Propriétés statistiques du barycentre dans l'espace de Wasserstein

Elsa Cazelles

Institut de Mathématiques de Bordeaux

université de BORDEAUX

Thèse

présentée à

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **Elsa Cazelles**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : **MATHÉMATIQUES APPLIQUÉES**

Statistical properties of barycenters in the Wasserstein space and fast algorithms for optimal transport of measures.

Soutenue publiquement le 21 Septembre 2018

après avis des rapporteurs :

Eustasio del Barrio Tellado	Professeur	Universidad de Valladolid	Rapporteur
Gabriel Peyré	DR CNRS	DMA - École normale supérieure	Rapporteur

devant la commission d'examen composée de :

Eustasio del Barrio Tellado	Professeur	Universidad de Valladolid	Rapporteur
Gérard Biau	Professeur	Université Paris-Sorbonne	Président de jury
Jérémie Bigot	Professeur	Université de Bordeaux	Directeur de thèse
Marco Cuturi	Professeur	ENSAE - Université Paris-Saclay	Examinateur
Claire Lacour	Maître de conférences	Université Paris-Sud	Examinatrice
Nicolas Papadakis	CR CNRS	Université de Bordeaux	Directeur de thèse
Gabriel Peyré	DR CNRS	DMA - École normale supérieure	Rapporteur

– 2018 –

Titre. Propriétés statistiques du barycentre dans l'espace de Wasserstein.

Résumé. Cette thèse se concentre sur l'analyse de données présentées sous forme de mesures de probabilité sur $\mathbb{R}^d$. L'objectif est alors de fournir une meilleure compréhension des outils statistiques usuels sur cet espace muni de la distance de Wasserstein. Une première notion naturelle est l'analyse statistique d'ordre un, consistant en l'étude de la moyenne de Fréchet (ou barycentre). En particulier, nous nous concentrons sur le cas de données (ou observations) discrètes échantillonnées à partir de mesures de probabilité absolument continues (*a.c.*) par rapport à la mesure de Lebesgue. Nous introduisons ainsi un estimateur du barycentre de mesures aléatoires, pénalisé par une fonction convexe, permettant ainsi d'imposer son *a.c.* Un autre estimateur est régularisé par l'ajout d'entropie lors du calcul de la distance de Wasserstein. Nous nous intéressons notamment au contrôle de la variance de ces estimateurs. Grâce à ces résultats, le principe de Goldenshluger et Lepski nous permet d'obtenir une calibration automatique des paramètres de régularisation. Nous appliquons ensuite ce travail au recalage de densités multivariées, notamment pour des données de cytométrie de flux. Nous proposons également un test d'adéquation de lois capable de comparer deux distributions multivariées, efficacement en terme de temps de calcul. Enfin, nous exécutons une analyse statistique d'ordre deux dans le but d'extraire les tendances géométriques globales d'un jeu de donnée, c'est-à-dire les principaux modes de variations. Pour cela nous proposons un algorithme permettant d'effectuer une analyse en composantes principales géodésiques dans l'espace de Wasserstein.

Mots-clés. Espace de Wasserstein, Barycentre, Transport optimal régularisé, ACP, Test d'hypothèse

Laboratoire d'accueil. Institut de Mathématiques de Bordeaux
351, cours de la Libération, 33 405, Talence, France.

Title. Statistical properties of barycenters in the Wasserstein space.

Abstract. This thesis focuses on the analysis of data in the form of probability measures on $\mathbb{R}^d$. The aim is to provide a better understanding of the usual statistical tools on this space endowed with the Wasserstein distance. The first order statistical analysis is a natural notion to consider, consisting of the study of the Fréchet mean (or barycentre). In particular, we focus on the case of discrete data (or observations) sampled from absolutely continuous probability measures (*a.c.*) with respect to the Lebesgue measure. We thus introduce an estimator of the barycenter of random measures, penalized by a convex function, making it possible to enforce its *a.c.* Another estimator is regularized by adding entropy when computing the Wasserstein distance. We are particularly interested in controlling the variance of these estimators. Thanks to these results, the principle of Goldenshluger and Lepski allows us to obtain an automatic calibration of the regularization parameters. We then apply this work to the registration of multivariate densities, especially for flow cytometry data. We also propose a test statistic that can compare two multivariate distributions, efficiently in terms of computational time. Finally, we perform a second-order statistical analysis to extract the global geometric tendency of a dataset, also called the main modes of variation. For that purpose, we propose algorithms allowing to carry out a geodesic principal components analysis in the space of Wasserstein.

Keywords. Wasserstein space, Barycenter, Regularized optimal transport, PCA, Hypothesis testing

Institute. Institut de Mathématiques de Bordeaux
351, cours de la Libération, 33 405, Talence, France.

Remerciements

Ces trois dernières années et demi au sein de l'IMB n'auraient pu que difficilement mieux se dérouler, et c'est en grande partie grâce à Jérémie et Nicolas. Je tiens donc à chaleureusement les remercier pour leur temps, leur écoute quant à mes préférences de recherche, leur prévenance, leur confiance, et globalement leur encadrement qui m'a permis de mûrir dans le travail. Merci aussi pour toutes les opportunités de rencontres à travers mes déplacements!

Je tiens à remercier Gabriel Peyré et Eustasio del Barrio, que je sais très occupés, d'avoir accepté de relire mon manuscrit, et pour leur amabilité lorsque nos chemins se sont croisés. Je remercie également les examinateurs-trice de prendre du temps aujourd'hui, Gérard Biau, Claire Lacour et Marco Cuturi (merci également d'avoir justifié un séjour au Japon!).

Tout au long de ces trois années, j'ai vraiment apprécié la vie à l'IMB, la bonne ambiance généralisée et la sympathie du personnel administratif. En particulier, merci à Ida, Muriel, la cellule info, Cathy et Nadia. Un remerciement spécial à Karine et Rachel, pour leur -et je pèse mes mots- exquise compagnie.

Aussi, un grand merci à l'équipe IOP toute entière, les anciens Charles De, Charles Do, Jean-François, Camille, Bernard, Marc, Adrien et les nouveaux Edoardo, Yann, Arthur, les présents et les absents, le voisinage (Michel), pour leur très grande bienveillance et bonne humeur. Merci d'avoir engendré un cadre de travail convivial -et bien entendu productif-.

Aux premiers doctorants que j'ai rencontrés, merci pour l'accueil ! Alice pour ta jovialité communicative, Zoé pour les premières sorties -et les suivantes-, Bruno pour tes râles, Marie, Pierre, Marc, Camille, pour les moments qu'on continue de passer ensemble, Thomas, pour ton amour du débat, Arnaud, et Fabien pour tes patates.

À tous les autres doctorants, Vassilis, coup de coeur de bureau, pour ta générosité (et Athina pour ta douceur). Baptiste, un héros du quotidien, Lara, le rayon de soleil, Thomas, la folie attendrissante, Alexandre, la très bonne deuxième impression, Corentin, ce soir il y a une soirée tu sais? t'es invité, Marco, la grande bouche, Antonin, l'amour de la mondialisation, les italiens, Edo, Marco, Sergio, Roberto, et leurs compatriotes, Corentin, Ibra. Bianca et Stéphane, une belle alternative à mes habitudes de l'IMB. Pierre et Rémi pour la famille IOP.

Momo et Zaza, pour tous les moments, jamais suffisants, passés à raconter nos vies, nos réussites et nos soucis. Thibault et Cannelle, Nathan et Cécilia, Louis-Marie et Luis, c'est toujours un grand plaisir de vous retrouver. Guillaume et Manon, toujours partants pour un verre, jamais à court de conversation. Diane et Stefan, qu'il fait bon rejoindre à Budva.

J'ai aussi eu la chance de tisser des amitiés lors de mes déplacements. Je pense en particulier à Aude, et à ta capacité à prendre du recul sur les choses de la vie. Et Vivien, tu me fais énormément rire, ce qui compense toutes les fois où tu es insupportable.

Les Bordelais avec qui j'ai passé le plus de temps, mes adorés, Jo, tu nous rends la vie plus facile et plus belle, à coup de pâté en croûte, Nico, pour toutes tes créations, bien mieux que Laurent Gerra, Niko, merci pour ton oreille attentive et tes conseils avisés (qui l'eût cru?), et merci pour ton appart et tes colocs, un refuge précieux à 50 mètres de chez moi. Roxane, le bol d'air frais/la tempête qui manquait à ce petit groupe. Les dimanches après-midi sont un vrai régal en votre compagnie.

De mes années à Toulouse, je retiens bien sûr Hugo et Clémentine, on se suit de loin, mais c'est toujours super de vous revoir. Et ça vaut bien sûr aussi pour vous Vincent, Tom, Hedi, Myriam et Marie.

Aussi, les autres Toulousains, mes premiers grands copains de fac. Mélanie et David, merci pour nos moments de profonde tristesse qui se transforment en beaux souvenirs (parc d'enfants à Ljubljana, la boîte à Krk, mon anniversaire dans un bar à chicha, les bières au miel, *sexual healing* sous la pluie, et j'en passe). Vincent, merci pour ta gentillesse et ton optimisme sans borne. Et merci à vous trois pour Sem, best WE ever. Achaya, merci de nous rappeler que tout ira bien, toujours; Léa, de m'apprendre tout plein de choses, et Xavier de rendre tout ça plus léger. Merci Ju... non rien, au temps pour moi.

Si on remonte à plus loin, beaucoup plus loin, merci à Éric (et Clara!) d'apporter de la spontanéité, de la fraîcheur et de la fantaisie dans mes habitudes, pensée pour Frédo aussi!

En redescendant juste un peu, merci aux copines, Amandine, Anne, Julia, Lucie, Marie, Mathilde et Océane. De loin je vous aime toujours autant, et chacun de vos caractères.

Sami et Philou, qui rendent la vie tellement plus marrante, et pleine de joie, toujours à l'écoute des potins. Vous savez déjà tout le bien que je pense de vous, on se le dit quatre fois par jour <3. J'ai hâte des prochaines aventures ensemble!

Léa, le sourire qui vit pour les petits plaisirs du quotidien, on adore (et Romain, c'est un plaisir de te compter parmi nous)! Vincent, merci d'être toujours présent dans les moments phares, pour ta bonne humeur infaillible, et pour ton manque total d'inhibition. Petite pensée pour Pierre, loin des yeux près du coeur! Charline, pour ton humour, ta générosité et ton intelligence, et Anthony, pour ton esprit de partage, ta curiosité et ton indulgence! À vous deux vous avez fait de Bordeaux ma maison. La vieillesse à nous tous, ça va être bien.

Merci à ma famille (et Ahmed), de veiller sur moi à distance. En toute objectivité, vous êtes les meilleurs d'entre nous!

Enfin en vrac, je souhaite remercier le TNT, de promouvoir le talent de jeunes comédiens; le lait de poule d'avoir accompagné les huit hivers derniers (et très certainement les huit prochains); merci au guide de Kotor de m'avoir réconciliée avec l'Histoire; merci à Miri pour ses punchlines ("Princesse E. dort chez grand-mère"); merci à Naples, tout simplement.

Contents

Introduction (Français)

Dans cette introduction, nous présentons une revue - non exhaustive - de la littérature sur le transport optimal, ainsi que ses nombreuses applications en analyse de données. Nous introduisons également les définitions et notations qui serviront tout au long de cette thèse. Pour finir, nous exposons un résumé détaillé de nos travaux et le contenu de ce manuscrit.

A. Transport optimal et applications

A.1. Monge, Kantorovich et Wasserstein

Gaspard Monge a été le premier à introduire en 1781 le problème de transfert de masses dans son *Mémoire sur la théorie des déblais et des remblais*. Son but était de trouver le moyen le plus efficace, c'est-à-dire demandant le moins d'effort possible, pour transporter un tas de sable dans un trou de même volume. Dans sa formulation moderne, le problème consiste à trouver l'application mesurable optimale permettant de transférer à moindre coût la masse d'une mesure de probabilité μ à support dans un espace mesuré $\mathcal{X}$, sur une autre ν à support dans $\mathcal{Y}$. Ainsi, le problème de Monge, revient à

$$\text{minimiser} \qquad \int_{\mathcal{X}} c(x, T(x)) d\mu(x) \tag{A.1}$$

sur l'ensemble des fonctions mesurables $T : \mathcal{X} \to \mathcal{Y}$ telles que $\nu = T\#\mu$. Cet opérateur *pushforward* $\#$ est défini tel que pour tout ensemble mesurable $B \subset \mathcal{Y}$, on a $\nu(B) = \mu(T^{-1}(B))$. La fonction $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ est une fonction de coût mesurable. Un exemple de transfert de masse, dans l'esprit du problème de déblais et remblais de Monge, est présenté dans la Figure A.1.

Toutefois, de telles applications T n'existent pas toujours, en particulier si la masse de μ en un point doit se scinder en plusieurs morceaux. Pour palier cette restriction, Kantorovich a étendu dans les années quarante le problème (A.1) en introduisant un plan de transport entre la mesure de départ et la mesure cible, qui contient le comportement du transfert de masse, c'est-à-dire

$$\text{minimiser} \qquad \iint_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) \tag{A.2}$$

sur l'ensemble des plans de transport π appartenant à $\Pi(\mu, \nu)$, *i.e.* l'ensemble des mesures produit sur $\mathcal{X} \times \mathcal{Y}$ de marginales respectives μ et ν. Dans le cas de mesures discrètes, *e.g.* Figure A.2, le plan de transport peut allouer de la masse d'un point du support de μ en

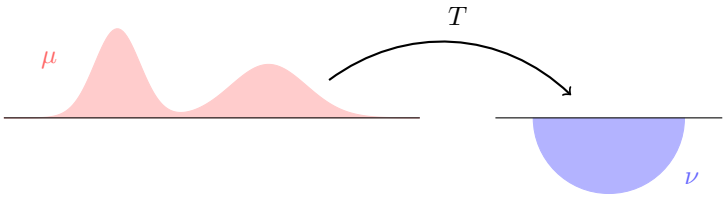


FIGURE A.1. Transfert de la masse de μ sur la masse de ν par l'application T telle que $\nu = T\#\mu$.

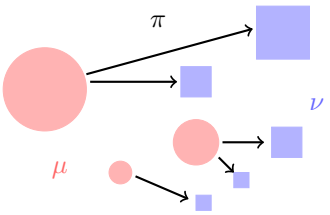


FIGURE A.2. Transfert de la masse de μ sur la masse de ν par un plan de transport $\pi \in \Pi(\mu, \nu)$.

différents points du support de ν, ce qu'une application T n'est pas capable de faire. Les notions de transport optimal, ainsi que les points de vue géométriques et différentiels de ces problèmes de minimisation, sont détaillés dans les ouvrages de Villani [**Vil03**, **Vil08**], Ambrosio et Gigli [**AG13**] et Ambrosio et al. [**AGS04**].

Un cadre particulièrement intéressant du transport optimal se détache lorsque $\mathcal{X}$ est un espace polonais muni de la distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. En effet dans ce cas, le problème de transfert de masse de Kantorovich entre deux mesures définit une distance pour un coût $c := d^p$, dès lors que les mesures appartiennent au bon espace. Plus précisément, pour $p \in [1, +\infty)$, nous dénotons $\mathcal{P}_p(\mathcal{X})$ l'ensemble des mesures de probabilité boréliennes (également appelées distributions) sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ à support dans $\mathcal{X}$, où $\mathcal{B}(\mathcal{X})$ est la σ-algèbre des sous-ensembles boréliens de $\mathcal{X}$, admettant un moment d'ordre p. Autrement dit,

$$\mu \in \mathcal{P}_p(\mathcal{X}) \text{ équivaut à } \int_{\mathcal{X}} d^p(x_0, x) d\mu(x) < +\infty \text{ pour n'importe quel } x_0 \in \mathcal{X}.$$

Remarquons que $\mathcal{P}_p(\mathcal{X})$ est inclu dans l'ensemble $\mathcal{M}(\mathcal{X})$ des mesures de Radon bornées. Nous obtenons alors la définition suivante :

DÉFINITION A.1. *La p-distance de Wasserstein (1969, Leonid Wasserstein) est donnée pour μ, ν dans $\mathcal{P}_p(\mathcal{X})$ par*

$$W_p(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \iint_{\mathcal{X}^2} d^p(x,y) d\pi(x,y) \right)^{1/p} \tag{A.3}$$

où l'infimum est pris sur l'ensemble $\Pi(\mu,\nu)$ des plans de transport sur $\mathcal{X} \times \mathcal{X}$ de marginales respectives μ et ν.

Cette distance a notamment l'avantage de caractériser la convergence faible de mesure sur l'espace métrique $(\mathcal{P}_p(\mathcal{X}), W_p)$ (voir *e.g.* Chapitre 7 de [**Vil03**]).

Kantorovich a également décrit le problème de minimisation (A.3) dans sa formulation duale, correspondant à une optimisation contrainte sur un espace de fonctions. Nous rappelons que l'espace $\mathbb{L}_p(\mu)$, pour $p \in [1,\infty)$ et $\mu \in \mathcal{M}(\mathcal{X})$, est l'espace des fonctions $f : \mathcal{X} \to \mathbb{R}$ telles que $|f|^p$ est μ-intégrable, et telles que toutes les fonctions égales μ-presque partout sont identifiées. Le problème dual de (A.3) est alors donné par le théorème suivant.

THÉORÈME A.2 (Théorème du dual de Kantorovich). *Soient $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, alors on a*

$$W_p(\mu,\nu) = \left(\sup_{(\phi,\psi) \in C_{W_p}} \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{X}} \psi(y) d\nu(y) \right)^{1/p}, \tag{A.4}$$

où C_{W_p} est l'ensemble des fonctions mesurables $(\phi,\psi) \in \mathbb{L}_1(\mu) \times \mathbb{L}_1(\nu)$ satisfaisant

$$\phi(x) + \psi(y) \leq d^p(x,y), \tag{A.5}$$

pour μ-presque tout $x \in \mathcal{X}$ et ν-presque tout $y \in \mathcal{X}$.

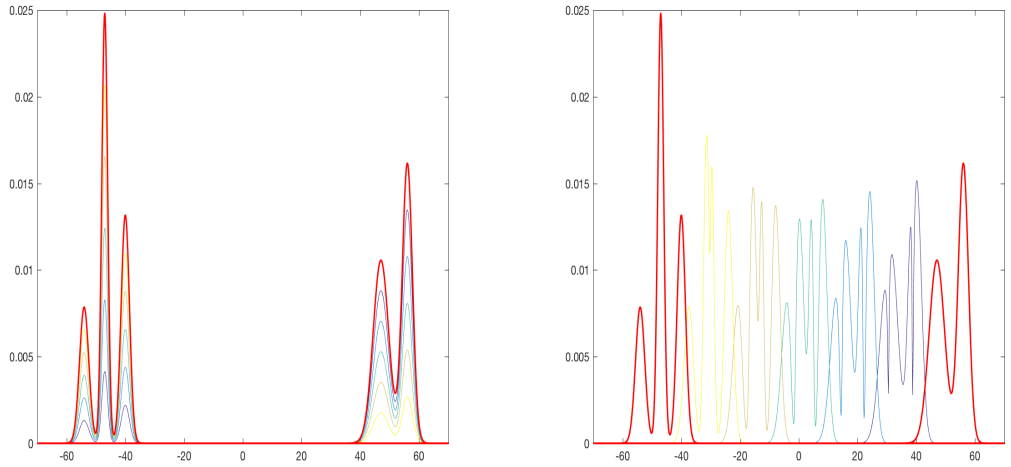
A.2. La distance de Wasserstein sur la droite réelle

Le cas de mesures à support sur la droite réelle, c'est-à-dire lorsque $\mathcal{X}$ est un intervalle (possiblement non borné) de $\mathbb{R}$, est particulièrement intéressant car la distance de Wasserstein W_p est alors égale à la distance $\mathbb{L}_p$ des fonctions quantiles. Formellement, en notant F_μ la fonction de répartition de μ et F_μ^- son quantile généralisé, la distance de Wasserstein devient, pour $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$,

$$W_p(\mu,\nu) = \left(\int_0^1 |F_\mu^-(t) - F_\nu^-(t)|^p dt \right)^{1/p}. \tag{A.6}$$

Si $\mu \in \mathcal{P}_p^{ac}(\mathbb{R})$, l'espace des mesures de $\mathcal{P}_p(\mathbb{R})$ qui sont absolument continues, alors $T^* := F_\nu^- \circ F_\mu$ est l'application *pushforward* optimale de μ à ν et dans ce cas, $W_p^p(\mu,\nu) = \int_{\mathbb{R}} |T^*(x) - x|^p d\mu(x)$.

La formulation de la distance de Wasserstein sur la droite réelle permet notamment de mieux comprendre ce qui la différencie des distances $\mathbb{L}_p$. Considérons deux mesures $\mu, \nu \in \mathcal{P}_p^{ac}(\mathbb{R})$ admettant des densités de probabilité $f_\mu, f_\nu : \mathbb{R} \to \mathbb{R}$. Alors une distance $\mathbb{L}_p$ est pertinente lorsque ces deux densités partagent le même support, et permet de comparer les variations en un point $x \in \mathbb{R}$. En revanche, dès lors que deux densités sont de support disjoints, leur distance $\mathbb{L}_p$ sera la même qu'elles soient proches ou non sur la droite réelle. Il est alors possible d'établir un parallèle entre la distance de Wasserstein (A.6) (définie comme une distance $\mathbb{L}_p([0,1])$ sur les quantiles) et la distance $\mathbb{L}_p(\mathbb{R})$. Afin d'illustrer les déplacements de masse entre deux mesures pour les distances W_2 et $\mathbb{L}_2$, nous présentons en Figure A.3 deux mélanges de gaussiennes, l'une comportant trois modes, l'autre deux, et nous réprésentons les chemins géodésiques entre ces deux mesures. Il apparaît que les métriques se comportent de façon tout à fait différente. La métrique $\mathbb{L}_2$ déplace la masse selon l'amplitude des mesures, et les densités de probabilité sur la géodésique présentent donc toutes cinq modes. La métrique W_2 quant à elle déplace la masse le long de la droite



Géodésique dans $\mathbb{L}_2(\mathbb{R})$ Géodésique dans $W_2(\mathbb{R})$

FIGURE A.3. Les densités rouges représentent deux mélanges de gaussiennes de support disjoints. Les dégradés de couleurs correspondent aux chemins géodésiques entre ces deux mesures, respectivement pour les métriques $\mathbb{L}_2(\mathbb{R})$ et $W_2(\mathbb{R})$.

réelle, et une densité le long de la géodésique transforme sa géométrie, passant de trois modes à deux modes, de gauche à droite.

De même, en dimensions supérieures, la distance de Wasserstein prend en compte la distance parcourue lors d'un transfert de masse, ce dont n'est pas capable une distance $\mathbb{L}_p$. L'importance du support est d'autant plus évidente lorsque l'on considère deux mesures de Dirac, pour lesquelles la distance de Wasserstein est donnée par la distance entre leur point de support.

A.3. La distance de Wasserstein sur un espace fini

Dans le cas discret, c'est-à-dire lorsque les mesures $\mu \in \mathcal{P}_p(\mathcal{X})$ sont à support sur un nombre fini de point, *i.e.* $\mathcal{X} = \{x_1, \dots, x_N\} \subset (\mathcal{X})^N$, on peut écrire $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ où $(a_1, \dots, a_N)$ est un vecteur de poids positifs appartenant au simplexe $\Sigma_N := \{a = (a_i)_{i=1,\dots,N} \in \mathbb{R}_+^N$ tel que $\sum_{i=1}^N a_i = 1\}$ et δ_{x_i} est la mesure de Dirac en x_i. Comme l'espace $\mathcal{X}$ est fixé, une mesure de probabilité sur $\mathcal{X}$ est entièrement caractérisée par un vecteur de poids dans le simplexe. Par abus de notation, nous identifions donc une mesure discrète $\mu \in \mathcal{P}_p(\mathcal{X})$ à son vecteur de poids $a = (a_1, \dots, a_n) \in \Sigma_N$ (et nous nous permettons d'écrire $a = \mu$). Le transfert de masse correspond alors à un problème d'optimisation linéaire et s'écrit pour $a, b \in \Sigma_N$

$$W_p(a,b) = \left(\min_{T \in U(a,b)} \langle T, C \rangle \right)^{1/p} \tag{A.7}$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire usuel entre matrices (*i.e.* soient A, B deux matrices réelles carrées, alors $\langle A, B \rangle = \mathrm{trace}(A^t B)$), et

- $U(a,b) = \{T \in \mathbb{R}_+^{N \times N} \ | T\mathbb{1}_N = a, T^T \mathbb{1}_N = b\}$ est l'ensemble des matrices de transport de marginales a et b (avec $\mathbb{1}_N$ représentant le vecteur de $\mathbb{R}^N$ dont les entrées sont toutes égales à 1) ,

- $C \in \mathbb{R}_+^{N\times N}$ est la matrice de coût éléments par éléments de $\mathcal{X}$ dont la (i,j)-ème coordonnée correpond à $C_{i,j} = d(x_i, x_j)^p$.

La version duale de ce problème est alors donnée par

$$W_p(a,b) = \left(\max_{\alpha,\beta\in\mathbb{R}^N,\ \alpha_i+\beta_j\leqslant C_{i,j}} \langle \alpha, a\rangle + \langle \beta, b\rangle \right)^{1/p}. \tag{A.8}$$

A.4. Le transport optimal régularisé par l'entropie

De nombreuses applications nécessitent de considérer des données sous la forme de mesures discrètes (ou d'histogrammes) sur un espace euclidien $\mathbb{R}^d$. La distance de Wasserstein s'est alors avérée être une mesure statistique pertinente dans différents domaines tels que le partitionnement de distributions discrètes [**YWWL17**], des modèles Bayésien non-paramétriques [**Ngu13**], la comparaison d'empreintes [**SM16**], l'apprentissage non supervisé [**ACB17**], l'analyse en composantes principales [**BGKL17**, **SC15**], le traitement d'images et l'apprentissage automatique [**FPPA14**, **BCC+15**, **CP16b**, **DPR16**], etc...

Dans ces cas, il est toujours possible de fixer une grille $\mathcal{X} = \{x_1, \dots, x_N\} \subset (\mathbb{R}^d)^N$ sur laquelle sont définies les mesures. Cependant, le coût de calcul d'une distance de transport (A.7) est de l'ordre de $\mathcal{O}(N^3 \log N)$. Il devient donc excessif pour des valeurs trop importantes de N. Régulariser un problème avec un terme d'entropie afin de réduire sa complexité est une approche classique en optimisation [**Wil69**]. Pour pallier le coût de calcul d'une distance de transport, Cuturi [**Cut13**] a donc proposé d'ajouter un terme de régularisation entropique au problème linéaire de transfert de masse, conduisant à la notion de transport optimal régularisé par l'entropie, ou divergence de Sinkhorn, entre mesures de probabilités discrètes. Initialement, le but de la régularisation était de calculer efficacement un terme proche de la distance de Wasserstein entre deux mesures de probabilité, via un algorithme itératif pour lequel chaque itération coûte $\mathcal{O}(N^2)$. Nous verrons par la suite que ce problème a aussi des effets de régularisation pouvant être bénéfique pour les données aberrantes (voir (B.19)).

Définition A.3. *La divergence de Sinkhorn est définie pour $a, b \in \Sigma_N$ et $\varepsilon > 0$ par*

$$W_{p,\varepsilon}^p(a,b) = \min_{U\in U(a,b)} \langle U, C\rangle - \lambda h(U) \tag{A.9}$$

où $h(U) = -\sum_{i,j} U_{ij} \log U_{ij}$ est l'entropie négative de la matrice de transport $U \in U(a,b)$.

Remarquons que la divergence de Sinkhorn ne définit pas une métrique sur l'espace de mesures discrètes dans $\mathcal{P}_p(\mathcal{X})$. En particulier, $W_{p,\varepsilon}^p(a,a)$ n'est pas nul. La formulation duale de (A.9) est alors donnée par [**Cut13**, **CD14**]

$$W_{p,\varepsilon}^p(a,b) = \max_{\alpha,\beta\in\mathbb{R}^N} \alpha^T a + \beta^T b - \sum_{i,j} \varepsilon e^{-\frac{1}{\varepsilon}(c_{ij}-\alpha_i-\beta_j)}, \tag{A.10}$$

Il existe une relation explicite entre les solutions optimales des problèmes primal (A.9) et dual (A.10) ci-dessus. Ces solutions peuvent par ailleurs être calculées par un algorithme itératif appelé algorithme de Sinkhorn [**CD14**].

Le transport régularisé par l'entropie a récemment gagné en popularité dans l'apprentissage automatique et les statistiques, car il rend possible l'utilisation d'une approximation de distances de transport pour l'analyse de données de grandes dimensions. Il a trouvé diverses applications telles que les modèles génératifs [**GPC17**], l'apprentissage multi-étiquettes [**FZM+15**], l'apprentissage de dictionnaires [**RCP16**] ou encore le traitement d'images, voir *e.g.* [**CP16b**, **RP15**], l'extraction de texte par comparaison de mots-clés [**GCB16**] et dans le moyennage de données de neuro-imagerie [**GPC15**]. Le livre de Cuturi et Peyré [**PFR12**]

présente une grande partie des applications propres au transport, et notamment au transport régularisé.

A.5. Inférence et distance de Wasserstein

A.5.1. Limite de mesures empiriques en distance de Wasserstein

Le cadre considéré est celui de n variables aléatoires $(\boldsymbol{X}_j)_{j=1,\dots,n}$ indépendantes et identiquement distribuées (iid) générées selon une mesure de probabilité inconnue $\mu \in \mathcal{P}(\mathbb{R}^d)$. On obtient alors la mesure dite empirique associée à l'échantillon d'observations, donnée par

$$\boldsymbol{\mu}_n = \sum_{j=1}^{n} \delta_{\boldsymbol{X}_j}.$$

En particulier, nous utiliserons des notations en gras $\boldsymbol{\nu}, \boldsymbol{X}, \boldsymbol{f}, \dots$ pour se référer à des objets aléatoires. La dérivation des limites distributionnelles de la mesure empirique $\boldsymbol{\mu}_n$ vers son équivalent en population μ en distance de Wasserstein, c'est-à-dire l'étude asymptotique de $W_p(\boldsymbol{\mu}_n, \mu)$ quand n tend vers l'infini, est bien comprise pour des mesures de probabilités à support sur $\mathbb{R}$, voir [**MC98**, **FM05**, **DBCAMRR99**, **DBGU05**, **DBCAM$^+$00**] pour n'en citer que quelques-uns. Ces résultats se basent sur la formulation quantile du transport en 1D. Par conséquent, les travaux menés ont permis de définir de nouvelles statistiques de test d'adéquation à une loi. Le cas unidimensionel est également traité dans le manuscrit de Bobkov et Ledoux [**BL14**], dans lequel ils fournissent une étude de la quantité $\mathbb{E}(W_p(\boldsymbol{\mu}_n, \mu))$. Ramdas et al. dans [**RTC17**] ont aussi étudié le lien entre les tests non-paramétriques et la distance de Wasserstein, mettant l'accent sur les distributions à support dans $\mathbb{R}$. Ces résultats ont été étendus à des distributions paramétriques spécifiques à support sur $\mathbb{R}^d$ et appartenant à une classe elliptique (cas gaussian en particulier), voir [**RMS16**] (et les références qui s'y trouvent). Panaretos et Zemel présentent notamment une revue de la littérature des outils statistiques dans l'espace de Wasserstein dans leur récent papier [**PZ18**]. Récemment, un théorème central limite a été établi en distance de Wasserstein dans [**DBL17**] pour des mesures empiriques échantillonées à partir de mesures absolument continues sur $\mathbb{R}^d$. Le cas de mesures discrètes à support sur un espace métrique fini a également été considéré dans [**SM16**], révélant la convergence (dans l'esprit du théorème central limite) des distances de Wasserstein empirique vers la valeur optimale d'un programme linéaire.

A.5.2. La moyenne de Fréchet dans l'espace de Wasserstein

Afin d'étudier un jeu de données composé de plusieurs sujets, le barycentre dans l'espace de Wasserstein $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, correspondant à la notion de moyenne de Fréchet [**Fré48**], est un outil statistique naturel. Cette moyenne est une extension du barycentre euclidien usuel à des espaces non linéaires. Comme introduit par Agueh et Carlier dans [**AC11**], un barycentre empirique de Wasserstein $\hat{\nu}_n$ d'un ensemble de n mesures de probabilité $\nu_1, \dots, \nu_n$ dans $\mathcal{P}_2(\mathbb{R}^d)$ est défini par

$$\hat{\nu}_n \in \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\Omega)} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i). \tag{A.11}$$

Une caractérisation détaillée de ces barycentres en termes d'existence, d'unicité et de régularité pour des mesures de probabilité dont le support est inclu dans $\mathbb{R}^d$ est disponible dans [**AC11**]. Le lien entre ces barycentres et la solution du problème de multi-marginales y est aussi étudié, ainsi que dans [**Pas13**].

Il est intéressant de remarquer que lorsque les mesures $\nu_1, \dots, \nu_n$ sont des distributions gaussiennes non dégénérées, leur barycentre est également une distribution gaussienne, et

cela est toujours vérifié pour un ensemble de mesures appartenant à des familles translations-dilatations (voir [ÁEdBCAM15b, ÁEdBCAM16]).

La notion de barycentre de Wasserstein a été en premier généralisée dans [LGL16] pour des mesures de probabilité aléatoires (voir aussi [ÁEdBCAM15a] pour des concepts similaires). Une mesure de probabilité $\boldsymbol{\nu}$ dans $\mathcal{P}_2(\mathbb{R}^d)$ est dite aléatoire si elle est générée à partir d'une distribution $\mathbb{P}$ sur $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{B}\left(\mathcal{P}_2(\mathbb{R}^d)\right)$, où $\mathcal{B}\left(\mathcal{P}_2(\mathbb{R}^d)\right)$ est la σ-algèbre de Borel générée par la topologie induite par la distance W_2.

Definition .1. *Soit $W_2(\mathcal{P}_2(\mathbb{R}^d))$ l'espace des distributions $\mathbb{P}$ sur $\mathcal{P}_2(\mathbb{R}^d)$ (muni de la distance de Wasserstein W_2) tel que pour un (et donc tout) $\mu \in \mathcal{P}_2(\Omega)$*

$$\mathcal{W}_2^2(\delta_\mu, \mathbb{P}) := \mathbb{E}_{\mathbb{P}}(W_2^2(\mu, \boldsymbol{\nu})) = \int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) < +\infty,$$

où $\boldsymbol{\nu} \in \mathcal{P}_2(\mathbb{R}^d)$ est une mesure aléatoire de distribution $\mathbb{P}$ et δ_μ est la mesure de Dirac au point μ. Le barycentre de Wasserstein d'une mesure de probabilité aléatoire de loi $\mathbb{P} \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ est donné par

$$\mu_{\mathbb{P}} \in \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{argmin} \int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu, \nu) d\mathbb{P}(\nu). \tag{A.12}$$

Les auteurs de [LGL16] ont établi l'existence, l'unicité et la consistance de barycentres de mesures de probabilité aléatoires à support sur un espace géodésique localement compact. Lorsque la mesure $\mathbb{P}_n = \frac{1}{n}\sum \delta_{\nu_i}$ est discrète sur $\mathcal{P}_2(\mathbb{R}^d)$, nous retrouvons bien $\mu_{\mathbb{P}_n}$ qui correspond au barycentre empirique (A.11). Dans le cas où $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ sont des mesures de probabilités aléatoires iid de loi $\mathbb{P}$, le barycentre $\mu_{\mathbb{P}}$ est appelé le barycentre de population.

Le cas plus général de mesures de probabilité à support sur une variété riemannienne a été étudié dans [KP17]. Par la suite, les *trimmed barycenters* dans l'espace de Wasserstein sur $\mathcal{P}_2(\mathbb{R}^d)$ ont été introduits dans [ÁEdBCAM15a], permettant de combiner des informations à partir d'unités expérimentales dans le cadre d'estimations parallélisées ou distribuées.

Le transport optimal est utilisé dans [PZ16] pour le recalage de processus ponctuels représentants un échantillon d'observations organisées en sujets indépendants. Les auteurs de [PZ16] ont proposé un estimateur consistent du barycentre de population de Wasserstein de processus ponctuels dans le cas $d = 1$, et une extension de leur méthodologie pour $d \geqslant 2$ est considérée dans [PZ17]. Leur méthode présente deux étapes. Un lissage par noyau est d'abord opéré sur les données, conduisant à un ensemble de mesures absolument contiues (*a.c.*) dont on calcule le barycentre de Wasserstein dans un second temps. Enfin, sont discutés dans [PZ16, BGKL18] des taux de convergence (pour la métrique de Wasserstein) du barycentre empirique de Wasserstein de mesures discrètes à support sur la ligne réelle.

En revanche, l'analyse statistique de distances de Wasserstein régularisées, et celle de barycentres également régularisés, est très peu présente dans la littérature.

A.6. Le recalage d'histogrammes

Les problèmes de recalade d'histogrammes trouvent des applications dans de nombreux domaines. En bio-informatique, les chercheurs veulent notamment normaliser automatiquement de grands jeux de données pour comparer et analyser des caractéristiques au sein d'une même population de cellules. Malheureusement, les informations acquises sont bruitées en raison d'un mauvais alignement, provoqué par des variations techniques de l'environnement. Le besoin de prendre en compte la variabilité de phase dans l'analyse statistique de tels jeux de données est un problème connu dans de nombreux domaines scientifiques. On trouve des exemples dans le cas unidimensionnel ($d = 1$) : études biodémographiques et génomiques [ZM11], économiques [KU01], analyse de l'activité neuronale en neurosciences [WS11] ou connectivité fonctionelle entre les régions du cerveau [PM+16b]. En

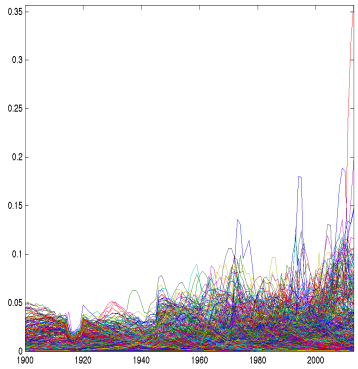


FIGURE A.4. Proportion d'enfants nés avec un certain prénom par an en France de 1900 à 2013. Chaque courbe représente un prénom.
Source : INSEE.

dimension supérieure, *i.e.* $d \geqslant 2$, le problème de recalage de données provient par exemple de l'analyse statistique des processus ponctuels spatiaux [**Ger16**, **PZ17**] ou des données de cytométrie de flux [**HKB$^+$10**, **PLW$^+$14**].

Le transport optimal permet de corriger les effets de mauvais alignements au sein d'un jeu de données, cependant, son utilité n'a été exploitée que par peu d'auteurs [**PZ17**].

A.7. L'analyse en composantes principales d'histogrammes

Il est toujours possible d'appliquer une ACP (fonctionelle) standard sur un ensemble de densités de probabilité $(f_i)_{i=1,\dots,n}$ vus comme des fonctions de $\mathbb{L}_2(\mathbb{R})$ en diagonalisant l'opérateur de covariance $\mathrm{Cov} : \mathbb{L}_2(\mathbb{R}) \mapsto \mathbb{L}_2(\mathbb{R})$ défini par

$$\mathrm{Cov}(h) = \frac{1}{n}\sum_{i=1}^{n}\langle f_i - \bar{f}_n, h\rangle(f_i - \bar{f}_n), \quad h \in \mathbb{L}_2(\mathbb{R})$$

où $\bar{f}_n$ est la moyenne euclidienne de $f_1, \dots, f_n \in \mathbb{L}_2(\mathbb{R})$. Les vecteurs propres de Cov associés aux plus grandes valeurs propres décrivent les principaux modes de variabilité des données autour de la moyenne $\bar{f}_n$. Ainsi, les premier et second modes de variations sont donnés par les courbes $g^{(j)} : \mathbb{R} \to \mathbb{L}_2(\mathbb{R}), j = 1, 2$ par

$$g_t^{(j)} = \bar{f}_n + tw_j, \quad t \in \mathbb{R}$$

où $w_1 \in \mathbb{L}_2(\mathbb{R})$ (resp. w_2) est le vecteur propre associé à la plus grande valeur (resp. seconde plus grande) propre de l'opérateur de covariance Cov. Afin d'illustrer ces variations, nous considérons le jeu de donnée de prénoms de la Figure A.4 (source : Insee). Ce jeu de donnée est composé d'histogrammes représentant le nombre d'enfants nés par an pour un prénom donné, entre 1900 et 2013 en France (chaque histogramme est normalisé sur le support [1900,2013]), voir Figure A.5 pour des exemples. Ainsi un histogramme est la donnée de 114 années. Nous disposons dans ce jeu de données de $n = 1060$ prénoms, pour un ensemble de personnes variant de 10077 à 1920210 par prénom. Les modes de variations dans $\mathbb{L}_2(\mathbb{R})$ obtenus sont présentés dans la Figure A.6.

Les résultats d'ACP fonctionelle sont très insatisfaisants pour plusieurs raisons. Premièrement, les fonctions obtenues $g_t^{(j)}$ ne sont pas des densités de probabilité, elles prennent notamment des valeurs négatives. Deuxièmement, la métrique $\mathbb{L}_2$ ne prend en compte que les variations d'amplitude des données.

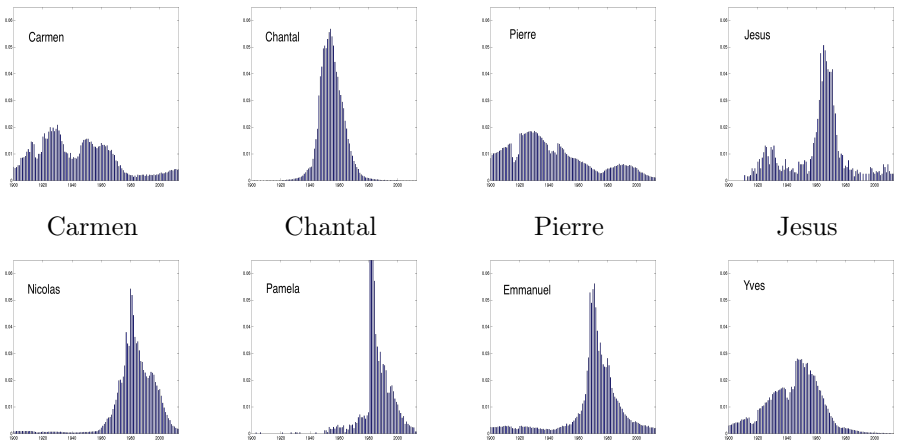
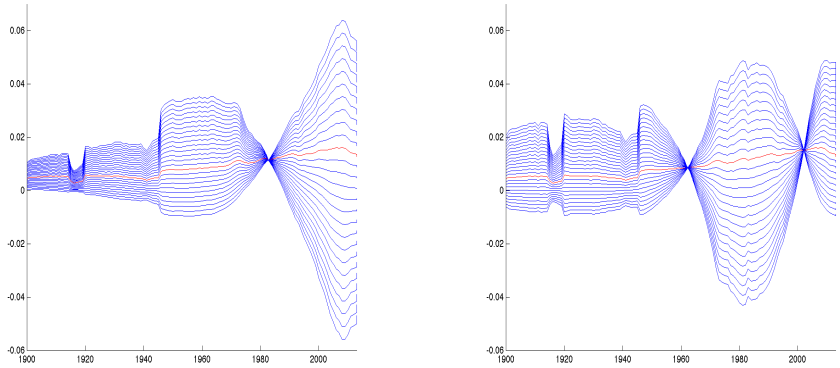


FIGURE A.5. Un histogramme représente la proportion d'enfants nés avec un certain prénom par an en France de 1900 à 2013. Source : INSEE



$g_t^{(1)} = \bar{f}_n + tw_1$, pour $-0.15 \leqslant t \leqslant 0.12$ $\quad g_t^{(2)} = \bar{f}_n + tw_2$, pour $-0.16 \leqslant t \leqslant 0.09$

FIGURE A.6. Les deux premiers modes de variations dans $\mathbb{L}_2(\mathbb{R})$ pour les données de prénoms obtenus via une ACP fonctionelle dans $\mathbb{L}_2(\mathbb{R})$. La courbe rouge représente le barycentre euclidien des données.

Afin de pallier ces deux inconvénients, il est intéressant de travailler directement sur l'espace de mesures de probabilité (admettant un moment d'ordre deux fini) muni de la distance 2-Wasserstein. Cet espace n'est cependant pas hilbertien. Par conséquent, l'ACP standard, qui implique le calcul d'une matrice de covariance, ne peut pas être appliquée directement pour calculer les modes de variation principaux au sens de Wasserstein. Néanmoins, une notion significative d'ACP peut encore être définie en s'appuyant sur la structure pseudo-riemannienne de l'espace de Wasserstein, qui a été largement étudiée dans [**AGS04**] et [**AGS08**]. Suivant ce principe, une structure pour l'analyse en composantes géodésiques principales (ACGP) de mesures de probabilités à support sur un intervalle $\Omega \subset \mathbb{R}$ a été introduite dans [**BGKL17**]. L'ACPG est définie comme le problème de l'estimation d'un sous-espace géodésique principal (d'une dimension donnée) qui maximise la variance de la projection des données dans ce sous-espace. Dans cette approche, le point de base du

sous-espace est le barycentre de Wasserstein $\hat{f}_n$ des données f_i tel qu'il a été introduit dans [**AC11**]. L'existence, la cohérence et une caractérisation détaillée de l'ACPG dans $\mathcal{P}_2(\Omega)$ ont été étudiées dans [**BGKL17**]. En particulier, les auteurs ont montré que cette approche équivaut à projeter les données dans l'espace tangent de $\mathcal{P}_2(\Omega)$ à la moyenne de Fréchet, puis d'effectuer une ACP dans cet espace de Hilbert, tout en contraignant le problème à un sous-ensemble convexe et fermé de fonctions. Projeter les données dans cet espace tangent n'est pas difficile dans le cas unidimensionnel puisqu'il s'agit de calculer un ensemble de cartes optimales T entre les données et leur barycentre de Wasserstein, pour lequel une forme explicite est disponible, via les fonctions de répartition et les fonctions quantiles (voir par exemple [**Vil03**, §2.2]). Les auteurs de [**BGKL17**] n'ont pas proposé d'algorithme pour résoudre le problème d'ACGP, qui consiste à minimiser une fonction non-convexe et non-différentiable, sur un espace de contraintes convexes. Dans [**BGKL17**], seule une approximation numérique du calcul des composantes géodésiques principales a été proposée. L'approche consiste à appliquer une log-ACP, à savoir une ACP standard de l'ensemble de données projetées préalablement dans l'espace tangent de $\mathcal{P}_2(\Omega)$ à sa moyenne de Fréchet $\bar{\nu}_n$. La log-ACP des données de prénoms est présentée dans la Figure A.7. Les résultats sont plus conformes aux attentes, représentant les effets de translation (première composante, gauche) et les effets d'amplitude (deuxième composante, droite) du jeu de données.

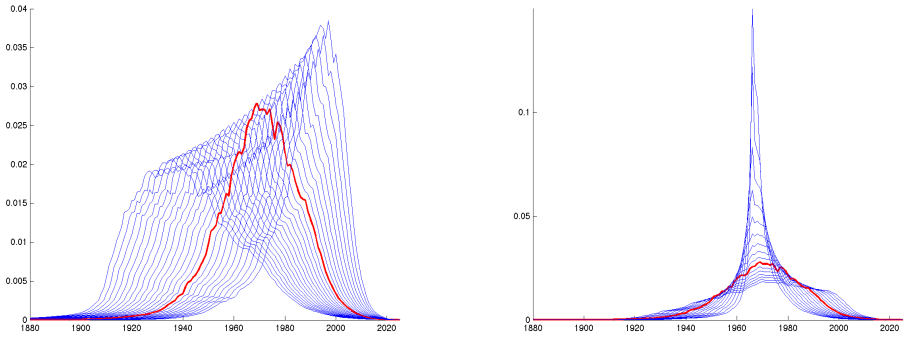


FIGURE A.7. Les deux premiers modes de variations pour les données de prénoms obtenus la log-ACP d'histogrammes. La courbe rouge représente le barycentre de Wasserstein des données.

B. Problématiques et principales contributions de la thèse

Le cadre suivi dans cette thèse est l'analyse d'éléments pouvant être décrits pas des mesures de probabilité (discrètes ou absolument continues) aléatoires à support sur $\mathbb{R}^d$. Nous étudions donc des jeux de données composés de n mesures discrètes $\boldsymbol{\nu}_{p_1}, \ldots, \boldsymbol{\nu}_{p_n}$ obtenues à partir d'observations aléatoires

$$\boldsymbol{X} = (\boldsymbol{X}_{i,j})_{1 \leqslant i \leqslant n;\ 1 \leqslant j \leqslant p_i}, \tag{B.13}$$

organisées sous la forme de n sujets (ou unités expérimentales), telles que $\boldsymbol{\nu}_{p_i}$ est définie par

$$\boldsymbol{\nu}_{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}. \tag{B.14}$$

B.1. Barycentres de Wasserstein pénalisés par une fonction convexe

Le barycentre de Wasserstein $\hat{\boldsymbol{\nu}}_n$, défini en (A.11), de mesures $(\boldsymbol{\nu}_{p_i})_{i=1,\dots,n}$ (B.14) construites à partir d'observations aléatoires $\boldsymbol{X}$ (B.13) peut souffrir d'irrégularités dues par exemple aux données aberrantes, ou encore au manque d'observations sur les mesures $\boldsymbol{\nu}_{p_i}$. Plus précisément, nous n'avons généralement accès qu'à un jeu de données de variables aléatoires générées à partir de distributions inconnues absolument continues. La théorie (voir [**AC11**]) assure que lorsqu'au moins une des mesures ν_i est absolument continue (*a.c.*) par rapport à la mesure de Lebesgue, alors le barycentre de Wasserstein le sera également. Cependant, il n'y a aucune raison que le barycentre obtenu à partir d'observations discrètes vérifie cette règle, et une régularisation s'avère nécessaire pour forcer l'absolue continuité.

Prenons l'exemple de données de flux de cytométrie obtenues à partir du réseau de recherche immune (Immune Tolerance Network[1]) et présentées dans la Figure B.8. Cet échantillon est constitué de 15 patients. Pour chacun d'eux, nous disposons d'un nombre réduit, entre 88 et 2185, de mesures de cellules, pour lesquelles les valeurs des marqueurs FSC et SSC sont récupérées (voir Section II.1.2.3 pour plus de détails concernant ce jeu de données). À partir d'un tel échantillon, nous aimerions retrouver la distribution bidimensionnelle sous-jacente, a priori absolument continue, des marqueurs FSC (*forward-scattered light*) et SSC (*side-scattered light*) des patients.

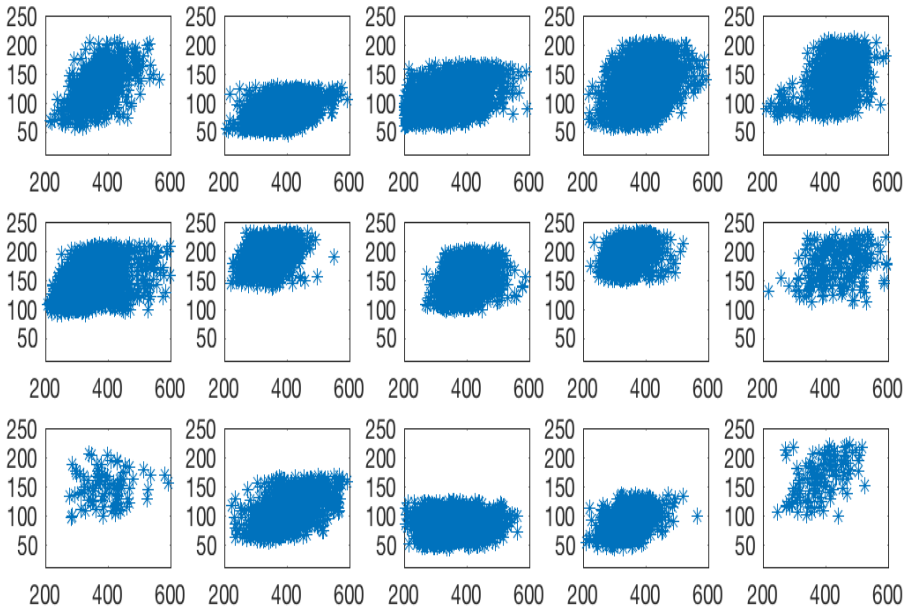


FIGURE B.8. Exemple de données de cytométrie de flux mesurées à partir de $n = 15$ patients. L'axe horizontal (resp. axe vertical) représente les valeurs du marqueur FSC (resp. SSC).

La première contribution de cette thèse réside dans l'introduction du barycentre de Wasserstein pénalisé par une fonction de pénalité convexe E pour des mesures définies sur

[1] http://bioconductor.org/packages/release/bioc/html/flowStats.html

un convexe Ω de $\mathbb{R}^d$:

$$\mu_{\mathbb{P}}^{\gamma} = \underset{\mu \in \mathcal{P}_2(\Omega)}{\arg\min} \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu), \tag{B.15}$$

pour $\gamma > 0, \mathbb{P} \in \mathcal{P}_2(\Omega)$ (possiblement discrète). Soulignons ici que l'on se concentre principalement sur des fonctions de pénalité E qui forcent les minimiseurs (B.15) à être *a.c.* et de fonction de densité lisse. Ce problème de pénalisation de barycentre est motivé par la méthode non-paramétrique introduite dans [**BFS12**] dans le cas classique d'estimation de densité à partir d'échantillons discrets.

Nous prouvons dans un premier travail, l'existence et l'unicité de ces minimiseurs pour une large classe de fonctions $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$, avec $\Omega \subset \mathbb{R}^d$ convexe. Dans un second temps, nous montrons que l'introduction d'un terme de pénalisation dans le calcul des barycentres de Wasserstein de mesures discrètes permet de construire un estimateur consistent d'un barycentre de population *a.c.* Plus précisément, nous nous plaçons dans le cadre de n mesures $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_n$ de loi $\mathbb{P}$ définies en (B.14) pour un échantillon d'observations $\boldsymbol{X}_{i,1}, \dots, \boldsymbol{X}_{i,p_i}$ dans Ω définies en (B.13). Le barycentre pénalisé empirique de Wasserstein est alors défini par

$$\hat{\boldsymbol{\mu}}_{n,p}^{\gamma} = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \boldsymbol{\nu}_{p_i}) + \gamma E(\mu). \tag{B.16}$$

Nous étudions alors la convergence de cet estimateur $\hat{\boldsymbol{\mu}}_{n,p}^{\gamma}$ vers son équivalent en population $\mu_{\mathbb{P}}^{\gamma}$ (B.15) en terme de divergence de Bregman d_E associé à la fonction de pénalité E. Les divergences de Bregman sont en effet largement utilisées pour comparer des mesures *a.c.* (par exemple en géométrie de l'information [**AN00**]). Ainsi on obtient le théorème suivant.

Théorème B.4 (Théorème I.17). *Pour $\Omega \subset \mathbb{R}^d$ compact, et pour tout $\gamma > 0$, on a*

$$\lim_{n \to \infty} \mathbb{E}(d_E^2(\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}, \mu_{\mathbb{P}}^{\gamma})) = 0 \tag{B.17}$$

Des résultats plus fins, donnant une borne sur la variance de l'estimateur $\hat{\boldsymbol{\mu}}_{n,p}^{\gamma}$ du barycentre pénalisé ont également été obtenus, en invoquant plus de régularité via la fonction de pénalité E.

En adaptant l'algorithme de [**CP16b**], qui se base sur une descente de gradient de la version duale du problème de transport, nous fournissons un algorithme permettant de calculer ces barycentres.

B.2. Barycentres de Wasserstein régularisés par l'entropie

Une autre façon de régulariser un barycentre de Wasserstein consiste à utiliser la régularisation entropique du transport présentée en (A.9). Cett approche mène au barycentre de Sinkhorn [**CD14**, **CP16b**, **CDPS17**, **BCC+15**]. Pour cela, on considère n mesures aléatoires *iid* discrètes $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n \in \Sigma_N$ générées à partir d'une distribution $\mathbb{P} \in \Sigma_N$. Ainsi, pour chaque $1 \leqslant i \leqslant n$, on suppose que les observations $(\tilde{X}_{i,j})_{1 \leqslant j \leqslant p_i}$ sont des variables aléatoires *iid* de loi $\boldsymbol{q}_i$. Nous définissons alors pour $\varepsilon > 0$

$$\begin{aligned} r^{\varepsilon} &= \underset{r \in \Sigma_N}{\arg\min}\, \mathbb{E}_{\boldsymbol{q} \sim \mathbb{P}}[W_{2,\varepsilon}^2(r, \boldsymbol{q})] && \text{le barycentre population de Sinkhorn} \\ \hat{\boldsymbol{r}}_{n,p}^{\varepsilon} &= \underset{r \in \Sigma_N}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} W_{2,\varepsilon}^2(r, \hat{\boldsymbol{q}}_i^{p_i}) && \text{le barycentre empirique de Sinkhorn} \end{aligned} \tag{B.18}$$

qui correspondent à des moyennes de Fréchet par rapport à la divergence de Sinkhorn.Comme on peut le voir sur la Figure B.9, le paramètre ε a un effet de lissage sur le barycentre empirique $\hat{\boldsymbol{r}}_{n,p}^{\varepsilon}$ obtenu à partir de deux mélanges de gaussiennes. Plus le paramètre ε augmente, plus la masse du barycentre s'étale. Ainsi la pénalisation entropique n'a plus seulement un intérêt calculatoire (afin d'accélérer le calcul d'une distance de transport), mais devient un

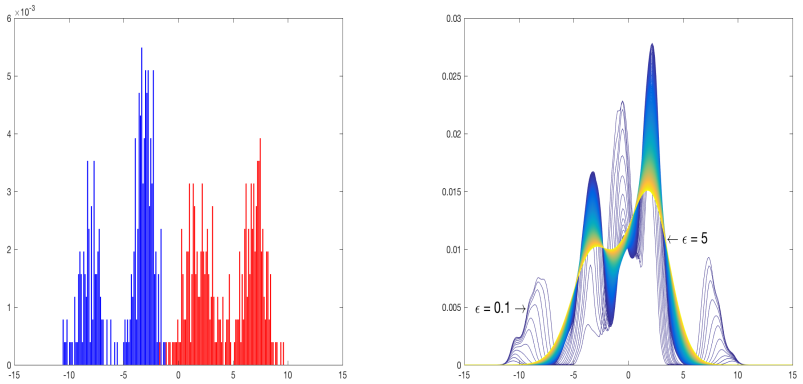


FIGURE B.9. Un exemple simulé de $n = 2$ sujets construits avec $p_1 = p_2 = 300$ observations générées à partir de mélanges de gaussiennes de moyennes et variances aléatoires. (Gauche) Les graphes bleu et rouge sont des histogrammes d'intervalles de variation égaux et petits. (Droite) 400 barycentres de Sinkhorn $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ pour ε variant de 1 à 5. Les couleurs encodent la variation de ε.

véritable outil de régularisation. Nous avons prouvé grâce à la forte convexité de la divergence de Sinkhorn qu'il est possible d'obtenir une borne sur la variance de l'estimateur de ce barycentre de Sinkhorn. Pour cela, il est nécessaire de restreindre l'analyse à des mesures discrètes appartenant à l'espace

$$\Sigma_N^{\rho} = \left\{ r \in \Sigma_N \ : \ \min_{1 \leqslant \ell \leqslant N} r_\ell \geqslant \rho \right\},$$

et de considérer un barycentre appartenant à cet espace.

THÉORÈME B.5 (Théorème I.22). *Soit* $p = \min_{1 \leqslant i \leqslant n} p_i$ *et* $\varepsilon > 0$. *Alors*

$$\mathbb{E}(|r^{\varepsilon} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}|^2) \leqslant \frac{32L^2}{\varepsilon^2 n} + \frac{2L}{\varepsilon}\sqrt{\frac{N}{p}},$$

avec

$$L_{\rho,\varepsilon} = \left(\sum_{1 \leqslant m \leqslant N} \left(2\varepsilon \log(N) + \sup_{1 \leqslant \ell,k \leqslant N} |C_{m\ell} - C_{k\ell}| - 2\varepsilon \log(\rho) \right)^2 \right)^{1/2}. \tag{B.19}$$

B.3. Application à l'alignement d'histogrammes

Les barycentres régularisés peuvent notamment être utilisés pour faire face au problème de recalage d'histogrammes, une mesure $\boldsymbol{\nu}_i$ représentant alors un histogramme de données. Cette approche diffère de celle de [**PZ16**, **PZ17**] présentée en sous-section A.6 car nous incluons ici directement l'étape de lissage dans le calcul du barycentre de Wasserstein.

Le problème qui se pose alors est le choix des paramètres de régularisation γ dans (B.16) et ε dans (B.18) lors du calcul des barycentres. Pour cela, la méthode de Goldenshuler-Lepski (GL) (comme formulée dans [**LM16**]) propose une solution en se basant sur les bornes de la variance des barycentres régularisés, permettant alors une calibration automatique des paramètres de régularisation. Dans la Figure B.10, la fonction de compromis biais-variance

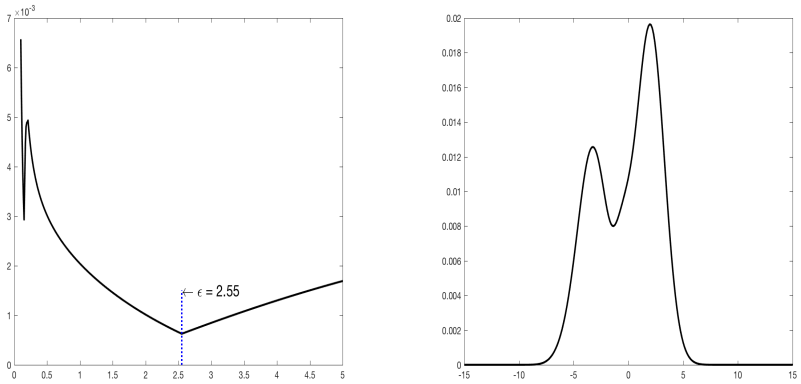


FIGURE B.10. (Gauche) Fonction de compromis biais-variance donnée par la méthode de Goldenshuler-Lepski, associé aux mélanges de gaussiennes de la Figure B.9. (Droite) Barycentre de Sinkhorn optimal associé à $\varepsilon = 2.55$.

GL est tracée pour les barycentres de Sinkhorn de la Figure B.9, nous donnant le paramètre optimal de régularisation $\varepsilon = 2.55$, et le barycentre optimal associé (Figure B.10, gauche).

B.4. Tests statistiques

Au cours de la thèse, nous avons également étudié la convergence en divergence de Sinkhorn des mesures de probabilité empiriques à support sur un espace métrique fini. Ce travail a été motivé par la nécessité de trouver des compromis aux tests statistiques basé sur la distance de transport. En effet, hormis pour le cas unidimensionnel ($d = 1$), les distances de transports mènent à des statistiques de test dont l'implémentation numérique peut devenir excessive pour des mesures empiriques à support sur $\mathbb{R}^d$ avec $d \geqslant 2$. Par conséquent, utiliser des statistiques de tests basées sur les divergences de Sinkhorn peut présenter un intérêt pratique. Les travaux menés se concentrent donc sur l'étude de l'inférence statistique de mesures discrètes en terme de transport régularisé par l'entropie.

Nous obtenons de nouveaux résultats sur la distribution asymptotique de ces divergences pour des données échantillonnées à partir de distributions (inconnues) à support sur un espace métrique fini. Nos résultats sont inspirés du travail de [**SM16**] sur la distribution asymptotique de la distance de Wasserstein empirique sur un espace fini en terme de coût de transport non régularisé. L'application principale consiste à obtenir de nouvelles statistiques de test (pour un ou deux échantillons) pour la comparaison de distributions de probabilité multivariées.

Enfin, pour illustrer l'applicabilité de cette approche, nous proposons également une procédure bootstrap pour estimer des quantités d'intérêt inconnues dans le calcul de ces statistiques de test.

B.5. Analyse en composantes géodésiques principales

Nous avons finalement proposé de comparer les méthodes de log-ACP et d'ACGP comme introduites dans [**BGKL17**,**SC15**]. Dans notre approche, les histogrammes sont vus comme des densités de probabilité constantes par morceaux à support sur un intervale $\Omega \subset \mathbb{R}$ donné. Dans ce contexte, les modes de variation d'un ensemble d'histogrammes peuvent être étudiés à travers la notion d'ACP géodésique de mesures de probabilité dans l'espace de Wasserstein

$\mathcal{P}_2(\Omega)$ admettant ces histogrammes pour densité. Comme précisé précédemment en sous-section (A.7), cette approche a été proposée dans la littérature statistique [**BGKL17**] pour les mesures de probabilité sur la droite réelle et dans l'apprentissage automatique [**SC15**, **WSB+13**] dans le cas de mesures de probabilité discrètes sur $\mathbb{R}^d$. Cependant, l'exécution de l'ACGP reste compliquée même dans le cas le plus simple de densités de probabilité à support sur $\mathbb{R}$.

Nous avons alors fourni un algorithme rapide pour effectuer l'ACGP de mesures définies sur la droite réelle, et nous comparons ses résultats à ceux de la log-ACP [**FLPJ04**,**PM16a**]. L'ACP géodésique consiste à résoudre un problème d'optimisation non convexe. Pour le résoudre approximativement, nous proposons un nouvel algorithme forward-backward. Nous présentons aussi une comparaison détaillée entre la log-ACP et l'ACP géodésique d'histogrammes unidimensionnels, pour différents ensembles de données en dimensions 1 et 2.

C. Contenu de la thèse

Chapitre I Dans un premier chapitre, nous introduisons deux barycentres régularisés, l'un via une fonction de pénalité convexe, l'autre en utilisant le transport optimal régularisé par l'entropie. Une étude de ces moyennes et des résultats de convergence de leurs estimateurs sont présentés, issus des papiers [**BCP18b**, **BCP18a**].

Chapitre II Les barycentres régularisés du Chapitre I sont alors utilisés pour le problème de recalage d'histogrammes et nous proposons une méthode pour calibrer automatiquement les paramètres de régularisation, [**BCP18a**].

Chapitre III Nous énonçons un théorème central limite du transport optimal régularisé par l'entropie dans ce troisième chapitre. Nous en déduisons des statistiques de test d'adéquation à des lois pour des histogrammes multivariés, [**BCP17**].

Chapitre IV Enfin, nous développons de nouveaux algorithmes pour le problème d'analyse en composantes géodésiques principales dans l'espace de Wasserstein, provenant du papier [**CSB+18**].

Introduction (English)

In this introduction, we present a - non exhaustive - review of the literature on optimal transport, as well as its many applications in data analysis. We also introduce the definitions and notations that will be used throughout this thesis. We then present a detailed summary of our work and the contents of this manuscript.

A. Optimal transport and applications

A.1. Monge, Kantorovich and Wasserstein

Gaspard Monge introduced in 1781 the problem of mass transfer in his *Mémoire sur la théorie des déblais et des remblais*. He aimed to find the most efficient way, that is requiring the least possible effort, to transport a pile of sand in a hole of the same volume. In its modern formulation, the problem consists in finding the optimal measurable map for transferring at a lower cost the mass of a probability measure μ supported on a measure space $\mathcal{X}$ on another measure ν supported on $\mathcal{Y}$. Then, Monge's problem boils down to

$$\text{minimize} \qquad \int_{\mathcal{X}} c(x, T(x)) d\mu(x) \tag{A.1}$$

over the set of measurable functions $T : \mathcal{X} \to \mathcal{Y}$ such that $\nu = T\#\mu$. This *pushforward* operator $\#$ is defined such that for any measurable set $B \subset \mathcal{Y}$, we have $\nu(B) = \mu(T^{-1}(B))$. The function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is a measurable cost function. An example of mass transfer, in the spirit of the *déblais et remblais* problem of Monge, is presented in Figure A.1.

However, such applications T do not always exist, especially if the mass of μ at a given point in $\mathcal{X}$ musts split into several pieces. To overcome this restriction, Leonid Kantorovich extended in the 1940s the Monge problem (A.1) by introducing a transport plan between the starting measure and the target measure, which contains the behavior of mass transfer. This corresponds to

$$\text{minimize} \qquad \iint_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) \tag{A.2}$$

over the set of transport plans π belonging to $\Pi(\mu, \nu)$, *i.e.* the set of product measures on $\mathcal{X} \times \mathcal{Y}$ with respective marginals μ and ν. When considering discrete measures, e.g. Figure A.2, the transport plan can allocate the mass of a point of the support of μ at different points of the support of ν, whereas it can not be done with a map T. The notion of optimal

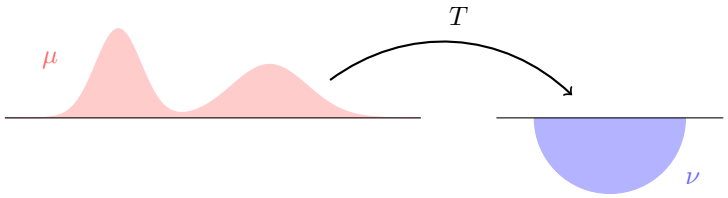


FIGURE A.1. Transfer of the mass of μ onto the mass of ν through the map T such that $\nu = T\#\mu$.

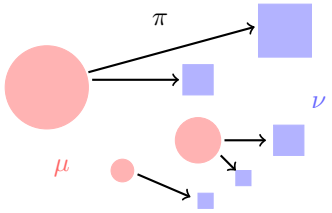


FIGURE A.2. Transfer of the mass of μ onto the mass of ν through a transport plan $\pi \in \Pi(\mu, \nu)$.

transport, as well as the geometric and differential points of view of these minimization problems, are detailed in the works of Villani [**Vil03**, **Vil08**], Ambrosio and Gigli [**AG13**] and Ambrosio and al. [**AGS04**].

A particularly interesting framework of optimal transport appears when $\mathcal{X}$ is a Polish space endowed with a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. Indeed, in this case, the Kantorovich optimal transport problem between two measures defines a distance for a cost $c := d^p$, as soon as the measures belong to a proper space. More precisely, for $p \in [1, +\infty)$, we denote $\mathcal{P}_p(\mathcal{X})$ the set of Borel probability measures (also called distributions) on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ supported on $\mathcal{X}$, where $\mathcal{B}(\mathcal{X})$ is the σ-algebra of the Borel subsets included in $\mathcal{X}$, admitting a moment of order p. In other words,

$$\mu \in \mathcal{P}_p(\mathcal{X}) \text{ is equivalent to } \int_{\mathcal{X}} d^p(x_0, x) d\mu(x) < +\infty \text{ for any } x_0 \in \mathcal{X}.$$

Note that $\mathcal{P}_p(\mathcal{X})$ is included in the set of bounded Radon measures $\mathcal{M}(\mathcal{X})$. We then get the following definition.

DEFINITION A.1. *The p-Wasserstein distance (1969, Leonid Wasserstein) is given for μ, ν in $\mathcal{P}_p(\mathcal{X})$ by*

$$W_p(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \iint_{\mathcal{X}^2} d^p(x,y) d\pi(x,y) \right)^{1/p} \tag{A.3}$$

where the infimum is taken over the set $\Pi(\mu,\nu)$ of transport plans on $\mathcal{X} \times \mathcal{X}$ with respective marginals μ and ν.

This distance has the advantage to characterize the weak convergence of measures on the metric space $(\mathcal{P}_p(\mathcal{X}), W_p)$ (see *e.g.* Chapter 7 in [**Vil03**]).

Kantorovich also described the problem of minimization (A.3) in its dual formulation, corresponding to a constrained optimization over a function space. We recall that the space $\mathbb{L}_p(\mu)$, for $p \in [1,\infty)$ and $\mu \in \mathcal{M}(\mathcal{X})$, is the space of functions $f : \mathcal{X} \to \mathbb{R}$ such that $|f|^p$ is μ-integrable, and such that all functions that are equal μ-almost everywhere are identified. The dual problem of (A.3) is then given by the following theorem.

THEOREM A.2 (Kantorovich's duality theorem). *Let $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, then we have*

$$W_p(\mu,\nu) = \left(\sup_{(\phi,\psi) \in C_{W_p}} \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{X}} \psi(y) d\nu(y) \right)^{1/p}, \tag{A.4}$$

where C_{W_p} is the set of measurable functions $(\phi,\psi) \in \mathbb{L}_1(\mu) \times \mathbb{L}_1(\nu)$ satisfying

$$\phi(x) + \psi(y) \leq d^p(x,y), \tag{A.5}$$

for μ-almost every $x \in \mathcal{X}$ and ν-almost every $y \in \mathcal{X}$.

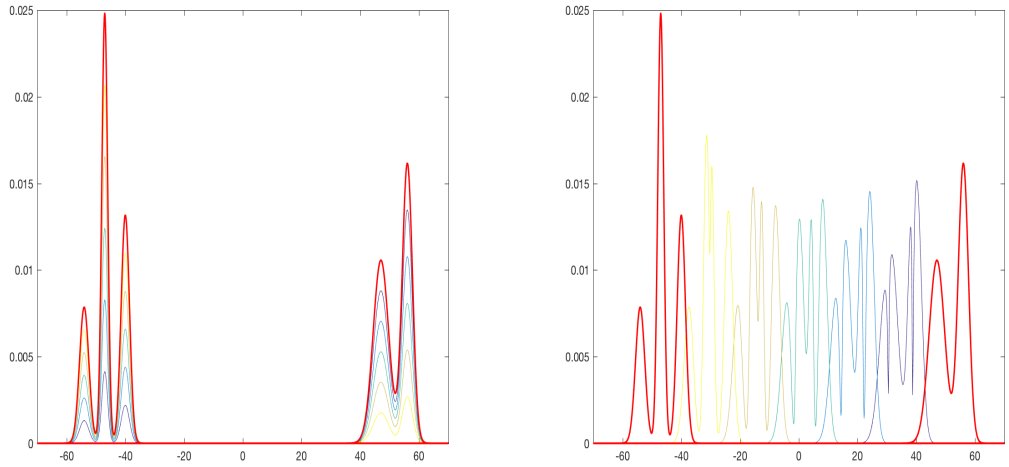
A.2. The Wasserstein distance on the real line

The case of measures supported on the real line, namely when $\mathcal{X}$ is an interval (possibly unbounded) of $\mathbb{R}$, is significant since the distance of Wasserstein W_p is then equal to the $\mathbb{L}_p$ distance of quantile functions. Formally, by denoting F_μ the distribution function of μ and F_μ^- its generalized quantile, the Wasserstein distance becomes, for $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$,

$$W_p(\mu,\nu) = \left(\int_0^1 |F_\mu^-(t) - F_\nu^-(t)|^p dt \right)^{1/p}. \tag{A.6}$$

If $\mu \in \mathcal{P}_p^{ac}(\mathbb{R})$, the space of measures in $\mathcal{P}_p(\mathbb{R})$ that are absolutely continuous, then $T^* := F_\nu^- \circ F_\mu$ is the optimal *pushforward* application from μ to ν, and in this case, $W_p^p(\mu,\nu) = \int_{\mathbb{R}} |T^*(x) - x|^p d\mu(x)$.

The formulation of Wasserstein distance on the real line makes it possible to better understand its distinctions with $\mathbb{L}_p$ distances. Consider two measures $\mu, \nu \in \mathcal{P}_p^{ac}(\mathbb{R})$ with probability density functions (*pdf*) $f_\mu, f_\nu : \mathbb{R} \to \mathbb{R}$. A $\mathbb{L}_p$ distance is relevant when these two densities share the same support, as it allows to compare the variations at a point $x \in \mathbb{R}$ of the support. On the other hand, if the supports of two densities are disjoint, their $\mathbb{L}_p$ distance will be the same whether they are close or not on the real line. It is possible to draw a parallel between the Wasserstein distance (A.6) (defined as a $\mathbb{L}_p([0,1])$ distance on the quantiles) and the distance $\mathbb{L}_p(\mathbb{R})$. In order to illustrate the mass displacements between two measures for the W_2 and $\mathbb{L}_2$ distances, we present in Figure A.3 two mixtures of Gaussian, one having three modes, the other two modes, and we represent the geodesic paths between these two measures. The metrics have a completely different behavior. The $\mathbb{L}_2$ metric moves the mass according to the amplitude of the pdf, and the pdfs on the geodesic thus all have 5 modes. On the other hand, the W_2 metric moves the mass along the real line, and the geometry of a density along the geodesic varies from three modes to two modes, from left to right.



Geodesic in $\mathbb{L}_2(\mathbb{R})$ Geodesic in $W_2(\mathbb{R})$

FIGURE A.3. The red probability density functions represent two mixtures of Gaussian with disjoint supports. Colors' gradations correspond to the geodesic paths between the two measures, for the $\mathbb{L}_2(\mathbb{R})$ and $W_2(\mathbb{R})$ metrics respectively.

Similarly, in higher dimensions, the Wasserstein distance takes into account the distance that a mass has to travel, which is not possible for a $\mathbb{L}_p$ distance. The importance of the support is all the more evident when one considers two Dirac measures, for which the Wasserstein distance is given by the distance between their support point.

A.3. The Wasserstein distance on a finite space

In the discrete setting, when the measures $\mu \in \mathcal{P}_p(\mathcal{X})$ are supported on a finite number of points, *i.e.* $\mathcal{X} = \{x_1, \dots, x_N\} \subset \mathcal{X}^N$, one can write $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ where $(a_1, \dots, a_N)$ is a vector of positive weights belonging to the simplex $\Sigma_N := \{a = (a_i)_{i=1,\dots,N} \in \mathbb{R}_+^N$ such that $\sum_{i=1}^N a_i = 1\}$ and δ_{x_i} is the Dirac measure at x_i. As the space $\mathcal{X}$ is considered to be fixed, a probability measure supported on $\mathcal{X}$ is entirely characterized by a vector of weights in the simplex. By a slight abuse of notation, we thus identify a measure $\mu \in \mathcal{P}_p(\mathcal{X})$ by its vector of weights $a = (a_1, \dots, a_n) \in \Sigma_N$ (and we sometimes write $a = \mu$). The optimal transport problem (A.3) then corresponds to a linear optimization problem and reads for $a, b \in \Sigma_N$

$$W_p(a,b) = \min_{T \in U(a,b)} \langle T, C \rangle^{1/p} \tag{A.7}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product between matrices (*i.e.* let A, B be two real squared matrices, then $\langle A, B \rangle = \mathrm{trace}(A^t B)$), and

- $U(a,b) = \{T \in \mathbb{R}_+^{N \times N} \ | T\mathbb{1}_N = a, T^T \mathbb{1}_N = b\}$ is the set of transport matrices with marginals a and b (with $\mathbb{1}_N$ denoting the vector of $\mathbb{R}^N$ with all entries equal to one),
- $C \in \mathbb{R}_+^{N \times N}$ is the pairwise cost matrix associated to the space $\mathcal{X}$ whose (i,j)-th entry is $C_{i,j} = d(x_i, x_j)^p$.

The dual version of this problem is then given by

$$W_p(a,b) = \left(\max_{\alpha, \beta \in \mathbb{R}^N, \ \alpha_i + \beta_j \leqslant C_{i,j}} \langle \alpha, a \rangle + \langle \beta, b \rangle \right)^{1/p}. \tag{A.8}$$

A.4. The entropy regularized optimal transport

Many applications need to consider data in the form of discrete measures (or histograms) on a Euclidean space $\mathbb{R}^d$. The Wasserstein distance then proved to be a relevant statistical measure in different domains such as clustering of discrete distributions [**YWWL17**], non-parametric Bayesian models [**Ngu13**], fingerprints comparison [**SM16**], unsupervised learning [**ACB17**], principal component analysis [**BGKL17**, **SC15**], image processing and machine learning [**FPPA14**, **BCC+15**, **CP16b**, **DPR16**], etc ...

In these cases, it is always possible to set a grid $\mathcal{X} = \{x_1, \dots, x_N\} \subset (\mathbb{R}^d)^N$ on which the measures are defined. However, the cost of computing a transport distance (A.7) is of order $\mathcal{O}(N^3 \log N)$. It thus becomes excessive for large values of N. Regularizing a problem with an entropy term to reduce its complexity is a classic approach in optimization [**Wil69**]. To overcome the cost of computing a transport distance, Cuturi [**Cut13**] has therefore proposed to add an entropy regularization term to the linear optimal transport problem, leading to the notion of entropy regularized optimal transport, or Sinkhorn divergence, between discrete measures. Initially, the goal of the regularization was to efficiently compute a term close to the Wasserstein distance between two probability measures, via an iterative algorithm for which each iteration costs $\mathcal{O}(N^2)$. We will see later that this problem also has regularization effects that may be beneficial for outliers (see (B.18)).

Definition A.3. *The Sinkhorn divergence is defined for $a, b \in \Sigma_N$ and $\varepsilon > 0$ by*

$$W^p_{p,\varepsilon}(a, b) = \min_{U \in U(a,b)} \langle U, C \rangle - \lambda h(U) \tag{A.9}$$

where $h(U) = -\sum_{i,j} U_{ij} \log U_{ij}$ is the negative entropy of the transport matrix $U \in U(a, b)$.

Let us notice that the Sinkhorn divergence does not define a metric on the space of discrete measures included in $\mathcal{P}_p(\mathcal{X})$. In particular, $W^p_{p,\varepsilon}(a, a)$ is not zero. The dual formulation of (A.9) is then given by [**Cut13**, **CD14**]

$$W^p_{p,\varepsilon}(a, b) = \max_{\alpha,\beta \in \mathbb{R}^N} \alpha^T a + \beta^T b - \sum_{i,j} \varepsilon e^{-\frac{1}{\varepsilon}(c_{ij} - \alpha_i - \beta_j)}. \tag{A.10}$$

There is an explicit relationship between the optimal solutions of the primal (A.9) and dual (A.10) problems above. These solutions can also be computed by an iterative algorithm called Sinkhorn algorithm [**CD14**].

Entropy regularized optimal transport has recently gained popularity in machine learning and statistics because it makes it possible to use an approximation of transport distances for the analysis of large dataset. It has found various applications such as generative models [**GPC17**] and more generally for high dimensional data analysis in multi-label learning [**FZM+15**], dictionnary learning [**RCP16**] and image processing (see e.g. [**CP16b**, **RP15**] and references therein), text mining via bag-of-words comparison [**GCB16**], averaging of neuroimaging data [**GPC15**]. The book by Cuturi and Peyré [**PFR12**] presents a large part of applications specific to optimal transport, and in particular to regularized transport.

A.5. Inference and Wasserstein distance

A.5.1. Limit of empirical measures in Wasserstein distance

One can consider n random variables $(\boldsymbol{X}_j)_{j=1,\dots,n}$ independent and identically distributed (*iid*) generated according to an unknown probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$. We then obtain the so-called empirical measure associated to the sample of observations, given

by

$$\boldsymbol{\mu}_n = \sum_{j=1}^{n} \delta_{\boldsymbol{X}_j}.$$

In particular, we will use notations in bold $\boldsymbol{\nu}, \boldsymbol{X}, \boldsymbol{f}, \ldots$ to refer to random objects. The derivation of distributional limits of the empirical measure $\boldsymbol{\mu}_n$ towards its equivalent in population μ in Wasserstein distance, namely the asymptotic study of $W_p(\boldsymbol{\mu}_n, \mu)$ when n tends to infinity, is well understood for probability measures with support on $\mathbb{R}$, see [**MC98**, **FM05**, **DBCAMRR99**, **DBGU05**, **DBCAM+00**] to name a few. These results are based on the quantile formulation of the one-dimensional transport. Therefore, these works have led to the development of new test statistics. The one-dimensional case is also treated in the paper of Bobkov and Ledoux [**BL14**], in which they provide a study of the quantity $\mathbb{E}(W_p(\boldsymbol{\mu}_n, \mu))$. Ramdas and al. in [**RTC17**] have also investigated the link between non-parametric tests and Wasserstein distance, with an emphasis on distributions supported on $\mathbb{R}$. These results have been extended to specific parametric distributions with support on $\mathbb{R}^d$ and belonging to an elliptic class (Gaussian case in particular), see [**RMS16**] (and references therein). Panaretos and Zemel present a review of the literature of statistical tools in the Wasserstein space in their recent paper [**PZ18**]. Also, a central limit theorem has been established in Wasserstein distance in [**DBL17**] for empirical measures sampled from absolutely continuous measures on $\mathbb{R}^d$. The case of discrete measures with support on a finite metric space has also been considered in [**SM16**], revealing the convergence (in the spirit of the central limit theorem) of the empirical Wasserstein distances towards the optimal value of a linear program.

A.5.2. The Fréchet mean in the Wasserstein space

In order to study a dataset composed of several subjects, the barycenter in the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, corresponding to the notion of Fréchet mean [**Fré48**], is a natural statistical tool. This average is an extension of the usual Euclidean barycenter to non-linear spaces. As introduced by Agueh and Carlier in [**AC11**], a Wasserstein empirical barycenter $\hat{\nu}_n$ of a set of n probability measures $\nu_1, \ldots, \nu_n$ in $\mathcal{P}_2(\mathbb{R}^d)$ is defined by

$$\hat{\nu}_n \in \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i). \tag{A.11}$$

A detailed characterization of these barycenters in terms of existence, uniqueness and regularity for probability measures with support included in $\mathbb{R}^d$ is available in [**AC11**]. The authors, as well as the author of [**Pas13**], study the link between these barycenters and the solutions of multi-marginals optimal transport problem.

It is interesting to note that when the measures $\nu_1, \ldots, \nu_n$ are non degenerate Gaussian distributions, their barycenter is also a Gaussian distribution, and this is still true for a set of measures belonging to translations-dilations families (see [**ÁEdBCAM15b**, **ÁEdBCAM16**]).

The notion of Wasserstein barycenter has been generalized in [**LGL16**] for random probability measures (see also [**ÁEdBCAM15a**] for similar concepts). A probability measure $\boldsymbol{\nu}$ in $\mathcal{P}_2(\mathbb{R}^d)$ is called random if it is generated from a distribution $\mathbb{P}$ on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{B}\left(\mathcal{P}_2(\mathbb{R}^d)\right))$, where $\mathcal{B}\left(\mathcal{P}_2(\mathbb{R}^d)\right)$ is the Borel σ-algebra generated by the topology induced by the distance W_2.

Definition A.4. *Let $W_2(\mathcal{P}_2(\mathbb{R}^d))$ be the space of distributions $\mathbb{P}$ on $\mathcal{P}_2(\mathbb{R}^d)$ (endowed with the Wasserstein distance W_2) such that for one (and thus for any) $\mu \in \mathcal{P}_2(\mathbb{R}^d)$*

$$\mathcal{W}_2^2(\delta_\mu, \mathbb{P}) := \mathbb{E}_{\mathbb{P}}(W_2^2(\mu, \boldsymbol{\nu})) = \int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) < +\infty,$$

where $\boldsymbol{\nu} \in \mathcal{P}_2(\mathbb{R}^d)$ *is a random measure with distribution* $\mathbb{P}$ *and* δ_μ *is a Dirac measure at the point* μ*. The Wasserstein barycenter of a random probability measure of law* $\mathbb{P} \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ *is given by*

$$\mu_{\mathbb{P}} \in \underset{\mu \in \mathcal{P}_2(\Omega)}{argmin} \int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu, \nu) d\mathbb{P}(\nu). \tag{A.12}$$

The authors of [**LGL16**] have established the existence, uniqueness, and consistency of barycenters of random probability measures supported on a locally compact geodesic space. When the measure $\mathbb{P}_n = \frac{1}{n}\sum \delta_{\nu_i}$ is discrete on $\mathcal{P}_2(\mathbb{R}^d)$, we get $\mu_{\mathbb{P}_n}$, which corresponds to the empirical barycenter (A.11). In the case where $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are random probability measures *iid* of law $\mathbb{P}$, the barycenter $\mu_{\mathbb{P}}$ is called the population barycenter. The more general case of probability measures with support on a Riemannian manifold has been studied in [**KP17**]. Subsequently, the trimmed barycenters in the Wasserstein space on $\mathcal{P}_2(\mathbb{R}^d)$ were introduced in [**ÁEdBCAM15a**] for the purpose of combining information from different experimental units in a parallelized or distributed estimation setting.

Tools from optimal transport are used in [**PZ16**] for the registration of point processes organized in samples from independent subjects (or experimental units). The authors of [**PZ16**] have proposed a consistent estimator of the population Wasserstein barycenter of point processes in the case $d = 1$, and an extension of their methodology is considered for $d \geqslant 2$ in [**PZ17**]. Their method contains two steps. First, a kernel smoothing is performed on the data, which leads to a set of *a.c.* measures. A Wasserstein barycenter is then computed from this set of measures. Finally, rates of convergence (for the Wasserstein metric) of the empirical Wasserstein barycenter or discrete measures supported on the real line are discussed in [**PZ16**, **BGKL18**].

However, there is a lack of statistical analysis of regularized Wasserstein distances and regularized Wasserstein barycenters in the literature.

A.6. Registration of histograms

The problem of registering histograms finds applications in many fields. In bio-informatics, researchers aims to automatically normalize large datasets to compare and analyze characteristics within a single cell population. Unfortunately, the information acquired is noisy due to misalignment, caused by technical variations of the environment. The need to take into account phase variability in the statistical analysis of such datasets is a known problem in many scientific fields. There are examples in the one-dimensional case ($d = 1$): biodemographic and genomic studies [**ZM11**], economics [**KU01**], spike trains analysis in neuroscience [**WS11**] or functional connectivity between the brain regions [**PM$^+$16b**]. In higher dimension, *i.e.* $d \geqslant 2$, the problem of data registration comes for example from the statistical analysis of spatial point processes [**Ger16**, **PZ17**] or flow cytometry data [**HKB$^+$10**, **PLW$^+$14**].

Optimal transport allows to correct mis-alignment effects in a dataset, but has only been exploited by a few authors [**PZ17**].

A.7. Principal component analysis of histograms

It is always possible to apply a standard (functional) PCA on a set of probability density functions (pdf) $(f_i)_{i=1,\ldots,n}$ seen as functions in $\mathbb{L}_2(\mathbb{R})$ by diagonalizing the covariance operator $\mathrm{Cov} : \mathbb{L}_2(\mathbb{R}) \mapsto \mathbb{L}_2(\mathbb{R})$ defined by

$$\mathrm{Cov}(h) = \frac{1}{n}\sum_{i=1}^{n} \langle f_i - \bar{f}_n, h\rangle (f_i - \bar{f}_n), \quad h \in \mathbb{L}_2(\mathbb{R})$$

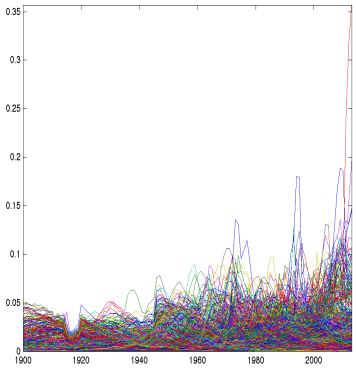


FIGURE A.4. Proportion of children born with a given name per year in France between 1900 and 2013. Each curve represents a first name. Source: INSEE

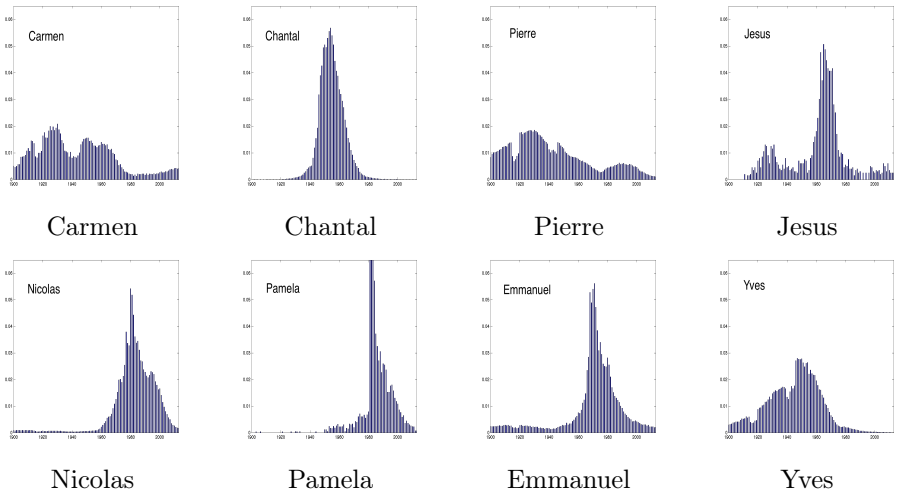


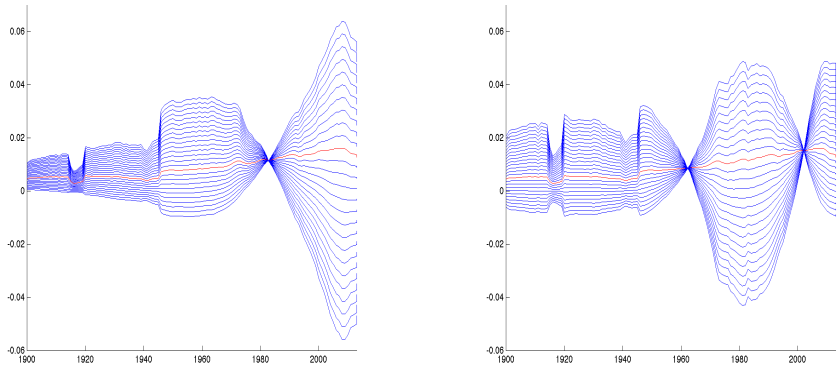
FIGURE A.5. A histogram represents the proportion of children born with a given name by year in France between 1900 and 2013. Source: INSEE.

where $\bar{f}_n$ is the Euclidean mean of $f_1, \ldots, f_n \in \mathbb{L}_2(\mathbb{R})$. The eigenvectors of Cov associated to the largest eigenvalues describe the main modes of variation of the data around the mean $\bar{f}_n$. Hence the first and second modes of variation are given by the curves $g^{(j)} : \mathbb{R} \to \mathbb{L}_2(\mathbb{R}), j = 1, 2$ with

$$g_t^{(j)} = \bar{f}_n + t w_j, \quad t \in \mathbb{R}$$

where $w_1 \in \mathbb{L}_2(\mathbb{R})$ (resp. w_2) is the eigenvector associated to the largest (resp. the second largest) eigenvalue of the covariance operator Cov. To illustrate these variations, we consider the dataset of given names in Figure A.4 (source: Insee). This dataset is composed of histograms representing the number of children born per year for a given name, between 1900 and 2013 in France (each histogram is normalized on the support [1900,2013]), see Figure A.5 for examples. A histogram thus contains information during 114 years. We have in this dataset $n = 1060$ first names, for a set of people ranging from 10077 to 1920210 per name. The mode of variations obtained with PCA in $\mathbb{L}_2(\mathbb{R})$ are presented in Figure A.6.

Functional PCA results are very unsatisfactory for several reasons. First, the functions obtained $g_t^{(j)}$ are not pdf, in particular they take negative values. Secondly, the $\mathbb{L}_2$ metric only takes into account the amplitude variation of the data.



$g_t^{(1)} = \bar{f}_n + tw_1$, pour $-0.15 \leqslant t \leqslant 0.12$ $\quad g_t^{(2)} = \bar{f}_n + tw_2$, pour $-0.16 \leqslant t \leqslant 0.09$

FIGURE A.6. The first two modes of variations in $\mathbb{L}_2(\mathbb{R})$ for the names dataset for a functional PCA in $\mathbb{L}_2(\mathbb{R})$. The red curve represent the Euclidean barycenter of the dataset.

In order to overcome these two drawbacks, the idea would be to work directly on the probability measures space (admitting a finite second order moment) endowed with the 2 -Wasserstein distance. Unfortunately, it is not a Hilbert space. Therefore, standard PCA, which involves computing a covariance matrix, can not be applied directly to compute principal mode of variations in a Wasserstein sense. Nevertheless, a meaningful notion of PCA can still be defined by relying on the pseudo-Riemannian structure of the Wasserstein space, which is extensively studied in [**AGS04**] and [**AGS08**]. Following this principle, a framework for geodesic principal conponent analysis (GPCA) of probability measures supported on a interval $\Omega \subset \mathbb{R}$ is introduced in [**BGKL17**]. GPCA is defined as the problem of estimating a principal geodesic subspace (of a given dimension) which maximizes the variance of the projection of the data to that subspace. In that approach the base point of that subspace is the Wasserstein barycenter $\bar{\nu}_n$ of the data f_i as introduced in [**AC11**]. Existence, consistency and a detailed characterization of GPCA in $\mathcal{P}_2(\Omega)$ are studied in [**BGKL17**]. In particular, the authors have shown that this approach is equivalent to map the data in the tangent space of $\mathcal{P}_2(\Omega)$ at the Fréchet mean, and then to perform a PCA in this Hilbert space, that is constrained to lie in a convex and closed subset of functions. Mapping the data to this tangent space is not difficult in the one-dimensional case as it amounts to computing a set of optimal maps T between the data and their Wasserstein barycenter, for which a closed form is available using their quantile functions (see for example [**Vil03**, §2.2]). To perform PCA on the mapped data, the authors of [**BGKL17**] fell short of proposing an algorithm to minimize that problem, which has a non-convex and non-differentiable objective function. Only a numerical approximation to the computation of GPCA has been proposed in [**BGKL17**], which amounts to applying log-PCA, namely a standard PCA of the dataset mapped beforehand to the tangent space of $\mathcal{P}_2(\Omega)$ at its Fréchet mean $\bar{\nu}_n$. The log-PCA of the names dataset is displayed in Figure A.7. The results correspond more to the expectations, as they include translation effects (first component, left) and amplitude effects (second component, right) of the dataset.

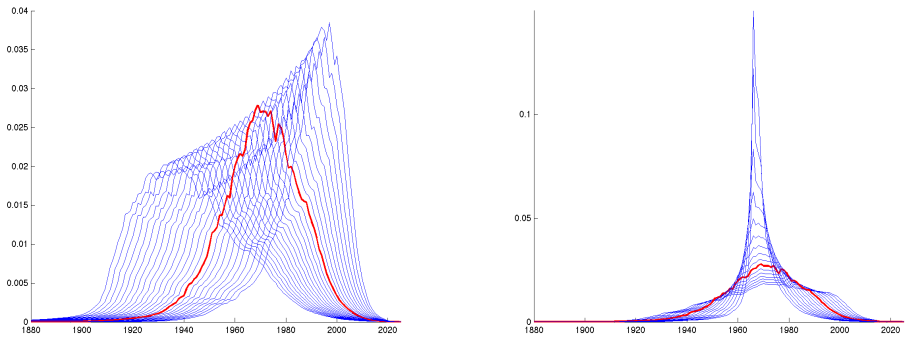


FIGURE A.7. The first two modes of variations for the names dataset obtained from the log-PCA of histograms. The red curve represent the Wasserstein barycenter of the dataset.

B. Problems and main contributions of the thesis

The framework followed in this thesis is the analysis of elements that can be described by random probability measures (discrete or absolutely continuous) with support on $\mathbb{R}^d$. We therefore study datasets composed of n discrete measures $\boldsymbol{\nu}_{p_1}, \ldots, \boldsymbol{\nu}_{p_n}$ obtained from random observations

$$\boldsymbol{X} = (\boldsymbol{X}_{i,j})_{1 \leqslant i \leqslant n;\ 1 \leqslant j \leqslant p_i}, \tag{B.13}$$

organized in the form of n subjects (or experimental units), such that $\boldsymbol{\nu}_{p_i}$ is defined by

$$\boldsymbol{\nu}_{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}. \tag{B.14}$$

B.1. Wasserstein barycenter penalized by a convex function

The Wasserstein barycenter $\hat{\boldsymbol{\nu}}_n$, defined in (A.11), of measures $(\boldsymbol{\nu}_{p_i})_{i=1,\ldots,n}$ (B.14) constructed from random observations $\boldsymbol{X}$ (B.13) may suffer from irregularities due for example to outliers, or lack of observations for the measures $\boldsymbol{\nu}_{p_i}$. Specifically, we generally only have access to a set of random variables generated from absolutely continuous unknown distributions. The theory (see [**AC11**]) ensures that when at least one of the measures ν_i is absolutely continuous (*a.c.*) with respect to the Lebesgue measure, then the Wasserstein barycenter will also be *a.c.*. However, there is no reason that the barycenter obtained from discrete observations satisfies this rule, and regularization is necessary to enforce absolute continuity.

Consider the example of flow cytometry data available from the Immune Tolerance Network[1] and presented in Figure B.8. This dataset consists of 15 patients. For each of them, we dispose of a small number, between 88 and 2185, of cell measurements, for which the values of the FSC and SSC markers are retrieved (see Section II.1.2.3 for more details on this dataset). From such a sample, we would like to find the underlying two-dimensional distribution, which should be absolutely continuous, of the FSC (forward-scattered light) and SSC (side-scattered light) cell markers of the patients.

[1] http://bioconductor.org/packages/release/bioc/html/flowStats.html

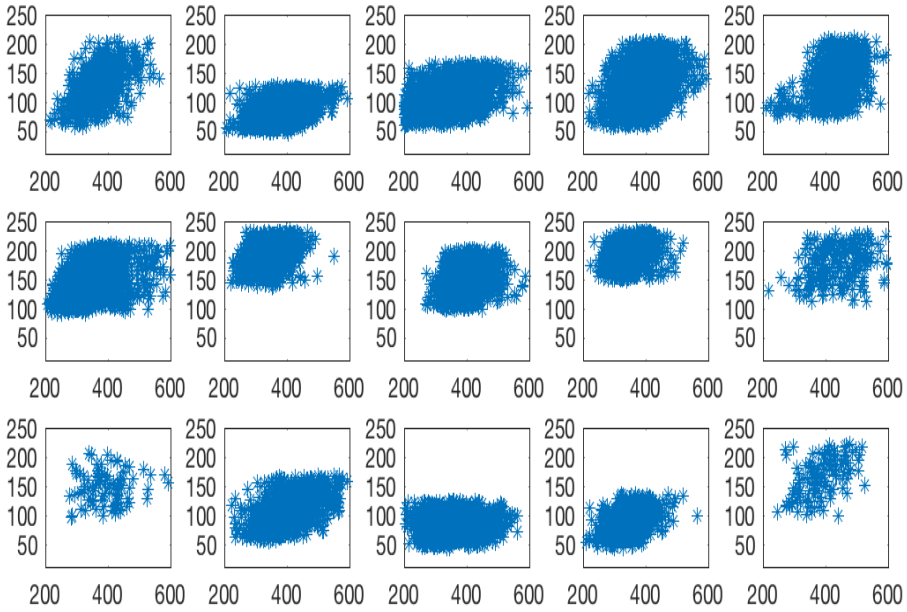


FIGURE B.8. Example of flow cytometry data measured from $n = 15$ patients (restricted to a bivariate projection). The horizontal axis (resp. vertical axis) represent the values of the FSC (resp. SSC) cell marker.

The first contribution of this thesis lies in the introduction of Wasserstein barycenter penalized by a convex penalty function E for measures defined on a convex set $\Omega \subset \mathbb{R}^d$:

$$\mu_{\mathbb{P}}^{\gamma} = \underset{\mu \in \mathcal{P}_2(\Omega)}{\arg\min} \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu), \tag{B.15}$$

for $\gamma > 0, \mathbb{P} \in \mathcal{P}_2(\Omega)$ (possibly discrete). Note that we mainly focus on penalty functions E that enforce the minimizers of (B.15) to be *a.c.* with a smooth pdf. This barycenter penalization problem is motivated by the non-parametric method introduced in [**BFS12**] in the classical case of density estimation from discrete samples.

First we prove the existence and uniqueness of these minimizers for a large class of functions $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$, with $\Omega \subset \mathbb{R}^d$ convex. Second, we demonstrate that the introduction of a penalization term in the computation of the Wasserstein barycenter of discrete measures allows to build a consistent *a.c.* estimator of a population barycenter. More precisely, we consider n measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ of law $\mathbb{P}$ defined in (B.14) for a sample of observations $\boldsymbol{X}_{i,1}, \ldots, \boldsymbol{X}_{i,p_i}$ in Ω defined in (B.13). The empirical penalized Wasserstein barycenter is then defined by

$$\hat{\boldsymbol{\mu}}_{n,p}^{\gamma} = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \boldsymbol{\nu}_{p_i}) + \gamma E(\mu). \tag{B.16}$$

We study the convergence of this estimator $\hat{\boldsymbol{\mu}}_{n,p}^{\gamma}$ towards its population counterpart $\mu_{\mathbb{P}}^{\gamma}$ (B.15) in terms of Bregman divergence d_E associated to the penalty function E. The Bregman divergences are indeed widely used to compare *a.c.* measures (for example in geometry of information [**AN00**]). Thus we obtain the following theorem.

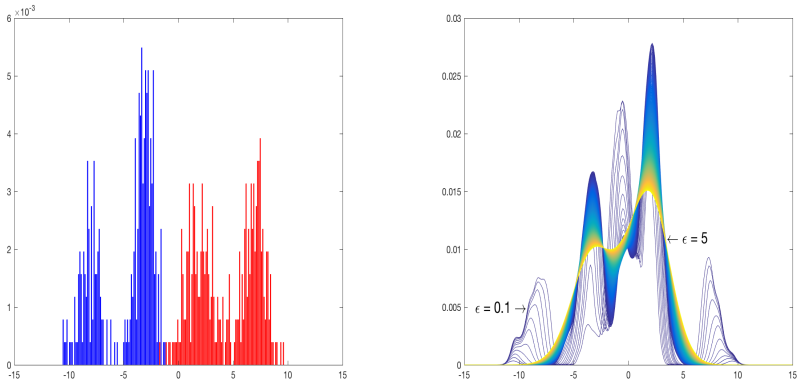


FIGURE B.9. A simulated example of $n = 2$ subjects obtained with $p_1 = p_2 = 300$ observations sampled from gaussian mixtures with random means and variances. (Left) The red and blue bar graphs are histograms with bins of equal and very small size to display the two sets of observations. (Right) 400 Sinkhorn barycenters $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ for ε ranging from 1 to 5. Colors encode the variation of ε.

THEOREM B.5 (Theorem I.17). *For $\Omega \subset \mathbb{R}^d$ compact, and for all $\gamma > 0$, we have*

$$\lim_{n\to\infty} \mathbb{E}(d_E^2(\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}, \mu^{\gamma}_{\mathbb{P}})) = 0 \tag{B.17}$$

Finer results, on the bound on the variance of the estimator $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ of the penalized barycenter, have also been obtained, by relying on more regularity through the penalty function E.

By adapting the algorithm of [**CP16b**], which is based on a gradient descent of the dual version of optimal transport, we provide an algorithm to compute these barycenters.

B.2. Entropy regularized Wasserstein barycenters

Another way of regularizing a Wasserstein barycenter consists in using the entropy regularization of the transport presented in (A.9). This approach leads to the Sinkhorn barycenter [**CD14**, **CP16b**, **CDPS17**, **BCC+15**]. For this purpose, we consider n discrete random measures $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n \in \Sigma_N$ *iid* generated from a distribution $\mathbb{P} \in \Sigma_N$. Thus, for each $1 \leqslant i \leqslant n$, we assume that the observations $(\tilde{X}_{i,j})_{1\leqslant j\leqslant p_i}$ are *iid* random variables of law $\boldsymbol{q}_i$. We then define for $\varepsilon > 0$

$$\begin{aligned} r^{\varepsilon} &= \operatorname*{arg\,min}_{r\in\Sigma_N} \mathbb{E}_{\boldsymbol{q}\sim\mathbb{P}}[W^2_{2,\varepsilon}(r, \boldsymbol{q})] && \text{the population Sinkhorn barycenter} \\ \hat{\boldsymbol{r}}^{\varepsilon}_{n,p} &= \operatorname*{arg\,min}_{r\in\Sigma_N} \frac{1}{n}\sum_{i=1}^{n} W^2_{2,\varepsilon}(r, \hat{\boldsymbol{q}}_i^{p_i}) && \text{the empirical Sinkhorn barycenter} \end{aligned} \tag{B.18}$$

which correspond to Fréchet means with respect to the Sinkhorn divergence. As it can be seen in Figure B.9, the parameter ε has a smoothing effect on the empirical barycenter $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ obtained from two Gaussian mixtures. The higher the parameter ε, the more the mass of the barycenter spreads. Thus the entropy penalization is no longer only of computational interest (in order to speed up the computation of a transport distance), but becomes a real tool of regularization. We have proved, thanks to the strong convexity of the Sinkhorn divergence, that it is possible to obtain a bound on the variance of the estimator of this

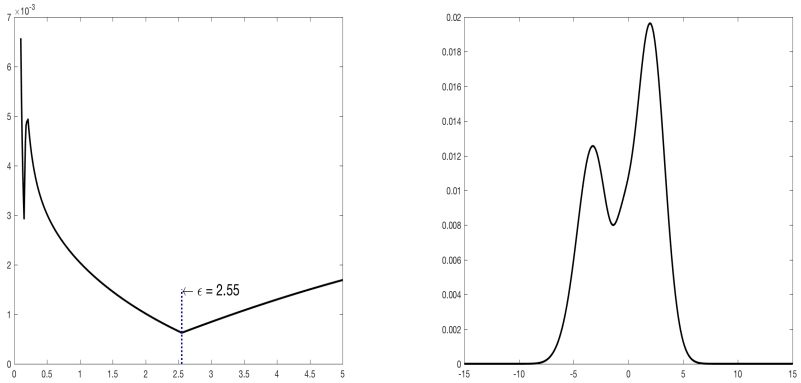


FIGURE B.10. (Left) Bias-variance trade-off function given by the Goldenshuler-Lepski method, associated to the Gaussian mixtures of Figure B.9. (Right) Optimal Sinkhorn barycenter in the GL sense, for which $\varepsilon = 2.55$.

Sinkhorn barycenter. For this purpose, it is necessary to restrict the analysis to discrete measures belonging to the space

$$\Sigma_N^\rho = \left\{ r \in \Sigma_N \ : \ \min_{1 \leqslant \ell \leqslant N} r_\ell \geqslant \rho \right\},$$

and to consider barycenter that lies on this space.

THEOREM B.6 (Theorem I.22). *Let $p = \min_{1 \leqslant i \leqslant n} p_i$ and $\varepsilon > 0$. Then*

$$\mathbb{E}(|r^\varepsilon - \hat{\boldsymbol{r}}^\varepsilon_{n,p}|^2) \leqslant \frac{32L^2}{\varepsilon^2 n} + \frac{2L}{\varepsilon}\sqrt{\frac{N}{p}},$$

with

$$L_{\rho,\varepsilon} = \left(\sum_{1 \leqslant m \leqslant N} \left(2\varepsilon \log(N) + \sup_{1 \leqslant \ell,k \leqslant N} |C_{m\ell} - C_{k\ell}| - 2\varepsilon \log(\rho) \right)^2 \right)^{1/2}. \tag{B.19}$$

B.3. Application to the registration of histograms

Regularized barycenters can be used to tackle the registration of histograms, a measure $\boldsymbol{\nu}_i$ then representing a histogram. This approach differs from the one of [**PZ16**, **PZ17**] presented in subsection A.6 since we directly include the smoothing step in the computation of the Wasserstein barycenter.

The problem that arises is the automatic choice of the regularizations parameters γ in (B.16) and ε in (B.18) for the computation of the barycenters. The Goldenshluger-Lepski (GL) method (as formulated in [**LM16**]) suggests a solution based on the derivation of upper bounds on the variance for the regularized barycenters, allowing a data-driven choice for the regularization parameters. In Figure B.10 (left), the bias-variance trade-off function is displayed for the Sinkhorn barycenters of Figure B.9, giving the optimal parameter $\varepsilon = 2.55$, and the associated barycenter (Figure B.10, right).

B.4. Statistic tests

During this thesis, we have also studied the convergence in entropy regularized optimal transport of empirical probability measures supported on a finite metric space. This work

was motivated by the need to find compromises for statistical tests based on transport distances. Indeed, except for the one-dimensional case ($d = 1$), transport distances lead to test statistics whose numerical implementation can become excessive for empirical measures with support on $\mathbb{R}^d$ with $d \geqslant 2$. Therefore, using test statistics based on Sinkhorn divergences may be of practical interest. The work carried out thus focuses on the study of the statistical inference of discrete measures in terms of Sinkhorn divergence.

We obtain new results on the asymptotic distribution of these divergences for data sampled from (unknowns) distributions supported on a finite metric space. Our results are inspired by the work of [**SM16**] on the asymptotic distribution of the empirical Wasserstein distance over a finite space in terms of un-regularized transport cost. The main application consists in developing new test statistics (for one or two samples) for the comparison of multivariate probability distributions.

Finally, to illustrate the applicability of this approach, we also propose a bootstrap procedure to estimate unknown quantities of interest in the computation of these test statistics.

B.5. Geodesic principal component analysis

We finally propose to compare the methods of log-PCA and geodesic PCA (GPCA) as introduced in [**BGKL17**, **SC15**]. In our approach, histograms are viewed as piecewise constant pdf with support on a given interval $\Omega \subset \mathbb{R}$. In this context, the modes of variation of a set of histograms can be studied through the notion of GPCA of probability measures in the Wasserstein space $\mathcal{P}_2(\Omega)$ admitting these histograms as pdf. As previously stated in subsection A.7, this approach has been proposed in the statistical literature [**BGKL17**] for probability measures on the real line and in machine learning [**SC15**, **WSB+13**] for discrete probability measures on $\mathbb{R}^d$. However, the computation of the GPCA remains difficult even in the simplest case of probability density with support on $\mathbb{R}$.

We then provided a fast algorithm to perform GPCA of measures defined on the real line, and we compare its results to those of the log-PCA [**FLPJ04**, **PM16a**]. Geodesic PCA consists in solving a non-convex optimization problem. To solve it approximately, we propose a new forward-backward algorithm. We also present a detailed comparison between the log-PCA and the GPCA of one-dimensional histograms, for different datasets in dimension 1 and 2.

C. Outline of the thesis

Chapter I We introduce two regularized barycenters. The first one is regularized by a convex penalty function, the second using entropy regularized optimal transport. We study some properties of these barycenters and we propose convergence results of their estimators. This chapter is related to the papers [**BCP18b**, **BCP18a**].

Chapter II The regularized barycenters of Chapter I are then used to tackle the histogram registration problem and we propose a method to automatically calibrate the regularization parameters, as developed in [**BCP18a**].

Chapter III A central limit theorem for entropy regularized optimal transport is stated. We also derive test statistics for multivariate histograms, [**BCP17**].

Chapter IV Finally, we develop new algorithms for geodesic principal component analyses in Wasserstein space, based on the paper [**CSB+18**].

Chapter I

Regularized barycenters in the Wasserstein space

In Section I.1 we analyze the existence, uniqueness, stability and consistency of penalized Wasserstein barycenters (B.16) for various penalty functions E and any parameter $\gamma > 0$. This section corresponds to the paper [**BCP18b**]. In Section I.2, we study the variance of the Sinkhorn barycenter defined in (B.18). These developments are stated in the paper [**BCP18a**]. The large majority of proofs is deferred in Section I.3.

I.1. Penalized barycenters in the Wasserstein space

Introducing a convex penalization term in the definition (A.11) of a Wasserstein barycenter for random measures supported on Ω, a convex subset of $\mathbb{R}^d$, is a way to incorporate some prior knowledge on the behavior of its population counterpart. The existence and uniqueness of penalized Wasserstein barycenters defined in (I.1) is first proved for a large class of penalization functions E and for either a discrete distribution $\mathbb{P}_n$ supported on $\mathcal{P}_2(\Omega)$ or its population counterpart $\mathbb{P}$. The Bregman divergence associated to the penalization term is then considered to obtain a stability result on penalized barycenters. Especially this allows us to compare the case of data made of n absolutely continuous (*a.c.*) probability measures $\nu_1, \ldots, \nu_n$, with the more realistic setting where we have only access to a dataset of random variables sampled from unknown distributions as in (B.13). The convergence of the penalized empirical barycenter of a set of n iid random probability measures towards its population counterpart is finally analyzed. This approach is shown to be appropriate for the statistical analysis of either discrete or absolutely continuous random measures. It also allows to construct, from a set of discrete measures, consistent estimators of population Wasserstein barycenters that are absolutely continuous.

In this section, in the purpose of obtaining a regularized Wasserstein barycenter, we consider the following convex minimization problem

$$\min_{\mu \in \mathcal{P}_2(\Omega)} \int W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu) \tag{I.1}$$

where E is a convex penalty function, $\gamma > 0$ is a penalization parameter and $\mathbb{P}$ is any distribution on $\mathcal{P}_2(\Omega)$ (possibly discrete or not).

I.1.1. Penalized barycenters of a random measure

The following assumptions are made on the penalty function E.

Assumption I.1. *A penalty function $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ is a proper and lower semicontinuous function (for the Wasserstein distance W_2) that is strictly convex on its domain*

$$\mathcal{D}(E) = \{\mu \in \mathcal{P}_2(\Omega) \text{ such that } E(\mu) < +\infty\}. \tag{I.2}$$

We will often rely on the class of relative G-functionals (see Chapter 9, Section 9.4 of [**AGS08**]) defined below.

Definition I.2. *The relative G-functional with respect to (w.r.t) a given positive measure $\lambda \in \mathcal{M}(\Omega)$ is the function $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ defined by*

$$E(\mu) = \begin{cases} \displaystyle\int_\Omega G\left(\frac{d\mu}{d\lambda}(x)\right) d\lambda(x), & \text{if } \mu \ll \lambda \\ +\infty & \text{otherwise,} \end{cases} \tag{I.3}$$

where $G : [0, +\infty) \to [0, +\infty]$ is a proper, lower semicontinuous and strictly convex function with superlinear growth.

*Thanks to Lemma 9.4.3 in [**AGS08**], a relative G-functional is a lower semicontinuous function for the Wasserstein distance W_2, so that it satisfies Assumption I.1.*

When λ is the Lebesgue measure on $\Omega \subset \mathbb{R}^d$, choosing such a penalty function enforces the Wasserstein barycenter to be a.c. Hence, a typical example of penalty function satisfying Assumption I.1 is the negative entropy [**BFS12**] (see e.g. Lemma 1.4.3 in [**DE97**]) defined as

$$E_e(\mu) = \begin{cases} \displaystyle\int_\Omega (f(x)(\log(f(x)) - 1) + 1)dx, & \text{if } \mu \text{ admits a density } f \text{ with respect to} \\ & \text{the Lebesgue measure } dx \text{ on } \Omega, \\ +\infty & \text{otherwise.} \end{cases} \tag{I.4}$$

It is of interest to use the negative entropy as a penalty function when one has only access to discrete observations, that is in the setting where each $\boldsymbol{\nu}_i$ is a discrete measure of the form (B.14). Indeed in this case, the resulting Wasserstein barycenter minimizing (A.11) will not necessary be a.c. (unless it is penalized) whereas we are interested in recovering a density from discrete measures. In this case, a discrete barycenter will not represent in a satisfying way the underlying measures ν_i.

Penalized Wasserstein barycenters of a random measure $\boldsymbol{\nu} \in \mathcal{P}_2(\Omega)$ are then defined as follows.

Definition I.3. *Let E be a penalizing function satisfying Assumption I.1. For a distribution $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ and a penalization parameter $\gamma \geqslant 0$, the functional $J^\gamma_\mathbb{P} : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ is defined as*

$$J^\gamma_\mathbb{P}(\mu) = \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu), \ \mu \in \mathcal{P}_2(\Omega). \tag{I.5}$$

If it exists, a minimizer $\mu^\gamma_\mathbb{P}$ of $J^\gamma_\mathbb{P}$ is called a penalized Wasserstein barycenter of the random measure $\boldsymbol{\nu}$ with distribution $\mathbb{P}$.

In particular, if $\mathbb{P}$ is the discrete (resp. empirical) measure defined by $\mathbb{P} = \mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\nu_i}$ (resp. $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{\nu}_i}$) where each $\nu_i \in \mathcal{P}_2(\Omega)$ (resp. $\boldsymbol{\nu}_i \in \mathcal{P}_2(\Omega)$ random), then $J^\gamma_\mathbb{P}$ becomes

$$J^\gamma_{\mathbb{P}_n}(\mu) = \frac{1}{n}\sum_{i=1}^n W_2^2(\mu, \nu_i) + \gamma E(\mu). \tag{I.6}$$

Note that $J^\gamma_\mathbb{P}$ is strictly convex on $\mathcal{D}(E)$ by Assumption I.1.

I.1.2. Subgradient's inequality

In order to analyze the stability of the minimizers of $J^\gamma_{\mathbb{P}}$ with respect to the distribution $\mathbb{P}$, the notion of Bregman divergence related to a sufficiently smooth penalizing function E will be needed. To simplify the presentation, we shall now restrict our analysis to relative G-functionals (I.3).

Definition I.4 (Subgradient). *Let $J : \mathcal{P}_2(\Omega) \to \mathbb{R}$ be a convex, proper and lower semicontinuous function. Any subgradient $\xi \in \partial J(\mu)$ of J at $\mu \in \mathcal{D}(J)$ satisfies the inequality*

$$J(\nu) \geq J(\mu) + \langle \xi, \nu - \mu \rangle \text{ for every } \nu \in \mathcal{P}_2(\Omega), \tag{I.7}$$

and the linear form in the right-hand side of (I.7) is understood as

$$\langle \xi, \nu - \mu \rangle = \int_\Omega \xi(x)(d\nu(x) - d\mu(x)).$$

If the function is differentiable, then the subdifferential $\partial J(\mu)$ is a singleton, and thus we have $\partial J(\mu) = \nabla J(\mu)$, the gradient of J at point μ.

In what follows, we will consider subgradients for two different purposes: (i) to define a Bregman divergence with respect to E and (ii) to obtain the main result of this section that involves subgradient of the Wasserstein distance.

Definition I.5. *A penalizing function E is said to be a smooth relative G-functional if the function G is differentiable on $[0, +\infty)$. We denote by $\nabla E(\mu)$ the subgradient of E at $\mu \in \mathcal{D}(E)$ taken as*

$$\nabla E(\mu)(x) = \nabla G\left(\frac{d\mu}{d\lambda}(x)\right),\ x \in \Omega.$$

Definition I.6 (Bregman divergence). *Let E be a smooth relative G-functional. For $\mu, \nu \in \mathcal{D}(E) \subset \mathcal{P}_2(\Omega)$ the (symmetric) Bregman divergence related to E is defined by*

$$d_E(\mu, \nu) = \langle \nabla E(\mu) - \nabla E(\nu), \mu - \nu \rangle. \tag{I.8}$$

Remark I.7. *More generally, the Bregman divergence between μ and ν related to a convex functional $J : \mathcal{P}_2(\Omega) \to \mathbb{R}$ is defined for two particular subgradients $\xi \in \partial G(u)$ and $\kappa \in \partial G(v)$ by*

$$d_J^{\xi,\kappa}(\mu, \nu) = \langle \xi - \kappa, \mu - \nu \rangle.$$

To illustrate the above definitions, let us assume that λ is the Lebesgue measure dx, and consider two a.c. measures $\mu = \mu_f$ and $\nu = \nu_g$ with density f and g. An example of a smooth relative G-functional is the case where $G(u) = u^2/2$ for which $E(\mu_f) = \frac{1}{2}\|f\|^2_{\mathbb{L}^2(\Omega)} = \frac{1}{2}\int_\Omega |f(x)|^2 dx$,

$$\nabla E(\mu_f)(x) = f(x),\ \nabla E(\nu_g)(x) = g(x) \quad \text{and} \quad d_E(\mu_f, \nu_g) = \int_\Omega (f(x) - g(x))^2 dx.$$

Remark I.8. *It should be noted that the case where E is the negative entropy E_e defined in (I.4) is critical. Indeed, the negative entropy is obviously a relative G-functional with $G(u) = u(\log(u) - 1) + 1$ and $\lambda = dx$. However, as this function is not differentiable at $u = 0$, it does not lead to a smooth relative G-functional. To use such a penalizing function, it is necessary to restrict the analysis of penalized Wasserstein barycenters to the set of a.c. measures in $\mathcal{P}_2(\Omega)$ with densities that are uniformly bounded from below by a positive constant α on the set Ω. In this setting, we have that*

$$\nabla E_e(\mu_f) = \log(f(x)) \quad \textit{and} \quad \nabla E_e(\nu_g) = \log(g(x)),\ x \in \Omega,$$

and the Bregman divergence is the symmetrized Kullback-Leibler divergence

$$d_{E_e}(\mu_f, \nu_g) = \int_\Omega (f(x) - g(x)) \log\left(\frac{f(x)}{g(x)}\right) dx,$$

where $f(x) \geqslant \alpha$ and $g(x) \geqslant \alpha$ for all $x \in \Omega$.

Then, a key result to study the stability of penalized Wasserstein barycenters with respect to the distribution $\mathbb{P}$ is stated below. It involves a subgradient ϕ of the Wasserstein distance. As detailed in the proof given in the Section I.3.1, this subgradient corresponds to the Kantorovich potential introduced in Theorem A.2.

THEOREM I.9 (Subgradient's inequality). *Let E be a smooth relative G-functional and thus satisfying Assumption I.1. Let ν be a probability measure in $\mathcal{P}_2(\Omega)$, and define the functional*

$$J : \mu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu, \nu) + \gamma E(\mu)$$

where $\gamma \geqslant 0$. If $\mu \in \mathcal{P}_2(\Omega)$ minimizes J, then there exists a subgradient $\phi^{\mu,\nu} \in \mathbb{L}_1(\mu)$ of $W_2^2(\cdot, \nu)$ at μ and a potential $\psi \in \mathbb{L}_1(\nu)$ verifying $\phi^{\mu,\nu}(x) + \psi(y) \leq |x - y|^2$ for all x, y in Ω such that $(\phi^{\mu,\nu}, \psi)$ is an optimal couple of the Kantorovich's dual problem associated to μ, ν (Theorem A.2). Moreover, for all $\eta \in \mathcal{P}_2(\Omega)$,

$$\gamma \, \langle \nabla E(\mu), \mu - \eta \rangle \leq - \int \phi^{\mu,\nu} d(\mu - \eta). \tag{I.9}$$

I.1.3. Existence, uniqueness and stability of penalized barycenters

In this section, we present some properties of the minimizers of the functional $J^\gamma_{\mathbb{P}}$ presented in Definition I.3 in terms of existence, uniqueness and stability.

We first consider the minimization problem (I.5) in the particular setting where $\mathbb{P}$ is a discrete distribution on $\mathcal{P}_2(\Omega)$. That is, we study the problem

$$\min_{\mu \in \mathcal{P}_2(\Omega)} J^\gamma_{\mathbb{P}_n}(\mu) = \int W_2^2(\mu, \nu) d\mathbb{P}_n(\nu) + \gamma E(\mu) = \frac{1}{n} \sum_{i=1}^n W_2^2(\mu, \nu_i) + \gamma E(\mu) \tag{I.10}$$

where $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\nu_i} \in W_2(\mathcal{P}_2(\Omega))$ where $\nu_1, \ldots, \nu_n$ are measures in $\mathcal{P}_2(\Omega)$.

THEOREM I.10. *Suppose that Assumption I.1 holds and that $\gamma > 0$. Then, the functional $J^\gamma_{\mathbb{P}_n}$ defined by (I.10) admits a unique minimizer on $\mathcal{P}_2(\Omega)$ which belongs to the domain $\mathcal{D}(E)$ of the penalizing function E, as defined in (I.2).*

The proof of Theorem I.10 is given in Section I.3.2. Thanks to this result, one may impose the penalized Wasserstein barycenter $\mu^\gamma_{\mathbb{P}_n}$ to be a.c. on Ω by choosing a penalization function E with value $+\infty$ outside of the space of a.c. distributions. For this choice, (I.10) becomes a problem of minimization over a set of pdf.

The existence and uniqueness of (I.5) can now be shown in a general case. Since any probability measure in $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ can be approximated by a sequence of finitely supported measures $\mathbb{P}_n$ (see Theorem I.28 in Section I.3.2), we can lean on Theorem I.10 for the proof of the following result, which is also detailed in the Section I.3.2.

THEOREM I.11. *Let $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$. Suppose that Assumption I.1 holds and that $\gamma > 0$. Then, the functional $J^\gamma_{\mathbb{P}}$ defined by (I.5) admits a unique minimizer.*

When $\gamma > 0$, we now study the stability of the minimizer of $J^\gamma_{\mathbb{P}}$ with respect to discrete distributions $\mathbb{P}$ and the symmetric Bregman divergence d_E (I.8) associated to a smooth relative G-functional E. Set $\nu_1, \ldots, \nu_n \in \mathcal{P}_2(\Omega)$ and $\eta_1, \ldots, \eta_n \in \mathcal{P}_2(\Omega)$. We denote by $\mathbb{P}^\nu_n$ (resp. $\mathbb{P}^\eta_n$) the discrete measure $\frac{1}{n} \sum_{i=1}^n \delta_{\nu_i}$ (resp. $\frac{1}{n} \sum_{i=1}^n \delta_{\eta_i}$) in $W_2(\mathcal{P}_2(\Omega))$.

THEOREM I.12. *Let E be a smooth relative G-functional thus satisfying Assumption I.1. Let $\mu_\nu, \mu_\eta \in \mathcal{P}_2(\Omega)$ with μ_ν minimizing $J^\gamma_{\mathbb{P}^\nu_n}$ and μ_η minimizing $J^\gamma_{\mathbb{P}^\eta_n}$ defined by (I.10). Then, the symmetric Bregman divergence associated to E can be upper bounded as follows*

$$d_E(\mu_\nu, \mu_\eta) \leq \frac{2}{\gamma n} \inf_{\sigma \in \mathcal{S}_n} \sum_{i=1}^n W_2(\nu_i, \eta_{\sigma(i)}), \tag{I.11}$$

where $\mathcal{S}_n$ is the permutation group of the set $\{1, \ldots, n\}$.

The proof of Theorem I.12 is given in Section I.3.3. To better interpret the upper bound (I.11), we need the notion of Kantorovich transport distance $\mathcal{T}_{W_2}$ on the metric space $(\mathcal{P}_2(\Omega), W_2)$, see [**Vil03**]. For $\mathbb{P}, \mathbb{Q} \in W_2(\mathcal{P}_2(\Omega))$ endowed with the Wasserstein distance W_2, we have that

$$\mathcal{T}_{W_2}(\mathbb{P}, \mathbb{Q}) := \inf_{\Pi} \int_{\mathcal{P}_2(\Omega) \times \mathcal{P}_2(\Omega)} W_2(\mu, \nu) d\Pi(\mu, \nu),$$

where the minimum is taken over all probability measures Π on the product space $\mathcal{P}_2(\Omega) \times \mathcal{P}_2(\Omega)$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. Since $\mathbb{P}^\nu_n$ and $\mathbb{P}^\eta_n$ are discrete probability measures supported on $\mathcal{P}_2(\Omega)$, it follows that the upper bound (I.11) in Theorem I.12 can also be written as (by Birkhoff's theorem for bi-stochastic matrices, see e.g. [**Vil03**])

$$d_E(\mu_\nu, \mu_\eta) \leq \frac{2}{\gamma} \mathcal{T}_{W_2}(\mathbb{P}^\nu_n, \mathbb{P}^\eta_n).$$

Hence the above upper bound means that the Bregman divergence between the penalized Wasserstein barycenters μ_ν and μ_η is controlled by the Kantorovich transport distance between the distributions $\mathbb{P}^\nu_n$ and $\mathbb{P}^\eta_n$.

Theorem I.12 is of particular interest in the setting where the ν_i's and η_i's are discrete probability measures on $\mathbb{R}^d$. If we assume that $\nu_i = \frac{1}{p}\sum_{j=1}^p \delta_{\boldsymbol{X}_{i,j}}$ and $\eta_i = \frac{1}{p}\sum_{j=1}^p \delta_{\boldsymbol{Y}_{i,j}}$ where $(\boldsymbol{X}_{i,j})_{1\leq i\leq n; 1\leq j\leq p}$ and $(\boldsymbol{Y}_{i,j})_{1\leq i\leq n; 1\leq j\leq p}$ are (possibly random) vectors in $\mathbb{R}^d$, then by (I.11),

$$d_E(\mu_\nu, \mu_\eta) \leq \frac{2}{\gamma n} \inf_{\sigma \in \mathcal{S}_n} \sum_{i=1}^n \left(\inf_{\lambda \in \mathcal{S}_p} \left\{ \frac{1}{p} \sum_{j=1}^p |\boldsymbol{X}_{i,j} - \boldsymbol{Y}_{\sigma(i),\lambda(j)}|^2 \right\} \right)^{1/2}$$

where computing W_2 becomes an assignment task through the estimation of permutations σ and λ.

Theorem I.12 is also useful to compare the penalized Wasserstein barycenters respectively obtained from data made of n a.c. probability measures $\nu_1, \ldots, \nu_n$ and from their empirical counterpart $\boldsymbol{\nu}_{p_i} = \frac{1}{p_i}\sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}$, where $(\boldsymbol{X}_{i,j})_{j=1,\ldots,p_i}$ are iid and generated from ν_i. Denoting as $\hat{\boldsymbol{\mu}}^\gamma_{n,p}$ the random density satisfying

$$\hat{\boldsymbol{\mu}}^\gamma_{n,p} = \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\Omega)} \frac{1}{n} \sum_{i=1}^n W_2^2 \left(\mu, \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}} \right) + \gamma E(\mu),$$

it follows from inequality (I.11) that

$$\mathbb{E}\left(d_E^2 \left(\mu^\gamma_{\mathbb{P}^\nu_n}, \hat{\boldsymbol{\mu}}^\gamma_{n,p} \right) \right) \leq \frac{4}{\gamma^2 n} \sum_{i=1}^n \mathbb{E}\left(W_2^2(\nu_i, \boldsymbol{\nu}_{p_i}) \right). \tag{I.12}$$

This result allows to discuss the rate of convergence (for the squared symmetric Bregman divergence) of $\hat{\boldsymbol{\mu}}^\gamma_{n,p}$ to $\mu^\gamma_{\mathbb{P}^\nu_n}$ as a function of the rate of convergence (for the squared Wasserstein distance) of the empirical measure $\boldsymbol{\nu}_{p_i}$ to ν_i, for each $1 \leq i \leq n$, in the asymptotic setting where $p = \min_{1\leq i\leq n} p_i$ is let going to infinity.

As an illustrative example, in the one-dimensional case $d = 1$ and for absolutely continuous measures, one may use the work in [**BL14**] on a detailed study of the variety of rates

of convergence of an empirical measure on the real line toward its population counterpart for the expected squared Wasserstein distance. For example, we obtain from Theorem 5.1 in [BL14], that

$$\mathbb{E}\left(W_2^2(\nu_i, \boldsymbol{\nu}_{p_i})\right) \leqslant \frac{2}{p_i+1} K(\nu_i), \text{ with } K(\nu_i) = \int_\Omega \frac{F_i(x)(1-F_i(x))}{f_i(x)} dx,$$

where f_i is the pdf of ν_i, and F_i denotes its cumulative distribution function. Therefore, provided that $K(\nu_i)$ is finite for each $1 \leqslant i \leqslant n$, one obtains the following rate of convergence of $\hat{\boldsymbol{\mu}}_{n,p}^\gamma$ to $\mu_{\mathbb{P}_n^\nu}^\gamma$ for $d = 1$

$$\mathbb{E}\left(d_E^2\left(\mu_{\mathbb{P}_n^\nu}^\gamma, \hat{\boldsymbol{\mu}}_{n,p}^\gamma\right)\right) \leq \frac{8}{\gamma^2 n} \sum_{i=1}^n \frac{K(\nu_i)}{p_i+1} \leqslant \frac{8}{\gamma^2}\left(\frac{1}{n}\sum_{i=1}^n K(\nu_i)\right) p^{-1}. \tag{I.13}$$

Note that by the results in Appendix A in [BL14], a necessary condition for $J_2(\nu_i)$ to be finite is to assume that f_i is almost everywhere positive on the interval Ω. Rates of convergence in W_2 distance between a discrete measure and its empirical counterpart are also given in one-dimension in [BL14].

I.1.4. Convergence properties of penalized empirical barycenters

In this subsection, we study, for $\Omega \subset \mathbb{R}^d$ compact, the convergence of the penalized Wasserstein barycenter of a set $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ of independent random measures sampled from a distribution $\mathbb{P}$ towards a minimizer of $J_{\mathbb{P}}^0$, i.e. a population Wasserstein barycenter of the probability distribution $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$. Throughout this section, it is assumed that E is a smooth relative G-functional so that it satisfies Assumption I.1. We first introduce and recall some notations.

Definition I.13. *For $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ iid random measures in $\mathcal{P}_2(\Omega)$ sampled from a distribution $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$, we set $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{\nu}_i}$. Moreover, we use the notation (with $\gamma > 0$)*

$$\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \; J_{\mathbb{P}_n}^\gamma(\mu) = \int W_2^2(\mu, \nu) d\mathbb{P}_n(\nu) + \gamma E(\mu) \tag{I.14}$$

$$\mu_{\mathbb{P}}^\gamma = \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \; J_{\mathbb{P}}^\gamma(\mu) = \int W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu) \tag{I.15}$$

$$\mu_{\mathbb{P}}^0 \in \underset{\mu \in \mathcal{P}_2(\Omega)}{\operatorname{argmin}} \; J_{\mathbb{P}}^0(\mu) = \int W_2^2(\mu, \nu) d\mathbb{P}(\nu), \tag{I.16}$$

that will be respectively referred as to the penalized empirical Wasserstein barycenter (I.14), the penalized population Wasserstein barycenter (I.15) and the population Wasserstein barycenter (I.16).

Remark I.14. *Thanks to Theorem I.10, one has that the penalized Wasserstein barycenters $\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma$ and $\mu_{\mathbb{P}}^\gamma$ are well defined in the sense that they are the unique minimizers of $J_{\mathbb{P}_n}^\gamma$ and $J_{\mathbb{P}}^\gamma$ respectively. By Theorem 2 in [LGL16], there exists a population Wasserstein barycenter $\mu_{\mathbb{P}}^0$ but it is not necessarily unique. Nevertheless, as argued in [AC17], a sufficient condition for the uniqueness of $\mu_{\mathbb{P}}^0$ is to assume that the distribution $\mathbb{P}$ gives a strictly positive mass to the set of a.c. measures with respect to the Lebesgue measure. Moreover, under such an assumption for $\mathbb{P}$, it follows that $\mu_{\mathbb{P}}^0$ is an a.c. measure.*

In what follows, we discuss some convergence results of the penalized Wasserstein barycenters $\mu_{\mathbb{P}}^\gamma$ as γ tends to 0 and $\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma$ as n tends to $+\infty$. To this end, we will need tools borrowed from the empirical process theory (see [VDVW96]).

DEFINITION I.15. *Let $\mathcal{F} = \{f : U \mapsto \mathbb{R}\}$ be a class of real-valued functions defined on a given set U, endowed with a norm $\|\cdot\|$. An envelope function F of $\mathcal{F}$ is any function $u \mapsto F(u)$ such that $|f(u)| \leq F(u)$ for every $u \in U$ and $f \in \mathcal{F}$. The minimal envelope function is $u \mapsto \sup_f |f(u)|$. The covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of balls $\{\|g - f\| < \epsilon\}$ of radius ϵ and center g needed to cover the set $\mathcal{F}$. The metric entropy is the logarithm of the covering number. Finally, we define*

$$I(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\epsilon \|F\|_{\mathbb{L}_2(Q)}, \mathcal{F}, \|\cdot\|_{\mathbb{L}_2(Q)})} d\epsilon \tag{I.17}$$

where the supremum is taken over all discrete probability measures Q supported on U with $\|F\|_{\mathbb{L}_2(Q)} = \left(\int |F(u)|^2 dQ(u)\right)^{1/2} > 0$. The term $I(\delta, \mathcal{F})$ is essentially the integral of the square root of the metric entropy along the radius of the covering balls of $\mathcal{F}$.

The proof of the following theorems are given in Section I.3.4.

I.1.4.1. Convergence of $\mu_{\mathbb{P}}^\gamma$ towards $\mu_{\mathbb{P}}^0$.

We here present convergence results of the penalized population Wasserstein barycenter $\mu_{\mathbb{P}}^\gamma$ toward $\mu_{\mathbb{P}}^0$ as $\gamma \to 0$. This is classically referred to as the convergence of the bias term in nonparametric statistic.

THEOREM I.16. *Suppose that Ω is a compact of $\mathbb{R}^d$. Then, every limit of a subsequence of $(\mu_{\mathbb{P}}^\gamma)_\gamma$ in the metric space $(\mathcal{P}_2(\Omega), W_2)$ is a population Wasserstein barycenter. If we further assume that $\mu_{\mathbb{P}}^0$ is unique, then one has that*

$$\lim_{\gamma \to 0} W_2(\mu_{\mathbb{P}}^\gamma, \mu_{\mathbb{P}}^0) = 0.$$

Moreover, if $\mu_{\mathbb{P}}^0 \in \mathcal{D}(E)$ and $\nabla E(\mu_{\mathbb{P}}^0)$ is a continuous function on Ω then

$$\lim_{\gamma \to 0} D_E(\mu_{\mathbb{P}}^\gamma, \mu_{\mathbb{P}}^0) = 0,$$

where D_E is the non-symmetric Bregman divergence defined by

$$D_E(\mu_{\mathbb{P}}^\gamma, \mu_{\mathbb{P}}^0) = E(\mu_{\mathbb{P}}^\gamma) - E(\mu_{\mathbb{P}}^0) - \langle \nabla E(\mu_{\mathbb{P}}^0), \mu_{\mathbb{P}}^\gamma - \mu_{\mathbb{P}}^0 \rangle. \tag{I.18}$$

I.1.4.2. Convergence of $\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma$ towards $\mu_{\mathbb{P}}^\gamma$.

We establish a general result about the convergence to zero of $\mathbb{E}(d_E^2(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma))$ that is referred to as the variance term. Complementary results on the rate of convergence of this variance term are then given. These additional results are shown to be useful to obtain a data-driven choice for the regularization paper γ as detailed in Chapter II where we provide numerical experiments illustrating the use of penalized Wasserstein barycenters for data analysis.

THEOREM I.17. *If Ω is a compact of $\mathbb{R}^d$, then, for any $\gamma > 0$, one has that*

$$\lim_{n \to \infty} \mathbb{E}(d_E^2(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma)) = 0 \tag{I.19}$$

The proof of this theorem leans on the subgradient's inequality I.9, and is deferred in Subsection I.3.4.2.

We can actually provide a rate of convergence for this variance term which deeply depends on compactness properties of the space of measures considered in the minimization problem (I.10). To this end, we introduce the class of functions

$$\mathcal{H} = \{h_\mu : \nu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu, \nu) \in \mathbb{R} \text{ for } \mu \in \mathcal{P}_2(\Omega)\}.$$

THEOREM I.18. *If Ω is a compact of $\mathbb{R}^d$, then one has that*

$$\mathbb{E}(d_E^2(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma)) \leq \frac{CI(1,\mathcal{H})\|H\|_{\mathbb{L}_2(\mathbb{P})}}{\gamma^2 n} \tag{I.20}$$

where C is a positive constant depending on Ω, H is an envelope function of $\mathcal{H}$ and $I(1,\mathcal{H})$ is defined in (I.17).

To complete this result in a satisfying way, one needs to prove that $I(1,\mathcal{H})$ is bounded, which depends on the rate of convergence of the metric entropy towards infinity as the radius ϵ of the covering balls tends to zero.

I.1.4.3. The one-dimensional case.

The special case of probability measures $\boldsymbol{\nu}_1, \dots \boldsymbol{\nu}_n$ supported in $\mathbb{R}$ allows to obtain a proper bound on $\mathbb{E}(d_E^2(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma))$. Studying the metric entropy of the class $\mathcal{H}$ boils down to studying the metric entropy of the space $(\mathcal{P}_2(\Omega), W_2)$. By approximating each measure by discrete ones, this corresponds to the metric entropy of the space of discrete distributions on Ω, which is of order $1/\epsilon^d$ where d is the dimension of Ω assumed to be compact (see e.g. [**Ngu13**]). The term $I(1,\mathcal{H})$ appearing in (I.20) is thus finite in the one dimensional case (see Section I.3.4.2 for a rigorous proof).

THEOREM I.19. *If Ω is a compact of $\mathbb{R}$, then there exists a finite constant $c > 0$ such that*

$$\mathbb{E}(d_E^2(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma)) \leq \frac{c}{\gamma^2 n}.$$

I.1.4.4. The d-dimensional case with additional penalization.

In the case $d \geq 2$, the class of functions $\mathcal{H} = \{h_\mu : \mu \in \mathcal{P}_2(\Omega)\}$ is too large to control the metric entropy in such a way that $I(1,\mathcal{H}) < +\infty$. To tackle this issue, we impose more smoothness on the penalized Wasserstein barycenter. More precisely, we assume that Ω is a smooth and uniformly convex set, and for a smooth relative G-functional with reference measure $\lambda = dx$ (that we denote by E_G) we choose the penalizing function

$$E(\mu) = \begin{cases} E_G(\mu) + \|f\|^2_{H^k(\Omega)} = \int_\Omega G(f(x))dx + \|f\|^2_{H^k(\Omega)}, & \text{if } f = \frac{d\mu}{dx} \text{ and } f \geqslant \alpha, \\ +\infty & \text{otherwise.} \end{cases} \tag{I.21}$$

where $\|\cdot\|_{H^k(\Omega)}$ denotes the Sobolev norm associated to the $\mathbb{L}^2(\Omega)$ space and $\alpha > 0$ is arbitrarily small. Remark that we could choose a linear combination with different weights for the relative G-functional and the squared Sobolev norm. Then, the following result holds.

THEOREM I.20. *Suppose that Ω is a compact and uniformly convex set with a C^1 boundary. Assume that the penalty function E is given by (I.21) for some $\alpha > 0$ and $k > d-1$. Then, there exists a finite constant $c > 0$ such that*

$$\mathbb{E}\left(d_{E_G}^2\left(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma\right)\right) \leq \mathbb{E}\left(d_E^2\left(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma\right)\right) \leq \frac{c}{\gamma^2 n}.$$

I.2. Entropy regularized Wasserstein barycenters

In this section, we consider measures defined on fixed and finite discrete grid $\Omega_N = \{x_1, \dots, x_N\} \subset (\mathbb{R}^d)^N$ and we analyze the variance of the Sinkhorn barycenter defined in (B.18).

I.2.1. Results on the variance of the Sinkhorn barycenters

For two discrete measures $r, q \in \Sigma_N$, the Sinkhorn divergence is defined in (A.9). We shall then use two key properties to analyze the variance of Sinkhorn barycenters which are the strong convexity (see Theorem I.24 below) and the Lipschitz continuity (see Lemma I.25 below) of the mapping $r \mapsto W^2_{2,\varepsilon}(r, q)$ (for a given $q \in \Sigma_N$).

However, to guarantee the Lipschitz continuity of this mapping, it is necessary to restrict the analysis to discrete measures r belonging to the convex set

$$\Sigma_N^\rho = \left\{ r \in \Sigma_N \ : \ \min_{1 \leqslant \ell \leqslant N} r_\ell \geqslant \rho \right\},$$

where $0 < \rho < 1$ is an arbitrarily small constant. This means that our theoretical results on the variance of the Sinkhorn barycenters hold for discrete measures with non-vanishing entries. Nevertheless, we obtain upper bounds on these variances which depend explicitly on the constant ρ, allowing to discuss its choice.

Hence, as it has been done for the penalized barycenters in Definition I.13, we introduce the definitions of empirical and population Sinkhorn barycenters (constrained to belong to the set Σ_N^ρ).

Definition I.21. *Let $0 < \rho < 1/N$, and $\mathbb{P}$ be a probability distribution on Σ_N^ρ. Let $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n \in \Sigma_N^\rho$ be an iid sample drawn from the distribution $\mathbb{P}$. For each $1 \leqslant i \leqslant n$, we assume that $(\tilde{X}_{i,j})_{1 \leqslant j \leqslant p_i}$ are iid random variables sampled from $\boldsymbol{q}_i$. For each $1 \leqslant i \leqslant n$, let us define the following discrete measures*

$$\tilde{\boldsymbol{q}}_i^{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\tilde{X}_{i,j}} \quad \text{and} \quad \hat{\boldsymbol{q}}_i^{p_i} = (1 - \rho N)\tilde{\boldsymbol{q}}_i^{p_i} + \rho \mathbb{1}_N,$$

where $\mathbb{1}_N$ is the vector of $\mathbb{R}^N$ with all entries equal to one. Thanks to the condition $0 < \rho < 1/N$, it follows that $\hat{\boldsymbol{q}}_i^{p_i} \in \Sigma_N^\rho$ for all $1 \leqslant i \leqslant n$, which may not be the case for some $\tilde{\boldsymbol{q}}_i^{p_i}$, $i = 1, \dots, n$. Then, we define

$$r^\varepsilon = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \mathbb{E}_{\boldsymbol{q} \sim \mathbb{P}}[W^2_{2,\varepsilon}(r, \boldsymbol{q})] \qquad \textit{the population Sinkhorn barycenter} \tag{I.22}$$

$$\hat{\boldsymbol{r}}^\varepsilon_{n,p} = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \frac{1}{n} \sum_{i=1}^{n} W^2_{2,\varepsilon}(r, \hat{\boldsymbol{q}}_i^{p_i}) \qquad \textit{the empirical Sinkhorn barycenter} \tag{I.23}$$

In the optimization problem (I.23), we choose to use the discrete measures $\hat{\boldsymbol{q}}_i^{p_i}$ instead of the empirical measures $\tilde{\boldsymbol{q}}_i^{p_i}$ to guarantee the use of discrete measures belonging to Σ_N^ρ in the definition of the empirical Sinkhorn barycenter $\hat{\boldsymbol{r}}^\varepsilon_{n,p}$.

The following theorem is the main result of this section which gives an upper bound on the variance of $\hat{\boldsymbol{r}}^\varepsilon_{n,p}$ in terms of the expected squared Euclidean norm between elements of Σ_N.

Theorem I.22. *Let $p = \min_{1 \leqslant i \leqslant n} p_i$ and let $\varepsilon > 0$. Then, one has that*

$$\mathbb{E}(|r^\varepsilon - \hat{r}^\varepsilon_{n,p}|^2) \leqslant \frac{32 L^2_{\rho,\varepsilon}}{\varepsilon^2 n} + \frac{2 L_{\rho,\varepsilon}}{\varepsilon} \left(\sqrt{\frac{N}{p}} + 2\rho(N + \sqrt{N}) \right), \tag{I.24}$$

with

$$L_{\rho,\varepsilon} = \left(\sum_{1 \leqslant m \leqslant N} \left(2\varepsilon \log(N) + \sup_{1 \leqslant \ell,k \leqslant N} |C_{m\ell} - C_{k\ell}| - 2\varepsilon \log(\rho) \right)^2 \right)^{1/2}. \tag{I.25}$$

A few remarks can be made about the above result. The bound in the right-hand side of (I.24) explicitly depends on the size N of the grid. This will be taken into account for the choice of the optimal parameter $\hat{\varepsilon}$ (see Chapter II). Moreover, it can be used to discuss the choice of ρ. First, if one take $\rho = \epsilon^\kappa$, the Lipschitz constant (Lemma I.25) $L_{\rho,\varepsilon} = L_\varepsilon$ becomes

$$L_\varepsilon = \left(\sum_{1 \leqslant m \leqslant N} \left(2\varepsilon(\log(N) - \kappa \log(\epsilon)) + \sup_{1 \leqslant \ell,k \leqslant N} |C_{m\ell} - C_{k\ell}| \right)^2 \right)^{1/2},$$

which is a constant (not depending on ρ) such that

$$\lim_{\epsilon \to 0} L_\varepsilon = \left(\sum_{1 \leqslant m \leqslant N} \left(\sup_{1 \leqslant \ell,k \leqslant N} |C_{m\ell} - C_{k\ell}| \right)^2 \right)^{1/2}.$$

If we further assume that $\rho = \epsilon^\kappa < \min(1/N, 1/p)$ we obtain the upper bound

$$\mathbb{E}(|r^\varepsilon - \hat{r}^\varepsilon_{n,p}|^2) \leqslant \frac{32 L_\varepsilon^2}{\varepsilon^2 n} + \frac{2L_\varepsilon}{\varepsilon} \left(\sqrt{\frac{N}{p}} + 2 \left(\frac{N}{p} + \sqrt{\frac{N}{p^2}} \right) \right). \tag{I.26}$$

Finally, it should be remarked that Theorem I.22 holds for general cost matrices C that are symmetric and non-negative.

I.2.2. Proof of the variance properties of the Sinkhorn barycenters

The proof of the upper bound (I.24) relies on the use of the strong convexity of the functional $r \mapsto W^2_{2,\varepsilon}(r,q)$ for $q \in \Sigma_N$, without constraint on its entries. This property can be derived by studying the Legendre transform of $r \mapsto W^2_{2,\varepsilon}(r,q)$. For a fixed distribution $q \in \Sigma_N$, using the notation in [**CP16b**], we define the function

$$H_q(r) := W^2_{2,\varepsilon}(r,q), \quad \text{for all } r \in \Sigma_N.$$

Its Legendre transform is given for $g \in \mathbb{R}^N$ by $H_q^*(g) = \max_{r \in \Sigma_N} \langle g, r \rangle - H_q(r)$ and its differentiation properties are presented in the following theorem.

THEOREM I.23 (Theorem 2.4 in [**CP16b**]). *For $\varepsilon > 0$, the Fenchel-Legendre dual function H_q^* is C^∞. Its gradient function ∇H_q^* is $1/\varepsilon$-Lipschitz. Its value, gradient and Hessian at $g \in \mathbb{R}^N$ are, writing $\alpha = \exp(g/\varepsilon)$ and $K = \exp(-C/\varepsilon)$,*

$$H_q^*(g) = \varepsilon(E(q) + \langle q, \log(K\alpha) \rangle), \ \nabla H_q^*(g) = \operatorname{diag}(\alpha) K \frac{q}{K\alpha} \in \Sigma_N$$
$$\nabla^2 H_q^*(g) = \frac{1}{\varepsilon} \left(\operatorname{diag}\left(\operatorname{diag}(\alpha) K \frac{q}{K\alpha} \right) \right) - \frac{1}{\varepsilon} \operatorname{diag}(\alpha) K \operatorname{diag}\left(\frac{q}{(K\alpha)^2} \right) K \operatorname{diag}(\alpha),$$

where the notation $\frac{q}{r}$ stands for the component-wise division of the entries of q and r.

From this result, we can deduce the strong convexity of the dual functional H_q as stated below.

THEOREM I.24. *Let $\varepsilon > 0$. Then, for any $q \in \Sigma_N$, the function H_q is ε-strongly convex for the Euclidean 2-norm.*

The proof of Theorem I.24 is deferred to Section I.3.5. We can also ensure the Lipschitz continuity of $H_q(r)$, when restricting our analysis to the set $r \in \Sigma_N^\rho$.

LEMMA I.25. *Let $q \in \Sigma_N$ and $0 < \rho < 1$. Then, one has that $r \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ with $L_{\rho,\varepsilon}$ defined in (I.25).*

The proof of this Lemma is given in Section I.3.6.

We can now proceed to the proof of Theorem I.22. Let us introduce the following Sinkhorn barycenter

$$\boldsymbol{r}_n^\varepsilon = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \frac{1}{n}\sum_{i=1}^n W_{2,\varepsilon}^2(r, \boldsymbol{q}_i) = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(r),$$

of the iid random measures $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ (assumed to belong to Σ_N^ρ). By the triangle inequality, we have that

$$\mathbb{E}(|r^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \leqslant 2\mathbb{E}(|r^\varepsilon - \boldsymbol{r}_n^\varepsilon|^2) + 2\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2). \tag{I.27}$$

To control the first term of the right hand side of the above inequality, we use that (for any $q \in \Sigma_N$) $r \mapsto H_q(r)$ is ε-strongly convex by Theorem I.24 and $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ by Lemma I.25 where $L_{\rho,\varepsilon}$ is the constant defined by equation (I.25). Under these assumptions, it follows from Theorem 6 in [**SSSSS09**] that

$$\mathbb{E}(|r^\varepsilon - \boldsymbol{r}_n^\varepsilon|^2) \leqslant \frac{16 L_{\rho,\varepsilon}^2}{\varepsilon^2 n}. \tag{I.28}$$

For the second term in the right hand side of (I.27), we obtain by the strong convexity of H_q that

$$\frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\hat{\boldsymbol{r}}_{n,p}^\varepsilon) \geqslant \frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\boldsymbol{r}_n^\varepsilon) + \frac{1}{n}\sum_{i=1}^n \nabla H_{\boldsymbol{q}_i}(\boldsymbol{r}_n^\varepsilon)^T(\hat{\boldsymbol{r}}_{n,p}^\varepsilon - \boldsymbol{r}_n^\varepsilon) + \frac{\varepsilon}{2}|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2.$$

Theorem 3.1 in [**CP16b**] ensures that $\frac{1}{n}\sum_i \nabla H_{q_i}(\boldsymbol{r}_n^\varepsilon) = 0$. The same inequality also holds for the terms $H_{\hat{\boldsymbol{q}}_i^{p_i}}$, and we therefore have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\hat{\boldsymbol{r}}_{n,p}^\varepsilon) &\geqslant \frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\boldsymbol{r}_n^\varepsilon) + \frac{\varepsilon}{2}|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2, \\
\frac{1}{n}\sum_{i=1}^n H_{\hat{\boldsymbol{q}}_i^{p_i}}(\boldsymbol{r}_n^\varepsilon) &\geqslant \frac{1}{n}\sum_{i=1}^n H_{\hat{\boldsymbol{q}}_i^{p_i}}(\hat{\boldsymbol{r}}_{n,p}^\varepsilon) + \frac{\varepsilon}{2}|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2.
\end{aligned}$$

Using the symmetry of the Sinkhorn divergence, Lemma I.25 also implies that the mapping $q \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ for any discrete distribution r. Hence, by summing the two above inequalities, and by taking the expectation on both sides, we obtain that

$$\varepsilon\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \leqslant \frac{2L_{\rho,\varepsilon}}{n}\sum_{i=1}^n \mathbb{E}(|\boldsymbol{q}_i - \hat{\boldsymbol{q}}_i^{p_i}|).$$

Using the inequalities

$$|\boldsymbol{q}_i - \hat{\boldsymbol{q}}_i^{p_i}| \leqslant |\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| + \rho N|\tilde{\boldsymbol{q}}_i^{p_i}| + \rho|\mathbb{1}_N| \leqslant |\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| + \rho(N + \sqrt{N}),$$

we finally have that

$$\varepsilon\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \leqslant 2L_{\rho,\varepsilon}\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}(|\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}|^2)} + \rho(N + \sqrt{N})\right). \tag{I.29}$$

Conditionally on $\boldsymbol{q}_i$, one has that $p_i\tilde{\boldsymbol{q}}_i^{p_i}$ is a random vector following a multinomial distribution $\mathcal{M}(p_i, \boldsymbol{q}_i)$. Hence, for each $1 \leqslant k \leqslant N$, denoting $\boldsymbol{q}_{i,k}$ (resp. $\tilde{\boldsymbol{q}}_{i,k}^{p_i}$) the k-th coordinate

of $\boldsymbol{q}_i$ (resp. $\tilde{\boldsymbol{q}}_i^{p_i}$), one has that

$$\mathbb{E}\left(\tilde{\boldsymbol{q}}_{i,k}^{p_i}|\boldsymbol{q}_i\right) = \boldsymbol{q}_{i,k} \quad \text{and} \quad \mathbb{E}\left[\left(\tilde{\boldsymbol{q}}_{i,k}^{p_i} - \boldsymbol{q}_{i,k}\right)^2 |\boldsymbol{q}_i\right] = \frac{\boldsymbol{q}_{i,k}(1-\boldsymbol{q}_{i,k})}{p_i} \leqslant \frac{1}{4p_i}.$$

Thus, we have

$$\mathbb{E}(|\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}|^2) = \sum_{k=1}^{N} \mathbb{E}(\boldsymbol{q}_{i,k} - \tilde{\boldsymbol{q}}_{i,k}^{p_i})^2 \leqslant \frac{1}{4}\sum_{k=1}^{N} p_i^{-1} \leqslant \frac{N}{4p} \tag{I.30}$$

and we obtain from (I.29) and (I.30) that

$$\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \leqslant \frac{L_{\rho,\varepsilon}}{\varepsilon}\left(\sqrt{\frac{N}{p}} + 2\rho(N+\sqrt{N})\right). \tag{I.31}$$

Combining inequalities (I.27), (I.28), and (I.31) concludes the proof of Theorem I.22.

I.3. Proofs of Chapter I

I.3.1. Proof of the subgradient's inequality, Theorem I.9

The proof of Theorem I.9 is based on the two succeeding lemmas.

LEMMA I.26. *The two following assertions are equivalent:*

(1) $\mu \in \mathcal{P}_2(\Omega)$ minimizes J over $\mathcal{P}_2(\Omega)$,
(2) there exists a subgradient $\phi \in \partial J(\mu)$ such that $\langle \phi, \eta - \mu\rangle \geq 0$ for all $\eta \in \mathcal{P}_2(\Omega)$.

PROOF OF LEMMA I.26. *2⇒1.* Let $\phi \in \partial J(\mu)$ such that $\langle \phi, \eta-\mu\rangle \geq 0$ for all $\eta \in \mathcal{P}_2(\Omega)$. By definition of the subgradient, $\forall\ \eta \in \mathcal{P}_2(\Omega)$, we have $J(\eta) \geq J(\mu) + \langle \phi, \eta - \mu\rangle$ which is greater than $J(\mu)$ by assertion. Hence μ minimizes J.
1⇒2. Take $\mu \in \text{int}(\text{dom } J)$ (that is $J(\mu) < +\infty$) such that μ is a minimum of J over $\mathcal{P}_2(\Omega)$. Then the directional derivative of J at the point μ along $(\eta - \mu)$ exists (Proposition 2.22 in [Cla13]) and satisfies

$$J'(\mu; \eta - \mu) := \lim_{\substack{t\to 0\\ t>0}} \frac{J(\mu + t(\eta-\mu)) - J(\mu)}{t} \geq 0. \tag{I.32}$$

Remark that $\mathcal{P}_2(\Omega)$ is a convex set. By Proposition 4.3 of [Cla13], since J is a proper convex function and $\mu \in \text{dom}(J)$, we obtain the equivalence

$$\phi \in \partial J(\mu) \ \Leftrightarrow\ \langle \phi, \Delta\rangle \leq J'(\mu; \Delta) \text{ for all } \Delta \in \mathcal{P}_2(\Omega).$$

Moreover, since J is proper convex and lower semi-continuous, so is $J'(f;\cdot)$. Given that $\mathcal{P}_2(\Omega)$ is a Hausdorff convex space, we get by Theorem 7.6 of [AB06], that for all $(\eta-\mu) \in \mathcal{P}_2(\Omega)$, $J'(\mu; \eta-\mu) = \sup\{\langle \phi, \eta-\mu\rangle$ where ϕ is such that $\langle \phi, \Delta\rangle \leq J'(\mu;\Delta), \forall \Delta$ in $\mathcal{P}_2(\Omega)\}$. Hence by (I.32) we get $\sup_{\phi\in\partial J(\mu)} \langle \phi, \eta-\mu\rangle \geq 0$. We then define the ball $B_\epsilon = \{\eta + \mu \in \mathcal{M}(\Omega)$ such that $\|\eta\|_{TV} \leq \epsilon\}$, where $\|\cdot\|_{TV}$ is the norm of total variation. We still have

$$\inf_{\eta\in B_\epsilon \cap \mathcal{P}_2(\Omega)} \sup_{\phi\in\partial J(\mu)} \langle \phi, \eta-\mu\rangle \geq 0.$$

Note that $\partial J(\mu)$ in a convex set. Moreover $B_\epsilon \cap \mathcal{P}_2(\Omega)$ is compact, and $(\phi, \eta) \mapsto \langle \phi, \eta - \mu\rangle$ is bilinear. Thus we can switch the infimum and the supremum by the Ky Fan's theorem (4.36 in [Cla13]). In that way, there exists $\phi \in \partial J(f)$ such that $\inf_{\eta\in B_\epsilon\cap\mathcal{P}_2(\Omega)} \langle \phi, \eta-\mu\rangle \geq 0$.

By convexity of $\mathcal{P}_2(\Omega)$, any $\zeta \in \mathcal{P}_2(\Omega)$ can be written as $t(\eta-\mu)+\mu$ for some $t \geq 0$ and $\eta \in B_\epsilon \cap \mathcal{P}_2(\Omega)$. This concludes the proof of the lemma. □

LEMMA I.27. *Let $\mu \in \mathcal{P}_2(\Omega)$ and $\phi \in \mathbb{L}_1(\mu)$, then*

$$\phi \in \partial_1 W_2^2(\mu,\nu) \Leftrightarrow \exists\ \psi \in \mathbb{L}_1(\nu) \text{ such that } \phi(x)+\psi(y) \leq |x-y|^2$$

and $W_2^2(\mu,\nu) = \int \phi d\mu + \int \psi d\nu$ where $\partial_1 W_2^2(\mu,\nu)$ denote the subdifferential of the function $W_2^2(\cdot,\nu)$ at μ.

PROOF OF LEMMA I.27. ($\Leftarrow$). We first assume that for $\phi^{\mu,\nu} \in \mathbb{L}_1(\mu)$, there exists $\psi^{\mu,\nu} \in \mathbb{L}_1(\nu)$ such that $W_2^2(\mu,\nu) = \int \phi^{\mu,\nu} d\mu + \int \psi^{\mu,\nu} d\nu$ and $\phi^{\mu,\nu}(x)+\psi^{\mu,\nu}(y) \leq |x-y|^2$. Then for all $\eta \in \mathcal{P}_2(\Omega)$, denoting $(\phi^{\eta,\nu}, \psi^{\eta,\nu})$ an optimal couple for η and ν, we get

$$\begin{aligned} W_2^2(\eta,\nu) &= \sup_{\phi(x)+\psi(y)\leq|x-y|^2} \int \phi d\eta + \int \psi d\nu = \int \phi^{\eta,\nu} d\eta + \int \psi^{\eta,\nu} d\nu \\ &\geqslant W_2^2(\mu,\nu) + \int \phi^{\mu,\nu} d(\eta-\mu). \end{aligned}$$

Hence, from the definition of a subgradient, we have $\phi^{\mu,\nu} \in \partial_1 W_2^2(\mu,\nu)$.
($\Rightarrow$). We denote by F the function $\mu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu)$. Let $\phi^* \in \partial F(\mu)$, then by the Legendre-Fenchel theory, we have that $F^*(\phi^*)+F(\mu) = \int \phi^* d\mu$, where F^* denote the Fenchel conjugate of F. We want to show that there exists $\psi \in \mathbb{L}_1(\nu)$ verifying $\phi^*(x)+\psi(y) \leq |x-y|^2$ such that

$$\int \phi^* d\mu - W_2^2(\mu,\nu) = -\int \psi d\nu,$$

which is equivalent to $F^*(\phi^*) = -\int \psi d\nu$. In this aim, we define $\psi^\phi(\cdot) := \inf_{y\in\Omega}\{|\cdot - y|^2 - \phi(y)\}$ and $H(\phi) := -\int \psi^\phi d\nu$, and we recall that $H^*(\mu) = \sup_{\phi\in Y}\{\int \phi d\mu - H(\phi)\}$. H is convex, l.s.c. on Y and proper since

$$\begin{aligned} H(\phi) = -\int \psi^\phi d\nu &= \int \sup_{y\in\Omega}\{\phi(y) - |x-y|^2\}\, d\nu(x) \\ &\geq \int (\phi(y_0) - 2|y_0| - 2|x|^2) d\nu(x) > -\infty \text{ by definition of } \nu, \end{aligned}$$

where $y_0 \in \Omega$ is such that $\phi(y_0)$ is finite. We get $H^{**}(\phi) = H(\phi)$ by Theorem 2.3.3. in [**Zal02**]. Moreover, for $\mu \in \mathcal{P}_2(\Omega)$, we have by the duality formulation of Kantorovich (e.g Lemma 2.1. of [**AC11**]) that

$$\begin{aligned} W_2^2(\mu,\nu) &= \sup\left\{\int_\Omega \phi d\mu + \int_\Omega \psi d\nu;\ \phi,\psi \in \mathcal{C}_b,\ \phi(x)+\psi(y) \leq |x-y|^2\right\} \\ &= \sup\left\{\int_\Omega \phi d\mu + \int_\Omega \psi d\nu;\ \phi,\psi \in \mathcal{C}_b,\ \psi(y) \leq \inf_x\{|x-y|^2 - \phi(x)\}\right\} \\ &= \sup_\phi\left\{\int_\Omega \phi d\mu + \int_\Omega \psi^\phi\right\} = H^*(\mu). \end{aligned} \tag{I.33}$$

We deduce that $H^{**}(\phi) = \sup_{f\in\mathcal{P}_2(\Omega)}\{\int \phi d\mu - W_2^2(\mu,\nu)\} = F^*(\phi)$, which implies $F^*(\phi^*) = H(\phi^*)$. Thus we end up with the equality $F(\mu) = \int \phi^* d\mu - F^*(\phi^*) = \int \phi^* d\mu + \int \psi^{\phi^*} d\nu$. This exactly means that for $\phi^* \in \partial_1 W_2^2(\mu,\nu)$, there exists ψ^{ϕ^*} such that $\phi^*(x)+\psi^{\phi^*}(y) \leq |x-y|^2$ and $W_2^2(\mu,\nu) = \int \phi^* d\mu + \int \psi^{\phi^*} d\nu$, which concludes the proof. □

From these lemmas, we directly get the proof of Theorem I.9.

PROOF OF THEOREM I.9. Let $\mu \in \mathcal{P}_2(\Omega)$ be a minimizer of J. From Lemma I.26, we know that there exists ϕ a subgradient of J in μ such that $\langle \phi, \eta-\mu\rangle \geq 0$ for all $\eta \in \mathcal{P}_2(\Omega)$. Since $\zeta \mapsto E(\zeta)$ is convex differentiable, $\zeta \mapsto W_2^2(\zeta,\nu)$ is a continuous convex function and μ minimizes J, we have by the subdifferential of the sum (Theorem 4.10 in [**Cla13**]) that

$\partial J(\mu) = \partial_1 W_2^2(\mu,\nu) + \gamma\, \nabla E(\mu)$. This implies that all $\phi \in \partial J(\mu)$ is written $\phi = \phi_1 + \phi_2$ with $\phi_1 = \phi^{\mu,\nu}$ optimal for the couple (μ,ν) (by Lemma I.27) and $\phi_2 = \gamma \nabla E(\mu)$. Finally, we have that $\langle \phi^{\mu,\nu} + \gamma \nabla E(\mu), \eta - \mu\rangle \geq 0$ for all $\eta \in \mathcal{P}_2(\Omega)$ that is $\gamma\, \langle \nabla E(\mu), \mu - \eta\rangle \leq -\int \phi^{\mu,\nu} d(\mu-\eta)$, $\forall \eta \in \mathcal{P}_2(\Omega)$. □

I.3.2. Proof of existence and uniqueness of penalized barycenters in I.3

For the sake of completeness, we introduce the functional space $Y := \{g \in \mathcal{C}(\Omega) : x \mapsto g(x)/(1+|x|^2) \text{ is bounded}\}$ endowed with the norm $\|g\|_Y = \sup_{x\in\Omega} |g(x)|/(1+|x|^2)$ where $\mathcal{C}(\Omega)$ is the space of continuous functions from Ω to $\mathbb{R}$. We finally denote as Z the closed subspace of Y given by $Z = \left\{g \in \mathcal{C}(\Omega) : \lim_{|x|\to\infty} g(x)/(1+|x|^2) = 0\right\}$. The space $\mathcal{M}(\Omega)$ of bounded Radon measures is identified with the dual of $\mathcal{C}_0(\Omega)$ (space of continuous functions that vanish at infinity). Finally, we denote by $\mathbb{L}_1(\mu)$ the set of integrable functions $g : \Omega \to \mathbb{R}$ with respect to the measure μ.

Proof of Theorem I.10. Let $(\mu^k)_k \subset \mathcal{P}_2(\Omega)$ a minimizing sequence of probability measures of $J^\gamma_{\mathbb{P}_n}$. Hence, there exists a constant $M \geq 0$ such that $\forall k$, $J^\gamma_{\mathbb{P}_n}(\mu^k) \leq M$. It follows that for all k, $\frac{1}{n}\sum_{i=1}^n W_2^2(\mu^k,\nu_i) \leq M$. By Lemma 2.1 of [AC11] we thus have

$$\frac{1}{n}\sum_{i=1}^n W_2^2(\nu^i,\mu^k) = 2\sum_{i=1}^n \sup_{f\in Z}\left\{\int_\Omega f d\mu^k + \int_\Omega Sf(x) d\nu^i(x)\right\} \leq M,$$

where $Sf(x) = \inf_{y\in\Omega}\{\frac{1}{2n}|x-y|^2 - f(y)\}$. Since the function $x \mapsto |x|^\alpha$ (with $1<\alpha<2$) belongs to Z, we have that $\int_{\mathbb{R}^d} |x|^\alpha d\mu^k(x)$ is bounded by a constant $L \geq 0$ for all k. We deduce that $(\mu^k)_k$ is tight (for instance, take the compact $K^c = \{x \in \Omega \text{ such that } |x|^\alpha > \frac{L}{\epsilon}\}$). Since $(\mu^k)_k$ is tight, by Prokhorov's theorem, there exists a subsequence of $(\mu^k)_k$ (still denoted $(\mu^k)_k$) which weakly converges to a probability measure μ. Moreover, one can prove that $\mu \in \mathcal{P}_2(\Omega)$. Indeed for all lower semicontinuous functions bounded from below by f, we have that $\liminf_{k\to\infty} \int_\Omega f(x) d\mu^k(x) \geq \int_\Omega f(x) d\mu(x)$ by weak convergence. Hence for $f : x \mapsto |x|^2$, we get $\int_\Omega |x|^2 d\mu(x) \leq \liminf_{k\to\infty} \int_\Omega |x|^2 d\mu^k(x) < +\infty$, and thus $\mu \in \mathcal{P}_2(\Omega)$.

Let $(\pi_i^k)_{1\leq i\leq n, 1\leq k}$ be a sequence of optimal transport plans where π_i^k is an optimal transport plan between μ^k and ν_i. Since $\sup_k W_2^2(\mu^k,\nu_i) = \sup_k \iint_{\Omega\times\Omega} |x-y|^2 d\pi_i^k(x,y) < +\infty$, we may apply Proposition 7.1.3 of [AGS08]: $(\pi_i^k)_k$ is weakly relatively compact on the probability space over $\Omega\times\Omega$ and every weak limit π_i is an optimal transport plan between μ and ν_i with, for all $1 \leq i \leq n$, $W_2^2(\mu,\nu_i) \leq \liminf_{k\to\infty} \int_{\Omega\times\Omega} |x-y|^2 d\pi_i^k(x,y) < +\infty$. Since E is lower semicontinuous, we get that

$$\liminf_{k\to\infty} J^\gamma_{\mathbb{P}_n}(\mu^k) = \liminf_{k\to\infty} \frac{1}{n}\sum_{i=1}^n W_2^2(\mu^k,\nu_i) + \gamma E(\mu^k) \geq \frac{1}{n}\sum_{i=1}^n W_2^2(\mu,\nu_i) + \gamma E(\mu) = J^\gamma_{\mathbb{P}_n}(\mu).$$

Hence $J^\gamma_{\mathbb{P}_n}$ admits at least $\mu \in \mathcal{P}_2(\Omega)$ as a minimizer. Finally, by the strict convexity of $J^\gamma_{\mathbb{P}_n}$ on its domain, the minimizer is unique and it belongs to $\mathcal{D}(E)$ as defined in (I.2), which completes the proof. □

Proof of Theorem I.11. First, let us prove the existence of a minimizer. For that purpose, we decide to follow the sketch of the proof of the existence of a Wasserstein barycenter given by Theorem 1 in [LGL16]. We suppose that $(\mathbb{P}_n)_{n\geq 0} \subseteq W_2(\mathcal{P}_2(\Omega))$ is a sequence of measures, such that $\mu^n \in \mathcal{P}_2(\Omega)$ is a probability measure minimizing $J^\gamma_{\mathbb{P}_n}$, for all n. Furthermore, we suppose that there exists $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ such that $\mathcal{W}_2(\mathbb{P},\mathbb{P}_n) \underset{n\to+\infty}{\longrightarrow} 0$. We then have to prove that $(\mu^n)_{n\geq 1}$ is precompact and that all limits minimize $J^\gamma_{\mathbb{P}}$. We denote $\tilde{\boldsymbol{\mu}}$ a random measure with distribution $\mathbb{P}$ and $\tilde{\boldsymbol{\mu}}^n$ a random measure with distribution $\mathbb{P}_n$.

Hence we get

$$\begin{aligned} W_2(\mu^n, \delta_x) &= \mathcal{W}_2(\delta_{\mu^n}, \delta_{\delta_x}) \leq \mathcal{W}_2(\delta_{\mu^n}, \mathbb{P}_n) + \mathcal{W}_2(\mathbb{P}_n, \delta_{\delta_x}) \\ &= \mathbb{E}(W_2^2(\mu^n, \tilde{\boldsymbol{\mu}}^n))^{1/2} + \mathbb{E}(W_2^2(\tilde{\boldsymbol{\mu}}^n, \delta_x))^{1/2}. \end{aligned}$$

Moreover, $\mathbb{E}(W_2^2(\mu^n, \tilde{\boldsymbol{\mu}}^n))^{1/2} \leq M$ for a constant $M \geq 0$ since μ_n minimizes $J^\gamma_{\mathbb{P}_n}$ and $\tilde{\boldsymbol{\mu}}^n$ is of law $\mathbb{P}_n$. Then for $x \in \Omega$

$$W_2(\mu^n, \delta_x) \leq M + \mathcal{W}_2(\mathbb{P}_n, \delta_{\delta_x}) \leq M + \mathcal{W}_2(\mathbb{P}_n, \mathbb{P}) + \mathcal{W}_2(\mathbb{P}, \delta_{\delta_x}) \leq L,$$

since $\mathcal{W}_2(\mathbb{P}_n, \mathbb{P}) \underset{n\to+\infty}{\longrightarrow} 0$ and $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ by hypothesis. By Markov inequality, we have for $r > 0$

$$\mu^n(B(x,r)^c) = \mathbb{P}_{\mu^n}(|X - x|^2 \geq r^2) \leq \frac{\mathbb{E}_{\mu^n}(|X-x|^2)}{r^2} = \frac{W_2^2(\mu^n, \delta_x)}{r^2},$$

and $\mu^n(B(x,r)^c) \leq \frac{L^2}{r^2}$. Hence $(\mu^n)_n$ is tight: it is possible to extract a subsequence (still denoted (μ^n)) which converges weakly to a measure μ by Prokhorov's theorem. Let us show that μ minimizes $J^\gamma_{\mathbb{P}}$. Let $\eta \in \mathcal{P}_2(\Omega)$ and $\nu \in \mathcal{P}_2(\Omega)$ with distribution $\mathbb{P}$.

$$\begin{aligned} J^\gamma_{\mathbb{P}}(\eta) &= \mathbb{E}_{\mathbb{P}}(W_2^2(\eta, \nu)) + \gamma E(\eta) \\ &= \mathcal{W}_2^2(\delta_\eta, \mathbb{P}) + \gamma E(\eta) \\ &= \lim_{n\to+\infty} \mathcal{W}_2^2(\delta_\eta, \mathbb{P}_n) + \gamma E(\eta) && \text{since by hypothesis } \mathcal{W}_2(\mathbb{P}_n, \mathbb{P}) \to 0, \\ &\geq \liminf_{n\to+\infty} \mathcal{W}_2^2(\delta_{\mu^n}, \mathbb{P}_n) + \gamma E(\mu^n) && \text{since } \mu^n \text{ minimizes } J^\gamma_{\mathbb{P}_n}. \end{aligned} \tag{I.34}$$

Moreover, we have by the inverse triangle inequality that

$$\liminf_{n\to+\infty} \mathcal{W}_2(\delta_{\mu^n}, \mathbb{P}_n) \geqslant \liminf_{n\to+\infty} (\mathcal{W}_2(\delta_{\mu^n}, \mathbb{P}) - \mathcal{W}_2(\mathbb{P}, \mathbb{P}_n)).$$

First, $\mathcal{W}_2(\mathbb{P}, \mathbb{P}_n) \to 0$ by assumption. Second, we have that

$$\begin{aligned} \liminf_{n\to+\infty} \mathcal{W}_2(\delta_{\mu^n}, \mathbb{P}) &\geqslant \int \liminf_{n\to+\infty} W_2^2(\mu_n, \nu) d\mathbb{P}(\nu) && \text{by Fatou's Lemma} \\ &\geqslant \int W_2^2(\mu, \nu) d\mathbb{P}(\nu) = \mathcal{W}_2^2(\delta_\mu, \mathbb{P}) && \text{by the equality (I.33)} \end{aligned}$$

Thus from (I.34) and by lower semicontinuity of E, we conclude that $J^\gamma_{\mathbb{P}}(\eta) \geq \mathcal{W}_2^2(\delta_\mu, \mathbb{P}) + \gamma E(\mu) = J^\gamma_{\mathbb{P}}(\mu)$. Hence μ minimizes $J^\gamma_{\mathbb{P}}$. To finish the proof of the existence of a minimizer, we need the following result which proof can be found in [**LGL16**].

THEOREM I.28. *For all $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$, there is a sequence of finitely supported distributions $\mathbb{P}_n$ (that is $\mathbb{P}_n = \sum_{k=1}^K \lambda_k \delta_{\kappa_k}$ where $\sum_{k=1}^K \lambda_k = 1$) such that $\mathcal{W}_2^2(\mathbb{P}_n, \mathbb{P}) \underset{n\to+\infty}{\longrightarrow} 0$.*

Now, by Theorem I.28 it follows that for a given distribution $\mathbb{P}$, one can find a sequence of finitely supported distributions $\mathbb{P}_n$ such that for all n there exists a unique measure $\mu^n \in \mathcal{P}_2(\Omega)$ minimizing $J^\gamma_{\mathbb{P}_n}$ using Theorem I.10 and such that $W_2^2(\mathbb{P}_n, \mathbb{P}) \underset{n\to+\infty}{\longrightarrow} 0$ thanks to Theorem I.28. Therefore there is a probability measure μ which minimizes $J^\gamma_{\mathbb{P}}$. Let us make sure that μ is indeed in the space $\mathcal{P}_2(\Omega)$. From Theorem I.10, we also have that $\mu^n \in \mathcal{P}_2(\Omega)$ for all n. Thus by weak convergence, $\int_\Omega |x|^2 d\mu(x) \leq \liminf_{n\to+\infty} \int_\Omega |x|^2 d\mu^n(x) < +\infty$. Finally, the uniqueness of the minimum is obtained by the strict convexity of the functional $\mu \mapsto \mathbb{E}_{\mathbb{P}}(W_2^2(\mu, \nu)) + \gamma E(\mu)$ on the domain $\mathcal{D}(E)$, which completes the proof. □

I.3.3. Proof of the stability's Theorem I.12

Proof of Theorem I.12. We denote by $\mu, \zeta \in \mathcal{P}_2(\Omega)$ the probability measures such that μ minimizes $J^{\gamma}_{\mathbb{P}^{\nu}_n}$ and ζ minimizes $J^{\gamma}_{\mathbb{P}^{\eta}_n}$. For each $1 \leqslant i \leqslant n$, one has that $\theta \mapsto \frac{1}{n} W_2^2(\theta, \nu_i)$ is a convex, proper and continuous function. Therefore, from Theorem 4.10 in [Cla13], we have that $\partial J_{\mathbb{P}^{\nu}_n}(\mu) = \frac{1}{n}\sum_{i=1}^n \partial_1 W_2^2(\mu, \nu_i) + \gamma \nabla E(\mu)$. Hence by Lemma I.27, any $\phi \in \partial J_{\mathbb{P}^{\nu}_n}(\mu)$ is of the form $\phi = \frac{1}{n}\sum_{i=1}^n \phi_i + \gamma\ \nabla E(\mu)$ where for all $i = 1, \dots, n$, $\phi_i = \phi^{\mu,\nu_i}$ is optimal in the sense that $(\phi^{\mu,\nu_i}, \psi^{\mu,\nu_i})$ is an optimal couple associated to (μ, ν_i) in the Kantorovich formulation of the Wasserstein distance (see Theorem A.2). Therefore by Lemma I.26, there exists $\phi = \frac{1}{n}\sum_{i=1}^n \phi^{\mu,\nu_i} + \gamma \nabla E(\mu)$ such that $\langle \phi, \theta - \mu \rangle \geq 0$ for all $\theta \in \mathcal{P}_2(\Omega)$. Likewise, there exists $\breve{\phi} = \frac{1}{n}\sum_{i=1}^n \phi^{\zeta,\eta_i} + \gamma \nabla E(\zeta)$ such that $\langle \breve{\phi}, \theta - \zeta \rangle \geq 0$ for all $\theta \in \mathcal{P}_2(\Omega)$. Finally, we obtain

$$\gamma \langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta \rangle \leq -\int_\Omega \left(\frac{1}{n}\sum_{i=1}^n (\phi^{\mu,\nu_i} - \phi^{\zeta,\eta_i}) \right) d(\mu - \zeta).$$

Following the proof of Kantorovich duality's theorem in [Vil03], we can restrict the supremum over $(\phi, \psi) \in C_W$ in Kantorovich's duality Theorem A.2 to the admissible pairs (ϕ^{cc}, ϕ^c) where $\phi^c(y) = \inf_x \{|x-y|^2 - \phi(x)\}$ and $\phi^{cc}(x) = \inf_y \{|x-y|^2 - \phi^c(y)\}$. Then, we replace ϕ^{μ,ν_i} by $(\phi^{\mu,\nu_i})^{cc}$ (resp. ϕ^{ζ,η_i} by $(\phi^{\zeta,\eta_i})^{cc}$) and ψ^{μ,ν_i} by $(\phi^{\mu,\nu_i})^c$ (resp. ψ^{ζ,η_i} by $(\phi^{\zeta,\eta_i})^c$) and obtain

$$\begin{aligned}
\gamma \langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta \rangle &\leq -\frac{1}{n}\sum_{i=1}^n \int_\Omega \left[(\phi^{\mu,\nu_i})^{cc}(x) - (\phi^{\zeta,\eta_i})^{cc}(x) \right] d(\mu - \zeta)(x) \\
&= -\frac{1}{n}\sum_{i=1}^n \iint_{\Omega\times\Omega} \left[(\phi^{\mu,\nu_i})^{cc}(x) - (\phi^{\zeta,\eta_i})^{cc}(x) \right] d(\pi^{\mu,\nu_i} - \pi^{\zeta,\eta_i})(x,y),
\end{aligned}$$

where π^{μ,ν_i} is an optimal transport plan on $\Omega \times \Omega$ with marginals μ and ν_i for $i \in \{1, \dots, n\}$ (and π^{ζ,η_i} optimal with marginals ζ and η_i). Developing the right-hand side expression in the above inequality, we get

$$\begin{aligned}
\gamma \langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta \rangle \leq &-\frac{1}{n}\sum_{i=1}^n \left[\iint (\phi^{\mu,\nu_i})^{cc}(x) d\pi^{\mu,\nu_i}(x,y) + \iint (\phi^{\zeta,\eta_i})^{cc}(x) d\pi^{\zeta,\eta_i}(x,y) \right] \\
&+ \frac{1}{n}\sum_{i=1}^n \left[\iint (\phi^{\mu,\nu_i})^{cc}(x) d\pi^{\zeta,\eta_i}(x,y) + \iint (\phi^{\zeta,\eta_i})^{cc}(x) d\pi^{\mu,\nu_i}(x,y) \right].
\end{aligned}$$

From the condition (A.5) in the Kantorovich's dual problem, we have that $(\phi^{\mu,\nu_i})^{cc}(x) \leq |x-y|^2 - (\phi^{\mu,\nu_i})^c(y)$ and $(\phi^{\zeta,\eta_i})^{cc}(x) \leq |x-y|^2 - (\phi^{\zeta,\eta_i})^c(y)$ for all $i \in \{1, \dots, n\}$. Moreover, we have that $(\phi^{\mu,\nu_i})^{cc}(x) d\pi^{\mu,\nu_i}(x,y) = \left[|x-y|^2 - (\phi^{\mu,\nu_i})^c(y) \right] d\pi^{\mu,\nu_i}(x,y)$ and likewise $(\phi^{\zeta,\eta_i})^{cc}(x) d\pi^{\zeta,\eta_i}(x,y) = \left[|x-y|^2 - (\phi^{\zeta,\eta_i})^c(y) \right] d\pi^{\zeta,\eta_i}(x,y)$. We therefore deduce that

$$\begin{aligned}
&\gamma \langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta \rangle \\
\leq &-\frac{1}{n}\sum_{i=1}^n \left[\iint \left[|x-y|^2 - (\phi^{\mu,\nu_i})^c(y) \right] d\pi^{\mu,\nu_i}(x,y) + \iint \left[|x-y|^2 - (\phi^{\zeta,\eta_i})^c(y) \right] d\pi^{\zeta,\eta_i}(x,y) \right] \\
&+ \frac{1}{n}\sum_{i=1}^n \left[\iint \left[|x-y|^2 - (\phi^{\mu,\nu_i})^c(y) \right] d\pi^{\zeta,\eta_i}(x,y) + \iint \left[|x-y|^2 - (\phi^{\zeta,\eta_i})^c(y) \right] d\pi^{\mu,\nu_i}(x,y) \right] \\
= &\frac{1}{n}\sum_{i=1}^n \int_\Omega \left[(\phi^{\mu,\nu_i})^c(y) - (\phi^{\zeta,\eta_i})^c(y) \right] d(\nu_i - \eta_i)(y).
\end{aligned}$$

For all $1 \leq i \leq n$, we have that $(\phi^{\mu,\nu_i})^c$ and $(\phi^{\zeta,\eta_i})^c$ are 1-Lipschitz by definition, which implies that $\frac{1}{2}\left[(\phi^{\mu,\nu_i})^c - (\phi^{\zeta,\eta_i})^c\right]$ is 1-Lipschitz for all $1 \leq i \leq n$. We then conclude

$$\begin{aligned}\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle \leq & \frac{2}{n}\sum_{i=1}^{n} \sup\left\{\int \phi \, d(\nu_i - \eta_i);\ \phi \in \cap \mathbb{L}^1(|\nu_i - \eta_i|), \|\phi\|_{Lip} \leq 1\right\} \\ = & \frac{1}{n}\sum_{i=1}^{n} W_1(\nu_i, \eta_i) \leq \frac{2}{n}\sum_{i=1}^{n} W_2(\nu_i, \eta_i),\end{aligned}$$

by the Kantorovich-Rubinstein theorem presented in [**Vil03**], while the last inequality above comes from Hölder inequality between the distance W_2 and the distance W_1 defined for θ_1, θ_2 (probability measures on Ω with moment of order 1) as

$$W_1(\theta_1, \theta_2) = \inf_{\pi} \int_{\Omega}\int_{\Omega} |x - y| d\pi(x, y),$$

where π is a probability measure on $\Omega \times \Omega$ with marginals θ_1 and θ_2. Since μ and ζ are independent, we can assign to ν_i any $\eta_{\sigma(i)}$ for $\sigma \in \mathcal{S}_n$ the permutation group of $\{1, \ldots, n\}$ to obtain $\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle \leq \frac{2}{n} \inf_{\sigma \in \mathcal{S}_n} \sum_{i=1}^{n} W_2(\nu_i, \eta_{\sigma(i)})$, which completes the proof. □

I.3.4. Proofs of penalized barycenters's convergence properties

I.3.4.1. Convergence of $\mu_{\mathbb{P}}^{\gamma}$ towards $\mu_{\mathbb{P}}^{0}$

Proof of Theorem I.16. By Theorem 2.1.(d) in [**Bra06**], $J_{\mathbb{P}}^{\gamma}$ Γ-converges to $J_{\mathbb{P}}^{0}$ in 2-Wasserstein metric. Indeed for every sequence $(\mu_\gamma)_\gamma \subset \mathcal{P}_2(\Omega)$ converging to $\mu \in \mathcal{P}_2(\Omega)$,

$$J_{\mathbb{P}}^{0}(\mu) \leq \liminf_{\gamma \to 0} J_{\mathbb{P}}^{\gamma}(\mu_\gamma),$$

by lower semicontinuity of $J_{\mathbb{P}}^{\gamma}$ with respect to the W_2 metric. Moreover, there exists a sequence $(\mu_\gamma)_\gamma$ converging to μ (for instance take $(\mu_\gamma)_\gamma$ constant and equal to μ) such that $\lim_{\gamma \to 0} J_{\mathbb{P}}^{\gamma}(\mu_\gamma) = \lim_{\gamma \to 0} J_{\mathbb{P}}^{\gamma}(\mu) = J_{\mathbb{P}}^{0}(\mu)$. One can also notice that $J_{\mathbb{P}}^{\gamma} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ is equi-coercive: for all $t \in \mathbb{R}$, the set $\{\nu \in \mathcal{P}_2(\Omega) \text{ such that } J_{\mathbb{P}}^{\gamma}(\nu) \leq t\}$ is included in a compact K_t since it is closed in the compact set $\mathcal{P}_2(\Omega)$ (by compactness of Ω). Therefore, we can apply the fundamental theorem of Γ-convergence (Theorem 2.10 in [**Bra06**]) in the metric space $(\mathcal{P}_2(\Omega), W_2)$ to obtain the first statement of Theorem I.16.

Let us now use this result to prove the convergence in non-symmetric Bregman divergence of $\mu_{\mathbb{P}}^{\gamma}$ under the assumption that the population Wasserstein barycenter is unique. By definition (I.15) of $\mu_{\mathbb{P}}^{\gamma}$, we get that

$$\int W_2^2(\mu_{\mathbb{P}}^{\gamma}, \nu) d\mathbb{P}(\nu) - \int W_2^2(\mu_{\mathbb{P}}^{0}, \nu) d\mathbb{P}(\nu) + \gamma(E(\mu_{\mathbb{P}}^{\gamma}) - E(\mu_{\mathbb{P}}^{0})) \leq 0, \tag{I.35}$$

and by definition (I.16) of $\mu_{\mathbb{P}}^{0}$, one has that $\int W_2^2(\mu_{\mathbb{P}}^{\gamma}, \nu) d\mathbb{P}(\nu) - \int W_2^2(\mu_{\mathbb{P}}^{0}, \nu) d\mathbb{P}(\nu) \geq 0$. Therefore, one has that $E(\mu_{\mathbb{P}}^{\gamma}) - E(\mu_{\mathbb{P}}^{0}) \leqslant 0$ and thus, by definition (I.18) of the non-symmetric Bregman divergence, it follows that

$$D_E(\mu_{\mathbb{P}}^{\gamma}, \mu_{\mathbb{P}}^{0}) \leq \langle \nabla E(\mu_{\mathbb{P}}^{0}), \mu_{\mathbb{P}}^{0} - \mu_{\mathbb{P}}^{\gamma}\rangle.$$

Since $\nabla E(\mu_{\mathbb{P}}^{0})$ is assumed to be a continuous function on the compact set Ω, the above inequality and the fact that $\lim_{\gamma \to 0} W_2(\mu_{\mathbb{P}}^{\gamma}, \mu_{\mathbb{P}}^{0}) = 0$ implies that $\lim_{\gamma \to 0} D_E(\mu_{\mathbb{P}}^{\gamma}, \mu_{\mathbb{P}}^{0}) = 0$ since convergence of probability measures for the W_2 metric implies weak convergence. □

I.3.4.2. Convergence of $\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}$ towards $\mu^{\gamma}_{\mathbb{P}}$

In what follows, C denotes a universal constant whose value may change from line to line.

Proof of Theorem I.17. From the subgradient's inequality (I.9) and following the arguments used in the proof of the stability's Theorem I.12, we have that, for each $\boldsymbol{\nu}_i$, $i = 1, \dots, n$, there exists $\phi^{\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}, \boldsymbol{\nu}_i}$ integrable with respect to $\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}(x)dx$ such that for all $\eta \in \mathcal{P}_2(\Omega)$:

$$\left\langle \frac{1}{n}\sum_{i=1}^{n} \phi^{\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}, \boldsymbol{\nu}_i} + \gamma \nabla E(\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}), \eta - \boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n} \right\rangle \geq 0. \tag{I.36}$$

By applying once again the subgradient's inequality, we get

$$\mu^{\gamma}_{\mathbb{P}} \text{ minimizes } J^{\gamma}_{\mathbb{P}} \;\Leftrightarrow \exists \phi \in \partial J^{\gamma}_{\mathbb{P}}(\mu^{\gamma}_{\mathbb{P}}) \text{ s. t. } \langle \phi, \eta - \mu^{\gamma}_{\mathbb{P}} \rangle \geq 0 \text{ for all } \eta \in \mathcal{P}_2(\Omega).$$

Let us explicit the form of a subgradient $\phi \in \partial J^{\gamma}_{\mathbb{P}}(\mu^{\gamma}_{\mathbb{P}})$ using the Theorem of the subdifferential of a sum. We have that $\mu \mapsto W_2^2(\mu, \nu)$ is continuous for all $\nu \in \mathcal{P}_2(\Omega)$. Moreover by symmetry, $\nu \mapsto W_2^2(\mu, \nu)$ is measurable for all $\mu \in \mathcal{P}_2(\Omega)$ and $W_2^2(\mu, \nu) \leq \iint |x-y|^2 d\mu(x)d\nu(y) \leq 2\int |x|^2 d\mu(x) + 2\int |y|^2 d\nu(y) \leq C$ is integrable with respect to $d\mathbb{P}(\nu)$. Hence, by the Theorem of continuity under integral sign, we deduce that $\mu \mapsto \mathbb{E}[W_2^2(\mu, \boldsymbol{\nu})]$ is continuous. Thus we can deal with the sum of the subdifferential and one has that $\partial J^{\gamma}_{\mathbb{P}}(\mu^{\gamma}_{\mathbb{P}}) = \partial_1[\mathbb{E}(W_2^2(\mu^{\gamma}_{\mathbb{P}}, \boldsymbol{\nu}))] + \gamma \nabla E(\mu^{\gamma}_{\mathbb{P}})$, where $\boldsymbol{\nu}$ is still a random measure with distribution $\mathbb{P}$. Also notice that the Theorem 23 in [**Roc74**] implies $\partial_1 \mathbb{E}[W_2^2(\mu^{\gamma}_{\mathbb{P}}, \boldsymbol{\nu})] = \mathbb{E}[\partial_1 W_2^2(\mu^{\gamma}_{\mathbb{P}}, \boldsymbol{\nu})]$. We sum up as

$$\mu^{\gamma}_{\mathbb{P}} \text{ minimizes } J^{\gamma}_{\mathbb{P}} \Leftrightarrow \left\langle \int \phi^{\mu^{\gamma}_{\mathbb{P}}, \nu} d\mathbb{P}(\nu) + \gamma \nabla E(\mu^{\gamma}_{\mathbb{P}}), \eta - \mu^{\gamma}_{\mathbb{P}} \right\rangle \geq 0, \; \forall \; \eta \in \mathcal{P}_2(\Omega). \tag{I.37}$$

In the sequel, to simplify the notation, we use $\boldsymbol{\mu} := \boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}$ and $\eta := \mu^{\gamma}_{\mathbb{P}}$. Therefore thanks to (I.36) and (I.37), we get

$$\begin{aligned}
&d_E(\boldsymbol{\mu}, \eta) = \langle \nabla E(\boldsymbol{\mu}) - \nabla E(\eta), \boldsymbol{\mu} - \eta \rangle \\
\leq &-\frac{1}{\gamma}\left\langle \frac{1}{n}\sum_{i=1}^{n} \phi^{\boldsymbol{\mu}, \boldsymbol{\nu}_i} - \int \phi^{\eta, \nu} d\mathbb{P}(\nu), \boldsymbol{\mu} - \eta \right\rangle \qquad \text{(I.38)} \\
= &\frac{1}{\gamma}\left(\frac{1}{n}\sum_{i=1}^{n} \left[\int \phi^{\boldsymbol{\mu}, \boldsymbol{\nu}_i}(x) d\eta(x) - \int \phi^{\boldsymbol{\mu}, \boldsymbol{\nu}_i}(x) d\boldsymbol{\mu}(x) \right] + \iint \phi^{\eta, \nu} d\mathbb{P}(\nu) d\boldsymbol{\mu}(x) - \iint \phi^{\eta, \nu} d\mathbb{P}(\nu) d\eta(x) \right).
\end{aligned}$$

We would like to switch integrals of the two last terms. In that purpose, we use that $\int W_2^2(\eta, \nu) d\mathbb{P}(\nu) < +\infty$, since $\mathbb{P} \in W_2(\mathbb{P}_2(\Omega))$.

As $0 \leq \int W_2^2(\eta, \nu) d\mathbb{P}(\nu) = \int \left(\int \phi^{\eta, \nu}(x) d\eta(x) + \int \psi^{\eta, \nu}(x) d\nu(y) \right) d\mathbb{P}(\nu)$, we also have that $\iint \phi^{\eta, \nu}(x) d\eta(x) d\mathbb{P}(\nu) < +\infty$. Since $x \mapsto \phi^{\eta, \nu}(x)$ and $\nu \mapsto \phi^{\eta, \nu}(x)$ are measurable, we obtain by Fubini's theorem $\int_{\Omega} \int_{\mathcal{P}_2(\Omega)} \phi^{\eta, \nu} d\mathbb{P}(\nu) d\eta(x) = \int_{\mathcal{P}_2(\Omega)} \int_{\Omega} \phi^{\eta, \nu} d\eta(x) d\mathbb{P}(\nu)$. By the same tools, since

$$\begin{aligned}
\int W_2^2(\boldsymbol{\mu}, \nu) d\mathbb{P}(\nu) &= \int \left(\int \phi^{\boldsymbol{\mu}, \nu}(x) d\boldsymbol{\mu}(x) + \int \psi^{\boldsymbol{\mu}, \nu}(x) d\nu(y) \right) d\mathbb{P}(\nu) \\
&\geq \int \left(\int \phi^{\eta, \nu}(x) d\boldsymbol{\mu}(x) + \int \psi^{\eta, \nu}(x) d\nu(y) \right) d\mathbb{P}(\nu),
\end{aligned}$$

we get $\int \left(\int \phi^{\eta, \nu}(x) d\mu(x) \right) d\mathbb{P}(\nu) < +\infty$, so $\int_{\Omega} \int_{\mathcal{P}_2(\Omega)} \phi^{\eta, \nu} d\mathbb{P}(\nu) d\boldsymbol{\mu}(x) = \int_{\mathcal{P}_2(\Omega)} \int_{\Omega} \phi^{\eta, \nu} d\boldsymbol{\mu}(x) d\mathbb{P}(\nu)$.

Therefore, by the dual formulation of Kantorovich, we have that

$$-\int \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i} d\boldsymbol{\mu}(x) = \int \psi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(y) d\boldsymbol{\nu}_i(y) - \iint |x-y|^2 d\pi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x,y) \tag{I.39}$$

$$-\int \phi^{\eta,\nu} d\eta(x) = \int \psi^{\eta,\nu}(y) d\nu(y) - \iint |x-y|^2 d\pi^{\eta,\nu}(x,y), \tag{I.40}$$

where $\pi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}$ and $\pi^{\eta,\nu}$ are optimal transport plans for the Wasserstein distance. Also, $\phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}$ and $\phi^{\eta,\nu}$ verify the Kantorovich condition, that is

$$\phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) \leq -\psi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(y) + |x-y|^2 \tag{I.41}$$

$$\phi^{\eta,\nu}(x) \leq -\psi^{\eta,\nu}(y) + |x-y|^2. \tag{I.42}$$

Next, the trick is to write $\int \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) d\eta(x) = \iint \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) d\pi^{\eta,\boldsymbol{\nu}_i}(x,y)$ and $\int \phi^{\eta,\nu}(x) d\boldsymbol{\mu}(x) = \iint \phi^{\eta,\nu}(x) d\pi^{\boldsymbol{\mu},\nu}(x,y)$. Thus, by using the equalities (I.39), (I.40) and the inequalities (I.41) and (I.42), the result (I.38) becomes

$$\begin{aligned}\gamma d_E(\boldsymbol{\mu},\eta) \leq & -\frac{1}{n}\sum_{i=1}^n \iint |x-y|^2 d\pi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x,y) + \frac{1}{n}\sum_{i=1}^n \iint |x-y|^2 d\pi^{\eta,\boldsymbol{\nu}_i}(x,y) \\ & + \int \iint |x-y|^2 d\pi^{\boldsymbol{\mu},\nu}(x,y) d\mathbb{P}(\nu) - \int \iint |x-y|^2 d\pi^{\eta,\nu}(x,y) d\mathbb{P}(\nu).\end{aligned} \tag{I.43}$$

We denote

$$S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}} := \int \iint |x-y|^2 d\pi^{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n},\nu}(x,y) d\mathbb{P}(\nu) \; - \frac{1}{n}\sum_{i=1}^n \iint |x-y|^2 d\pi^{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n},\boldsymbol{\nu}_i}(x,y) \tag{I.44}$$

$$S^n_{\mu^\gamma_{\mathbb{P}}} := \frac{1}{n}\sum_{i=1}^n \iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}_i}(x,y) - \mathbb{E}\left(\iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}}(x,y)\right), \tag{I.45}$$

and the previous inequality (I.43) finally writes

$$\gamma d_E(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}}) \leq S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}} + S^n_{\mu^\gamma_{\mathbb{P}}}. \tag{I.46}$$

Taking the expectation with respect to the random measures, (I.46) implies

$$\gamma^2 \mathbb{E}(d^2_E(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}})) \leq 2\mathbb{E}(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2) + 2\mathbb{E}(|S^n_{\mu^\gamma_{\mathbb{P}}}|^2). \tag{I.47}$$

The first term related to $\mu^{\gamma n}_{\mathbb{P}}$ is easy to handle, since for $i = 1, \ldots, n$ the random variables $\iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}_i}(x,y)$ are independent and identically distributed. From the law of large numbers, we can notice that $S^n_{\mu^\gamma_{\mathbb{P}}} \longrightarrow 0$ almost surely when $n \to +\infty$. In particular, we observe that

$$\mathbb{E}\left(|S^n_{\mu^\gamma_{\mathbb{P}}}|^2\right) = \frac{1}{n}\mathrm{Var}\left(\iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}}(x,y)\right) \leqslant \frac{C}{n}. \tag{I.48}$$

Let us now study $\mathbb{E}(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2)$ thanks to the empirical process theory. We recall that the class of functions $\mathcal{H}$ on $\mathcal{P}_2(\Omega)$ is defined as

$$\mathcal{H} = \{h_\mu : \nu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu) \in \mathbb{R}; \mu \in \mathcal{P}_2(\Omega)\}, \tag{I.49}$$

ant its associated norm is $\|G\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} |G(h)|$ where $G : \mathcal{H} \to \mathbb{R}$.
Therefore we obtain

$$S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}} = \int_{\mathcal{P}_2(\Omega)} h_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}(\nu) d\mathbb{P}(\nu) - \int_{\mathcal{P}_2(\Omega)} h_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}(\nu) d\mathbb{P}_n(\nu) := (\mathbb{P}-\mathbb{P}_n)(h_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}) \leq \sup_{h\in\mathcal{H}} |(\mathbb{P}-\mathbb{P}_n)(h)|. \tag{I.50}$$

We define the envelope function of $\mathcal{H}$ by $H : \nu \in \mathcal{P}_2(\Omega) \mapsto \sup_{\mu\in\mathcal{P}_2(\Omega)}\{W_2(\mu,\nu); W_2^2(\mu,\nu)\}$, which is integrable with respect to $\mathbb{P}$ by compacity of Ω. Let then $\mathcal{H}_M$ be the class of

functions $\tilde{h}_\mu := h_\mu \mathbb{1}_{H \leqslant M}$ when h_μ ranges over $\mathcal{H}$. By the triangle reverse inequality, we have for $\tilde{h}_\mu, \tilde{h}_{\mu'} \in \mathcal{H}_M$

$$\|\tilde{h}_\mu - \tilde{h}_{\mu'}\|_{\mathbb{L}1(\mathbb{P}_n)} = \frac{1}{n}\sum_{i=1}^n |W_2(\mu,\nu_i) - W_2(\mu',\nu_i)|\ (W_2(\mu,\nu_i) - W_2(\mu',\nu_i))\mathbb{1}_{H\leqslant M}$$
$$\leqslant W_2(\mu,\mu')\frac{2}{n}\sum_{i=1}^n H(\nu_i)\mathbb{1}_{H\leqslant M} \leqslant 2MW_2(\mu,\mu').$$

We deduce that $N(\varepsilon, \mathcal{H}_M, \mathbb{L}_1(\mathbb{P}_n)) \leqslant N(\frac{\varepsilon}{2M}, K_M, W_2)$ where $K_M = \{\mu \in \mathcal{P}_2(\Omega)\}$ is compact. Then from Borel-Lebesgue, we deduce that $\log N(\varepsilon, \mathcal{H}_M, \mathbb{L}_1(\mathbb{P}_n))$ can be bounded from above by a finite number which does not depend on n. Theorem 2.4.3 in [**VDVW96**] allows us to conclude that $|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|$ tends to 0 almost surely. By the mapping theorem, $|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2$ also tends to 0 a.s. Since it is bounded by a constant only depending on the diameter of Ω, we have that it is bounded by an integrable function. By the theorem of dominated convergence, we get $\mathbb{E}\left(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2\right) \underset{n\to\infty}{\longrightarrow} 0$. Gathering (I.47), (I.48), we get for all $\gamma > 0$

$$\mathbb{E}(d_E^2(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}})) \underset{n\to\infty}{\longrightarrow} 0.$$

□

Rate of convergence between $\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}$ and $\mu^\gamma_{\mathbb{P}}$. In order to have a rate a convergence, we will need existing results on the notion of bracketing number, that is defined below.

Definition I.29. *Given two real-valued functions l and r, the bracket $[l, r]$ is the set of all functions f with $l \leq f \leq r$. An ϵ-bracket is a bracket $[l, r]$ with $\|l - r\| < \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.*

Proof of Theorem I.18. This proof follows from the proof of Theorem I.17. We recall from (I.47) that

$$\gamma^2\mathbb{E}(d_E^2(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}})) \leq 2\mathbb{E}(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2) + 2\mathbb{E}(|S^n_{\mu^\gamma_{\mathbb{P}}}|^2), \text{ for } S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}, S^n_{\mu^\gamma_{\mathbb{P}}} \text{ defined in (I.44), (I.45)} \quad \text{(I.51)}$$

where by (I.48), $\mathbb{E}\left(|S^n_{\mu^\gamma_{\mathbb{P}}}|^2\right) \leqslant \frac{C}{n}$ and by (I.50) we have for $\mathcal{H}$ given in (I.49)

$$|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}| \leqslant \sup_{h\in\mathcal{H}} |(\mathbb{P} - \mathbb{P}_n)(h)|.$$

Rewriting this term, we get $|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}| \leqslant \frac{1}{\sqrt{n}}\|\mathbb{G}_n\|_{\mathcal{H}}$ where $\mathbb{G}_n(h) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})(h)$. We obtain

$$\mathbb{E}\left(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2\right) \leq \frac{1}{n}\mathbb{E}\left(\|\mathbb{G}_n\|^2_{\mathcal{H}}\right) = \frac{1}{n}\|\ \|\mathbb{G}_n\|_{\mathcal{H}}\|^2_{\mathbb{L}_2(\mathbb{P})}. \quad \text{(I.52)}$$

We then use the following Theorem 2.14.1. of [**VDVW96**] to control the last term in (I.52).

Theorem I.30. *Let $\mathcal{H}$ be a Q-measurable class of measurable functions with measurable envelope function H. Then for $p \geq 1$, we have*

$$\|\ \|\mathbb{G}_n\|_{\mathcal{H}}\|_{\mathbb{L}_p(Q)} \leq CI(1,\mathcal{H})\|H\|_{\mathbb{L}_{2\vee p}(Q)} \quad \text{(I.53)}$$

with C a constant, $I(1,\mathcal{H})$ defined in (I.17) and H an envelope function.

Gathering the results of (I.47), (I.48), (I.52) and (I.53), we get

$$\mathbb{E}(d_E^2\left(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}}\right)) \leq \frac{1}{\gamma^2 n}\left(C + CI(1,\mathcal{H})\|H\|_{\mathbb{L}_2(\mathbb{P})}\right), \quad \text{(I.54)}$$

which is completely valid for any Ω compact in $\mathbb{R}^d$. The norm $\|H\|_{\mathbb{L}_2(\mathbb{P})}$ is clearly finite since for all $\nu \in \mathcal{P}_2(\Omega)$, $|h_\mu(\nu)| \leqslant 4c^2_\Omega$, with $c^2_\Omega = \sup_{x\in\Omega}|x|^2$. □

PROOF OF THE THEOREM I.19. We assume here that $\Omega \subset \mathbb{R}$ is compact. It remains to study the term $I(1,\mathcal{H})$ defined in (I.17) for $\mathcal{H}$ in (I.49). By the triangle reverse inequality, we have

$$|h_\mu(\nu) - h_{\mu'}(\nu)| = |W_2(\nu,\mu) - W_2(\nu,\mu')|\ (W_2(\nu,\mu) + W_2(\nu,\mu')) \leq\ W_2(\mu,\mu')\ 2H(\nu).$$

Then, from Theorem 2.7.11 in [**VDVW96**], and since Theorem 4 in [**KT59**] allows us to bound the metric entropy by the bracket entropy, we get

$$\begin{aligned}\log N(\epsilon\|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \|\cdot\|_{\mathbb{L}_2(Q)}) &\leq \log N_{[]}(\epsilon\|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \|\cdot\|_{\mathbb{L}_2(Q)}) \\ &\leq \log N(\epsilon, \mathcal{P}_2(\Omega), W_2) \leq \log N_{[]}(\epsilon, \mathcal{P}_2(\Omega), W_2). \end{aligned} \tag{I.55}$$

Also, for $d = 1$, we have

$$W_2(\mu,\mu') = \left(\int_0^1 |F_\mu^-(t) - F_{\mu'}^-(t)|^2 dt\right)^{1/2} = \|F_\mu^- - F_{\mu'}^-\|_{\mathbb{L}_2([0,1])} \tag{I.56}$$

where F_μ^- is the quantile function of the cumulative distribution function F_μ of μ. We denote by $\mathcal{G} = \{F_\mu^-, \mu \in \mathcal{P}_2(\Omega)\}$ the class of quantile functions of probability measures μ in $\mathcal{P}_2(\Omega)$, which are monotonic functions. Moreover, we can observe that $F_\mu^- : [0,1] \to [F_\mu^-(0), F_\mu^-(1)] \subseteq \Omega$, where Ω is a compact included in $\mathbb{R}$. Hence, $\mathcal{G}$ is uniformly bounded, say by a constant $M > 0$. By Theorem 2.7.5. in [**VDVW96**] on the bracket entropy of the class of monotonic functions, we obtain that $\log N_{[]}(\epsilon, \mathcal{G}, \mathbb{L}_2[0,1]) \leq \frac{CM}{\epsilon}$, for some constant $C > 0$. Finally, from relations (I.55) and (I.56), we can deduce that

$$I(1,\mathcal{H}) = \sup_Q \int_0^1 \sqrt{1 + \log N(\epsilon\|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \mathbb{L}_2(Q))}d\epsilon \leq \int_0^1 \sqrt{1 + \frac{CM}{\epsilon}}d\epsilon < \infty.$$

□

PROOF OF THE THEOREM I.20. We here consider that Ω is a compact of $\mathbb{R}^d$ and that E is given by (I.21). Let us begin by underlining that since the norm of a Sobolev space is weakly* lower semicontinuous, E is indeed lower semicontinuous for the Wasserstein metric. Supposing that Ω has a C^1 boundary, we have by the Sobolev embedding theorem that $H^k(\Omega)$ is included in the Hölder space $C^{m,\beta}(\bar{\Omega})$ for any integer m and $\beta \in]0,1]$ satisfying $m + \beta = k - d/2$. Hence, the densities of $\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma$ and $\mu_{\mathbb{P}}^\gamma$ given by (I.14) and (I.15) belong to $C^{m,\beta}(\bar{\Omega})$.
From the Theorem I.18, we will use that:

$$\mathbb{E}(d_E^2\left(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma\right)) \leq \frac{1}{\gamma^2 n}\left(C + CI(1,\mathcal{H})\|H\|_{\mathbb{L}_2(\mathbb{P})}\right).$$

Similarly, since Ω is compact, we have $\|H\|_{\mathbb{L}_2(Q)} < \infty$, where $H(\nu) = \sup_{\mu\in\mathcal{D}(E)} \{W_2(\mu,\nu); W_2^2(\mu,\nu)\}$ where $\mathcal{D}(E)$ is defined by (I.2). Thus, instead of controlling the metric entropy $N(\epsilon\|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \|\cdot\|_{\mathbb{L}_2(Q)})$, it is enough to bound the metric entropy $N(\epsilon, \mathcal{D}(E), W_2)$ thanks to Theorem 2.7.11 in [**VDVW96**].

To this end, since $\mu, \mu' \in \mathcal{D}(E)$ are a.c. measures, one has that

$$W_2(\mu,\mu') \leq \left(\int_\Omega |T(x) - T'(x)|^2 dx\right)^{1/2} \text{ where } T\#\lambda^d = \mu \text{ and } T'\#\lambda^d = \mu',$$

where λ^d denotes the Lebesgue measure on Ω. Thanks to Theorem 3.3 in [**DPF14**] on the regularity of optimal maps (results initally due to Caffarelli, [**Caf92**] and [**Caf96**]), the coordinates of T and T' are $C^{m+1,\beta}(\bar{\Omega})$ functions $\lambda^d-a.e.$ Thus, we can bound $N(\epsilon, \mathcal{D}(E), W_2)$ by the bracket entropy $N_{[]}(\epsilon, C^{m+1,\beta}(\bar{\Omega}), \mathbb{L}_2(\Omega))$ since $|T(x)-T'(x)|^2 = \sum_{j=1}^d |T_j(x_j)-T_j'(x_j)|^2$

where $T_j, T_j' : \Omega \to \mathbb{R}$. Now, by Corollary 2.7.4 in [VDVW96], we get

$$\log N_{[]}(\epsilon, C^{m+1,\beta}(\bar{\Omega}), \mathbb{L}_2(\Omega)) \leq K \left(\frac{1}{\epsilon}\right)^V$$

for any $V \geq d/(m+1)$. Hence, as soon as $V/2 < 1$ (for which the condition $k > d-1$ is sufficient if $V = d/(m+1)$), the upper bound in (I.20) is finite for $\mathcal{H} = \{h_\mu : \nu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu) \in \mathbb{R}; \mu \in \mathcal{D}(E)\}$, which yields the result of Theorem I.20 by finally following the arguments in the proof of Theorem I.19 and since $d_{E_G} \leqslant d_E$. □

I.3.5. Strong convexity of the Sinkhorn divergence, Proof of Theorem I.24

The proof of Theorem I.24 relies on the analysis of the eigenvalues of the Hessian matrix $\nabla^2 H_q^*(g)$ of the functional H_q^*.

Proposition I.31. *For all $g \in \mathbb{R}^N$, $\nabla^2 H_q^*(g)$ admits $\lambda_N = 0$ as eigenvalue with its associated normalized eigenvector $v_N := \frac{1}{\sqrt{N}}\mathbb{1}_N \in \mathbb{R}^N$, which means that rank$(\nabla^2 H_q^*(g)) \leqslant N-1$ for all $g \in \mathbb{R}^N$ and $q \in \Sigma_N$.*

Proof. Let $g \in \mathbb{R}^N$, then by Theorem I.23

$$\begin{aligned}\nabla^2 H_q^*(g) v_N &= \frac{1}{\varepsilon}\operatorname{diag}(\alpha) K \frac{q}{K\alpha} - \frac{1}{\varepsilon}\operatorname{diag}(\alpha) K \operatorname{diag}\left(\frac{q}{(K\alpha)^2}\right) K\alpha \\ &= \frac{1}{\varepsilon}\operatorname{diag}(\alpha) K \frac{q}{K\alpha} - \frac{1}{\varepsilon}\operatorname{diag}(\alpha) K \frac{q}{K\alpha} = 0,\end{aligned}$$

and $\lambda_N = 0$ is an eigenvalue of $\nabla^2 H_q^*(g)$. □

Let $(v_k)_{1\leqslant k\leqslant N}$ be the eigenvectors of $\nabla^2 H_q^*(g)$, depending on both q and g, with their respective eigenvalues $(\lambda_k)_{1\leqslant k\leqslant N}$. As the Hessian matrix is symmetric and diagonalizable, let us now prove that the eigenvalues associated to the eigenvectors $(v_k)_{1\leqslant k\leqslant N-1}$ of $\nabla^2 H_q^*(g)$ are all positive.

Proposition I.32. *For all $q \in \Sigma_N$ and $g \in \mathbb{R}^N$, we have that*

$$0 = \lambda_N < \lambda_k \qquad \text{for all} \quad 1 \leqslant k \leqslant N-1.$$

Proof. The eigenvalue $\lambda_N = 0$ associated to v_N has been treated in Proposition I.31. Let $v \in V = (\mathrm{Vect}(v_N))^\perp$ (i.e. v does not have constant coordinates) an eigenvector of $\nabla^2 H_q^*(g)$. Hence we can suppose that, let say $v^{(j)}$, is its larger coordinate, and that there exists $i \neq j$ such that $v^{(j)} > v^{(i)}$. Without loss of generality, we can assume that $v^{(j)} > 0$. Then

$$\begin{aligned}[\nabla^2 H_q^*(g) v]_j &= \left[\frac{1}{\varepsilon}\left(\operatorname{diag}\left(\operatorname{diag}(\alpha) K \frac{q}{K\alpha}\right)\right) v\right]_j - \left[\frac{1}{\varepsilon}\operatorname{diag}(\alpha) K \operatorname{diag}\left(\frac{q}{(K\alpha)^2}\right) K \operatorname{diag}(\alpha) v\right]_j \\ &= \frac{1}{\varepsilon}\alpha_j v^{(j)} \sum_{i=1}^N K_{ji}\frac{q_i}{[K\alpha]_i} - \frac{1}{\varepsilon}\sum_{i=1}^N\sum_{m=1}^N \alpha_j K_{jm}\frac{q_m}{[K\alpha]_m^2}\alpha_i K_{mi} v^{(i)} \\ &> \frac{1}{\varepsilon}\alpha_j v^{(j)} \sum_{i=1}^N K_{ji}\frac{q_i}{[K\alpha]_i} - \frac{1}{\varepsilon}\sum_{i=1}^N\sum_{m=1}^N \alpha_j K_{jm}\frac{q_m}{[K\alpha]_m^2}\alpha_i K_{mi} v^{(j)} \text{ since } v^{(j)} \geqslant v^{(i)}, \forall i \\ &= 0 \qquad \text{since} \quad \sum_{i=1}^N \alpha_i K_{im} = [K\alpha]_m.\end{aligned}$$

Thus $\lambda v^{(j)} = [\nabla^2 H_q^*(g) v]_j > 0$, and we necessarily have that $\lambda > 0$. □

The set of eigenvalues of $\nabla^2 H_q^*(g)$ is also bounded from above.

PROPOSITION I.33. *For all $q \in \Sigma_N$ and $g \in \mathbb{R}^N$ we have that $Tr(\nabla^2 H_q^*(g)) \leqslant \frac{1}{\varepsilon}$ and thus $\lambda_k \leqslant 1/\varepsilon$ for all $k = 1, \dots, N$.*

PROOF. We directly get from Theorem I.23 that

$$\mathrm{Tr}(\nabla^2 H_q^*(g)) \leqslant \frac{1}{\varepsilon}\mathrm{Tr}\left(\mathrm{diag}\underbrace{\left(\mathrm{diag}(\alpha)K\frac{q}{K\alpha}\right)}_{\in\,\Sigma_N}\right) = \frac{1}{\varepsilon}.$$

□

We can now provide the proof of Theorem I.24. Since H_q is convex, proper and lower-semicontinuous, we know by the Fenchel-Moreau theorem that $H_q^{**} = H_q$. Hence by Corollary 12.A in the Rockafellar's book [**Roc74**], we have that

$$\nabla H_q = (\nabla H_q^*)^{-1}, \tag{I.57}$$

in the sense that $\nabla H_q^* \circ \nabla H_q(r) = r$ for any $r \in \Sigma_N$.

To continue the proof, we focus on a definition of the function H_q restricted to the linear subspace V. Let $(v_1, \dots, v_{N-1})$ be an orthonormal basis of $V = (\mathrm{Vect}(v_N))^\perp$ and $P = [v_1 \ \cdots \ v_{N-1}] \in \mathbb{R}^{N\times(N-1)}$ the matrix of the basis. Remark that PP^T is the matrix of the orthogonal projection onto V, and that $PP^T = I_N - v_N v_N^T$. If we define $\tilde{\Sigma}_{N-1} := P^T\Sigma_N \in \mathbb{R}^{N-1}$, then for $r \in \Sigma_N$, there exists $\tilde{r} \in \tilde{\Sigma}_{N-1}$ such that $r = P\tilde{r} + \frac{1}{\sqrt{N}}v_N$. Hence we can introduce the functional $\tilde{H}_q : \tilde{\Sigma}_{N-1} \to \mathbb{R}$ defined by

$$\tilde{H}_q(\tilde{r}) := H_q\left(P\tilde{r} + \frac{1}{\sqrt{N}}v_N\right).$$

For $\tilde{g} \in \mathbb{R}^{N-1}$ we have that

$$\begin{aligned}
\tilde{H}_q^*(\tilde{g}) &= \max_{\tilde{r}\in\tilde{\Sigma}_{N-1}} \langle \tilde{g}, \tilde{r}\rangle - \tilde{H}_q(\tilde{r}) \\
&= \max_{r\in\Sigma_N} \langle \tilde{g}, P^T r - u_N\rangle - H_q(r) \quad \text{where } u_N = \frac{1}{N}\left(\sum_{i=1}^N v_1^{(i)}, \dots, \sum_{i=1}^N v_{N-1}^{(i)}\right) \\
&= H_q^*(P\tilde{g}) - \langle \tilde{g}, u_N\rangle.
\end{aligned}$$

Since H_q^* is C^∞ (see Theorem I.23), we can differentiate $\tilde{H}_q^*$ with respect to $\tilde{g}$ to obtain that

$$\begin{aligned}
\nabla\tilde{H}_q^*(\tilde{g}) &= P^T\nabla H_q^*(P\tilde{g}) - u_N \\
\nabla^2\tilde{H}_q^*(\tilde{g}) &= P^T\nabla^2 H_q^*(P\tilde{g})P.
\end{aligned}$$

By Proposition I.32, we know that $\nabla^2 H_q^*(P\tilde{g}) \in \mathbb{R}^{N\times N}$ admits a unique eigenvalue equals to 0 which is associated to the eigenvector v_N. All other eigenvalues are positive (Proposition I.32) and bounded from above by $1/\varepsilon$ (Proposition I.33). Since $\nabla\tilde{H}_q^* : \mathbb{R}^{(N-1)} \to \mathbb{R}^{(N-1)}$ is a C^∞-diffeomorphism, using equality (I.57) (that is also valid for $\tilde{H}_q$), we have that

$$\begin{aligned}
\nabla^2\tilde{H}_q(\tilde{r}) &= \nabla\left((\nabla\tilde{H}_q^*)^{-1}(\tilde{r})\right) \\
&= [\nabla^2\tilde{H}_q^*((\nabla\tilde{H}_q^*)^{-1}(\tilde{r}))]^{-1} \\
&= [\nabla^2\tilde{H}_q^*(\nabla\tilde{H}_q(\tilde{r}))]^{-1},
\end{aligned}$$

where the second equality follows from the global inversion theorem, and the last one uses again equality (I.57). Thus we get

$$\lambda_{\min}(\nabla^2\tilde{H}_q(\tilde{r})) \geqslant \varepsilon.$$

The above inequality implies the strong convexity of $\tilde{H}_q$ which reads for $\tilde{r}_0, \tilde{r}_1 \in \tilde{\Sigma}_{n-1}$

$$\tilde{H}_q(\tilde{r}_1) \geqslant \tilde{H}_q(\tilde{r}_0) + \nabla \tilde{H}_q(\tilde{r}_0)^T(\tilde{r}_1 - \tilde{r}_0) + \frac{\varepsilon}{2}\|\tilde{r}_1 - \tilde{r}_0\|_2,$$

and this translates for H_q and $r_0, r_1 \in \Sigma_N$ to

$$H_q(r_1) \geqslant H_q(r_0) + \nabla H_q(r_0)^T PP^T(r_1 - r_0) + \frac{\varepsilon}{2}\|PP^T(r_1 - r_0)\|^2.$$

To conclude, we remark that $(r_1 - r_0) \in V$ (indeed one has that $r_1 - r_0 = \sum_{j=1}^{N-1}\langle v_j, r_1 - r_0\rangle v_j$ since $\langle v_N, r_1 - r_0\rangle = 0$ and thus $PP^T(r_1 - r_0) = r_1 - r_0$). Hence, we finally obtain the strong convexity of H_q

$$H_q(r_1) \geqslant H_q(r_0) + \nabla H_q(r_0)^T(r_1 - r_0) + \frac{\varepsilon}{2}\|(r_1 - r_0)\|^2.$$

This completes the proof of Theorem I.24.

I.3.6. Lipschitz constant of H_q, Proof of Lemma I.25

We recall the dual version of the Sinkhorn minimization problem, stated in (A.10):

$$W_{2,\varepsilon}^2(r, q) = \max_{\alpha,\beta\in\mathbb{R}^N} \alpha^T r + \beta^T q - \sum_{1\leqslant m,\ell\leqslant N} \varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m - \beta_\ell)} \tag{I.58}$$

where $C_{m\ell}$ are the entries of the matrix cost C. We also recall the Lemma I.25 of interest.

Lemma I.34. *Let $q \in \Sigma_N$ and $0 < \rho < 1$. Then, one has that $r \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ with*

$$L_{\rho,\varepsilon} = \left(\sum_{1\leqslant m\leqslant N}\left(2\varepsilon\log(N) + \sup_{1\leqslant \ell,k\leqslant N}|C_{m\ell} - C_{k\ell}| - 2\varepsilon\log(\rho)\right)^2\right)^{1/2}. \tag{I.59}$$

Proof. Let $r, s, q \in \Sigma_N$. We denote by $(\alpha^{q,r}, \beta^{q,r})$ a pair of optimal dual variables in the problem (I.58). Then, we have that

$$\begin{aligned}
|H_q(r) - H_q(s)| &= (H_q(r) - H_q(s))\mathbb{1}_{H_q(r)\geqslant H_q(s)} + (H_q(s) - H_q(r))\mathbb{1}_{H_q(r)\leqslant H_q(s)} \\
&\leqslant \left(\langle\alpha^{q,r}, r\rangle + \langle\beta^{q,r}, q\rangle - \sum_{m,\ell}\varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,r} - \beta_\ell^{q,r})} - \langle\alpha^{q,r}, s\rangle - \langle\beta^{q,r}, q\rangle + \right. \\
&\qquad \left. \sum_{m,\ell}\varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,r} - \beta_\ell^{q,r})}\right)\mathbb{1}_{(H_q(r)\geqslant H_q(s))} \\
&\quad + \left(\langle\alpha^{q,s}, s\rangle + \langle\beta^{q,s}, q\rangle - \sum_{m,\ell}\varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,s} - \beta_\ell^{q,s})} - \langle\alpha^{q,s}, r\rangle - \langle\beta^{q,s}, q\rangle + \right. \\
&\qquad \left. \sum_{m,\ell}\varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,s} - \beta_\ell^{q,s})}\right)\mathbb{1}_{(H_q(r)\leqslant H_q(s))} \\
&\leqslant \sup_{\alpha\in\{\alpha^{q,r},\alpha^{q,s}\}}|\langle\alpha, r - s\rangle| \leqslant \sup_{\alpha\in\{\alpha^{q,r},\alpha^{q,s}\}}|\alpha|\ |r - s|.
\end{aligned} \tag{I.60}$$

Let us now prove that the norm of the dual variable $\alpha^{q,r}$ (resp. $\alpha^{q,s}$) is bounded by a constant not depending on q and r (resp. q and s). To this end, we follow some of the arguments in the

proof of Proposition A.1 in [GCB16]. Since the dual variable $\alpha^{q,r}$ achieves the maximum in equation (I.58), we have that for any $1 \leqslant m \leqslant N$

$$r_m - \sum_{1 \leqslant \ell \leqslant N} e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,r} - \beta_\ell^{q,r})} = 0.$$

Let $r \in \Sigma_N^\rho$. Hence, $r_m \neq 0$, and thus one may define $\lambda_m = \varepsilon \log(r_m)$. Then, it follows from the above equality that $\sum_{1 \leqslant \ell \leqslant N} e^{-\frac{1}{\varepsilon}(C_{m\ell} + \lambda_m - \alpha_m^{q,r} - \beta_\ell^{q,r})} = 1$ which implies that

$$\alpha_m^{q,r} = -\varepsilon \log \left(\sum_{1 \leqslant \ell \leqslant N} e^{-\frac{1}{\varepsilon}(C_{m\ell} + \lambda_m - \beta_\ell^{q,r})} \right).$$

Now, for each $1 \leqslant m \leqslant N$, we define

$$\tilde{\alpha}_m^{q,r} = \min_{1 \leqslant \ell \leqslant N} \{C_{m\ell} + \lambda_m - \beta_\ell^{q,r}\} = \min_{1 \leqslant \ell \leqslant N} \{C_{m\ell} - \beta_\ell^{q,r}\} + \lambda_m, \tag{I.61}$$

and we consider the inequality

$$|\alpha_m^{q,r} - \alpha_k^{q,r}| \leqslant |\alpha_m^{q,r} - \tilde{\alpha}_m^{q,r}| + |\tilde{\alpha}_m^{q,r} - \tilde{\alpha}_k^{q,r}| + |\tilde{\alpha}_k^{q,r} - \alpha_k^{q,r}|. \tag{I.62}$$

By equation (I.61) one has that $\tilde{\alpha}_m^{q,r} + \beta_\ell^{q,r} - C_{m\ell} - \lambda_m \leqslant 0$. Hence we get

$$-\alpha_m^{q,r} = \varepsilon \log \left(\sum_{1 \leqslant \ell \leqslant N} e^{-\frac{1}{\varepsilon}\tilde{\alpha}_m^{q,r}} e^{\frac{1}{\varepsilon}(\tilde{\alpha}_m^{q,r} + \beta_\ell^{q,r} - C_{m\ell} - \lambda_m)} \right) \leqslant -\tilde{\alpha}_m^{q,r} + \varepsilon \log(N). \tag{I.63}$$

On the other hand, using the inequality

$$\sum_{1 \leqslant \ell \leqslant N} e^{-\frac{1}{\varepsilon}(C_{m\ell} + \lambda_m - \beta_\ell^{q,r})} \geqslant e^{-\frac{1}{\varepsilon}(C_{m\ell_*} + \lambda_m - \beta_{\ell_*}^{q,r})} = e^{-\frac{1}{\varepsilon}\tilde{\alpha}_m^{q,r}},$$

where ℓ_* is a value of $1 \leqslant \ell \leqslant N$ achieving the minimum in (I.61), we obtain that

$$-\alpha_m^{q,r} \geqslant -\tilde{\alpha}_m^{q,r}. \tag{I.64}$$

By combining inequalities (I.63) and (I.64), we finally have

$$|\tilde{\alpha}_m^{q,r} - \alpha_m^{q,r}| \leqslant \varepsilon \log(N). \tag{I.65}$$

To conclude, it remains to remark that, by equation (I.61), the vector $(\tilde{\alpha}_m^{q,r} - \lambda_m)_{1 \leqslant m \leqslant N}$ is the c-transform of the vector $(\beta_\ell^{q,r})_{1 \leqslant \ell \leqslant N}$ for the cost matrix C. Therefore, by using standard results in optimal transport which relate c-transforms to the modulus of continuity of the cost (see e.g. [San15], p. 11) one obtains that

$$|\tilde{\alpha}_m^{q,r} - \tilde{\alpha}_k^{q,r} + \lambda_k - \lambda_m| \leqslant \sup_{1 \leqslant \ell \leqslant N} |C_{m\ell} - C_{k\ell}|,$$

which implies that

$$|\tilde{\alpha}_m^{q,r} - \tilde{\alpha}_k^{q,r}| \leqslant \sup_{1 \leqslant \ell \leqslant N} |C_{m\ell} - C_{k\ell}| + \varepsilon |\log(r_m) - \log(r_k)|. \tag{I.66}$$

By combining the upper bounds (I.65) and (I.66) with the decomposition (I.62) we finally come to the inequality

$$|\alpha_m^{q,r} - \alpha_k^{q,r}| \leqslant 2\varepsilon \log(N) + \sup_{1 \leqslant \ell \leqslant N} |C_{m\ell} - C_{k\ell}| + \varepsilon |\log(r_m) - \log(r_k)|.$$

Since the dual variables achieving the maximum in equation (I.58) are defined up to an additive constant, one may assume that $\alpha_k^{q,r} = 0$. Under such a condition, we finally obtain that

$$|\alpha| \leqslant \left(\sum_{1 \leqslant m \leqslant N} \left(2\varepsilon \log(N) + \sup_{1 \leqslant k \leqslant N} \left\{ \sup_{1 \leqslant \ell \leqslant N} |C_{m\ell} - C_{k\ell}| + \varepsilon |\log(r_m) - \log(r_k)| \right\} \right)^2 \right)^{1/2}.$$

Using inequality (I.60) and the assumption that $r \in \Sigma_N^\rho$, we can thus conclude that $r \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ for

$$L_{\rho,\varepsilon} = \left(\sum_{1 \leqslant m \leqslant N} \left(2\varepsilon \log(N) + \sup_{1 \leqslant \ell, k \leqslant N} |C_{m\ell} - C_{k\ell}| - 2\varepsilon \log(\rho) \right)^2 \right)^{1/2}.$$

□

Chapter II

Use of regularized barycenters in alignment tasks

This chapter is based on the preprint [**BCP18a**].

II.1. Introduction

II.1.1. Motivations

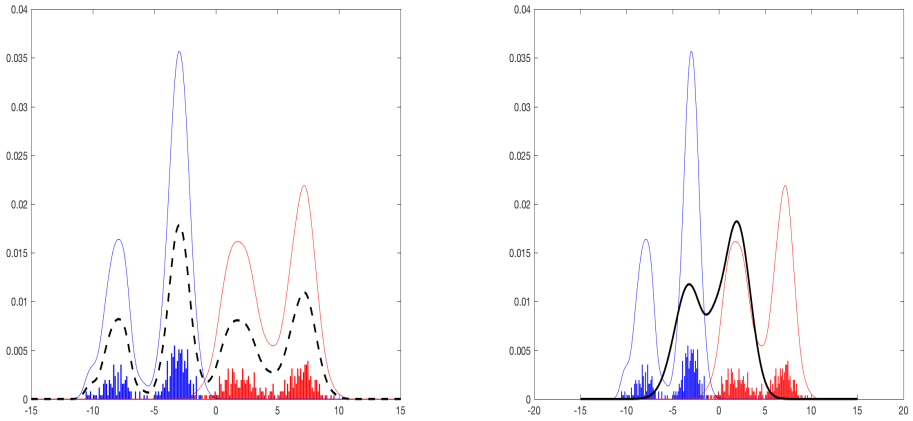
This paper is concerned with the problem of aligning (or registering) elements of a dataset that can be modeled as n random densities, or more generally, probability measures supported on $\mathbb{R}^d$. As raw data in the form of densities are generally not directly available, we focus on the setting where one has access to a set of random vectors $(X_{i,j})_{1\leqslant j\leqslant p_i;\ 1\leqslant i\leqslant n}$ in $\mathbb{R}^d$ organized in the form of n subjects (or multiple point clouds), such that $X_{i,1},\dots,X_{i,p_i}$ are iid observations sampled from a random density $\boldsymbol{f}_i$ for each $1\leqslant i\leqslant n$. In the presence of phase variation in the observations due to mis-alignment in the acquisition process, it is necessary to use a registration step to obtain meaningful notions of mean and variance from the analysis of the dataset. In Figure II.1(a), we display a simulated example of $n=2$ random distributions made of observations sampled from Gaussian mixtures $\boldsymbol{f}_i$ with two components whose means and variances are randomly chosen for each distribution. Certainly, one can estimate a mean density using a preliminary smoothing step (with a Gaussian kernel K and data-driven choices of the bandwidth parameters $(h_i)_{i=1,\dots,n}$) followed by standard averaging, that is considering

$$\bar{f}_{n,p}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i h_i}\sum_{j=1}^{p_i}K\left(\frac{x-X_{i,j}}{h_i}\right),\ x\in\mathbb{R}^d. \tag{II.1}$$

Unfortunately this leads to an estimator which is not consistent with the shape of the $\boldsymbol{f}_i$'s. Indeed, the estimator $\bar{f}_{n,p}$ (Euclidean mean) has four modes due to mis-alignment of the data from different subjects.

II.1.2. Contributions

In this work, in order to simultaneously align and smooth multiple point clouds (in the idea of recovering the underlying density function), we average the data using the notion



(a) (b)

FIGURE II.1. A simulated example of $n = 2$ subjects made of $p_1 = p_2 = 300$ observations sampled from Gaussian mixtures with random means and variances. The red and blue bar graphs are histograms with bins of equal and very small size to display the two sets of observations. The red and blue curves represent the kernel density estimators associated to each subject with data-driven choices (using cross-validation) of the bandwidths. (a) The dashed black curve is the Euclidean mean $\bar{f}_{n,p}$ of the red and blue densities. (b) The solid black curve is the entropy regularized Wasserstein barycenter $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ (defined in (II.4)) of the raw data using a Sinkhorn divergence and the numerical approach from [**CP16b**], with a data-driven choice for $\hat{\varepsilon} = 2.55$.

of Wasserstein barycenter (A.11), as it has been shown to be a relevant tool to account for phase variability in density registration [**BGKL18**, **PZ16**, **PZ17**]. In what follows, we consider two approaches for the computation of a regularized Wasserstein barycenter of n discrete probability measures given by

$$\hat{\boldsymbol{\nu}}_i^{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{X_{i,j}} \qquad \text{for} \quad 1 \leqslant i \leqslant n, \tag{II.2}$$

from observations $(X_{i,j})_{1 \leqslant j \leqslant p_i;\, 1 \leqslant i \leqslant n}$.

The first one is presented in Section I.1 and consists in adding a convex penalization term to the definition of an empirical Wasserstein barycenter [**AC11**]. We recall that it leads to the estimator

$$\hat{\boldsymbol{\mu}}^{\gamma}_{n,p} = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\Omega)} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \hat{\boldsymbol{\nu}}_i^{p_i}) + \gamma E(\mu), \tag{II.3}$$

where $\gamma > 0$ is a regularization parameter, and $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ is a smooth and convex penalty function which enforces the measure $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ to be absolutely continuous. In this chapter, we discuss the choice of the penalty function E, as well as the numerical computation of $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ (using an appropriate discretization of Ω and a binning of the data), and its benefits for statistical data analysis.

Another way to regularize an empirical Wasserstein barycenter is to use the notion of entropy regularized optimal transportation to compute an entropy regularized Wasserstein

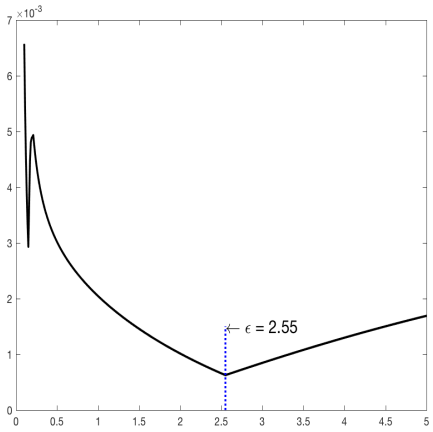


FIGURE II.2. The GL's trade-off function associated to the entropy regularized Wasserstein barycenters of the dataset in Figure II.1, for ε ranging from 0.1 to 5

barycenter, presented in Section I.2 and given by

$$\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} = \underset{r\in\Sigma^{\rho}_{N}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} W^2_{2,\varepsilon}(r, \hat{\boldsymbol{q}}^{p_i}_i) \tag{II.4}$$

The selection of the regularization parameters γ or ϵ is the main issue for computing adequate penalized or Sinkhorn barycenters in practice. We here rely on the Goldenshluger-Lepski (GL) principle in order to perform an automatic calibration of such parameters.

II.1.2.1. Data-driven choice of the regularizing parameters

The main contribution in this work is to propose a data-driven choice for the regularization parameters γ in (II.3) and ε in (II.4) using the Goldenshluger-Lepski (GL) method (as formulated in [**LM16**]), which leans on a bias-variance trade-off function, described in details in Section II.3.1. The method consists in comparing estimators pairwise, for a given range of regularization parameters, with respect to a given loss function. It provides an optimal regularization parameter that minimizes a bias-variance trade-off function. We displayed in Figure II.2 this functional for the dataset of Figure II.1, which leads to an optimal (in the sense of GL's strategy) parameter choice $\hat{\varepsilon} = 2.55$. The entropy regularized Wasserstein barycenter in Figure II.1(b) is thus chosen accordingly.

From the results on simulated data displayed in Figure II.1(b), it is clear that computing the regularized Wasserstein barycenter $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ (with an appropriate choice for ε) leads to the estimation of mean density whose shape is consistent with the distribution of the data for each subject. In some sense, the regularization parameters γ and ε may also be interpreted as the usual bandwidth parameter in kernel density estimation, and their choice greatly influences the shape of the estimators $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ and $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ (see Figure II.7 and Figure II.8 in Section II.3).

To choose the optimal parameter, the GL's strategy requires some variance information through the knowledge of an upper bound on the decay to zero of the expected $\mathbb{L}_2(\Omega)$ distance between a regularized empirical barycenter (computed from the data) and its population counterpart. These bounds have been proven in Chapter I, Section I.1 pour the penalized Wasserstein barycenters and Section I.2 for the entropy regularized Wasserstein barycenters.

II.1.2.2. Computation issues: binning of the data and discretization of Ω

In our numerical experiments we consider algorithms for computing regularized barycenters from a set of discrete measures (or histograms) defined on possibly different grids of points of $\mathbb{R}^d$ (or different partitions). They are numerical approximations of the regularized Wasserstein barycenters $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ and $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ by a discrete measure of the form $\sum_{k=1}^{N} w_k \delta_{x_k}$ using a fixed grid $\Omega_N = \{x_1, \dots, x_N\}$ of N equally spaced points $x_k \in \mathbb{R}^d$ (bin locations). For simplicity, we adopt a binning of the data (II.2) on the same grid, leading to a dataset of discrete measures (with supports included in Ω_N) that we denote

$$\tilde{\boldsymbol{q}}_i^{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\tilde{X}_{i,j}}, \text{ where } \tilde{X}_{i,j} = \operatorname*{arg\,min}_{x \in \Omega_N} |x - X_{i,j}|, \tag{II.5}$$

for $1 \leqslant i \leqslant n$. In this paper, we rely on the smooth dual approach proposed in [**CP16b**] to compute penalized and Sinkhorn barycenters on a grid of equi-spaced points in Ω (after a proper binning of the data).

Binning (i.e. choosing the grid Ω_N) surely incorporates some sort of additional regularization. A discussion on the influence of the grid size N on the smoothness of the barycenter is proposed in Section II.3.1 where we describe the GL's strategy. Besides, in our simulations, the choice of N is mainly guided by numerical issues on the computational cost of the algorithms used to approximate $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ and $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$.

II.1.2.3. Registration of flow cytometry data

In biotechnology, flow cytometry is a high-throughput technique that can measures a large number of surface and intracellular markers of single cell in a biological sample. With this technique, one can assess individual characteristics (in the form of multivariate data) at a cellular level to determine the type of cell, their functionality and their differentiation. At the beginning of flow cytometry, the analysis of such data was performed manually by visually separating regions or gates of interest on a series of sequential bivariate projection of the data, a process known as gating. However, the development of this technology now leads to datasets made of multiple measurements (e.g. up to 18) of millions of individuals cells. A significant amount of work has thus been carried out in recent years to propose automatic statistical methods to overcome the limitations of manual gating (see e.g. [**HKB+10**, **HAG+17**, **LMP16**, **PLW+14**] and references therein).

When analyzing samples in cytometry measured from different patients, a critical issue is data registration across patients. As carefully explained in [**HKB+10**], the alignment of flow cytometry data is a preprocessing step which aims at removing effects coming from technological issues in the acquisition of the data rather than significant biological differences. In this chapter, we use data analyzed in [**HKB+10**] that are obtained from a renal transplant retrospective study conducted by the Immune Tolerance Network (ITN). This dataset is freely available from the `flowStats` package of Bioconductor [**GCB+04**] that can be downloaded from `http://bioconductor.org/packages/release/bioc/html/flowStats.html`. It consists of samples from 15 patients. After an appropriate scaling trough an arcsinh transformation and an initial gating on total lymphocytes to remove artefacts, we focus our analysis on the FSC (forward-scattered light) and SSC (side-scattered light) cell markers which are of interest to measure the volume and morphological complexity of cells. The number of considered cells by patient varies from 88 to 2185. The resulting dataset is displayed in Figure II.3. It clearly shows a mis-alignment issue between measurements from different patients.

The last contribution of the paper is thus to demonstrate the usefulness of regularized Wasserstein barycenters to correct mis-alignment effects in the analysis of data produced by flow cytometers.

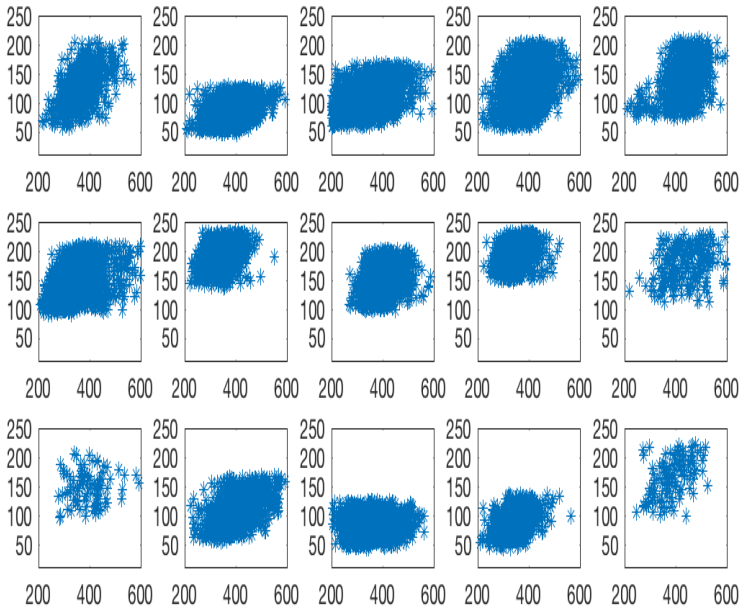


FIGURE II.3. Example of flow cytometry data measured from $n = 15$ patients (restricted to a bivariate projection). The horizontal axis (resp. vertical axis) represent the values of the FSC (resp. SSC) cell marker.

II.2. Penalized Wasserstein barycenters

In this section, we adopt the framework from Section I.1 in which the Wasserstein barycenter is regularized through a convex penalty function as presented in (II.3).

II.2.1. Choice of the function E that penalized the barycenter $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ in (II.3)

The choice of a specific function E is driven by the need to retrieve an absolutely continuous measure from discrete observations (X_{ij}), as it is often done when approximating data through kernel smoothing in density estimation. More precisely, assume that $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ is a distribution which gives mass one to the set of a.c. measures (with respect to the Lebesgue measure dx). Then, thanks to Remark I.14 page 36, there exists a unique population Wasserstein barycenter $\mu^0_{\mathbb{P}}$ which is an *a.c.* measure. We also assume that Ω is a compact and uniformly convex set with a C^1 boundary.

Thus we define the penalizing functional E by

$$E(\mu) = \begin{cases} \|f\|^2_{H^k(\Omega)}, & \text{if } f = \frac{d\mu}{dx} \text{ and } f \geqslant \alpha, \\ +\infty & \text{otherwise.} \end{cases} \tag{II.6}$$

where we recall that $\|\cdot\|_{H^k(\Omega)}$ denotes the Sobolev norm associated to the $\mathbb{L}^2(\Omega)$ space, $\alpha > 0$ is arbitrarily small and $k > d-1$.

For two *a.c.* measures $\mu = \mu_f$ and $\nu = \nu_g$ with density f and g, it is easily seen that the non-symmetric and symmetric Bregman divergences (I.18) related to E satisfy

$$d_E(\mu_f, \nu_g) \geqslant \|f-g\|^2_{\mathbb{L}^2(\Omega)} \quad \text{and} \quad D_E(\mu_f, \nu_g) \geqslant \frac{1}{2}\|f-g\|^2_{\mathbb{L}^2(\Omega)}.$$

Let us now discuss the convergence of the measure $\hat{\boldsymbol{\mu}}^\gamma_{n,p}$ towards $\mu^0_{\mathbb{P}}$ with respect to the squared $\mathbb{L}^2(\Omega)$ distance between their respective densities $\hat{\boldsymbol{f}}^\gamma_{n,p}$ and $f^0_{\mathbb{P}}$, when both n and p tend to infinity and γ tends to 0. To this end, it is necessary to assume that $f^0_{\mathbb{P}} \geqslant \alpha$, and we consider the decomposition

$$\mathbb{E}\left(\|\hat{\boldsymbol{f}}^\gamma_{n,p} - f^0_{\mathbb{P}}\|^2_{\mathbb{L}^2(\Omega)}\right) \leqslant 3\sqrt{\underbrace{\mathbb{E}\left(d^2_E\left(\hat{\boldsymbol{\mu}}^\gamma_{n,p}, \mu^\gamma_{\mathbb{P}^\nu_n}\right)\right)}_{\text{Stability term}}} + 3\sqrt{\underbrace{\mathbb{E}\left(d^2_E\left(\mu^\gamma_{\mathbb{P}^\nu_n}, \mu^\gamma_{\mathbb{P}}\right)\right)}_{\text{Variance term}}} + 6\underbrace{D_E\left(\mu^\gamma_{\mathbb{P}}, \mu^0_{\mathbb{P}}\right)}_{\text{Bias term}}.$$

Then, we gather the results on the stability's Theorem I.12 combined with Theorem 1 in [**FG15**], Theorem I.18 (convergence of the variance term) and Theorem I.16 (convergence of the bias term) to prove the convergence to zero of the three terms in the right-hand side of the above inequality.

Stability term. Recall that by inequality (I.12) one has that

$$\mathbb{E}\left(d^2_E\left(\hat{\boldsymbol{\mu}}^\gamma_{n,p}, \mu^\gamma_{\mathbb{P}^\nu_n}\right)\right) \leq \frac{4}{\gamma^2 n}\sum_{i=1}^n \mathbb{E}\left(W_2^2(\boldsymbol{\nu}_i, \boldsymbol{\nu}_{p_i})\right).$$

For each $1 \leqslant i \leqslant n$ and conditionally on $\boldsymbol{\nu}_i$, the convergence to zero of $\mathbb{E}\left(W_2^2(\boldsymbol{\nu}_i, \boldsymbol{\nu}_{p_i})\right)$ as p_i tends to infinity can be controlled using the results in [**FG15**]. For instance, if the measure $\boldsymbol{\nu}_i$ has a moment of order $q > 4$ then, by Theorem 1 in [**FG15**], it follows that there exists a constant $C_{q,d} > 0$ (depending only on q and d) such that

$$\mathbb{E}\left(W_2^2(\boldsymbol{\nu}_i, \boldsymbol{\nu}_{p_i})\right) \leqslant C_{q,d}\mathbb{E}\left(M_q^{2/q}(\boldsymbol{\nu}_i)\right) p_i^{-1/2}$$

provided that $d < 4$, and where $M_q(\boldsymbol{\nu}_i) = \int_\Omega |x|^q d\boldsymbol{\nu}_i(x)$. Hence, under such assumptions on q and d, it follows that

$$\mathbb{E}\left(d^2_E\left(\hat{\boldsymbol{\mu}}^\gamma_{n,p}, \mu^\gamma_{\mathbb{P}^\nu_n}\right)\right) \leq 4C_{q,d}\mathbb{E}\left(M_q^{2/q}(\boldsymbol{\nu}_1)\right)\frac{1}{\gamma^2 p^{1/2}}. \tag{II.7}$$

Variance term. By Theorem I.18 , one obtains that $\mathbb{E}\left(d^2_E\left(\mu^\gamma_{\mathbb{P}^\nu_n}, \mu^\gamma_{\mathbb{P}}\right)\right) \leq \frac{C}{\gamma^2 n}$.

Bias term. By Theorem I.16, $\lim_{\gamma\to 0} D_E(\mu^\gamma_{\mathbb{P}}, \mu^0_{\mathbb{P}}) = 0$.

Let us finally assume that the distribution $\mathbb{P}$ is such that $\mathbb{E}\left(M_q^{2/q}(\boldsymbol{\nu}_1)\right) < +\infty$. Therefore, under the various assumptions made in this discussion, and by combining the above results, the expected squared $\mathbb{L}^2(\Omega)$ error $\mathbb{E}\left(\|\hat{\boldsymbol{f}}^\gamma_{n,p} - f^0_{\mathbb{P}}\|^2_{\mathbb{L}^2(\Omega)}\right)$ converges to zero provided that $\gamma = \gamma_{n,p}$ is a sequence of regularizing parameters converging to zero such that

$$\lim_{\min(n,p)\to\infty} \gamma^2_{n,p} n = +\infty \qquad \text{and} \qquad \lim_{\min(n,p)\to\infty} \gamma^2_{n,p} p^{1/2} = +\infty.$$

We finally get

THEOREM II.1. *Let $\hat{\boldsymbol{f}}^\gamma_{n,p}$ and $f^\gamma_{\mathbb{P}}$ be the density functions of $\hat{\boldsymbol{\mu}}^\gamma_{n,p}$ and $\mu^\gamma_{\mathbb{P}}$, induced by the choice (II.6) of the penalty function E. Then there exists a constant $c > 0$ such that*

$$\mathbb{E}\left(\|\hat{\boldsymbol{f}}^\gamma_{n,p} - f^\gamma_{\mathbb{P}}\|^2_{\mathbb{L}_2(\Omega)}\right) \leqslant c\left(\frac{1}{\gamma p^{1/4}} + \frac{1}{\gamma n^{1/2}}\right) \tag{II.8}$$

where $p = \min_{1\leqslant i\leqslant n} p_i$ and provided that $d < 4$ and $\mathbb{E}\left(\int_\Omega |x|^q d\boldsymbol{\nu}_1(x)\right) < +\infty$ for some $q > 4$.

II.2.2. Numerical computation

We provide in the last Section II.4 of this chapter efficient minimization algorithms for the computation of $\hat{\boldsymbol{f}}^{\gamma}_{n,p}$, after a binning of the data on a fixed grid Ω_N. For Ω included in the real line, a simple subgradient descent is considered. When data are histograms supported on $\mathbb{R}^d$, $d \geqslant 2$, we rely on a smooth dual approach based on the work of [**CP16b**].

II.3. Numerical experiments

In this section, we first present a method to automatically choose the parameters γ in (II.3) and ε in (II.4), that we illustrate with one-dimensional datasets. Then, we report the results from numerical experiments on simulated Gaussian mixtures and flow cytometry dataset in $\mathbb{R}^2$.

II.3.1. Goldenshluger-Lepski method

By analogy with the work in [**LM16**] based on the Goldenshluger-Lepski (GL) principle, we propose to compute a bias-variance trade-off functional which will provide an automatic selection method for the regularization parameters for either penalized or Sinkhorn barycenters. The method consists in comparing estimators pairwise, for a given range of regularization parameters, with respect to a loss function.

Since the formulation of the GL's principle is similar for both estimators, we only present the theory for the Sinkhorn barycenter described in Section I.2. The trade-off functional is composed of a term measuring the disparity between two estimators and of a penalty term that is chosen according to the upper bounds of the variance in Theorem I.22. More precisely, assume that we dispose of a finite collection of estimators $(\hat{\boldsymbol{r}}^{\varepsilon}_{n,p})_{\varepsilon}$ for ε ranging in a space Λ depending on the data at hand. The GL method consists in choosing a value $\hat{\varepsilon}$ which minimizes the following bias-variance trade-off function:

$$\hat{\varepsilon} = \underset{\varepsilon \in \Lambda}{\arg\min} \;\; B(\varepsilon) + bV(\varepsilon) \tag{II.9}$$

for which we set the "bias term" as

$$B(\varepsilon) = \sup_{\tilde{\varepsilon} \leqslant \varepsilon} \left[|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - \hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p}|^2 - bV(\tilde{\varepsilon}) \right]_+ \tag{II.10}$$

where $x_+ = \max(x, 0)$ denotes the positive part. The authors in [**LM16**] propose a few suggestions to properly choose both the parameter $b > 0$ and the functional V. This leads to a "variance term" V chosen proportional to the right-hand side of (I.24). Following [**LM16**] and from our numerical experiments, we observed that bV has to depend on the size N of the grid Ω_N in order to fit the scaling of the disparity term $|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - \hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p}|^2$.

II.3.2. Simulated data: one-dimensional Gaussian mixtures

We illustrate GL's principle as well as the choice of the parameter b for the one-dimensional example of Gaussian mixtures that is displayed in Figure II.4. Our dataset consists of observations $(X_{i,j})_{1 \leqslant i \leqslant n \; ; 1 \leqslant p}$ sampled from $n = 15$ random distributions $\boldsymbol{\nu}_i$ that are mixtures of two Gaussian distributions with weights $(0.35, 0.65)$, random means respectively belonging to the intervals $[-6, -2]$ and $[2, 6]$ and random variances both belonging to the interval $(0, 2]$. For each random mixture distribution, we sample $p = 50$ observations.

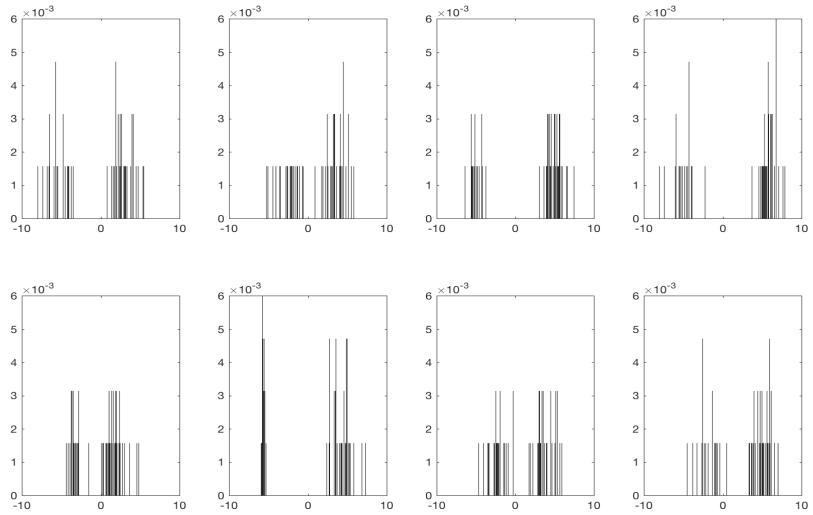


FIGURE II.4. A subset of 8 histograms (out of $n = 15$) obtained with random variables sampled from one-dimensional Gaussian mixtures distributions $\boldsymbol{\nu}_i$ (with random means and variances). Histograms are constructed by binning the data $(X_{i,j})_{1\leqslant i\leqslant n\ ;1\leqslant p}$ on a grid Ω_N of size $N = 2^8$.

Thanks to inequality (I.24), we choose to take the function V defined by

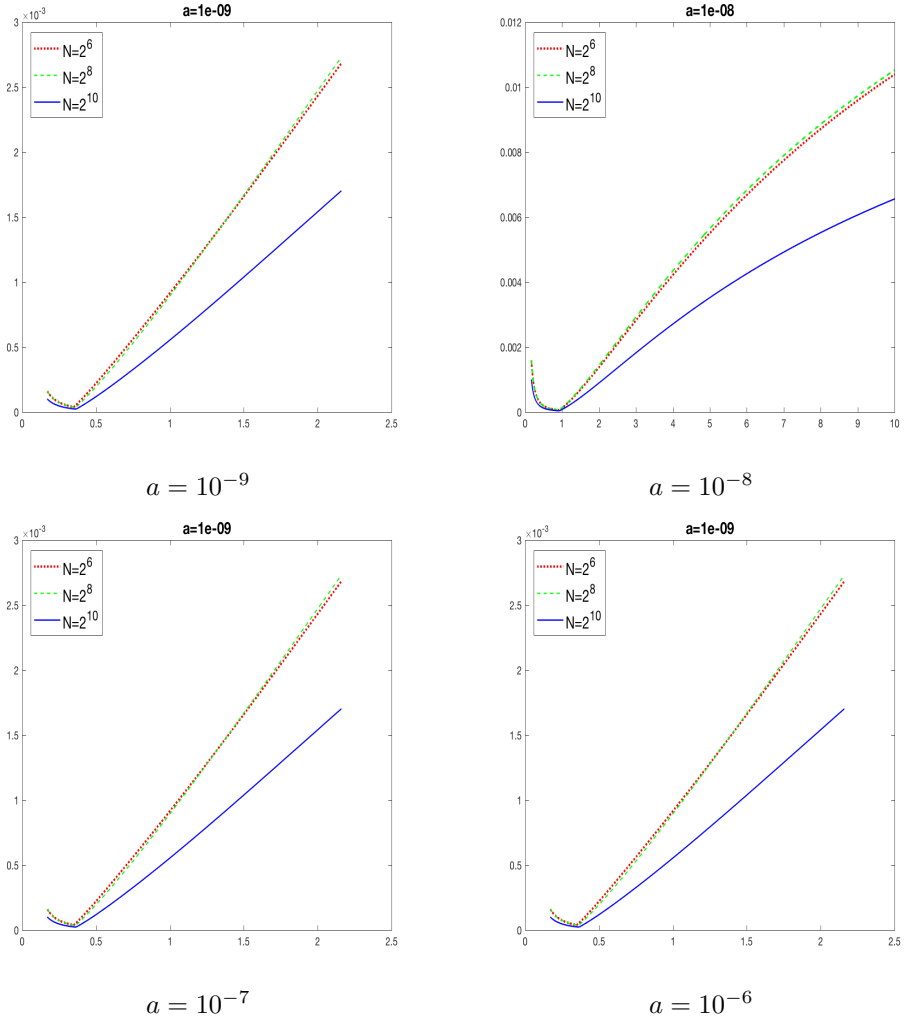
$$V(\varepsilon) = \frac{32L_{\rho,\varepsilon}^2}{\varepsilon^2 n} + \frac{2L_{\rho,\varepsilon}}{\varepsilon}\left(\sqrt{\frac{N}{p}} + 2\rho(N+\sqrt{N})\right),$$

and $\rho = \min(1/N, 1/p)$.

By definition, see equation (I.25), the Lipschitz constant $L_{\rho,\varepsilon}$ is of order $\sqrt{N}$ since the term $\sup_{1\leqslant\ell,k\leqslant N}|C_{m\ell} - C_{k\ell}| = sup_{1\leqslant\ell,k\leqslant N}|\ |x_m - x_\ell|^2 - |x_k - x_\ell|^2\ |$ is not influenced by the size of the grid but rather by the largest distances between points in the support. Hence the variance function V clearly scales polynomially fast with N. To compensate this scaling effect, we choose the parameter $b = a/N^2$ for some constant $a > 0$, as our experiments suggest. Using this choice for b, we obtain data-driven regularization parameters $\hat{\epsilon}$ that are of the same order for different grid size N as it can be seen from Figure II.5, where we display the function $\varepsilon \mapsto B(\varepsilon) + bV(\varepsilon)$ for different values of $a = bN^2$ ranging from 10^{-9} to 10^{-6} and grid sizes $N = 2^6, 2^8, 2^{10}$ (using the same data sampled from random Gaussian mixtures before binning). For a better representation, we normalize the trade-off functions since we are only interested in their minimizer. We also present in Figure II.6 the Sinkhorn barycenters associated to the regularization parameters $\hat{\epsilon}$ that minimize the trade-off functions displayed in Figure II.5. Note that for a better visualization, we have again normalized these barycenters with respect to the grid of size $N = 2^8$. The shapes of these barycenters are very similar despite the change of grid size. Finally, we suggest to choose a such that the trade-off curve has a minimum that is roughly in the center of the parameter's range of interest.

To define the variance function in the case of penalized barycenters $\hat{\boldsymbol{f}}_{n,p}^{\gamma}$ of Section II.2, we use the upper bound in inequality (II.8) leading to the choice

$$V(\gamma) = \frac{1}{\gamma p^{1/4}} + \frac{1}{\gamma n^{1/2}}$$



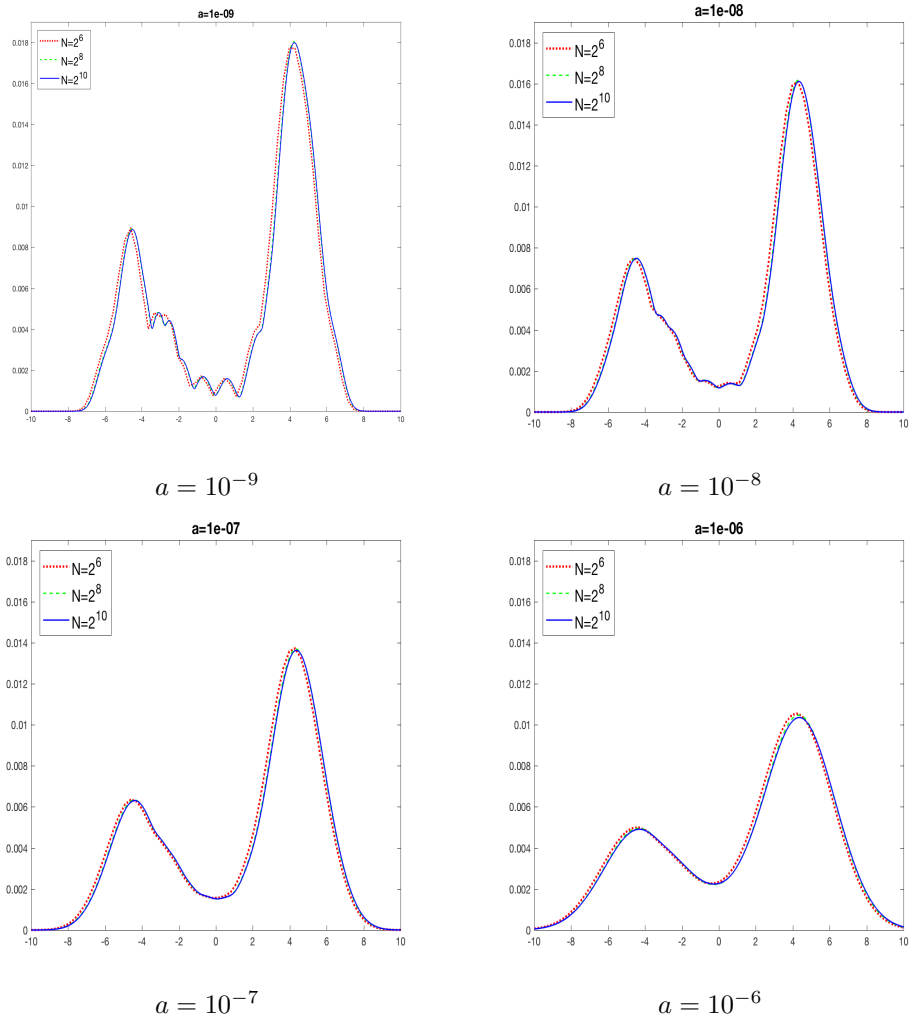
$a = 10^{-9}$ $a = 10^{-8}$

$a = 10^{-7}$ $a = 10^{-6}$

FIGURE II.5. Influence of the parameter $b = a/N^2$ on the shape of the bias-variance trade-off function $\varepsilon \mapsto B(\varepsilon) + bV(\varepsilon)$ for Sinkhorn barycenters of the dataset in Figure II.4. The range of ϵ is the interval $[0.1, 10]$. A zoom is performed for the two figures (c) and (d). From left to right and top to bottom, $a = 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}$. The dotted red curve corresponds to a grid size $N = 2^6$, the dashed green curve to $N = 2^8$ and the solid blue curve to $N = 2^{10}$.

We remark that the size of the grid does not appear in the above variance function. Thus, the parameter b is chosen independent of N in the trade-off function $\gamma \mapsto B(\gamma) + bV(\gamma)$.

From now on, we fix the size $N = 2^8$ of the grid, and we comment the choice of the parameters $\hat{\varepsilon}$ and $\hat{\gamma}$. We display in Figure II.7(a) the trade-off function $B(\varepsilon)+bV(\varepsilon)$, and we discuss the influence of ε on the smoothness and support of the Sinkhorn barycenter. From Figure II.7(b), we observe that, when the parameter $\varepsilon = 0.18$ is small (dotted blue curve), then the corresponding Sinkhorn barycenter $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ is irregular, and it presents spurious peaks. On the contrary, too much regularization, e.g. $\varepsilon = 9.5$, implies that the barycenter (dashed green curve) is flattened and its mass is spread out. Here, the optimal barycenter (solid red



$a = 10^{-9}$ $a = 10^{-8}$

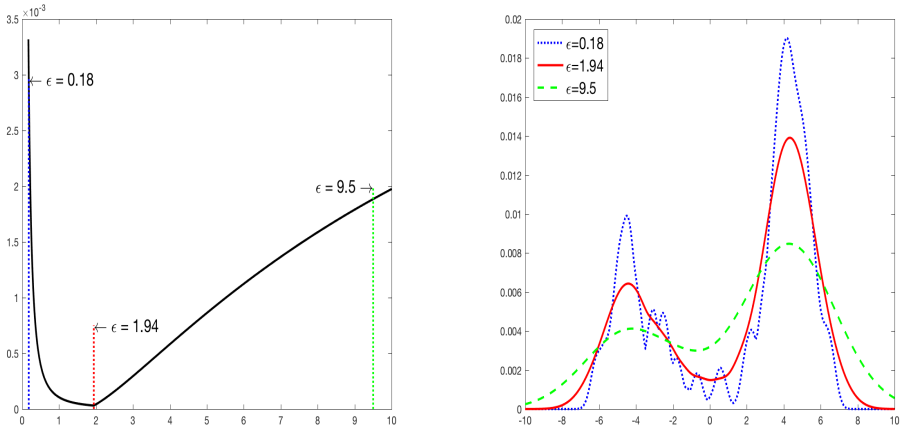
$a = 10^{-7}$ $a = 10^{-6}$

FIGURE II.6. Sinkhorn barycenters of the dataset in Figure II.4 associated to the regularization parameters $\hat{\varepsilon}$ minimizing the trade-off functions displayed in Figure II.5 for different values of a and different grid sizes N.

curve), that is $\hat{\boldsymbol{r}}_{n,p}^{\hat{\varepsilon}}$ for $\hat{\varepsilon} = 1.94$ minimizing the trade-off function (II.9), seems to be a good compromise between under and over-smoothing.

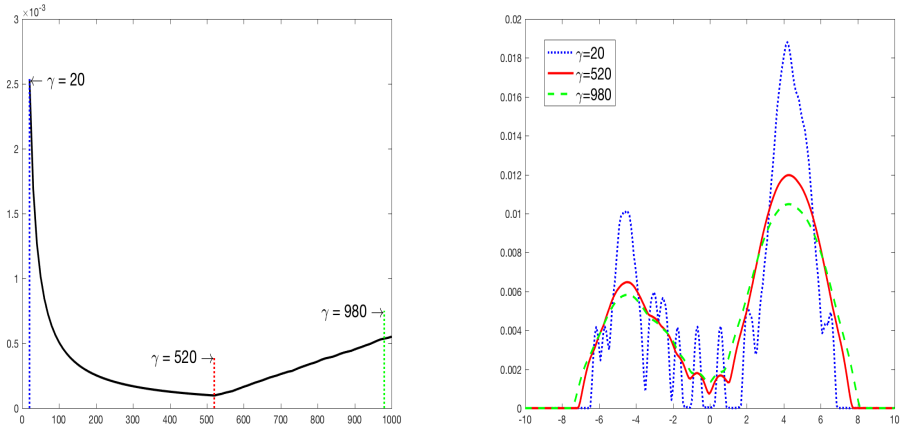
We repeat the same experiment for the penalized barycenter of Section II.2 with the Sobolev norm $H^1(\Omega)$ in the penalization function E (II.6) . The results are displayed in Figure II.8. The advantage of choosing a Sobolev penalty function is that the mass of the barycenter is overall less spread out and the spikes are sharper. However, for a small regularization parameter $\gamma = 20$ (dotted blue curve), the barycenter $\boldsymbol{f}_{n,p}^{\gamma}$ presents a lot of irregularities as the penalty function tries to minimize its $\mathbb{L}_2$-norm. When the regularization parameter increases in a significant way ($\gamma = 980$ associated to the dashed green curve), the irregularities disappear and the support of the penalized barycenter becomes wider. The GL's principle leads to the choice $\hat{\gamma} = 520$ which corresponds to a penalized barycenter (solid red curve) that is satisfactory.

We compare these Wasserstein barycenters to the Euclidean mean $\bar{f}_{n,p}$ (II.1), obtained after a pre-smoothing step of the data for each subject using the kernel method. From



(a) (b)

FIGURE II.7. One dimensional Gaussian mixtures dataset and Sinkhorn barycenters. (a) The trade-off function $\varepsilon \mapsto B(\varepsilon) + bV(\varepsilon)$ which attains its optimum at $\hat{\varepsilon} = 1.94$. (b) Three Sinkhorn barycenters $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ associated to the parameters $\varepsilon = 0.18, 1.94, 9.5$.



(a) (b)

FIGURE II.8. One dimensional Gaussian mixtures dataset and penalized barycenters. (a) The trade-off function $\gamma \mapsto B(\gamma) + bV(\gamma)$ which attains its optimum at $\hat{\gamma} = 520$. (b) Three penalized barycenters $\boldsymbol{f}^{\gamma}_{n,p}$ associated to the parameters $\gamma = 20, 520, 980$.

Figure II.9, the density $\bar{f}_{n,p}$ is very irregular and it suffers from mis-alignment issues. The irregularity of this estimator mainly comes from the low-dimensional sampling per subject ($p = 50$).

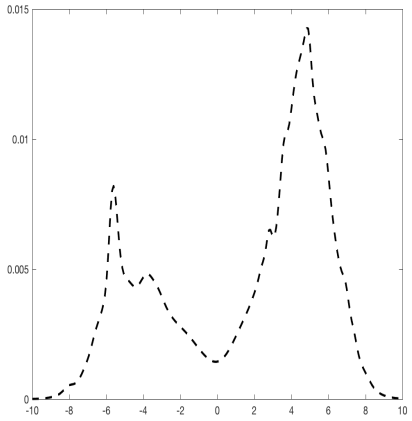


FIGURE II.9. Euclidean mean density $\bar{f}_{n,p}$ of the one dimensional Gaussian mixtures dataset using a preliminary smoothing step of each subject with a Gaussian kernel.

II.3.3. Simulated data: two-dimensional Gaussian mixtures

In this section, we illustrate the validity of our methods for two-dimensional data. We consider a simulated example of observations $(X_{i,j})_{1\leqslant i\leqslant n\ ;1\leqslant j\leqslant p}$ sampled from $n = 15$ random distributions $\boldsymbol{\nu}_i$ that are a mixture of three multivariate Gaussian distributions $\boldsymbol{\nu}_i = \sum_{j=1}^{3} \theta_j \mathcal{N}(\boldsymbol{m}_j^i, \boldsymbol{\Gamma}_j^i)$ with fixed weights $\theta = (1/6, 1/3, 1/2)$. The means $\boldsymbol{m}_j^i$ and covariance matrices $\boldsymbol{\Gamma}_j^i$ are random variables with expectation given by (for $j = 1, 2, 3$)

$$m_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad m_2 = \begin{pmatrix} 7 \\ 4 \end{pmatrix}, \quad m_3 = \begin{pmatrix} 1 \\ 9 \end{pmatrix}, \quad \text{and} \quad \Gamma_1 = \Gamma_2 = \Gamma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

For each $i = 1, \ldots, n$, we simulate a sequence $(X_{i,j})_{1\leqslant j\leqslant p}$ of $p = 50$ iid random variables sampled from $\boldsymbol{\nu}_i = \sum_{j=1}^{3} \theta_j \mathcal{N}(\boldsymbol{m}_j^i, \boldsymbol{\Gamma}_j^i)$ where $\boldsymbol{m}_j^i$ (resp. $\boldsymbol{\Gamma}_j^i$) are random vectors (resp. matrices) such that each of their coordinate follows a uniform law centered in m_j with amplitude ± 2 (resp. each of their diagonal elements follows a uniform law centered in Γ_j with amplitude ± 0.95). We display in Figure II.10 the dataset $(X_{i,j})_{1\leqslant j\leqslant p\ ;1\leqslant i\leqslant n}$. Each $X_{i,j}$ is then binned on a grid of size 64×64 (thus $N = 4096$).

First, we compute 60 Sinkhorn barycenters by letting ε ranging from 0.1 to 6. We draw in Figure II.11(a) the trade-off function, and we plot a zoom of this function in Figure II.11(b) around the minimizer $\hat{\varepsilon} = 3.6$. The corresponding Sinkhorn barycenter $\hat{\boldsymbol{r}}_{n,p}^{\hat{\varepsilon}}$ is displayed in Figure II.12(a). We also present the Euclidean mean $\bar{f}_{n,p}$ (after a preliminary smoothing step) in Figure II.12(b). The Sinkhorn barycenter has three distinct modes. Hence, this approach handles in a very efficient way the scaling and translation variations in the dataset (which corresponds to the correction of the mis-alignment issue). On the other hand, the Euclidean mean mixes the distinct modes of the Gaussian mixtures. It is thus less robust to outliers since the support of the barycenter is significantly spread out.

Finally, we display the penalized barycenter of this dataset in Figure II.13(b) for the Sobolev penalty function (II.6) and the data-driven choice of γ. In Figure II.13(a), we plot the trade-off function for γ ranging from 1 to 140. This curve suggests to choose $\hat{\gamma} = 80$. From Figure II.13(b), we observe that the mass of $\hat{\boldsymbol{f}}_{n,p}^{\hat{\gamma}}$ is concentrated on three main modes. The approach thus manages to keep the geometry of the underlying Gaussian mixtures.

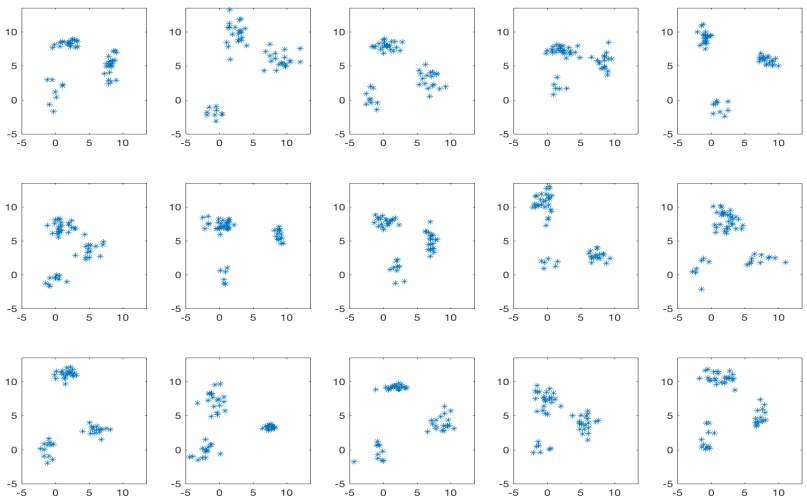
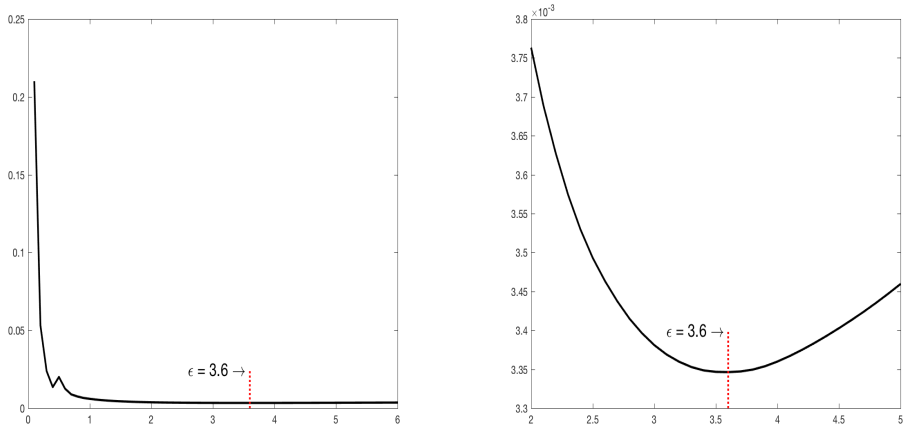


FIGURE II.10. Dataset $(X_{i,j})_{1\leqslant j\leqslant p\ ;1\leqslant i\leqslant n}$ generated from $n = 15$ two-dimensional random Gaussian mixtures $\boldsymbol{\nu}_i$ with $p = 50$.

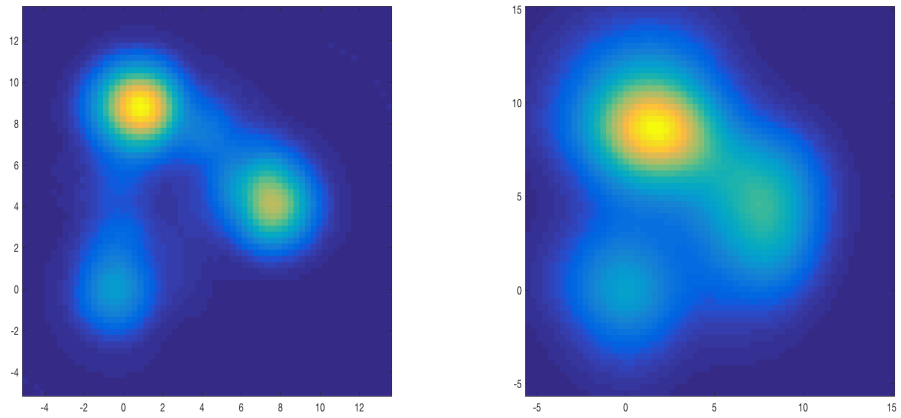


(a) (b)

FIGURE II.11. Two-dimensional Gaussian mixtures dataset. (a) The trade-off function $\varepsilon \mapsto B(\varepsilon) + bV(\varepsilon)$ which attains its optimum at $\hat{\varepsilon} = 3.6$. (b) A zoom of the trade-off function around the minimizer $\hat{\varepsilon}$.

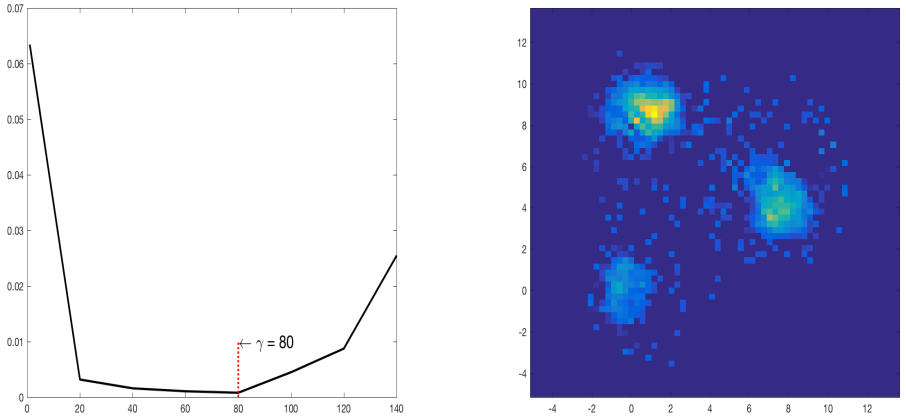
II.3.4. Sinkhorn versus penalized barycenters

To conclude these numerical experiments with simulated data, we would like to point out that computing the Sinkhorn barycenter is much faster than computing the penalized barycenter. Indeed, entropy regularization of the transport plan in the Wasserstein distance has been first introduced in order to reduce the computational cost of a transport distance. Its computation requires $\mathcal{O}(N^3 \log N)$ operations for discrete probability measures with a



(a) (b)

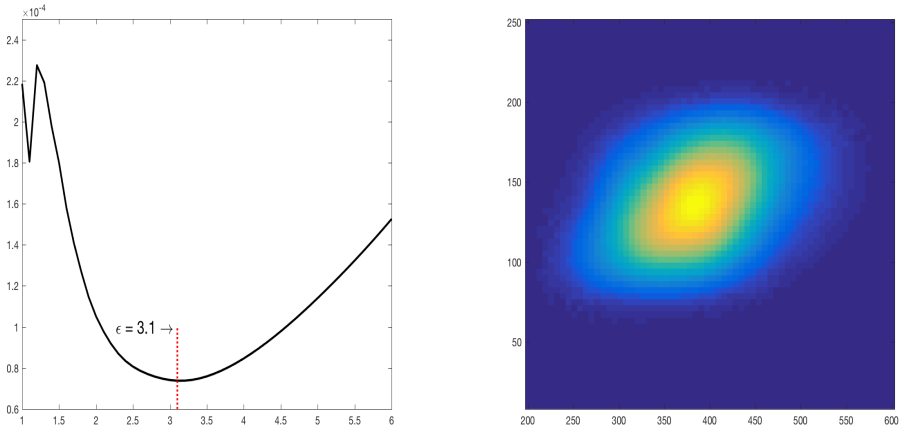
FIGURE II.12. Two-dimensional Gaussian mixtures dataset. (a) The Sinkhorn barycenter $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ for $\hat{\varepsilon} = 3.6$ chosen by the GL's principle. (b) The Euclidean mean $\bar{f}_{n,p}$ (after a preliminary smoothing step).



(a) (b)

FIGURE II.13. Two-dimensional Gaussian mixtures dataset. (a) The trade-off function $\gamma \mapsto B(\gamma) + bV(\gamma)$ which attains its optimum at $\hat{\gamma} = 80$. (b) Penalized barycenter $\hat{\boldsymbol{f}}^{\hat{\gamma}}_{n,p}$ associated to the parameter $\hat{\gamma} = 80$.

support of size N when the computation of a Sinkhorn divergence only takes $\mathcal{O}(N^2)$ operations at each iteration of a gradient descent (see e.g. [Cut13]). We have also found that the Sinkhorn barycenter yields more satisfying estimators in terms of smoothness. Therefore, in the rest of this section, we do not consider the penalized barycenter anymore.



(a) (b)

FIGURE II.14. Two dimensional flow cytometry dataset and Sinkhorn barycenter. (a) The trade-off function $\varepsilon \mapsto B(\varepsilon) + bV(\varepsilon)$ which attains its optimum at $\hat{\varepsilon} = 3.1$. (b) Sinkhorn barycenter $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ associated to the parameter $\hat{\varepsilon} = 3.1$.

II.3.5. Real data: flow cytometry

We have at our disposal data from flow cytometry that have been described in Section II.1.2.3, and we focus on the FSC and SSC cell markers resulting in the dataset that is displayed in Figure II.3. We again apply a binning of the data on a two-dimensional grid of size $N = 64 \times 64$. In Figure II.14(a) we plot the trade-off function related to the Sinkhorn barycenters for the parameter ε ranging from 1 to 6. Its minimum is attained for $\hat{\varepsilon} = 3.1$. We display the corresponding Sinkhorn barycenter in Figure II.14(b). This barycenter clearly allows to correct mis-alignment issues of the data.

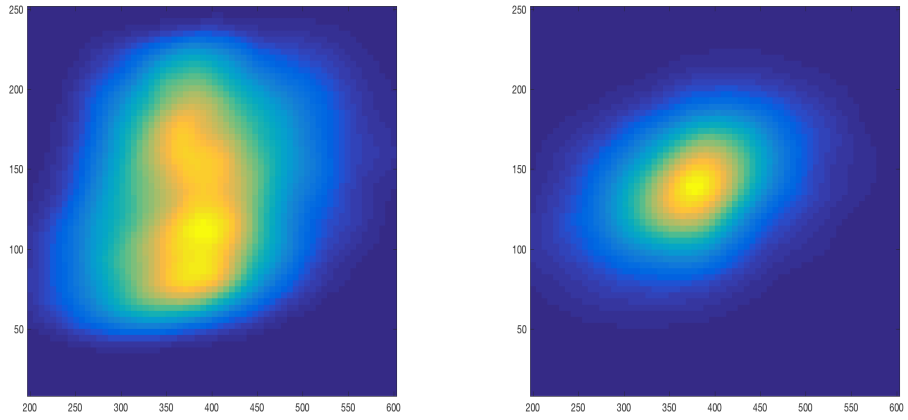
Notice that we have also conducted experiments for Sinkhorn barycenters with non-equal weights, corresponding to the proportion of measurements for each patient. The result being analogous, we do not report them.

II.4. Algorithms to compute penalized Wasserstein barycenters presented in Section I.1 and Section II.2

In this section we describe how the minimization problem

$$\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i) + \gamma E(\mu) \text{ over } \mu \in \mathcal{P}_2(\Omega), \tag{II.11}$$

can be solved numerically by using an appropriate discretization and the work of [**CP16b**]. In our numerical experiments, we focus on the case where $E(\mu) = +\infty$ if μ is not a.c. to enforce the regularized Wasserstein barycenter to have a smooth pdf (we write $E(f) = E(\mu_f)$ if μ has a density f). In this setting, if the grid of points is of sufficiently large size, then the weights f^k yield a good approximation of this pdf. A discretization of the minimization problem (II.11) is used to compute a numerical approximation of a regularized Wasserstein



(a) (b)

FIGURE II.15. Two dimensional flow cytometry dataset. (a) Euclidean mean $\bar{f}_{n,p}$ of the data (after smoothing but without registration), (b) $\mathbb{L}_2$-mean of pre-processed data using kernel smoothing and density registration by landmark alignment with the method in [**HKB+10**].

barycenter $\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}$. A fixed grid $\{x^k\}_{k=1}^N$ of equally spaced points $x^k \in \mathbb{R}^d$ is considered and $\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}$ is approximated by the discrete measure $\sum_{k=1}^N f^k \delta_{x^k}$ where the f^k are positive weights summing to one.

In what follows, we first describe an algorithm that is specific to the one-dimensional case, and then we propose another algorithm that is valid for any $d \geqslant 1$.

Discrete algorithm for $d = 1$ and data defined on the same grid. We first propose to compute a regularized empirical Wasserstein barycenter for a dataset made of discrete measures $\nu_1, \ldots, \nu_n$ (or one-dimensional histograms) defined on the same grid of reals $\{x^k\}_{k=1}^N$ that the one chosen to approximate $\boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}$. Since the grid is fixed, we identify a discrete measure ν with the vector of weights $\nu = (\nu(x^1), \ldots, \nu(x^N))$ in $\mathbb{R}^N_+$ (with entries that sum up to one) of its values on this grid.

The estimation of the regularized barycenter onto this grid can be formulated as:

$$\min_f \frac{1}{n} \sum_{i=1}^n W_2^2(f, \nu_i) + \gamma E(f) \text{ s.t } \sum_k f^k = 1, \text{ and } f^k = f(x^k) \geqslant 0, \tag{II.12}$$

with the obvious abuse of notation $W_2^2(f, \nu_i) = W_2^2(\mu_f, \nu_i)$ and $E(f) = E(\mu_f)$.

Then, to compute a minimizer of the convex optimization problem (II.12), we perform a subgradient descent. We denote by $(f^{(\ell)})_{\ell \geqslant 1}$ the resulting sequence of discretized regularized barycenters in $\mathbb{R}^N$ along the descent. Hence, given an initial value $f^{(1)} \in \mathbb{R}^N_+$ and for $\ell \geq 1$, we thus have

$$f^{(\ell+1)} = \Pi_S \left(f^{(\ell)} - \tau^{(\ell)} \left[\gamma \nabla E(f^{(\ell)}) + \frac{1}{n} \sum_{i=1}^n \nabla_1 W_2^2(f^{(\ell)}, \nu_i) \right] \right) \tag{II.13}$$

where $\tau^{(\ell)}$ is the ℓ-th step time, and Π_S stands for the projection on the simplex $S = \{y \in \mathbb{R}^N_+ \text{ such that } \sum_{j=1}^N y^j = 1\}$. Thanks to Proposition 5 in [**PFR12**], we are able to compute a sub-gradient of the squared Wasserstein distance $W_2^2(f^{(\ell)}, \nu_i)$ with respect to its first argument (for discrete distributions). For that purpose, we denote by $R_f(s) = \sum_{x^j \leq s} f(x^j)$ the cdf of $\mu_f = \sum_{k=1}^N f(x^k)\delta_{x^k}$ and by $R_f^-(t) = \inf\{s \in \mathbb{R} : R_f(s) \geq t\}$ its pseudo-inverse.

PROPOSITION II.2 ([PFR12]). *Let $f = (f(x^1), f(x^2), \dots, f(x^N))$ and $\nu = (\nu(x^1), \nu(x^2), \dots, \nu(x^N))$ be two discrete distributions defined on the same grid of values $x^1, \dots, x^N$ in $\mathbb{R}$. For $p \geq 1$, the subgradients of $f \mapsto W_p^p(f, \nu)$ can be written as*

$$\nabla_1 W_p^p(f, \nu) : x_j \mapsto \sum_{m \geq j} |x^m - \tilde{x}^m|^p - |x^{m+1} - \tilde{x}^m|^p, \tag{II.14}$$

where

$$\begin{cases} \tilde{x}^m = x^k \text{ if } R_g(x^{k-1}) < R_f(x^m) < R_\nu(x^k) \\ \tilde{x}^m \in [x^{k-1}, x^k] \text{ if } R_f(x^m) = R_\nu(x^k). \end{cases}$$

Even if subgradient descent is only shown to converge with diminishing time steps [BM06], we observed that using a small fixed step time (of order 10^{-5}) is sufficient to obtain in practice a convergence of the iterates $(f^{(\ell)})_{\ell \geqslant 1}$. Moreover, we have noticed that the principles of FISTA (Fast Iterative Soft Thresholding, see e.g. [BT09]) accelerate the speed of convergence of the above described algorithm.

Discrete algorithm for $d \geqslant 1$ in the general case.

We assume that data $\nu_1, \dots, \nu_n$ are given in the form of n discrete probability measures (histograms) supported on $\mathbb{R}^d$ (with $d \geqslant 1$) that are not necessarily defined on the same grid. More precisely, we assume that

$$\nu_i = \sum_{j=1}^{p_i} \nu_i^j \delta_{y_i^j},$$

for $1 \leqslant i \leqslant n$ where the y_i^j's are arbitrary locations in $\Omega \subset \mathbb{R}^d$, and the ν_i^j's are positive weights (summing up to one for each i).

The estimation of the regularized barycenter onto a given grid $\{x^k\}_{k=1}^N$ of $\mathbb{R}^d$ can then be formulated as the following minimization problem:

$$\min_f \frac{1}{n} \sum_{i=1}^n W_2^2(f, \nu_i) + \gamma E(f) \text{ s.t } \sum_k f^k = 1, \text{ and } f^k \geqslant 0, \tag{II.15}$$

with the notation $f = (f^1, f^2, \dots, f^N)$ and the convention that $W_2^2(f, \nu_i)$ denotes the squared Wasserstein distance between $\mu_f = \sum_{k=1}^N f^k \delta_{x^k}$ and ν_i.

Problem (II.15) could be exactly solved by considering the discrete $p_i \times N$ transport matrices S_i between the barycenter μ_f to estimate and the data ν_i. Indeed, problem (II.15) is equivalent to the convex problem

$$\min_f \min_{S_1 \cdots S_n} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{p_i} \sum_{k=1}^N ||y_i^j - x^k||^2 S_i^{j,k} + \gamma E(f) \tag{II.16}$$

under the linear constraints

$$\forall i = 1, \dots, n, \sum_{j=1}^{p_i} S_i^{j,k} = f^k, \sum_{k=1}^N S_i^{j,k} = \nu_i^j, \text{ and } S_i^{j,k} \geqslant 0.$$

However, optimizing over the $p_i \times N$ transport matrices S_i for $1 \leqslant i \leqslant n$ involves memory issues when using an accurate discretization grid $\{x^k\}_{k=1}^N$ with a large value of N. For this reason, we consider subgradient descent algorithms that allow dealing directly with problem (II.15).

To this end, we rely on the dual approach introduced in [COO15] and the numerical optimisation scheme proposed in [CP16b]. Following these works, one can show that the dual problem of (II.15) with a regularization of the form $E(Kf)$ and K a discrete linear operator reads as

$$\min_{\phi_0, \cdots \phi_n} \sum_{i=1}^n H_{\nu_i}(\phi_i) + E_\gamma^*(\phi_0) \text{ s.t } K^T \phi_0 + \sum_{i=1}^n \phi_i = 0, \tag{II.17}$$

where the ϕ_i's are dual variables (vectors in $\mathbb{R}^N$) defined on the discrete grid $\{x^k\}_{k=1}^N$, E_γ^* is the Legendre transform of γE and $H_{\nu_i}(.)$ is the Legendre transform of $W_2^2(.,\nu_i)$ that reads:

$$H_{\nu_i}(\phi_i) = \sum_{j=1}^{p_i} \nu_i^j \min_{k=1\cdots N} \left(\frac{1}{2}||y_i^j - x^k||^2 - \phi_i^k \right).$$

Barycenter estimations f_i can finally be recovered from the optimal dual variables ϕ_i solution of (II.17) as:

$$f_i \in \partial H_{\nu_i}(\phi_i), \text{ for } i = 1\cdots n. \tag{II.18}$$

Following [**COO15**], one value of the above subgradient can be obtained at point x^k as:

$$\partial H_{\nu_i}(\phi_i)_k = \sum_{j=1}^{p_i} \nu_i^j S_i^{j,k}, \tag{II.19}$$

where $S_i^{j,k}$ is any row stochastic matrix of size $p_i \times N$ checking:

$$S_i^{j,k} \neq 0 \text{ iff } k \in \arg\min_{k=1\cdots N} \left(\frac{1}{2}||y_i^j - x^k||^2 - \phi_i^k \right).$$

From the previous expressions, we see that $f_i^k = \sum_{j=1}^{p_i} \nu_i^j S_i^{j,k}$ corresponds to the discrete pushforward of data ν_i with the transport matrix S_i with the associated cost:

$$H_{\nu_i}(\phi_i) = \sum_{j=1}^{p_i} \sum_{k=1}^{N} \left(\frac{1}{2}||y_i^j - x^k||^2 - \phi_i^k \right) S_i^{j,k} \nu_i^j.$$

Numerical optimization. Following [**CP16b**], the dual problem (II.17), can be simplified by removing one variable and thus discarding the linear constraint $K^T\phi_0 + \sum_{i=1}^n \phi_i = 0$. In order to inject the regularity given by ϕ_0 in all the reconstructed barycenters obtained by ϕ_i, $i = 1\cdots n$, we modified the change of variables of [**CP16b**] by setting $\psi_i = \phi_i + K^T\phi_0/n$ for $i = 1\cdots n$ and $\psi_0 = \phi_0$, leading to $\sum_{i=1}^n \psi_i = 0$. One variable, say ψ_n, can then be directly obtained from the other ones. Observing that $\phi_n = -K^T\psi_0 - \sum_{i=1}^{n-1} \psi_i/n$, we thus obtain:

$$\min_{\psi_0,\cdots\psi_{n-1}} \sum_{i=1}^{n-1} H_{\nu_i}(\psi_i - K^T\psi_0/n) + H_{\nu_n}(-K^T\psi_0 - \sum_{i=1}^{n-1} \psi_i/n) + E_\gamma^*(\psi_0). \tag{II.20}$$

The subgradient (II.19) can then be used in a descent algorithm over the dual problem (II.20). For differentiable penalizers E, we consider the L-BFGS algorithm [**ZBLN97**, **Bec11**] that integrates a line search method (see e.g. [**BV04**]) to select the best time step $\tau^{(\ell)}$ at each iteration ℓ of the subgradient descent:

$$\begin{cases} \psi_0^{(\ell+1)} &= \psi_0^{(\ell)} - \tau^{(\ell)}(\nabla E_\gamma^*(\psi_0^{(\ell)}) + d_0^\ell) \\ \psi_i^{(\ell+1)} &= \psi_i^{(\ell)} - \tau^{(\ell)} d_i^\ell \qquad\qquad i = 1\cdots n-1, \end{cases} \tag{II.21}$$

where:

$$\begin{aligned} d_0^\ell &= K\left(\partial H_{\nu_n}\left(-K^T\psi_0^{(\ell)}/n - \textstyle\sum_{i=1}^{n-1}\psi_i^{(\ell)}\right) - \textstyle\sum_{i=1}^{n-1} \partial H_{\nu_i}\left(\psi_i^{(\ell)} - K^T\psi_0^{(\ell)}/n\right)\right) \\ d_i^\ell &= \partial H_{\nu_i}\left(\psi_i^{(\ell)} - K^T\psi_0^{(\ell)}/n\right) - \partial H_{\nu_n}\left(-K^T\psi_0^{(\ell)}/n - \textstyle\sum_{i=1}^{n-1}\psi_i^{(\ell)}\right). \end{aligned}$$

The barycenter is finally given by (II.18), taking $\phi_i = \psi_i - K^T\psi_0/n$. Even if we only treated differentiable functions E in the theoretical part of this paper, we can numerically consider non differentiable penalizers E, such as Total Variation ($K = \nabla$, $E = |.|_1$). In this case, we make use of the Forward-Backward algorithm. This just modifies the update of ψ_0 in (II.21), by changing the explicit scheme involving ∇E_γ^* onto an implicit one through the proximity operator of E_γ^*:

$$\psi_0^{(\ell+1)} = \text{Prox}_{\tau^{(\ell)} E_\gamma^*}\left(\psi_0^{(\ell)} - \tau^{(\ell)} d_0^\ell\right) = \arg\min_\psi \frac{1}{2\tau^{(\ell)}}||\psi_0^{(\ell)} - \tau^{(\ell)} d_0^\ell - \psi||^2 + E_\gamma^*(\psi).$$

Algorithmic issues and stabilization. As detailed in [COO15], the computation of one subgradient in (II.19) relies on the look for Euclidean nearest neighbors between vectors $(y_i^j, 0)$ and $(x^k, \sqrt{c - \phi_i^k})$, with $c = \max_k \phi_i^k$. Selecting only one nearest neighbor leads to bad numerical results in practice as subgradient descent may not be stable. For this reason, we considered the $K = 10$ nearest neighbors for each j to build the row stochastic matrices S_i at each iteration as: $S_i^{j,k} = w_i^{jk} / \sum_{k'} w_i^{jk'}$, with $w_i^{jk} = \exp(-(\frac{1}{2}\|y_i^j - x^k\|^2 - \phi_i^k)/\varepsilon)$ if k is within the K nearest neighbors for j and data i and $w_i^{jk} = 0$ otherwise.

Chapter III

Central limit theorem for entropy regularized optimal transport on finite spaces

This chapter corresponds to the preprint [**BCP17**].

III.1. Introduction

We discuss in Section III.2 the notion of directional derivative of the Sinkhorn divergences in order to obtain our main result on a central limit theorem, for data sampled from one or two unknown probability distributions, via an appropriate adaptation of the delta-method. We also propose a bootstrap procedure in Section III.3 in order to obtain new test statistics for measuring the discrepancies between multivariate probability distributions. The proof uses the notions of directional Hadamard differentiability and delta-method. It is inspired by the results in the work of Sommerfeld and Munk in [**SM16**] on the asymptotic distribution of empirical Wasserstein distance on finite space using un-regularized transportation costs. Numerical experiments are respectively presented in Section III.4 and Section III.5 for synthetic data and real data. We also illustrate the benefits of a bootstrap procedure. A comparison with existing methods to measure the discrepancy between multivariate distributions is finally proposed.

III.2. Distribution limits for empirical Sinkhorn divergences

In this chapter we consider probability measures distributed on the finite space $\Omega_N = \{x_1, \dots, x_N\}$.

There exists an explicit relation between the optimal solutions of the primal (A.9) and dual (A.10) problems of entropy regularized optimal transport in page 21, and they can be computed through an iterative method called Sinkhorn's algorithm [**CD14**].

Proposition III.1 (Sinkhorn's algorithm). *Let $K = \exp(-C/\varepsilon)$ be the element wise exponential of the matrix cost C divided by $-\varepsilon$. Then, there exists a pair a vectors $(u, v) \in \mathbb{R}_+^N \times \mathbb{R}_+^N$ such that the optimal solutions T_λ^* and $(\alpha_\varepsilon^*, \beta_\varepsilon^*)$ of problems (A.9) and (A.10) are*

respectively given by

$$T_\varepsilon^* = \mathrm{diag}(u)K\,\mathrm{diag}(v),\ \textit{and}\ \alpha_\varepsilon^* = -\lambda\log(u),\ \beta_\varepsilon^* = -\lambda\log(v).$$

Moreover, such a pair (u, v) is unique up to scalar multiplication, and it can be recovered as a fixed point of the Sinkhorn map

$$S_{\{a,b\}} : (u, v) \in \mathbb{R}^N \times \mathbb{R}^N \mapsto (a/(Kv), b/(K^T u)). \tag{III.1}$$

where K^T is the transpose of K and $/$ stands for the component-wise division.

III.2.1. Directional derivative of $W_{2,\varepsilon}^2$

We recall that, if it exists, the Hadamard directional derivative of a function $g : D_g \subset \mathbb{R}^d$ at $z \in D_g$ in the direction h is defined as

$$g'_h(z) = \lim_{n\to\infty} \frac{g(z + t_n h_n) - g(z)}{t_n}$$

for any sequences $(t_n)_n$ such that $t_n \searrow 0$ and $h_n \to h$ with $z + t_n h_n \in D_g$ for all n. As recalled in [SM16], the derivative $h \mapsto g'_h(z)$ is not necessarily a linear map contrary to the usual notion of Hadamard differentiability. A typical example being the function $g(z) = |z|$ (with $D_g = \mathbb{R}$) which is not Hadamard differentiable at $z = 0$ in the usual sense, but directionally differentiable with $g'_h(0) = |h|$.

Theorem III.2. *The functional $(a, b) \mapsto W_{2,\varepsilon}^2(a, b)$ is directionally Hadamard differentiable at all $(a, b) \in \Sigma_N \times \Sigma_N$ in the direction $(h_1, h_2) \in \Sigma_N \times \Sigma_N$, with derivative*

$$(W_{2,\varepsilon}^2)'_{h_1,h_2}(a, b) = \max_{(\alpha,\beta)\in N_\varepsilon(a,b)} \langle\alpha, h_1\rangle + \langle\beta, h_2\rangle$$

where $N_\varepsilon(a, b) \subset \mathbb{R}^N \times \mathbb{R}^N$ is the set of optimal solutions of the dual problem (A.10).

Proof. From Proposition 1 in [CD14], a subgradient of the convex function $(a, b) \mapsto W_{2,\varepsilon}^2(a, b)$ is any optimal solution (α, β) of the dual problem (A.10). From Theorem 11 in [Roc74], we directly get the statement of the theorem. □

III.2.2. Central limit theorem

We denote by $\xrightarrow{\mathcal{L}}$ the convergence in distribution of a random variable and $\xrightarrow{\mathbb{P}}$ the convergence in probability. We also recall that notation $G \overset{\mathcal{L}}{\sim} a$ means that G is a random variable taking its values in $\mathcal{X}$ with law $a = (a_1, \ldots, a_n) \in \Sigma_N$ (namely that $\mathbb{P}(G = x_i) = a_i$ for each $1 \leqslant i \leqslant N$). Let $a, b \in \Sigma_N$. We denote by $\hat{a}_n$ and $\hat{b}_m$ the empirical measures respectively generated by iid samples $X_1, \ldots, X_n \overset{\mathcal{L}}{\sim} a$ and $Y_1, \ldots, Y_m \overset{\mathcal{L}}{\sim} b$:

$$\hat{a}_n = (\hat{a}_n^x)_{x\in\mathcal{X}},\ \text{where}\ \hat{a}_n^{x_i} = \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{X_j = x_i\}} = \frac{1}{n}\#\{j : X_j = x_i\}\ \text{for all}\ 1 \leqslant i \leqslant N.$$

We also define the multinomial covariance matrix

$$\Sigma(a) = \begin{bmatrix} a_{x_1}(1 - a_{x_1}) & -a_{x_1}a_{x_2} & \cdots & -a_{x_1}a_{x_N} \\ -a_{x_2}a_{x_1} & a_{x_2}(1 - a_{x_2}) & \cdots & -a_{x_2}a_{x_N} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{x_N}a_{x_1} & -a_{x_N}a_{x_2} & \cdots & a_{x_N}(1 - a_{x_N}) \end{bmatrix}$$

and the independent Gaussian random vectors $G \sim \mathcal{N}(0, \Sigma(a))$ and $H \sim \mathcal{N}(0, \Sigma(b))$. As classically done in statistics, we say that

$$\begin{cases} H_0 & a = b \text{ is the null hypothesis,} \\ H_1 & a \neq b \text{ is the alternative hypothesis.} \end{cases}$$

The following theorem is our main result on distribution limits of empirical Sinkhorn divergences.

THEOREM III.3. *Recall that $K = \exp(-C/\varepsilon)$ is the matrix obtained by element wise exponential of $-C/\varepsilon$. Then, the following central limit theorems holds for empirical Sinkhorn divergences.*

(1) Null hypothesis, i.e. $a = b$. Let $(u, v) \in \mathbb{R}_+^{N\times N}$ be a fixed point of the Sinkhorn map $S_{\{a,a\}}$ defined in (III.1)

(a) H_0 - One sample.

$$\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, a) - W_{2,\varepsilon}^2(a, a)) \xrightarrow{\mathcal{L}} \langle G, \varepsilon \log(u)\rangle. \tag{III.2}$$

(b) H_0 - Two samples. Let $\rho_{n,m} = \sqrt{(nm/(n+m))}$. If n and m tend to infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to \theta \in (0,1)$, then

$$\rho_{n,m}(W_{2,\varepsilon}^2(\hat{a}_n, \hat{b}_m) - W_{2,\varepsilon}^2(a, a)) \xrightarrow{\mathcal{L}} \langle G, \varepsilon \log(u)\rangle. \tag{III.3}$$

(2) Alternative case, i.e. $a \neq b$. Let $(u, v) \in \mathbb{R}_+^{N\times N}$ be a fixed point of the Sinkhorn map $S_{\{a,b\}}$

(a) H_1 - One sample.

$$\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, b) - W_{2,\varepsilon}^2(a, b)) \xrightarrow{\mathcal{L}} \langle G, \varepsilon \log(u)\rangle. \tag{III.4}$$

(b) H_1 - Two samples. For $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \theta \in (0,1)$,

$$\rho_{n,m}(W_{2,\varepsilon}^2(\hat{a}_n, \hat{b}_m) - W_{2,\varepsilon}^2(a, b)) \xrightarrow{\mathcal{L}} \sqrt{\theta}\langle G, \varepsilon \log(u)\rangle + \sqrt{1-\theta}\langle H, \varepsilon \log(v)\rangle. \tag{III.5}$$

PROOF. Following the proof of Theorem 1 in [**SM16**], we have that (e.g. thanks to Theorem 14.6 in [**Was11**])

$$\sqrt{n}(\hat{a}_n - a) \xrightarrow{\mathcal{L}} G, \text{ where } G \overset{\mathcal{L}}{\sim} \mathcal{N}(0, \Sigma(a)),$$

since $n\hat{a}_n$ is a sample of a multinomial probability measure with probability a.

Therefore, for the one sample case, we apply the Delta-method for directionally differentiable functions in the sense of Hadamard (see Theorem 1 of Romisch in [**Röm05**]). Thanks to Theorem III.2, we directly get:

$$\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, a) - W_{2,\varepsilon}^2(a, a)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\varepsilon(a,a)} \langle G, \alpha\rangle \tag{III.6}$$

$$\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, b) - W_{2,\varepsilon}^2(a, b)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\varepsilon(a,b)} \langle G, \alpha\rangle \text{ for } a \neq b. \tag{III.7}$$

For the two samples case, we use that

$$\rho_{n,m}((\hat{a}_n, \hat{b}_m) - (a, b)) \xrightarrow{\mathcal{L}} (\sqrt{t}G, \sqrt{1-t}H),$$

where $\rho_{n,m}$ and t are given in the statement of the Theorem. Then, applying again the delta-method for Hadamard directionally diffenrentiable functions, we obtain that for n and m tending to infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to t \in (0,1)$,

$$\rho_{n,m}(W_{2,\varepsilon}^2(\hat{a}_n, \hat{b}_m) - W_{2,\varepsilon}^2(a, b)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\varepsilon(a,b)} \sqrt{\theta}\langle G, \alpha\rangle + \sqrt{1-\theta}\langle H, \beta\rangle. \tag{III.8}$$

In the null hypothesis case ($a = b$), this simplifies into

$$\rho_{n,m}(W_{2,\varepsilon}^2(\hat{a}_n, \hat{b}_m) - W_{2,\varepsilon}^2(a, a)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\varepsilon(a,a)} \langle G, \alpha\rangle. \tag{III.9}$$

Now, thanks to Proposition III.1, we know that there exists positive vectors $u \in \mathbb{R}_+^N$ and $v \in \mathbb{R}_+^N$ (unique up to scalar multiplication) such that an optimal solution in $N_\varepsilon(a, b)$ of $W_{2,\varepsilon}^2$ is given by

$$\alpha^* = -\varepsilon \log(u), \quad \beta^* = -\varepsilon \log(v)$$

for a and b equal or not. From such results, for (u, v) obtained through Sinkhorn's algorithm (III.1), we can deduce that

$$\max_{(\alpha,\beta)\in N_\varepsilon(a,b)} \langle G, \alpha\rangle \overset{\mathcal{L}}{\sim} \max_{t\in\mathbb{R}} \langle G, -\varepsilon\log(u) + t\mathbb{1}_N\rangle \overset{\mathcal{L}}{\sim} \max_{t\in\mathbb{R}} \left(\langle G, -\varepsilon\log(u)\rangle + \langle G, t\mathbb{1}_N\rangle\right).$$

Moreover, we have

$$\langle G, t\mathbb{1}_N\rangle \overset{\mathcal{L}}{\sim} \mathcal{N}(t\mathbb{1}'_N\,\mathbb{E}(G), t\mathbb{1}'_N\Sigma(a)t\mathbb{1}_N) \overset{\mathcal{L}}{\sim} \mathcal{N}(0, t^2\mathbb{1}'_N\Sigma(a)\mathbb{1}_N) \overset{\mathcal{L}}{\sim} \mathcal{N}(0,0) \overset{\mathcal{L}}{\sim} \delta_0$$

since G is centered in 0 and $\mathbb{1}'_N\Sigma(a)\mathbb{1}_N = 0$ for a in the simplex. Notice that $\langle G, -\varepsilon\log(u)\rangle \overset{\mathcal{L}}{\sim} \langle G, \varepsilon\log(u)\rangle$. Hence, let Y be a random variable of law δ_0. By independence, we have that $\langle G, -\varepsilon\log(u)\rangle + Y$ follows the same law as $\langle G, \varepsilon\log(u)\rangle$ since G is centered in 0. By the same process, we get

$$\max_{(\alpha,\beta)\in N_\varepsilon(a,b)} \sqrt{\theta}\langle G, \alpha\rangle + \sqrt{1-\theta}\langle H, \beta\rangle \;\overset{\mathcal{L}}{\sim}\; \sqrt{\theta}\langle G, \varepsilon\log(u)\rangle + \sqrt{1-\theta}\langle H, \varepsilon\log(v)\rangle.$$

Therefore we apply this result to the convergence in distribution obtained previously in (III.8) and (III.9), which concludes the proof. □

Distribution limits of empirical Sinkhorn divergences may also be characterized by the following result which follows from Theorem 1 of Romisch [**Röm05**] using the property that $\Sigma_N \times \Sigma_N$ is a convex set.

THEOREM III.4. *The following asymptotic result holds for empirical Sinkhorn divergences.*

(1) One sample

$$\sqrt{n}\left(W^2_{2,\varepsilon}(\hat{a}_n, b) - W^2_{2,\varepsilon}(a, b) - \max_{(\alpha,\beta)\in N_\varepsilon(a,b)} \langle \hat{a}_n - a, \alpha\rangle\right) \xrightarrow{\mathbb{P}} 0.$$

(2) Two samples - For $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \theta \in (0,1)$,

$$\rho_{n,m}\left(W^2_{2,\varepsilon}(\hat{a}_n, \hat{b}_m) - W^2_{2,\varepsilon}(a, b) - \max_{(\alpha,\beta)\in N_\varepsilon(a,b)} (\langle \hat{a}_n - a, \alpha\rangle + \langle \hat{b}_m - b, \beta\rangle)\right) \xrightarrow{\mathbb{P}} 0.$$

III.3. Use of bootstrap for statistical inference

The results obtained in Section III.2 on the distribution of empirical Sinkhorn divergences are only asymptotic, and it is thus of interest to estimate their non-asymptotic distribution using a bootstrap procedure. Bootstrap consists in drawing new samples from an empirical distribution $\mathbb{P}_n$ that has been obtained from an unknown distribution $\mathbb{P}$. Therefore, conditionally on $\mathbb{P}_n$, it allows to obtain new observations (considered as approximately sampled from $\mathbb{P}$) that can be used to approximate the distribution of a test statistics using Monte-Carlo experiments. We refer to [**ET93**] for a general introduction to the bootstrap procedure.

Nevertheless, as carefully explained in [**SM16**], for a test statistic based on functions that are only Hadamard directionally differentiability, a classical bootstrap procedure is not consistent. To overcome this issue, we decide to choose α and β in $N_\varepsilon(a, b)$ (defined in Theorem III.2) such that their components sum up to zero. In this way the optimal solution of the dual problem (A.10) becomes unique as initially remarked in [**CD14**]. We denote this solution by $(\alpha^0_\varepsilon, \beta^0_\varepsilon)$, and we let $N^0_\varepsilon(a, b) = \{(\alpha^0_\varepsilon, \beta^0_\varepsilon)\}$. Under this additional normalization, the previous results remain true. In particular, the directional derivative of $W^2_{2,\varepsilon}$ at (a, b) becomes

$$(W^2_{2,\varepsilon})'_\varepsilon(a, b) : (h_1, h_2) \mapsto \langle \alpha^0_\varepsilon, h_1\rangle + \langle \beta^0_\varepsilon, h_2\rangle,$$

which is a linear map. Hence, by Proposition 2.1 in [FS14], the functional $(a,b) \mapsto W^2_{2,\varepsilon}(a,b)$ is Hadamard differentiable in the usual sense on $\Sigma_N \times \Sigma_N$. We can thus apply the Delta-method to prove consistency of the bootstrap in our setting using the bounded Lipschitz metric defined below.

DEFINITION III.5. *The Bounded Lipschitz (BL) metric is defined for μ, ν probability measures on Ω by*

$$d_{BL}(\mu,\nu) = \sup_{h \in BL_1(\Omega)} \left| \int h d\mu - \int h d\nu \right|$$

where $BL_1(\Omega)$ is the set of real functions $\Omega \to \mathbb{R}$ such that $\|h\|_\infty + \|h\|_{Lip} \leqslant 1$.

Our main result adapted on the use of bootstrap samples can be stated as follows.

THEOREM III.6. *For $X_1, \dots, X_n \overset{\mathcal{L}}{\sim} a$ and $Y_1, \dots, Y_m \overset{\mathcal{L}}{\sim} b$, let $\hat{a}^*_n$ (resp. $\hat{b}^*_m$) be bootstrap versions of $\hat{a}_n$ (resp. $\hat{b}_m$) of size n (resp. m).*

(1) One sample case: $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}^*_n, b) - W^2_{2,\varepsilon}(\hat{a}_n, b))$ converges in distribution (conditionally on $X_1, \dots, X_n$) to $\langle G, \alpha^0_\varepsilon \rangle$ for the BL metric, in the sense that

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}(h(\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}_n, b) - W^2_{2,\varepsilon}(\hat{a}^*_n, b)))|X_1, \dots, X_n] - \mathbb{E}[h\langle G, \alpha^0_\varepsilon \rangle]| \xrightarrow{\mathbb{P}} 0$$

(2) Two samples case: $\rho_{n,m}(W^2_{2,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - W^2_{2,\varepsilon}(\hat{a}_n, \hat{b}_m))$ converges in distribution (conditionally on $X_1, \dots, X_n$, $Y_1, \dots, Y_m$) to $\sqrt{\theta}\langle G, \alpha^0_\varepsilon \rangle + \sqrt{1-\theta}\langle H, \beta^0_\varepsilon \rangle$ for the BL metric, in the sense that

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}(h(\rho_{n,m}(W^2_{2,\varepsilon}(\hat{a}_n, \hat{b}_m) - W^2_{2,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m)))|X_1, \dots, X_n, Y_1, \dots, Y_m]$$
$$- \mathbb{E}[h(\sqrt{\theta}\langle G, \alpha^0_\varepsilon \rangle + \sqrt{1-\theta}\langle H, \beta^0_\varepsilon \rangle)]| \xrightarrow{\mathbb{P}} 0$$

PROOF. We only prove the one sample case since both convergence can be shown by similar arguments. We know that $\sqrt{n}(\hat{a}_n - a)$ tends in distribution to $G \sim \mathcal{N}(0, \Sigma(a))$. Moreover $\sqrt{n}(\hat{a}^*_n - \hat{a}_n)$ converges (conditionally on $X_1, \dots, X_n$) in distribution to G by Theorem 3.6.1 in [VDVW96]. Theorem 3.9.11 in the same book, on the consistency of the Delta-method combined with bootstrap, allows us to conclude. □

III.4. Numerical experiments with synthetic data

We propose to illustrate Theorem III.3 and Theorem III.6 with simulated data consisting of random measures supported on a $p \times p$ square grid of regularly spaced points $(x_i)_{i=1,\dots,N}$ in $\mathbb{R}^2$ (with $N = p^2$) for p ranging from 5 to 20. We use the squared Euclidean distance. Therefore, the cost C scales with the size of the grid. The range of interesting values for ε is thus closely linked to the size of the grid (as it can be seen in the expression of $K = \exp(-C/\varepsilon)$. Hence, $\varepsilon = 100$ for a 5×5 grid corresponds to more regularization than $\varepsilon = 100$ for a 20×20 grid.

We ran our experiments on Matlab using the accelerate version [TCDP17][1] of the Sinkhorn transport algorithm [Cut13]. Furthermore, we considered the numerical logarithmic stabilization described in [SHB+18a] which allows to handle small values of ε.

[1] http://www.math.u-bordeaux.fr/~npapadak/GOTMI/codes.html

III.4.1. Convergence in distribution

We first illustrate the convergence in distribution of empirical Sinkhorn divergences (as stated in Theorem III.3) for either the hypothesis H_0 with one sample, or the hypothesis H_1 with two samples.

Hypothesis H_0 - One sample. We consider the case where a is the uniform distribution on a square grid. We generate $M = 10^3$ empirical distributions $\hat{a}_n$ (such that $n\hat{a}_n$ follows a multinomial distribution with parameter a) for different values of n and grid size. In this way, we obtain M realizations of $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}_n, a) - W^2_{2,\varepsilon}(a, a))$, and we use a kernel density estimate (with a data-driven bandwidth) to compare the distribution of these realizations to the density of the Gaussian distribution $\langle G, \varepsilon \log(u)\rangle$. The results are reported in Figure III.1.

It can be seen that the convergence of empirical Sinkhorn divergences to its asymptotic distribution ($n \to \infty$) is relatively slow. Moreover, for a fixed number n of observations, the convergence becomes slower as ε increases. We can also notice that for various values of (n, ε), the non-asymptotic distribution of $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}_n, a) - W^2_{2,\varepsilon}(a, a))$ seems to be non-Gaussian. This justifies the use of the bootstrap procedure described in Section III.3.

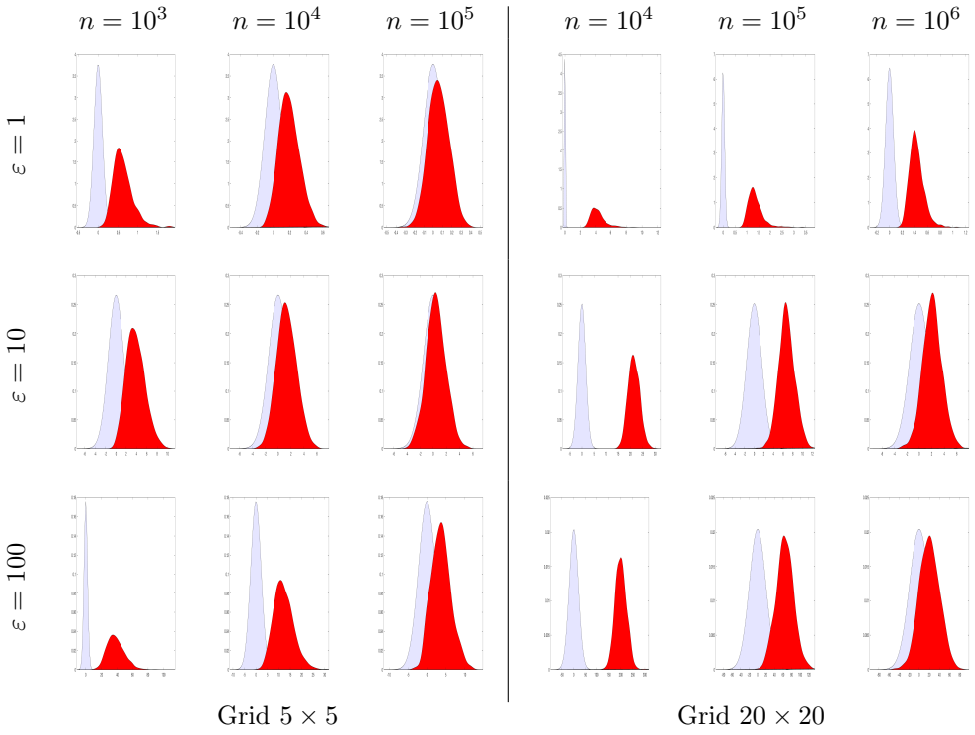


FIGURE III.1. Hypothesis H_0 with one sample. Illustration of the convergence in distribution of empirical Sinkhorn divergences for a 5×5 grid (left) and a 20×20 grid (right), for $\varepsilon = 1, 10, 100$ and n ranging from 10^3 to 10^6. Densities in red (resp. light blue) represent the distribution of $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}_n, a) - W^2_{2,\varepsilon}(a, a))$ (resp. $\langle G, \varepsilon \log(u)\rangle$).

Let us now shed some light on the bootstrap procedure described in Section III.3. The results on bootstrap experiments are reported in Figure III.2. From the uniform distribution a, we generate one random distribution $\hat{a}_n$. The value of the realization $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}_n, a) - W^2_{2,\varepsilon}(a, a))$ is represented by the red vertical lines in Figure III.2.

Besides, we generate from $\hat{a}_n$, a sequence of $M = 10^3$ bootstrap samples of random measures denoted by $\hat{a}_n^*$ (such that $n\hat{a}_n^*$ follows a multinomial distribution with parameter $\hat{a}_n$). We use again a kernel density estimate (with a data-driven bandwidth) to compare the distribution of $\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n^*, a) - W_{2,\varepsilon}^2(\hat{a}_n, a))$ to the distribution of $\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, a) - W_{2,\varepsilon}^2(a, a))$ displayed in Figure III.1. The green vertical lines in Figure III.2 represent a confidence interval of level 95%. The observation represented by the red vertical line is consistently located with respect to this confidence interval, and the density estimated by bootstrap decently captures the shape of the non-asymptotic distribution of Sinkhorn divergences.

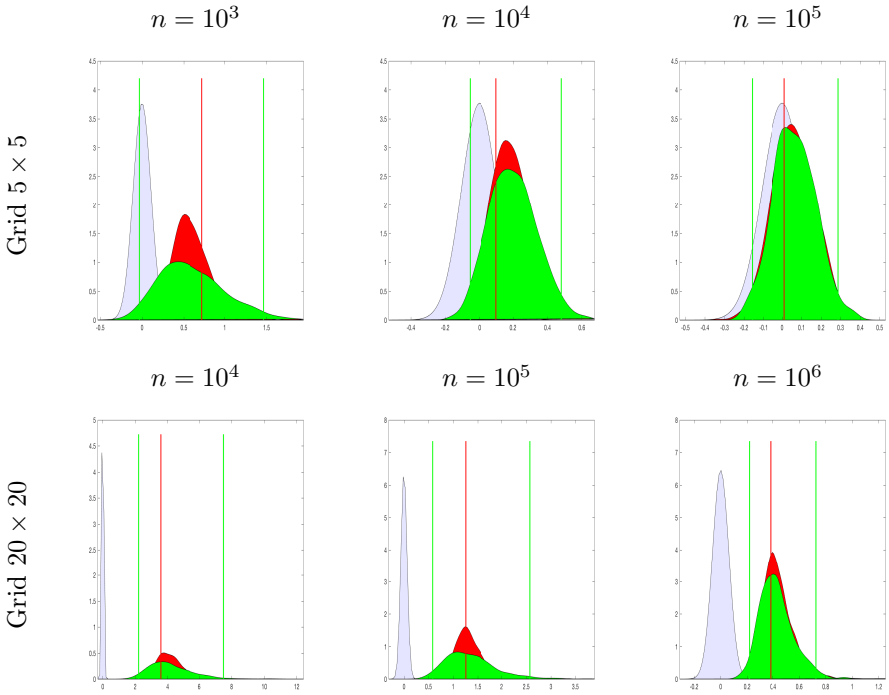


FIGURE III.2. Hypothesis H_0 with one sample. Illustration of the bootstrap with $\varepsilon = 1$ and two grids of size 5×5 and 20×20 to approximate the non-asymptotic distribution of empirical Sinkhorn divergences. Densities in red (resp. light blue) represent the distribution of $\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, a) - W_{2,\varepsilon}^2(a, a))$ (resp. $\langle G, \varepsilon \log(u) \rangle$). The green density represents the distribution of the random variable $\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n^*, a) - W_{2,\varepsilon}^2(\hat{a}_n, a))$ in Theorem III.6.

Hypothesis H_1 - Two samples. We consider now the setting where a is still a uniform distribution, and

$$b \propto \mathbb{1}_N + \theta(1, 2, \ldots, N)$$

is a distribution with linear trend depending on a slope parameter $\theta \geqslant 0$ that is fixed to 0.5, see Figure III.3.

As previously, we run $M = 10^3$ experiments to obtain a kernel density estimation of the distribution of

$$\rho_{n,m}(W_{2,\varepsilon}^2(\hat{a}_n, \hat{b}_m) - W_{2,\varepsilon}^2(a, b)),$$

that we compare to the density of the Gaussian variable with mean 0 and variance

$$\varepsilon \sqrt{\theta \log(u)^t \Sigma(a) \log(u) + (1-\theta) \log(v)^t \Sigma(b) \log(v)}.$$

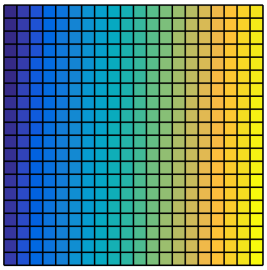


FIGURE III.3. Example of a distribution b with linear trend (with slope parameter $\theta = 0.5$ on a 20×20 grid).

The results are reported in Figure III.4. The convergence of empirical Sinkhorn divergences to their asymptotic distribution seems to be much faster under the hypothesis H_1, but increasing the regularization parameter still makes this convergence slower.

REMARK III.7. *A possible explanation for the slow convergence under the hypothesis H_0 is that, in this setting, the Sinkhorn divergence $W_{2,\varepsilon}^2(a,a)$ is very close to 0, but as soon as we generate an empirical measure $\hat{a}_n$, the value of $W_{2,\varepsilon}^2(\hat{a}_n,a)$ seems to explode in comparison to the divergence between a and itself.*

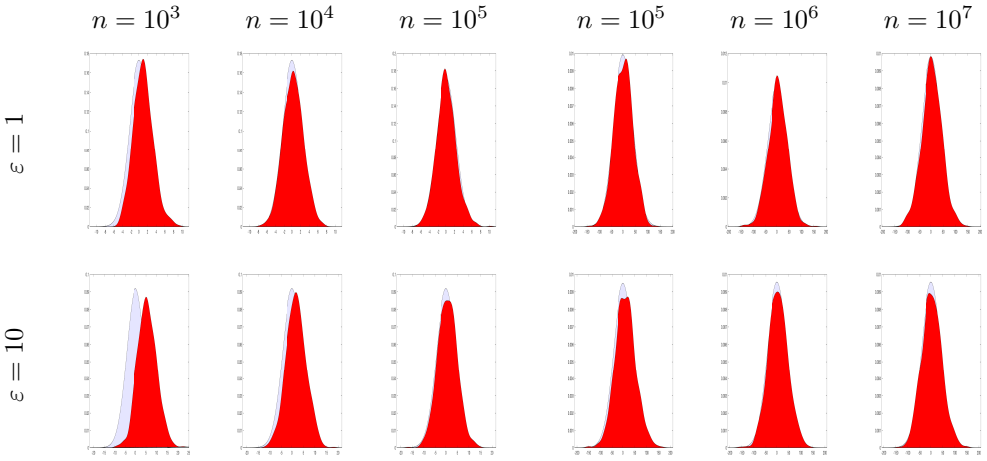


FIGURE III.4. Hypothesis H_1 - two samples. Illustration of the convergence in distribution of empirical Sinkhorn divergences for a 5×5 grid (left) and a 20×20 grid (right), for $\varepsilon = 1, 10$, $n = m$ and n ranging from 10^3 to 10^7. Densities in red (resp. blue) represent the distribution of $\rho_{n,m}(W_{2,\varepsilon}^2(\hat{a}_n,\hat{b}_m) - W_{2,\varepsilon}^2(a,b))$ (resp. $\sqrt{\theta}\langle G, \varepsilon\log(u)\rangle + \sqrt{1-\theta}\langle H, \varepsilon\log(v)\rangle$ with $\theta = 1/2$).

We also report in Figure III.5 results on the consistency of the bootstrap procedure under the hypothesis H_1 with two samples. From the distributions a and b, we generate two random distributions $\hat{a}_n$ and $\hat{b}_m$. The value of the realization $\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n,\hat{b}_n) - W_{2,\varepsilon}^2(a,b))$ is represented by the red vertical lines in Figure III.5. Then, we generate from $\hat{a}_n$ and $\hat{b}_m$, two sequences of $M = 10^3$ bootstrap samples of random measures denoted by $\hat{a}_n^*$

and $\hat{b}^*_m$. We use again a kernel density estimate (with a data-driven bandwidth) to compare the green distribution of $\rho_{n,m}(W^2_{2,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - W^2_{2,\varepsilon}(\hat{a}_n, \hat{b}_m))$ to the red distribution of $\rho_{n,m}(W^2_{2,\varepsilon}(\hat{a}_n, \hat{b}_m) - W^2_{2,\varepsilon}(a, b))$ displayed in Figure III.5. The green vertical lines in Figure III.5 represent a confidence interval of level 95%. The observation represented by the red vertical line is consistently located with respect to this confidence interval, and the green density estimated by bootstrap captures very well the shape and location of the non-asymptotic distribution of Sinkhorn divergences.

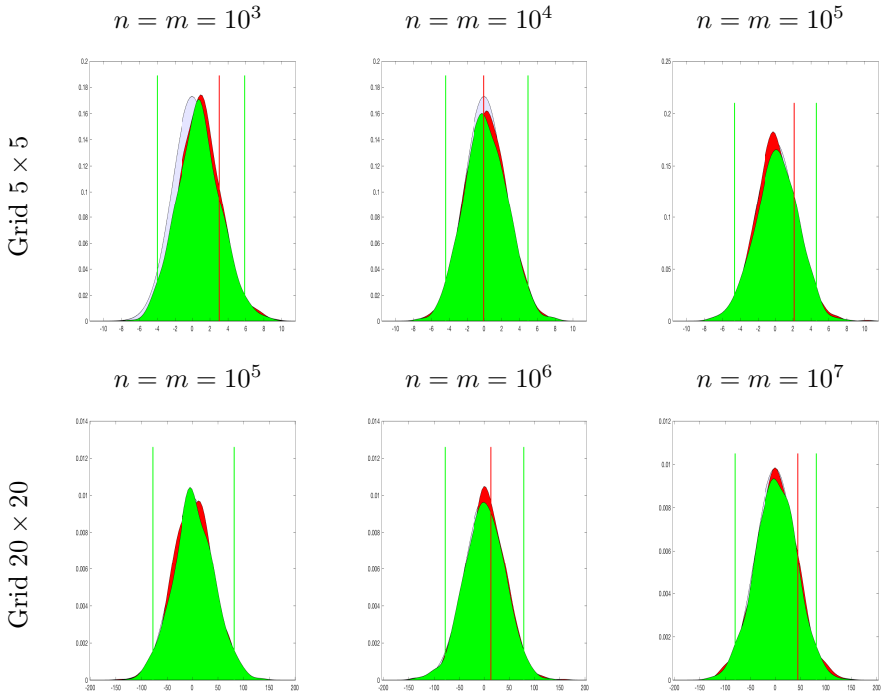


FIGURE III.5. Hypothesis H_1 - two samples. Illustration of the bootstrap with $\varepsilon = 1$ and two grids of size 5×5 and 20×20 to approximate the non-asymptotic distribution of empirical Sinkhorn divergences. Densities in red (resp. blue) represent the distribution of $\rho_{n,m}(W^2_{2,\varepsilon}(\hat{a}_n, \hat{b}_m) - W^2_{2,\varepsilon}(a, b))$ (resp. $\sqrt{\theta}\langle G, \varepsilon \log(u)\rangle + \sqrt{1-\theta}\langle H, \varepsilon \log(v)\rangle$). The green density is the distribution of the random variable $\rho_{n,m}(W^2_{2,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - W^2_{2,\varepsilon}(\hat{a}_n, \hat{b}_m))$ in Theorem III.6.

III.4.2. Estimation of test power using the bootstrap

One sample - distribution with linear trend and varying slope parameter. We illustrate the consistency and usefulness of the bootstrap procedure by studying the statistical power (that is $\mathbb{P}(\text{Reject } H_0 | H_1 \text{ is true})$) of statistical tests (at level 5%) based on empirical Sinkhorn divergences. For this purpose, we choose a to be a distribution with linear trend whose slope parameter θ is ranging from 0 to 0.15 on a 5×5 grid and b to be uniform. We assume that we observe a single realization of an empirical measure $\hat{a}_n$ sampled from a with $n = 10^3$. Then, we generate $M = 10^3$ bootstrap samples of random measures

$\hat{a}^*_{n,j}$ from $\hat{a}_n$ (with $1 \leqslant j \leqslant M$), which allows the computation of the p-value

$$p\text{-value} = \#\{j \text{ such that } \sqrt{n}|W^2_{2,\varepsilon}(\hat{a}^*_{n,j}, b) - W^2_{2,\varepsilon}(\hat{a}_n, b)| \geqslant \sqrt{n}|W^2_{2,\varepsilon}(\hat{a}_n, b) - W^2_{2,\varepsilon}(a, b)|\}/M.$$

This experiments is repeated 100 times, in order to estimate the power (at level $\alpha = 5\%$) of a test based on $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}_n, b) - W^2_{2,\varepsilon}(a, b))$ by comparing the resulting sequence of p-values to the value α. The results are reported in Figure III.6.

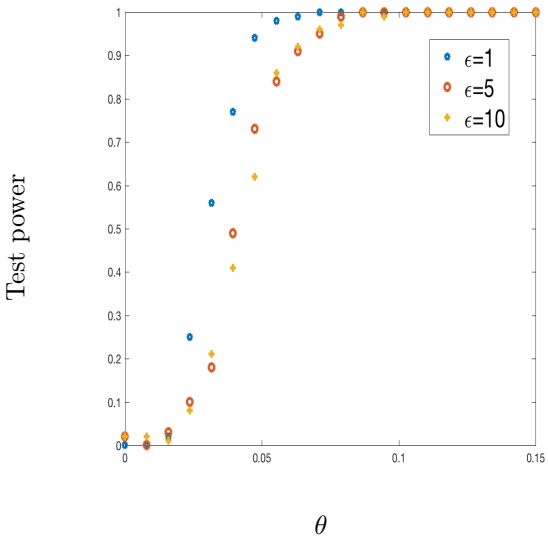


FIGURE III.6. Test power (probability of rejecting H_0 knowing that H_1 is true) on a 5×5 grid in the one sample case, as a function of the slope parameter θ ranging from 0 to 0.15 for $\varepsilon = 1$ (blue), $\varepsilon = 5$ (orange) and $\varepsilon = 10$ (yellow), with $n = 10^3$.

It can be seen that this test is a good discriminant, especially when ε is small. As soon as the slope θ increases and b sufficiently differs from a, then the probability of rejecting H_0 increases. Moreover, for a fixed value of the slope parameter θ of distribution b, the test power becomes larger as ε gets smaller. This suggests the use of a small regularization parameter ε to be more accurate for discriminating two measures using statistical testing based on empirical Sinkhorn divergences.

III.5. Analysis of real data

We consider a dataset containing the locations of reported incidents of crime (with the exception of murders) in Chicago in 2014 which is publicly available[2], and that has been recently studied in [**BCP18b**] and [**Ger16**]. Victims' addresses are shown at the block level only (specific locations are not identified) in order to (i) protect the privacy of victims and (ii) have a sufficient amount of data for the statistical analysis. The city of Chicago is represented as a two-dimensional grid $\mathcal{X} = \{x_1, \ldots, x_N\}$ of size $N = 27 \times 18 = 486$ of equi-spaced points $x_i = (x_i^{(1)}, x_i^{(2)}) \in [1, 27] \times [1, 18] \subset \mathbb{R}^2$. For each month $1 \leqslant k \leqslant 12$ of

[2] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

the year 2014, the spatial locations of reported incidents of crime in Chicago are available. This yields to a dataset made of 12 empirical measures

$$\hat{\eta}_k = \sum_{i=1}^{N} \hat{a}_i^{(k)} \delta_{x_i} \text{ for } 1 \leqslant k \leqslant 12,$$

where $\hat{a}_i^{(k)}$ is the relative frequency of reported crimes for month k at location x_i. We denote by $n = n_k$ the number of reported crimes for month k. This dataset is displayed in Figure III.7 and III.8. To compute the cost matrix C, we use the squared Euclidean distance between the spatial locations $x_i \in \mathbb{R}^2$.

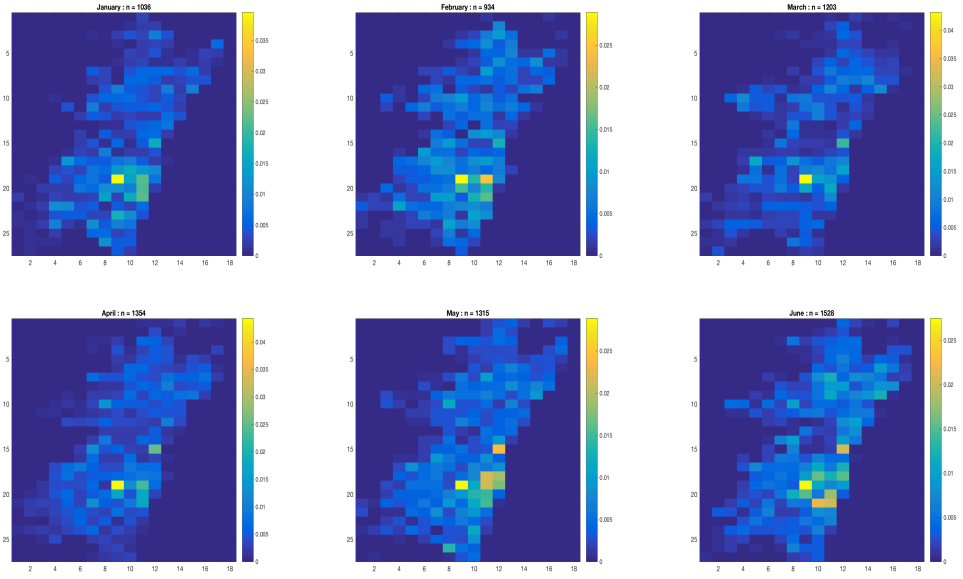


FIGURE III.7. Spatial locations of reported incidents (relative frequencies) of crime in Chicago for the first 6 months of 2014 over a two-dimensional grid of size 27×18.

III.5.1. Testing the hypothesis of uniform distribution of crimes locations

We first test the null hypothesis that the distribution of crimes locations over the whole year 2014 is uniform. To this end, we consider the Euclidean barycenter of the dataset $(\hat{\eta}_k)_{1 \leqslant k \leqslant n}$ defined as

$$\bar{\eta}_{12} = \frac{1}{12} \sum_{k=1}^{12} \hat{\eta}_k = \sum_{i=1}^{N} \bar{a}_i \delta_{x_i}$$

which represents the locations of crime in 2014. This discrete measure is displayed in Figure III.9(a). It can be seen that $\bar{\eta}_{12}$ is a discrete empirical measure consisting of $n = 16104$ observations such that $\bar{a}_i = 0$ for many locations x_i. We use the one sample testing procedure described previously, and a bootstrap approach to estimate the distribution of the test statistics

$$\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, a) - W_{2,\varepsilon}^2(a, a))$$

with $\hat{a}_n = \bar{\eta}_{12}$ and a the uniform distribution over the support of $\bar{\eta}_{12}$ defined as $\{x_i : \bar{a}_i \neq 0, 1 \leqslant i \leqslant N\}$, see Figure III.9(b). We report results for $\varepsilon = 1$ and $\varepsilon = 5$ by displaying in

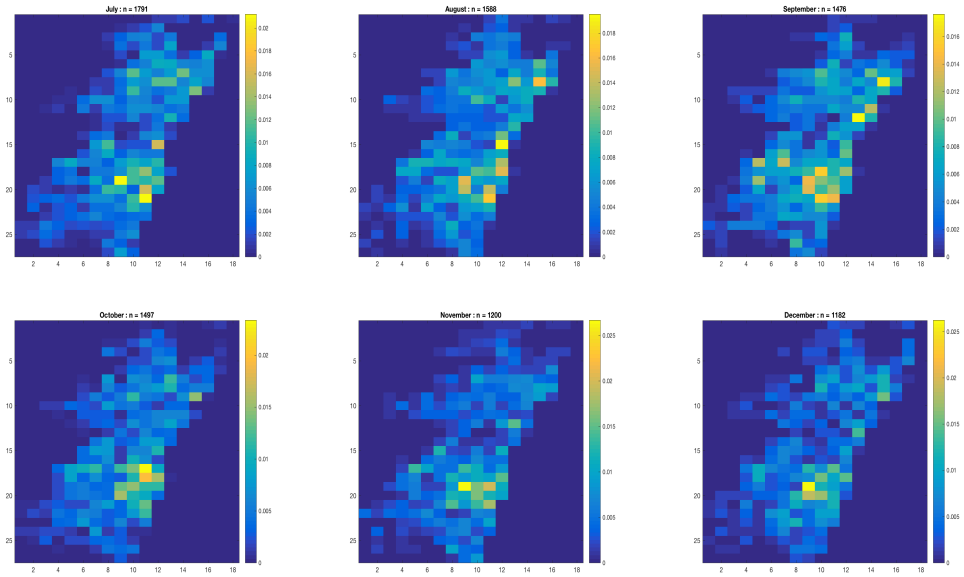


FIGURE III.8. Spatial locations of reported incidents (relative frequencies) of crime in Chicago for the last 6 months of 2014 over a two-dimensional grid of size 27×18.

Figure III.9(cd) an estimation of the density of the bootstrap statistics $\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n^*, a) - W_{2,\varepsilon}^2(\hat{a}_n, a))$. For both values of ε, the value of $\sqrt{n}(W_{2,\varepsilon}^2(\hat{a}_n, a) - W_{2,\varepsilon}^2(a, a))$ is outside the support of this density, and the null hypothesis that crimes are uniformly distributed (over the support of $\bar{\eta}_{12}$ is thus rejected.

III.5.2. Testing equality across months

We propose now to investigate the possibility of equal distributions of crime locations between different months. To this end, we first compute a reference measure using data from the first 6 months. Under the assumption that the distribution of crime locations does not change from one month to another, it is natural to consider the Euclidean barycenter

$$\bar{\eta}_6 = \frac{1}{6} \sum_{k=1}^{6} \hat{\eta}_k,$$

as a reference measure to which the data from the last 6 months of 2014 can be compared. The measure $\bar{\eta}_6$ is displayed in Figure III.10(a) and Figure III.11(a).

One sample testing. We use the one sample testing procedure described previously, and a bootstrap approach to estimate the distribution of the test statistics

$$\sqrt{n_k}(W_{2,\varepsilon}^2(\hat{a}_{n_k}, a) - W_{2,\varepsilon}^2(a, a))$$

with $a = \bar{\eta}_6$ and $\hat{a}_{n_k} = \hat{\eta}_k$, for $7 \leqslant k \leqslant 12$. We report results for $\varepsilon = 1$ by displaying in Figure III.10 an estimation of the density of the bootstrap statistics $\sqrt{n_k}(W_{2,\varepsilon}^2(\hat{a}_{n_k}^*, a) - W_{2,\varepsilon}^2(\hat{a}_{n_k}, a))$, and the values of the observations $\sqrt{n_k}(W_{2,\varepsilon}^2(\hat{a}_{n_k}, a) - W_{2,\varepsilon}^2(a, a))$ for the last 6 months of 2014. It can be seen that, at level 5%, the null hypothesis that the distribution of crime locations is equal to the reference measure $\bar{\eta}_6$ is accepted for the months of September, October, November and December, but that it is rejected for the months of July and August.

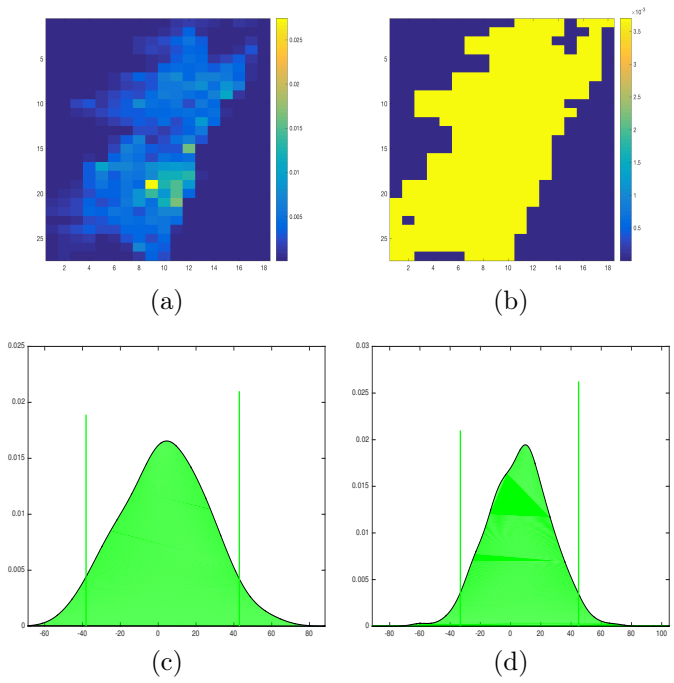


FIGURE III.9. Testing uniform distribution of crimes locations. (a) Euclidean barycenter $\bar{\eta}_{12}$ (empirical measure corresponding to locations of crime in Chicago for the whole year 2014 over a two-dimensional grid of size 27×18), (b) Uniform distribution a over the support of $\bar{\eta}_{12}$. Green densities represent the distribution of the bootstrap statistics $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}^*_n, a) - W^2_{2,\varepsilon}(\hat{a}_n, a))$ (vertical bars represent a confidence interval of level 95%) for (c) $\varepsilon = 1$ and (d) $\varepsilon = 5$. The value of $\sqrt{n}(W^2_{2,\varepsilon}(\hat{a}_n, a) - W^2_{2,\varepsilon}(a, a))$ (with $\hat{a}_n = \bar{\eta}_{12}$) is outside the support $[-100, 100]$ for each value of ε, and it is thus not represented.

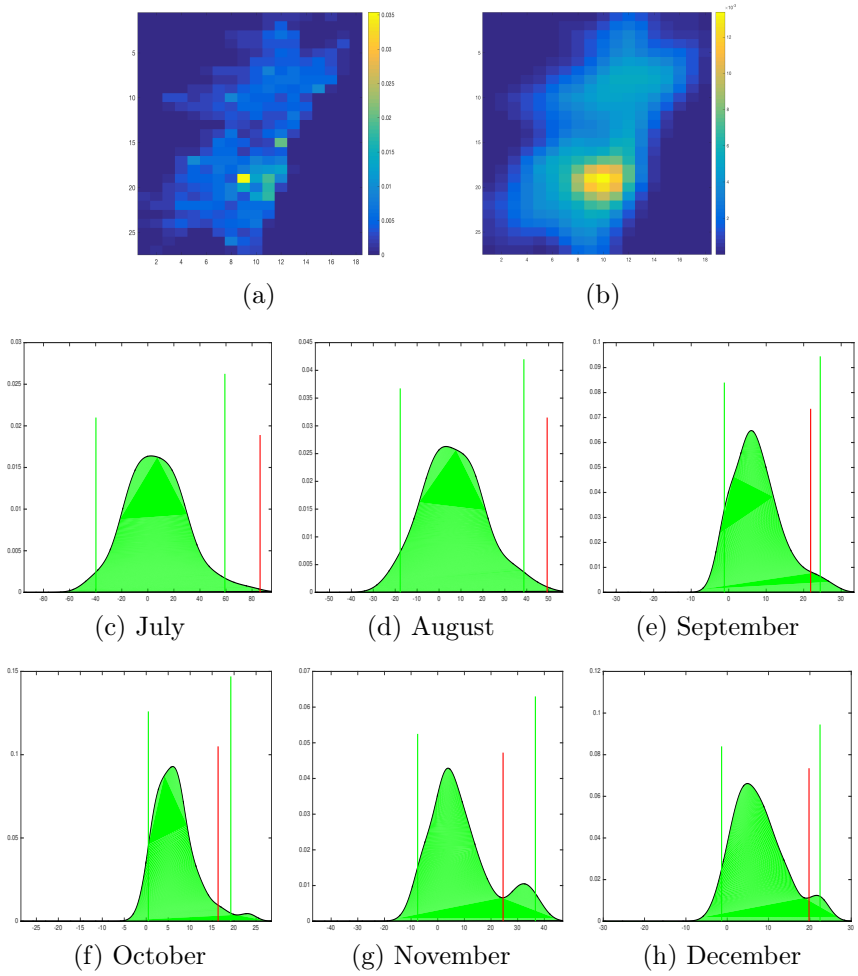
Alternatively, one may think of using a smoothed Wasserstein barycenter $\bar{\eta}^\varepsilon_6$ of the data $(\hat{\eta}_k)_{1 \leqslant k \leqslant 6}$ as a reference measure that is defined as

$$\bar{\eta}^\varepsilon_6 = \operatorname*{arg\,min}_{\eta \in \mathcal{P}_p(\mathcal{X})} \frac{1}{6} \sum_{k=1}^{6} p_\varepsilon(\hat{\eta}_k, \eta).$$

To compute such a smoothed Wasserstein barycenter, we use the algorithmic approach proposed in [**CP16b**], and we display $\bar{\eta}^\varepsilon_6$ for $\varepsilon = 1$ in Figure III.10(b) and $\varepsilon = 0.3$ in Figure III.11(b).

For $\varepsilon = 1$, this smoothed Wasserstein barycenters is visually quite different from the measures $(\hat{\eta}_k)_{7 \leqslant k \leqslant 12}$ that are displayed in Figure III.8. For $\varepsilon = 1$, we found that using $\bar{\eta}^\varepsilon_6$ as a reference measure in one sample testing (with $\hat{a}_{n_k} = \hat{\eta}_k$ and $a = \bar{\eta}^\varepsilon_6$) leads to reject the null hypothesis that the distribution of crime locations is equal to $\bar{\eta}^\varepsilon_6$ for all $7 \leqslant k \leqslant 12$ (last 6 months of 2014). As a consequence we do not display the corresponding results.

For $\varepsilon = 0.3$, the Wasserstein barycenter $\bar{\eta}^\varepsilon_6$ is a slightly smoothed version of the Euclidean one $\bar{\eta}_6$. We display in Figure III.11 an estimation of the density of the bootstrap statistics $\sqrt{n_k}(W^2_{2,\varepsilon}(\hat{a}^*_{n_k}, a) - W^2_{2,\varepsilon}(\hat{a}_{n_k}, a))$, and the values of the observations $\sqrt{n_k}(W^2_{2,\varepsilon}(\hat{a}_{n_k}, a) - W^2_{2,\varepsilon}(a, a))$ for the last 6 months of 2014, with $a = \bar{\eta}^\varepsilon_6$ and $\varepsilon = 0.3$. At level 5%, the null hypothesis that the distribution of crime locations is equal to the reference measure $\bar{\eta}^\varepsilon_6$ is



(a) (b)

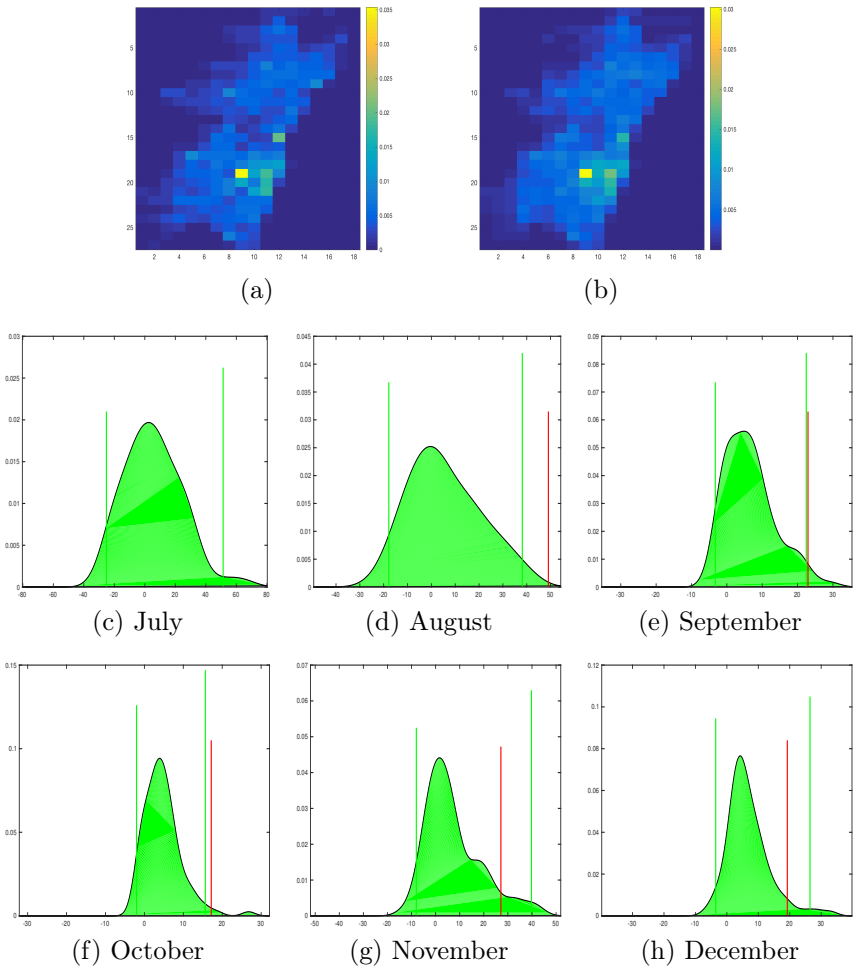
(c) July (d) August (e) September

(f) October (g) November (h) December

FIGURE III.10. Testing equality of distributions over months for $\varepsilon = 1$ with the Euclidean barycenter as a reference measure. (a) Euclidean barycenter $\bar{\eta}_6$ (empirical measure corresponding to locations of crime in Chicago for the first 6 months of 2014). (b) Smoothed Wasserstein barycenter $\bar{\eta}_6^\varepsilon$ of the measures $(\hat{\eta}_k)_{1\leqslant k\leqslant 6}$ for $\varepsilon = 1$. (c)-(h) Green densities represent the distribution of the bootstrap statistics $\sqrt{n_k}(W^2_{2,\varepsilon}(\hat{a}^*_{n_k}, a) - W^2_{2,\varepsilon}(\hat{a}_{n_k}, a))$ for the last 6 months of 2014, with $a = \bar{\eta}_6$ and $\hat{a}_{n_k} = \hat{\eta}_k$, for $7 \leqslant k \leqslant 12$. The green vertical bars represent a confidence interval of level 95% for each density. The red vertical bars represent the value of $\sqrt{n_k}(W^2_{2,\varepsilon}(\hat{a}_{n_k}, a) - W^2_{2,\varepsilon}(a, a))$.

accepted for the months of November and December, just as in the case where the Euclidean barycenter $\bar{\eta}_6$ is the reference measure. However, the null hypothesis is rejected for the four others months July, August, September and October.

Two samples testing. We finally consider the problem of testing the hypothesis that the distributions of crime locations between two months (from July to December) are equal to the reference measure $a = \bar{\eta}_6$ (Euclidean barycenter over the first 6 months of 2014) using the two samples test statistic based on Sinkhorn divergence for $\varepsilon = 1$ and $\varepsilon = 5$ combined with a bootstrap procedure. We report in Table 1 and Table 2 the estimated



(a) (b)

(c) July (d) August (e) September

(f) October (g) November (h) December

FIGURE III.11. Testing equality of distributions over months for $\varepsilon = 0.3$ with the smoothed Wasserstein barycenter as a reference measure. (a) Euclidean barycenter $\bar{\eta}_6$ (empirical measure corresponding to locations of crime in Chicago for the first 6 months of 2014). (b) Smoothed Wasserstein barycenter $\bar{\eta}_6^\varepsilon$ of the measures $(\hat{\eta}_k)_{1\leqslant k\leqslant 6}$ for $\varepsilon = 0.3$. (c)-(h) Green densities represent the distribution of the bootstrap statistics $\sqrt{n_k}(W_{2,\varepsilon}^2(\hat{a}_{n_k}^*, a) - W_{2,\varepsilon}^2(\hat{a}_{n_k}, a))$ for the last 6 months of 2014, with $a = \bar{\eta}_6^\varepsilon$ and $\hat{a}_{n_k} = \hat{\eta}_k$, for $7 \leqslant k \leqslant 12$. The green vertical bars represent a confidence interval of level 95% for each density. The red vertical bars represent the value of $\sqrt{n_k}(W_{2,\varepsilon}^2(\hat{a}_{n_k}, a) - W_{2,\varepsilon}^2(a, a))$.

p-values corresponding to such tests for all pairs of different months from July to December 2014. For both values of ε the interpretation of the results is similar. They tend to support the hypothesis that the distribution of crime locations is the same when comparing two months among September, October, November and December, and that this distribution is different when the comparison is done with the month of July. The results for August are more difficult to interpret, as it can be concluded that the distribution of crime locations for this month is equal to that of July, September, October and December.

As remarked in [**SM16**], there exists a vast literature for two-sample testing using univariate data. However, in a multivariate setting, it is difficult to consider that there exist standard methods to test the equality of two distributions. We compare the results that have been obtained using our approach with those given by a kernel based test proposed in [**AHT94**] that is implemented in the R package `ks`. The test statistics in [**AHT94**] uses the integrated square distance between two kernel-based density estimates computed from two empirical measures with a data-based choice of bandwidth. We report in Table 3 the p-values corresponding to this test for all pairs of different months from July to December 2014. It can be seen that the p-values obtained with this test are larger than those obtained with our testing procedure. Nevertheless, the conclusions on the equality of distributions of crime locations between different months are roughly the same than previously.

	July	August	September	October	November	December
July	1	0.07	0.04	0.01	$< 10^{-2}$	0.08
August		1	0.16	0.14	0.01	0.12
September			1	0.18	0.07	0.20
October				1	0.06	0.05
November					1	0.10
December						1

TABLE 1. Two samples testing of equal distributions between pairs of different months from July to December using a test statistic based on Sinkhorn divergence for $\varepsilon = 1$ with reference measure $a = \bar{\eta}_6$ (Euclidean barycenter over the first 6 months of 2014). The table reports estimated p-values using a bootstrap procedure for the test statistics $\rho_{n_k,n_\ell}(W^2_{2,\varepsilon}(\hat{a}_{n_k}, \hat{b}_{n_\ell}) - W^2_{2,\varepsilon}(a,a))$ (with $\hat{a}_{n_k} = \hat{\eta}_k$ and $\hat{b}_{n_\ell} = \hat{\eta}_\ell$) for $7 \leqslant k \leqslant \ell \leqslant 12$.

	July	August	September	October	November	December
July	1	0.12	0.04	0.01	$< 10^{-2}$	0.05
August		1	0.25	0.11	0.01	0.10
September			1	0.40	0.06	0.20
October				1	0.06	0.05
November					1	0.06
December						1

TABLE 2. Two samples testing of equal distributions between pairs of different months from July to December using a test statistics based on Sinkhorn divergence for $\varepsilon = 5$ with reference measure $a = \bar{\eta}_6$ (Euclidean barycenter over the first 6 months of 2014). The table reports estimated p-values using a bootstrap procedure for the test statistics $\rho_{n_k,n_\ell}(W^2_{2,\varepsilon}(\hat{a}_{n_k}, \hat{b}_{n_\ell}) - W^2_{2,\varepsilon}(a,a))$ (with $\hat{a}_{n_k} = \hat{\eta}_k$ and $\hat{b}_{n_\ell} = \hat{\eta}_\ell$) for $7 \leqslant k \leqslant \ell \leqslant 12$.

	July	August	September	October	November	December
July	1	0.14	0.04	0.04	0.10	0.09
August		1	0.25	0.06	0.12	0.16
September			1	0.11	0.30	0.30
October				1	0.16	0.14
November					1	0.43
December						1

TABLE 3. Two samples testing with kernel smoothing. The table reports p-values using the kernel based test proposed in [**AHT94**] for testing equality of distributions between different pairs of months from July to December 2014.

Chapter IV

Principal component analysis in the Wasserstein space

This chapter corresponds to the published paper [CSB+18]. The code to reproduce the results of this chapter is available online [1].

IV.1. Introduction

Most datasets describe multivariate data, namely vectors of relevant features that can be modeled as random elements sampled from an unknown distribution. In that setting, Principal Component Analysis (PCA) is certainly the simplest and most widely used approach to reduce the dimension of such datasets. We consider in this chapter the statistical analysis of data sets whose elements are histograms supported on the real line, and also discuss extensions to the general case of probability measures supported on the d-dimensional Euclidean space. Just as with PCA, our main goal in that setting is to compute the principal modes of variation of histograms around their mean element and therefore facilitate the visualization of such datasets. However, since the number, size or locations of significant bins in the histograms of interest may vary from one histogram to another, using standard PCA on histograms (with respect to the Euclidean metric) is bound to fail (see for instance Figure IV.1). For the purpose of learning principal modes of variation, we consider the issue of computing the PCA of histograms with respect to the 2-Wasserstein distance between probability measures.

Previous work in the one-dimensional case. PCA of histograms with respect to the Wasserstein metric has also been proposed in [VIB15] in the context of symbolic data analysis. Their approach consists in computing a standard PCA in the Hilbert space $L^2([0,1])$ of the quantile functions associated to the histograms. Therefore, the algorithm in [VIB15] corresponds to log-PCA of probability measures as suggested in [BGKL17], but it does not solve the problem of convex-constrained PCA in a Hilbert space associated to an exact GPCA in $\mathcal{P}_2(\Omega)$. A related problem, which can be referred to as geodesic regression (considered in [Fle11, Fle13] for data on a Riemannian manifold), has been considered by Jiang et al. in [JLG12] where the authors fit a single geodesic g_t to indexed histograms in order to model nonstationary time series. In the problem of finding principal geodesics, we do not assume that the dataset is indexed.

[1] https://github.com/ecazelles/2017-GPCA-vs-LogPCA-Wasserstein

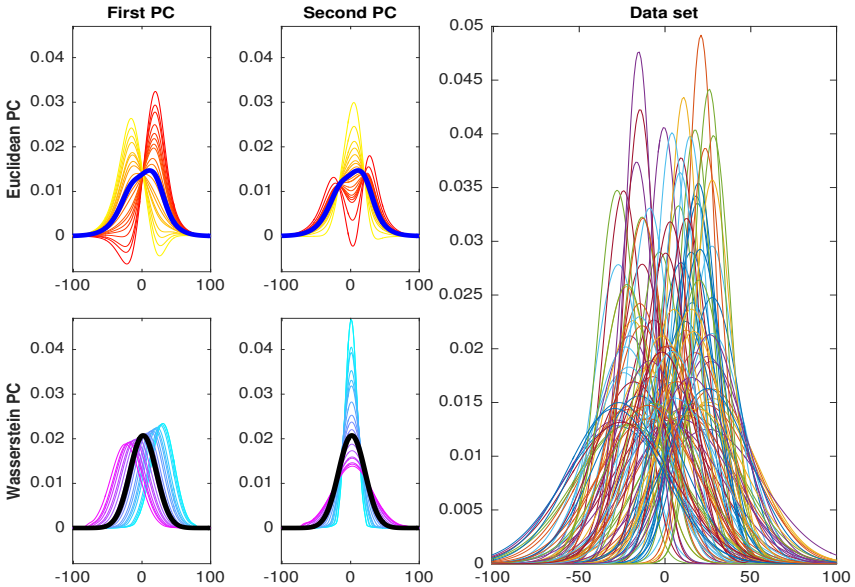


FIGURE IV.1. Synthetic example. (Right) A data set of $n = 100$ Gaussian histograms randomly translated and scaled. (Top-left) Standard PCA of this data set with respect to the Euclidean metric. The Euclidean barycenter of the data set is depicted in blue. (bottom-left) Geodesic PCA with respect to the Wasserstein metric using the iterative geodesic algorithm (IV.13). The black curve represents the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $\mathcal{P}_2(\Omega)$.

PGA and log-PCA on Riemannian manifolds. The method of GPCA proposed in [**BGKL17**] clearly shares similarities with analogs of PCA for data belonging to a Riemannian manifold $\mathcal{M}$ of finite dimension. These methods, generally referred to as Principal Geodesic Analysis (PGA), extend the notion of classical PCA in Euclidean spaces for the purpose of analyzing data belonging to curved Riemannian manifolds (see e.g. [**FLPJ04**, **SLHN10**]). This generalization of PCA proceeds by replacing Euclidean concepts of vector means, lines and orthogonality by the more general notions in Riemannian manifolds of Fréchet mean, geodesics, and orthogonality in tangent spaces. In [**FLPJ04**], linearized PGA, which we refer to as log-PCA, is defined as follows. In a first step, data are mapped to the tangent space $T_{\bar{\nu}}\mathcal{M}$ at their Fréchet mean $\bar{\nu}$ by applying the logarithmic map $\log_{\bar{\nu}}$ to each data point. Then, in a second step, standard PCA in the Euclidean space $T_{\bar{\nu}}\mathcal{M}$ can be applied. This provides a family of orthonormal tangent vectors. Principal components of variation in $\mathcal{M}$ can then be defined by back-projection of these tangent vectors on $\mathcal{M}$ by using the exponential map at $\bar{\nu}$, that is known to parameterize geodesics at least locally. Log-PCA has low computational cost, but this comes at the expense of two simplifications and drawbacks:

(1): First, log-PCA amounts to substituting geodesic distances between data points by the linearized distance in $T_{\bar{\nu}}\mathcal{M}$, which may not always be a good approximation because of the curvature of $\mathcal{M}$, see e.g. [**SLHN10**].

(2): Secondly, the exponential map at the Fréchet mean parameterizes geodesics only locally, which implies that principal components in $\mathcal{M}$ obtained with log-PCA may not be geodesic along the typical range of the dataset.

Numerical approaches to GPCA and log-PCA in the Wasserstein space. Computational methods have been introduced in [SC15, WSB+13] to extend the concepts of PGA on Riemannian manifolds to that of the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures supported on $\mathbb{R}^d$ endowed with the Wasserstein metric. [WSB+13] propose to compute a notion of template measure (using k-means clustering) of a set of discrete probability measures, and to consider then the optimal transport plans from that template measure to each measure in the data set. Computation of the barycentric projection of each optimal transport plan leads to a set of Monge maps over which a standard PCA can be applied, resulting in an orthonormal family of tangent vectors defined on the support of the template measure. Principal components of variation in $\mathbb{R}^d$ can then be obtained through the push-forward operator, namely by moving the mass along these tangent vectors. This approach, analog to log-PCA on Riemannian manifolds, suffers from the main drawbacks mentioned above: for $d > 1$, the linearized Wasserstein distance may be a crude approximation of the Wasserstein distance, and there is no guarantee that the computed tangent vectors parameterize geodesics of sufficient length to summarize most of the variability in the dataset. Losing geodesicity means that the principal components are curves in $\mathcal{P}_2(\mathbb{R}^d)$ along which the mass may not be transported optimally, which may significantly reduce the interpretability of these principal components. A different approach was proposed in [SC15], in which the notion of generalized geodesics in $\mathcal{P}_2(\mathbb{R}^d)$ (see e.g. Chapter 9 in [AGS08]) is used to define a notion of PGA of discrete probability measures. In [SC15], generalized geodesics are parameterized using two velocity fields defined on the support of the Wasserstein barycenter. The authors proposed to minimize directly the distances from the measures in the dataset to these generalized geodesics, by updating these velocity fields which are constrained to be in opposite directions. This approach is more involved computationally than log-PCA, but it avoids some of the drawbacks highlighted above. Indeed, the resulting principal components yield curves in $\mathcal{P}_2(\mathbb{R}^d)$ that are approximately geodesics. Nevertheless, the computational method in [SC15] uses a heuristic projection on the set of optimal velocity fields, which results in a algorithm which has no convergence guarantees. Moreover, by optimizing over generalized geodesics rather than geodesics, it does not solve exactly the problem of computing geodesic PCA in $\mathcal{P}_2(\mathbb{R}^d)$.

In this chapter, we focus on computing an exact GPCA on probability measures supported on $\Omega \subset \mathbb{R}^d$. We mainly focus on the case $d = 1$ (discussing extensions in the last section), which has the advantage that the linearized Wasserstein distance in the tangent space is equal to the Wasserstein distance in the space $\mathcal{P}_2(\Omega)$. The main challenge is thus to obtain principal curves which are geodesics along the range of the dataset.

The first work in this chapter is to propose two fast algorithms for GPCA in $\mathcal{P}_2(\Omega)$. The first algorithm finds iteratively geodesics such that the Wasserstein distance between the dataset and the parameterized geodesic is minimized with respect to $\mathcal{P}_2(\Omega)$. This approach is thus somewhat similar to the one in [SC15]. However, a heuristic barycentric projection is used in [SC15] to remain in the feasible set of constraints during the optimization process. In our approach, we rely on proximal operators of both the objective function and the constraints to obtain an algorithm which is guaranteed to converge to a critical point of the objective function. Moreover, we show that the global minimum of our objective function for the first principal geodesic curve corresponds indeed to the solution of the exact GPCA problem defined in [BGKL17]. While this algorithm is able to find iteratively orthogonal principal geodesics, there is no guarantee that several principal geodesics parameterize a surface which is also geodesic. This is the reason we also propose a second algorithm which computes all the principal geodesics at once by parameterizing a geodesic surface as a convex combination of optimal velocity fields and relaxing the orthogonality constraint between principal geodesics. Both algorithms are a variant of the proximal Forward-Backward algorithm. They converge to a stationary point of the objective function, as shown by recent results in non-convex optimization based on proximal methods [ABS13, OCBP14]. Our

second contribution is a numerical comparison of log-PCA in $\mathcal{P}_2(\Omega)$, as done in [**BGKL17**] (for $d = 1$) or [**WSB+13**], with our approach which solves the exact Wasserstein GPCA problem. Finally, we discuss extensions to the case of probability measures supported on the d-dimensional Euclidean space, providing detailed calculations in the two-dimensional case, and perform computation of GPCA on a two-dimensional example, comparing results with the ones obtained with the log-PCA approach.

In all our experiments, data are normalized in order to have a suitable representation as probability measures. We believe this preprocessing does not affect any useful properties of the histogram datasets considered in the present article, in the same way as centering or whitening are often used as a preprocessing step in many data-analysis tasks. Yet, if the total mass of a given histogram matters for some application, we could consider the use of unbalanced optimal transport [**CPSV18b**, **LMS18**, **CPSV18a**] which provides a distance between unnormalized measures. This generalization is out of the scope of this work and may be an interesting line of research in the future.

In Section IV.2, we provide some background on GPCA in the Wasserstein space $\mathcal{P}_2(\Omega)$, borrowing material from previous work in [**BGKL17**]. Section IV.3 describes log-PCA in $\mathcal{P}_2(\Omega)$, and some of its limitations are discussed. Section IV.4 contains the main results of our work, namely two algorithms for computing GPCA. In Section IV.5, we provide a comparison between GPCA and log-PCA using statistical analysis of real datasets of histograms. In the last Section IV.6 we discuss extensions of our algorithms to the case $d > 1$, and perform GPCA computation on a two-dimensional example, comparing again results with the log-PCA approach. Some perspectives on this work are also given. Finally, various details on the implementation of the algorithms are deferred to technical Appendices.

IV.2. Background on Geodesic PCA in the Wasserstein space

IV.2.1. The pseudo Riemannian structure of the Wasserstein space

In what follows, μ_r denotes a reference measure in $\mathcal{P}_2^{ac}(\Omega)$, whose choice will be discussed later on. The space $\mathcal{P}_2(\Omega)$ has a formal Riemannian structure described, for example, in [**AGS04**]. The tangent space at μ_r is defined as the Hilbert space $\mathbb{L}^2_{\mu_r}(\Omega)$ of real-valued, μ_r-square-integrable functions on Ω, equipped with the inner product $\langle \cdot, \cdot \rangle_{\mu_r}$ defined by $\langle u, v \rangle_{\mu_r} = \int_\Omega u(x)v(x)d\mu_r(x)$, $u, v \in \mathbb{L}^2_{\mu_r}(\Omega)$, and associated norm $\| \cdot \|_{\mu_r}$. We define the exponential and the logarithmic maps at μ_r, as follows.

Definition IV.1. *Let* $\mathrm{id} : \Omega \to \Omega$ *be the identity mapping. The exponential* $\exp_{\mu_r} : \mathbb{L}^2_{\mu_r}(\Omega) \to \mathcal{P}_2(\Omega)$ *and logarithmic* $\log_{\mu_r} : \mathcal{P}_2(\Omega) \to \mathbb{L}^2_{\mu_r}(\Omega)$ *maps are defined respectively as*

$$\exp_{\mu_r}(v) = (\mathrm{id} + v)\#\mu_r \quad \text{and} \quad \log_{\mu_r}(\nu) = F_\nu^- \circ F_{\mu_r} - \mathrm{id}. \tag{IV.1}$$

Contrary to the setting of Riemannian manifolds, the "exponential map" $\exp_{\mu_r}$ defined above is not a local homeomorphism from a neighborhood of the origin in the "tangent space" $\mathbb{L}^2_{\mu_r}(\Omega)$ to the space $\mathcal{P}_2(\Omega)$, see e.g. [**AGS04**]. Nevertheless, it is shown in [**BGKL17**] that $\exp_{\mu_r}$ is an isometry when restricted to the following specific set of functions

$$V_{\mu_r}(\Omega) := \log_{\mu_r}(\mathcal{P}_2(\Omega)) = \left\{ \log_{\mu_r}(\nu) \; ; \; \nu \in \mathcal{P}_2(\Omega) \right\} \subset \mathbb{L}^2_{\mu_r}(\Omega), \tag{IV.2}$$

and that the following results hold (see [**BGKL17**]).

Proposition IV.2. *The subspace* $V_{\mu_r}(\Omega)$ *satisfies the following properties :*

(P1) *the exponential map* $\exp_{\mu_r}$ *restricted to* $V_{\mu_r}(\Omega)$ *is an isometric homeomorphism, with inverse* $\log_{\mu_r}$. *We have hence* $W_2(\nu, \eta) = ||\log_{\mu_r}(\nu) - \log_{\mu_r}(\eta)||_{\mathbb{L}^2_{\mu_r}(\Omega)}$.

(P2) *the set* $V_{\mu_r}(\Omega) := \log_\mu(\mathcal{P}_2(\Omega))$ *is closed and convex in* $\mathbb{L}^2_{\mu_r}(\Omega)$.

(P3) *the space* $V_{\mu_r}(\Omega)$ *is the set of functions* $v \in \mathbb{L}^2_{\mu_r}(\Omega)$ *such that* $T := \mathrm{id} + v$ *is* μ_r*-almost everywhere non decreasing and that* $T(x) \in \Omega$*, for* $x \in \Omega$.

Moreover, it follows, from [BGKL17], that geodesics in $\mathcal{P}_2(\Omega)$ are exactly the image under $\exp_{\mu_r}$ of straight lines in $V_{\mu_r}(\Omega)$. This property is stated in the following lemma.

LEMMA IV.3. *Let* $\gamma : [0,1] \to \mathcal{P}_2(\Omega)$ *be a curve and let* $v_0 := \log_{\mu_r}(\gamma(0))$, $v_1 := \log_{\mu_r}(\gamma(1))$. *Then* $\gamma = (\gamma_t)_{t\in[0,1]}$ *is a geodesic if and only if* $\gamma_t = \exp_{\mu_r}((1-t)v_0 + tv_1)$, *for all* $t \in [0,1]$.

IV.2.2. GPCA for probability measures

Let $\nu_1, \ldots, \nu_n$ be a set of probability measures in $\mathcal{P}_2^{ac}(\Omega)$. Assuming that each ν_i is absolutely continuous simplifies the following presentation, and it is in line with the purpose of statistical analysis of histograms. We define now the notion of (empirical) GPCA of this set of probability measures by following the approach in [BGKL17]. The first step is to choose the reference measure μ_r. The natural choice is then the Wasserstein barycenter $\mu_r = \bar{\nu}$ of the ν_i's, defined in (A.11), which represents an average location in the data around which can be computed the principal sources of geodesic variability. Note that it immediately follows from results in [AC11] that $\bar{\boldsymbol{\nu}} \in \mathcal{P}_2^{ac}(\Omega)$, and that its cdf satisfies

$$F^-_{\bar{\boldsymbol{\nu}}} = \frac{1}{n}\sum_{i=1}^n F^-_{\nu_i}. \tag{IV.3}$$

To introduce the notion of a principal geodesic subspace of the measures $\nu_1, \ldots, \nu_n$, we need to introduce further notation and definitions. Let G be a subset of $\mathcal{P}_2(\Omega)$. The distance between $\mu \in \mathcal{P}_2(\Omega)$ and the set G is $W_2(\nu, G) = \inf_{\lambda\in G} W_2(\nu, \lambda)$, and the average distance between the data and G is taken as

$$\mathrm{D}_W(G) := \frac{1}{n}\sum_{i=1}^n W_2^2(\nu_i, G). \tag{IV.4}$$

DEFINITION IV.4. *Let* K *be some positive integer. A subset* $G \subset \mathcal{P}_2(\Omega)$ *is said to be a geodesic set of dimension* $\dim(G) = K$ *if* $\log_{\mu_r}(G)$ *is a convex set such that the dimension of the smallest affine subspace of* $\mathbb{L}^2_{\mu_r}(\Omega)$ *containing* $\log_{\mu_r}(G)$ *is of dimension* K.

The notion of principal geodesic subspace (PGS) with respect to the reference measure $\mu_r = \bar{\boldsymbol{\nu}}$ can now be presented below.

DEFINITION IV.5. *Let* $\mathrm{CL}(W)$ *be the metric space of nonempty, closed subsets of* $\mathcal{P}_2(\Omega)$, *endowed with the Hausdorff distance, and*

$$\mathrm{CG}_{\bar{\boldsymbol{\nu}},K}(W) = \{G \in \mathrm{CL}(W) \mid \bar{\boldsymbol{\nu}} \in G,\ G \text{ is a geodesic set and } \dim(G) \leqslant K\},\ K \geq 1.$$

A principal geodesic subspace (PGS) of $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ *of dimension* K *with respect to* $\bar{\boldsymbol{\nu}}$ *is a set*

$$G_K \in \operatorname*{arg\,min}_{G\in\mathrm{CG}_{\bar{\boldsymbol{\nu}},K}(W)} \mathrm{D}_W(G) = \operatorname*{arg\,min}_{G\in\mathrm{CG}_{\bar{\boldsymbol{\nu}},K}(W)} \frac{1}{n}\sum_{i=1}^n W_2^2(\nu_i, G). \tag{IV.5}$$

When $K = 1$, searching for the first PGS of $\boldsymbol{\nu}$ simply amounts to search for a geodesic curve $\gamma^{(1)}$ that is a solution of the following optimization problem:

$$\tilde{\gamma}^{(1)} := \operatorname*{arg\,min}_{\gamma} \left\{ \frac{1}{n}\sum_{i=1}^n W_2^2(\nu_i, \gamma) \mid \ \gamma \text{ is a geodesic in } \mathcal{P}_2(\Omega) \text{ passing through } \mu_r = \bar{\boldsymbol{\nu}}. \right\}$$

We remark that this definition of $\tilde{\gamma}^{(1)}$ as the first principal geodesic curve of variation in $\mathcal{P}_2(\Omega)$ is consistent with the usual concept of PCA in a Hilbert space in which geodesic are straight lines.

For a given dimension k, the GPCA problem consists in finding a nonempty closed geodesic subset of dimension k which contains the reference measure μ_r and minimizes Eq. (IV.4). We describe in the next section how we can parameterize such sets G.

IV.2.3. Geodesic PCA parameterization

GPCA can be formulated as an optimization problem in the Hilbert space $L^2_{\bar{\nu}}(\Omega)$. To this end, let us define the functions $\omega_i = \log_{\bar{\nu}}(\nu_i)$ for $1 \leqslant i \leqslant n$ that corresponds to the data mapped in the tangent space. It can be easily checked that this set of functions is centered in the sense that $\frac{1}{n}\sum_{i=1}^n \omega_i = 0$. Note that, in a one-dimensional setting, computing ω_i (mapping of the data to the tangent space) is straightforward since the optimal maps $T_i^* = F^-_{\nu_i} \circ F_{\bar{\nu}}$ between the data and their Fréchet mean are available in a simple and closed form.

For $\mathcal{U} = \{u_1, \dots, u_K\}$ a collection of $K \geqslant 1$ functions belonging to $\mathbb{L}^2_{\bar{\nu}}(\Omega)$, we denote by $\mathrm{Sp}(\mathcal{U})$ the subspace spanned by $u_1, \dots, u_K$. Defining $\Pi_{\mathrm{Sp}(\mathcal{U})} v$ as the projection of $v \in \mathbb{L}^2_{\bar{\nu}}(\Omega)$ onto $\mathrm{Sp}(\mathcal{U})$, and $\Pi_{\mathrm{Sp}(\mathcal{U})\cap V_{\bar{\nu}}(\Omega)} v$ as the projection of v onto the closed convex set $\mathrm{Sp}(\mathcal{U}) \cap V_{\bar{\nu}}(\Omega)$, then we have

Proposition IV.6. *Let $\omega_i = \log_{\bar{\nu}}(\nu_i)$ for $1 \leqslant i \leqslant n$, and $\mathcal{U}^* = \{u_1^*, \dots, u_k^*\}$ be a minimizer of*

$$\frac{1}{n}\sum_{i=1}^n \|\omega_i - \Pi_{\mathrm{Sp}(\mathcal{U})\cap V_{\bar{\nu}}(\Omega)}\omega_i\|^2_{\bar{\nu}}, \tag{IV.6}$$

over orthonormal sets $\mathcal{U} = \{u_1, \dots, u_K\}$ of functions in $\mathbb{L}^2_{\bar{\nu}}(\Omega)$ of dimension K (namely such that $\langle u_j, u_{j'}\rangle_{\bar{\nu}} = 0$ if $j \neq j'$ and $\|u_j\|_{\bar{\nu}} = 1$). If we let

$$G_{\mathcal{U}^*} := \exp_{\bar{\nu}}(\mathrm{Sp}(\mathcal{U}^*) \cap V_{\bar{\nu}}(\Omega)),$$

then $G_{\mathcal{U}^*}$ is a principal geodesic subset (PGS) of dimension k of the measures $\nu_1, \dots, \nu_n$, meaning that $G_{\mathcal{U}^*}$ belongs to the set of minimizers of the optimization problem (IV.5).

Proof. For $v \in \mathbb{L}^2_{\bar{\nu}}(\Omega)$ and a subset $C \in \mathbb{L}^2_{\bar{\nu}}(\Omega)$, we define $d_{\bar{\nu}}(v, C) = \inf_{u\in C} \|v - u\|_{\bar{\nu}}$. Remark that $\sum_i \omega_i = 0$. Hence by Proposition 3.3 in [**BGKL17**], if $\mathcal{U}^*$ minimizes

$$\frac{1}{n}\sum_{i=1}^n d^2_{\bar{\nu}}(\omega_i, \mathrm{Sp}(\mathcal{U}^*) \cap V_{\bar{\nu}}(\Omega)) = \frac{1}{n}\sum_{i=1}^n \|\omega_i - \Pi_{\mathrm{Sp}(\mathcal{U}^*)\cap V_{\bar{\nu}}(\Omega)}\omega_i\|^2_{\bar{\nu}},$$

then $\mathrm{Sp}(\mathcal{U}^*)\cap V_{\bar{\nu}}(\Omega) \in \arg\min_C \frac{1}{n}\sum_{i=1}^n d^2_{\bar{\nu}}(\omega_i, C)$, where C is taken over all nonempty, closed, convex set of $V_{\bar{\nu}}(\Omega)$ such that $dim(C) \leq K$ and $0 \in C$. By Proposition 4.3 in [**BGKL17**], and since $\log_{\bar{\nu}}(\bar{\nu}) = 0$, we can conclude that G^* is a geodesic subset of dimension K which minimizes (IV.4). □

Thanks to Proposition IV.6, it follows that GPCA in $\mathcal{P}_2(\Omega)$ corresponds to a mapping of the data into the Hilbert space $\mathbb{L}^2_{\bar{\nu}}(\Omega)$ which is followed by a PCA in $\mathbb{L}^2_{\bar{\nu}}(\Omega)$ that is constrained to lie in the convex and closed subset $V_{\bar{\nu}}(\Omega)$. This has to be interpreted as a geodesicity constraint coming from the definition of a PGS in $\mathcal{P}_2(\Omega)$. Because this geodesicity constraint is nontrivial to implement, recent works about geodesic PCA in $\mathcal{P}_2(\Omega)$ relied on a heuristic projection on the set of optimal maps [**SC15**], or relaxed the geodesicity constraint by solving a linearized PGA [**WSB+13**, **BGKL17**]. We describe the latter approach in the following section.

IV.3. The log-PCA approach

For data in a Riemannian manifold, we recall that log-PCA consists in solving a linearized version of the PGA problem by mapping the whole data set to the tangent space at the Fréchet mean through the logarithmic map [**FLPJ04**]. This approach is computationally attractive since it boils down to computing a standard PCA. [**WSB+13**] used this idea to define a linearized PGA in the Wasserstein space $W_2(\mathbb{R}^d)$, by defining the logarithmic map of a probability measure as the barycentric projection of an optimal transport plan with respect to a template measure. This approach has the two drawbacks (1) and (2) of log-PCA mentioned in the introduction. A third limitation inherent to the Wasserstein space is that when this template probability measure is discrete, the logarithmic map cannot be defined straightforwardly as there is no guarantee about the existence of an optimal map solution of the optimal transport problem. This is why the authors of [**WSB+13**] had to compute the barycentric projection of each optimal transport plan, which is obtained by simply averaging the locations of the split mass defined by this plan. This averaging process is however lossy as distinct probability measures can have the same barycentric projection.

We consider as usual a subset $\Omega \subset \mathbb{R}$. In this setting, $\mathcal{P}_2(\Omega)$ is a flat space as shown by the isometry property (P1) of Proposition IV.2. Moreover, if the Wasserstein barycenter $\bar{\boldsymbol{\nu}}$ is assumed to be absolutely continuous, then Definition IV.1 shows that the logarithmic map at $\bar{\boldsymbol{\nu}}$ is well defined everywhere. Under such an assumption, log-PCA in $\mathcal{P}_2(\Omega)$ corresponds to the following steps:

(1) compute the log-maps (see Definition IV.1) $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, $i = 1, \dots, n$,
(2) perform the PCA of the projected data $\omega_1, \cdots, \omega_n$ in the Hilbert space $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ to obtain K orthogonal directions $\tilde{u}_1, \dots, \tilde{u}_K$ in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ of principal variations,
(3) recover a principal subspace of variation in $\mathcal{P}_2(\Omega)$ with the exponential map $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}))$ of the principal eigenspace $\mathrm{Sp}(\tilde{\mathcal{U}})$ in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ spanned by $\tilde{u}_1, \dots, \tilde{u}_K$.

For specific datasets, log-PCA in $\mathcal{P}_2(\Omega)$ may be equivalent to GPCA, in the sense that the set $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega))$ is a principal geodesic subset of dimension K of the measures $\nu_1, \dots, \nu_n$, as defined by (IV.5). Informally, this case corresponds to the setting where the data are sufficiently concentrated around their Wasserstein barycenter $\bar{\boldsymbol{\nu}}$ (we refer to Remark 3.5 in [**BGKL17**] for further details). However, carrying out a PCA in the tangent space of $W_2(\mathbb{R})$ at $\bar{\boldsymbol{\nu}}$ is a relaxation of the convex-constrained GPCA problem (IV.6), where the elements of the sought principal subspace do not need to be in $V_{\bar{\boldsymbol{\nu}}}$. Indeed, standard PCA in the Hilbert space $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ amounts to finding $\tilde{\mathcal{U}} = \{\tilde{u}_1, \dots, \tilde{u}_K\}$ minimizing,

$$\frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \Pi_{\mathrm{Sp}(\mathcal{U})} \omega_i\|^2_{\bar{\boldsymbol{\nu}}}, , \tag{IV.7}$$

over orthonormal sets $\mathcal{U} = \{u_1, \dots, u_k\}$ of functions in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$. It is worth noting that the three steps of log-PCA in $\mathcal{P}_2(\Omega)$ are simple to implement and fast to compute, but that performing log-PCA or GPCA (IV.6) in $\mathcal{P}_2(\Omega)$ are not necessarily equivalent.

Log-PCA is generally used for two main purposes. The first one is to obtain a low dimensional representation of each data measure $\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ through the coefficients $\langle \omega_i, \tilde{u}_k \rangle_{L^2_{\bar{\boldsymbol{\nu}}}}$. From this low dimensional representation, the measure $\nu_i \in \mathcal{P}_2(\Omega)$ can be approximated through the exponential mapping $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\mathcal{U})} \omega_i)$. The second one is to visualize each mode of variation in the dataset, by considering the evolution of the curve $t \mapsto \exp_{\bar{\boldsymbol{\nu}}}(t\tilde{u}_k)$ for each $\tilde{u}_k \in \tilde{\mathcal{U}}$.

However, relaxing the convex-constrained GPCA problem (IV.6) when using log-PCA results in several issues. Indeed, as shown in the following paragraphs, not taking into account this geodesicity constraint makes difficult the computation and interpretation of $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}))$ as a principal subspace of variation, which may limit its use for data analysis.

Numerical implementation of pushforward operators. A first downside to the log-PCA approach is the difficulty of the numerical implementation of the pushforward operator in the exponential map $\exp_{\bar{\boldsymbol{\nu}}}(v) = (\mathrm{id} + v)\#\bar{\boldsymbol{\nu}}$ when the mapping $\mathrm{id} + v$ is not a strictly increasing function for a given vector $v \in \mathrm{Sp}(\tilde{\mathcal{U}})$. This can be shown with the following proposition, which provides a formula for computing the density of a pushforward operator.

PROPOSITION IV.7. *(Density of the pushforward) Let $\mu \in \mathcal{P}_2(\mathbb{R})$ be an absolutely continuous measure with density ρ (that is possibly supported on an interval $\Omega \subset \mathbb{R}$). Let $T : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $|T'(x)| > 0$ for almost every $x \in \mathbb{R}$, and define $\nu = T\#\mu$. Then, ν admits a density g given by,*

$$g(y) = \sum_{x \in T^{-1}(y)} \frac{\rho(x)}{|T'(x)|}, \; y \in \mathbb{R}. \tag{IV.8}$$

When T is injective, this simplifies to,

$$g(y) = \frac{\rho(T^{-1}(y))}{|T'(T^{-1}(y))|}. \tag{IV.9}$$

PROOF. Under the assumptions made on T, the coarea formula (which is a more general form of Fubini's theorem, see e.g. [**KP08**] Corollary 5.2.6 or [**EG15**] Section 3.4.3) states that, for any measurable function $h : \mathbb{R} \to \mathbb{R}$, one has

$$\int_{\mathbb{R}} h(x)|T'(x)|dx = \int_{\mathbb{R}} \sum_{x \in T^{-1}(y)} h(x)dy. \tag{IV.10}$$

Let B a Borel set and choose $h(x) = \frac{\rho(x)}{|T'|}\mathbf{1}_{T^{-1}(B)}, \; x$. Hence, using (IV.10), one obtains that

$$\int_{T^{-1}(B)} \rho(x)dx = \int_{\mathbb{R}} \sum_{x \in T^{-1}(y)} \frac{\rho(x)}{|T'(x)|}\mathbf{1}_{T^{-1}(B)}(x)dy = \int_B \sum_{x \in T^{-1}(y)} \frac{\rho(x)}{|T'(x)|}dy.$$

The definition of the pushforward $\nu(B) = \mu(T^{-1}(B))$ then completes the proof. □

The numerical computation of formula (IV.8) or (IV.9) is not straightforward. When T is not injective, computation of the formula (IV.8) must be done carefully by partitioning the domain of T in sets on which T is injective. Such a partitioning depends on the method of interpolation for estimating a continuous density ρ from a finite set of its values on a grid of reals. More importantly, when $T'(x)$ is very small, $\frac{\rho(x)}{|T'(x)|}$ may become very irregular and the density of $\nu = T\#\mu$ may exhibit large peaks, see Figure IV.2 for an illustrative example.

Pushforward of the barycenter outside the support Ω. A second downside of log-PCA in $\mathcal{P}_2(\Omega)$ is that the range of the mapping $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i$ may be larger than the interval Ω. This implies that the density of the pushforward of the Wasserstein barycenter $\bar{\boldsymbol{\nu}}$ by this mapping, namely $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)$, may have a support which is not included in Ω. This issue may be critical when trying to estimate the measure $\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ by its projected measure $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)$. For example, in a dataset of histograms with bins necessarily containing only nonnegative reals, a projected distribution with positive mass on negative reals would be hard to interpret.

A higher Wasserstein reconstruction error. Finally, relaxing the geodesicity constraint (IV.6) may actually increase the Wasserstein reconstruction error with respect to the Wasserstein distance. To state this issue more clearly, we define the reconstruction error of log-PCA as

$$\tilde{r}(\tilde{\mathcal{U}}) = \frac{1}{n}\sum_{i=1}^{n} W_2^2\big(\nu_i, \exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)\big). \tag{IV.11}$$

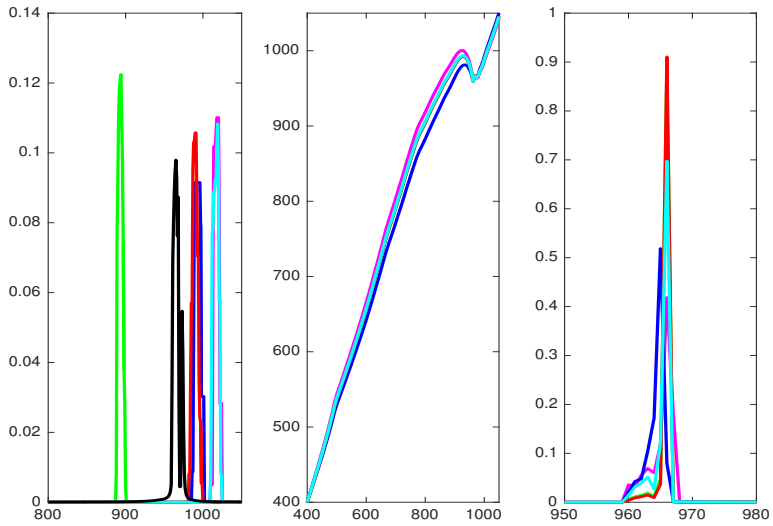


FIGURE IV.2. (Left) Distribution of the total precipitation (mm) collected in a year in $1 \leqslant i \leqslant 5$ stations among 60 in China - Source : Climate Data Bases of the People's Republic of China 1841-1988 downloaded from http://cdiac.ornl.gov/ndps/tr055.html. The black curve is the density of the Wasserstein barycenter of the 60 stations. (Middle) Mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\tilde{u}_2)}\omega_i$ obtained from the projections of these 5 distributions onto the second eigenvector $\tilde{u}_2$ given by log-PCA of the whole dataset. (Right) Pushforward $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_2)}\omega_i) = T_i \# \bar{\boldsymbol{\nu}}$ of the Wasserstein barycenter $\bar{\boldsymbol{\nu}}$ for each $1 \leqslant i \leqslant 5$. As the derivative T_i' take very small values, the densities of the pushforward barycenter $T_i \# \bar{\boldsymbol{\nu}}$ for $1 \leqslant i \leqslant 5$ exhibit large peaks (between 0.4 and 0.9) whose amplitude is beyond the largest values in the original data set (between 0.08 and 0.12).

and the reconstruction error of GPCA as

$$r(\mathcal{U}^*) = \frac{1}{n}\sum_{i=1}^{n} W_2^2\big(\nu_i, \exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\mathcal{U}^*)\cap V_{\bar{\boldsymbol{\nu}}}(\Omega)}\omega_i)\big). \tag{IV.12}$$

where $\mathcal{U}^*$ is a minimizer of (IV.6). Note that in (IV.11), the projected measures $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i)$ might have a support that lie outside Ω. Hence, the Wasserstein distance W_2 in (IV.11) has to be understood for measures supported on $\mathbb{R}$ (with the obvious extension to zero of ν_i outside Ω).

The Wasserstein reconstruction error $\tilde{r}(\tilde{\mathcal{U}})$ of log-PCA is the sum of the Wasserstein distances of each data point ν_i to a point on the surface $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\tilde{\mathcal{U}}))$ which is given by the decomposition of ω_i on the orthonormal basis $\tilde{\mathcal{U}}$. However, by Proposition IV.2, the isometry property (P1) only holds between $\mathcal{P}_2(\mathbb{R})$ and the convex subset $V_{\bar{\boldsymbol{\nu}}} \subset L^2_{\bar{\boldsymbol{\nu}}}(\mathbb{R})$. Therefore, we may not have $W_2\big(\nu_i, \exp_{\bar{\boldsymbol{\nu}}}\big(\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i\big)\big) = \|\omega_i - \Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i\|_{\bar{\boldsymbol{\nu}}}$ as $\Pi_{\mathrm{Sp}(\tilde{\mathcal{U}})}\omega_i$ is a function belonging to $L^2_{\bar{\boldsymbol{\nu}}}(\mathbb{R})$ which may not necessarily be in $V_{\bar{\boldsymbol{\nu}}}$. In this case, the minimal Wasserstein distance between ν_i and the surface $\exp_{\bar{\boldsymbol{\nu}}}(\mathrm{Sp}(\mathcal{U}^*))$ is not equal to $\|\omega_i - \Pi_{\mathrm{Sp}(\mathcal{U})}\omega_i\|_{\bar{\boldsymbol{\nu}}}$, and this leads to situations where $\tilde{r}(\tilde{\mathcal{U}}) > r(\mathcal{U}^*)$ as illustrated in Figure IV.3.

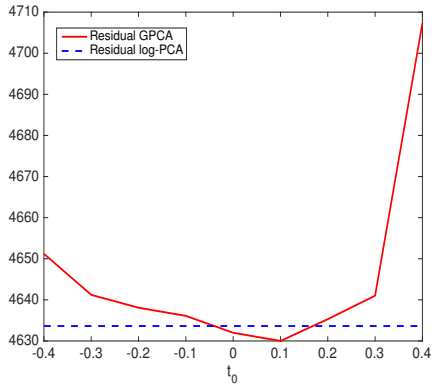


FIGURE IV.3. Comparison of the Wasserstein reconstruction error between GPCA and log-PCA on the synthetic dataset displayed in Figure IV.1 for the first component, with an illustration of the role of the parameter t_0 in (IV.14).

IV.4. Two algorithmic approaches for GPCA in $\mathcal{P}_2(\Omega)$, for $\Omega \subset \mathbb{R}$

In this section, we introduce two algorithms which solve some of the issues of log-PCA that have been raised in Section IV.3. First, the output of the proposed algorithms guarantees that the computation of mappings to pushforward the Wassertein barycenter to approximate elements in the data set are strictly increasing (that is they are optimal). As a consequence, the resulting pushforward density behaves numerically much better. Secondly, the geodesic curve or surface are constrained to lie in $\mathcal{P}_2(\Omega)$, implying that the projections of the data are distributions whose supports do not lie outside Ω.

IV.4.1. Iterative geodesic approach

In this section, we propose an algorithm to solve a variant of the convex-constrained GPCA problem (IV.6). Rather than looking for a geodesic subset of a given dimension which fits well the data, we find iteratively orthogonal principal geodesics (i.e. geodesic set of dimension one). Assuming that that we already know a subset $\mathcal{U}^{k-1} \subset \mathbb{L}^2_{\boldsymbol{\nu}}(\Omega)$ containing $k-1$ orthogonal principal directions $\{u_l\}_{l=1}^{k-1}$ (with $\mathcal{U}^0 = \emptyset$), our goal is to find a new direction $u_k \in \mathbb{L}^2_{\boldsymbol{\nu}}(\Omega)$ of principal variation by solving the optimization problem:

$$u_k \in \underset{v \perp \mathcal{U}^{k-1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \Pi_{\mathrm{Sp}(v) \cap V_{\boldsymbol{\nu}}(\Omega)} \omega_i\|^2_{\boldsymbol{\nu}}, \tag{IV.13}$$

where the infimum above is taken over all $v \in \mathbb{L}^2_{\boldsymbol{\nu}}(\Omega)$ belonging to the orthogonal of $\mathcal{U}^{k-1}$. This iterative process is not equivalent to the GPCA problem (IV.6), with the exception of the first principal geodesic ($k = 1$). Nevertheless, it computes principal subsets $\mathcal{U}^k$ of dimension k such that the projections of the data onto every direction of principal variation lie in the convex set $V_{\boldsymbol{\nu}}$.

The following proposition is the key result to derive an algorithm to solve (IV.13) on real data.

PROPOSITION IV.8. *Introducing the characteristic function of the convex set $V_{\boldsymbol{\nu}}(\Omega)$ as:*

$$\chi_{V_{\boldsymbol{\nu}}(\Omega)}(v) = \begin{cases} 0 & \text{if } v \in V_{\boldsymbol{\nu}}(\Omega) \\ +\infty & \text{otherwise} \end{cases}$$

the optimization problem (IV.13) *is equivalent to*

$$u_k = \underset{v \perp \mathcal{U}^{k-1}}{\arg\min} \min_{t_0 \in [-1;1]} H(t_0, v), \tag{IV.14}$$

where

$$H(t_0, v) = \frac{1}{n} \sum_{i=1}^{n} \min_{t_i \in [-1;1]} \|\omega_i - (t_0 + t_i)v\|_{\bar{\boldsymbol{\nu}}}^2 + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0 - 1)v) + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0 + 1)v). \tag{IV.15}$$

Proof. We first observe that $\Pi_{\mathrm{Sp}(u) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)} \omega_i = \beta_i u$, with $\beta_i \in \mathbb{R}$ and $\beta_i u \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$. Hence, for u_k solution of (IV.13), we have:

$$\frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \Pi_{\mathrm{Sp}(u_k) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)} \omega_i\|_{\bar{\boldsymbol{\nu}}}^2 = \frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \beta_i u_k\|_{\bar{\boldsymbol{\nu}}}^2.$$

such that $\beta_i \in \mathbb{R}$ and $\beta_i u_k \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$ for all $i \in \{1, \dots, n\}$. We take $M \in \mathrm{argmax}_{1 \leqslant i \leqslant n} \beta_i$ and $m \in \underset{1 \leqslant i \leqslant n}{\arg\min} \beta_i$. Without loss of generality, we can assume that $\beta_M > 0$ and $\beta_m < 0$. We then define $v = (\beta_M - \beta_m)u_k/2$ and $t_0 = (\beta_M + \beta_m)/(\beta_M - \beta_m)$, that checks $|t_0| < 1$. Hence, for all $i = 1, \dots, n$, there exists $t_i \in [-1; 1]$ such that: $\beta_i u_k = (t_0 + t_i)v \in V_{\bar{\boldsymbol{\nu}}}$. In particular, one has $t_M = 1$ and $t_m = -1$, which means that $(t_0 \pm 1)v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$. Reciprocally, $(t_0 \pm 1)v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$ ensures us by convexity of $V_{\bar{\boldsymbol{\nu}}}(\Omega)$ that for all $t_i \in [-1; 1]$, $(t_0 + t_i)v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$. □

Proposition IV.8 may be interpreted as follows. For a given $t_0 \in [-1; 1]$, let $v \in \perp \mathcal{U}^{k-1}$ satisfying $(t_0 - 1)v \in V_{\bar{\boldsymbol{\nu}}}$ and $(t_0 + 1)v \in V_{\bar{\boldsymbol{\nu}}}$. Then, if one defines the curve

$$g_t(t_0, v) = (\mathrm{id} + (t_0 + t)v) \# \bar{\boldsymbol{\nu}} \text{ for } t \in [-1; 1], \tag{IV.16}$$

it follows, from Lemma IV.3, that $(g_t(t_0, v))_{t \in [-1;1]}$ is a geodesic since it can be written as $g_t(t_0, v) = \exp_{\bar{\boldsymbol{\nu}}}((1 - u)w_0 + uw_1), u \in [0, 1]$ with $w_0 = (t_0 - 1)v$, $w_1 = (t_0 + 1)v$, $u = (t + 1)/2$, and with w_0 and w_1 belonging to $V_{\bar{\boldsymbol{\nu}}}$ for $|t_0| < 1$. From the isometry property (P1) in Proposition IV.2, one has

$$\min_{t_i \in [-1;1]} \|\omega_i - (t_0 + t_i)v\|_{\bar{\boldsymbol{\nu}}}^2 = \min_{t_i \in [-1;1]} W_2^2(\nu_i, g_{t_i}(v)), \tag{IV.17}$$

and thus the objective function $H(t_0, v)$ in (IV.14) is equal to the sum of the squared Wasserstein distances between the data set and the geodesic curve $(g_t(t_0, v))_{t \in [-1;1]}$.

The choice of the parameter t_0 corresponds to the location of the mid-point of the geodesic $g_t(t_0, v)$, and it plays a crucial role. Indeed, the minimization of $H(t_0, v)$ over $t_0 \in [-1; 1]$ in (IV.14) cannot be avoided to obtain an optimal Wasserstein reconstruction error. This is illustrated by the Figure IV.3, where the Wasserstein reconstruction error $\tilde{r}(\tilde{\mathcal{U}})$ of log-PCA (see equation (IV.11)) is compared with the ones of GPCA, for different t_0, obtained for $k = 1$ as

$$t_0 \in [-1; 1] \mapsto H(t_0, u_1^{t_0})$$

with $u_1^{t_0} = \underset{v}{\arg\min} H(t_0, v)$. This shows that GPCA can lead to a better low dimensional data representation than log-PCA in term of Wasserstein residual errors.

IV.4.2. Geodesic surface approach

Once a family of vectors $(v_1, \cdots, v_k)$ has been found through the minimization of problem (IV.13), one can recover a geodesic subset of dimension k by considering all convex combinations of the vectors $((t_0^1 + 1)v_1, (t_0^1 - 1)v_1, \cdots, (t_0^k + 1)v_k, (t_0^k - 1)v_k)$. However, this subset may not be a solution of (IV.6) since we have no guarantee that a data point ν_i is actually close to this geodesic subset. This discussion suggests that we may consider solving the GPCA problem (IV.6) over geodesic set parameterized as in Proposition IV.13. In order to find principal geodesic subsets which are close to the data set, we consider a family

$V^K = (v_1, \cdots, v_K)$ of linearly independant vectors and $\mathbf{t}_0^K = (t_0^1, \cdots, t_0^K) \in [-1,1]^K$ such that $(t_0^1 - 1)v_1, (t_0^1 + 1)v_1, \cdots, (t_0^K - 1)v_K, (t_0^K + 1)v_K$ are all in $V_{\bar{\boldsymbol{\nu}}}$. Convex combinations of the latter family provide a parameterization of a geodesic set of dimension K by taking the exponential map $\exp_{\bar{\boldsymbol{\nu}}}$ of

$$\hat{V}_{\bar{\boldsymbol{\nu}}}(V^K, \mathbf{t}_0^K) = \left\{ \sum_{k=1}^{K} (\alpha_k^+(t_0^k + 1) + \alpha_k^-(t_0^k - 1))v_k, \alpha^{\pm} \in A \right\} \tag{IV.18}$$

where A is a simplex constraint: $\alpha^{\pm} \in A \Leftrightarrow \alpha_k^+, \alpha_k^- \geqslant 0$ and $\sum_{k=1}^K (\alpha_k^+ + \alpha_k^-) \leqslant 1$. We hence substitute the general sets $\mathrm{Sp}(\mathcal{U}) \cap V_{\bar{\boldsymbol{\nu}}}(\Omega)$ in the definition of the GPCA problem (IV.6) to obtain,

$$\begin{aligned}
(u_1, \cdots, u_K) &= \underset{V^K, \mathbf{t}_0^K}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|\omega_i - \Pi_{\hat{V}_{\bar{\boldsymbol{\nu}}}(V^K, t^K)} \omega_i\|_{\bar{\boldsymbol{\nu}}}^2, \\
&= \underset{v_1, \cdots, v_K}{\arg\min} \min_{\mathbf{t}_0^K \in [-1,1]^K} \frac{1}{n} \sum_{i=1}^{n} \min_{\alpha_i^{\pm} \in A} \|\omega_i - \sum_{k=1}^{K} (\alpha_{ik}^+(t_0^k + 1) + \alpha_{ik}^-(t_0^k - 1))v_k\|_{\bar{\boldsymbol{\nu}}}^2 \\
&\quad + \sum_{k=1}^{K} \left(\chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0^k + 1)v_k) + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0^k - 1)v_k) \right) + \sum_{i=1}^{n} \chi_A(\alpha_i^{\pm}).
\end{aligned} \tag{IV.19}$$

IV.4.3. Discretization and Optimization

In this section we follow the framework of the iterative geodesic algorithm. We provide additional details when the optimization procedure of the geodesic surface approach differs from the iterative one.

IV.4.3.1. Discrete optimization problem

Let $\Omega = [a; b]$ be a compact interval, and consider its discretization over N points $a = x_1 < x_2 < \cdots < x_N = b$, $\Delta_j = x_{j+1} - x_j$, $j = 1, \ldots, N-1$. We recall that the functions $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$ for $1 \leqslant i \leqslant n$ are elements of $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ which correspond to the mapping of the data to the tangent space at the Wasserstein barycenter $\bar{\boldsymbol{\nu}}$. In what follows, for each $1 \leqslant i \leqslant n$, the discretization of the function ω_i over the grid reads $\mathbf{w}_i = (w_i^j)_{j=1}^N \in \mathbb{R}^N$. We also recall that $\chi_A(u)$ is the characteristic function of a given set A, namely $\chi_A(u) = 0$ if $u \in A$ and $+\infty$ otherwise. Finally, the space $\mathbb{R}^N$ is understood to be endowed with the following inner product and norm $\langle \mathbf{u}, \mathbf{v} \rangle_{\bar{\boldsymbol{\nu}}} = \sum_{j=1}^N \bar{\boldsymbol{f}}(x_j) u_j v_j$ and $\|\mathbf{v}\|_{\bar{\boldsymbol{\nu}}}^2 = \langle \mathbf{v}, \mathbf{v} \rangle_{\bar{\boldsymbol{\nu}}}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, where $\bar{\boldsymbol{f}}$ denotes the density of the measure $\bar{\boldsymbol{\nu}}$. Let us now suppose that we have already computed $k-1$ orthogonal (in the sense $\langle \mathbf{u}, \mathbf{v} \rangle_{\bar{\boldsymbol{\nu}}} = 0$) vectors $\mathbf{u}_1, \cdots \mathbf{u}_{k-1}$ in $\mathbb{R}^N$ which stand for the discretization of orthonormal functions $u_1, \ldots, u_{k-1}$ in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ over the grid $(x_j)_{j=1}^N$.

Discretizing problem (IV.14) for a fixed $t_0 \in]-1; 1[$, our goal is to find a new direction $\mathbf{u}_k \in \mathbb{R}^N$ of principal variations by solving the following problem over all $\mathbf{v} = \{v_j\}_{j=1}^N \in \mathbb{R}^N$:

$$\mathbf{u}_k \in \underset{\mathbf{v} \in \mathbb{R}^N}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(\min_{t_i \in [-1;1]} \|\mathbf{w}_i - (t_0 + t_i)\mathbf{v}\|_{\bar{\boldsymbol{\nu}}}^2 \right) + \chi_S(\mathbf{v}) + \chi_V((t_0 - 1)\mathbf{v}) + \chi_V((t_0 + 1)\mathbf{v}), \tag{IV.20}$$

where $S = \{\mathbf{v} \in \mathbb{R}^N \text{ s.t. } \langle \mathbf{v}, \mathbf{u}_l \rangle_{\bar{\boldsymbol{\nu}}} = 0, l = 1 \cdots k-1\}$ is a convex set that deals with the orthogonality constraint $\mathbf{v} \perp \mathcal{U}^{k-1}$ and V corresponds to the discretization of the constraints contained in $V_{\bar{\boldsymbol{\nu}}}(\Omega)$. From Proposition IV.2 (P3), we have that $\forall v \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$, $T := \mathrm{id} + v$ is non decreasing and $T(x) \in \Omega$ for all $x \in \Omega$. Hence the discrete convex set V is defined as

$$V = \{\mathbf{v} \in \mathbb{R}^N \text{ s.t. } x_{j+1} + v_{j+1} \geqslant x_j + v_j,\ j = 1 \cdots N-1 \text{ and } x_j + v_j \in [a; b],\ j = 1 \cdots N\}$$

and can be rewritten as the intersection of two convex sets dealing with each constraint separately.

PROPOSITION IV.9. *One has*

$$\chi_V((t_0-1)\mathbf{v}) + \chi_V((t_0+1)\mathbf{v}) = \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}),$$

where the convex sets D and E respectively deal with the domain constraints $x_j+(t_0+1)v_j \in [a;b]$ and $x_j+(t_0-1)v_j \in [a;b]$, i.e.:

$$D = \{\mathbf{v} \in \mathbb{R}^N, \text{ s.t. } m_j \leqslant v_j \leqslant M_j\}, \tag{IV.21}$$

with $m_j = \max\left(\frac{a-x_j}{t_0+1}, \frac{b-x_j}{t_0-1}\right)$ and $M_j = \min\left(\frac{a-x_j}{t_0-1}, \frac{b-x_j}{t_0+1}\right)$, and the non decreasing constraint of $id + (t_0 \pm 1)\mathbf{v}$:

$$E = \{\mathbf{z} \in \mathbb{R}^N \text{ s.t. } -1/(t_0+1) \leqslant z_j \leqslant 1/(1-t_0)\}. \tag{IV.22}$$

with the differential operator $K : \mathbb{R}^N \to \mathbb{R}^N$ computing the discrete derivative of $\mathbf{v} \in \mathbb{R}^n$ as

$$(K\mathbf{v})_j = \begin{cases} (v_{j+1}-v_j)/(x_{j+1}-x_j) & \text{if } 1 \leqslant j < N \\ 0 & \text{if } j = N, \end{cases} \tag{IV.23}$$

Having D and E both depending on t_0 is not an issue since problem (IV.20) is solved for fixed t_0.

Introducing $\mathbf{t} = \{t_i\}_{i=1}^n \in \mathbb{R}^n$, problem (IV.20) can be reformulated as:

$$\min_{\mathbf{v}\in\mathbb{R}^N} \min_{\mathbf{t}\in\mathbb{R}^n} J(\mathbf{v},\mathbf{t}) := \underbrace{\sum_{i=1}^n \|\mathbf{w}_i - (t_0+t_i)\mathbf{v}\|_{\boldsymbol{\nu}}^2}_{F(\mathbf{v},\mathbf{t})} + \underbrace{\chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) + \chi_{B_1^n}(\mathbf{t})}_{G(\mathbf{v},\mathbf{t})}. \tag{IV.24}$$

where B_1^n is the L^∞ ball of $\mathbb{R}^n$ with radius 1 dealing with the constraint $t_i \in [-1;1]$. Notice that F is differentiable but non-convex in $(\mathbf{v},\mathbf{t})$ and G is non-smooth and convex.

Geodesic surface approach. For fixed $(t_0^1,\dots,t_0^K) \in \mathbb{R}^K$ and $\alpha^\pm = \{\alpha_k^+, \alpha_k^-\}_{k=1}^K$, the discretized version of (IV.19) is then

$$\min_{\mathbf{v_1},\dots,\mathbf{v_K}\in\mathbb{R}^N} \min_{\alpha_\mathbf{1}^\pm,\dots,\alpha_\mathbf{n}^\pm \in \mathbb{R}^{2K}} F'(\mathbf{v},\mathbf{t}) + G'(\mathbf{v},\mathbf{t}), \tag{IV.25}$$

where $F'(\mathbf{v},\mathbf{t}) = \sum_{i=1}^n \|\mathbf{w}_i - \sum_{k=1}^K (\alpha_{ik}^+(t_0^k+1) + \alpha_{ik}^-(t_0^k-1))\mathbf{v_k}\|_{\boldsymbol{\nu}}^2$ is still non-convex and differentiable, $G'(\mathbf{v},\mathbf{t}) = \sum_{k=1}^K \left(\chi_E(K\mathbf{v_k}) + \chi_{D_k}(\mathbf{v_k})\right) + \sum_{i=1}^n \chi_A(\alpha_\mathbf{i}^\pm)^2$ is convex and non smooth, A is the simplex of $\mathbb{R}^{2K}$ and D_k is defined as in (IV.21), depending on t_0^k. We recall that the orthogonality between vectors $\mathbf{v_k}$ is not taken into account in the geodesic surface approach.

IV.4.3.2. Optimization through the Forward-Backward Algorithm

Following [**ABS13**], in order to compute a critical point of problem (IV.24), one can consider the Forward-Backward algorithm (see also [**OCBP14**] for an acceleration using inertial terms). Denoting as $X = (\mathbf{v},\mathbf{t}) \in \mathbb{R}^{N+n}$, taking $\tau > 0$ and $X^{(0)} \in \mathbb{R}^{N+n}$, it reads:

$$X^{(\ell+1)} = \mathrm{Prox}_{\tau G}(X^{(\ell)} - \tau\nabla F(X^{(\ell)})), \tag{IV.26}$$

where $\mathrm{Prox}_{\tau G}(\tilde{X}) = \arg\min_X \frac{1}{2\tau}||X-\tilde{X}||^2 + G(X)$ with the Euclidean norm $||\cdot||$. In order to guarantee the convergence of this algorithm, the gradient of F has to be Lipschitz continuous with parameter $M > 0$ and the time step should be taken as $\tau < 1/M$. The details of computation of ∇F and $\mathrm{Prox}_{\tau G}$ for the two algorithms are given in Appendix IV.7.

IV.5. Statistical comparison between log-PCA and GPCA on synthetic and real data

IV.5.1. Synthetic example - Iterative versus geodesic surface approaches

First, for the synthetic example displayed in Figure IV.1, we compare the two algorithms (iterative and geodesic surface approaches) described in Section IV.4. The results are reported in Figure IV.4 by comparing the projection of the data onto the first and second geodesics computed with each approach. We also display the projection of the data onto the two-dimensional surface generated by each method. It should be recalled that the principal surface for the iterative geodesic algorithm is not necessarily a geodesic surface but each $g_t(t_0^k, u_k)_{t\in[-1;1]}$ defined by (IV.16) for $k = 1, 2$ is a geodesic curve for $\mathcal{U} = \{u_1, u_2\}$. For data generated from a location-scale family of Gaussian distributions, it appears that each algorithm provides a satisfactory reconstruction of the data set. The main divergence concerns the first and second principal geodesic. Indeed enforcing the orthogonality between components in the iterative approach enables to clearly separate the modes of variation in location and scaling, whereas searching directly a geodesic surface in the second algorithm implies a mixing of these two types of variation.

Note that the barycenter of Gaussian distributions $\mathcal{N}(m_i, \sigma_i^2)$ can be shown to be Gaussian with mean $\sum m_i$ and variance $(\sum \sigma_i)^2$.

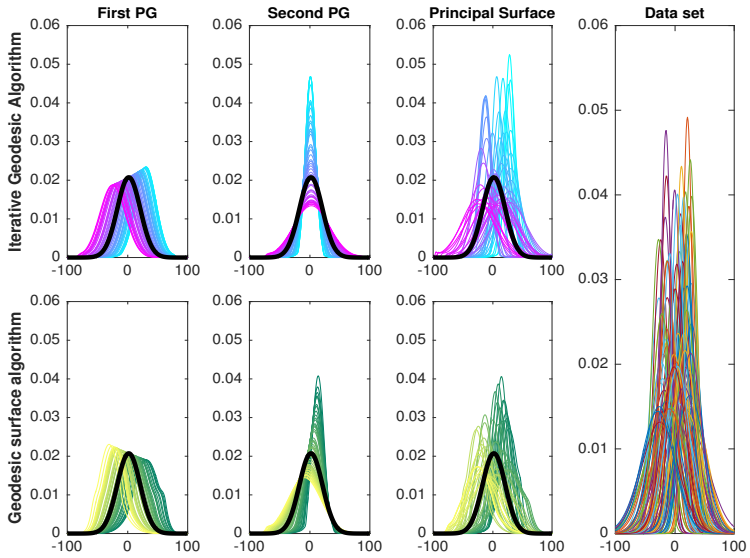


FIGURE IV.4. Synthetic example - Data sampled from a location-scale family of Gaussian distributions. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $\mathcal{P}_2(\Omega)$.

IV.5.2. Population pyramids

As a first real example, we consider a real dataset whose elements are histograms representing the population pyramids of $n = 217$ countries for the year 2000 (this dataset is produced by the International Programs Center, US Census Bureau (IPC, 2000), available at https://www.census.gov/programs-surveys/international-programs.html). Each histogram in the database represents the relative frequency by age, of people living in a given country. Each bin in a histogram is an interval of one year, and the last interval corresponds to people older than 85 years. The histograms are normalized so that their area is equal to one, and thus they represent a set of pdf. In Figure IV.5, we display the population pyramids of 4 countries, and the whole dataset. Along the interval $\Omega = [0, 84]$, the variability in this dataset can be considered as being small.

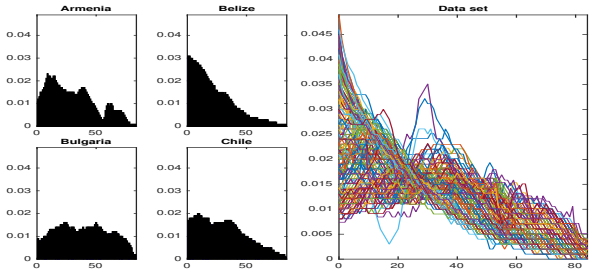


FIGURE IV.5. Population pyramids. A subset of population pyramids for 4 countries (left) for the year 2000, and the whole dataset of $n = 217$ population pyramids (right) displayed as pdf over the interval $[0, 84]$.

For $K = 2$, log-PCA and the iterative GPCA algorithm lead to the same principal orthogonal directions in $\mathbb{L}^2_{\boldsymbol{\nu}}(\Omega)$, namely that $\tilde{u}_1 = u_1^*$ and $\tilde{u}_2 = u_2^*$ where $(\tilde{u}_1, \tilde{u}_2)$ minimizes (IV.7) and (u_1^*, u_2^*) are minimizers of (IV.14). In this case, all projections of data $\omega_i = \log_{\boldsymbol{\nu}}(\nu_i)$ for $i = 1, \ldots, n$ onto $\mathrm{Sp}(\{\tilde{u}_1, \tilde{u}_2\})$ lie in $V_{\boldsymbol{\nu}}(\Omega)$, which means that log-PCA and the iterative geodesic algorithm lead exactly the same principal geodesics. Therefore, population pyramids is an example of data that are sufficiently concentrated around their Wasserstein barycenter so that log-PCA and GPCA are equivalent approaches (see Remark 3.5 in [**BGKL17**] for further details). Hence, we only display in Figure IV.6 the results of the iterative and geodesic surface algorithms.

In the iterative case, the projection onto the first geodesic exhibits the difference between less developed countries (where the population is mostly young) and more developed countries (with an older population structure). The second geodesic captures more subtle divergences concentrated on the middle age population. It can be observed that the geodesic surface algorithm gives different results since the orthogonality constraint on the two principal geodesics is not required. In particular, the principal surface mainly exhibit differences between countries with a young population with countries having an older population structure, but the difference between its first and second principal geodesic is less contrasted.

IV.5.3. Children's first name at birth

In a second example, we consider a dataset of histograms which represent, for a list of $n = 1060$ first names, the distribution of children born with that name per year in France between years 1900 and 2013. In Figure IV.7, we display the histograms of four different names, as well as the whole dataset. Along the interval $\Omega = [1900, 2013]$, the variability in this dataset is much larger than the one observed for population pyramids. This dataset has been provided by the INSEE (French Institute of Statistics and Economic Studies).

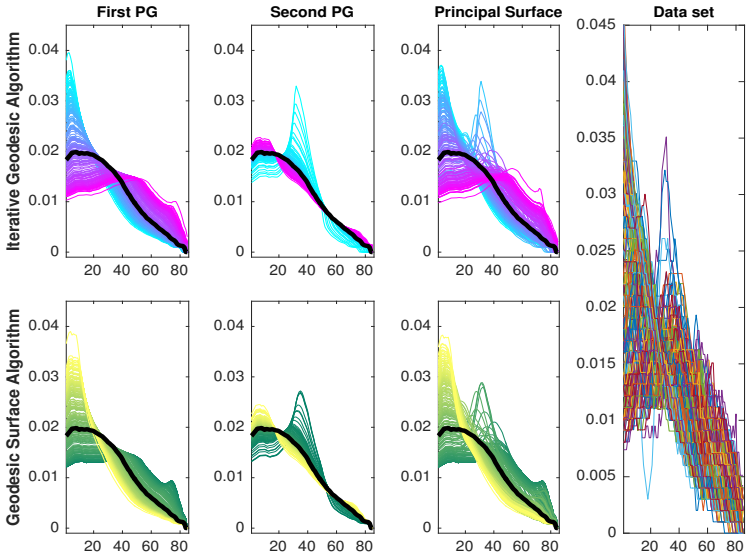


FIGURE IV.6. Population pyramids. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The first (resp. second) column is the projection of the data into the first (resp. second) principal direction. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $\mathcal{P}_2(\Omega)$.

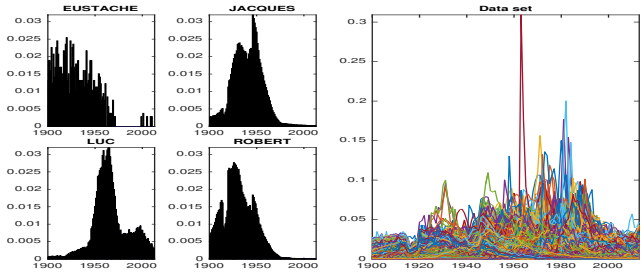


FIGURE IV.7. Children's first name at birth. An subet of 4 histograms representing the distribution of children born with that name per year in France, and the whole dataset of $n = 1060$ histograms (right), displayed as pdf over the interval $[1900, 2013]$

This is an example of real data where log-PCA and GPCA are not equivalent procedures for $K = 2$ principal components. We recall that log-PCA leads to the computation of principal orthogonal directions $\tilde{u}_1, \tilde{u}_2$ in $\mathbb{L}^2_{\boldsymbol{\nu}}(\Omega)$ minimizing (IV.7). First observe that in the left column of Figure IV.8, for some data $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, the mappings $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ are decreasing, and their range is larger than the interval Ω (that is, for some $x \in \Omega$, one has that $\tilde{T}_i(x) \notin \Omega$). Hence, such $\tilde{T}_i$ are not optimal mappings. Therefore, the condition $\Pi_{\mathrm{Sp}(\tilde{U})}\omega_i \in V_{\bar{\boldsymbol{\nu}}}(\Omega)$ for all $1 \leqslant i \leqslant n$ (with $\tilde{U} = \{\tilde{u}_1, \tilde{u}_2\}$) is not satisfied, implying that log-PCA does not lead to a solution of GPCA thanks to Proposition 3.5 in [**BGKL17**].

Hence, for log-PCA, the corresponding histograms displayed in the right column of Figure IV.8 are such that $\Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i \notin V_{\bar{\boldsymbol{\nu}}}(\Omega)$. This implies that the densities of the projected measures $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ have a support outside $\Omega = [1900, 2013]$. Hence, the

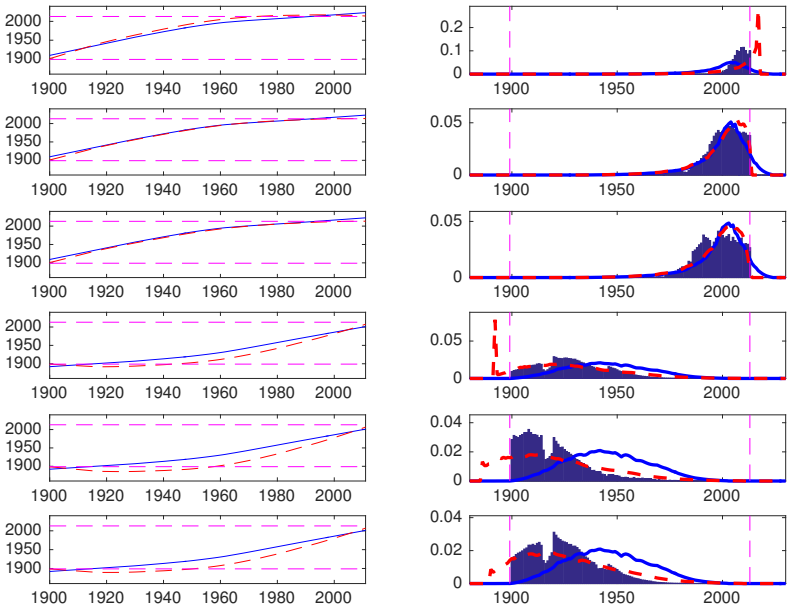


FIGURE IV.8. Children's first name at birth with support $\Omega = [1900, 2013]$. (Left) The dashed red curves represent the mapping $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ where $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, and $\tilde{u}_1$ is the first principal direction in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via log-PCA. The blue curves are the mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{u_1^*\})}\omega_i$, where u_1^* is the first principal direction in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via the iterative algorithm. (Right) The histogram stands for the pdf of measures ν_i that have a large Wasserstein distance with respect to the barycenter $\bar{\boldsymbol{\nu}}$. The red curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ with log-PCA, while the blue curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ with GPCA.

estimation of the measure $\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ by its projection onto the first mode of variation obtained with log-PCA is not satisfactory.

In Figure IV.8, we also display the results given by the iterative geodesic algorithm, leading to orthogonal directions u_1^*, u_2^* in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ that are minimizers of (IV.14). Contrary to the results obtained with log-PCA, one observes in Figure IV.8 that all the mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{u_1^*\})}\omega_i$ are non-decreasing, and such that $T_i(x) \in \Omega$ for all $x \in \Omega$. Nevertheless, by enforcing these two conditions, one has that a good estimation of the measure $\nu_i = \exp_{\bar{\boldsymbol{\nu}}}(\omega_i)$ by its projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ is made difficult as most of the mass of ν_i is located at either the right or left side of the interval Ω which is not the case for its projection. The histograms displayed in the right column of Figure IV.8 correspond to the elements in the dataset that have a large Wasserstein distance with respect to the barycenter $\bar{\boldsymbol{\nu}}$. This explains why it is difficult to have good projected measures with GPCA. For elements in the dataset that are closest to $\bar{\boldsymbol{\nu}}$, the projected measures $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ and $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ are much closer to ν_i and for such elements, log-PCA and the iterative geodesic algorithm lead to similar results in terms of data projection.

To better estimate the extremal data in Figure IV.8, a solution is to increase the support of the data to the interval $\Omega_0 = [1850, 2050]$, and to perform log-PCA and GPCA in the Wasserstein space $W_2(\Omega_0)$. The results are reported in Figure IV.9. In that case, it can be observed that both algorithms lead to similar results, and that a better projection is obtained for the extremal data. Notice that with this extended support, all the mappings

$\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ obtained with log-PCA are optimal in the sense that they are non-decreasing with a range inside Ω_0.

Finally, we display in Figure IV.10 and Figure IV.11 the results of the iterative and geodesic surface algorithms with either $\Omega = [1900, 2013]$ or with data supported on the extended support $\Omega_0 = [1850, 2050]$. The projection of the data onto the first principal geodesic suggests that the distribution of a name is deeply dependent on the part of the century. The second geodesic expresses a popular trend through a spike effect. In Figure IV.10, the artefacts in the principal surface that are obtained with the iterative algorithm at the end of the century, correspond to the fact that the projection of the data ω_i onto the surface spanned by the first two components is not ensured to belong to the set $V_{\bar{\boldsymbol{\nu}}}(\Omega)$.

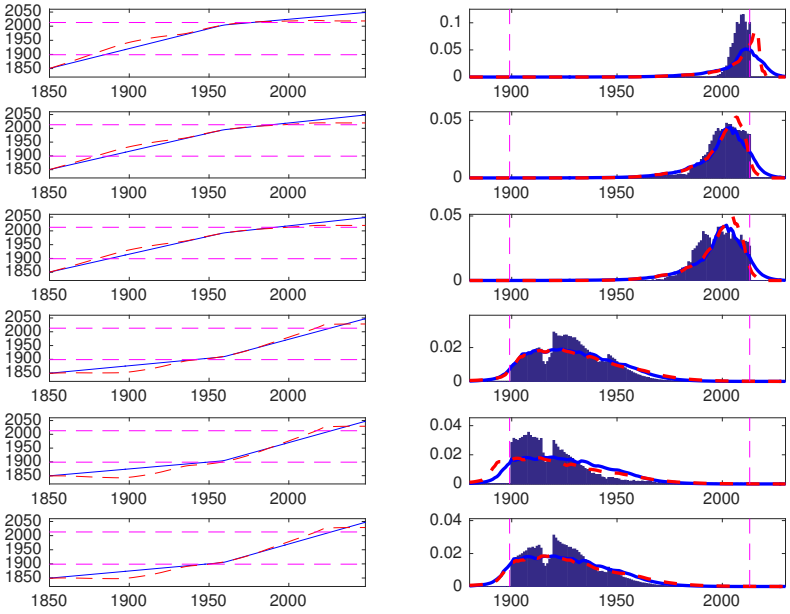


FIGURE IV.9. Children's first name at birth with extended support $\Omega_0 = [1850, 2050]$. (Left) The dashed red curves represent the mapping $\tilde{T}_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{\tilde{u}_1\})}\omega_i$ where $\omega_i = \log_{\bar{\boldsymbol{\nu}}}(\nu_i)$, and $\tilde{u}_1$ is the first principal direction in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via log-PCA. The blue curves are the mapping $T_i = \mathrm{id} + \Pi_{\mathrm{Sp}(\{u_1^*\})}\omega_i$, where u_1^* is the first principal direction in $\mathbb{L}^2_{\bar{\boldsymbol{\nu}}}(\Omega)$ obtained via the iterative algorithm. (Right) The histogram stands for the pdf of measures ν_i that have a large Wasserstein distance with respect to the barycenter $\bar{\boldsymbol{\nu}}$. The red curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(\tilde{u}_1)}\omega_i)$ with log-PCA, while the blue curves are the pdf of the projection $\exp_{\bar{\boldsymbol{\nu}}}(\Pi_{\mathrm{Sp}(u_1^*)}\omega_i)$ with GPCA.

IV.6. Extensions beyond $d > 1$ and some perspectives

We now briefly show that our iterative algorithm for finding principal geodesics can be adapted to the general case $d > 1$. This requires to take into account two differences with the one-dimensional case. First, the definition of the space $V_{\mu_r}(\Omega)$ in IV.1 relies on the explicit close-form solution (IV.1) of the optimal transport problem which is specific to the

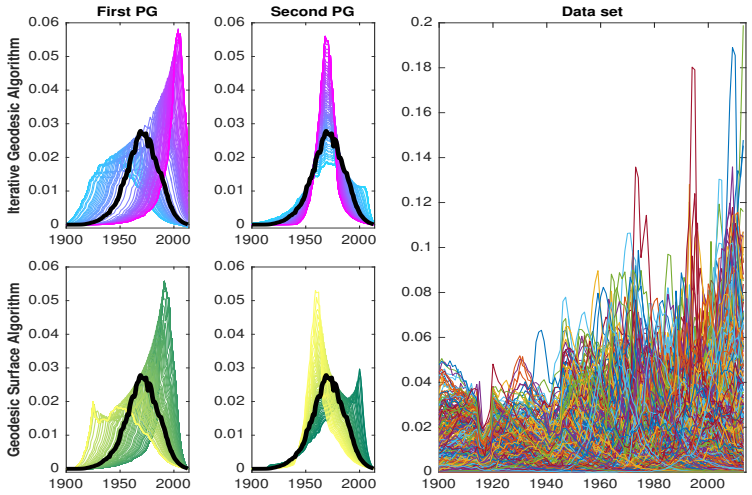


FIGURE IV.10. Children's first name at birth with support $\Omega = [1900, 2013]$. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The first (resp. second) column is the projection of the data into the first (resp. second) principal direction. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $\mathcal{P}_2(\Omega)$.

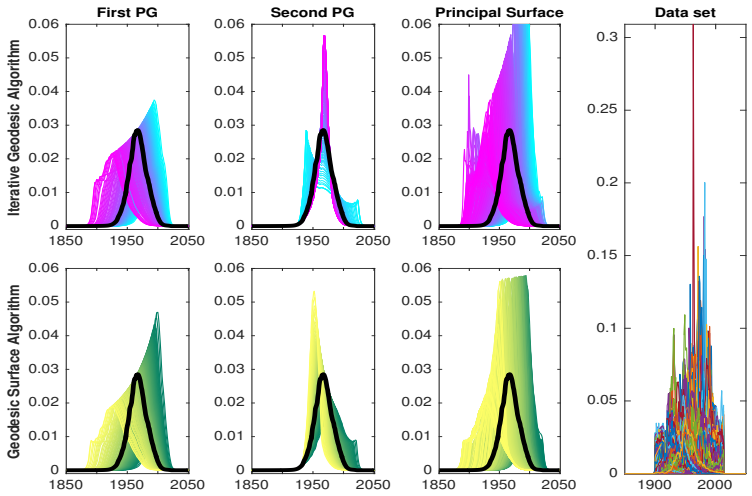


FIGURE IV.11. Children's first name at birth with extended support $\Omega_0 = [1850, 2050]$. The first row is the GPCA of the data set obtained with the iterative geodesic approach. The second row is the GPCA through the geodesic surface approach. The first (resp. second) column is the projection of the data into the first (resp. second) principal direction. The black curve is the density of the Wasserstein barycenter. Colors encode the progression of the pdf of principal geodesic components in $\mathcal{P}_2(\Omega)$.

one-dimensional case. We must hence provide a more general definition of $V_{\mu_r}(\Omega)$. Second, the isometry property $(P1)$ does not hold for $d > 1$, so that Wasserstein distances cannot be replaced by the $L^2_{\bar{\boldsymbol{\nu}}}$ norm between log-maps as in (IV.17) and must be explicitly computed and differentiated.

Definition of $V_{\mu_r}(\Omega)$ in the general case. In the one dimensional case, $V_{\mu_r}(\Omega)$ is characterized in Proposition IV.2 (P3) as the set of functions $v \in \mathbb{L}^2_{\mu_r}(\Omega)$ such that $T := \mathrm{id}+v$ is μ_r-almost everywhere non decreasing. A important result by Brenier [**Bre91**] is that, in any dimension, if μ_r does not give mass to small sets, there exists an optimal mapping $T \in \mathbb{L}^2_{\mu_r}(\Omega)$ between μ_r and any probability measure ν, and T is equal to the gradient of a convex function u, i.e. $T = \nabla u$. Therefore we define the set $V_{\mu_r}(\Omega)$ as the set of functions $v \in \mathbb{L}^2_{\mu_r}(\Omega)$ such that $\mathrm{id} + v = \nabla u$ for an arbitrary convex function u.

In order to deal with the latter constraint, we note it implies that $div(v) \geqslant -1$. Indeed, assuming that $\mathrm{id} + v = \nabla u$, then u being a convex potential involves $div(\nabla u) \geqslant 0$ which is equivalent to $div(\mathrm{id} + v) = div(v) + 1 \geqslant 0$. We therefore choose to substitute this constraint by the constraint $div(v) \geqslant -1$.

General objective function. Without the isometry property (P1), the objective function $H(t_0, v)$ in (IV.15) must be written with the explicit Wasserstein distance W_2,

$$H(t_0, v) = \frac{1}{n}\sum_{i=1}^{n} \min_{t_i \in [-1;1]} W_2^2(\nu_i, g_{t_i}(t_0, v)) + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0 - 1)v) + \chi_{V_{\bar{\boldsymbol{\nu}}}(\Omega)}((t_0 + 1)v), \quad \text{(IV.27)}$$

where $g_t(t_0, v) = (\mathrm{id} + (t_0 + t)v)\#\bar{\boldsymbol{\nu}}$ for $t \in [-1; 1]$ as defined in (IV.16). Optimizing over both the functions $\mathbf{v} \in (\mathbb{R}^d)^N$ and the projection times $\mathbf{t}$, the discretized objective function to minimize is,

$$\min_{\mathbf{v}\in\mathbb{R}^N} \min_{\mathbf{t}\in\mathbb{R}^n} J(\mathbf{v}, \mathbf{t}) := \underbrace{\sum_{i=1}^{n} W_2^2(\nu_i, g_{t_i}(t_0, \mathbf{v}))}_{F(\mathbf{v},\mathbf{t})} + \underbrace{\chi_S(v) + \chi_E(K\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_{B_1^n}(\mathbf{t})}_{G(\mathbf{v},\mathbf{t})}. \quad \text{(IV.28)}$$

where K is a discretized divergence operator, and $E = \{\mathbf{z} \in \mathbb{R}^N \ : \ \frac{-1}{t_0+1} \leqslant \mathbf{z} \leqslant \frac{1}{1-t_0}\}$, $D = \{\mathbf{v} \ : \ \mathrm{id} + (t_0 \pm 1)\mathbf{v} \in \Omega\}$ deals with the domain constraint and S deals with the orthogonality constraint w.r.t. to the preceding principal components. As for the one-dimensional case, we minimize J through the Forward-Backward algorithms detailed in the appendix IV.7.2.

Extension to higher dimensions is straightforward. However, considering that we have to discretize the support of the Wasserstein mean $\bar{\boldsymbol{\nu}}$, the approach becomes intractable for $d > 3$.

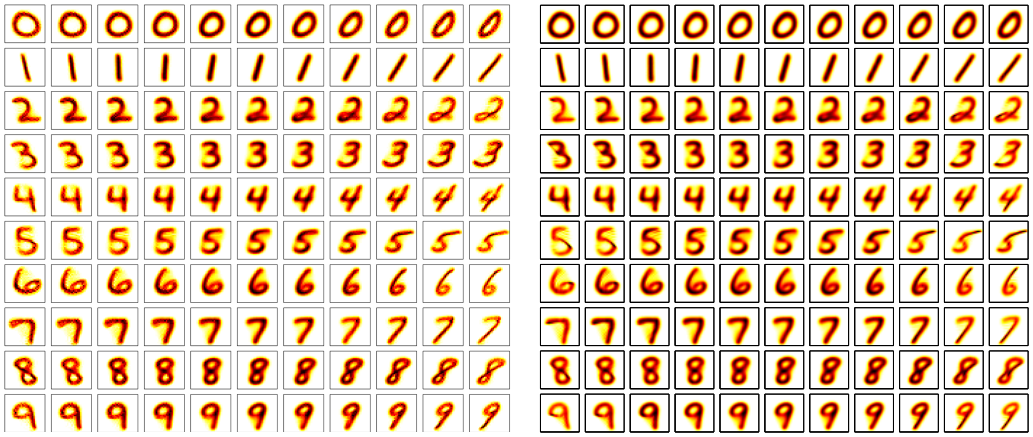
IV.6.1. Application to grayscale images

We consider the MNIST dataset [**LeC98**] which contains grayscale images of handwritten digits. All the images have identical size 28×28 pixels. Each grayscale image, once normalized so that the sum of pixel grayscale values sum to one, can be interpreted as a discrete probability measure, which is supported on the 2D grid of size 28×28. The ground metric for the Wasserstein distance is then the 2D squared Euclidean distance between the locations of the pixels of the two-dimensional grid. We compute the first principal components on 1000 images of each digit. Wasserstein barycenters, which are required as input to our algorithm, are approximated efficiently through iterative Bregman projections as proposed in [**BCC+15**]. We use the network simplex algorithm[2] to compute Wasserstein distances.

Figure IV.12 displays the results obtained with our proposed forward-backward algorithm (with t_0 set to 0 for simplicity), and the ones given by Log-PCA as described in section

[2] http://liris.cnrs.fr/~nbonneel/FastTransport/

IV.3. These two figures are obtained by sampling the first principal components. We then use kernel smoothing to display the discrete probability measures back to the original grid and present the resulting grayscale image with an appropriate colormap.



GPCA log-PCA

FIGURE IV.12. First principal geodesics for 1000 images of each digit from the MNIST dataset, computed through the proposed Forward-Backward algorithm (left) and log-PCA (right).

Visually, both the Log-PCA and GPCA approaches capture well the main source of variability of each set of grayscale images. We observe variations in the slant of the handwritten digits for all digits, the most obvious case being digit '1'. As a principal component is parameterized by a whole velocity field on the support of the Wasserstein mean of the data, single principal components can capture more interesting patterns, such as changes in the shape of the '0' or the presence or absence of the lower loop of the '2'. From purely visual inspection, it is difficult to tell which approach, Log-PCA or GPCA, provides a "better" principal component. For this purpose we compute the reconstruction error of each digit. This reconstruction error is computed in the same way for both Log-PCA and GPCA principal components: We sample the principal components at many times t and find for each image in a given dataset, the time at which the geodesic is the closest to the image sample. This provides an approximation of $\min_{t\in[-1;1]} W_2^2(\nu_i, g_t(v))$ for each image $i = 1, \dots, n$, where $(g_t)_{t\in[-1,1]}$ is the principal component. For the Log-PCA principal component, we take $\tilde{g}_t = (\mathrm{id} + t1.25\lambda v)\#\bar{\boldsymbol{\nu}}$, where λ is the eigenvalue corresponding to the first principal component. The 1.25 factor is useful to consider a principal curve which goes through the whole range of the dataset. For the GPCA principal geodesic, we have $g_t^* = (\mathrm{id} + tv)\#\bar{\boldsymbol{\nu}}$. The reconstruction errors are shown in Table 1. We see that, for each digit, we obtain a better, i.e. smaller, reconstruction error when using the proposed forward-backward algorithm. This result is not surprising, since the reconstruction error is explicitly minimized through the Forward-Backward algorithm. As previously mentioned, Log-PCA rather computes linearized Wasserstein distances. In one-dimension, the isometry property (P1) states that these quantities are equal. In dimension two or larger, that property does not hold.

IV.6.2. Discussion

The proposed forward-backward algorithm minimizes the same objective function as defined in [SC15]. The first difference with the algorithm provided [SC15] is that we take

MNIST digit	Log-PCA RE ($\cdot 10^3$)	GPCA RE ($\cdot 10^3$)
0	2.0355	**1.9414**
1	3.1426	**1.0289**
2	3.4221	**3.3575**
3	2.6528	**2.5869**
4	2.8792	**2.8204**
5	2.9391	**2.9076**
6	2.1311	**1.9864**
7	4.7471	**2.8205**
8	2.0741	**2.0222**
9	1.9303	**1.8728**

TABLE 1. Reconstruction Errors (RE) computed on 1000 sample images of each digit of the MNIST dataset. (center) Reconstruction error w.r.t. the first principal component computed with the Log-PCA algorithm. (right) Reconstruction error w.r.t. the first principal geodesic computed with the proposed Forward-Backward algorithm.

gradient steps with respect to both $\mathbf{v}$ and $\mathbf{t}$, while the latter first attempts to find the optimal t (by sampling the geodesics at many time t), before taking a gradient step of $\mathbf{v}$. Our approach reduces the cost of computing a gradient step by one order of magnitude. Secondly, [**SC15**] relied on barycentric projections of optimal plans to preserve the geodesicity of the principal curves in between gradient steps. That heuristic does not guarantee a decrease in the objective after a gradient step. Moreover, the method in [**SC15**] considered two velocity fields $\mathbf{v}_1, \mathbf{v}_2$ rather than a single $\mathbf{v}$ since the optimality of both $\mathbf{v}$ and $-\mathbf{v}$ could not be preserved through the barycentric projection.

IV.7. Algorithms

IV.7.1. Dimension $d = 1$

We here detail the application of Algorithm (IV.26) to the iterative GPCA procedure that consists in solving the problem (IV.24):

$$\min_{\mathbf{v}\in\mathbb{R}^N} \min_{\mathbf{t}\in\mathbb{R}^n} J(\mathbf{v},\mathbf{t}) := \underbrace{\sum_{i=1}^{n} \|\mathbf{w}_i - (t_0+t_i)\mathbf{v}\|^2_{\boldsymbol{\nu}}}_{F(\mathbf{v},\mathbf{t})} + \underbrace{\chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) + \chi_{B_1^n}(\mathbf{t})}_{G(\mathbf{v},\mathbf{t})}.$$

IV.7.1.1. Lipschitz constant of ∇F

Let us now look at the Lipschitz constant of $\nabla F(\mathbf{v},\mathbf{t})$ on the restricted acceptable set $D \times B_1^n$. We first denote as $\mathcal{H}$ the hessian matrix (of size $(N+n)\times(N+n)$) of the $\mathcal{C}^2$ function $F(X)$. We know that if the spectral radius of $\mathcal{H}$ is bounded by a scalar value M, i.e. $\rho(\mathcal{H}) \leqslant M$, then ∇F is a Lipschitz continuous function with constant M. Hence, we look at the eigenvalues of the Hessian matrix of $F = \sum_{i=1}^{n}\sum_{j=1}^{N} \bar{\boldsymbol{f}}_n(x_j)(w_i^j - (t_0+t_i)v_j)^2$ that is

$$\frac{\partial^2 F}{\partial t_i^2} = \sum_{j=1}^{N} 2v_j^2 \bar{\boldsymbol{f}}_n(x_j), \quad \frac{\partial^2 F}{\partial v_j^2} = \sum_{i=1}^{n} 2(t_0+t_i)^2 \bar{\boldsymbol{f}}_n(x_j), \quad \frac{\partial^2 F}{\partial t_i \partial v_j} = 2\bar{\boldsymbol{f}}_n(x_j)(2(t_0+t_i)v_j - w_i^j)$$

and $\frac{\partial^2 F}{\partial t_i \partial t_{i'}} = \frac{\partial^2 F}{\partial v_j \partial v_{j'}} = 0$, for all $i \neq i'$ or $j \neq j'$. Being $\{\mu_k\}_{k=1}^{n+N}$ the eigenvalues of $\mathcal{H}$, we have $\rho(\mathcal{H}) = \max_k |\mu_k| \leqslant \max_k \sum_l |\mathcal{H}_{kl}|$. We denote as $f_\infty = \max_j |\bar{\boldsymbol{f}}_n(x_j)|$ and likewise $w_\infty = \max_{i,j} |w_i^j|$. Since $|t_0| < 1$, $t_i^2 \leqslant 1$, $\forall \mathbf{t} \in B_1^n$ and $v_j^2 \leqslant \alpha^2 = (b-a)^2$, $\forall \mathbf{v} \in D$, by defining $\gamma = 2(1+|t_0|)\alpha + w_\infty$, we thus have

$$\rho(\mathcal{H}) \leqslant 2 f_\infty \max\left\{ n\alpha^2 + N\gamma, n\gamma + N(1+|t_0|)^2 \right\} := M. \tag{IV.29}$$

IV.7.1.2. Computing $\mathbf{Prox}_{\tau G}$

In order to implement the algorithm (IV.26), we finally need to compute the proximity operator of G defined as:

$$\begin{aligned}(\mathbf{v}^*, \mathbf{t}^*) &= \mathrm{Prox}_{\tau G}(\tilde{\mathbf{v}}, \tilde{\mathbf{t}}) \\ &= \arg\min_{\mathbf{v},\mathbf{t}} \frac{1}{2\tau}(||\mathbf{v}-\tilde{\mathbf{v}}||^2 + ||\mathbf{t}-\tilde{\mathbf{t}}||^2) + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) + \chi_{B_1^n}(\mathbf{t}).\end{aligned}$$

This problem can be solved independently on $\mathbf{v}$ and $\mathbf{t}$. For $\mathbf{t}$, it can be done pointwise as $t_i^* = \arg\min_{t_i} \frac{1}{2\tau}||t_i - \tilde{t}_i||^2 + \chi_{B_1^1}(t_i) = \mathrm{Proj}_{[-1;1]}(\tilde{t}_i)$. Unfortunately, there is no closed form expression of the proximity operator for the component $\mathbf{v}$. It requires to solve the following intern optimization problem at each extern iteration (ℓ) of the algorithm (IV.26):

$$\mathbf{v}^* = \arg\min_{\mathbf{v}} \frac{1}{2\tau}||\mathbf{v}-\tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}), \tag{IV.30}$$

where, to avoid confusions, we denote by $\mathbf{v}$ the variable that is optimized within the intern optimization problem (IV.30).

Remark IV.10. *The Lipschitz constant of $\nabla F(\mathbf{v}, \mathbf{t})$ in (IV.29) relies independently on $\mathbf{v}$ and $|t_0|$, thus we can choose the optimal gradient descent step τ for $\mathbf{v}^*$ and $\mathbf{t}^*$.*

Primal-Dual reformulation. Using duality (through Fenchel transform), one has:

$$\begin{aligned}&\min_{\mathbf{v}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v}-\tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) \\ &= \min_{\mathbf{v}\in\mathbb{R}^N} \max_{\mathbf{z}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v}-\tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \langle K\mathbf{v}, \mathbf{z}\rangle - \chi_E^*(\mathbf{z}),\end{aligned} \tag{IV.31}$$

where $\mathbf{z} = \{\mathrm{z}_j\}_{j=1}^N \in \mathbb{R}^N$ is a dual variable and $\chi_E^* = \sup_{\mathbf{v}} \langle \mathbf{v}, \mathbf{z}\rangle - \chi_E(\mathbf{v})$ is the convex conjugate of χ_E that reads:

$$(\chi_E^*(\mathbf{z}))_j = \begin{cases} -\mathrm{z}_j/(1+t_0) & \text{if } \mathrm{z}_j \leqslant 0, \\ \mathrm{z}_j/(1-t_0) & \text{if } \mathrm{z}_j > 0. \end{cases}$$

Hence, one can use the Primal-Dual algorithm proposed in [**CP16a**] to solve the problem (IV.31). For two parameters $\sigma, \theta > 0$ such that $||K||^2 \leqslant \frac{1}{\sigma}(\frac{1}{\theta} - \frac{1}{\tau})$ and given $\mathbf{v}^0, \bar{\mathbf{v}}^0, \mathbf{z}^0 \in \mathbb{R}^N$, the algorithm is:

$$\begin{cases} \mathbf{z}^{(m+1)} &= \mathrm{Prox}_{\sigma\chi_E^*}(\mathbf{z}^{(m)} + \sigma K \bar{\mathbf{v}}^{(m)}) \\ \mathbf{v}^{(m+1)} &= \mathrm{Prox}_{\theta(\chi_D + \chi_S)}(\mathbf{v}^{(m)} - \theta(K^*\mathbf{z}^{(m+1)} + \frac{1}{\tau}(\mathbf{v}^{(m)} - \tilde{\mathbf{v}})) \\ \bar{\mathbf{v}}^{(m+1)} &= 2\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)} \end{cases} \tag{IV.32}$$

where K^* is defined as $\langle K\mathbf{v}, \mathbf{z}\rangle = \langle \mathbf{v}, K^*\mathbf{z}\rangle$. Using the operator K defined in (IV.23), we thus have:

$$(K^*\mathbf{z})_j = \begin{cases} -\mathrm{z}_1/\Delta_1 & \text{if } j = 1 \\ \mathrm{z}_{j-1}/\Delta_{j-1} - \mathrm{z}_j/\Delta_j & \text{if } 1 < j < N. \\ \mathrm{z}_{N-1}/\Delta_{N-1} & \text{if } j = N, \end{cases} \tag{IV.33}$$

where $\Delta_j = x_{j+1} - x_j$. We have that $||K||^2 = \rho(K^*K)$, the largest eigenvalue of K^*K. With the discrete operators (IV.23) and (IV.33), $\rho(K^*K)$ can be bounded by

$$\delta^2 = 2 \max_j (1/\Delta_j^2 + 1/\Delta_{j+1}^2). \tag{IV.34}$$

One can therefore for instance take $\sigma = \frac{1}{\delta}$ and $\theta = \tau/(1+\delta\tau)$.

Proximity operators in (IV.32). The proximity operator of $\chi_D + \chi_S$ is obtained as:

$$(\text{Prox}_{\theta(\chi_D+\chi_S)}(\mathbf{v}))_j = (\text{Proj}_{D\cap S}(\mathbf{v}))_j = \text{Proj}_{[m_j;M_j]}\left(\left(\mathbf{v} - \sum_{l=1}^{k-1} \frac{\langle \mathbf{u}_l, \mathbf{v}\rangle_{\bar{\boldsymbol{\nu}}}}{||\mathbf{u}_l||^2_{\bar{\boldsymbol{\nu}}}}\mathbf{u}_l\right)_j\right), \tag{IV.35}$$

since projecting onto $D \cap S$ is equivalent to first project onto the orthogonal of $\text{Sp}(\mathcal{U}^{k-1})$ and then onto D. One can finally show that the proximity operator of χ_E^* can be computed pointwise as:

$$(\text{Prox}_{\sigma\chi_E^*}(\mathbf{z}))_j = \begin{cases} z_j - \sigma/(1-t_0) & \text{if } z_j > \sigma/(1-t_0) \\ z_j + \sigma/(1+t_0) & \text{if } z_j < -\sigma/(1+t_0) \\ 0 & \text{otherwise.} \end{cases} \tag{IV.36}$$

IV.7.1.3. Algorithms for GPCA

Gathering all the previous elements, we can finally find a critical point of the non-convex problem (IV.24) using the Forward-Backward (FB) framework (IV.26), as detailed in Algorithm 1.

Algorithm 1 Resolution with FB of problem (IV.24): $\min_{\mathbf{v},\mathbf{t}} F(\mathbf{v},\mathbf{t}) + G(\mathbf{v},\mathbf{t})$

Require: $\boldsymbol{w}_i \in \mathbb{R}^N$ for $i = 1 \cdots n$, $\mathbf{u}_1, \cdots \mathbf{u}_{k-1}$, $t_0 \in]-1;1[$, $\alpha = (b-a) > 0$, $\eta > 0$, $\delta > 0$ (defined in (IV.34)) and $M > 0$ (defined in (IV.29)).
Set $(\mathbf{v}^{(0)}, \mathbf{t}^{(0)}) \in D \times B_1^n$
Set $\tau < 1/M$, $\sigma = 1/\delta$ and $\theta = \tau/(1+\delta\tau)$.
%Extern loop:
while $||\mathbf{v}^{(\ell)} - \mathbf{v}^{(\ell-1)}||/||\mathbf{v}^{(\ell-1)}|| > \eta$ **do**
  % FB on $\mathbf{t}$ with $\mathbf{t}^{(\ell+1)} = Prox_{\tau G}(\mathbf{t}^{(\ell)} - \tau\nabla F(\mathbf{v}^{(\ell)}, \mathbf{t}^{(\ell)}))$:
  ${t_i}^{(\ell+1)} = \text{Proj}_{[-1;1]}\left({t_i}^{(\ell)} - \tau\sum_{j=1}^N {v_j}^{(\ell)} \bar{\boldsymbol{f}}_n(x_j)\left((t_0 + {t_i}^{(\ell)}){v_j}^{(\ell)} - w_i^j\right)\right)$
  % Gradient descent on $\mathbf{v}$ with $\tilde{\mathbf{v}} = \mathbf{v}^{(\ell)} - \tau\nabla F(\mathbf{v}^{(\ell)}, \mathbf{t}^{(\ell)})$:
  $\tilde{v}_j = {v_j}^{(\ell)} - \tau\bar{\boldsymbol{f}}_n(x_j)\sum_{i=1}^n (t_0 + {t_i}^{(\ell)})\left((t_0 + {t_i}^{(\ell)})\ {v_j}^{(\ell)} - w_i^j\right)$
  %Intern loop for $\mathbf{v}^{(\ell+1)} = Prox_{\tau G}(\tilde{\mathbf{v}})$:
  Set $\mathbf{z}^{(0)} \in E$, $\mathbf{v}^{(0)} = \tilde{\mathbf{v}}$, $\bar{\mathbf{v}}^{(0)} = \tilde{\mathbf{v}}$
  while $||\mathbf{v}^{(m)} - \mathbf{v}^{(m-1)}||/||\mathbf{v}^{(m-1)}|| > \eta$ **do**
    $\mathbf{z}^{(m+1)} = \text{Prox}_{\sigma\chi_E^*}\left(\mathbf{z}^{(m)} + \sigma K\bar{\mathbf{v}}^{(m)}\right)$ (using (IV.36))
    $\mathbf{v}^{(m+1)} = \text{Prox}_{\theta(\chi_D+\chi_S)}\left(\mathbf{v}^{(m)} - \theta(K^*\mathbf{z}^{(m+1)} + \frac{1}{\tau}(\mathbf{v}^{(m)} - \tilde{\mathbf{v}})\right)$ (using (IV.35))
    $\bar{\mathbf{v}}^{(m+1)} = 2\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)}$
    $m := m + 1$
  end while
  $\mathbf{v}^{(\ell+1)} = \mathbf{v}^{(m)}$
  $\ell := \ell + 1$
end while
return $\mathbf{u}_k = \mathbf{v}^{(\ell)}$

Geodesic surface approach. In order to solve the problem (IV.25), we follow the same steps as in the section IV.7.1.1-IV.7.1.2. First we obtain the Lipchitz constant of the

function $\tilde{F}$ by the same computations performed for the iterative algorithm. Then, since the constraints' problem in G' are separable, we can compute each component $\mathbf{v_k}$ and each $\alpha_{\mathbf{i}}^{\pm}$ independently. The only difference with the iterative algorithm concerns the proximal operator of the function χ_A, which is the projection into the simplex of $\mathbb{R}^{2K}$.

IV.7.2. Dimension $d = 2$

We now show how to generalize the algorithm to the two-dimensional case.

Gradients of F. We write $X = (x_1, \cdots, x_N) \in (\mathbb{R}^2)^N$ the discretized support of $\bar{\boldsymbol{\nu}}$, $Z_t = (x_1 + (t_0 + t)v_1, \cdots, x_N + (t_0 + t)v_N)$ the support $g_t(t_0, \mathbf{v})$, the geodesic sampled at time t. Let P^* be an optimal transport plan between $\bar{\boldsymbol{\nu}}$ and $g_t(t_0, \mathbf{v})$. The function $F(\mathbf{v}, \mathbf{t})$ is differentiable almost everywhere. Gradients can be computed in the same fashion as [**SC15**] to obtain,

$$\nabla_{\mathbf{v}} F = 2\sum_{i=1}^{n}(t_0 + t_i)(Z_{t_i} - XP^{*T}\mathrm{diag}(1/\bar{\boldsymbol{f}}_n)), \qquad \nabla_{t_i} F = 2\langle Z_{t_i}\mathrm{diag}(\bar{\boldsymbol{f}}_n), \mathbf{v}\,\rangle - 2\langle P^*, \mathbf{v}^T X\,\rangle, \tag{IV.37}$$

Proximal operator of G. The only difference between the one-dimensional case and the two-dimensional case considered here concerns the projection step of $\mathbf{v}$,

$$\mathbf{v}^* = \arg\min_{\mathbf{v}} \frac{1}{2\tau}||\mathbf{v} - \tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}). \tag{IV.38}$$

Primal-Dual reformulation. As for the on-dimensional case, one has,

$$\begin{aligned} &\min_{\mathbf{v}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v} - \tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \chi_E(K\mathbf{v}) \\ &= \min_{\mathbf{v}\in\mathbb{R}^N}\max_{\mathbf{z}\in\mathbb{R}^N} \frac{1}{2\tau}||\mathbf{v} - \tilde{\mathbf{v}}||^2 + \chi_S(\mathbf{v}) + \chi_D(\mathbf{v}) + \langle K\mathbf{v}, \mathbf{z}\rangle - \chi_E^*(\mathbf{z}), \end{aligned} \tag{IV.39}$$

where $\mathbf{z} = \{z_j\}_{j=1}^N \in \mathbb{R}^N$ is a dual variable and $\chi_E^* = \sup_{\mathbf{v}}\langle \mathbf{v}, \mathbf{z}\rangle - \chi_E(\mathbf{v})$ is the convex conjugate of χ_E. This can be solve with the same iterative steps as described in IV.7.1.2,

$$\begin{cases} \mathbf{z}^{(m+1)} &= \mathrm{Prox}_{\sigma\chi_E^*}(\mathbf{z}^{(m)} + \sigma K\bar{\mathbf{v}}^{(m)}) \\ \mathbf{v}^{(m+1)} &= \mathrm{Prox}_{\theta(\chi_D+\chi_S)}(\mathbf{v}^{(m)} - \theta(K^*\mathbf{z}^{(m+1)} + \frac{1}{\tau}(\mathbf{v}^{(m)} - \tilde{\mathbf{v}})) \\ \bar{\mathbf{v}}^{(m+1)} &= 2\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)} \end{cases} \tag{IV.40}$$

Here the definition of the divergence operator K and the transpose of the divergence operator K^* are specific to the dimension. For $d = 2$, with a regular grid discretizing Ω in $M \times N$ points, we take

$$K^T\mathbf{z} = -\nabla\mathbf{z} = -\begin{bmatrix}\partial_x^+\mathbf{z} \\ \partial_y^+\mathbf{z}\end{bmatrix},$$

with

$$\partial_x^+\mathbf{z}(i,j) = \begin{cases} \mathbf{z}(i+1,j) - \mathbf{z}(i,j) & if\, i < M \\ 0 & otherwise, \end{cases}$$

$$\partial_y^+\mathbf{z}(i,j) = \begin{cases} \mathbf{z}(i,j+1) - \mathbf{z}(i,j) & if\, j < N \\ 0 & otherwise \end{cases}$$

so that

$$K\mathbf{u} = K\begin{bmatrix}\mathbf{u}_x \\ \mathbf{u}_y\end{bmatrix} = \partial_x^-\mathbf{u}_x + \partial_y^-\mathbf{u}_y,$$

with

$$\partial_x^-\mathbf{u}(i,j) = \begin{cases} \mathbf{u}(i,j) - \mathbf{u}(i-1,j) & if\, 1 < i < M \\ \mathbf{u}(i,j) & if\, i = 1 \\ -\mathbf{u}(i-1,j) & if\, i = M. \end{cases}$$

To ensure convergence of IV.40, one can take $1/\sigma.(1/\theta - 1/\tau) = ||K||^2$. See [**CP16a**, **LP15**] for more details. Since we have $||K||^2 = 8$, the parameters can be taken as $\sigma = 1/4$ and $\theta = \tau/(1 + 2\tau)$.

Conclusion and perspectives

Using optimal transport, the aim of this thesis is to provide new tools to deal with statistical problems involving multivariate probability distributions. As we also consider implementation issues and practical problems, this thesis is at the boundary of theory, applied mathematics and computer sciences.

Chapter I. We strongly believe that the results on the variance of the penalized barycenters in Section I.1 could be improve, by relaxing the hypothesis on both the penalty function E, and the compact convex set $\Omega \subset \mathbb{R}^d$. This would need a finer theory than the objects used from the empirical process theory. Another interesting work would be to include regularity into the empirical barycenter, depending on the regularity of the measures $\nu_1, \dots, \nu_n$.

As seen in Chapter II, the entropy regularized Wasserstein barycenters $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ in Section I.2, behaves like a density estimator with a bandwidth parameter ε. In particular, Figure B.9, page 28 suggests that depending on the regularization parameter ε, the number of modes of the barycenter varies. Therefore, a thorough study of this estimator and its modes could be very interesting.

Chapter II. We become aware of an another method for the choice of the regularization parameter, using the recent paper of Spokoiny and Willrich [**SW15**], which is based on bootstrap techniques. We would like to apply this principle to the regularized barycenters.

Chapter III. We intend to further investigate the benefits of the use of Sinkhorn divergences to propose novel testing procedure to compare multivariate distributions for real data analysis. A first perspective is to apply the methodology developed in Chapter III to more than two samples using the notion of entropy regularized Wasserstein barycenters for the analysis of variance of multiple and multivariate random measures (MANOVA). However, as pointed out in [**CP16b**], a critical issue in this setting will be the choice of the regularization parameter ε, as it has a large influence on the shape of the estimated Wasserstein barycenter. Our simulations in Section III.5, page 86 show that using a smoothed Wasserstein barycenter as a reference measure may lead to different results than using an Euclidean barycenter when testing the hypothesis of equal distributions. We thus plan to study the behavior of the central limit theorem when the regularization parameter ε tends to zero, expecting that we could recover the results from [**SM16**] that stand for un-regularized transport.

Another issue is that, for one or two samples testing, the use of entropy regularized transport leads to a biased statistics in the sense that its expectation $W^2_{2,\varepsilon}(a,b)$ is not equal to zero under the hypothesis that $a = b$. A possible alternative to avoid this issue would be to use the so-called notion of Sinkhorn loss defined as

$$\bar{W}^2_{2,\varepsilon}(a,b) := 2W^2_{2,\varepsilon}(a,b) - W^2_{2,\varepsilon}(a,a) - W^2_{2,\varepsilon}(b,b),$$

that has been recently introduced in [**GPC17**], and which satisfies the property that $\bar{W}^2_{2,\varepsilon}(a,b) = 0$ when $a = b$. An interesting extension of the results in this Chapter would thus be to develop

test statistics based on the Sinkhorn loss for the comparison of multivariate distributions. We believe this can be done using similar tools.

Chapter IV. When considering probability measures over high-dimensional space ($d > 3$), our GPCA algorithm becomes intractable since we need to discretize the support of the Wasserstein mean of the data with a regular grid, whereas it is still possible to apply the method of [**SC15**], since an arbitrary support for the Wasserstein mean is used. A remaining challenge for computing principal geodesics in the Wasserstein space is then to propose an algorithm for GPCA which is still tractable in higher dimensions while not relying on barycentric projections of optimal transport plans as in [**SC15**].

Some others ideas for the future. Optimal transport is still a growing field today and we deeply think that important and new results can be employed. Indeed, recent results as in [**FHN+18**] propose new ideas to compute efficiently the optimal transport between measures, allowing in particular statistical data analysis for high-dimensional data. Also, the work of [**BPC16**,**SHB+18b**] on dictionary learning and regression study, could be extended to define new regression models in the Wasserstein space, allowing to explore measures outside the convex set defined by the dataset. Finally, in the recent paper [**RBVFT18**], the authors relate to Wasserstein barycenters for Bayesian learning approaches. Therefore new contributions could arise from the relation between the well known Bayesian statistic and the newest entropy regularized optimal transport.

Glossary

$\mathbb{1}_N$	Vector of $\mathbb{R}^N$ with all entries equal to one
$\|\cdot\|_{H^k(\mathbb{R}^d)}$	Sobolev norm of order k associated to the $\mathbb{L}^2(\mathbb{R}^d)$ space
$\|\cdot\|$	Usual Euclidean norm in $\mathbb{R}^d$
a.c.	Absolutely continuous with respect to Lebesgue measure
$\mathbb{E}(\boldsymbol{X})$	Expectation of a random variable $\boldsymbol{X}$
$h(U)$	Negative entropy of a matrix $U \in \mathbb{R}^{N\times N}, h(U) = -\sum_{i,j} U_{ij} \log U_{ij}$
iid	Independant and absolutely continuous
$\mathcal{L}$	Represent the law of a random variable
$\mathbb{L}_p(\mathbb{R}^d), p \in [1,\infty)$	Space of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\|f\|^p$ is Lebesgue integrable, and such that all functions that are equal dx-almost everywhere are identified
$\mathbb{L}_p(\mu), p \in [1,\infty)$	Space of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\|f\|^p$ is μ-integrable, and such that all functions that are equal μ-almost everywhere are identified
μ^γ	Penalized population Wasserstein barycenter, Definition I.13
$\hat{\boldsymbol{\mu}}^\gamma_{n,p}$	Penalized empirical Wasserstein barycenter, Definition I.13
Ω_N	Finite space $\Omega_N = \{x_1, \dots, x_N\} \in \Omega^N$
$\mathbb{P}$	Probability measure
pdf	Probability density function
$\Pi(\mu,\nu)$	Set of product measures on $supp(\mu) \times supp(\nu)$ with respective marginals μ and ν
$\mathcal{P}_p(\mathbb{R}^d)$	Set of Borel probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ supported on $\mathbb{R}^d$, admitting a moment of order p (see Introduction)
$\mathcal{P}_p^{ac}(\mathbb{R}^d)$	Set of measures in $\mathcal{P}_p(\mathbb{R}^d)$ that are absolutely with respect to Lebesgue measure on $\mathbb{R}^d$ (see Introduction)
r^ε	Entropy regularized population Wasserstein barycenter (or Sinkhorn population barycenter), Definition I.21

$\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$	Entropy regularized empirical Wasserstein barycenter (or Sinkhorn empirical barycenter), Definition I.21
Σ_N	Simplex $\Sigma_N = \{r \in \mathbb{R}_+^N \text{ such that } \sum_{i=1}^N r_i = 1\}$
Σ_N^ρ	Bounded simplex $\Sigma_N^\rho = \{r \in \Sigma_N \ : \ \min_{1 \leqslant \ell \leqslant N} r_\ell \geqslant \rho\}$
$supp(\mu)$	Support of the measure μ
$U(a, b)$	Set of transport matrices with marginals $a \in \Sigma_N$ and $b \in \Sigma_N$, $U(a,b) = \{T \in \mathbb{R}_+^{N \times N} \ \vert T\mathbb{1}_N = a, T^T\mathbb{1}_N = b\}$
$V_{\mu_r}(\Omega)$	Closed and convex set of functions $V_{\mu_r}(\Omega) := \log_{\mu_r}(\mathcal{P}_2(\Omega)) \subset \mathbb{L}^2_{\mu_r}(\Omega)$, for a reference probability measure μ_r, Definition IV.2
$W_p, p \in [1, \infty)$	p-Wasserstein distance, Definition A.1
$W_{p,\varepsilon}, p \in [1, \infty)$	Entropy regularized Wasserstein distance (or Sinkhorn divergence), Definition A.3

Bibliography

[AB06] C. D. Aliprantis and K. Border. *Infinite dimensional analysis: a Hitchhiker's guide*. Springer Science & Business Media, 2006.

[ABS13] H. Attouch, J. Bolte, and B. Svaiter. Convergence of descent methods for semi-algebraic and tame problems. *Mathematical Programming*, 137(1-2), 2013.

[AC11] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[AC17] M. Agueh and G. Carlier. Vers un théorème de la limite centrale dans l'espace de Wasserstein? *Comptes Rendus Mathématique*, 355(7):812–818, 2017.

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 214–223, 2017.

[ÁEdBCAM15a] P. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Wide consensus for parallelized inference. *ArXiv e-prints*, 1511.05350, 2015.

[ÁEdBCAM15b] P.C. Álvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A note on the computation of Wasserstein barycenters. *ArXiv e-prints*, 1511.05355, 2015.

[ÁEdBCAM16] P.C. Álvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[AG13] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[AGS04] L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei (9), Matematica e Applicazioni*, 15(3-4), 2004.

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[AHT94] N.H. Anderson, P. Hall, and D.M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1), 1994.

[AN00] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, USA, 2000.

[BA97] A. W. Bowman and Adelchi Azzalini. *Applied smoothing techniques for data analysis : the kernel approach with S-Plus illustrations*. Clarendon Press ; Oxford University Press, 1997.

[BCC+15] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[BCP17] J. Bigot, E. Cazelles, and N. Papadakis. Central limit theorems for sinkhorn divergence between probability distributions on finite spaces and statistical applications. *ArXiv e-prints*, 1711.08947, 2017.

[BCP18a] J. Bigot, E. Cazelles, and N. Papadakis. Data-driven regularization of wasserstein barycenters with an application to multivariate density registration. *ArXiv e-prints*, 1804.08962, 2018.

[BCP18b] J. Bigot, E. Cazelles, and N. Papadakis. Penalized barycenters in the Wasserstein space. *ArXiv e-prints*, 1606.01025v2, 2018.

[Bec11] S. Becker. Matlab wrapper and C implementation of L-BFGS-B-C, 2011. https://github.com/stephenbeckr/L-BFGS-B-C.

[BFS12] M. Burger, M. Franek, and C.-B. Schönlieb. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 2012(2):209–253, 2012.

[BGKL17] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic pca in the Wasserstein space by convex PCA. *Annales de l'Institut H. Poincaré, Probabilités et Statistiques*, 53(1), 2017.

[BGKL18] J. Bigot, R. Gouet, T. Klein, and A. Lopez. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line,. *Electronic Journal of Statistics*, To be published, 2018.

[BL14] S. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics and Kantorovich transport distances*. 2014.

[BM06] S. Boyd and A. Mutapcic. Subgradient methods. *Lecture notes of EE364, Stanford University*, 2007, 2006.

[BM07] F. Bauer and A. Munk. Optimal regularization for ill-posed problems in metric spaces. *Journal of Inverse and Ill-posed Problems*, 15(2):137–148, 2007.

[BO04] M. Burger and S. Osher. Convergence rates of convex variational regularization. *Inverse problems*, 20(5), 2004.

[Bor14] J.M. Borwein. A very complicated proof of the minimax theorem. *Minimax Theory and its Applications*, 2014.

[BPC16] N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

[Bra06] A. Braides. A handbook of γ-convergence. *Handbook of Differential Equations: stationary partial differential equations*, 3:101–213, 2006.

[Bre91] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4), 1991.

[BRPP15] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Bur16] M. Burger. *Bregman distances in inverse problems and partial differential equations*, pages 3–33. Springer International Publishing, Cham, 2016.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Caf92] L. A Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[Caf96] L. A Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496, 1996.

[CD14] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning 2014, PMLR W&CP*, volume 32, pages 685–693, 2014.

[CD15] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

[CDPS17] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Analysis*, 49(2):1385–1418, 2017.

[Cla13] F. Clarke. *Functional analysis, calculus of variations and optimal control*, volume 264. Springer Science & Business Media, 2013.

[COO15] G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, 2015.

[CP16a] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2), 2016.

[CP16b] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[CPSV18a] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[CPSV18b] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamics and kantorovich formulations. *Journal of Functional Analysis*, 2018.

[CSB+18] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis. Geodesic pca versus log-pca of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456, 2018.

[Cut13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[CZ97] Y. Censor and S. A. Zenios. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press on Demand, 1997.

[DBCAM+00] E. Del Barrio, J.A. Cuesta-Albertos, C. Matrán, S. Csörgö, C.M. Cuadras, T. de Wet, E. Giné, R. Lockhart, A. Munk, and W. Stute. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, 9(1):1–96, 2000.

[DBCAMRR99] E. Del Barrio, J.A. Cuesta-Albertos, C. Matrán, and J.M. Rodriguez-Rodriguez. Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics*, 27(4), 1999.

[DBGU05] E. Del Barrio, E. Giné, and F. Utzet. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1), 2005.

[DBL17] E. Del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. *ArXiv e-prints*, 1705.01299, 2017.

[DE97] P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1997.

[DPF14] G. De Philippis and A. Figalli. The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

[DPR16] A. Dessein, N. Papadakis, and J.-L. Rouas. Regularized Optimal Transport and the Rot Mover's Distance. *ArXiv e-prints*, 1610.06447, 2016.

[EG15] L.C. Evans and R.F. Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.

[ET76] I. Ekeland and R. Temam. *Convex analysis and 9 variational problems*. SIAM, 1976.

[ET93] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[Eud96] T. L. Eudey. Statistical considerations in DNA flow cytometry. *Statistical Science*, pages 320–334, 1996.

[FG10] A. Figalli and N. Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

[FG15] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[FGM09] J. Fontbona, H. Guérin, and S. Méléard. Measurability of optimal transportation and convergence rate for Landau type interacting particle systems. *Probability Theory and Related Fields*, 143(3-4):329–351, 2009.

[FGM10] J. Fontbona, H. Guérin, and S. Méléard. Measurability of optimal transportation and strong coupling of martingale measures. *Electronic Communications in Probability*, 15:124–133, 2010.

[FHN+18] A. Forrow, J-C. Hütter, M. Nitzan, G. Schiebinger, P. Rigollet, and J. Weed. Statistical optimal transport via geodesic hubs. *ArXiv e-prints*, 1806.07348, 2018.

[Fle11] T. Fletcher. Geodesic regression on Riemannian manifolds. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability*, 2011.

[Fle13] T. Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International journal of computer vision*, 105(2), 2013.

[FLPJ04] T. Fletcher, C. Lu, Stephen M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8), 2004.

[FM05] G. Freitag and A. Munk. On Hadamard differentiability in k-sample semiparametric models with applications to the assessment of structural relationships. *Journal of Multivariate Analysis*, 94(1), 2005.

[FPPA14] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[Fré48] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut H.Poincaré, Sect. B, Probabilités et Statistiques*, 10:235–310, 1948.

[FS14] Z. Fang and A. Santos. Inference on directionally differentiable functions. *ArXiv e-prints*, 1404.3763, 2014.

[FZM+15] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T.A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, 2015.

[GCB+04] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, Sep 2004.

[GCB16] A. Genevay, G. Cuturi, M. andPeyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Proc. NIPS'16*. Curran Associates, Inc., 2016.

[Ger16] D. Gervini. Independent component models for replicated point processes. *Spatial Statistics*, 18:474 – 488, 2016.

[GPC15] A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*. Springer, 2015.

[GPC17] A. Genevay, G. Peyré, and M Cuturi. Sinkhorn-autodiff: Tractable Wasserstein learning of generative models. Technical report, 2017.

[HAG+17] B.P. Hejblum, C. Alkhassim, R. Gottardo, F. Caron, and R. Thiébaut. Sequential dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. *ArXiv e-prints*, 1702.04407, 2017.

[HKB+10] F. Hahne, A.H. Khodabakhshi, A. Bashashati, C.-J. Wong, R.D. Gascoyne, A.P. Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, R. Gentleman, and R.R. Brinkman. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, 77(2):121–131, 2010.

[JLG12] X. Jiang, Z.-Q. Luo, and T.T. Georgiou. Geometric methods for spectral analysis. *IEEE Transactions on Signal Processing*, 60(3), 2012.

[KP08] S.G. Krantz and H.R. Parks. *Geometric integration theory*. Springer Science & Business Media, 2008.

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

[KT59] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[KU01] A. Kneip and K.J. Utikal. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96(454):519–542, 2001.

[LeC98] Y. LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[LGL16] T. Le Gouic and J.-M. Loubes. Existence and Consistency of Wasserstein Barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2016.

[LM16] C. Lacour and P. Massart. Minimal penalty for goldenshluger–lepski method. *Stochastic Processes and their Applications*, 126(12):3774–3789, 2016.

[LMP16] S.X. Lee, G.J. McLachlan, and S. Pyne. Modeling of inter-sample variation in flow cytometric data with the joint clustering and matching procedure. *Cytometry Part A*, 89(1):30–43, 2016.

[LMS18] M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3), 2018.

[LP15] D. A. Lorenz and T. Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2), 2015.

[MC98] A. Munk and C. Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):223–241, 1998.

[Ngu13] X. L. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[OCBP14] P. Ochs, Y. Chen, T. Brox, and T. Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2), 2014.

[Pas13] B. Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264(4):947–963, 2013.

[PC17] G. Peyré and M. Cuturi. Computational optimal transport. Technical report, 2017.

[PFR12] G. Peyré, J. Fadili, and J. Rabin. Wasserstein active contours. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2012.

[PLW+14] S. Pyne, S.X. Lee, K. Wang, J. Irish, P. Tamayo, M.-D. Nazaire, T. Duong, S.-K. Ng, D. Hafler, R. Levy, G.P. Nolan, J. Mesirov, and G.J. McLachlan. Joint modeling and

registration of cell populations in cohorts of high-dimensional flow cytometric data. *PloS one*, 9(7), 2014.

[PM16a] A. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a hilbert space. *The Annals of Statistics*, 44(1), 2016.

[PM+16b] A. Petersen, H.-G. Müller, et al. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

[PZ16] V. M. Panaretos and Y. Zemel. Amplitude and phase variation of point processes. *Annals of Statistics*, 44(2):771–812, 2016.

[PZ17] V. M. Panaretos and Y. Zemel. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, to be published, 2017.

[PZ18] V.M. Panaretos and Y. Zemel. Statistical aspects of wasserstein distances. *ArXiv e-prints*, 1806.05500, 2018.

[RBVFT18] G. Rios, J. Backhoff-Veraguas, J. Fontbona, and F. Tobar. Bayesian learning with wasserstein barycenters. *ArXiv e-prints*, 1805.10833, 2018.

[RCP16] A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[Rei16] M. Reid. Meet the Bregman divergences, 2016.

[RMS16] T. Rippl, A. Munk, and A. Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151, 2016.

[Roc74] R.T. Rockafellar. *Conjugate duality and optimization*. Siam, volume 16, 1974.

[Röm05] W. Römisch. Delta method, infinite dimensional. *Encyclopedia of statistical sciences*, 2005.

[RP15] J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2015.

[RTC17] A. Ramdas, N.G. Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[RV74] A.W. Roberts and D.E. Varberg. Another proof that convex functions are locally Lipschitz. *The American Mathematical Monthly*, 81(9):1014–1016, 1974.

[San10] F. Santambrogio. Introduction to optimal transport theory. *ArXiv e-prints*, 1009.3856, 2010.

[San15] F. Santambrogio. *Optimal Transport for Applied Mathematicians - Calculus of Variations, PDEs, and Modeling*. Springer Verlag Italia, 2015.

[SC15] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015.

[Sha90] A. Shapiro. On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3), 1990.

[SHB+18a] M.A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[SHB+18b] M.A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[SLHN10] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In *Computer Vision ECCV 2010*, volume 6316 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010.

[SM16] M. Sommerfeld and A. Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

[SSSSS09] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[Str96] T. Strömberg. The operation of infimal convolution. 1996.

[SW15] V. Spokoiny and N. Willrich. Bootstrap tuning in ordered model selection. *ArXiv e-prints*, 1507.05034, 2015.

[Sza97] S. J. Szarek. Metric entropy of homogeneous spaces. *ArXiv e-prints*, math/9701213, 1997.

[TCDP17] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis. Overrelaxed Sinkhorn-Knopp algorithm for regularized optimal transport. Technical report, 2017.

[VDVW96] A.W. Van Der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[VIB15] R. Verde, A. Irpino, and A. Balzanella. Dimension reduction techniques for distributional symbolic data. *IEEE Transactions on Cybernetics*, 2(46), 2015.

[Vil03] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[Vil08] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[Was11] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2011.

[Wil69] A.G. Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, 1969.

[WS11] W. Wu and A. Srivastava. An information-geometric framework for statistical inferences in the neural spike train space. *Journal of Computational Neuroscience*, 31(3):725–748, 2011.

[WSB+13] W. Wang, D. Slepcev, S. Basu, J.A. Ozolek, and G.K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2), 2013.

[WZ94] C.Y. Wang and F. Zhao. Directional derivatives of optimal value functions in mathematical programming. *Journal of optimization theory and applications*, 82(2), 1994.

[YWWL17] J. Ye, P. Wu, J.Z. Wang, and J. Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Trans. Signal Processing*, 65(9), 2017.

[Zal02] C. Zalinescu. *Convex analysis in general vector spaces*. World Scientific, 2002.

[ZBLN97] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, 1997.

[ZM11] Z. Zhang and H.-G. Müller. Functional density synchronization. *Computational Statistics & Data Analysis*, 55(7):2234–2249, 2011.