r^2 U_r^T$ and thus with $\Sigma_1 - S$. Moreover,

$$\begin{aligned}
\mathrm{tr}(P^T K) = \mathrm{tr}(C_0) &= \mathrm{tr}((\Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r)^{\frac{1}{2}}) \\
&= \mathrm{tr}(U_r^T U_r (\Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r)^{\frac{1}{2}}) \\
&= \mathrm{tr}(U_r (\Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r)^{\frac{1}{2}} U_r^T) \\
&= \mathrm{tr}((U_r \Lambda_r U_r^T P^T \Sigma_0 P U_r \Lambda_r U_r^T)^{\frac{1}{2}}) \\
&= \mathrm{tr}(((\Sigma_1 - S)^{\frac{1}{2}} P^T \Sigma_0 P (\Sigma_1 - S)^{\frac{1}{2}})^{\frac{1}{2}}) .
\end{aligned}$$

As before, using the Courant-Fischer min-max theorem (Courant, 1920; Fischer, 1905) to characterize the eigenvalues of $\Sigma_1 - S$, we get that $\mathrm{tr}(P^T K)$ is maximized when $S = 0$. Thus we get that

$$PW_2^2(\mu, \nu) = \min_{P^T \Sigma_0 P \Sigma_1 = \Sigma_1 P^T \Sigma_0 P} \mathrm{tr}(P^T \Sigma_0 P) + \mathrm{tr}(\Sigma_1) - \mathrm{tr}((\Sigma_1^{\frac{1}{2}} P^T \Sigma_0 P \Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}) .$$

Since $P^T \Sigma_0 P$ and Σ_1 commute, this expression can be reduced to

$$\begin{aligned}
PW_2^2(\mu, \nu) &= \min_{P^T \Sigma_0 P \Sigma_1 = \Sigma_1 P^T \Sigma_0 P} \mathrm{tr}\left(\left((P^T \Sigma_0 P)^{\frac{1}{2}} - \Sigma_1^{\frac{1}{2}}\right)^2\right) \\
&= \min_{P^T \Sigma_0 P \Sigma_1 = \Sigma_1 P^T \Sigma_0 P} \| D_{0,P}^{\frac{1}{2}} - D_1^{\frac{1}{2}} \|_{\mathcal{F}}^2 ,
\end{aligned} \tag{A.4}$$

where (P_1, D_1) is any diagonalization of Σ_1 $(= P_1 D_1 P_1^T)$, and where $D_{0,P} = P_1^T P^T \Sigma_0 P P_1$. Thus, the problem is reduced to find the $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ such that the eigenvalues of $(P^T \Sigma_0 P)^{\frac{1}{2}}$ are the closest possible (in term of $l2$ distance) to the eigenvalues of $\Sigma_1^{\frac{1}{2}}$. Let $(\alpha_1, \ldots, \alpha_d)^T$ and $(\beta_1, \ldots, \beta_{d'})^T$ denotes the respective eigenvalues of Σ_0 and Σ_1 sorted in decreasing order. Observe that for all $P \in \mathbb{V}_{d'}(\mathbb{R}^d)$ and for all $y \in \mathbb{R}^{d'}$, we have

$$\alpha_d \|y\|^2 = \alpha_d \|Py\|^2 \leq y^T P^T \Sigma_0 P y \leq \alpha_1 \|Py\|^2 = \alpha_1 \|y\|^2 ,$$

and so, for $k \leq d'$, if λ_k is an eigenvalue of $P^T \Sigma_0 P$, then

$$\alpha_d \leq \lambda_k \leq \alpha_1 .$$

(i) Suppose $\alpha_d > \beta_1$: since the eigenvalues of $P^T \Sigma_0 P$ are necessarily greater than α_d, we set the minimum eigenvalue of $P^T \Sigma_0 P$ to $\lambda_{d'} = \alpha_d$ in order to minimize (A.4). We denote $y_{d'} \in \mathbb{R}^{d'}$ the eigenvector associated. We have $y_{d'} P^T \Sigma_0 P y_{d'} = \alpha_d$ and $\|P y_{d'}\| = \|y_{d'}\| = 1$ so using Lemma A.2.3, we get that $\|P y_{d'}\|$ is an eigenvector of Σ_0 associated with α_d. Let λ_k and y_k be any other eigenvalue and its associated eigenvector in the orthonormal basis in which $P^T \Sigma_0 P$ is diagonal. We have $y_k^T y_{d'} = 0$ and so $y_k^T P^T P y_{d'} = 0$. Thus $P y_k$ is orthogonal to $P y_{d'}$. Since $\|P y_k\| = 1$, we get that $P y_k$ is also an eigenvector of Σ_0 and so it exists $i \leq d - 1$ such that $\lambda_k = y_k^T P^T \Sigma_0 P y_k = \alpha_i$. Thus, we conclude that the eigenvalues of the optimal $P^T \Sigma_0 P$ are the d' smallest eigenvalues of Σ_0. Now we determine their order. We have

$$\| D_{0,P}^{\frac{1}{2}} - D_1^{\frac{1}{2}} \|_{\mathcal{F}}^2 = \mathrm{tr}(D_{0,P}) + \mathrm{tr}(D_1) - 2\mathrm{tr}(D_{0,P}^{\frac{1}{2}} D_1^{\frac{1}{2}}) .$$

Since only $\mathrm{tr}(D_{0,P}^{\frac{1}{2}} D_1^{\frac{1}{2}})$ depends of the order of the eigenvalues, one can observe that setting $D_{0,P} = D_{0\uparrow}^{(d')}$ and $D_1 = D_{1\uparrow}$ minimizes the expression. The optimal value is achieved for any P of the form

$$P^* = P_{0\uparrow} \widetilde{\mathrm{Id}}_{d'}^{[d,d']} P_{1\uparrow}^T ,$$

Since for a given P, the expression of the optimal map which minimizes $W_2(P_{\#}^T \mu, \nu)$ is given by, for all $x \in \mathbb{R}^{d'}$,

$$T_{W_2}(x) = (P^T \Sigma_0 P)^{-1} (P^T \Sigma_0 P \Sigma_1)^{\frac{1}{2}} x .$$

Re-injecting P^* in the above gives the expression of T in that case.

(ii) Suppose that $\alpha_1 < \beta_d$: since the eigenvalues of $P^T \Sigma_0 P$ are necessarily smaller than α_1, we set $\lambda_1 = \alpha_1$ in order to minimize (A.4). We denote $y_1 \in \mathbb{R}^{d'}$ the eigenvector associated. We have $y_1 P^T \Sigma_0 P y_1 = \alpha_1$ and $\|P y_1\| = \|y_1\| = 1$ so using Lemma A.2.3, we get that $\|P y_1\|$ is an eigenvector of Σ_0 associated with α_1. Applying the same reasoning as before, we get that the eigenvalues of the optimal $P^T \Sigma_0 P$ are the d' largest eigenvalues of Σ_0. Thus, setting $D_{0,P} = D_{0\downarrow}^{(d')}$ and $D_1 = D_{1\downarrow}$ minimizes (A.4). This is achieved for any P of the form

$$P^* = P_{0\downarrow} \widetilde{\mathrm{Id}}_{d'}^{[d,d']} P_{1\downarrow}^T .$$

By re-injecting this in the expression of T_{W_2}, we get the expression of T in that case, which concludes the proof.

□

A.3 Proofs of the claims of Chapter 5

A.3.1 Proof of Lemma 5.4.2

Proof of Lemma 5.4.2. First note that in this proof, we denote $\mathbb{R}^{d \times d'}$ the set of matrices of size $d \times d'$ that we distinct of the set $\mathbb{R}^{dd'}$ of vector with $d \times d'$ coordinates. We set $g : P \in \mathbb{R}^{d \times d'} \mapsto \Sigma_1^{\frac{1}{2}} P^T \Sigma_0 P \Sigma_1^{\frac{1}{2}}$ and $h : Q \in \mathbb{S}_+^{d'} \mapsto Q^{\frac{1}{2}}$ such that for all matrix P of size $d \times d'$, we have

$$f(P) = \mathrm{tr}(h(g(P))) .$$

For any matrix $A \in \mathbb{R}^{d \times d'}$, we denote $\mathrm{vec}(A) \in \mathbb{R}^{dd'}$ the vector obtained by stacking the columns of A. Observe that, see (Magnus and Neudecker, 2019) for details, for any function $\phi : \mathbb{R}^{d \times d'} \to \mathbb{R}^{r \times s}$, the Jacobian matrix $J[\phi]$ of ϕ can be defined as, for all $P \in \mathbb{R}^{d \times d'}$,

$$J[\phi](P) = \frac{\partial \mathrm{vec}(f(P))}{\partial \mathrm{vec}(P)} .$$

Moreover, observe that since $f : \mathbb{R}^{d \times d'} \to \mathbb{R}$, $J[f][P] \in \mathbb{R}^{dd'}$ and

$$\frac{\partial f(P)}{\partial P} = \mathrm{vec}^{-1}(J^T[f](P)) ,$$

where vec^{-1} is the inverse vector operator, i.e. such that for any $A \in \mathbb{R}^{d \times d'}$, $\mathrm{vec}^{-1}(\mathrm{vec}(A)) = A$. Applying the chain rule to derive f, we have

$$J[f](P) = J[\mathrm{tr}]((h \circ g)(P)) J[h](g(P)) J[g](P) .$$

- First, we compute $J[g](P)$. It follows, using formula provided by Petersen et al. (2008) and Magnus and Neudecker (2019),

$$\partial(\Sigma_1^{\frac{1}{2}} P^T \Sigma_0 P \Sigma_1^{\frac{1}{2}}) = \Sigma_1^{\frac{1}{2}} \partial P^T \Sigma_0 P \Sigma_1^{\frac{1}{2}} + \Sigma_1^{\frac{1}{2}} P^T \Sigma_0 \partial P \Sigma_1^{\frac{1}{2}},$$

and so

$$\partial \mathrm{vec}(\Sigma_1^{\frac{1}{2}} P^T \Sigma_0 P \Sigma_1^{\frac{1}{2}}) = (\Sigma_1^{\frac{1}{2}} P^T \Sigma_0 \otimes_K \Sigma_1^{\frac{1}{2}}) \partial \mathrm{vec}(P^T) + (\Sigma_1^{\frac{1}{2}} \otimes_K \Sigma_1^{\frac{1}{2}} P^T \Sigma_0) \partial \mathrm{vec}(P)$$

$$= (\Sigma_1^{\frac{1}{2}} P^T \Sigma_0 \otimes_K \Sigma_1^{\frac{1}{2}}) K_{dd'} \partial \mathrm{vec}(P) + (\Sigma_1^{\frac{1}{2}} \otimes_K \Sigma_1^{\frac{1}{2}} P^T \Sigma_0) \partial \mathrm{vec}(P)$$
$$= (I_{d'^2} + K_{d'^2})(\Sigma_1^{\frac{1}{2}} \otimes_K \Sigma_1^{\frac{1}{2}} P^T \Sigma_0) \partial \mathrm{vec}(P) \ ,$$

where $\otimes_K$ denotes the Kronecker product and for any r, K_r is the commutation matrix of size $r \times r$, see (Magnus and Neudecker, 2019) for details. Thus,

$$J[g](P) = (I_{d'^2} + K_{d'^2})(\Sigma_1^{\frac{1}{2}} \otimes_K \Sigma_1^{\frac{1}{2}} P^T \Sigma_0) \ .$$

- Now we compute $J[h](Q)$. Observe that we have for any $Q \in \mathbb{S}_+^{d'}$,

$$Q^{\frac{1}{2}} Q^{\frac{1}{2}} = Q \ .$$

Thus it follows, denoting $s : Q \mapsto Q^{1/2}$,

$$\partial s(Q) Q^{\frac{1}{2}} + Q^{\frac{1}{2}} \partial s(Q) = \partial Q \ .$$

This latter equation is a Sylvester equation with variable $\partial s(Q)$, which is equivalent to the following linear system:

$$(Q^{\frac{1}{2}} \oplus_K Q^{T\frac{1}{2}}) \partial \mathrm{vec}(s(Q)) = \partial \mathrm{vec}(Q) \ ,$$

where $\oplus_K$ stands for the Kronecker sum. If Q is non-degenerate, $Q^{\frac{1}{2}} \oplus_K Q^{T\frac{1}{2}}$ is also non-degenerate and so in that case

$$J[h](Q) = (Q^{\frac{1}{2}} \oplus_K Q^{T\frac{1}{2}})^{-1} \ .$$

- Finally, it is easy to see that for $R \in \mathbb{R}^{d' \times d'}$ we have

$$J[\mathrm{tr}](R) = \mathrm{vec}^T(\mathrm{Id}_{d'}).$$

Thus, denoting $A = \Sigma_1^{\frac{1}{2}} P^T \Sigma_0 P \Sigma_1^{\frac{1}{2}}$ and observing that A is symmetric and full-rank when P is full-rank (since we supposed that Σ_0 and Σ_1 are full rank), it follows that for all full-rank matrix P of size $d \times d'$,

$$J^T[f](P) = (\Sigma_1^{\frac{1}{2}} \otimes_K \Sigma_0 P \Sigma_1^{\frac{1}{2}})(I_{d'^2} + K_{d'^2})(A^{\frac{1}{2}} \oplus_K A^{\frac{1}{2}})^{-1} \mathrm{vec}(\mathrm{Id}_{d'}) \ ,$$

where we used that $K_{d'^2}$ and $(A \oplus_K A)^{-1}$ were symmetric. Observe now that $(A^{\frac{1}{2}} \oplus_K A^{\frac{1}{2}})^{-1} \mathrm{vec}(\mathrm{Id}_{d'}) = \mathrm{vec}(X)$, where $X \in \mathbb{R}^{d' \times d'}$ is the unique solution of the following Sylvester equation

$$A^{\frac{1}{2}} X + X A^{\frac{1}{2}} = \mathrm{Id}_{d'} \ .$$

Since A is symmetric, one can set $A = QDQ^T$ where $Q \in \mathbb{O}(\mathbb{R}^{d'})$ and D is a diagonal matrix of size d'. The Sylvester equation can be rewritten

$$D^{\frac{1}{2}} Y + Y D^{\frac{1}{2}} = \mathrm{Id}_{d'} \ ,$$

where $Y = Q^T X Q$. Since A is full-rank, D is invertible and it is easy to see that the unique solution of this latter equation is $Y = (1/2) D^{-\frac{1}{2}}$ and so $X = (1/2) A^{-\frac{1}{2}}$ and thus

$$(A^{\frac{1}{2}} \oplus_K A^{\frac{1}{2}})^{-1} \mathrm{vec}(\mathrm{Id}_{d'}) = \frac{1}{2} \mathrm{vec}(A^{-\frac{1}{2}}) \ .$$

Moreover, since A is symmetric, we have $K_{d'^2} \mathrm{vec}(A^{-\frac{1}{2}}) = \mathrm{vec}(A^{-\frac{1}{2}})$ and so it follows that

$$J^T[f](P) = (\Sigma_1^{\frac{1}{2}} \otimes_K \Sigma_0 P \Sigma_1^{\frac{1}{2}}) \mathrm{vec}(A^{-\frac{1}{2}})$$
$$= \mathrm{vec}(\Sigma_0 P \Sigma_1^{\frac{1}{2}} A^{-\frac{1}{2}} \Sigma_1^{\frac{1}{2}}) \ ,$$

which concludes the proof. □

Appendix B

Supplementary materials of Part II

Contents

B.1 Proofs of the claims of Chapter 7

B.1.1 Proof of Corollary 7.3.7

To prove Corollary 7.3.7, we will first need to prove the following result.

Lemma B.1.1. *Let* $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ *and* $r > 0$. *We denote* $\mathsf{B} = (\mathsf{A}_r)^c$. *Then*

$$\mathsf{B}_r \subset \bar{\mathsf{A}^c} \,,$$

where $\bar{\mathsf{A}^c}$ *denotes the closure of the complementary of* A.

Proof. Let $x \in \mathsf{B}_r$. There exists $b \in \mathsf{B}$ such that $\|x - b\| \leq r$. Moreover, since $\mathsf{B} = (\mathsf{A}_r)^c$, it follows that for all $a \in \mathsf{A}$,

$$\|b - a\| > r \,.$$

Then

$$r < \|b - x\| + \|x - a\| \,,$$

and so, it follows that for all $a \in \mathsf{A}$,

$$\|x - a\| > 0 \,.$$

Thus $x \in \bar{\mathsf{A}^c}$. □

Now we are ready to turn to the proof of Corollary 7.3.7.

Proof of Corollary 7.3.7. We set $r = d(\mathsf{M}_1, \mathsf{M}_2)/2$ and $\mathsf{A} = (\mathsf{M}_1)_r$. Using Theorem 7.3.6, we have

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \alpha_g(\mathsf{A}, r) - \min\{g_{\#}\mu_{d'}(\mathsf{A}), \nu(\mathsf{A})\} - \nu(\mathsf{A}_r \setminus \mathsf{A}) \,.$$

First we suppose that $g_{\#}\mu_{d'}(\mathsf{A}) \geq \nu(\mathsf{A})$: since Φ is a non-decreasing function, it follows that

$$\alpha_g(\mathsf{A}, r) = \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{A}))\right) \geq \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(\nu(\mathsf{A}))\right) \,.$$

Moreover, $\min\{g_{\#}\mu_{d'}(\mathsf{A}), \nu(\mathsf{A})\} = \nu(\mathsf{A}) = \lambda = \Phi(\Phi^{-1}(\lambda))$ and so it follows

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \Phi\left(\tfrac{d(\mathsf{M}_1,\mathsf{M}_2)}{2\mathrm{Lip}(g)} + \Phi^{-1}(\lambda)\right) - \Phi(\Phi^{-1}(\lambda)) \geq \textstyle\int_{\Phi^{-1}(\lambda)}^{r/\mathrm{Lip}(g)+\Phi^{-1}(\lambda)} \varphi(t)dt \;,$$

since ν has no mass on $\mathsf{A}_r \setminus \mathsf{A}$. Now we suppose that $g_{\#}\mu_{d'}(\mathsf{A}) \leq \nu(\mathsf{A})$: we then set $\mathsf{B} = \mathsf{A}^c$. Since $g_{\#}\mu_{d'}(\mathsf{A}) \leq \nu(\mathsf{A})$, we have $g_{\#}\mu_{d'}(\mathsf{B}) \geq \nu(\mathsf{B})$. Applying Theorem 7.3.6, and the same reasoning as before we get

$$\begin{aligned} D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) &\geq \alpha_g(\mathsf{B}, r) - \min\{g_{\#}\mu_{d'}(\mathsf{B}), \nu(\mathsf{B})\} - \nu(\mathsf{B}_r \setminus \mathsf{B}) \\ &\geq \Phi\left(\frac{d(\mathsf{M}_1,\mathsf{M}_2)}{2\mathrm{Lip}(g)} + \Phi^{-1}(1-\lambda)\right) - \Phi(\Phi^{-1}((1-\lambda)) - \nu(\mathsf{B}_r \setminus \mathsf{B}) \;. \end{aligned}$$

Using Lemma B.1.1, we get that $\nu(\mathsf{B}_r \setminus \mathsf{B}) \leq \nu(\bar{\mathsf{A}^c} \setminus (\mathsf{A}_r)^c)$ but $\nu(\bar{\mathsf{A}^c} \setminus (\mathsf{A}_r)^c) = 0$ since ν has no mass on $\bar{\mathsf{A}^c} \setminus (\mathsf{A}_r)^c$ except on its boundary and so its follows that

$$\begin{aligned} D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) &\geq \Phi\left(\tfrac{d(\mathsf{M}_1,\mathsf{M}_2)}{2\mathrm{Lip}(g)} + \Phi^{-1}(1-\lambda)\right) - \Phi(\Phi^{-1}((1-\lambda)) \\ &\geq \Phi\left(\tfrac{d(\mathsf{M}_1,\mathsf{M}_2)}{2\mathrm{Lip}(g)} - \Phi^{-1}(\lambda)\right) - \Phi(-\Phi^{-1}(\lambda)) \\ &\geq \int_{-\Phi^{-1}(\lambda)}^{r/\mathrm{Lip}(g)-\Phi^{-1}(\lambda)} \varphi(t)dt \;, \end{aligned}$$

since $\Phi^{-1}(1-\lambda) = -\Phi^{-1}(\lambda)$. Since $\lambda \geq 1/2$, it follows that $\Phi^{-1}(\lambda) \geq 0$ and so

$$\int_{-\Phi^{-1}(\lambda)}^{r/\mathrm{Lip}(g)-\Phi^{-1}(\lambda)} \varphi(t)dt \geq \int_{\Phi^{-1}(\lambda)}^{r/\mathrm{Lip}(g)+\Phi^{-1}(\lambda)} \varphi(t)dt \;,$$

which concludes the proof. □

B.1.2 Proof of Corollary 7.3.8

Proof of Corollary 7.3.8. As previously, we prove the corollary when $\nu = \frac{1}{2}[\mathrm{N}(-m, \sigma^2 \mathrm{Id}_d) + \mathrm{N}(m, \sigma^2 \mathrm{Id}_d)]$ since the problem can always be reduced to that case by translation and setting $m = (m_2 - m_1)/2$. Since the problem is invariant by rotation, we can assume without any loss of generality that $m = (\|m\|, 0, \ldots, 0)$. Let H be the half-space of $\mathbb{R}^d$ defined by $\mathsf{H} = (-\infty, 0] \times \mathbb{R}^{d-1}$ and we set $r = \|m\|/2\sigma$. First we suppose that $g_{\#}\mu_{d'}(\mathsf{H}) \geq \nu(\mathsf{H})$: using Theorem 7.3.6, we get that

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \alpha_g(\mathsf{H}, r) - \min\{g_{\#}\mu_{d'}(\mathsf{H}), \nu(\mathsf{H})\} - \nu(\mathsf{H}_r \setminus \mathsf{H}) \;,$$

with $\mathsf{H}_r = (-\infty, \|m\|/2\sigma] \times \mathbb{R}^{d-1}$. On one hand we have that $\nu = \nu_1 \otimes \mathrm{N}(0, \sigma^2 \mathrm{Id}_{d-1})$, where $\nu_1 = \frac{1}{2}[\mathrm{N}(-\|m\|, \sigma^2) + \mathrm{N}(\|m\|, \sigma^2)]$ and so $\nu(\mathsf{H}_r \setminus \mathsf{H}) = \nu_1([0, \|m\|/2\sigma])$. On the other hand we have that $\min\{g_{\#}\mu_{d'}(\mathsf{H}), \nu(\mathsf{H})\} = \nu(\mathsf{H})$ and $g_{\#}\mu_{d'}(\mathsf{H}) \geq 1/2$ since $g_{\#}\mu_{d'}(\mathsf{H}) \geq \nu(\mathsf{H})$. Hence it follows that

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \Phi(r/\mathrm{Lip}(g)) - \tfrac{1}{2} - \nu_1([0, \|m\|/2\sigma]) \;.$$

Now we suppose that $g_{\#}\mu_{d'}(\mathsf{H}) \leq \nu(\mathsf{H})$: we then set $\mathsf{H}_2 = (0, +\infty] \times \mathbb{R}^{d-1}$. Since $g_{\#}\mu_{d'}(\mathsf{H}) \leq 1/2$, we get that $g_{\#}\mu_{d'}(\mathsf{H}_2) \geq 1/2$ and so $g_{\#}\mu_{d'}(\mathsf{H}_2) \geq \nu(\mathsf{H}_2)$. Hence we retrieve the previous case and so it follows that

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \Phi(r/\mathrm{Lip}(g)) - \tfrac{1}{2} - \nu_1([-\|m\|/2\sigma, 0]) \;.$$

Since $\nu_1([-\|m\|/2\sigma, 0]) = \nu_1([0, \|m\|/2\sigma])$, we get in both cases

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \Phi(r/\mathrm{Lip}(g)) - \tfrac{1}{2} - \nu_1([0, \|m\|/2\sigma]) \;.$$

Now we derive the value of $\nu_1([0, \|m\|/2\sigma])$:

$$\begin{aligned} \nu_1([0, \|m\|/2\sigma]) = \frac{1}{2}\int_0^{m/2\sigma} (2\pi\sigma^2)^{-1/2} \exp[-(x+m)^2/2\sigma^2]dx \\ + \frac{1}{2}\int_0^{m/2\sigma} (2\pi\sigma^2)^{-1/2} \exp[-(x-m)^2/2\sigma^2]dx \end{aligned}$$

$$= \frac{1}{2}\int_{-m/2\sigma}^{m/2\sigma} (2\pi\sigma^2)^{-1/2}\exp[-(x+m)^2/2\sigma^2]dx$$
$$= \frac{1}{2}\int_{\|m\|(2\sigma-1)/2\sigma^2}^{\|m\|(2\sigma+1)/2\sigma^2} \varphi(x)dx \ ,$$

which concludes the proof. □

B.1.3 Proof of Corollary 7.3.10

Proof of Corollary 7.3.10. As previously, we prove the corollary when $\nu = \frac{1}{2}[\mathrm{N}(-m, \sigma^2\,\mathrm{Id}_d)+\mathrm{N}(m, \sigma^2\,\mathrm{Id}_d)]$ since the problem can always be reduced to that case by translation and setting $m = (m_2-m_1)/2$. Since the problem is invariant by rotation, we can assume without any loss of generality that $m = (\|m\|, 0, \dots, 0)$. Furthermore, observe that the half-space $\{(m_2-m_1)^T\,(x-(m_1+m_2)/2) \le 0 \ : \ x \in \mathbb{R}^d\}$ becomes $(-\infty, 0]\times\mathbb{R}^{d-1}$ in that case, and that the condition $\lambda \in (0, 1/2]$ is indeed non-restrictive since the problem is invariant by rotation. We set as before $\mathsf{H} = (-\infty, 0]\times\mathbb{R}^{d-1}$ and $r = \|m\|/2\sigma$. Applying Theorem 7.3.9, we get

$$D_{\mathrm{KL}}(g_{\#}\mu_{d'}||\nu) \ge \beta_g(\mathsf{H}, r)\log\left(\frac{\beta_g(\mathsf{H},r)}{\nu(\mathsf{H}_r\setminus\mathsf{H})}\right) + (1-\beta_g(\mathsf{H},r))\log\left(\frac{1-\beta_g(\mathsf{H},r)}{1-\nu(\mathsf{H}_r\setminus\mathsf{H})}\right) \ .$$

On one hand, we get

$$\begin{aligned}
\beta_g(\mathsf{H}, r) &= \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{H})\right) - g_{\#}\mu_{d'}(\mathsf{H}) \\
&= \Phi\left(\tfrac{r}{\mathrm{Lip}(g)} + \Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{H})\right) - \Phi\left(\Phi^{-1}(g_{\#}\mu_{d'}(\mathsf{H})\right) \\
&= \int_{\Phi^{-1}(\lambda)}^{\|m\|/2\sigma\mathrm{Lip}(g)+\Phi^{-1}(\lambda)} \varphi(t)\mathrm{d}t \\
&= \int_{-\Phi^{-1}(1-\lambda)}^{\|m\|/2\sigma\mathrm{Lip}(g)-\Phi^{-1}(1-\lambda)} \varphi(t)\mathrm{d}t \ ,
\end{aligned}$$

denoting $\lambda = g_{\#}\mu_{d'}(\mathsf{H})$. We replaced $\Phi^{-1}(\lambda)$ by $-\Phi^{-1}(1-\lambda)$ in order to emphasize that $\Phi^{-1}(\lambda) \le 0$ since $\lambda \le 1/2$. Observe that if we supposed $\lambda \ge 1/2$, we would have $\beta_g(\mathsf{H}^c, r) \ge \beta_g(\mathsf{H}, r)$ and so the bound that we would have found by reasoning on H would have been sub-optimal. On the other hand, observing as before that $\nu = \nu_1 \otimes \mathrm{N}(0, \sigma^2\,\mathrm{Id}_{d-1})$, where $\nu_1 = \frac{1}{2}[\mathrm{N}(-\|m\|, \sigma^2) + \mathrm{N}(\|m\|, \sigma^2)]$, we get that

$$\begin{aligned}
\nu(\mathsf{H}_r \setminus \mathsf{H}) &= \nu_1([0, \|m\|/2\sigma]) \\
&= \frac{1}{2}\int_{\|m\|(2\sigma-1)/2\sigma^2}^{\|m\|(2\sigma+1)/2\sigma^2} \varphi(t)\mathrm{d}t \ ,
\end{aligned}$$

which concludes the proof. □

B.2 Additional theoretical result

In this section we derive a generalization of Corollary 7.3.7 when ν is a distribution whose support is composed of more than two disconnected manifolds.

Proposition B.2.1. *Let ν be a measure on $\mathbb{R}^d$ on N disconnected manifolds $(\mathsf{M}_1, \dots, \mathsf{M}_N)$, and let $g : \mathbb{R}^{d'} \to \mathbb{R}^d$ be a Lipschitz function. Then,*

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \ge \max_{I\subset[\![1,N]\!]} \int_{\Phi^{-1}(\lambda)}^{d(\bigsqcup_{i\in I}\mathsf{M}_i,\ \bigsqcup_{j\in[\![1,N]\!]\setminus I}\mathsf{M}_j)/2\mathrm{Lip}(g)+\Phi^{-1}(\lambda)} \varphi(t)dt \ ,$$

where for $\mathsf{A}, \mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$, $d(\mathsf{A}, \mathsf{B}) = \inf\{\|a-b\| \ : \ a \in \mathsf{A}, b \in \mathsf{B}\}$, and $\lambda = \nu\left(\bigsqcup_{i\in I}\mathsf{M}_i\right)$ if $\nu\left(\bigsqcup_{i\in I}\mathsf{M}_i\right) \ge 1/2$ and $\lambda = 1-\nu\left(\bigsqcup_{i\in I}\mathsf{M}_i\right)$ otherwise.

Proof. Let $I \subset [\![1, N]\!]$. First, we suppose that $\nu\left(\bigsqcup_{i\in I} \mathsf{M}_i\right) \geq 1/2$. Since ν can be seen as a bi-modal distribution on the two disconnected sets $\bigsqcup_{i\in I} \mathsf{M}_i$ and $\bigsqcup_{j\in[\![1,N]\!]\setminus I} \mathsf{M}_j$, we can apply Corollary 7.3.7. Thus we get

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) \geq \int_{\Phi^{-1}(\lambda)}^{d(\bigsqcup_{i\in I} \mathsf{M}_i,\ \bigsqcup_{j\in[\![1,N]\!]\setminus I} \mathsf{M}_j)/2\mathrm{Lip}(g)+\Phi^{-1}(\lambda)} \varphi(t)dt \ .$$

If $\nu\left(\bigsqcup_{i\in I} \mathsf{M}_i\right) \leq 1/2$, we can still apply Corollary 7.3.7 by interchanging the roles of $\bigsqcup_{i\in I} \mathsf{M}_i$ and $\bigsqcup_{j\in[\![1,N]\!]\setminus I} \mathsf{M}_j$, thus we get also Inequality (B.2) in that case, which concludes the proof. □

B.3 Experimental details

We detail our experiments in dimension 1 in Appendix B.3.1. In Appendix B.3.2, we give details on our experiment on the synthetic mixture of two Gaussians derived from MNIST. Finally, we detail the experiment on the subset of all 3 and 7 of MNIST in Appendix B.3.3. We trained our models using 2 NVIDIA Titan Xp from the proprietary server of our institution with an estimated total training time of approximately 175 GPU hours. Code is available here [1].

B.3.1 Univariate case

In the univariate case we use a simple 3-layer Multi Layer Perceptron (MLP) of shape $(1, 128, 256, 1)$ as decoder for the VAE and as generator for the GAN. The network has a total of 33537 learnable parameters. The score network uses also a a 3-layer MLP block, this time of shape $(1, 96, 196, 1)$, in which at each layer is injected the noise information transformed by a positional encoding (Vaswani et al., 2017) and then by another MLP block size $(16, 32, 64)$, see Figure B.1. The score network has a total of 34665 learnable parameters. In all three models, we use LeakyReLU (Maas et al., 2013) as non-linearity with a negative slope of 0.2. The three models are trained during 400 epochs with a batch size of 1000 using ADAM (Kingma and Ba, 2015) with a momentum of 0.9 and a learning rate of 10^{-4}. In the following, we give more specific details for each model.

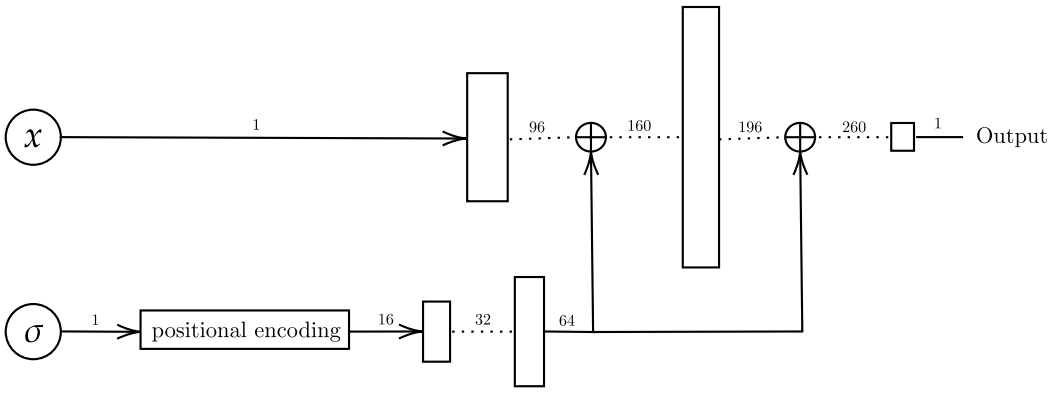

Figure B.1: Architecture of the score network used for the univariate experiments. The "positional encoding" block applies the sine transform described in Vaswani et al. (2017). $\bigoplus$ corresponds to concatenation, the vertical blocks correspond to the fully connected layers and the numbers over the arrows correspond to the size of the vectors.

Variational autoencoder. We use the vanilla VAE model as described in Kingma and Welling (2014). In the following, we denote θ and λ the respective parameters of the decoder and the encoder. The decoder f_ϕ is composed of an MLP block of size $(1, 256, 128)$ followed by two parallel fully connected layers of shape $(128, 1)$ which gives two outputs $f_{1\phi}(x)$ and $f_{2\phi}(x)$. Then the input z of the decoder

[1] https://github.com/AntoineSalmona/Push-forward-Generative-Models

g_θ is obtained by the so-called reparametrization trick, which consists in sampling $z \sim q_\phi^{z|x}$, where $q_\phi^{z|x} = \mathrm{N}(f_{1\phi}(x), \exp[f_{2\phi}(x)])$. During training, the model minimizes the following loss function:

$$\mathcal{L}_{\mathrm{VAE}}(\theta, \phi) = \mathbb{E}_{x\sim\nu}[\mathrm{ELBO}_{\theta,\phi}(x, q_\phi^{z|x}, p_\theta^{x|z})] \,,$$

where $p_\theta^{x|z} = \mathrm{N}(g_\theta(z), c^2 \,\mathrm{Id}_d)$ and ELBO is the Evidence Lower Bound (Blei et al., 2017), defined as follows:

$$\mathrm{ELBO}_{\theta,\phi}(x, q_\phi^{z|x}, p_\theta^{x|z}) = \mathbb{E}_{z\sim q_\phi^{z|x}}[\log(p_\theta(x|z))] - D_{\mathrm{KL}}(q_\phi^{z|x}||\mathrm{N}(0, \mathrm{Id}_{d'})) \,.$$

The standard deviation c in $p_\theta^{x|z}$ is an hyperparameter of the model. For our experiments, we observed that $c = 0.1$ gave good results.

Generative adversarial network. As for the VAE, we use the vanilla GAN model as described in Goodfellow et al. (2014). The discriminator is 4-layer MLP of shape $(1, 512, 256, 128, 1)$ with spectral normalization (Miyato et al., 2018) in order to reduce as much as possible mode collapse. We train the model using the vanilla adversarial loss, that the discriminator d_ϕ tries to maximize and that the generator g_θ tries to minimize

$$\mathcal{L}_{\mathrm{GAN}}(\theta, \phi) = \mathbb{E}_{x\sim\nu}[\log(d_\phi(x)] + \mathbb{E}_{z\sim\mathrm{N}(0,\mathrm{Id}_{d'})}[\log(1 - d_\phi(g_\theta(z)))] \,.$$

We also tried with the hinge version of the adversarial loss, as proposed in Lim and Ye (2017) and Tran et al. (2017) and we obtained similar results.

Score-based generative modeling. Our diffusion model is similar to the model introduced by Song and Ermon (2019). The neural network s_θ learns to approximate, for a given x and a given σ, the score $\nabla_x p_\nu(x, \sigma)$ of the data distribution convoluted with a Gaussian distribution of standard deviation σ. This is done by first defining a geometrical progression $\{\sigma_i\}_{i=1}^L$ where $L = 10$ and where the ratio is chosen such that $\sigma_L \approx 0.01$, and then minimizing the Fischer divergence (Vincent, 2011)

$$\mathcal{L}_{\mathrm{SGM}}(\theta) = \mathbb{E}_{\sigma\sim 1/L\sum\delta_{\sigma_i}}\left[\sigma^2 \mathbb{E}_{x\sim\nu}\left[\mathbb{E}_{y\sim\mathrm{N}(x,\sigma^2\,\mathrm{Id}_d)}\left[\left\|s_\theta(y,\sigma) + (y-x)/\sigma^2\right\|^2\right]\right]\right] .$$

Then, in order to generate data, we use an annealed Langevin dynamic scheme as defined in Song and Ermon (2019). In the Langevin dynamic, we set the step size to 2×10^{-5} and the number of step for each value of σ to 100 as in Song and Ermon (2019).

Influence of generator depth. For this experiment, we increase the number of layers of the VAE decoder and the GAN generator from 2 to 6. At each new layer, we double the number of neurons at the previous layer. For instance, the generative network with 2 layers is thus an MLP of shape $(1, 128, 1)$ and the one with 6 layers is an MLP of shape $(1, 128, 256, 512, 1024, 2048, 1)$. Specifically to the GAN model, we also increase the number of layers in the discriminator in order to keep the dynamic between this latter and the generator balanced. As in the 3-layers case, the discriminator is one layer deeper than the generator. For instance, the discriminator associated to the generator with 2 layers is an MLP of shape $(1, 256, 128, 1)$.

Influence of generator architecture. For this experiment, we use a feed-forward MLP of shape $(1, 256, 256, 256, 1)$ as backbone. Then we add two additive pre-activation skip-connections of type "resnet" between the first and the second hidden layers and between the second and the third hidden layers. Finally, we replace the two previous additive skip-connections of type "resnet" by concatenation pre-activation skip-connections of type "densenet".

B.3.2 Synthetic mixture of Gaussians on MNIST

Models details. We adapt our three models to MNIST, changing mainly the networks architectures and making small modifications that we describe in what follows. We base the architecture of the GAN and the VAE on DCGAN (Radford et al., 2015), using the generator as decoder and the discriminator as encoder for our VAE. This is done by doubling the last layer of the discriminator in order that the VAE encoder has two outputs as in the univariate case. For the GAN model, we replaced the convolutional discriminator by a simple MLP of shape $(784, 512, 256, 128, 1)$ because the dynamic between the generator

and the discriminator seemed unbalanced otherwise. We also update our GAN model using some features of SAGAN (Zhang et al., 2019): applying spectral normalization on the discriminator and using the unconditional hinge version of the adversarial loss (Lim and Ye, 2017; Tran et al., 2017):

$$\mathcal{L}_{\mathrm{GAN}}^{d_\phi} = -\mathbb{E}_{x\sim\nu}[\min\{0, -1 + d_\phi(x)\}] - \mathbb{E}_{z\sim\mathrm{N}(0,\mathrm{Id}_{d'})}[\min\{0, -1 - d_\phi(g_{\theta(z)}\}] \,,$$
$$\mathcal{L}_{\mathrm{GAN}}^{g_\theta} = -\mathbb{E}_{z\sim\mathrm{N}(0,\mathrm{Id}_{d'})}[d_\phi(g_\theta(z))] \,.$$

Such loss function is equivalent to minimize the Kullback-Leibler divergence between the generated distribution and the data distribution. The VAE decoder and the GAN generator have 1713088 learnable parameters. For the score network architecture, we use the vanilla U-Net architecture (Ronneberger et al., 2015) in which we double the number of channels at each layer, we add group normalization (Wu and He, 2018) after each convolution and we replace the ReLU non-linearies by SiLU (Elfwing et al., 2018). As in the univariate case, we use positional encoding (Vaswani et al., 2017) followed by a MLP block of shape $(1, 16, 32)$ to incorporate the noise information at each layer. The score network has 1607392 learnable parameters. For inference, we use the same Langevin dynamic scheme as above with the same hyperparameters as in the univariate case. The three models are trained during 100 epochs with a batch size of 128 using ADAM with a momentum of 0.9 and a learning rate of 2×10^{-4}.

Additional details. The histograms of projection on the line passing through the mean of each Gaussians are obtained using 20000 generated samples. To assign a color to each bin of the histograms, we train a simple MLP of shape $(784, 1024, 50, 10)$ as classifier on MNIST. The classifier is trained during 10 epochs using again ADAM with a momentum of 0.9 and a learning rate of 2×10^{-4} and reaches an accuracy of 0.98 on the test set.

B.3.3 Subset of MNIST

Models details. Since the dataset is more complex than before, we use bigger models. For the score network, we use the architecture defined in Ho et al. (2020), in which we set the number of channels to 64 instead to 128 and we remove the self attention layers (Wang et al., 2018) for computational resource purposes. The score network has 6072065 learnable parameters. Again, we use an annealed Langevin dynamic scheme for inference with the same hyperparameters as before. For the VAE and the GAN, we use the same architecture as before, using this time the convolutional discriminator of DCGAN, and quadrupling the number of channels at each layer. This is mainly done in order to scale the generator/decoder to the score network. Hence the VAE decoder/GAN generator has 7151104 learnable parameters. We train all three models during 600 epochs with a batch size of 128 using ADAM with a momentum of 0.9 and a learning rate of 2×10^{-4}.

Additional details. We use the deep Wasserstein embedding proposed by Courty et al. (2018) in order to visualize histograms of projection in the Wasserstein space. We use the exact same network architecture and the same training procedure that in Courty et al. (2018): first, one million pairs of digits of MNIST are chosen randomly, in which 700000 are kept for the training set, 200000 for the test set, and 100000 for the validation set. We normalize each image in order to consider it as a two-dimensional distribution and we compute the 1-Wasserstein distance for each pair. Then, we train an autoencoder in a supervised manner in a way that the images at output of the autoencoder are close to the images in input, and that the euclidean distance between two vectors in the latent space is close to the 1-Wassertein distance between the two corresponding images of MNIST. As in Courty et al. (2018), the latent Wasserstein space is of dimension 50 and the autoencoder is trained during 100 epochs with a batch size of 100 and with an early stopping criterion. Again, we use ADAM with a momentum of 0.9 and a learning rate of 10^{-3}. We use the same classifier as before to assign color to each bin of the histograms. Finally, the histograms of projection on the line passing through the deep Wasserstein barycenters of all 3 and 7 are obtained using 20000 generated samples.

B.4 Additional experimental results

In the following, we provide additional experimental results. First, we compare estimates of the bounds of Theorem 7.3.6, Corollary 7.3.8, and Theorem 7.3.9 to estimates of the total variation distance and the Kullback-Leibler divergence in the univariate case. Then we study the possible correlation between

the size of the score network and the tendency of the score-based model to generate unbalanced modes. Finally, we provide additional visualizations of histograms of generated distributions for the univariate case and generated samples for the experiments on MNIST.

B.4.1 Bounds on TV distance and KL divergence in the univariate case

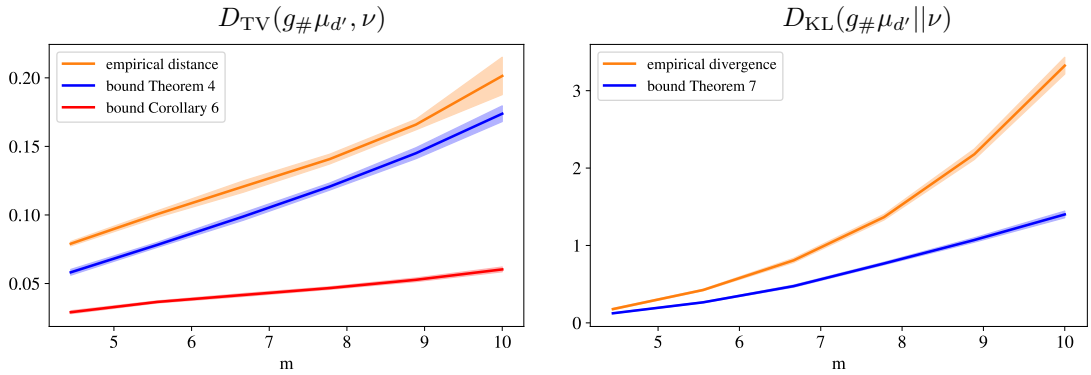

Figure B.2: total variation distance (left) and Kullback-Leibler divergence (right) for the VAE (in orange) and estimates of the respective lower bounds from Theorem 7.3.6 and Theorem 7.3.9 in blue. The lower bound of Corollary 7.3.8 is also plotted in red for the total variation. The experiments are averaged over 10 runs and the colored bands correspond to +/- the standard deviation.

In this experiment, we compare estimates of the bounds of Theorem 7.3.6, Corollary 7.3.8, and Theorem 7.3.9 to estimates of the total variation distance and the Kullback-Leibler divergence. We only provide results for the VAE since the bounds are not interesting for the GAN since they are consequences of interpolation between modes due to a small Lipschitz constant of the generative network. Yet this latter in the GAN case achieves a large Lipschitz constant so does not interpolate significantly. To estimate empirically the total variation distance and the Kullback-Leibler divergence, we used their respective analytical formula

$$D_{\mathrm{TV}}(g_{\#}\mu_{d'}, \nu) = (1/2)\int_{\mathbb{R}} |p_{g_{\#}\mu_{d'}}(x) - p_{\nu}(x)|dx \ ,$$
$$D_{\mathrm{KL}}(g_{\#}\mu_{d'}||\nu) = \int_{\mathbb{R}} p_{g_{\#}\mu_{d'}}(x) \log\left(p_{g_{\#}\mu_{d'}}(x)/p_{\nu}(x)\right) dx \ ,$$

where $p_{g_{\#}\mu_{d'}}$ and p_{ν} are the respective densities of $g_{\#}\mu_{d'}$ and ν. In order to estimate the lower bounds of Theorem 7.3.6 and Theorem 7.3.9, we set A of the form $(-\infty, -r/2]$ and we perform a grid search on r. In Figure B.2, we can observe that the estimates of the bounds provided by Theorem 7.3.6 and Theorem 7.3.9 are not tights. This is possibly because we selected a sub-optimal A but it most likely follows from the fact that the bounds don't take into account that $g_{\#}\mu_{d'}$ has automatically less mass on the modes than ν since a significant amount of its total mass is between them. One can also observe that the explicit lower bound of Corollary 7.3.8 is much smaller than the bound of Theorem 7.3.6. This can be explained by the facts that $\|m\|/2\sigma$ is probably a sub-optimal choice of r and that the bound of Corollary 7.3.8 minimizes the interpolation between modes over all the mappings with Lipschitz constant $\mathrm{Lip}(g)$, regardless whether these mappings approximate well ν on its modes or not. Since there is less interpolation if the modes are unbalanced (see Section 7.3.3), it is likely that the mappings g such that $g_{\#}\mu_{d'}$ is unbalanced are affecting the value of this bound in a bad way.

B.4.2 Additional examples

B.4.2.1 Univariate histograms

We provide additional visualizations of histograms of generated data with the three models for various values of m in Figure B.3. We can observe that the score-based model already generates unbalanced modes, but the phenomenon is globally less visible than in higher dimensions. Secondly, we provide additional visualizations of histograms of generated data with GANs trained with an additional gradient penalty term in the generator loss for various values of $L \approx \mathrm{Lip}(g)$ in Figure B.4.

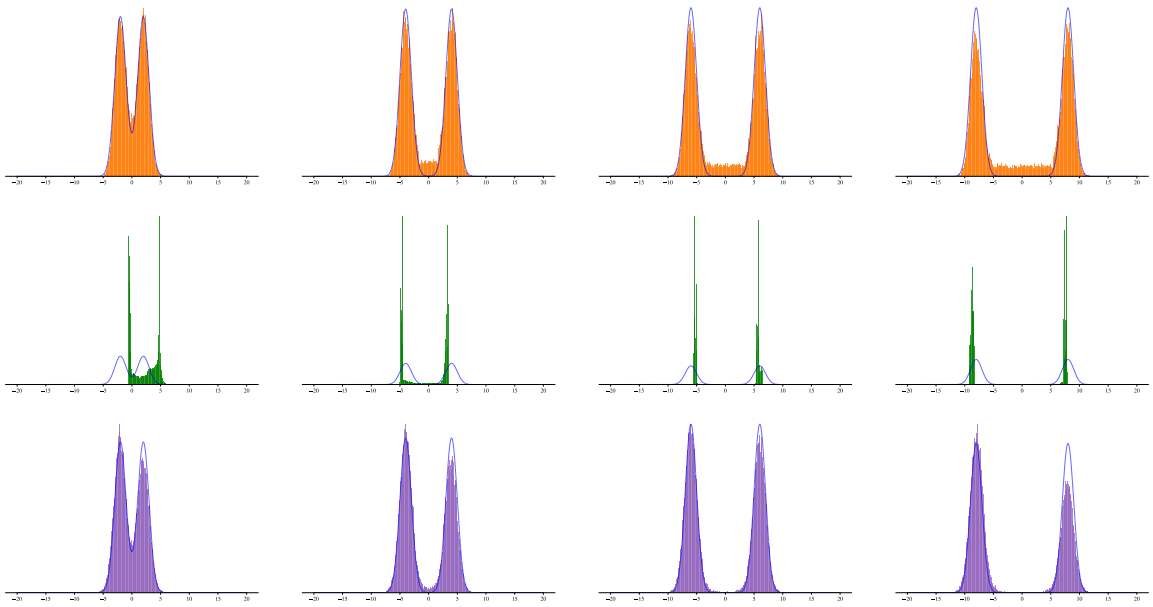

Figure B.3: Histograms of distributions generated with VAE (top, in orange), GAN (middle, in green), and with SGM (bottom, in purple) for $m = 2$, $m = 4$, $m = 6$ and $m = 8$. The data distribution densities are plotted in blue.

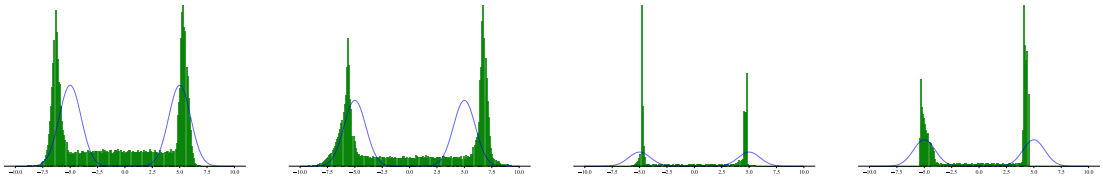

Figure B.4: histograms of distributions generated with GANs with with gradient penalty for $\text{Lip}(g) \approx L = 11$, $\text{Lip}(g) \approx L = 15$, $\text{Lip}(g) \approx L = 19$ and $\text{Lip}(g) \approx L = 23$. The data distribution densities are plotted in blue.

B.4.2.2 Visualization of generated data

Finally, we show randomly chosen generated samples with VAE, GAN and SGM on the synthetic mixture of Gaussian on MNIST and the subset of all 3 and 7 of MNIST in Figure B.5

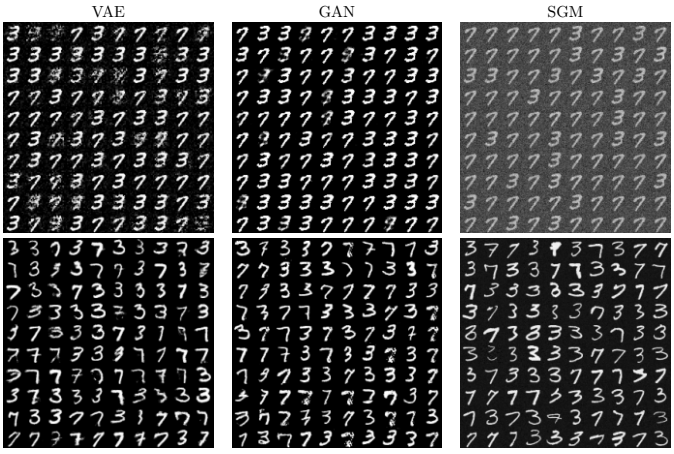

Figure B.5: Generated samples with VAE, GAN and SGM on the synthetic mixture of Gaussian on MNIST (top) and the subset of all 3 and 7 of MNIST (bottom). The samples have been randomly chosen.

Bibliography

Altschuler, J., Bach, F., Rudi, A., and Niles-Weed, J. (2019). Massively scalable Sinkhorn distances via the Nyström method. *Advances in neural information processing systems*, 32.

Altschuler, J., Bach, F., Rudi, A., and Weed, J. (2018). Approximating the quadratic transportation metric in near-linear time. *arXiv preprint arXiv:1810.10046*.

Altschuler, J., Niles-Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in neural information processing systems*, 30.

Alvarez-Melis, D. and Jaakkola, T. (2018). Gromov–Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890.

Alvarez-Melis, D., Jegelka, S., and Jaakkola, T. S. (2019). Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR.

Ambrosio, L., Fusco, N., and Pallara, D. (2000). *Functions of bounded variation and free discontinuity problems*. Courier Corporation.

Anderson, B. D. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, 50:5–43.

Anstreicher, K. and Wolkowicz, H. (2000). On Lagrangian relaxation of quadratic matrix constraints. In *Journal on Matrix Analysis and Applications*, volume 22, pages 41–55. SIAM.

Antoniou, A., Storkey, A., and Edwards, H. (2018). Augmenting image classifiers using data augmentation generative adversarial networks. In *International Conference on Artificial Neural Networks*, pages 594–603. Springer.

Arjovsky, M. and Bottou, L. (2017). Towards principled methods for training generative adversarial networks. *stat*, 1050:17.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR.

Baker, J. (1971). Isometries in normed spaces. *The American Mathematical Monthly*, 78(6):655–658.

Barratt, S. and Sharma, R. (2018). A note on the inception score. *arXiv preprint arXiv:1801.01973*.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research*, 18:1–43.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., and Jacobsen, J.-H. (2019). Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR.

Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R., and Jacobsen, J.-H. (2021). Understanding and mitigating exploding inverses in invertible neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1800. PMLR.

Beinert, R., Heiss, C., and Steidl, G. (2022). On assignment problems related to Gromov–Wasserstein distances on the real line. *arXiv preprint arXiv:2205.09006*.

Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Bertsekas, D. P. and Eckstein, J. (1988). Dual coordinate step methods for linear network flow problems. *Mathematical Programming*, 42(1-3):203–243.

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA.

Besag, J. (1994). Comments on "representations of knowledge in complex systems" by u. grenander and mi miller. *J. Roy. Statist. Soc. Ser. B*, 56(591-592):4.

Birkhoff, G. (1946). Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blumberg, A. J., Carriere, M., Mandell, M. A., Rabadan, R., and Villar, S. (2020). MREC: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. *stat*, 1050:20.

Bolley, F. (2008). Separability and completeness for the Wasserstein distance. *Lecture Notes in Mathematics-Springer-Verlag-*, 1934:371.

Bonnotte, N. (2013). *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie).

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: theory and applications*. Springer Science & Business Media.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. In *Communications on Pure and Applied Mathematics*, volume 44, pages 375–417. Wiley.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 6.

Brogat-Motte, L., Flamary, R., Brouard, C., Rousu, J., and d'Alché Buc, F. (2022). Learning to predict graphs with fused Gromov–Wasserstein barycenters. In *International Conference on Machine Learning*, pages 2321–2335. PMLR.

Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. (2019). Learning generative models across incomparable spaces. In *International conference on machine learning*, pages 851–861. PMLR.

Bures, D. (1969). An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w-algebras. *Transactions of the American Mathematical Society*, 135:199–212.

Burkard, R. E., Cela, E., Pardalos, P. M., and Pitsoulis, L. S. (1998). *The quadratic assignment problem*. Springer.

Caffarelli, L. A. (1996). Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496.

Caffarelli, L. A. (2003). The Monge-Ampère equation and optimal transportation, an elementary review. In *Optimal Transportation and Applications*.

Cai, Y. and Lim, L.-H. (2022). Distances between probability distributions of different dimensions. *IEEE Transactions on Information Theory*, 68(6):4020–4031.

Calamai, P. H. and Moré, J. J. (1987). Projected gradient methods for linearly constrained problems. *Mathematical programming*, 39(1):93–116.

Carlier, G., Chernozhukov, V., and Galichon, A. (2016). Vector quantile regression: an optimal transport approach. *Annals of Statistics*, 44(3):1165–1192.

Chen, T., Fox, E., and Guestrin, C. (2014). Stochastic gradient Hamiltonian Monte Carlo. In *International conference on machine learning*, pages 1683–1691. PMLR.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Variational lossy autoencoder. In *International Conference on Learning Representations*.

Chen, Y., Georgiou, T. T., and Tannenbaum, A. (2018). Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278.

Child, R. (2020). Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*.

Chowdhury, S. and Mémoli, F. (2019). The Gromov–Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787.

Chowdhury, S., Miller, D., and Needham, T. (2021). Quantized Gromov–Wasserstein. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 811–827. Springer.

Cohen, S. and Guibasm, L. (1999). The Earth mover's distance under transformation sets. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1076–1083. IEEE.

Cornish, R., Caterini, A., Deligiannidis, G., and Doucet, A. (2020). Relaxing bijectivity constraints with continuously indexed normalising flows. In *International Conference on Machine Learning*, pages 2133–2143. PMLR.

Courant, R. (1920). Über die eigenwerte bei den differentialgleichungen der mathematischen physik. *Mathematische Zeitschrift*, 7(1-4):1–57.

Courty, N., Flamary, R., and Ducoffe, M. (2018). Learning Wasserstein embeddings. In *ICLR 2018-6th International Conference on Learning Representations*, pages 1–13.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 39, pages 1853–1865. IEEE.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: an overview. *IEEE signal processing magazine*, 35(1):53–65.

Cuesta, J. A. and Matrán, C. (1989). Notes on the Wasserstein metric in Hilbert spaces. *The Annals of Probability*, pages 1264–1276.

Cunningham, W. H. (1976). A network simplex method. *Mathematical Programming*, 11:105–116.

Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. (2018). Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614.

Dai, B. and Wipf, D. (2018). Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*.

Dantzig, G. B. (1951). Application of the simplex method to a transportation problem. *Activity analysis and production and allocation*.

De Bortoli, V. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709.

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970.

Delon, J., Gozlan, N., and Saint-Dizier, A. (2022). Generalized Wasserstein barycenters between probability measures living on different subspaces. *Annals of Applied Probability*.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. (2020). Gromov–Wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, pages 2020–04.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. (2022). SCOT: single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Deshpande, I., Zhang, Z., and Schwing, A. G. (2018). Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34.

Doersch, C. (2016). Tutorial on variational autoencoders. *stat*, 1050:13.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dowson, D. and Landau, B. (1982). The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455.

Du, Y. and Mordatch, I. (2019). Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*.

Dumont, T., Lacombe, T., and Vialard, F.-X. (2022). On the existence of Monge maps for the Gromov–Wasserstein problem.

Dumoulin, V., Shlens, J., and Kudlur, M. (2016). A learned representation for artistic style. In *International Conference on Learning Representations*.

Dupuis, P. and Ellis, R. S. (2011). *A weak convergence approach to the theory of large deviations*. John Wiley & Sons.

Elfwing, S., Uchibe, E., and Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11.

Fan, J., Liu, S., Ma, S., Chen, Y., and Zhou, H.-M. (2022). Scalable computation of Monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. (2019). Efficient and accurate estimation of Lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., and Goodfellow, I. (2018). Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *International Conference on Learning Representations*.

Feng, J., Song, L., Huo, X., Yang, X., and Zhang, W. (2013). Image restoration via efficient Gaussian mixture model learning. In *2013 IEEE International Conference on Image Processing*, pages 1056–1060. IEEE.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. (2017). Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 291–299. Springer.

Figalli, A. (2007). Existence, uniqueness, and regularity of optimal transport maps. *SIAM journal on mathematical analysis*, 39(1):126–137.

Figalli, A. (2009). Regularity of optimal transport maps. *Séminaire Bourbaki*, 2008:997–1011.

Fischer, E. (1905). Über quadratische formen mit reellen koeffizienten. *Monatshefte für Mathematik und Physik*, 16:234–249.

Fix, E. and Hodges, J. (1951). Discriminatory analysis: nonparametric discrimination: consistency properties. report. 4. *T. USAF School of Aviation Medicine*.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. (2021). POT: Python Optimal Transport. *The Journal of Machine Learning Research*, 22(1):3571–3578.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. (2019). Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR.

Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192.

Galerne, B., Leclaire, A., and Rabin, J. (2017). Semi-discrete optimal transport in patch space for enriching Gaussian textures. In *Geometric Science of Information: Third International Conference, GSI 2017, Paris, France, November 7-9, 2017, Proceedings 3*, pages 100–108. Springer.

Genevay, A., Peyre, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617. PMLR.

Ghosh, P., Sajjadi, M. S., Vergari, A., Black, M., and Schölkopf, B. (2019). From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*.

Givens, C. R., Shortt, R. M., et al. (1984). A class of Wasserstein metrics for probability distributions. In *Michigan Mathematical Journal*, volume 31, pages 231–240. the University of Michigan.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

González-Sanz, A., De Lara, L., Béthune, L., and Loubes, J.-M. (2022). GAN estimation of Lipschitz optimal transport maps. *arXiv preprint arXiv:2202.07965*.

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. *Stat*, 1050:20.

Gramfort, A., Peyré, G., and Cuturi, M. (2015). Fast optimal transport averaging of neuroimaging data. In *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24*, pages 261–272. Springer.

Gromov, M. (1981). Groups of polynomial growth and expanding maps (with an appendix by jacques tits). *Publications Mathématiques de l'IHÉS*, 53:53–78.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2021). A review on generative adversarial networks: algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 35(4):3313–3332.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30.

Gulrajani, I., Raffel, C., and Metz, L. (2018). Towards GAN benchmarks which require generalization. In *International Conference on Learning Representations*.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31.

Gurumurthy, S., Kiran Sarvadevabhatla, R., and Venkatesh Babu, R. (2017). DELIGAN: generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition*, pages 166–174.

Gutierrez, J., Rabin, J., Galerne, B., and Hurtut, T. (2017). Optimal patch assignment for statistically constrained texture synthesis. In *Scale Space and Variational Methods in Computer Vision: 6th International Conference, SSVM 2017, Kolding, Denmark, June 4-8, 2017, Proceedings 6*, pages 172–183. Springer.

Hagemann, P. and Neumayer, S. (2021). Stabilizing invertible neural networks using mixture models. *Inverse Problems*, 37(8):085002.

Haker, S. and Tannenbaum, A. (2001). Optimal mass transport and image registration. In *Proceedings IEEE Workshop on Variational and Level Set Methods in Computer Vision*, pages 29–36. IEEE.

Hanin, B. (2019). Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10):992.

Haradal, S., Hayashi, H., and Uchida, S. (2018). Biosignal data augmentation based on generative adversarial networks. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 368–371. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Hitchcock, F. L. (1941). The distribution of a product from several sources to numerous localities. *Journal of mathematics and physics*, 20(1-4):224–230.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Hosseini, R. and Sra, S. (2020). An alternative to EM for Gaussian mixture models: batch and stochastic Riemannian optimization. *Mathematical programming*, 181(1):187–223.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

Ioffe, S. and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134.

Issenhuth, T., Tanielian, U., Picard, D., and Mary, J. (2020). Learning disconnected manifolds: avoiding the no GAN's land by latent rejection.

Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139.

James, I. M. (1976). *The topology of Stiefel manifolds*, volume 24. Cambridge University Press.

Jin, H., Yu, Z., and Zhang, X. (2022). Certifying robust graph classification under orthogonal Gromov–Wasserstein threats. *Advances in Neural Information Processing Systems*, 35:1737–1750.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., Kachman, T., and Mitliagkas, I. (2021). Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*.

Jolicoeur-Martineau, A., Piché-Taillefer, R., Mitliagkas, I., and des Combes, R. T. (2020). Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations*.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Karp, R. M., Vazirani, U. V., and Vazirani, V. V. (1990). An optimal algorithm for on-line bipartite matching. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 352–358.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

Khayatkhoei, M., Elgammal, A., and Singh, M. (2018). Disconnected manifold learning for generative adversarial networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7354–7364.

Kingma, D. P. and Ba, J. (2015). Adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. *Stat*, 1050:1.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392.

Kloeckner, B. (2010). A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9(2):297–323.

Kloeden, P. E., Platen, E., Kloeden, P. E., and Platen, E. (1992). *Stochastic differential equations*. Springer.

Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: an introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. (2017). On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). Generalized sliced Wasserstein distances. *Advances in neural information processing systems*, 32.

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. (2018a). Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Kolouri, S., Rohde, G. K., and Hoffmann, H. (2018b). Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436.

Koopmans, T. C. and Beckmann, M. (1957). Assignment problems and the location of economic activities. *Econometrica: journal of the Econometric Society*, pages 53–76.

Korotin, A., Kolesov, A., and Burnaev, E. (2022). Kantorovich strikes back! Wasserstein GANs are not optimal transport? *Advances in Neural Information Processing Systems*, 35:13933–13946.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Kumar, A. and Poole, B. (2020). On implicit regularization in β-VAEs. In *International Conference on Machine Learning*, pages 5480–5490. PMLR.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. (2019). Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32.

Lacoste-Julien, S. (2016). Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2022). Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems*.

Leclaire, A., Delon, J., and Desolneux, A. (2022). Optimal transport between GMMs for texture synthesis.

Leclaire, A. and Rabin, J. (2021). A stochastic multi-layer algorithm for semi-discrete optimal transport with applications to texture synthesis and style transfer. *Journal of Mathematical Imaging and Vision*, 63:282–308.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690.

Lee, H., Lu, J., and Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882.

Li, P., Wang, Q., and Zhang, L. (2013). A novel Earth mover's distance methodology for image matching with Gaussian mixture models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1696.

Lim, J. H. and Ye, J. C. (2017). Geometric GAN. *arXiv preprint arXiv:1705.02894*.

Lin, G., Milan, A., Shen, C., and Reid, I. (2017). RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934.

Liu, F., Zhang, G., and Lu, J. (2020). Heterogeneous domain adaptation: an unsupervised approach. *IEEE transactions on neural networks and learning systems*, 31(12):5588–5602.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. (2019). Sliced-Wasserstein flows: nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113. PMLR.

Loiola, E. M., De Abreu, N. M. M., Boaventura-Netto, P. O., Hahn, P., and Querido, T. (2007). A survey for the quadratic assignment problem. *European journal of operational research*, 176(2):657–690.

Lu, C., Chen, J., Li, C., Wang, Q., and Zhu, J. (2020). Implicit normalizing flows. In *International Conference on Learning Representations*.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. (2018). Are GANs created equal? a large-scale study. *Advances in neural information processing systems*, 31.

Luhman, E. and Luhman, T. (2021). Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*.

Luise, G., Pontil, M., and Ciliberto, C. (2020). Generalization properties of optimal transport gans with latent distribution learning. *arXiv preprint arXiv:2007.14641*.

Luo, Y., Zhu, L.-Z., Wan, Z.-Y., and Lu, B.-L. (2020). Data augmentation for enhancing EEG-based emotion recognition with deep generative models. *Journal of Neural Engineering*, 17(5):056021.

Luzi, L., Marrero, C. O., Wynar, N., Baraniuk, R. G., and Henry, M. J. (2023). Evaluating generative networks using Gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 279–288.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. (2019). BIVA: a very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32.

Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. (2013). Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, volume 30, page 3. PMLR.

Magnus, J. R. and Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Mallasto, A., Montúfar, G., and Gerolin, A. (2019). How well do WGANs estimate the Wasserstein metric? *arXiv preprint arXiv:1910.03875*.

Mazur, S. and Ulam, S. (1932). Sur les transformations isométriques d'espaces vectoriels normés. *CR Acad. Sci. Paris*, 194(946-948):116.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309.

McCann, R. J. and Guillen, N. (2011). Five lectures on optimal transportation: geometry, regularity and applications. *Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal*, pages 145–180.

Mehr, E., Jourdan, A., Thome, N., Cord, M., and Guitteny, V. (2019). DiscoNet: shapes learning on disconnected manifolds for 3d editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3474–3483.

Melzi, S., Marin, R., Rodolà, E., Castellani, U., Ren, J., Poulenard, A., Wonka, P., and Ovsjanikov, M. (2019). SHREC 2019: matching humans with different connectivity. In *Eurographics Workshop on 3D Object Retrieval*, volume 7, page 3. The Eurographics Association.

Mémoli, F. (2009). Spectral Gromov–Wasserstein distances for shape matching. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 256–263. IEEE.

Mémoli, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. In *Foundations of Computational Mathematics*, volume 11, pages 417–487. Springer.

Mémoli, F. and Needham, T. (2018). Gromov–Monge quasi-metrics and distance distributions. *arXiv*, 2018.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. (2017). Unrolled generative adversarial networks. In *International Conference on Learning Representations*.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 6.

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27.

Morgan, N. and Bourlard, H. (1989). Generalization and parameter estimation in feedforward nets: some experiments. *Advances in neural information processing systems*, 2.

Mulayoff, R., Michaeli, T., and Soudry, D. (2021). The implicit bias of minima stability: a view from function space. *Advances in Neural Information Processing Systems*, 34:17749–17761.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. (2020). Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812.

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. (2019). Asymptotic guarantees for learning generative models with the sliced-Wasserstein distance. *Advances in Neural Information Processing Systems*, 32.

Nagarajan, V., Raffel, C., and Goodfellow, I. J. (2018). Theoretical insights into memorization in GANs. In *Neural Information Processing Systems Workshop*, volume 1.

Nazeri, K., Ng, E., and Ebrahimi, M. (2018). Image colorization using generative adversarial networks. In *Articulated Motion and Deformable Objects: 10th International Conference, AMDO 2018, Palma de Mallorca, Spain, July 12-13, 2018, Proceedings 10*, pages 85–94. Springer.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR.

Nijkamp, E., Hill, M., Han, T., Zhu, S.-C., and Wu, Y. N. (2020). On the anatomy of MCMC-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5272–5280.

Odena, A., Buckman, J., Olsson, C., Brown, T., Olah, C., Raffel, C., and Goodfellow, I. (2018). Is generator conditioning causally related to GAN performance? In *International Conference on Machine Learning*, pages 3849–3858. PMLR.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Parmar, G., Zhang, R., and Zhu, J.-Y. (2022). On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420.

Pasande, M., Hosseini, R., and Araabi, B. N. (2022). Stochastic first-order learning for large-scale flexibly tied Gaussian mixture model. *arXiv preprint arXiv:2212.05402*.

Paty, F.-P., d'Aspremont, A., and Cuturi, M. (2020). Regularity as regularization: smooth and strongly convex brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pele, O. and Taskar, B. (2013). The tangent Earth mover's distance. In *Geometric Science of Information: First International Conference, GSI 2013, Paris, France, August 28-30, 2013. Proceedings*, pages 397–404. Springer.

Pennington, J., Schoenholz, S., and Ganguli, S. (2017). Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in Neural Information Processing Systems*, 30.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Petric Maretic, H., El Gheche, M., Chierchia, G., and Frossard, P. (2019). GOT: an optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems*, 32.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: with applications to data science. In *Foundations and Trends in Machine Learning*, volume 11, pages 355–607. Now Publishers Inc.

Peyré, G., Cuturi, M., and Solomon, J. (2016). Gromov–Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR.

Philippis, G. (2013). *Regularity of optimal transport maps and applications*, volume 17. Springer Science & Business Media.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. (2020). The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*.

Rabin, J., Ferradans, S., and Papadakis, N. (2014). Adaptive color transfer with relaxed optimal transport. In *2014 IEEE international conference on image processing (ICIP)*, pages 4852–4856. IEEE.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012). Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677.

Razavi, A., Van den Oord, A., and Vinyals, O. (2019). Generating diverse high-fidelity images with VQ–VAE–2. *Advances in neural information processing systems*, 32.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR.

Robbins, H. (1955). *An empirical Bayes approach to statistics*. Office of Scientific Research, US Air Force.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rout, L., Korotin, A., and Burnaev, E. (2021). Generative modeling with optimal transport maps. In *International Conference on Learning Representations*.

Rubner, Y., Tomasi, C., and Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The Earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Rüschendorf, L. and Rachev, S. T. (1990). A characterization of random variables with minimum L2-distance. *Journal of multivariate analysis*, 32(1):48–54.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Rustamov, R. M., Ovsjanikov, M., Azencot, O., Ben-Chen, M., Chazal, F., and Guibas, L. (2013). Map-based exploration of intrinsic shape differences and variability. *ACM Transactions on Graphics (TOG)*, 32(4):1–12.

Ryner, M., Kronqvist, J., and Karlsson, J. (2023). Globally solving the Gromov–Wasserstein problem for point clouds in low dimensional euclidean spaces. *arXiv preprint arXiv:2307.09057*.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. (2022). Palette: image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems*, 31.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training GANs. *Advances in neural information processing systems*, 29.

Salimans, T. and Kingma, D. P. (2016). Weight normalization: a simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29.

Salmona, A., Bouza, L., and Delon, J. (2022a). Deoldify: a review and implementation of an automatic colorization method. *Image Processing On Line*, 12:347–368.

Salmona, A., De Bortoli, V., Delon, J., and Desolneux, A. (2022b). Can push-forward generative models fit multimodal distributions? *Advances in Neural Information Processing Systems*, 35:10766–10779.

Salmona, A., Delon, J., and Desolneux, A. (2021). Gromov–Wasserstein distances between Gaussian distributions. *Journal of Applied Probability*, 59(4).

Salmona, A., Delon, J., and Desolneux, A. (2023). Gromov–Wassertein-like distances in the Gaussian mixture models space. *preprint*.

San-Roman, R., Nachmani, E., and Wolf, L. (2021). Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*.

Sandfort, V., Yan, K., Pickhardt, P. J., and Summers, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in ct segmentation tasks. *Scientific Reports*, 9(1):1–9.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Scetbon, M. and Cuturi, M. (2020). Linear time Sinkhorn divergences using positive features. *Advances in Neural Information Processing Systems*, 33:13468–13480.

Scetbon, M., Cuturi, M., and Peyré, G. (2021). Low-rank Sinkhorn factorization. In *International Conference on Machine Learning*, pages 9344–9354. PMLR.

Scetbon, M., Peyré, G., and Cuturi, M. (2022). Linear-time Gromov–Wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR.

Schmitzer, B. and Schnörr, C. (2013). Modelling convex shape priors and matching based on the Gromov–Wasserstein distance. *Journal of mathematical imaging and vision*, 46:143–159.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. (2017). Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*.

Sembach, L., Burgard, J. P., and Schulz, V. (2022). A Riemannian Newton trust-region method for fitting Gaussian mixture models. *Statistics and Computing*, 32(1):8.

Shao, S., Wang, P., and Yan, R. (2019). Generative adversarial networks for data augmentation in machine fault diagnosis. *Computers in Industry*, 106:85–93.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. (2016). Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016). Ladder variational autoencoders. *Advances in neural information processing systems*, 29.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32.

Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems*, 33:12438–12448.

Song, Y., Shen, L., Xing, L., and Ermon, S. (2021). Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Stanczuk, J., Etmann, C., Kreusser, L. M., and Schönlieb, C.-B. (2021). Wasserstein GANs work because they fail (to approximate the Wasserstein distance). *arXiv preprint arXiv:2103.01678*.

Stéphanovitch, A., Tanielian, U., Cadre, B., Klutchnikoff, N., and Biau, G. (2022). Optimal 1-Wasserstein distance for WGANs. *arXiv preprint arXiv:2201.02824*.

Strand, O. N. (1974). Theory and methods related to the singular-function expansion and landweber's iteration for integral equations of the first kind. *SIAM Journal on Numerical Analysis*, 11(4):798–825.

Sturm, K.-T. (2006). On the geometry of metric measure spaces. I. *Acta Math*, 196:65–131.

Sturm, K.-T. (2012). The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*.

Su, Z., Wang, Y., Shi, R., Zeng, W., Sun, J., Luo, F., and Gu, X. (2015). Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259.

Subakan, Y. C. and Smaragdis, P. (2018). Generative adversarial source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 26–30. IEEE.

Sudakov, V. N. and Tsirelson, B. S. (1978). Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Takatsu, A. (2010). On Wasserstein geometry of Gaussian measures. In *Probabilistic approach to geometry*, pages 463–472. Mathematical Society of Japan.

Tanielian, U., Issenhuth, T., Dohmatob, E., and Mary, J. (2020). Learning disconnected manifolds: a no GAN's land. In *International Conference on Machine Learning*, pages 9418–9427. PMLR.

Teodoro, A., Almeida, M., and Figueiredo, M. (2015). Single-frame image denoising and inpainting using Gaussian mixtures. In *International Conference on Pattern Recognition Applications and Methods*, volume 2, pages 283–288. SCITEPRESS.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schölkopf, B. (2018). Wasserstein auto-encoders. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview. net.

Tran, D., Ranganath, R., and Blei, D. (2017). Hierarchical implicit models and likelihood-free variational inference. *Advances in Neural Information Processing Systems*, 30.

Vahdat, A. and Kautz, J. (2020). NVAE: a deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679.

Vahdat, A., Kreis, K., and Kautz, J. (2021). Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34.

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Vayer, T. (2020). *A contribution to Optimal Transport on incomparable spaces*. PhD thesis, Lorient.

Vayer, T., Courty, N., Tavenard, R., and Flamary, R. (2019a). Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR.

Vayer, T., Flamary, R., Courty, N., Tavenard, R., and Chapel, L. (2019b). Sliced Gromov–Wasserstein. *Advances in Neural Information Processing Systems*, 32.

Villani, C. (2003). *Topics in optimal transportation*. American Mathematical Soc.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. (2021). Semi-relaxed Gromov–Wasserstein divergence with applications on graphs. *arXiv preprint arXiv:2110.02753*.

Virmaux, A. and Scaman, K. (2018). Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31.

Wang, C. and Mahadevan, S. (2011). Heterogeneous domain adaptation using manifold alignment. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, page 1541.

Wang, W., Ozolek, J. A., Slepčev, D., Lee, A. B., Chen, C., and Rohde, G. K. (2010). An optimal transportation approach for nuclear structure-based pathology. *IEEE transactions on medical imaging*, 30(3):621–631.

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., and Rohde, G. K. (2013). A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101:254–269.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803.

Watson, D., Ho, J., Norouzi, M., and Chan, W. (2021). Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*.

Wenliang, L. K. and Kanagawa, H. (2020). Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*.

Wu, H., Köhler, J., and Noé, F. (2020). Stochastic normalizing flows. *Advances in Neural Information Processing Systems*, 33:5933–5944.

Wu, Y. and He, K. (2018). Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19.

Xu, H., Luo, D., and Carin, L. (2019a). Scalable Gromov–Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32.

Xu, H., Luo, D., Zha, H., and Duke, L. C. (2019b). Gromov–Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR.

Xu, H., Wang, W., Liu, W., and Carin, L. (2018). Distilled Wasserstein learning for word embedding and topic modeling. *Advances in Neural Information Processing Systems*, 31.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214.

Yang, Z., Chen, W., Wang, F., and Xu, B. (2018). Improving neural machine translation with conditional sequence generative adversarial nets. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:1346–1355.

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. (2017). Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493.

Yeh, Y.-R., Huang, C.-H., and Wang, Y.-C. F. (2014). Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Transactions on Image Processing*, 23(5):2009–2018.

Yu, G., Sapiro, G., and Mallat, S. (2011). Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity. *IEEE Transactions on Image Processing*, 21(5):2481–2499.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR.

Zhang, Q. and Chen, J. (2020). A unified framework for Gaussian mixture reduction with composite transportation distance. *arXiv preprint arXiv:2002.08410*.

Zhang, S., Jiao, L., Liu, F., and Wang, S. (2017). Global low-rank image restoration with Gaussian mixture model. *IEEE transactions on cybernetics*, 48(6):1827–1838.

Zhu, L., Yang, Y., Haker, S., and Tannenbaum, A. (2007). An image morphing technique based on optimal mass preserving mapping. *IEEE transactions on image processing*, 16(6):1481–1495.

Zoran, D. and Weiss, Y. (2011). From learning models of natural image patches to whole image restoration. In *2011 international conference on computer vision*, pages 479–486. IEEE.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: Transport de mesures de probabilités à travers des espaces euclidiens de dimensions différentes

Mots clés: apprentissage, transport optimal, modèles génératifs, Gromov-Wasserstein

Résumé: Dans cette thèse, nous étudions trois problèmes liés au transport de mesures qui vivent dans des espaces euclidiens différents, les deux premiers dans le contexte du transport optimal et le dernier dans le contexte de la modélisation générative. Dans la partie sur le transport optimal, nous étudions d'abord le comportement des généralisations communes du transport optimal, dont celle dite de Gromov-Wasserstein, entre des distributions gaussiennes qui vivent sur des espaces non directement comparables. Ensuite, nous concevons une distance de transport optimal entre des mélanges de gaussiennes de dimensions différentes. Finalement dans la partie sur la modélisation générative, nous étudions l'expressivité des modèles génératifs en relation avec la constante de Lipschitz de leur fonction de transport.

Title: Transport of probability distributions across different Euclidean spaces

Keywords: machine learning, optimal transport, generative modeling, Gromov-Wasserstein

Abstract: In this thesis, we study three problems related to the transport of measures lying on different Euclidean spaces, the first two being in the context of optimal transport and the last one being in the context of generative modeling. In the optimal transport part, we first study the behavior of the common generalizations of optimal transport, including the so-called Gromov-Wasserstein distance, between Gaussian distributions in incomparable spaces. Secondly, we design a computationally efficient and scalable OT distance between Gaussian mixtures possibly living in different Euclidean spaces. Finally, in the generative modeling part, we study the expressivity of generative models relatively to the Lipschitz constant of their push-forward mapping.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France